HHS Public Access

Author manuscript

Brain Comput Interfaces (Abingdon). Author manuscript; available in PMC 2016 March 03.

Published in final edited form as:

Brain Comput Interfaces (Abingdon). 2015 ; 2(4): 161–173. doi:10.1080/2326263X.2015.1063363.

Identifying the Attended Speaker Using Electrocorticographic (ECoG) Signals

K. Dijkstra[a,b,c], **P. Brunner**[a,b], **A. Gunduz**[a,d], **W. Coon**[a,e], **A.L. Ritaccio**[b], **J. Farquhar**[c], and **G. Schalk**[a,b,e,*]

[a]Ctr for Adapt Neurotech, Wadsworth Center, New York State Department of Health, Albany, NY [b]Dept of Neurology, Albany Medical College, Albany, NY [c]Donders Inst for Brain, Cognition and Behaviour, Radboud Univ Nijmegen, The Netherlands [d]J. Crayton Pruitt Family Dept of Biomed Eng, Univ of Florida, Gainesville, FL [e]Dept of Biomed Sci, State Univ of New York at Albany, Albany, NY

Abstract

People affected by severe neuro-degenerative diseases (e.g., late-stage amyotrophic lateral sclerosis (ALS) or locked-in syndrome) eventually lose all muscular control. Thus, they cannot use traditional assistive communication devices that depend on muscle control, or brain-computer interfaces (BCIs) that depend on the ability to control gaze. While auditory and tactile BCIs can provide communication to such individuals, their use typically entails an artificial mapping between the stimulus and the communication intent. This makes these BCIs difficult to learn and use.

In this study, we investigated the use of selective auditory attention to natural speech as an avenue for BCI communication. In this approach, the user communicates by directing his/her attention to one of two simultaneously presented speakers. We used electrocorticographic (ECoG) signals in the gamma band (70–170 Hz) to infer the identity of attended speaker, thereby removing the need to learn such an artificial mapping.

Our results from twelve human subjects show that a single cortical location over superior temporal gyrus or pre-motor cortex is typically sufficient to identify the attended speaker within 10 s and with 77% accuracy (50% accuracy due to chance). These results lay the groundwork for future studies that may determine the real-time performance of BCIs based on selective auditory attention to speech.

Keywords

Brain-Computer Interface (BCI); Electrocorticography (ECoG); Auditory Attention; Cocktail Party

*Corresponding author. gerwin.schalk@health.ny.gov.

1. Introduction

Communication is an essential part of being human. It allows us to interact with each other, to establish relationships, and to express needs and desires. This fundamental human ability can become compromised in people affected by paralysis, as they are no longer able to control the muscles that allow us to gesture or speak. Conventional assistive devices (e.g., eye trackers or tongue/cheek switches) re-establish communication, but generally rely on some residual muscle control. In contrast, brain-computer interfaces (BCIs) re-establish communication by using brain signals, effectively circumventing muscular pathways [1]. However, BCIs still depend on perceptual modalities, such as auditory, tactile or, most frequently, visual perception, for stimulation or feedback.

In one common BCI approach called the P300 matrix speller, the user communicates by directing attention to one of many visual stimuli. These systems are interesting in part because they preserve the identity between the stimulus (e.g., a highlighted ‘A’) and the symbol the user wants to communicate (e.g., the letter ‘A’). Unfortunately, recent studies have shown that the communication performance of the matrix speller depends considerably on the ability to control eye gaze [2, 3], which is lost or diminished in people affected by severe neuro-degenerative diseases (e.g., late-stage amyotrophic lateral sclerosis (ALS) or locked-in syndrome).

This important issue has led to an increased interest in BCI paradigms that use non-visual sensory modalities, such as auditory [4–8] or tactile stimulation ([9, 10]; see [11] for review). In these paradigms, the user selectively attends to one of multiple auditory or tactile stimuli. Each attended stimulus elicits event-related potentials (ERPs) that are different from those elicited by unattended stimuli. This difference in evoked responses allows the BCI system to infer the attended stimulus. Nevertheless, to use this effect for BCI communication, the user still has to learn a relatively artificial mapping between a stimulus (e.g., a particular but arbitrary sound) and a communication output (e.g., a particular but arbitrary letter or word). This mapping is simple when there are only few possible outputs (e.g., a yes or no command). Unfortunately, when the number of possible outputs is larger, such as with a spelling device, this mapping is not only arbitrary but also complex. Thus, such BCI systems are cumbersome to learn and use. One way to address this issue is to exploit the natural human ability to selectively attend to one of several speakers in a ‘cocktail party’ environment. Using this approach, a BCI could learn which speaker a subject is attending to, or which speech stimulus represents the intention of the subject, simply by the subject paying attention to one of several speech stimuli.

There are two avenues to identify the attended speech stimulus. In the first avenue, speech stimuli are designed (e.g., altered and broken up [12]) such that they elicit a particular and discriminable evoked response to identify the attended speech stimulus. However, such altered speech stimuli are difficult to understand, which makes such a BCI system difficult to use. More importantly, this approach does not scale well beyond two simultaneously presented speech stimuli. In the second avenue, the BCI detects the specific details of the attended speech stimulus (e.g., its spectrotemporal structure) in the neural response. The spectrotemporal structure of speech is of particular utility in this context, because it results

from the combination of linguistic elements at different levels (e.g., phonemes, syllables, words and phrases) that give speech a variation in sound intensity over time. The integrity of this amplitude variation (or envelope) is necessary to understand speech [13].

While evoked responses can readily be detected in scalp-recorded electroencephalography (EEG), electrocorticography (ECoG) is better suited to capture the detailed spectrotemporal structure of neurological processes relating to speech. Of particulate note, it provides ready access to signals in the broadband gamma (70–170 Hz) range, which have been shown to be a reliable indicator of local cortical population activity [14] and which cannot readily be detected in scalp-recorded EEG [15]. In fact, recent studies using subdurally recorded electrocorticography (ECoG) have shown that indeed, ECoG signals in the high gamma band and at specific cortical locations (e.g., superior temporal gyrus, STG) track the envelope of perceived speech [16–20].

These findings have recently been extended to simultaneously presented streams of speech, i.e., a cocktail party situation in ECoG [21, 22] and EEG [23]. These studies found that the neural tracking of speech envelope is more pronounced for the attended stimulus than for the unattended stimulus. This finding has led to an emerging interest in determining the usefulness of this effect in the BCI context [24].

In the present study, we set out to fully characterize the ECoG correlates of attended/unattended speech and to determine the potential communication performance of a cocktail party-based BCI. To do this, we recorded ECoG signals from twelve human subjects while they selectively attended to one of two simultaneously presented speech stimuli. Our results show that ECoG responses from a single cortical location over STG or pre-motor cortex is typically sufficient to identify attended speech within 5 s of selective auditory attention with an accuracy of at least 70%. By using multiple cortical locations, this performance can be improved to at least 81%. With additional real-time validation of this approach, our work could lay the basis for a BCI that would allow people to communicate their intent simply by attending to different simultaneously presented auditory stimuli.

2. Methods

2.1. Subjects

We recruited twelve human subjects who underwent temporary placement of subdural electrodes as part of their clinical treatment for epilepsy. This clinical treatment included the localization of epileptogenic zones and their delineation from functional cortical areas.

The subjects had 57–133 subdural electrodes implanted over their left or right hemisphere. Cortical coverage varied across subjects (Figure 1) and included frontal, temporal, parietal and occipital cortical areas. Electrodes consisted of platinum-iridium discs (4 mm in diameter, 2.3 mm exposed), embedded in silicon and spaced 6–10 mm apart (Ad-Tech Medical Instrument Corp., Racine, WI). We used post-operative radiographs (anterior-posterior and lateral) and computed tomography (CT) scans to verify the cortical location of the electrodes. We then used Curry software (Neuroscan Inc, El Paso, TX) to create subject-specific 3D cortical brain models from high-resolution pre-operative magnetic resonance

imaging (MRI) scans. We co-registered the MRIs by means of the post-operative CT and extracted the electrode coordinates according to the Talairach Atlas [25]. These electrode coordinates are depicted on Talairach template brains in Figure 1.

