Radboud University Nijmegen

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Artificial Intelligence

Identifying Attended Speech from Electrocorticographic Signals in a 'Cocktail Party' Setting

Author:
Karen Dijkstra

Supervisors:
Dr. Peter Brunner[1]
Dr. Jason Farquhar[2]
Dr. Gerwin Schalk[1]
External examiner:
Prof.dr.ir Peter Desain[2]

[1] Center for Adaptive Neurotechnologies, Wadsworth Center, Albany, NY
[2] Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour

July 1, 2014

Identifying Attended Speech from Electrocorticographic Signals in a 'Cocktail Party' Setting

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. These people are unable to use traditional assistive communication devices that depend on residual muscle control, or brain-computer interfaces (BCIs) that rely on the ability to control gaze. Auditory and tactile BCIs are considered as some of the few remaining communication options for such individuals.

In this study we aimed to determine the viability of auditory attention to speech as a paradigm for BCI. We analyzed data from an experiment in which subjects attended to one of two speakers, to determine if the attended speech can be identified with better than chance performance in single trials.

Our results show that we can correctly identify the attended speech in 7 out of 12 subjects, with an accuracy of 80% over segments of data between 4-6 seconds in length, using a regularized logistic regression. Additionally, with segments as short as 2 seconds, the average accuracy for these subjects was 70%, commonly regarded as sufficient accuracy for BCI communication. When only a single ECoG channel (i.e., cortical location) was used for classification, the attended speech could be identified in 5/12 subjects, averaging to 77% accuracy across segments 4-6 segments in length.

Even though we were unable to determine why this approach failed to produce results for 5 participants, we believe that these results demonstrate the potential of this paradigm for BCI. Obvious next steps for this research include a further investigation of the large subject variability observed, the development of an online implementation of this paradigm and/or an expansion of the current experimental set up to determine how the obtained classification accuracy scales with an increased number of simultaneously presented speech stimuli.

Introduction

Communication is an essential part of being human, allowing us to interact with each other, establish relationships and express needs and desires.

This fundamental human ability can become compromised in people affected by paralysis, as they are no longer able to gesture or speak. Conventional assistive devices (e.g., eye trackers or tongue/cheek switches) reestablish communication but generally rely on some residual muscle control (eye or mouth movements). In contrast, Brain-Computer Interfaces (BCIs) translate brain signals directly into communication output, effectively circumventing the otherwise necessary muscular pathways (Wolpaw et al., 2002). However, BCIs still depend on perceptual modalities, such as auditory, tactile or, most frequently, visual perception, for stimulation or feedback.

This visual modality is popular as the resulting BCIs are generally most usable and intuitive. However, recent studies have shown that visual BCIs, such as the popular 'P300' matrix speller, still depend on eye-gaze (Brunner et al., 2010; Treder and Blankertz, 2010), which is lost in people affected by severe neuro-degenerative diseases (e.g., late-stage ALS or locked-in syndrome).

This has led to an increased interest in BCI paradigms that use non-visual sensory modalities, such as auditory (Hill and Scholkopf, 2012; Belitski et al., 2011; Furdea et al., 2009; Klobassa et al., 2009; Halder et al., 2010; Schreuder et al., 2010) or tactile stimulation (Brouwer and van Erp 2010; van der Waal et al. 2012; see Riccio et al. 2012 for review). A successful paradigm for these auditory BCI's allows users to make a binary decision by attending to one out of two simultaneously presented streams of tones (Hill and Scholkopf, 2012). From the brain activity the stream that was attended is inferred, allowing the user to encode their intention by attending to a specific stream.

Drawbacks of this approach are the fact that the stimuli streams used for this BCI are artificial, requiring cognitive effort to continiously attend to the streams, and the application is limited in the number of streams that can be presented simultaneously. To address this first issue, there have been attempts to develop auditory BCIs that use speech stimuli instead of artificial tones (Lopez-Gordo et al., 2012). However in this approach, the speech stimuli (i.e., words or phonemes) are presented in a specific temporal pattern, to elicit brain responses in a predictable pattern. Following speech is a natural human ability and attending to one speaker while there are speakers or other noise in the background is a task we perform almost daily (known as the cocktail party effect). Yet, if this speech structure is altered, the intuitiveness of this approach is likely reduced. Ideally, we would want to use natural speech as stimulation for such a BCI.

Recent advancements in the neuroscience of speech perception lead us to believe that it should be possible to identify which speech was attended out of multiple, simultaneously presented speech stimuli. Specifically, these advancements have shown that the neural tracking of speech can be measured with electrocorticography (ECoG) (Martin et al., 2014; Potes et al., 2012, 2014; Pasley et al., 2012; Kubanek et al., 2013), and that this neural tracking is selective to the attended speech (Zion Golumbic et al., 2013; Kerlin et al., 2010; Mesgarani and Chang, 2012). However, whether the attended speech can be identified in single trials and to what extent this effect may support BCI communication remains to be determined. To answer these questions we analyzed a dataset from an auditory attention experiment, in which subjects attended to one out of two speakers, while their brain activity was measured with ECoG.

Organization of this thesis

In the next section, the topics relevant to this thesis will be described in more detail. These topics include the research on the neural tracking of speech that prompted this research and the measuring of brain activity using ECoG. At the end of this background section I will state the questions we aimed to answer with this research. In the subsequent section I will describe the auditory attention experiment that was performed to obtain the data we analyzed. The section 'Data analysis' will then describe the analysis we used to obtain our results. These results will then be detailed in the following section. Finally, in the 'Discussion', the research questions will be addressed on the basis of these results, together with the conclusion that the results show that auditory attention to speech indeed has potential as a paradigm for BCI. Additionally some remaining questions will be discussed, together with suggestions for follow-up research.

Background

Neural tracking of speech

Perception of speech is a task people perform daily with relatively little effort. However, speech perception is not a trivial task, evidenced by the fact that the first steps in automated speech recognition were taken in the 1950s, but even now, more than 60 years later, speech recognition only plays a limited role in our interaction with computers (e.g., SIRI; Apple Inc. Cupertino, CA).

For human speech perception, the integrity of the temporal structure of speech is crucial to the understanding of speech. If this temporal structure is altered by slowing or increasing the speed, speech can become unintelligible (Ahissar et al., 2001). On the other hand, when non-temporal properties are removed, and only modulation of the speech amplitude over time is maintained, speech remains intelligible (Shannon et al., 1995). This specific spectro-temporal structure of speech is a result of 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.

Recent studies have shown that the neural response to speech is reflective of this spectro-temporal structure of the perceived speech. This has been demonstrated by reconstructing speech from a multitude of neural features and by showing that the envelope of specific frequency bands (e.g., high gamma), at specific cortical locations (e.g. Superior Temporal Gyrus, STG), track the envelope of perceived speech (Martin et al., 2014; Potes et al., 2012, 2014; Pasley et al., 2012; Kubanek et al., 2013).

On top of that, these findings have recently been extended to simultaneously presented streams of speech (Zion Golumbic et al., 2013; Kerlin et al., 2010; Mesgarani and Chang, 2012). This is of interest for neuroscience as the precise mechanism that allows us to perform this task where we attend to a single speaker in a multi-speaker environment is unknown. This ability is called the 'cocktail party' effect, and it is this ability that makes attention to speech a promising approach for BCIs. These studies shown, that in such a cocktail party situation the neural tracking is stronger for attended than for the unattended speech.

These findings form the basis for the research in this thesis, as they indicate that these measures of neural tracking could allow for the identification of the attended speech in single trials. Note that the majority of the cited research on speech perception in this section was obtained from ECoG experiments, similar to the experiment analysed in this study.

Electrocorticography (ECoG) ECoG is a method of measuring brain activity, in which electrodes are placed directly on the surface of the brain. Compared to regular EEG, in which electrodes are placed on the scalp, this has the advantage that the measured activity has not yet passed through the skull. Ths is relevant, as the skull is not very conductive, which reduces the strength of the signal and makes it more difficult to determine the original source of the activity in EEG. ECoG therefore has much higher spatial resolution than EEG, though it is limited in its spatial resolution to those areas of the cortex on which the grids were placed. This high spatial resolution, combined with the high temporal resolution that is inherent to electrical measurements of brain activity, makes ECoG particularly interesting for BCI.

However, ECoG is highly invasive, as it requires a part of the skull to be removed surgically (a craniotomy), to place the electrodes on the cortex. For this reason, ECoG research is performed only in cases of medical necessity. This is, for instance, the case for epilepsy patients who suffer from severe epileptic seizures and do not respond sufficiently to medication. For these patients, ECoG electrodes are used to localize the epileptogenic zones, and to identify important functional cortical areas, prior to surgical resection.

Research has shown that BCI's using ECoG perform much better than their EEG counterparts (Brunner et al., 2011). While ECoG is highly invasive, and is currently not widely used for BCIs, for people that have few to no remaining options for communication, such an invasive method might be an acceptable trade off.

Research Questions

In this thesis, we analyzed a dataset from a previously conducted auditory attention experiment, in which subjects attended to one of two speakers in a cocktail party setting. The goal of this research was to analyze the data to answer the following research questions. The main question for this thesis:

Q1: Can the attended speech be identified from the brain activity, in single trials, with better than chance performance?

