Understanding the neural basis of speech production using Machine Learning

Otilia Stretcu

November 15, 2017

Background. Understanding how neurons act together to produce speech is still an open problem. Several studies have attempted to decode different aspects of speech from the cortex neural activity, while a person is speaking [1, 8], but evidence from the medical domain [2] suggests that other brain regions, such as the subthalamic nucleus (STN), may also be involved in speech production.

Aim. We explore the problem of decoding properties of speech (e.g. volume, manner, used articulators) from neural activity recordings. From a neuroscience perspective, the goal is to understand which properties are encoded in different parts of the cortex and STN. From a machine learning (ML) perspective, we are interested discover what models are best at extracting relevant information from the scarce and noisy neural activity data.

Data. The brain signals are recorded from 14 human subjects, while reading out loud words. The data consists of ECoG recordings from the ventral primary motor and primary sensory cortical areas, and Local Field Potentials from the STN.

Proposed Approach. We approach several decoding tasks: (1) predicting when a person is speaking or not, from their neural activity, (2) predicting the manner of speech (e.g. nasal, plosive) and what articulators (e.g. tongue, lips) are used, and (3) predicting the volume/loudness of the speech. For each of these tasks, we apply a series of ML methods, from simple regression models (e.g. ridge regression) to deep learning models (e.g. recurrent neural networks). We apply these models on different levels of preprocessing of the data, from electrode signals in time domain to particular frequency bands. The goal is to understand which models are able to discover interesting patterns, with different levels of domain knowledge required for preprocessing. Finally, we use our best models to discover which areas of the brain encode different kinds of information about speech production.

Results. Our analysis shows that machine learning models are able to discover different speech features from the neural activity. We were able to classify when a subject is speaking or not from their neural activity in the primary motor and primary sensory cortex with up to 96% accuracy, and up to 80% accuracy from the STN (an area whose connection to speech production is not entirely understood). We were also able to decode certain features of speech (e.g. voicing, manner) and inspect the brain regions and time points that contribute to the prediction. From a ML perspective, we observe that even simple models overfit easily on our dataset due to the low-sample, high-dimensionality problem, and that parameter tuning and proper regularization methods are crucial in making accurate predictions. Finally, we recommend neural network based models if time and computation resources are available for tuning the parameters, and simple models otherwise.

Broader impacts. Our work has important practical applications in the medical domain. For example, a neural decoder can be used in neuroprosthetics to enable communication for the impaired. Moreover, understanding the involvement of the STN in speech can improve the deep brain stimulation techniques used for treating Parkinson's disease patients. More generally, our work provides an useful overview of which ML models are suitable in dealing with different modalities of brain data, which can facilitate further neuroscience studies.

Keywords: neuroscience, speech production, ECoG, Local Field Potentials, machine learning.

1 Introduction

Understanding how the brain controls speech is an open problem in neuroscience, with far reaching consequences for the betterment of humanity. So far, we know some parts of the brain that are involved when producing speech (there is evidence that the motor cortex has areas specialized for tongue, lips, etc.). However, the details of which exact locations in the brain control different articulators, or different properties of speech (e.g. loudness, pitch) are still not entirely understood. Succeeding at such a task would not only bring science closer to understanding how speech production is controlled by the brain, but it would also have important consequences in medical applications. For example, a neural decoder can be used in neuroprosthetics to enable communication for the impaired by translating brain signals into spoken language.

Our work, however, has been driven by a different problem: speech production is disrupted in several neurological diseases, including Parkinson's disease (PD). PD is the second most common neurodegenerative disease in America [5]. Although it is mostly known as a motor disorder, PD also affects speech production and control. A common method of treatment that does not have the side effects of drug medications is deep brain stimulation (DBS) [3]. DBS consists of implanting electrodes in a certain area of the brain, that produce electrical impulses which aim to regulate the abnormal neural impulses. In PD, electrical stimulation of the Subthalamic Nucleus (STN) has shown major reductions of the tremors, allowing PD patients to live a more normal life. However, neither DBS, nor other medication are able to adequately treat the speech disruption in PD. In fact, one of the most common side effects of DBS is a decrease in verbal fluency [6, 7, 2]. This work is part of a joint project between University of Pittsburgh, Carnegie Mellon University, and Johns Hopkins University, with the goal of understanding how the STN is involved in the production of speech, and thus facilitating more targeted treatment for speech disorders.

The dataset used in this project was collected at the University of Pittsburgh and Johns Hopkins University, and it is the first dataset that contains simultaneous recordings from the cortex (ECoG) and STN (local field potentials and micro-electrode recordings) while the subjects are speaking. This allows us to apply machine learning methods on multiple data modalities, and look for patterns of speech both at cortex level and at STN level.

Therefore, the goal of this project is twofold. From a neuroscience perspective, the goal is to understand what properties of speech production are encoded in different parts of the cortex and STN. From a machine learning perspective, we are interested discover what models are best at extracting relevant information from the scarce and noisy brain data, with as little domain knowledge supervision as possible.

We approach a series of decoding tasks: (1) predicting when a person is speaking versus not speaking, from the neural activity, (2) predicting the manner of speech (e.g. nasal, plosive) and what articulators (e.g. tongue, lips) are used, (3) predicting the volume/loudness of the speech. Prior work in the field has shown that different areas of the sensorimotor cortex show an increase in power in the high γ frequency band (85 - 175 Hz) during speech [1], measured using ECoG. Furthermore, [1] illustrate that different cortex regions show a difference in power depending on the used articulator. [8] also show that it is possible to classify accurately spoken phonemes from the sensorimotor cortex, and that the activity patterns in the sensorimotor cortex reflect the sequences of muscle contractions during speech. In our work, we aim to validate these results and extend them to the study of more linguistic features, as well as investigate if similar results can be discovered in the STN. To the best of our knowledge, no other work has attempted these analyses on STN data.