The subdural electrodes were implanted for a duration of 5–7 days. During this period, subjects volunteered to participate in our study. Both grid placement and duration of clinical monitoring were based solely on the requirements of the clinical evaluation. The twelve subjects (7 males, 5 females) were 15–60 years old (median 45), with an IQ higher than 75 (median 95). None of the subjects had a history of hearing impairment. IQ and handedness were assessed in a neuropsychological evaluation [26] and language dominance was determined through a pre-operative Wada test [27]. The results of this test and additional subject information are summarized in Table 1. All subjects provided informed consent, and the study was approved by the Institutional Review Board of Albany Medical College.

2.2. Data collection

We recorded ECoG signals from the implanted electrodes using g.USBamp or g.HIamp (g.tec, Graz, Austria) amplifier/digitizer systems, which sampled the data at 1200 Hz. Control of data acquisition and stimulus presentation were accomplished using the BCI2000 software platform [28–30]. Clinical monitoring occurred simultaneously by using a connector that split the cables coming from the patient into one set that was connected to the clinical monitoring system and another set that was connected to the amplifiers. This ensured that clinical data collection was not compromised at any time. Two electrocorticographically silent electrodes (i.e., locations that were not identified as eloquent cortex by electrocortical stimulation mapping) served as electrical ground and reference, respectively.

2.3. Stimuli and task

The subjects' task was to selectively attend to one of two simultaneously presented speakers (see diagram in Figure 2A). The two speakers were John F. Kennedy and Barack Obama, each delivering his inauguration address. Thus, both speeches featured similar linguistic features, but were uncorrelated in their sound intensities ($r = -0.02$, $p = 0.9$). To simulate a cocktail party situation, we mixed the two (monaural) speeches into a binaural presentation. This allowed us to manipulate the aural location of each speaker. The speech stream that was presented in each ear contained 20% : 80% of the volume of one speaker and 80% : 20% of the other, respectively. We broke these combined streams into segments of 15–25 s in length, which resulted in a total of 10 segments of 187 s combined length.

Throughout the experiment, we presented each segment four times through in-ear monitoring earphones. Over these four presentations, we permuted the aural location (i.e., left and right) and the identity (i.e., JFK and Obama) of the attended speaker. In other words, over these four trials, the subjects had to attend to each of the two speakers at each of the two aural locations.

Each trial began with an auditory cue that indicated the ear to which the subject should attend. For the purposes of this study, we complemented this auditory cue with a visual cue that indicated the identity and aural location (e.g., 'JFK in LEFT ear'). The visual cue

remained on the screen throughout the trial. Each trial consisted of a 4 s cue and a 15–25 s stimulus, and was followed by a 5 s inter-stimulus period. This resulted in a total of 40 trials (i.e., 10 segments, each presented 4 times) of 12.5 min total length that were presented in a counter-balanced order. These 40 trials were divided into 5 blocks of 8 trials each with a 3 min break between each block.

2.4. Features

The data consisted of the ECoG signals and the corresponding attended and unattended speech streams as shown in Figure 2B. From these data, we extracted the high gamma band envelope at each cortical location and the envelopes of the covertly attended and unattended speech (i.e., JFK and Obama). From these signal envelopes, we extracted two sets of features that reflected the neural tracking of the attended or unattended speech, respectively. For this, we correlated the high gamma band envelope at each cortical location, once with the attended and once with the unattended speech envelope. This resulted in one Spearman's r-value for each feature set, each cortical location and each trial. An example of this is shown in Figure 2C.

2.4.1. Signal processing—We first pre-processed the ECoG signals from the 58–133 channels to remove noise and common mode activity. To do this, we high-pass filtered the signals at 0.5 Hz and re-referenced them to a common average reference that we composed from only those 58–133 channels for which the 60 Hz line noise was within 1.5 standard deviations of the average. Finally, we used a notch filter to remove any remaining 60 Hz line noise.

We then extracted the signal envelope in the high gamma band using these pre-processed ECoG signals. To do this, we applied a 70–170 Hz Butterworth filter and then extracted the envelope of the filtered signals using the Hilbert transform. As low frequency components dominate the ECoG signal [31], the utilized Butterworth filter featured a high attenuation in the stopbands (i.e., 18th order, 30dB attenuation below 64 Hz and above 200 Hz). We low-pass filtered the resulting signal envelope at 6 Hz for anti-aliasing while retaining the temporal information at the syllabic level [32]. Finally, to reduce the computational effort, we decimated the sampling rate of the signal by a factor of 10, to 120 Hz.

For each auditory stimulus, we extracted the time course of the sound intensity, i.e., the envelope of the signal waveform in the speech band (80–6000 Hz). To do this, we applied a 80–6000 Hz Butterworth filter to each audio signal, and then extracted the envelope of the filtered signals using the Hilbert transform. The characteristics of the Butterworth filter were chosen to ensure that high amplitude low frequency components were sufficiently removed (i.e., 10th order, 30dB attenuation below 40 Hz and above 8000 Hz). Finally, we low-pass filtered the speech envelopes at 6 Hz and downsampled them to 120 Hz.

2.4.2. Feature extraction—For each of the three signals (i.e., ECoG gamma envelope and right/left speech envelope) and each trial, we extracted features that reflected the neural tracking of the attended or unattended speech, respectively. We defined neural tracking of speech as the correlation between the gamma envelope (of a given cortical location) and the speech envelope. We calculated this correlation separately for the attended and unattended

speech, thereby obtaining two sets of r-values labeled ‘attended’ and ‘unattended,’ respectively.

We expected there to be a delay between the audio presentation and resulting cortical processing, i.e., the time from presentation of the audio stimuli to the observation of the cortical change. To account for this delay, we measured the neural tracking of the sound intensity across different delays (0 to 250 ms, see Figure 3).

On the basis of these results, we selected a delay of 100 ms across all subjects. We corrected for this delay by shifting the speech envelopes relative to the ECoG envelopes prior to calculating the correlation values.

2.5. Classification

We quantified the extent to which we could identify the attended speech from the extracted features in single trials. To do this, we applied two different classification methods on the extracted features (i.e., the correlation values). The classifier's goal was to predict the identity of the attended stimulus from the neural features. The first method used features from a single electrode in a univariate classifier; the second method combined features from multiple electrodes in a multivariate regularized logistic classifier. We evaluated the classification accuracy on 100 ms to 10 s long trial segments using 10 iterations of a 10-fold cross-validation. Finally, we determined the significance of these results using a permutation test.

For the univariate classification method, we assumed that in specific single locations the neural tracking of attended speech would be more pronounced than that of unattended speech. In other words, we assumed that the sign of the difference between two features (i.e., the attended and unattended r-values) directly predicted the attended stimulus. We applied this classification to all 58-133 cortical locations. In our cross-validation procedure, we selected the cortical location for which this assumption was most consistent and evaluated its predictive performance.

With the multivariate classifier, we used an elastic net regularization to select a linear combination of features from multiple cortical locations to predict the attended stimulus. In a nested cross-validation procedure, we determined the parameters of the elastic net feature selection (i.e., shrinkage $\lambda = 0.01–0.35$, trade-off between lasso and ridge $\alpha = 0.5$). Our main cross-validation procedure then used a logistic regression combined with an the elastic net regularization to determine this linear combination (i.e., the 1–76 selected features), and to quantify the classifier's predictive performance.

For both classification methods, we used a permutation test to determine the significance of the classification accuracy and thus to statistically validate our results. In this permutation test, we reversed the speech envelope to remove the temporal relationship between the speech and neural envelopes while keeping their autocorrelation intact. To determine the distribution of random performance, we repeated this analysis 100 times on data for which we shifted the reversed speech envelope by random amounts of time. We then applied the feature extraction and classification steps explained above on this permuted data. This

resulted in a distribution of classification values for each of the two methods. Finally, we determined the likelihood (i.e., the p-value) that our cross-validated performance was different from this random performance distribution.