While better than chance performance is in theory sufficient to extract information from a persons brain activity, in practice it does not necessarily lead to a useable BCI. Commonly, for a binary BCI, classification accuracy of 70% is considered sufficient for communication (Kübler et al., 2001). Furthermore, the accuracy is not the only relevant factor, which leads to the following additional questions.

Q1a: Is this performance sufficient for communication (>=70%)?

Q1b: What is the minimum length of stimulation required for this performance?

An additional factor that is relevant for a potential BCI application is the invasiveness of ECoG. The cortical electrodes implanted in the epilepsy patients that participated in this experiment consisted of grids of a few up to a few dozed of electrodes, covering a range of cortical areas. The size of the craniotomy for the placement of these electrodes, depends to a degree on the size of the grid that is placed. Notably, if only a single electrode would be required, a single drill hole in the skull could suffice. This leads to an additional question:

Q1c: How well can the attended speech be predicted, when only a single ECoG channel (i.e.,cortical location) is taken into account?

Auditory Attention Experiment

Subjects

Twelve subjects participated in this auditory attention experiment. Each of them underwent temporary placement of subdural electrodes as part of their clinical treatment for epilepsy. These electrodes were implanted for a duration of 5–7 days. During this period, subjects voluntarily participated in our study. 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 between 15–60 years old (median 45), each with an IQ higher than 75 (median 95). None of the subjects had a history of hearing impairment. A Wada test was performed to determine the language dominance of the subjects (Wada and Rasmussen, 1960). In the Wada test, the hemisphere that is responsible for language is determined by alternatively shutting down one of the hemispheres by injecting a sedative. The degree to which this degrades language abilities in the patients signifies the importance of this hemisphere in language function. This is used in the clinical treatment of these epilepsy patients, to determine whether language function is present in the hemisphere targeted for surgical intervention.

As language dominance may be relevant for this study, the results of this Wada test are summarized in Table 1, together with other relevant subject information. All subjects provided informed consent, and the study was approved by the institutional review board of Albany Medical College.

The subjects had between 57 and 133 subdural electrodes implanted over their left or right hemisphere. These 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). The cortical locations of these electrodes were verified using post-operative radiographs (anterior-posterior and lateral) and computed tomography (CT) scans. Subject-specific 3D cortical brain models were created from high-resolution pre-operative magnetic resonance imaging (MRI) scans, using Curry software (Neuroscan Inc, El Paso, TX). The MRIs were co-registered by means of the post-operative CT and the electrode coordinates were extracted according to the Talairach Atlas (Talairach and Tournoux, 1988). These electrode coordinates are depicted on Talairach template brains in Figure 1. Across all subjects, electrode coverage included frontal, temporal, parietal and occipital cortical areas.

Data Collection

ECoG signals were recorded from the implanted electrodes using g.USBamp or g.HIamp (g.tec, Graz, Austria), at a sampling rate of 1200 Hz. Data acquisition and stimulus presentation were accomplished using the BCI2000 software platform (Schalk et al., 2004; Mellinger and Schalk, 2007; Schalk and Mellinger, 2010). Clinical monitoring occurred simultaneously with data acquisition for this experiment by using a connector that split the cables coming from the patient into two sets, one that was connected to the clinical monitoring system and another set that was connected to the g.tec amplifiers. This ensured that clinical care or 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 ground and reference.

Stimuli and Task

Auditory stimuli were created from fragments of speeches from two speakers (John F. Kennedy and Barack Obama; each delivering their inauguration address). For a given stimulus, a fragment from John F. Kennedy was paired with a fragment from Barack

Table 1: Subject information including age, sex, handedness, hemispheric language dominance, hemisphere of the implanted grid and the total number of electrodes. The corresponding electrode locations are depicted 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

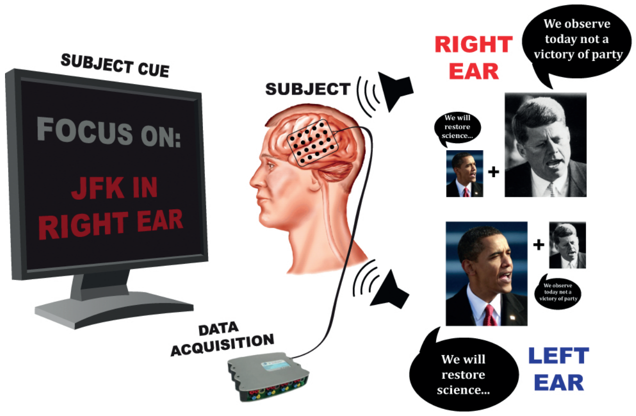
SUBJECT CUE
FOCUS ON:
JFK IN RIGHT EAR
SUBJECT
DATA ACQUISITION
RIGHT EAR
We observe today not a victory of party
We will restore science...
+
+
We observe today not a victory of party
We will restore science...
LEFT EAR

Figure 2: **Experimental setup and method.** Subjects selectively directed auditory attention to one of two speakers (John F. Kennedy and Barack Obama) in a cocktail party setting.

Obama. Fragments consisted of up to a few sentences of speech, between 15 and 25 seconds in length, with each fragment pair matched in length (10 fragment pairs in total). To simulate a cocktail party setting the paired fragments were mixed into a binaural presentation. This binaural presentation consisted of two auditory streams, one presented to the left ear and one presented to the right ear. A given stimulus contained 20% of the volume of one speaker and 80% of the other speaker for one ear, with the opposite volume configuration for the other ear. From each pair of fragments, two different stimuli were created, so that each fragment was once presented predominantly left and once predominantly right, resulting in 20 different stimuli.

The subjects' task consisted of selectively directing auditory attention to one of the two speakers in the stimuli. These stimuli were presented through in-ear earphones. Each trial started with an auditory and visual cue indicating the target speaker and side, followed by a stimulus and ending in a rest period of 5 s. This experimental set up is depicted in Figure 2

In the experiment each stimulus was presented two times, once with Obama as the target speaker and once with John F. Kennedy. Over these four presentations, the aural location (left and right) and the identity of the attended speaker (JFK and Obama) were permuted. In other words, over these four trials, the subjects were required to attend to each of the two speakers at each of the two aural locations.

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.

Data analysis

The data from the experiment consisted of the ECoG data for each of the subjects, the auditory stimuli (left and right) and the original audio fragments of each speaker from which the auditory stimuli were constructed.

The goal of the analysis was to determine whether it is possible to identify which speech was attended on the basis of the ECoG data and the auditory data. we measured the neural tracking of each of the individual fragments (from which the stimuli for a given trial had been composed) rather than measuring the tracking of the presented stimuli, as the audio fragments had been mixed together for this. From these audio fragments we created two data vectors, one labeled attended and one labeled unattended, that contained for each trial the fragment that was attended or unattended, respectively.

The main steps taken in this data analysis consisted of the pre-processing of the ECoG data, the extraction of features from the ECoG data and the auditory data, and

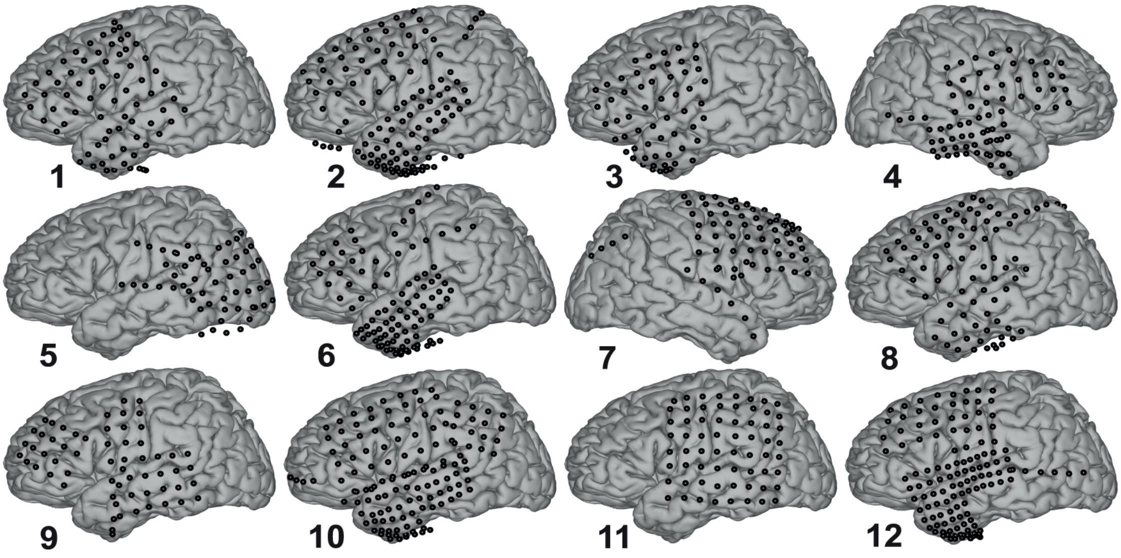

Figure 1: **Electrode coverage.** Electrode coverage and density varied across subjects. Electrode locations (in black) included frontal, temporal, parietal and occipital cortical areas. Four subjects (2, 6, 10 and 12) were implanted with high-density grids (electrodes spaced 6 mm apart).

the classification procedure in which the predictiveness of these features was measured.