Also different from prior research, for each of these tasks, we employ a series of machine learning

methods, from simple regression models (e.g. ridge regression) to deep learning models (e.g. recurrent neural networks, fully connected network). We apply these models on different levels of preprocessing of the data, from raw electrode signals to particular frequency bands in the spectrogram. The goal is to understand which models are able to discover interesting patterns in the data, with different levels of domain knowledge required for preprocessing. Finally, we use our best models to make discoveries about the neural mechanisms of speech production, thus hoping to push the field of neuroscience further.

2 Problem statement

In this project, we investigate how the neural activity in the cortex and subthalamic nucleus is involved in speech production. We approach this problem as machine learning prediction task: what features of speech (e.g. volume, used articulators, manner) can be predicted from the neural activity in the primary motor and primary sensory cortex (measured using ECoG), and in the subthalamic nucleus (measured using local field potentials)?

3 Data

Subjects. The data used in our analysis comes from 14 human subjects suffering from Parkinson's disease, and was collected during a Deep Brain Stimulation (DBS) implant procedure. All participants provided written, informed consent in accordance with a protocol approved by the Institutional Review Board of the University of Pittsburgh (IRB Protocol #PRO13110420).

Stimuli content. During the experiment, the subjects were asked to read a set of 120 stimuli from a computer screen and speak them out loud, one by one. All stimuli are either English words or non-words that consist of 3 phonemes in the order *consonant - vowel - consonant* (CVC), such as *fought* (read as */fɔt/*) or *van* (read as */væn/*).

Stimuli presentation. The data collection is divided in 4 recording sessions, with a break in between. During each session, the subject is shown a list of 60 stimuli, one by one. We call the presentation of a stimulus and the corresponding response from the subject a *trial*. Each trial consists of a sequence of states: (1) a green fixation cross is shown on screen for 250 ms, (2) a black screen is shown for variable period if time (500-1000 ms), (3) a unique CVC syllable stimulus appears and it stays on screen until the subjects finishes reading it out loud, (4) a white fixation cross is shown on screen between different trials. When the white cross turns green again, and we repeat from step (1). This process is illustrated in Figure 1.

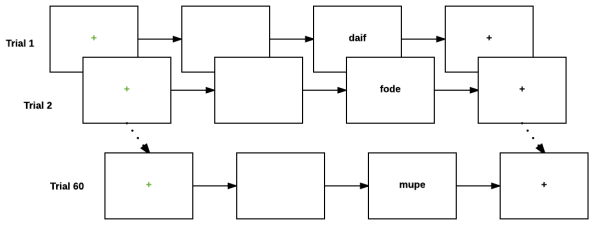


Figure 1: Stimulus presentation procedure

Neural activity data. We record two types of neural activity data: **(1)** Electrocorticography (ECoG) recordings from the ventral primary motor and primary sensory cortical areas. For some subjects, a few ECoG electrodes may also cover the temporal lobe near the auditory cortex. ECoG records electrical activity from the cortex using electrodes placed directly on the surface of the brain. Our recordings use between 6-36 electrodes (depending on the subject). For most subjects,

the ECoG is placed on the left side of the brain. **(2)** local field potentials (LFP) from the subthalamic nucleus. LFP measures the electrical potential collected from the neurons in a small brain area. Figure 2 illustrates the two recorded modalities.

Audio data. We also record audio data of the spoken works, at sampling frequencies between 30000Hz-96000Hz.

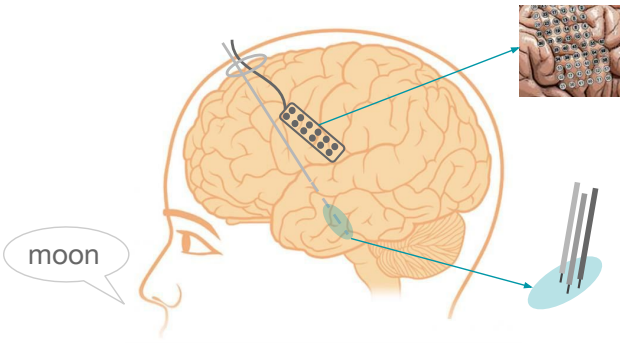


Figure 2: Brain activity recording paradigm.

Preprocessing. The following preprocessing steps have been applied to all our experiments (further preprocessing is mentioned in each experiment separately). For ECoG: remove DC offset; notch filter for 60Hz noise and harmonics; downsample to 1200Hz; reject bad channels; unreferenced; lowpass filtered at 400Hz. For LFP: filter using a built-in filter 500Hz; resample from 1375Hz to 1200Hz. All modalities, including audio, are aligned in time.

Exploring the data

As an example, we show in Figure 3 the data collected for a single trial, chosen randomly. We annotate on the figure the beginning of each of the trial states described above (also specified in the legend). We notice that the data does not show a visible pattern associated with the presentation of the stimulus, or with speech.

However, in neuroscience, typically the information lies in the frequency domain. Therefore we convert our time series into spectrograms. To convert the ECoG and LFP signals into a spectrogram, we used a Tukey window with shape parameter 0.25, over segments of 256 time points, and slide 32. The results are shown in Figures 24 and 25 in Appendix C. Note that the lower frequencies, 0-25 Hz, dominate the spectrogram, and do not allow us to visualize the higher frequencies, where speech-related information could potentially lie. The features not being on the same scale can also be a problem for the machine learning methods. Therefore we resort to a standard practice used in neuroscience of normalizing the each frequency band relative to a baseline period. For each frequency band, we compute the mean and standard deviation of all the time points between the *Green cross* cue and the *Stimulus* cue, and then we z-score all time points relative to this mean and standard deviation. The results are shown in Figures 4 and 5. With this preprocessing step we are now able to see all spectrogram bands on the same scale. We can even observe a higher activity in the high γ frequency band (85-175Hz) in the ECoG plot during speech time. This suggests that meaningful information about speech can be found in the γ band, and is consistent with other neuroscience studies [1, 8].