2.5.1. Performance evaluation—We determined the potential communication performance of a BCI based on selective auditory attention to speech. To do this, we calculated the information transferred in each trial as a function of the classification accuracy (P = 0%–100%) and number of simultaneous choices (N = 2) [33]. We then calculated the information transfer rate in bits/min as a function of the trial length.

$$IT = log_2 N + P\, log_2 P + (1 - P)\, log_2 \frac{1 - P}{N - 1}$$

2.5.2. Dependence on length of data—We were interested in determining how classification accuracy and information transfer rate depended on the length of the data segments. To address this question, we applied the two classification methods to data segments of different length (100 ms to 10 s). We extracted these data segments from each trial beginning at 2 s into the trial. For each segmentation length, we performed 10 iterations of a 10-fold cross-validation. This resulted in 100 cross-validated classification accuracies for each subject and segmentation length.

We used a permutation test, described in the previous section, to determine the significance of this classification accuracy on 5 s long segments.

2.6. Temporal evolution

We were also interested in determining how the difference between the ‘attended’ and ‘unattended’ correlation, and thus classification performance, develops during the trial. To address this question, we extracted 1 s long segments in increments of 100 ms starting from 1 s before the onset of each trial (i.e., [−1000–0 ms, −900–100 ms, . . . , 9000–10000 ms]). For each of these segments, we extracted the correlation with the ‘attended’ and ‘unattended’ speech and then applied the univariate classifier as described previously.

3. Results

3.1. Neural correlates of attended and unattended speech

The results in Figure 4 show the neural tracking of the attended and unattended speech in the form of an activation index. For each cortical location, and each subject (1–12), this activation index expresses the negative logarithm of the p-value (−log(p)) of the correlation between the high gamma ECoG envelope and the attended (●, top) or unattended (○, bottom) speech envelope. The neural tracking is focused predominantly on areas on or around STG (all subjects), but also on discrete areas in superior pre-motor cortex (subjects 2, 3, 4 and 8).

3.2. Identification of attended speech

We then determined the extent to which we can identify the attended speech stimulus from the ECoG signals in individual trials using 5 s of data. The results are shown in Figure 5A and give the single-trial accuracy for each subject and for the two investigated classification methods (univariate (blue) and multivariate (orange) regression). The subjects are presented in the order of their average classification accuracy; an asterisk indicates significance (determined in the permutation test, adjusted for multiple comparisons by using a false discovery rate ([34], with a 5% probability that a false discovery is accepted). Figure 5B shows the average classification accuracy across subjects for which statistical significance was obtained for at least one method (subjects 1–7, from here on referred to as 'significant subjects'). A comparison between the two methods shows that on average, multivariate regression results in 11% higher classification accuracy compared to univariate regression (81% vs. 70%, paired t-test: $p < 0.0003$).

To further expand on the results from section 3.1, we averaged the activation index topographies across significant and non-significant subjects. These topographies are shown in Figure 6. In these results, the significant subjects (Figure 6A) show a stronger and more distributed response to the attended (●) than to the unattended speech (○). In contrast, non-significant subjects (Figure 6B) show only a marginal difference in their response to the attended (●) compared to the unattended speech (○).

The results in Figure 7 show the neural tracking measured as the correlation between the ECoG envelope (at center frequencies ranging from 2.5 to 250 Hz in steps of 5 Hz) and the attended or unattended speech envelope (orange or blue traces, respectively), averaged across significant and non-significant subjects. In these results, the significant subjects (Figure 7A) show neural tracking of the attended speech that is stronger across all frequency bands, especially in the broadband gamma band (70–170 Hz). The tracking shows a negative relationship in the low frequency band (10–30 Hz). For the non-significant subjects (Figure 7B), this negative relationship at low frequencies is not apparent and the tracking of attended and unattended speech at higher frequencies is at the same low level.

3.3. Relationship between segment length and classification accuracy

In the previous section, we determined the classification accuracy on 5 s long data segments. In this section, we examine the relationship between the segment length and classification accuracy. The results in Figure 8 show the classification accuracy for variable segments lengths (0.1 to 10 s) for all significant subjects (Figure 8A) and non-significant subjects (Figure 8B). For the significant subjects (Figure 8A), accuracy rises steadily and reaches 86.5% at 10 s. Throughout the investigated segment length, the ~10% advantage of the multivariate over the univariate classification method persists. In contrast, classification for the non-significant subjects (Figure 8B) stays around chance level for both classification methods.

Next, we were interested in determining the effect that this difference in accuracy between univariate and multivariate classifiers has on the information throughput of a BCI. To do this, we calculated the information transfer rate (ITR) for our significant subject group,

omitting any inter-trial period and not accounting for the 2 s tuning-in period that was excluded from the classification data. The results in Supplementary Figure S1 show that for the multivariate classifier, ITR reaches 6.2 bits/min for 1.5 s long segments. In contrast, the univariate classifier achieves 2.6 bits/min for 4 s long segments.

3.4. Effect of 'tuning in' on correlation and classification accuracy

In the previous analyses, we focused on determining the classification accuracy that we achieved given trial segments of variable length without considering the start point of the trial relative to the task. Indeed, we explicitly omitted any potential 'tuning-in' period by removing the first 2 s of each trial. With the following analyses, we determined the effect that tuning in to the attended speech has on the correlation and the classification accuracy over the course of the trial.

The results in Figure 9 show the difference between the 'attended' and 'unattended' correlation (9A), as well as the resulting univariate classification accuracy (9B) averaged across subjects 1–5.

These results show that it takes ~1 s for a difference between the 'attended' and 'unattended' correlation to develop. Consequently, for the first second, the classification accuracy remains around chance level (i.e., 50%). From there on, two effects are visible. First, the correlation and classification accuracy show an upward trend. Second, this trend is superimposed with cycles of higher and lower correlation and classification accuracy.

4. Discussion

This study shows that it is possible to identify the speaker that a subject selectively attends to when he/she is presented with two simultaneously presented speeches. The identification accuracy depends on several factors. First, we found that only a subset of our subjects (7/12) showed a difference in neural tracking that allowed us to infer the attended speech with statistical significance. Second, we found that within this group, generally a single electrode is sufficient to identify attended speech within 5 s of selective auditory attention and with 70% accuracy. Third, using multiple electrodes in a multivariate approach, this performance can be improved to an average of 81%. Fourth, this 11% increase in accuracy results in a two-fold increase in the Information Transfer Rate (ITR).

Detailed analysis of the ECoG signals indicates that neural tracking of the attended speech is stronger and more widely distributed than that of unattended speech. This is in line with a previous ECoG study that investigated auditory attention [21]. Furthermore our analyses localized the most informative locations to STG and pre-motor cortex. While pre-motor cortex locations were found to be informative in only two of twelve subjects, they do echo findings of previous studies [17, 18, 35].

4.1. Feasibility as a BCI

The results shown in this paper indicate that the presented method could support BCI communication. While being invasive, it may be justified for those affected by severe neuro-degenerative diseases (e.g., late-stage ALS, locked-in syndrome) who have lost all muscular

control and therefore cannot use conventional assistive devices or BCIs that depend on visual stimulation or feedback.

Additionally, the results suggest that sufficient communication performance (> 70%, [36]) could be achieved with a single electrode placed over STG. This finding is important, because placement of ECoG grids as used in this study requires a large craniotomy. In contrast, a single electrode could be placed through a burr hole [37]. Furthermore, the electrodes in this study were placed subdurally (i.e., the electrodes are placed underneath the dura). Penetration of the dura increases the risk of bacterial infection [38–42]. Epidural electrodes (i.e., electrodes placed on top of the dura) provide signals of approximately comparable fidelity [43, 44]. A single electrode placed epidurally could reduce risk, which should make this approach more clinically practical.

Further advances in clinically practical recordings of ECoG signals from multiple cortical locations [45, 46] could support an even higher communication performance. Our results showed that combining signals from multiple cortical locations results in a two-fold increase in the Information Transfer Rate (ITR).

Beyond BCI communication, the presented auditory attention-based approach may eventually support the detection of consciousness in comatose and minimally conscious patients without them having to learn a complicated task. Currently, communication in such populations is accomplished by using a functional magnetic resonance imaging (fMRI) approach [47]. In this approach, the patient receives instructions that are intended to activate specific brain regions. Over the course of this task, the fMRI BOLD signal measures this activity. Activity changes then suggest whether the patient performs the task, and thus whether he/she performs activities that typically are associated with consciousness. As this approach requires an fMRI scanner, it is not well suited to provide communication on a daily basis. If consent can be obtained and the risk of this procedure can be justified, then the approach presented in this paper could o er a viable option for communication once the level of consciousness has been determined.