Preprocessing

The ECoG data from each subject was preprocessed to remove noise. The data was first high-pass filtered at 0.5 Hz to remove drift. We then re-referenced all channels, using a common average reference, to remove the noise common across channels. This reference was composed from only those channels for which the 60 Hz line noise was within 1.5 standard deviations of the average. In other words, the amount of 60 Hz line noise was used as an indicator of the quality of the data from that channel and channels that were considered noisy by this indicator were excluded from the common average, to avoid spreading this noise to other channels. Finally, we used a notch filter to remove any 60 Hz line noise that remained after the common average reference.

Feature extraction

For the purpose of this research, neural tracking is defined as the correlation between a speech envelope and a high-gamma ECoG envelope that was measured during the presentation of this speech envelope. This was chosen based on existing studies that used this as a measure of neural tracking (Potes et al., 2012; Kubanek et al., 2013).

To compute this correlation between the ECoG envelope and the speech envelope, specific frequency bands needed to be extracted from the ECoG signals and the speech data, followed by an envelope extraction and a resampling to the same sampling rate.

For this correlation analysis there are three factors that are relevant: the frequencies represented in neural signal, the temporal relationship between the neural signal and the speech envelope and the spatial distribution of the neural signal across the cortex . Therefore there are three relevant dimensions to this correlation analysis: frequency, time and space.

Exploratory research on feature dimensions To determine which of those dimensions were relevant for our analysis we performed some exploratory research.

With regard to the frequency band, the previous studies used a high-gamma frequency band: 70-150 Hz Kubanek et al. and 70-170 Hz Potes et al.. To determine the frequency band that was optimal for the measuring of neural tracking we extracted frequency bands, in bins of 5 Hz, from 0 to 250 Hz, and correlated the envelopes of these frequency bands with the attended and unattended speech envelopes. The results of this analysis, across different ECoG channels (i.e.,cortical locations), can be found in Figure 5, for subject 1. Inspection of these channel x frequency plots across subjects, showed that generally the highest correlations were obtained somewhere in the range of 70-170 Hz, with additional positive correlations around the edges of this band, in some cases. Additionally, this analysis showed a negative correlation, between the attended speech envelope and the ECoG envelopes in the lower frequencies (5-30Hz). However, these lower frequencies did not appear to encode additional information in preliminary classification analyses and were therefore not included in our final analysis pipeline. Note that to perform this analysis, we already corrected for the temporal delay between the speech envelope and the ECoG envelope.

The temporal relationship between the presented speech envelope and the ECoG envelope is relevant, because there is a delay between this presentation and the cortical processing of this stimulus. To get an estimate of the delay between the presentation of speech and the cortical processing of this speech, we crosscorrelated the two signals with each other. This crosscorrelation calculates the correlation between the two signals, while one signal is shifted repeatedly in relation to the other. Where this crosscorrelation is the highest, the speech and the ECoG signal are optimally aligned, and from this the optimal shift can be identified. The existing research indicated that this delay can vary both per subject, and per channel within a subject. We therefore analyzed the crosscorrelations for the different EcoG channels, for each subject. The results of this analysis can be found, in Figure 3, for 2 of our subjects. These results show that while there is a variation in the delay across cortical locations and subjects, the peaks of this correlation are relatively wide (in the order of 100ms), and the correlation with the attended speech is on average larger than with the unattended speech, even if the crosscorrelations are not measure at their peak. From this we concluded that it would most likely be sufficient to correct for this delay across subjects and channels, rather than adding this as a parameter or additional feature dimension to our classification. This delay was set as a shift of 100ms of the ECoG signals for each subject and channel, relative to the speech envelope (see Figure 4.

For each subject, the brain acitivity was measured with multiple ECoG electrodes, distributed across the cortex. By calculating the correlation between the speech envelopes and the ECoG envelopes for each ECoG channel, we obtained a spatial distribution of neural tracking across the cortex. These results can be found in Figure 6, for each of the subjects. We included the different ECoG channels as a dimension in the classification features, so that the classifier could take advantage of the spatial patterns of neural tracking for identification of the attended speech.

Classification features Based on the exploratory research we analyzed the preprocessed ECoG data and the auditory data as follows:

From the preprocessed ECoG signals, we extracted the high gamma frequency by applying an 18th order 70–170 Hz Butterworth filter. We obtained the envelope of the filtered signals by taking the absolute values of the Hilbert transform of the filtered signal. Finally, we low-pass filtered the envelopes at 6 Hz and downsampled them from their original sampling rate of 1200 Hz to 120 Hz.

From the audio of the two speakers we extracted the envelope of the speech band (approximately 80–6000 Hz) by applying a 10th order 0.08–6 kHz Butterworth filter to the two audio signals. The envelope of the speech in this frequency band was obtained from the Hilbert transform of the filtered signal. We then also low-pass filtered these at 6 Hz and downsampled them from their original sampling rate of 44100 Hz to 120 Hz.

At this point, we corrected for the delay between the speech envelope and the audio envelope, by shifting the speech envelopes 12 samples (100 ms at 120 Hz), relative to the ECoG envelopes.

Because data from the experiment was obtained in trials of 15-25 seconds of length, and we were interested in analyzing the data at much shorter timescales, we segmented the data into trial segments, excluding the first 2 s of each trial. The first 2 seconds were excluded from this analysis to exclude any onset effects from tuning in to the target speaker. To be able to determine the relationship between stimulation length and the classification performance, we repeatedly applied segmentation process to obtain trial segments sets that varied in segment length from 100 ms to 10 s. For each trial segment, we have two speech envelopes A and B. For each speech envelope A, B, we have an assignment $\ell \in \{0, 1\}$, that denotes whether an envelope was attended (1) or unattended (0). For each speech envelope we extract a feature vector by calculating the spearman's rank correlation. This spearman's rank correlation was chosen over the standard Pearsons correlation as it is robust against outliers and was previously used in this context (Potes et al., 2012). The correlation for a speech envelope P and an ECoG envelope Q, is then defined as:

$$Corr(P, Q) = \frac{\sigma(p, q)}{\sigma(p, p)\sigma(q, q)} \tag{1}$$

where

$p = rank(P)$, $q = rank(Q)$, and $rank(x)$, is a function that ranks the samples in x from large to small.

$\sigma(x, y)$ calculates the covariance between variables x and y:

$$\sigma(x, y) = \frac{\sum_{i=1}^{n}(x_n - \bar{n})(y_n - \bar{y})}{N - 1} \tag{2}$$

with N as the number of samples in x (and y).

By taking the correlation for each trial segment and each ECoG channel, we obtain two feature matrices (R_A and R_B) of size T x C, where T is the total number of trials and C the number of ECoG channels. A feature matrix R_P with $P \in \{A, B\}$ is thus formally defined as:

$$R_P = \begin{pmatrix} Corr(P_1, Q_{1,1}) & \cdots & Corr(P_1, Q_{1,C}) \\ \vdots & & \vdots \\ Corr(P_T, Q_{T,1}) & \cdots & Corr(P_T, Q_{T,C}) \end{pmatrix} \tag{3}$$

Each of these feature matrices has a corresponding vector L of length T, that denotes for each trial segment if the correlations in the feature matrix from this trial segment were obtained from the attended ($\ell_t = 1$) or unattended speech envelope ($\ell_t = 0$).

Classification

The extracted features were used to predict which of the two speech envelopes was attended in a given trial. Because the number of channels used for classification would be relevant for future applications of this approach for BCI, we performed two classification analyses: one in which all channels were passed to the classifier, and one where only a single channel was used to predict which speech envelope the subject attended (a multivariate and a univariate classifier respectively).

Multivariate classification For the multivariate classification we used regularized (elastic net) logistic regression.

Given a vector of class labels y with $y \in \{1, 0\}$, and a matrix of features x, logistic regression estimates a vector of weights β, such that the following formula can be used to predict the probability that $y_i = 1$, for some an instance i out of all instances N.

$$p(Y = 1|x_i, \beta) = \frac{1}{1 + \exp^{-\beta^T x_i}} \tag{4}$$

The β-weights are estimated during training of the classifier, in which they are chosen such that the difference between the predicted class $p(Y = 1|x_i, \beta)$ and the actual class label y_i is minimized.

$$\min_{\beta} \sum_{n=1}^{N} -\log p(y^{(n)}|x^{(n)}, \beta) \tag{5}$$

Regularization of a classifier means to impose restrictions on the classifier weights, by including a penalty term on the size of the weights during estimation. For elastic net regularization this involves the combination of two other regularization approaches, a lasso regularization and ridge regression regularization. This elastic net regularization is defined as follows:

$$F_a(\beta) = \frac{(1-\alpha)}{2}||\beta||_2^2 + \alpha||\beta||_1 \tag{6}$$

Here the $||\beta||_2^2$ term is designed to encourage the weights to be 0 (lasso), while the $||\beta||_1$ encourages weights to be small (ridge regression). The α parameter controls to which degree these regularization terms individually affect the β weights. If $\alpha = 0$ this regularization method approaches ridge regression regularization, while if $\alpha = 1$, this regularization is effectively a lasso regularization. The degree of regularization is controlled by a λ parameter. In our elastic net regularized logistic regression this leads to the following estimation of weights.

$$\min_{\beta} \sum_{i=1}^{N} -\log p(y^{(n)}|x^{(n)}, \beta) + \lambda(\frac{(1-\alpha)}{2}||\beta||_2^2 + \alpha||\beta||_1) \tag{7}$$

For our purposes, the logistic regression was used to predict, for a given trial segment t, whether or not feature vector R_t was obtained from correlation with the attended speech of a given trial. We used regularization with this logistic regression, as the number of ECoG channels was larger than the number of channels that we expected to contribute to the prediction. Regularization effectively encourages the classifier to exclude channels from the classification, thereby reducing the chance that the logistic regression will produce an overfitted model. To estimate the β weights, a subset of the trials was assigned to training ($train \subset T$). The training set R, was then obtained by vertically concatenating $R_{A,train}$ and $R_{B,train}$. Analogously, the corresponding class label vector L was obtained by concatenating $L_{A,train}$ and $L_{B,train}$. The β weights were then estimated as follows:

$$\min_{\beta} \sum_{i=1}^{N} -\log p(\ell^{(n)}|R^{(n)}, \beta) + \lambda(\frac{(1-\alpha)}{2}||\beta||_2^2 + \alpha||\beta||_1) \tag{8}$$

Where N is the number of instances in R (= 2 * the total number of trials in *train*). α was set to 0.5, and λ was a hyperparameter that was estimated using a 5-fold crossvalidation across a range of λ-values (as the classification procedure includes a crossvalidation this was essentially the inner loop of a nested crossvalidation). This weight estimation was performed by a matlab implementation of regularized logistic regression (*glmlasso*, with a 'logit' link function).