To verify whether this activation is consistent across trials, despite the noise in the data, we average all spectrograms, for each electrode separately, across all trials. Since the trial duration is not of fixed length, as the speech duration is different in every trial, we align all trials at speech onset before averaging. We show this for ECoG and LFP in Figures 6 and 8, respectively. In these plots, the activation in the γ frequency band at speech onset is even more prominent, especially in the ECoG signal. For LFP, note that the electrodes change position within the STN in each of the 4 recording sessions, such that each session records the electrical activity from different neurons. Since the STN is involved in several processes that happen in our body, it is possible that some of the neurons we record from are not involved in modulating speech, and thus might not show any signal related to speech.

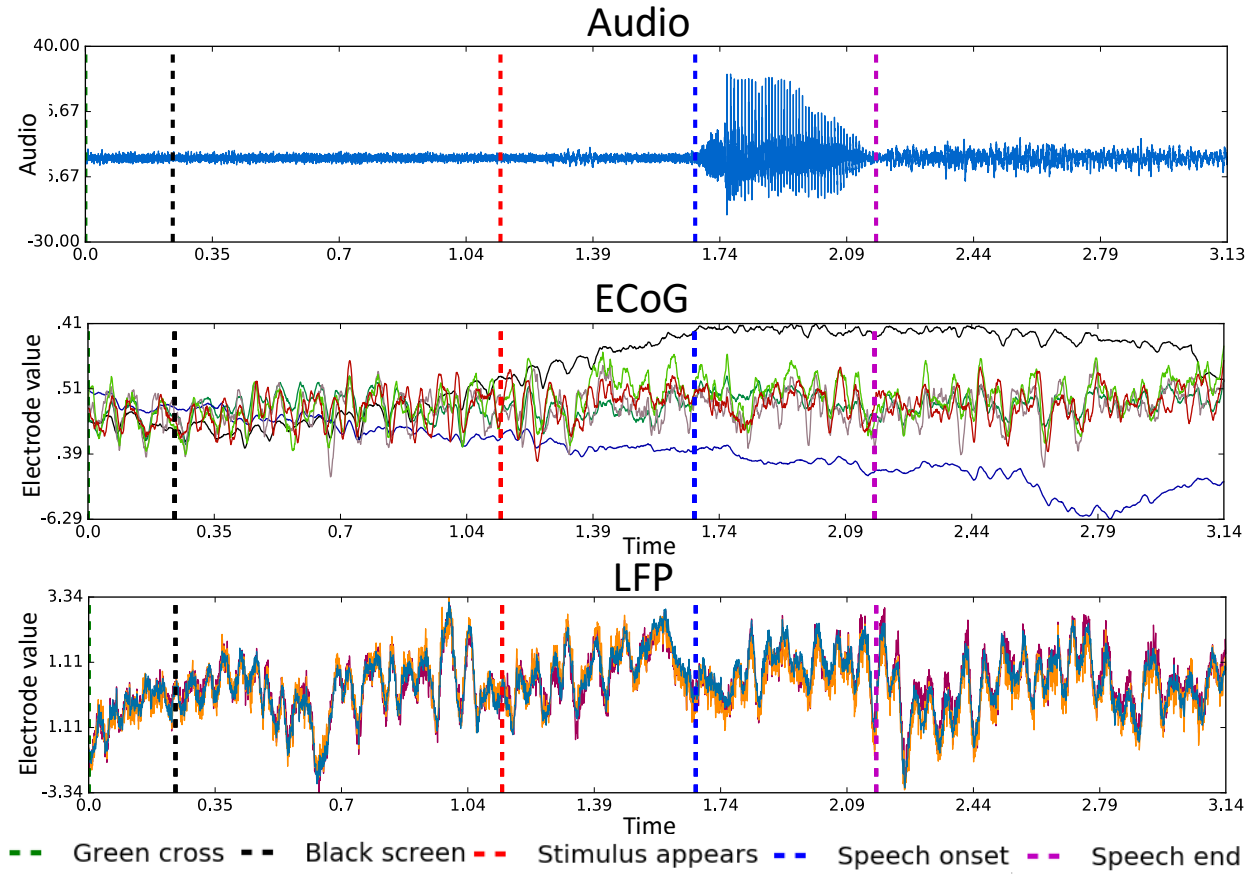


Figure 3: Audio, ECoG and LFP data for a single trial. Each plot shows multiple electrodes in different color.

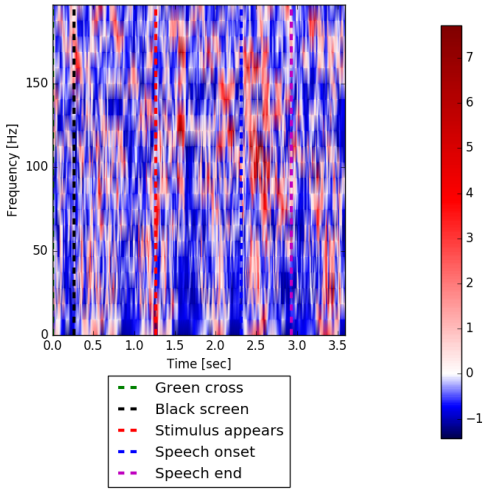


Figure 4: ECoG spectrogram for a single trial and a single electrode - normalized to baseline

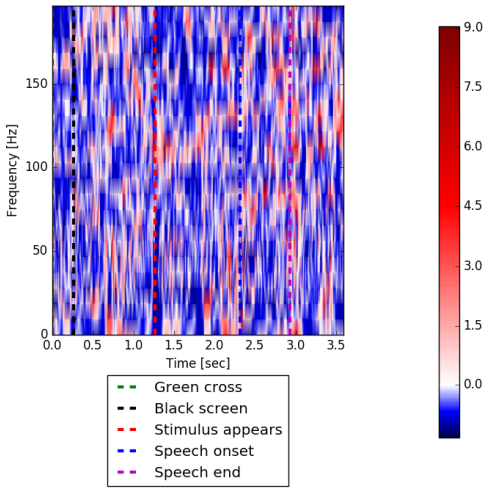


Figure 5: LFP spectrogram for a single trial and a single electrode - normalized to baseline