4.2. Comparison to other auditory BCIs

The performance obtained in our ECoG study did not reach the communication performance reported in some other studies that investigated the use of attention-related EEG responses to specifically designed auditory stimuli. For example, Hill et al., [48] achieved an online performance of 84.8% for 5 s trials, with an ITR of 4.98 bits/min ± 2.3. In their study, subjects attended to one of two concurrent streams of tones. In our study, we obtained a performance of 81% and an ITR of 4.2 bits/min ± 2.7 for the significant subjects. Results across all of our subjects were lower, with a performance of 70% at 5 s and an ITR of 2.5 bits/min ± 2.9. This may appear surprising, as ECoG is thought to have a better signal fidelity than EEG [15]. This apparent contradiction may be explained by four differences between our study and the other studies. First, the two streams of natural speech used in this study are similar in many respects (e.g., overlapping spectral representations), and thus cannot be expected to result in optimally differing neural responses. Second and similarly, because the stimuli are similar, some of our subjects may simply not have been able to perform the task properly, which likely explains the difference between significant and non-

significant subjects. Third, we were confined to recordings from frontal, parietal, and temporal areas within a single hemisphere. In contrast, in these EEG studies, signals were recorded from electrodes that fully covered both hemispheres. Fourth, our subjects were clinical patients rather than the healthy subjects used in the other studies. In comparison to many other auditory BCIs, the present approach has the unique advantage in that it uses natural speech without any alteration. This aspect may be particularly relevant for those who are already at a stage where learning how to use a BCI has become difficult.

On top of that, our approach may not have fully exploited the communication performance that could be obtained by increasing the number of simultaneously presented communication options. For two reasons, this is easier for the speech stimuli used here than for the streams of relatively artificial stimuli used in other approaches. First, communication based on streams of relatively artificial stimuli uses the response to the stimuli (ERPs) to identify the attended stream. To preserve the identity between stimulus and ERP, these streams need to be without temporal overlap. In contrast, natural speech stimuli can have temporal overlap as long as they remain fairly uncorrelated. In fact, while the natural speech stimuli used in this study were similar in many respects (e.g., overlapping spectral representations), they were fairly uncorrelated (i.e., correlation between sound intensities, $r = -0.07$). Second, the mental effort required to map a stream of relatively artificial stimuli to a communication intent limits the number of simultaneously presented streams. In contrast, speech stimuli can be identical with the communication intent, effectively removing the need for an otherwise needed mapping.

4.3. Comparison to other auditory attention studies

Our results show that using the gamma band to track auditory attention has limitations. First, while in our study the correlation between speech and gamma band envelopes reached up to 0.59 in single trials, the average across all trials never exceeded 0.35 for individual subjects. Second, both single-electrode correlation and classification accuracy (univariate and multivariate) level off after approximately 5 s. These two observations indicate that the gamma band envelope only holds limited information about the neural tracking of auditory attention. Other research has suggested that low-frequency amplitude and phase might encode additional information about selective auditory attention [21]. Future studies could explore whether combining gamma-band and low-frequency features can improve classification accuracy. As low-frequency features can be observed in EEG, this could eventually lead to a non-invasive BCI that uses auditory attention to natural speech.

In fact, results presented by Horton et al., [24] confirm the viability of alpha in the EEG for the identification of the attended speaker. While their study reported alpha amplitude at three latencies (90, 200, and 340 ms) to be informative, we found gamma power to be informative only at a latency of 100 ms. This may relate to a fundamental difference between their EEG study and our ECoG study. As these three latencies relate to the peaks of three well known auditory evoked responses [49], the observed alpha signal may represent discrete responses to the onset of the attended speech stimulus. In contrast, our results together with results from other ECoG studies [16–20] support the notion that the gamma band envelope tracks the envelope of perceived speech rather than just its onset.

4.4. Limitations

While our results indicate that the presented method could support BCI communication, the reported performance metrics may have been limited by our study design and the enrolled subjects.

For instance, in this study, subjects did not receive feedback on how well they performed the task. This is relevant, as many BCI studies have shown that providing feedback ensures that the subjects remain attentive to the task and that their performance improves over time [50, 51]. In addition, the lack of any behavioral verification in this task makes it difficult to determine whether the subjects properly attended as instructed. For this reason, we could not exclude any subjects or reject any trials on this basis. A similar study that used a behavioral verification reported that, on average, ~25% of the trials [22] were not attended. These trials then did not exhibit a neural tracking of the attended speech. For our study, this might account for the high variability that we saw in the performance across subjects.

One factor that could have made this task more difficult is our selection of the auditory stimuli. In our study, we chose speakers with similar linguistic features (i.e., male voices with similar cadence). Thus, our subjects could have had difficulties in performing the selective auditory attention task. Hence, subject and classification performance may be improved with speech stimuli that are have dissimilar linguistic features (e.g., male and female voices).

Another aspect that might have affected the subjects' performance in this study was cortical coverage. Grid placement was solely determined by clinical need. As a consequence, cortical coverage varied in location and density across subjects. Because there are other confounding variables (e.g., behavioral compliance and language dominance), and the low number of participants, we could not determine whether certain grid configurations yield better performance than others. At the same time, our average topographies (Figure 6) clearly show that coverage over STG or pre-motor cortex appears to be essential. This is important, as one consideration for our approach in a clinical application is where to implant the electrode and how to present the stimuli.

In our analyses, we assumed that the delay between speech stimuli and the elicited neural responses is constant across all cortical locations and subjects. However, detailed analysis of this delay (see Figure 3) revealed a standard deviation of 58 ms across cortical locations and subjects. This suggests that this delay varies across subjects and cortical locations as previously shown by Potes et al., [17], who reported a 110 ms delay in the neural tracking between STG and pre-motor cortex.

This is relevant, as the presented method depends on the correction of this delay to measure the proper correlation between speech stimuli and the neural response. In this initial study, we did not perform this correction individually for each cortical location and subject, as we wanted to keep the number of parameters as low as possible. Subsequent studies could further explore how correcting the delay for each subject and each electrode individually could improve communication performance.

In this study, we were unable to infer the attended speaker in 5 out of 12 human subjects. It is currently unclear whether this is because these 5 subjects simply did not or could not execute the task, similar to BCI illiteracy [52].

Because of these limitations, future studies are required to compare the presented approach with other established BCIs in their e cacy to provide BCI communication to the target population, i.e., people affected by severe neuro-degenerative diseases (e.g., late-stage amyotrophic lateral sclerosis (ALS) or locked-in syndrome).

5. Conclusion

This study confirms and extends earlier reports that showed that the envelope of an attended speech stimulus is preferentially represented in ECoG signals in the high gamma range and recorded over STG and a region of superior pre-motor cortex. We used this preferential representation of attended stimuli to identify to which speech the subjects attended when they were presented with two simultaneous speeches. Thus, our study provides evidence that attentional modulation of these neural signals should allow people to indicate a choice in a BCI context. At the same time, it is currently unclear whether this approach, which depends on invasively recorded ECoG, may have distinct advantages over other auditory attention-based BCIs that rely on scalp-recorded EEG.

In summary, our study shows that an auditory attention-based BCI that uses simultaneously presented natural speech stimuli could provide BCI communication without depending on other sensory modalities or a mapping between the stimulus and the communication intent. This provides the groundwork for future studies that could explore the practical utility of this approach for real-time BCI applications.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge Marcia Sanders for her invaluable assistance in editing the manuscript.

Funding

This work was supported by the NIH (EB006356 (GS), EB00856 (GS) and EB018783 (GS)), the US Army Research O ce (W911NF-07-1-0415 (GS), W911NF-08-1-0216 (GS) and W911NF-14-1-0440 (GS)) and Fondazione Neurone.