The β weights obtained from this optimization, effectively weighted the contribution of each channel (i.e., cortical location) to the prediction. For testing, the two instances from the same trial segment t, R_{At} and R_{Bt} were paired. A prediction for each was obtained using the equation from (4):

$$p(y_t = 1|R_t, \beta) = \frac{1}{1 + \exp^{-\beta^T R_t}} \tag{9}$$

If the $p(y_t = 1|R_t, \beta)$ that corresponded to the attended speech envelope was larger than the $p(y_t = 1|R_t, \beta)$ that corresponded to the unattended speech envelope, thenthe trial segment t was considered classified correctly (if $p(y_t = 1|R_{At}, \beta) = p(y_t = 1|R_{Bt}, \beta)$, the trial segment would be considered classified correctly with probability 0.5).

Univariate classification For the univariate classification, the prediction of the attended and unattended stream relied on the assumption that the correlation of the ECoG envelopes with the attended speech envelope

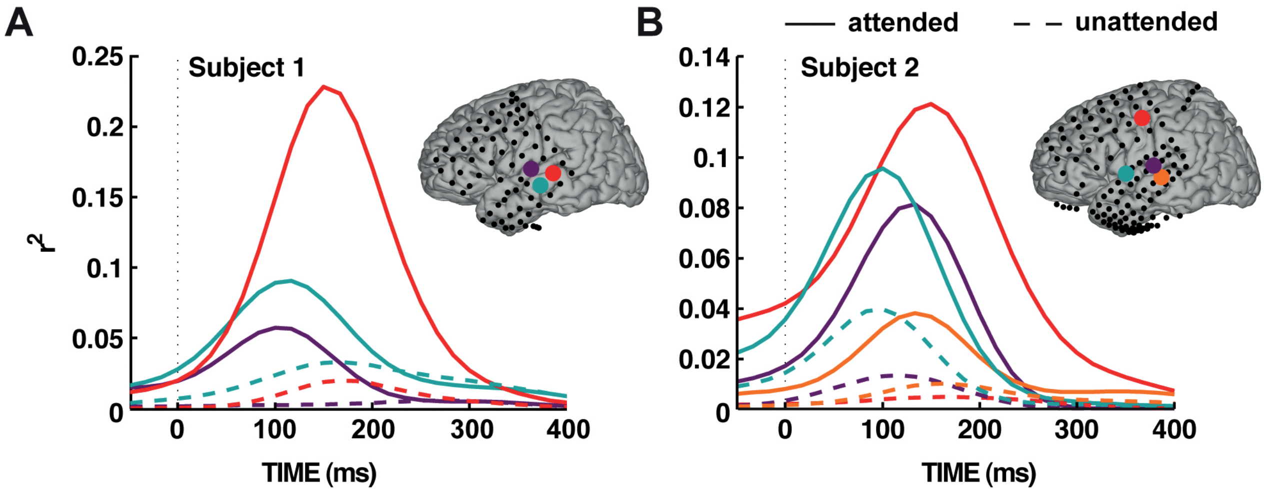

Figure 3: **Temporal relationship between speech envelope and neural signal.** The graphs show the crosscorrelation between the speech envelope and a subset of the ECoG envelopes, for two different subjects: **(A)** subject 1 and **(B)** subject 2. The brain models show the corresponding cortical locations. For these subjects the delay ranges between 100 to 200 ms.

would be higher than with the unattended speech envelope. The training of this classification consisted of the identification of the best channel c (i.e., the channel for which this assumption was most consistently true). As with the multivariate classification, the two instances (A and B) from the same trial segment t, for the channel c, R_{Atc} and R_{Btc} were paired. If the R_{tc} that corresponded to the attended speech envelope was larger than the R_{tc} that corresponded to the unattended speech envelope, the trial segment t was considered classified correctly (if $R_{Atc} = R_{Btc}$, the trial segment was considered classified correctly with probablity 0.5).

Classification procedure These two classification methods were applied to each of the sets of trial segments (varying in length from 100 ms to 10 s). The classification procedure consisted of 10 iterations of a 10-fold crossvalidation, resulting in 100 cross-validated performance values for each subject for each of the segmentation lengths. Note that the partitioning of the training and test set for the different crossvalidation folds ensured that two feature vectors for each trial segment, were either both in the training set or both in the test set.

To determine the significance of this performance, a randomization test was performed on the overall analysis for the 5 s segmentation length. The goal of such a randomization test is to verify that the ability of the classifier to predict the correct class is a result of the intended property of the data, and not some other factor in the data or property of the classification procedure. For the randomization test, the speech envelope vectors were reversed, effectively removing the temporal relationship between the speech and the neural envelopes, while keeping other properties intact (e.g., autocorrelation between the signals). To determine a distribution of random performance, we repeated this analysis 100 times on data for which the reversed envelope was shifted by random amounts of time. We then determined the likelihood (i.e., the p-value) that our cross-validated performance was different from random performance.

Results

Temporal relationship between speech presentation and neural response

As described in the 'feature extraction' section, to get an accurate measure of the neural tracking of the speech, it was necessary to correct for the delay between the presentation of the speech and the neural tracking. The extent of this delay was determined by calculating, for each subject and each channel, the cross-correlation between the ECoG envelopes and the attended and unattended speech envelope. Figure 3 shows these results for two different subjects, and a few selected channels. This figure shows that the delay is in the order of 100 to 200 ms and can vary per cortical location. For these subjects and the selected channels, the correlation with the unattended speech is generally smaller than the correlation with the attended speech, even when not correcting for the delay optimally.

For our classification analysis we opted for selecting a single delay correction across subjects and channels, as explained in the feature extraction section of 'Data Analysis'. The results in Figure 4 (blue trace) shows the correlation with the attended, averaged across all subjects. The results from a univariate classification for each subject (red trace), also peak at 100 ms.

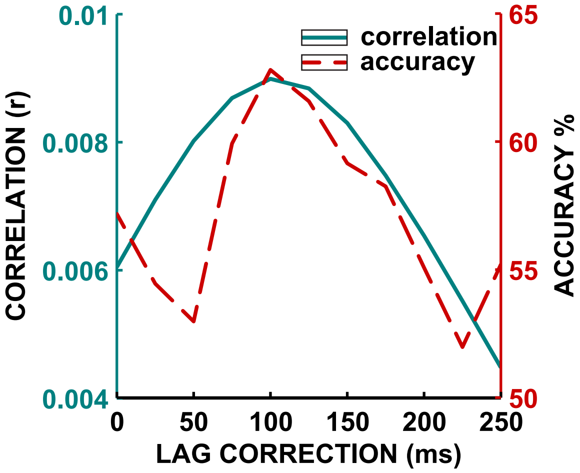

Figure 4: **Lag between speech presentation and neural response.** The correlation with the attended speech (blue) and classification accuracy (red) is shown, averaged across subjects, for latencies between 0 and 250 ms. Both correlation and univariate accuracy peak at 100 ms.

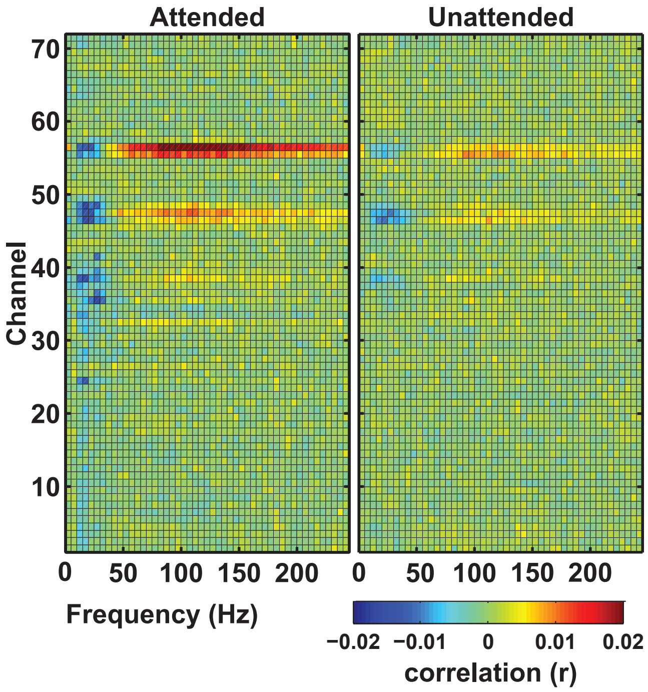

Figure 5: **Frequency response of neural tracking** The correlation of the ECoG envelopes with the attended and unattended speech envelopes is displayed in frequency bins of 5 Hz, from 0 to 200, for each of the ECoG channels of subject 1. The higher frequencies correlate positively with the attended speech, while the lower frequencies show a negative correlation. This effect is stronger for the attended speech, than it is for the unattended speech.