The signal observed in the γ band seems noisy even in the trial average plot. For this reason, we

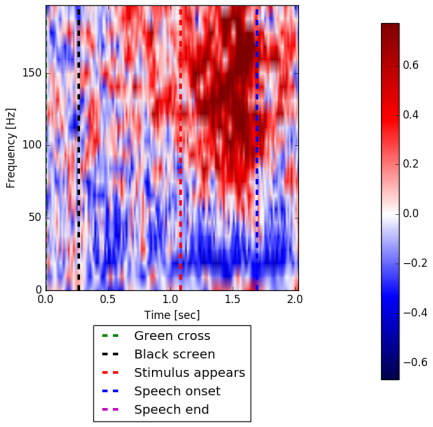


Figure 6: ECoG spectrogram, average over trials, for a single electrode

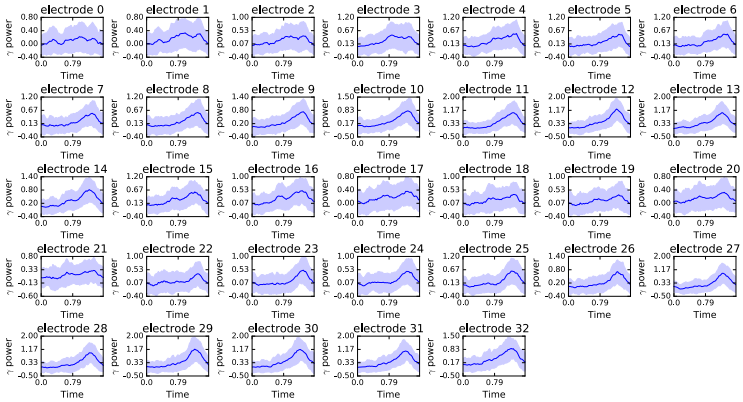


Figure 7: ECoG spectrogram, average over trials and over all γ frequencies (85-175Hz), per electrode

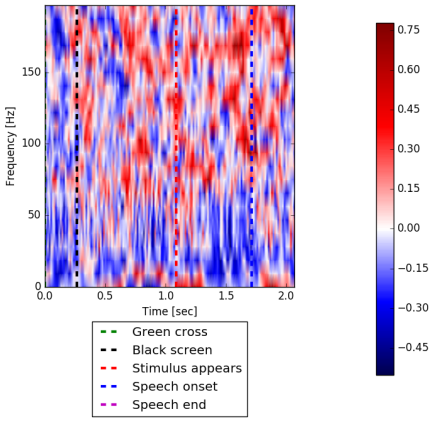


Figure 8: LFP spectrogram, average over trials, for a single electrode

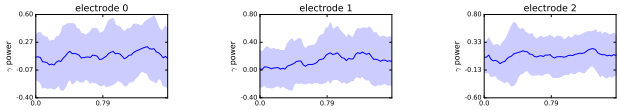


Figure 9: LFP spectrogram, average over trials and over all γ frequencies (85-175Hz), per electrode

apply another preprocessing step, and we average all the frequencies in the γ band. We obtain thus a single time series per electrode, where the value of the signal represents the average power of the frequencies in the γ band (85-175 Hz). This is shown in Figures 7 and 9.
In our experiments, we use as input to the machine learning models the neural activity at each of these preprocessing levels, and we analyze what models can extract meaningful information for the task at hand from these multiple views of neural activity.

4 Methods

We pose several machine learning (ML) prediction tasks. In all tasks, the input to the model is the neural activity from either cortex (ECoG) or STN (LFP), either as electrode measurements in time domain, or converted to frequency domain (and potentially band-passed), as discussed in Section 3. We apply our models on different levels of preprocessing of the data and discuss how this impacts the classification results, and the conclusions we can draw about brain activity. In the this section we describe the ML models we used, and what assumptions they make on the data.

We adopt the following convention for data format: for all data modalities and all types of preprocessing, the neural activity is converted to a 3D form and stored in a data structure of shape $trials \times electrodes \times time$. In the case of the spectrogram which has shape $trials \times electrodes \times frequencies \times time$, we consider each frequency in each electrode as a different electrode, and convert the data to the same 3D format: $trials \times (electrodes \times frequencies) \times time$.

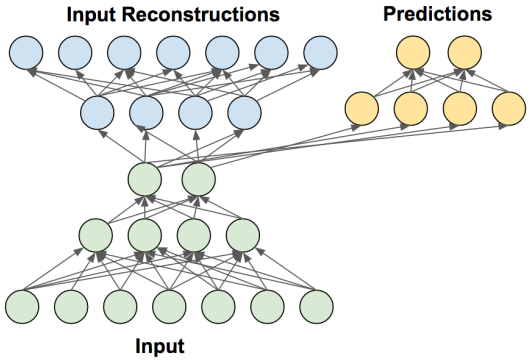


Figure 10: Multilayer Perceptron + Autoencoder

We perform either classification or regression, depending on the neuroscience task we want to explore, with the goal of predicting some speech property (e.g. classify when the subject speaking or not, classify which articulator is used (tongue, lips, teeth), predict the volume of the audio data using regression). We can generalize over these different cases, and denote the output as a structure of length $outputs$, or $outputs \times time$ (if we want to make a prediction per time point).

We present a short description of all the classification and regression models used in our experiments:

4.1 Regression Models

- Linear Regression (LR) – models the output as a linear combination of the input features. Since LR does not model time dependencies, in our case the inputs are the flattened time series of all electrodes (the value of each electrode at each time point constitutes a different feature).