References

1. Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain-computer interfaces for communication and control. Clin Neurophysiol. 2002; 113(6):767–791. [PubMed: 12048038]
2. Brunner P, Joshi S, Briskin S, Wolpaw JR, Bischof H, Schalk G. Does the ’P300’ speller depend on eye gaze? J Neural Eng. 2010; 7(5):056013.
3. Treder MS, Blankertz B. (C)overt attention and visual speller design in an ERP-based brain-computer interface. Behav Brain Funct. 2010; 6(1):28–28. [PubMed: 20509913]
4. Belitski A, Farquhar J, Desain P. P300 audio-visual speller. J Neural Eng. 2011; 8(2):025022. [PubMed: 21436523]

5. Furdea A, Halder S, Krusienski DJ, Bross D, Nijboer F, Birbaumer N, Kübler A. An auditory oddball (P300) spelling system for brain-computer interfaces. Psychophysiology. 2009; 46(3):617–625. [PubMed: 19170946]
6. Klobassa DS, Vaughan TM, Brunner P, Schwartz NE, Wolpaw JR, Neuper C, Sellers EW. Toward a high-throughput auditory P300-based brain-computer interface. Clin Neurophysiol. 2009; 120(7): 1252–1261. [PubMed: 19574091]
7. Halder S, Rea M, Andreoni R, Nijboer F, Hammer EM, Kleih SC, Birbaumer N, Kübler A. An auditory oddball brain-computer interface for binary choices. Clin Neurophysiol. 2010; 121(4):516–523. [PubMed: 20093075]
8. Schreuder M, Blankertz B, Tangermann M. A new auditory multi-class brain-computer interface paradigm: spatial hearing as an informative cue. PLoS One. 2010; 5(4)
9. Brouwer AM, van Erp JB. A tactile P300 brain-computer interface. Front Neurosci. 2010; 4:19–19. [PubMed: 20582261]
10. van der Waal M, Severens M, Geuze J, Desain P. Introducing the tactile speller: an ERP-based brain-computer interface for communication. J Neural Eng. 2012; 9(4):045002. [PubMed: 22831906]
11. Riccio A, Mattia D, Simione L, Olivetti M, Cincotti F. Eye-gaze independent EEG-based brain-computer interfaces for communication. J Neural Eng. 2012; 9(4):045001. [PubMed: 22831893]
12. Lopez-Gordo MA, Fernandez E, Romero S, Pelayo F, Prieto A. An auditory brain–computer interface evoked by natural speech. J Neural Eng. 2012; 9(3):036013. [PubMed: 22626956]
13. Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. Science. 1995; 270(5234):303–304. [PubMed: 7569981]
14. Ray S, Crone NE, Niebur E, Franaszczuk PJ, Hsiao SS. Neural correlates of high-gamma oscillations (60–200 Hz) in macaque local field potentials and their potential implications in electrocorticography. J Neurosci. 2008; 28(45):11526–11536. [PubMed: 18987189]
15. Ball T, Kern M, Mutschler I, Aertsen A, Schulze-Bonhage A. Signal quality of simultaneously recorded invasive and non-invasive EEG. NeuroImage. 2009; 46(3):708–716. [PubMed: 19264143]
16. Martin S, Brunner P, Holdgraf C, Heinze HJ, Crone NE, Rieger J, Schalk G, Knight RT, Pasley B. Decoding spectrotemporal features of overt and covert speech from the human cortex. Front Neuroeng. 2014; 7(14)
17. Potes C, Gunduz A, Brunner P, Schalk G. Dynamics of electrocorticographic (ECoG) activity in human temporal and frontal cortical areas during music listening. NeuroImage. 2012; 61(4):841–848. [PubMed: 22537600]
18. Potes C, Brunner P, Gunduz A, Knight RT, Schalk G. Spatial and temporal relationships of electrocorticographic alpha and gamma activity during auditory processing. NeuroImage. 2014; 97:188–195. [PubMed: 24768933]
19. Pasley 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(1):e1001251. [PubMed: 22303281]
20. Kubanek J, Brunner P, Gunduz A, Poeppel D, Schalk G. The tracking of speech envelope in the human cortex. PloS One. 2013; 8(1):e53398. [PubMed: 23408924]
21. Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE. Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. Neuron. 2013; 77(5):980–991. [PubMed: 23473326]
22. Mesgarani N, Chang EF. Selective cortical representation of attended speaker in multi-talker speech perception. Nature. 2012; 485(7397):233–236. [PubMed: 22522927]
23. Kerlin JR, Shahin AJ, Miller LM. Attentional gain control of ongoing cortical speech representations in a ”Cocktail party”. J Neurosci. 2010; 30(2):620–628. [PubMed: 20071526]
24. Horton C, Srinivasan R, D'Zmura M. Envelope responses in single-trial EEG indicate attended speaker in a cocktail party. J Neural Eng. 2014; 11(4):046015. [PubMed: 24963838]
25. Talairach, J.; Tournoux, P. Co-planar sterotaxic atlas of the human brain. Thieme Medical Publishers, Inc.; New York: 1988.

26. Wechsler, D. Weschsler Adult Intelligence Scale-III. The Psychological Corporation; San Antonio, TX: 1997.
27. Wada J, Rasmussen T. Intracarotid injection of sodium amytal for the lateralization of cerebral speech dominance. J Neurosurg. 1960; 17:266–282.
28. Schalk G, McFarland DJ, Hinterberger T, Birbaumer N, Wolpaw JR. BCI2000: a general-purpose brain-computer interface (BCI) system. IEEE Trans Biomed Eng. 2004; 51(6):1034–1043. [PubMed: 15188875]
29. Mellinger, J.; Schalk, G. BCI2000: A general-purpose software platform for BCI.. In: Dornhege, G.; del R Millan, J.; Hinterberger, T.; McFarland, D.; Müller, K., editors. Toward brain-computer interfacing. MIT Press; Cambridge, MA, USA: 2007. p. 359-367.
30. Schalk, G.; Mellinger, J. A practical guide to brain-computer interfacing with BCI2000. 1st ed.. Springer; London, UK: 2010.
31. Miller KJ, Sorensen LB, Ojemann JG, den Nijs M. Power-law scaling in the brain surface electric potential. PLoS Comput Biol. 2009; 5(12):e1000609. [PubMed: 20019800]
32. Zion Golumbic EM, Poeppel D, Schroeder CE. Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective. Brain Lang. Sep; 2012 122(3): 151–161. [PubMed: 22285024]
33. Wolpaw JR, Ramoser H, McFarland DJ, Pfurtscheller G. EEG-based communication: Improved accuracy by response verification. IEEE Trans Rehab Engin. 1998; 6:326–333.
34. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995:289–300.
35. Wilson SM, Iacoboni M. Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception. NeuroImage. 2006; 33(1):316–325. [PubMed: 16919478]
36. Kübler A, Kotchoubey B, Kaiser J, Wolpaw JR, Birbaumer N. Brain-computer communication: unlocking the locked in. Psychol Bull. 2001; 127(3):358–375. [PubMed: 11393301]
37. Leuthardt EC, Freudenberg Z, Bundy D, Roland J. Microscale recording from human motor cortex: implications for minimally invasive electrocorticographic brain-computer interfaces. Neurosurg Focus. 2009; 27(1):E10. [PubMed: 19569885]
38. Davson H. Review lecture. The blood-brain barrier. J Physiol. 1976; 255(1):1–28. [PubMed: 1255511]
39. Hamer HM, Morris HH, Mascha EJ, Karafa MT, Bingaman WE, Bej MD, Burgess RC, Dinner DS, Foldvary NR, Hahn JF, Kotagal P, Najm I, Wyllie E, Lüders HO. Complications of invasive video-EEG monitoring with subdural grid electrodes. Neurology. 2002; 58(1):97–103. [PubMed: 11781412]
40. Fountas KN, Smith JR. Subdural electrode-associated complications: a 20-year experience. Stereotact Funct Neurosurg. 2007; 85(6):264–272. [PubMed: 17709978]
41. Van Gompel JJ, Worrell GA, Bell ML, Patrick TA, Cascino GD, Raffel C, Marsh WR, Meyer FB. Intracranial electroencephalography with subdural grid electrodes: techniques, complications, and outcomes. Neurosurgery. 2008; 63(3):498–505. [PubMed: 18812961]
42. Wong CH, Birkett J, Byth K, Dexter M, Somerville E, Gill D, Chaseling R, Fearnside M, Bleasel A. Risk factors for complications during intracranial electrode recording in presurgical evaluation of drug resistant partial epilepsy. Acta Neurochir (Wien). 2009; 151(1):37–50. [PubMed: 19129963]
43. Torres Valderrama A, Oostenveld R, Vansteensel MJ, Huiskamp GM, Ramsey NF. Gain of the human dura in vivo and its effects on invasive brain signal feature detection. J Neurosci Methods. 2010; 187(2):270–279. [PubMed: 20109492]
44. Bundy DT, Zellmer E, Gaona CM, Sharma M, Szrama N, Hacker C, Freudenburg ZV, Daitch A, Moran DW, Leuthardt EC. Characterization of the effects of the human dura on macro- and micro-electrocorticographic recordings. J Neural Eng. 2014; 11(1):016006. [PubMed: 24654268]
45. Sillay KA, Rutecki P, Cicora K, Worrell G, Drazkowski J, Shih JJ, Sharan AD, Morrell MJ, Williams J, Wingeier B. Long-term measurement of impedance in chronically implanted depth and subdural electrodes during responsive neurostimulation in humans. Brain Stimul. 2013; 6(5):718–726. [PubMed: 23538208]