Frequency response for neural tracking of speech

To determine the frequencies in which neural tracking took place, we correlated the attended and unattended speech envelopes with the ECoG envelopes in different frequency bands (bands of 5 Hz between 0 and 250 Hz). Figure 5 shows the correlation with the attended and unattended speech envelopes for the different channels or cortical locations for each of the frequency bins, for subject 1. This figure shows a few different effects: First, there are only a number of channels for which a response can be identified (with the 3 main channels corresponding to those from Figure 3). Secondly, there appear to be two different responses across frequencies: one, a negative correlation with the speech envelope in the low frequencies (approximately 5-30 Hz), which is stronger and distributed across more channels for the attended speech, and two, a positive correlation in the high frequencies, that is similarly more prominent for the attended speech. The strongest correlations in this frequency range, for the responsive channels, correspond roughly to the chosen frequency band of 70-170 Hz, surrounded by some additional frequencies, especially for channel 56.

Spatial distribution of neural tracking of speech

To get an insight on the cortical locations that track speech, we displayed the correlations for each of the cortical locations onto the 3D template brain model (previously shown in Figure 1). For this we used an existing matlab package that, instead of simply plotting a value for each electrode, projects these values back onto the cortical surface as an estimation of the original cortical areas that contributed to the activity measured by a given electrode (Kubanek et al., 2013).

The results in Figure 6 show the neural tracking of the attended (•) and unattended speech (◦), for each subject. This neural tracking is represented as an activation index that expresses the correlation between the high gamma ECoG envelope and the speech envelope, for each cortical location. For this analysis the correlations were determined across all trials.

These projections show that the neural tracking is focused around two cortical areas. The first cortical area is the Superior Temporal Gyrus (STG), which shows activation in all 12 subjects. An additional cortical area, the superior pre-motor cortex, is activated in only a number of subjects (i.e., in subjects 2, 3, 4 and 8).

A comparison between the attended and unattended activation indices indicates that neural tracking is selective to the attended speech envelope, as the activation index is generally stronger and more widely distributed for the attended speech.

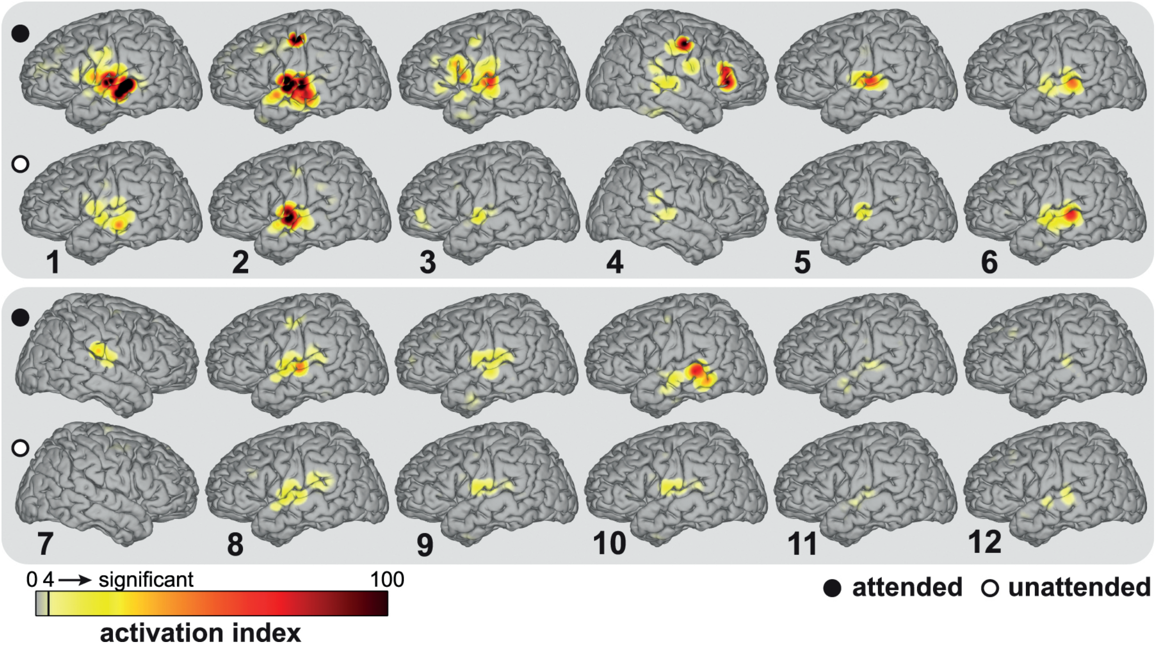

Figure 6: **Neural tracking of attended (•) and unattended (◦) speech.** Neural tracking is measured as the correlation between the high gamma ECoG envelope and the attended or unattended speech envelope. An activation index that expresses the $-\log(p)$ of this correlation is projected on the brain model for each subject. The comparison between the attended and unattended activation indices shows that neural tracking is selective to the attended speech envelope, as the activation index is generally stronger and more widely distributed in this case.

Identification of attended speech

The main question of this thesis concerns the identification of the attended speech in single trials. Figure 7A shows results from the classification of single trials segments. Here, the classification accuracy is shown for each subject, for both the univariate (blue) and the multivariate (orange) classification method. This accuracy was obtained from averaging over 4, 5, and 6 second trial segments, to give a stable estimate of performance. The subjects are presented in the order of their average classification accuracy. Performance significantly better than chance is indicated with an asterisk (determined through the randomization test, adjusted for multiple comparisons by using a false discovery rate with $q = 0.05$).

For 5 subjects, both methods could predict the attended speech better than chance, and for 2 other subjects only the univariate performance performed at chance level. For the remaining 5 subjects neither method achieved significant performance. From here on, those first 7 subjects will be referred to as 'significant subjects', while the latter will be referred to as 'non-significant subjects'. It is useful to make a distinction between these two groups, not to simply ignore these non-significant subjects, but to draw comparisons across results for these two groups.

Figure 7B shows the average classification accuracy across significant subjects only. For these subjects, a comparison between the two methods shows an 11% higher classification accuracy for multivariate regression compared to univariate regression (70% vs. 81%, paired t-test: $p < 0.0003$).

To tie these results to the results from the previous section, we averaged the activation index topographies across significant and non-significant subjects. These topographies are shown in Figure 8. In these results, the significant subjects (Figure 8A) show a stronger and more distributed response to the attended (•) than to the unattended speech (◦). In contrast, non-significant subjects (Figure 8B) show only a marginal difference in their response to the attended (•) compared to the unattended speech (◦).

Relationship between trial segment length and classification accuracy

The classification analysis was performed on trial segments ranging in length from 100 ms to 10 s, to determine the relationship between the trial segment length and the classification accuracy. The results in Figure 9 show the classification accuracy for the different trial lengths for significant subjects (Figure 9A) and non-significant subjects (Figure 9B). For the significant subjects, the univariate (blue) and multivariate (orange) accuracy rises steadily, and for the multivariate classfier the performance reaches 86.5% for 10 s segments. Throughout the investigated trial length,

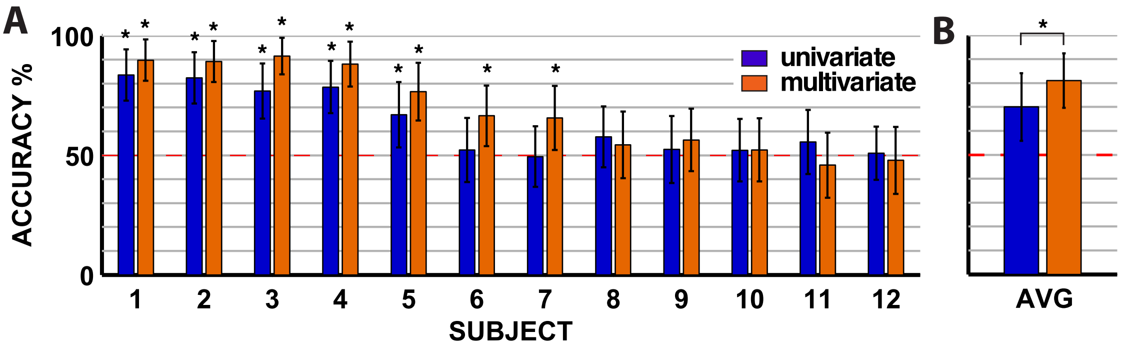

Figure 7: **Classification accuracy using a univariate (blue) or multivariate (orange) classification method, for trial segments 4-6 s in length.** **(A)** Accuracy per subject, sorted by average performance. For subjects 1-7 ('significant subjects') the accuracy is significant for at least one classification 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. A comparison between the two methods shows an 11% higher classification accuracy for multivariate regression compared to univariate regression (70% vs. 81%, paired t-test: $p < 0.0003$).