- Ridge Regression – LR with a L2 sparsity constraint on the weights. This helps in preventing overfitting, and is particularly useful when there are few observations.

- Lasso Regression – LR with a L1 sparsity constraint on the weights. This helps in preventing overfitting, and encourages the weights to be sparse.

- Multilayer Perceptron (MLP) – a neural network consisting of multiple fully connected layers, that can model non-linear dependencies between inputs and outputs.

- Recurrent Neural Network (RNN) – a neural network architecture that models time dependencies. In our experiments we use a variant of RNNs called Long Short-Term Memory (LSTM) networks [4], which ameliorate the vanishing gradient problem, which often occurs in RNNs.

- Multilayer Perceptron + Autoencoder (MLP+A): We propose a neural network architecture that combines an MLP with a Denoising Autoencoder. An Autoencoder is a neural network consisting of two components: an encoder network (an MLP with layers of decreasing size that learns a lower dimensional representation of the input), and a decoder network (an MLP with layers of increasing size that reconstruct the input from the encoded representation). A Denoising Autoencoder (DA) [9] is an variant of Autoencoder that tries to avoid learning the identity function by randomly corrupting the inputs at train time. The architecture we propose consists of an MLP and a DA that share the encoder component, as shown in Figure 10. The idea behind it is that the common encoded representation will aim to satisfy two goals simultaneously: (1) it is a meaningful low dimensional representation of the inputs (from which we can decode a denoised version of the original signal), and (2) it encodes information

useful for the MLP prediction task. Depending on how we balance the two loss functions at train time (the DA loss and the MLP loss), the MLP+A could focus more on reconstructing the inputs, or on predicting well. Therefore the architecture inherently enforces a form of regularization for the MLP, which is particularly useful when we have few observations of noisy inputs, as in our case. With the right balance of the losses, we hope this network can converge to better local minima, and obtain better prediction results than an MLP trained alone.

- Baselines:
  - Average Predictor: for each output feature, it always predicts the average of that feature in the training set.
  - Zero Predictor: always predicts 0.

4.2 Classification Models

- Logistic Regression (LR) – it is a special case of the generalized linear model. Binary LR is used to estimate the probability of a binary variable based on a set of features, but it can be extended to multiple classes. As in the case of Linear Regression, we can also apply L1 and L2 regularization.
- Naive Bayes – classification method based on the Bayes Theorem, making the *naive* assumption that the features are independent. It is particularly useful when the dimensionality of the inputs is high.
- Multilayer Perceptron (MLP) – similar to the regression MLP, plus a softmax layer on the outputs that ensures that the outputs are in range [0,1] and sum to 1, and can thus represent the probabilities of each class.
- Recurrent Neural Network – similar to the regression RNN, plus a softmax layer on the outputs.
- Autoencoder + Multilayer Perceptrons – similar to the regression Autoencoder + MLP, but in this case we have a classification MLP with softmax on the outputs.
- Baselines: methods that use only the labels in the training data to make predictions based on the statistics of the labels (without using the neural activity):
  - Mode Predictor: it counts the occurrences of each class in the training set, and always predicts the most common class at test time.
  - Random Classifier: predicts at random one of the labels in the training set, chosen uniformly.
  - Random Classifier with Counts: predicts at random one of the labels in the training set, chosen from the distribution of the labels in the training set.

5 Experiments and Results

We approach the following prediction tasks:

1. Classifying speech versus silence
2. Classifying articulation and linguistic features
3. Predicting speech volume

For all these experiments, we consider as input the different views of neural activity described in Section 3, and we apply the machine learning models described in Section 4.

5.1 Predicting speech versus silence

We pose the following question: given a short recording of neural activity from either cortex or STN, are we able to tell by looking only at the brain activity, whether the subject is speaking or not? As we observed in Section 3, by visual inspection of a single trial at a time (as opposed to the trial average), it is not easy to tell when speech is occurring. The question that we raise here is whether a machine learning classifier is able to discover any patterns in the data that can classify such neural activity segments accurately, and if so, what models work best.

5.1.1 Experimental setting

We treat this task as a binary classification problem, where class 1 corresponds to speech, and class 0 corresponds to silence. As inputs to the classifier, we crop from the neural data sequences of fixed length t seconds (we experiment with different t's between [200ms, 400ms]) when the subject is either speaking throughout the whole sequence, or not speaking at all. This is illustrated in Figure 11. Depending on the duration of speech in each trial and the chosen t, this results in at least one sample of speech and 2 or more samples of silence per trial. We ensure that both the train and test datasets contain the same number of samples for speech and silence by resampling the speech samples, such that a random classifier would obtain 50% accuracy.

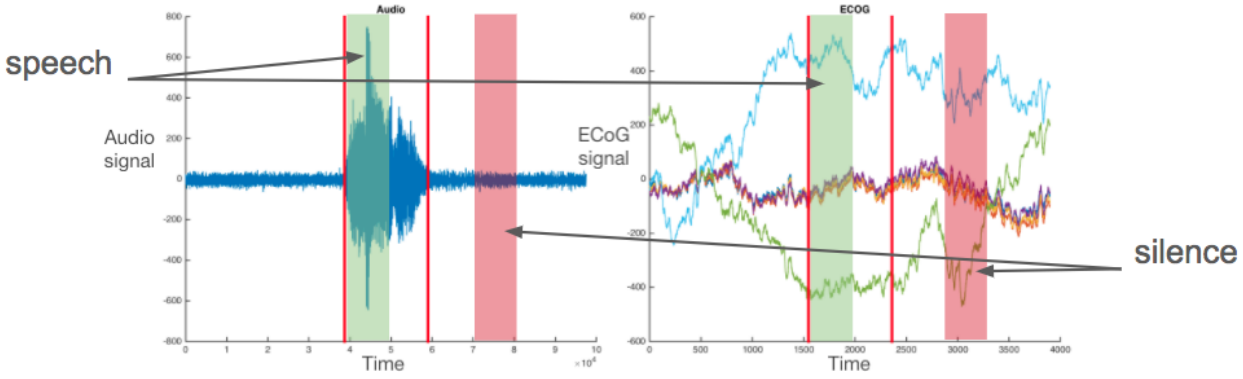


Figure 11: Selecting samples of speech and silence.