46. Stieglitz, T. Miniaturized neural interfaces and implants in neurological rehabilitation.. In: Jensen, W.; Andersen, OKs; Akay, M., editors. Replace, repair, restore, relieve bridging clinical and engineering solutions in neurorehabilitation. Vol. 7 of Biosystems & Biorobotics. Springer International Publishing; 2014. p. 9-14.

47. Owen AM, Coleman MR, Boly M, Davis MH, Laureys S, Pickard JD. Detecting awareness in the vegetative state. Science. 2006; 313(5792):1402. [PubMed: 16959998]

48. Hill NJ, Scholkopf B. An online brain-computer interface based on shifting attention to concurrent streams of auditory stimuli. J Neural Eng. 2012; 9(2):026011. [PubMed: 22333135]

49. Picton TW, Hillyard SA, Krausz HI, Galambos R. Human auditory evoked potentials. I: Evaluation of components. Electroen Clin Neuro. 1974; 36:179–190.

50. McFarland DJ, McCane LM, Wolpaw JR. EEG-based communication and control: short-term role of feedback. IEEE Trans Rehabil Eng. 1998; 6(1):7–11. [PubMed: 9535518]

51. Miller KJ, Schalk G, Fetz EE, den Nijs M, Ojemann JG, Rao RP. Cortical activity during motor execution, motor imagery, and imagery-based online feedback. Proc Natl Acad Sci U S A. 2010; 107(9):4430–4435. [PubMed: 20160084]

52. Allison, BZ.; Neuper, C. Could Anyone Use a BCI?. In: Tan, DS.; Nijholt, A., editors. Brain-Computer Interfaces - Applying our Minds to Human-Computer Interaction. Springer; London: 2010. p. 35-54.

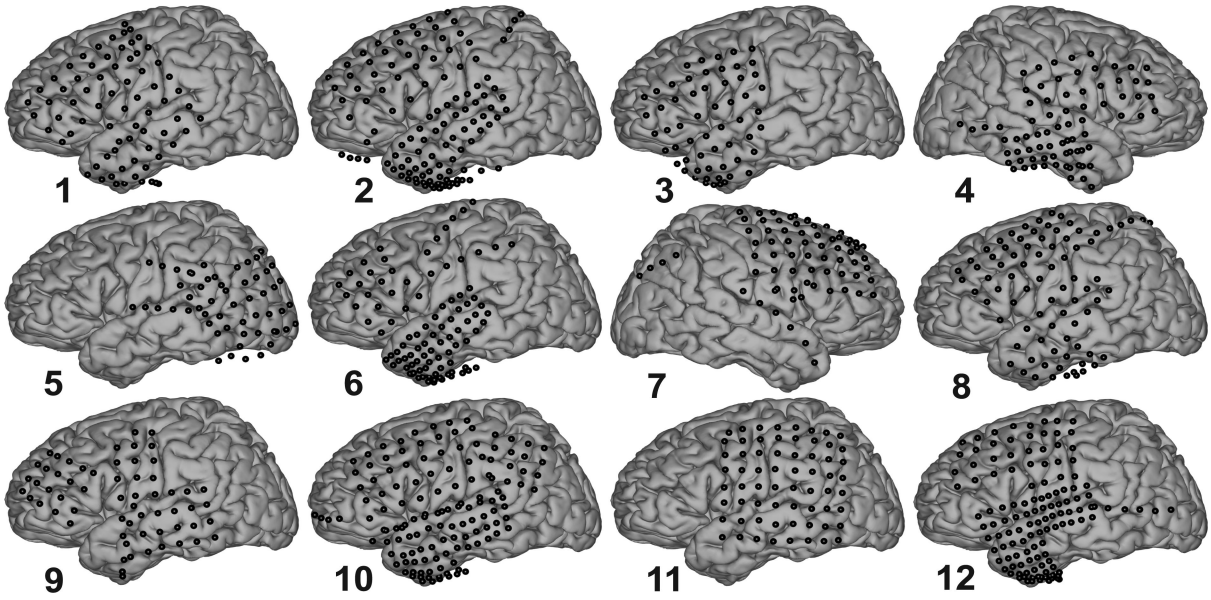

Figure 1. Electrode coverage

Electrode coverage and density varied across subjects. Electrode locations (black dots) included frontal, temporal, parietal and occipital cortical areas. Four subjects (4, 6, 8 and 12) were implanted with high-density grids (electrodes spaced 6 mm apart).

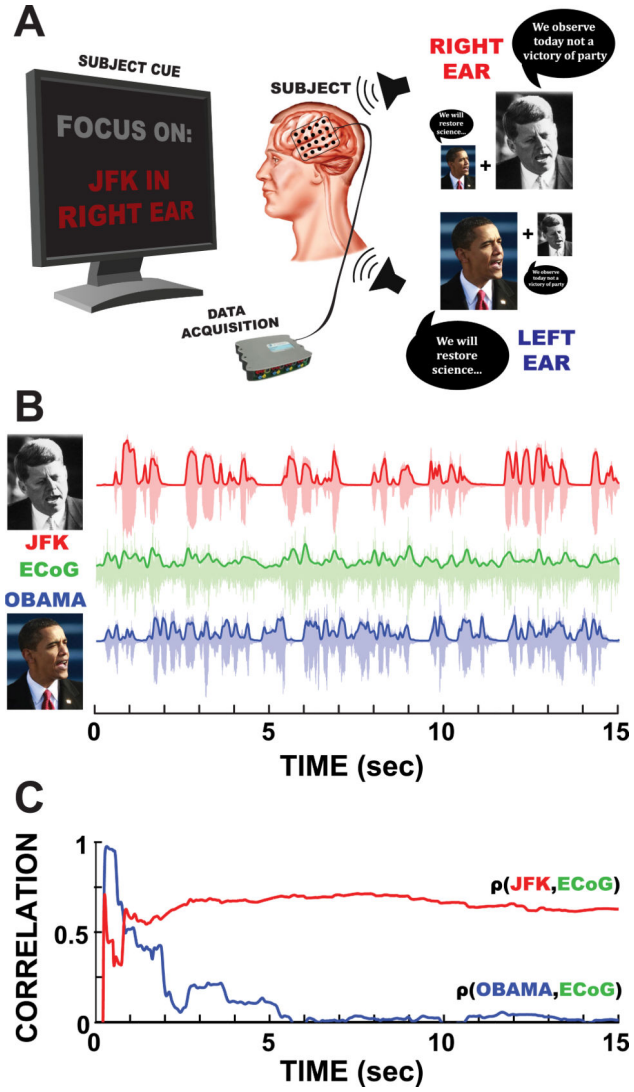

Figure 2. Experimental setup and methods

(A) Subjects selectively directed auditory attention to one of two simultaneously presented speakers. **(B)** We extracted the envelope of ECoG signals in the high gamma band, as well as the envelopes of the attended and unattended speech stimuli (i.e., JFK and Obama). **(C)** The correlation between the envelopes of the ECoG gamma band and the attended speech stimulus, accumulated over time, is markedly larger than the accumulated correlation between the envelopes of the ECoG gamma band and the unattended speech stimulus.