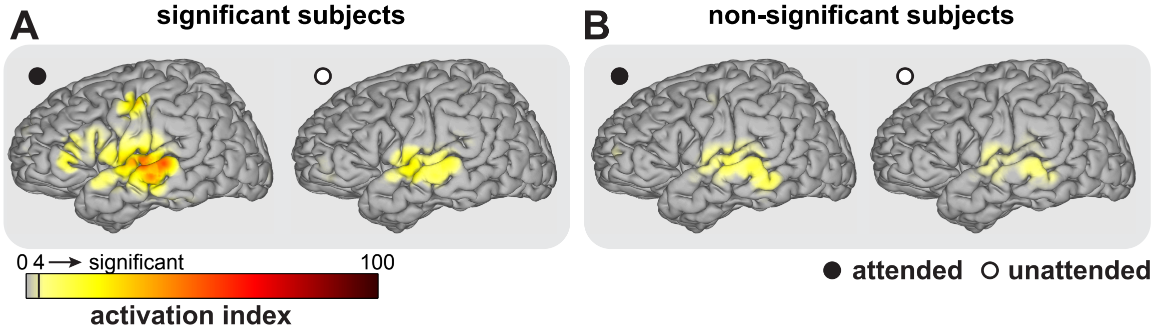

Figure 8: **Cortical tracking of attended (•) and unattended (◦) speech.** Two averages are displayed: **(A)** Subjects for which performance was significantly better than chance for at least one classification method and **(B)** subjects for which performance was at chance level. For the significant 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.

the ~10% advantage of the multivariate over the univariate classification method persists. Classification for the non-significant subjects stays around chance level for both classification methods, indicating that the non-significant results obtained on the 5 s segments are not specific to that segment length.

The results show a clear relationship between trial segment length and accuracy, where for longer trial segments it is easier to predict the attended speech. However, while classification accuracy is important for a BCI application, there is usually a trade-off in the amount of information transferred between the classification accuracy and the length of the stimulation.

The Information Transfer Rate (ITR) of a BCI measures the amount of information (measured in bits) that is transferred by the user to the system. A frequently used definition is that by WolpawWolpaw et al. (1998).

$$ITR = \log_2 N + P \log_2 P + (1-P) log_2 \frac{1-P}{N-1} \qquad (10)$$

Where N is the number of classes (here, N = 2) and P is the probability of an intended class being classified correctly (i.e., accuracy).

This ITR is measured in bits per trial, which we normalized by the length of the segment, so that we can compare the amount of information transferred per time unit.

From this formula we can identify three factors that affect the ITR: the probability of identifying the correct class, the number of classes and the length of the stimulus.

The results in Figure 10 show the bit rate across segment lengths for the two different classifiers. For the multivariate classifier, ITR reaches 6.2 bits/min for 1.5 s

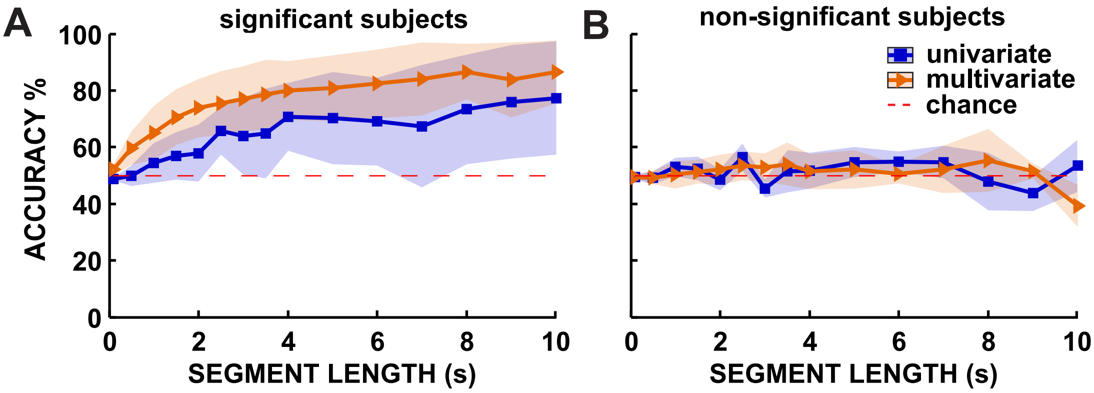

Figure 9: **Classification accuracy for different segment lengths for univariate (blue) and multivariate methods (orange).** Results are shown for **(A)** the significant subjects and **(B)** the non-significant subjects. For the significant subjects, the classification accuracy increases steadily with segment length for both classification methods. Throughout the investigated trial length, the ∼10% advantage of the multivariate over the univariate classification method persists.

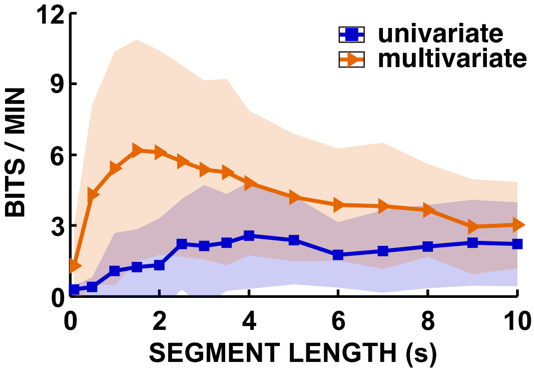

Figure 10: **Relationship between segment length and the Information Transfer Rate (ITR).** For the multivariate classifier, ITR reaches 6.2 bits/min for 1.5 s long segments. The the univariate classifier achieves its maximum ITR at 4 s segments, for an ITR of less than 3 bits/min.

long segments. These results illustrate that from a BCI perspective, a longer stimulus is not always better. The univariate classifier achieves less than half of this ITR and requires it requires segments twice as long to obtain its maximum (maximum at 2.6 bits/min at 4 s segment length).

Effect of ‘tuning in’ on correlation and classification accuracy

For the classification analysis we excluded the first 2s of each trial to avoid any effects of a ‘tuning-in’ period. Such a ‘tuning in’ period would be relevant for determining the overall ITR of a specific BCI. To estimate the actual ‘tuning in’ period we analyzed the trials from the experiment with a sliding window of 1 s, overlapping by 900 ms, from one second before the start to 10 s into the trial.

The results in Figure 11 show the difference in correlation of the neural signal with the attended and unattended speech, (11A) and classification accuracy (11B; non crossvalidated). Each data point reflects the correlation or accuracy of the preceding second of data. For this analysis, only those subjects for which we had previously established a significant performance for the univariate method (i.e., subjects 1-5), were taken into account.

This figure shows that at the onset of the trial, there is no difference in correlation, and the accuracy is at chance level. After about a second the correlation difference and the accuracy rise sharply, indicating that it takes on average about a second after onset of speech for the neural tracking of the attended speech to become detectable. An unexpected effect is the variation in correlation and accuracy across the trial, as we would expect these measures to become stable as the trial goes on.

Effect of laterality on classification accuracy

In our analysis, we did not consider any potential effects that the aural location of presentation may have on the neural response and the ability to identify the attended speech. Such an effect could potentially exist, as we recorded data from only a single hemisphere in each subject, and there are known lateralization effects in auditory processing. For example, stimuli are processed more strongly in the hemisphere contra-lateral to their presentation (Woldorff et al., 1999). For that reason, we determined whether trials presented ipsi-laterally to the implant could be classified as well, as trials presented contra-laterally to the implant.

The results in Figure 12 show the classification accuracy for univariate (top) and multivariate (bottom)

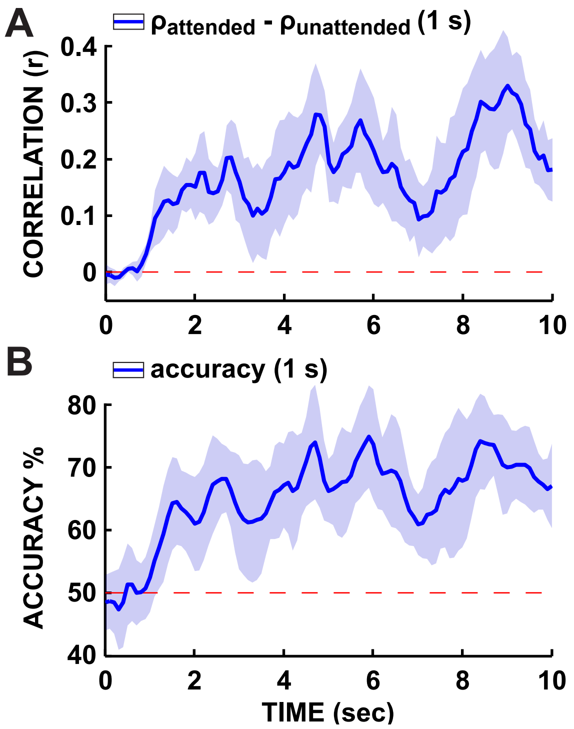

Figure 11: **Effect of 'tuning in' on correlation and classification accuracy.** The development of measurements of neural tracking during a trial measured as, **(A)** the difference in correlation of the neural signal with the attended and unattended speech, and **(B)** the corresponding univariate classification accuracy.

classification. For each subject, this figure shows the accuracy for ipsi-lateral (left) and contra-lateral trials (right). The results indicate that, on average, there is no statistically significant difference between ipsi- and contra-lateral classification accuracy, for the two classification methods (paired t-test: p = 0.5 and p = 0.8, respectively). Additionally, we performed a within-subject analysis to determine if there was a significant difference between these for any subjects. To analyze this we partitioned the data repeatedly into two random sets, to which we applied our classifiers and recorded the performance difference between the sets. The distribution of these classification differences was compared to the difference in performance obtained for the ipsi versus contra-lateral classification accuracy. This resulted in one statistically significant difference for one subject (subject 3), for the univariate analysis. While there are observable differences in the univariate classification rates for other subjects (e.g. subjects 1, 4 and 5), and these results were significantly different in a paired t-test, these results were not significant in this randomization test. This indicates that these differences are a result of the variability in the data samples. This is potentially due to the low number of training examples that are available to the classifier after dividing the available data into two sets.