We attempt all the classification models described in Section 4, using as input each of the following: the electrode data in time domain ($trials \times electrodes \times time$), the spectrogram ($trials \times (electrodes \times frequencies) \times time$) with frequencies 0-200 Hz, the spectrogram averaged over the high γ band 85-175 Hz ($trials \times electrodes \times time$).

5.1.2 Results

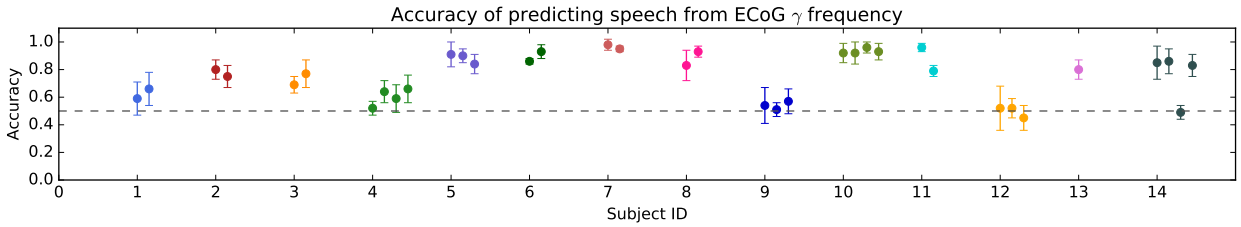
First, we evaluate the performance of different classifiers on a part of the subjects to identify the ones that work well for this task, and then we apply these models on all data.

Model comparison Due to space constraints, we show the results of all described classifiers on 4 recording sessions of a single subject, for each of the input configurations specified above. The parameters for each classifier have been chosen using 5-fold cross validation. For each recording session, we split the data into train and test using 5-fold cross validation, and we evaluate using classification accuracy and precision-recall area under the curve (AUC). We calculate the mean and standard deviation over the 5 folds for both metrics. Then we average these over all recording sessions. The full table of results for both ECoG and LFP are reported in Appendix A.

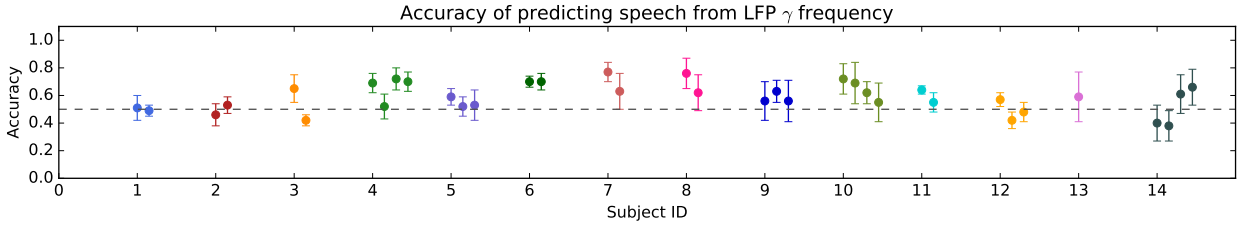
During our experiments we observed that all methods are prone to overfitting on our small dataset (except for Naive Bayes, which is not always able to fit the training data perfectly). Therefore proper regularization played a crucial role in obtaining results better than chance level, for all classifiers. Similarly, for more complex models, such as the neural network models, the choice of architecture also played a major role. With a careful balance between architecture size and regularization, neural network models (Multilayer Perceptron, Multilayer Perceptron + Autoencoder, Recurrent Neural Network) can perform better than simpler models like Logistic Regression or Naive Bayes (as can be seen for example in Appendix A for *ECoG spectrogram, LFP time domain, LFP spectrogram*). However finding the right parameter configuration is very time-consuming, and seems impractical to perform for every subject and every recording session. Among the neural network models, Multilayer Perceptron + Autoencoder worked best in terms of balance between performance and tuning efforts. This is probably due to the regularization implicitly enforced by the architecture: in order for the Autoencoder to reconstruct well the original signal, it forces the hidden layer to encode meaningful information about the signal (removing the random noise), while the Multilayer Perceptron part ensures that the hidden layer keeps the part of the signal that is relevant for the classification task at hand.

Therefore, based on our experiments, if not limited by time and computation constraints, we recommend the use of MLP+Autoencoder with proper parameter tuning. If such resources are not available and the MLP+Autoencoder cannot be properly tuned, we recommend instead simpler models such as Logistic Regression with L1 or L2 regularization on the weights.

Multi-subject results in frequency domain Since the Logistic Regression classifier with L1 regularization has shown good results consistently across subjects and is easy to train, we summarize the results of this classifier over all subjects and recording sessions in Figure 12. Note that for each subject we have between 1-4 recording sessions, and we train a different classifier per session. This is because the LFP electrodes change location in the STN after every session, and thus the neurons from which they record are different across sessions. For ECoG we can train the 4 sessions together. Our experiments show that training ECoG sessions together improves performance only slightly for most subjects, while for a few subjects some sessions have artifacts that affect the average performance. For these reasons, and for facilitating the comparison between ECoG and LFP results, we do classification per session for both modalities. The classifier for each session is trained using 5-fold cross validation. The input to the classifier is an array of shape $(trials \times electrodes \times time)$ representing the average power in the gamma frequency band over a time window of $t = 400ms$. Each trial represents a sample, and we have approximately 60 trials per session. For ECoG we have between 6-60 electrodes (depending on the subject), and for LFP 3 electrodes.