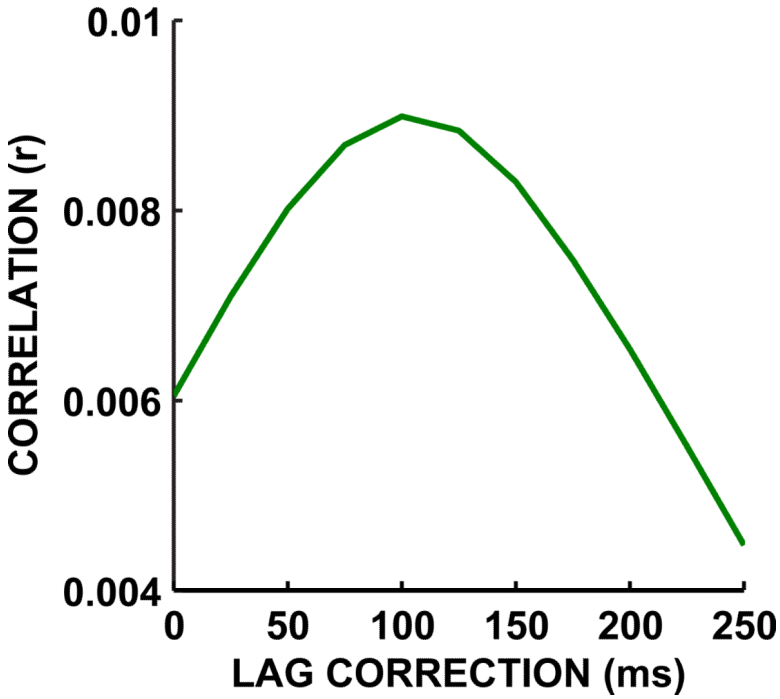

Figure 3. Lag between speech presentation and neural response

This figure shows the correlation between neural response and the attended speech (green), averaged across subjects, for corrected lags between 0 and 250 ms to peak at 100 ms.

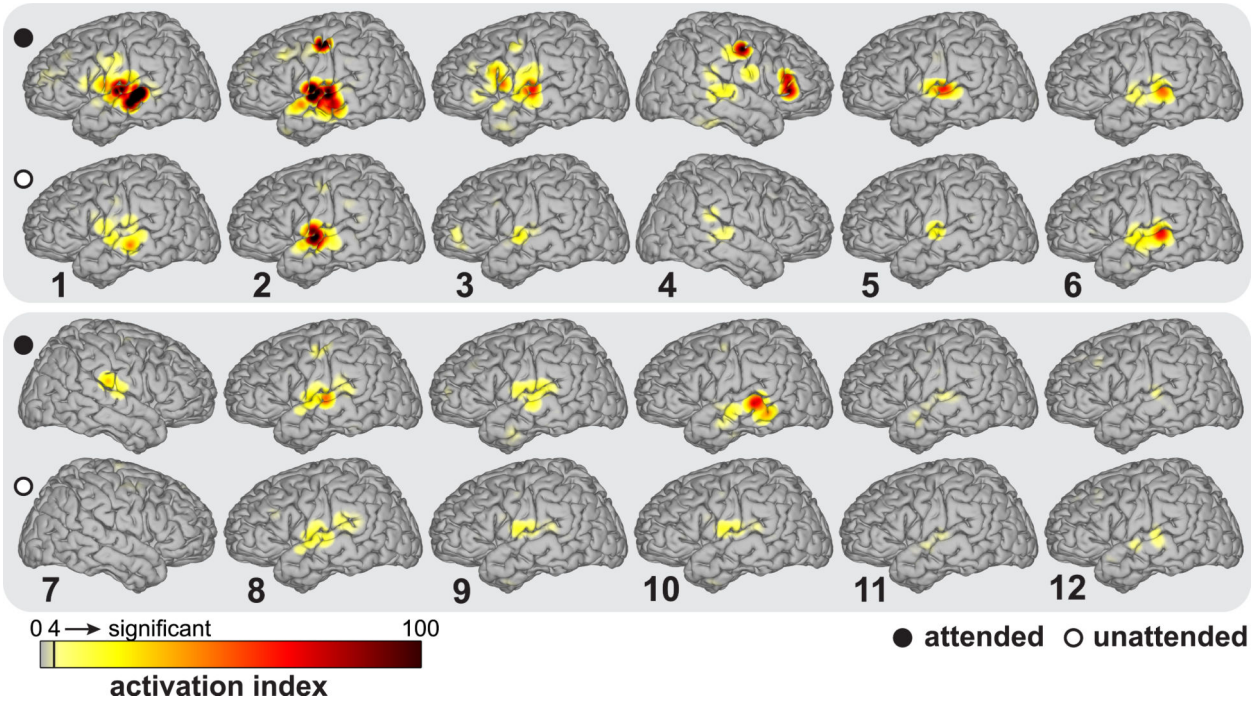

Figure 4. Neural tracking of attended (●) and unattended (○) speech

Neural tracking is measured as the correlation between the high gamma ECoG envelope and the attended or unattended speech envelope. Color gives the magnitude of this correlation expressed as an activation index (−log(p)).

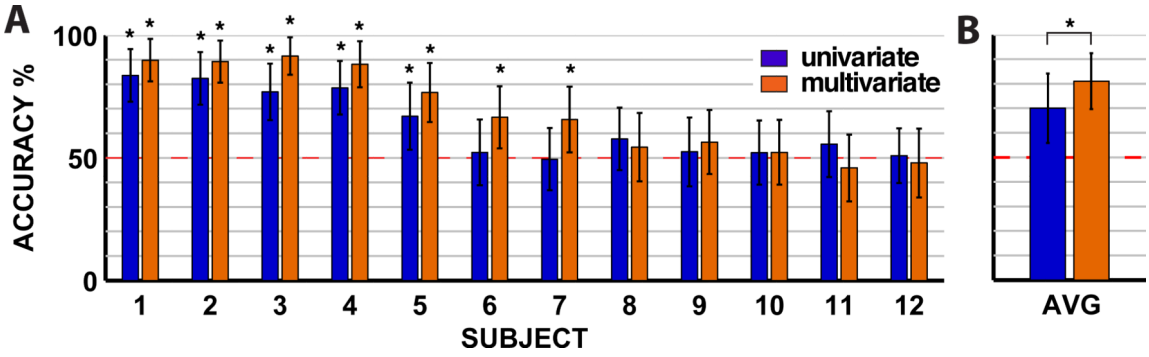

Figure 5. Classification accuracy to which the attended speech could be identified, using a univariate (blue) or multivariate (orange) classification method

(A) Accuracy per subject, sorted by average performance. For subjects 1-7 ('significant subjects'), accuracy is significantly larger than chance for at least one classification method (adjusted for multiple comparisons using a false discovery rate with q = 0.05). Significance is marked with an asterisk. **(B)** Average accuracy across subjects for subjects with statistically significant performance.