Discussion

With these results we can now answer the research questions stated at the start of this thesis.

Q1: Can the attended speech be identified from the brain activity, in single trials, with better than chance performance? For 7 out of 12 participants we can correctly identify the attended speech with better than chance accuracy. For the 5 remaining participants, the attended speech could not be identified, regardless of the length of the trial segments. For these subjects there was also little difference in the correlations of the ECoG with the attended and unattended speech (see Figure 8. This indicates that this is not just a failure of our classification method to pick up on any effects.

There are two explanations we offer for this effect: Firstly, it is possible that subjects were not able or did not perform the task consistently. We cannot exclude this possibility, as the experiment contained no behavioral measure to ascertain that subjects attended the target speaker and understood the content. A similar study by Mesgarani and Chang (2012), did use a behavioral verification and reported that across subjects ~25% of the trials were not attended. Subsequent analyses across these trials only showed that these did not exhibit a neural tracking of the attended speech. Secondly, there is a variation in electrode coverage between subjects. While it is unlikely that the neural tracking is so sensitive that, given that all subjects had cortical coverage of the STG, the electrodes for these 5 subjects did not capture the neural tracking, it is possible that this was a factor that played a role.

Q1a: Is this performance sufficient for communication (>=70%)? Across the 7 subjects for whom the attended speech could be identified, the average performance across 4-6 s trial segments was 81%. This would qualify as performance sufficient for communication. However, the performance obtained in our ECoG study did not reach the communication performance reported in some other auditory attention (EEG) BCI studies. For example, the previously discussed BCI by Hill and Scholkopf (2012) achieved an online performance of 84.8% for 5 s trials, with an ITR of 4.98 bits/min (± 2.3). In our study, we obtained a performance of 81% and an ITR of 4.2 bits/min (± 2.7) for the significant subjects. However, Hill and Scholkopf did not exclude any subjects from their analysis. 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).

While there are advantages to using natural speech instead of the artificial stimuli streams from their ex-

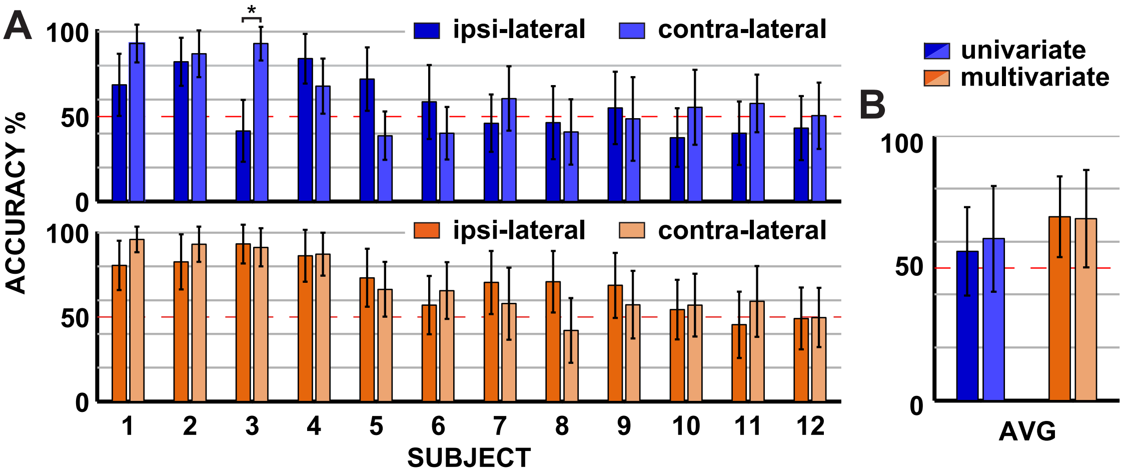

Figure 12: **Classification accuracy for auditory attention to the ear ipsi- and contra-lateral to the cortical implant.** The results from Figure 7 are shown for trials in which the attended speech stimulus was presented either ipsi- or contra-laterally to the cortical implant. **(A)** For a univariate approach, the aural location affects the classification accuracy for some subjects. **(B)** A multivariate regression is less susceptible to this effect. **(C)** On average, there is no clear relationship between the laterality of the aural location and the cortical implant (p = 0.5 and p = 0.8, respectively).

periment, it remains a question whether this approach should be pursued if it reaches lower communication performances, even when using a more sensitive measure of brain activity (ECoG). At the same time, it is important to note that their study was performed with healthy subjects. In contrast, the only subjects eligible for this type of ECoG research are those under medical care. This has the consequence, for instance, that subjects are frequently on medication during the experiment. To exclude the possibility that this method performs poorly in comparison due to difficulty with performing the task, the study should be repeated with a behavioral verification.

Additionally, the approach from Hill and Scholkopf does not scale easily beyond two simultaneous stimuli, as the tones between two streams should not overlap. As the variability across different speech stimuli is fairly uncorrelated, it is easier to increase the number of simultaneous stimuli for natural speech stimuli than it is for structured or altered stimuli that are correlated with each other. However, it remains to be determined how communication performance would scale with this increase.

Finally, in our 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 (McFarland et al., 1998; Miller et al., 2010). This indicates that performance may increase with an online implementation.

Q1b: What is the minimum length of stimulation required for this performance? For the 'significant' subjects, the multivariate classifier obtains an average accuracy of 70% with 2 s long segments. This shows that, even with very short segments, it is possible to identify the attended speech reliably. Additionally, the analysis of the ITR across trial segment lengths showed that the optimal trade-off between segment length and classification accuracy, was with segments shorter than 2 seconds (1.8 s). It should be noted that with the 'tuning in' effects, that we determined to be in the order of 1 s, this would require a total stimulation of 2.8 s.

Q1c: How well can the attended speech be predicted, when only a single ECoG channel (i.e.,cortical location) is taken into account? Univariately, the attended speech could be identified correctly with better than chance performance for only 5 out of the 7 'significant' subjects. The accuracy across these subjects, for trial segments 4-6 s in length, was approximately 77% (70%, with the two other 'significant' subjects included). For comparison the average multivariate performance for these 5 subjects alone, was approximately 90%. The results between the univariate and multivariate method were thus relatively small. This indicates that a single electrode might be a viable option for this approach. The location of this electrode could be determined using pre-operative fMRI (Vansteensel et al., 2010).

Another note relevant to the invasiveness of this approach: the electrodes in this study were placed subdurally (i.e., electrodes placed underneath the dura). Penetration of the dura increases the risk of bacterial infection (Davson, 1976; Hamer et al., 2002; Fountas and Smith, 2007; Van Gompel et al., 2008; Wong et al., 2009). Epidural electrodes (i.e., electrodes placed on top of the dura) provide signals of approximately comparable fidelity (Torres Valderrama et al., 2010). A single electrode placed epidurally could reduce risk and cost, making this approach more realistic despite its invasive nature.

Consistency with existing neuroscience

In our study, we found that neural tracking of the attended speech is stronger and more widely distributed across the cortex, than that of unattended speech. This confirms findings by Zion Golumbic et al. (2013), who investigated auditory attention in ECoG. We further found locations over the STG and the pre-motor cortex to be informative of the attended speech. The role of the pre-motor cortex in speech perception confirms similar findings from Wilson and Iacoboni (2006); Potes et al. (2012, 2014), who investigated the neural tracking of non-simultaneous speech. Although speech tracking of the pre-motor cortex was only visible in 4 out of 12 subjects (see Figure 6, it is of specific interest, as this pre-motor tracking was especially selective of the attended speech, with only very limited tracking of unattended speech detected in these areas. This would make it a good target for single electrode applications, if this pre-motor area could be determined beforehand in pre-operative fMRIs.

In our study, we determined the delay between the speech stimuli and the elicited neural signal, across all subjects and all electrodes. However, an analysis of the individual traces that were averaged (see Figure 4) revealed a standard deviation of 72 ms in this delay for univariate performance and a standard deviation of 58 ms for the correlation across subjects. In addition, our exploratory results (see Figure 3, indicated that there is a variance between subjects and across subjects. This is confirmed by results, shown by Potes et al. 2012, who reported a 110 ms delay between the neural tracking over the STG and the pre-motor cortex. For our study we did not include this information during classification, as we expected small returns. Subsequent studies could explore whether correcting the delay per ECoG electrode, for each subject, could improve communication performance.