(a) ECoG



(b) LFP

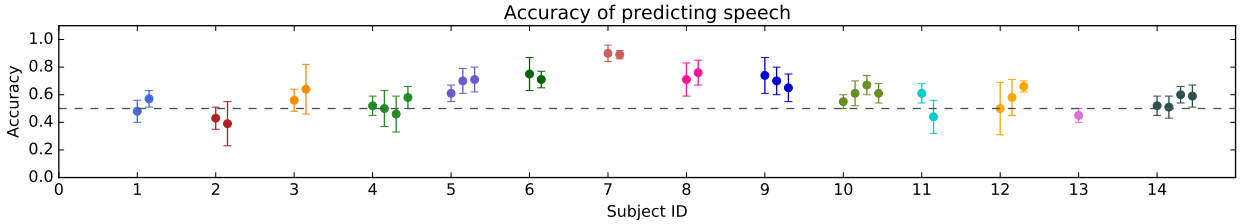
Figure 12: Accuracy of predicting speech vs. silence. The subjects are represented along the x axis, with a different color per subject. Each subject has 1-4 recording sessions. We mark the mean and standard deviation of the accuracy across 5 cross-validation folds for each session as a dot with error bars. The dashed horizontal line marks the 0.5 chance level.

From Figure 12, we observe that for almost all sessions the classification accuracy is above chance (50%). Since we have a balanced number of samples for each class *speech/silence*, predicting the most frequent label would also obtain 50% accuracy. Notice that for several subjects (e.g. IDs 5, 6, 7, 8, 10) the accuracy is close to 100% for ECoG, and >60% for LFP. This is an indicate that both cortex areas recorded with ECoG (ventral primary motor and primary sensory cortical areas) and the STN recorded with LFP exhibit patterns of neural activity connected to speech.

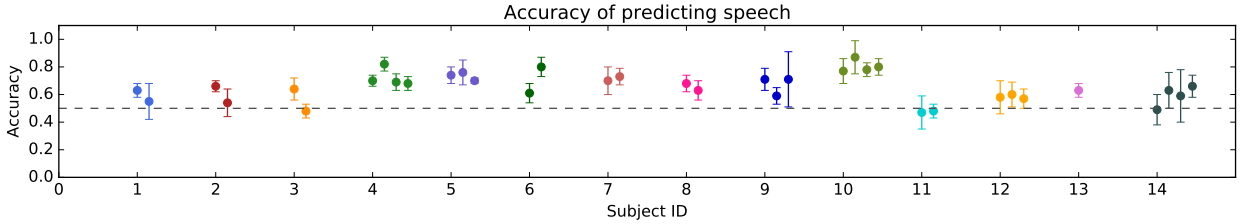
Multi-subject results in time domain We performed the same analysis in time domain. The questions we address are the following: (1) how will the same classifier perform without being guided by domain knowledge to look only at the γ frequency band, and (2) do other classifiers work better in time domain, than frequency domain. We show in Figure 13 the accuracy of predicting speech from ECoG and LFP, using a Logistic Regression classifier with L2 regularization. Classification using ECoG in time domain seems to perform worse than in frequency domain, which is not surprising giving our expectation that the information lies in the frequency domain. For LFP data, however, the results surprisingly show better accuracies in time domain, often surpassing the ECoG results for the same subject. This could potentially indicate that the STN encodes information about speech in a different frequency band than γ, and calls for further analysis of frequency bands. We can also compare the performance of different ML classifiers in time domain versus frequency domain. We use the results in Tables 2 and 3 of Appendix A for comparison. Logistic Regression (LR) works well in both time and frequency domain, and the Multilayer Perceptron + Autoencoder is close in accuracy to LR, or even surpasses it.

5.2 Predicting articulation and linguistic features

The next question we address is whether any articulation and linguistic features can be decoded from the primary sensory cortex, primary motor cortex, or STN. For instance, we would like to know



(a) ECoG



(b) LFP

Figure 13: Accuracy of predicting speech vs. silence from time domain using a Logistic Regression classifier with L2 regularization.

if you can predict what articulators (tongue, lips, teeth) are used when saying certain consonants, or the manner in which they are pronounced (fricative, plosive, nasal, etc.). Even more interestingly, we would like to know which parts of the brain and at what time points (relative to speech onset or stimulus presentation) contain information useful for decoding. Prior research has shown that the sensorimotor cortex exhibits some topographic organization of articulation features [1], and that certain phonemes (*/p/, /u/, /a:/, /k/*) can be discriminated with different degrees of precision [8]. Our goals in this experiment are the following: (1) validate existing results, (2) create a more granular map of when and where different features are encoded in the cortex, (3) discover if any such features are encoded in the STN, and if so, what part of the STN and at what time points.

5.2.1 Experimental setting

Each consonant can be described by several articulation and linguistic features. In our experiments, we use the features presented in Table 1. For a particular consonant, each feature takes a binary value: 1, if the consonant has that feature, or 0 otherwise. A complete list of consonants and the values of all features is shown in Table 4 in Appendix B. In this experiment, we use machine

Articulator			**Voicing**	**Manner**				
tongue	lips	teeth	voice (larynx)	fricative	lateral	nasal	plosive	trill