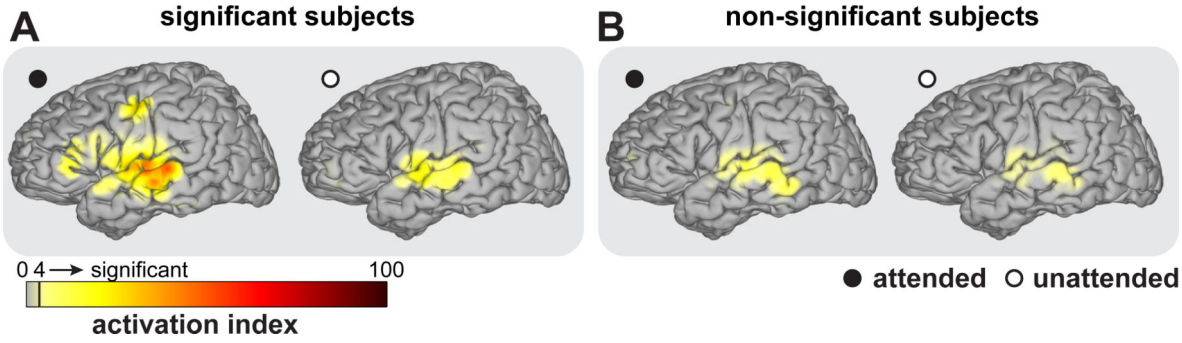

Figure 6. Neural tracking of attended (●) and unattended (○) speech

Two averages are displayed: **(A)** Subjects for which performance was significantly better than chance for at least one classification method and **(B)** subjects for which performance was at chance level. For the significant subjects, the tracking of the attended speech is both stronger and more widely distributed than the tracking of the unattended speech. For the non-significant subjects, the overall activation index is smaller. In addition, there is only a marginal difference in spatial distribution between attended and unattended stimuli.

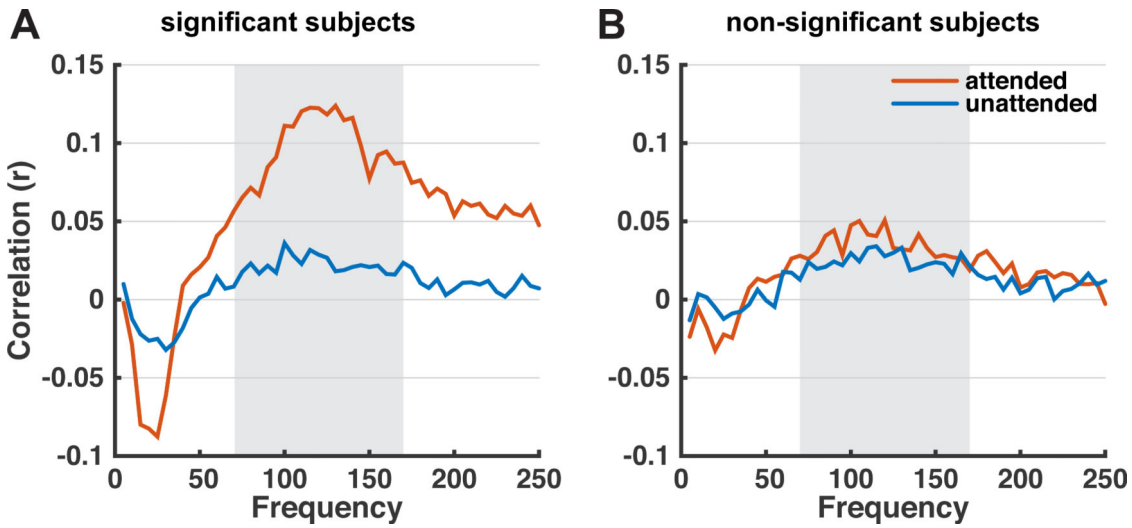

Figure 7. Neural tracking of attended and unattended speech across different frequencies

(A) and **(B)** show correlation coefficients across different frequencies, averaged across subjects, and for attended (orange trace) and unattended speech (blue trace). **(A)** Subjects for which performance was significantly better than chance for at least one classification method. **(B)** Subjects for which performance was at chance level. For the significant subjects, the tracking of the attended speech is stronger across all frequency bands, especially in the high gamma band (70–170 Hz, gray shaded). The tracking shows a negative relationship in the low frequency band (10–30 Hz). For the non-significant subjects, this negative relationship at lower frequencies is not apparent and the tracking of attended and unattended speech at higher frequencies is at the same low level.

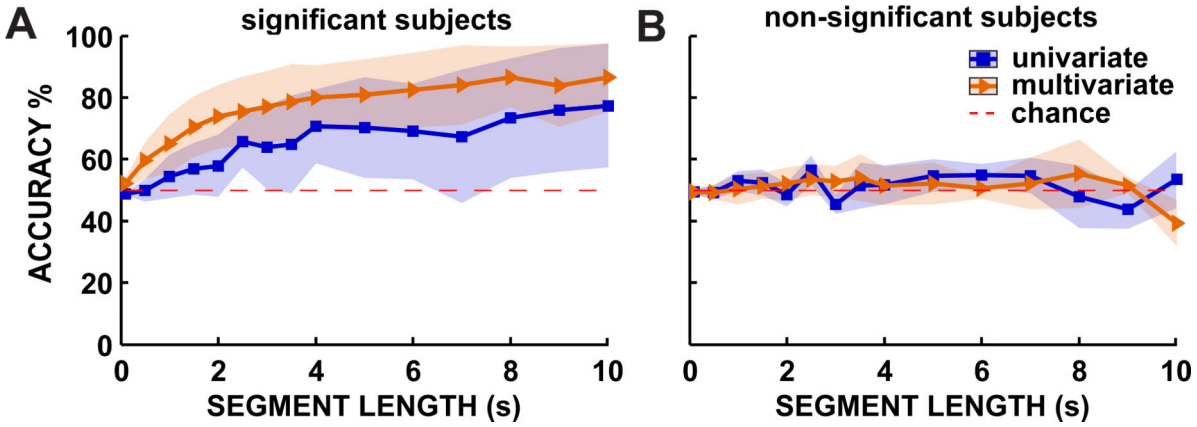

Figure 8. Accuracy for different segment lengths for univariate (blue) and multivariate methods (orange)

The classification accuracy increases steadily with segment length for both classification methods. Multivariate classification results in higher average accuracy than univariate classification for all segment lengths.

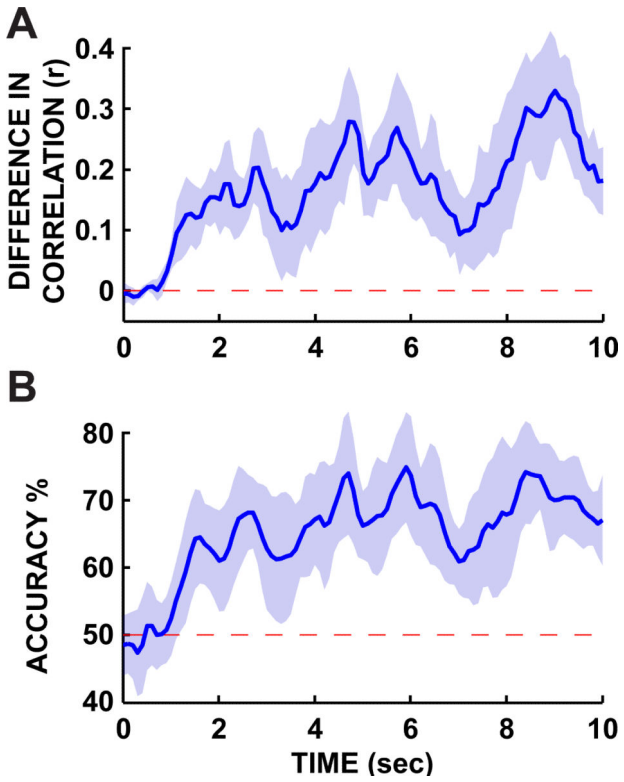

Figure 9. Effect of 'tuning-in' on correlation (A) and classification accuracy (B)

For the first second, the difference between the 'attended' and 'unattended' correlation remains zero resulting in a classification accuracy around chance level (i.e., 50%). Subsequently, correlation and classification accuracy trend upwards while being superimposed with cycles of higher and lower correlation and classification accuracy.

Table 1

Subject information

The corresponding electrode locations are shown in Figure 1.

Subject	Age	Sex	Handedness	Language dominance	Grid hemisphere	Number of electrodes
1	49	F	Left	Left	Left	72
2	28	F	Right	Bilateral	Left	120
3	45	M	Right	Left	Left	58
4	54	M	Left	Left	Right	75
5	60	M	Right	Left	Lef	59
6	25	F	Right	Left	Left	98
7	15	F	Right	N/A	Right	71
8	45	M	Right	N/A	Left	81
9	45	M	Left	Left	Left	61
10	28	M	Right	Left	Left	133
11	52	M	Left	Left	Left	64
12	24	F	Right	Bilateral	Left	128