Other research has suggested that low-frequency phase information might encode additional information about selective auditory attention (Zion Golumbic et al., 2013). 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 cocktail party BCI.

Future Research

A number of avenues for follow up research have been discussed above. To reiterate, the main follow-up-questions we identified are:

- Do we find this subject variability (failure to identify the attended speech for 40% of the subjects) in future instances of this experiment. And if so, can this be explained by a lack of attention to the target speaker (verified with a behavioral measure)?

- How does the classification performance scale, when the number of simultaneously presented speakers is increased?

- Can this approach be extended to an online application?

- Does the inclusion of additional information for the classifier (e.g. low-frequency phase information or delays optimized for each ECoG channel) lead to better classification accuracy?

Summary

To summarize, our study shows that an auditory attention based BCI that uses natural speech stimuli in a cocktail party setting could provide reasonable communication performance. While our results compare unfavorably with results obtained from existing auditory attention paradigms, this is mainly due to the large number of subjects for whom the attended speech could not be identified. Future research should investigate the possible causes of this, in order to derive a more concrete conclusion regarding the viability of natural language attention as a BCI pardigm. In other words, this provides the groundwork for future studies that would explore the usability of this approach for BCI applications in people affected by severe neuro-degenerative diseases.

References

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., and Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences*, 98(23):13367–13372.

Belitski, A., Farquhar, J., and Desain, P. (2011). P300 audio-visual speller. *J Neural Eng*, 8(2):025022.

Brouwer, A. M. and van Erp, J. B. (2010). A tactile P300 brain-computer interface. *Front Neurosci*, 4:19–19.

Brunner, P., Joshi, S., Briskin, S., Wolpaw, J. R., Bischof, H., and Schalk, G. (2010). Does the 'P300' speller depend on eye gaze? *J Neural Eng*, 7(5):056013.

Brunner, P., Ritaccio, A. L., Emrich, J. F., Bischof, H., and Schalk, G. (2011). Rapid communication with a "P300" matrix speller using electrocorticographic signals (ECoG). *Front Neurosci*, 5(5).

Davson, H. (1976). Review lecture. The blood-brain barrier. *J Physiol*, 255(1):1–28.

Fountas, K. N. and Smith, J. R. (2007). Subdural electrode-associated complications: a 20-year experience. *Stereotact Funct Neurosurg*, 85(6):264–272.

Furdea, A., Halder, S., Krusienski, D. J., Bross, D., Nijboer, F., Birbaumer, N., and Kübler, A. (2009). An auditory oddball (P300) spelling system for brain-computer interfaces. *Psychophysiology*, 46(3):617–625.

Halder, S., Rea, M., Andreoni, R., Nijboer, F., Hammer, E. M., Kleih, S. C., Birbaumer, N., and Kübler, A. (2010). An auditory oddball brain-computer interface for binary choices. *Clin Neurophysiol*, 121(4):516–523.

Hamer, H. M., Morris, H. H., Mascha, E. J., Karafa, M. T., Bingaman, W. E., Bej, M. D., Burgess, R. C., Dinner, D. S., Foldvary, N. R., Hahn, J. F., Kotagal, P., Najm, I., Wyllie, E., and Lüders, H. O. (2002). Complications of invasive video-EEG monitoring with subdural grid electrodes. *Neurology*, 58(1):97–103.

Hill, N. J. and Scholkopf, B. (2012). An online brain-computer interface based on shifting attention to concurrent streams of auditory stimuli. *J Neural Eng*, 9(2):026011.

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(2):620–628.

Klobassa, D. S., Vaughan, T. M., Brunner, P., Schwartz, N. E., Wolpaw, J. R., Neuper, C., and Sellers, E. W. (2009). Toward a high-throughput auditory P300-based brain-computer interface. *Clin Neurophysiol*, 120(7):1252–1261.

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., and Schalk, G. (2013). The tracking of speech envelope in the human cortex. *PloS one*, 8(1):e53398.

Kübler, A., Neumann, N., Kaiser, J., Kotchoubey, B., Hinterberger, T., and Birbaumer, N. P. (2001). Brain-computer communication: Self-regulation of slow cortical potentials for verbal communication. *Arch Phys Med Rehabil*, 82(11):1533–1539.

Lopez-Gordo, M. A., Fernandez, E., Romero, S., Pelayo, F., and Prieto, A. (2012). An auditory brain–computer interface evoked by natural speech. *J Neural Eng*, 9(3):036013.

Martin, S., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N. E., Rieger, J., Schalk, G., Knight, R. T., and Pasley, B. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. *Front Neuroeng*, 7(14).

McFarland, D. J., McCane, L. M., and Wolpaw, J. R. (1998). EEG-based communication and control: short-term role of feedback. *IEEE Trans Rehabil Eng*, 6(1):7–11.

Mellinger, J. and Schalk, G. (2007). BCI2000: A general-purpose software platform for BCI. In Dornhege, G., del R. Millan, J., Hinterberger, T., McFarland, D., and Müller, K., editors, *Toward Brain-Computer Interfacing*, pages 359–367, Cambridge, MA, USA. MIT Press.

Mesgarani, N. and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233–236.

Miller, K. J., Schalk, G., Fetz, E. E., den Nijs, M., Ojemann, J. G., and Rao, R. P. (2010). Cortical activity during motor execution, motor imagery, and imagery-based online feedback. *Proc Natl Acad Sci U S A*, 107(9):4430–4435.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., and Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS biology*, 10(1):e1001251.

Potes, C., Brunner, P., Gunduz, A., Knight, R. T., and Schalk, G. (2014). Spatial and temporal relationships of electrocorticographic alpha and gamma activity during auditory processing. *NeuroImage*.

Potes, C., Gunduz, A., Brunner, P., and Schalk, G. (2012). Dynamics of electrocorticographic (ECoG) activity in human temporal and frontal cortical areas during music listening. *NeuroImage*, 61(4):841–848.

Riccio, A., Mattia, D., Simione, L., Olivetti, M., and Cincotti, F. (2012). Eye-gaze independent EEG-based brain-computer interfaces for communication. *J Neural Eng*, 9(4):045001.

Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N., and Wolpaw, J. R. (2004). BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Trans Biomed Eng*, 51(6):1034–1043.

Schalk, G. and Mellinger, J. (2010). *A Practical Guide to Brain-Computer Interfacing with BCI2000*. Springer, London, UK, 1st edition.

Schreuder, M., Blankertz, B., and Tangermann, M. (2010). A new auditory multi-class brain-computer interface paradigm: spatial hearing as an informative cue. *PLoS One*, 5(4).

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304.

Talairach, J. and Tournoux, P. (1988). *Co-Planar Sterotaxic Atlas of the Human Brain*. Thieme Medical Publishers, Inc., New York.

Torres Valderrama, A., Oostenveld, R., Vansteensel, M. J., Huiskamp, G. M., and Ramsey, N. F. (2010).

Gain of the human dura in vivo and its effects on invasive brain signal feature detection. *J Neurosci Methods*, 187(2):270–279.

Treder, M. S. and Blankertz, B. (2010). (C)overt attention and visual speller design in an ERP-based brain-computer interface. *Behav Brain Funct*, 6(1):28–28.

van der Waal, M., Severens, M., Geuze, J., and Desain, P. (2012). Introducing the tactile speller: an ERP-based brain-computer interface for communication. *J Neural Eng*, 9(4):045002.

Van Gompel, J. J., Worrell, G. A., Bell, M. L., Patrick, T. A., Cascino, G. D., Raffel, C., Marsh, W. R., and Meyer, F. B. (2008). Intracranial electroencephalography with subdural grid electrodes: techniques, complications, and outcomes. *Neurosurgery*, 63(3):498–505.

Vansteensel, M. J., Hermes, D., Aarnoutse, E. J., Bleichner, M. G., Schalk, G., van Rijen, P. C., Leijten, F. S., and Ramsey, N. F. (2010). Brain-computer interfacing based on cognitive control. *Ann Neurol*, 67(6):809–816.

Wada, J. and Rasmussen, T. (1960). Intracarotid injection of sodium amytal for the lateralization of cerebral speech dominance. *J Neurosurg*, 17:266–282.

Wilson, S. M. and Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception. *NeuroImage*, 33(1):316–325.

Woldorff, M. G., Tempelmann, C., Fell, J., Tegeler, C., Gaschler-Markefski, B., Hinrichs, H., Heinze, H.-J., and Scheich, H. (1999). Lateralized auditory spatial perception and the contralaterality of cortical processing as studied with functional magnetic resonance imaging and magnetoencephalography. *Hum Brain Mapp*, 7(1):49–66.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clin Neurophysiol*, 113(6):767–791.

Wolpaw, J. R., Ramoser, H., McFarland, D. J., and Pfurtscheller, G. (1998). EEG-based communication: Improved accuracy by response verification. *IEEE Trans Rehab Engin*, 6:326–333.

Wong, C. H., Birkett, J., Byth, K., Dexter, M., Somerville, E., Gill, D., Chaseling, R., Fearnside, M., and Bleasel, A. (2009). Risk factors for complications during intracranial electrode recording in presurgical evaluation of drug resistant partial epilepsy. *Acta Neurochir (Wien)*, 151(1):37–50.

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D., and Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a Cocktail party. *Neuron*, 77(5):980–991.