Table 1: Consonant articulation and linguistic features

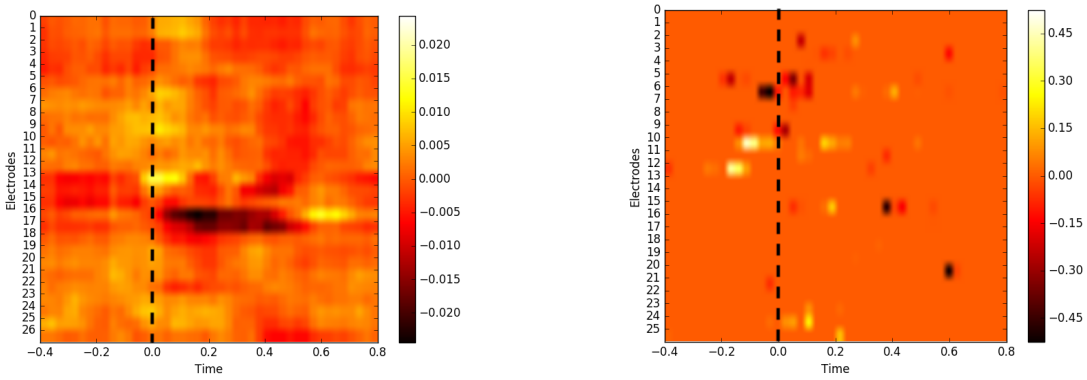
learning methods to discover which of these features can be decoded from the brain activity. We define several classification tasks, in which the input is the brain activity in the same format as described in section 5.1.1, and the output is a feature, or a set of features from Table 1. In terms of methods, we will focus on regression models whose parameters can be easily interpreted, as our goal is to learn where the information is encoded. Logistic regression (LR) with L1 or L2 regularization are especially attractive due the fact that they are fast to train, are interpretable, and, as we show

in the next section, they obtain good results across different subjects and classification tasks. In our setting, we use LR to learn a weight for each electrode at every time point, which allows us to understand where and when the information useful for decoding lies.

5.2.2 Results

Predicting Articulator and Voicing We start by analyzing the first set of features: predicting the used articulator (tongue, lips, teeth, larynx). Note from Table 4 that *tongue* and *lips* are mutually exclusive, but none of them is mutually exclusive with *teeth* or *larynx*. We show the accuracy of predicting *lips*, *teeth*, and *larynx (voicing)* for each of the subjects in Figures 19, 20, and 21 in Appendix B, respectively. The accuracies for *tongue* are similar to *lips* due to the mutual exclusion. Unlike the results of predicting speech versus silence, where almost all subjects showed accuracies above chance level, here we observe a large range of values across subjects. Some features can be decoded for some subjects with up to 95% accuracy, while others are at chance level. This is not surprising, because the ECoG electrodes are located in different positions for each subject, and the areas of the brain that show a change in neural activity for different articulators are very small, as suggested by prior research [1]. Therefore it es expected that for some subjects the ECoG electrodes will not overlap with the areas corresponding to some of the features. What is interesting to investigate is what brain areas contribute most to the classification task for the subjects for which a particular articulator can be decoded. To this purpose, we plot the LR weights for every electrode and every time point. We show as an example the weights learnt with LR for *voicing* and *lips* using L2 and L1 regularization, respectively, in Figure 14. The elements we want to consider are the ones that have highest absolute value (i.e. contribute most to the classification task). Note that this interpretation is valid because all inputs have been normalized, and are on the same scale. With L1 regularization it is even more clear what elements contribute, since the rest are set to 0. It is interesting to observe that in both cases only a small subset of electrodes contributes, and the timing is around speech onset time, which confirms the discoveries in [1], even though their experimental setting is somewhat different (they are looking at consonant-vowel transitions). With this analysis we can now look at the locations of the electrodes on the cortex and identify the regions of interest. We illustrate this in Figure 15 for a subject for which we obtain 86% accuracy at predicting when *lips* are used. It is interesting to notice that the weights useful for decoding lips correspond to electrodes in the sensorimotor cortex right before speech onset, and to electrodes in the auditory cortex after the subject has started speaking and can hear themselves.
We also attempted to predict the articulation features together, in a multi-task learning fashion, using a multilayer perceptron, but our results did not show any improvement from sharing information.
We did a preliminary similar analysis for LFP data from the STN. Our results showed that articulation features can be predicted better than chance in some cases, but with lower accuracies than from the cortex. We show in Figure 22 in Appendix B the weight plot for predicting voicing from LFP. A more thorough analysis is planned for future work, but the current results look very encouraging.

Predicting Manner We attempt to decode the next set of features, manner features, which refer to the configuration and interaction of the articulators. We have 5 mutually exclusive manner classes: *alveolar*, *bilabial*, *dental*, *labiodental*, and *palato-alveolar*. We show in Figure 16 the result of classifying 1-out-of-5 manner features using Logistic Regression with L2 regularization. Note from Table 4 that the 5 classes are not equally distributed among the consonants, which makes the dataset imbalanced. As baselines we compare the accuracy of our classifier with the Random Classifier and



(a) Weights for **voicing** from LR + L2 regularization. (b) Weights for **lips** from LR + L1 regularization.

Figure 14: Model weights trained with Logistic Regression (LR) when predicting articulation features. The vertical dashed line marks the speech onset time.

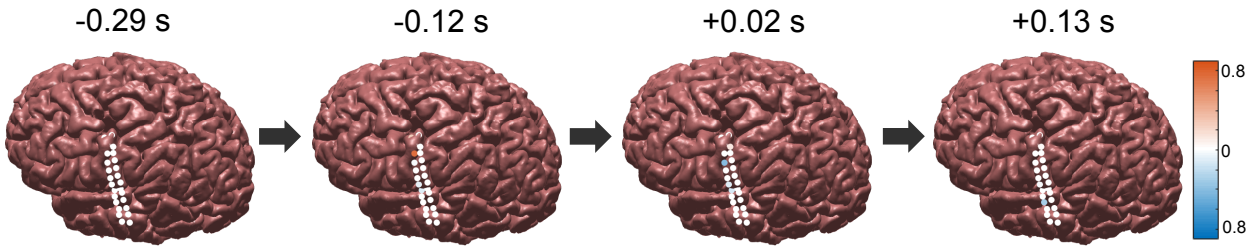


Figure 15: Logistic Regression weights trained for lips prediction, assigned to ECoG electrodes at different time points relative to speech onset.

Mode Predictor models described in Section 4. The results show that for most subjects this hard to predict better than chance, and only for few subjects the results are better than predicting the mode (i.e. always predict most common of the 5 classes). This may be due to the fact that some of the classes have very few samples in the training set. However, creating a binary classifier for the most common class, *fricative*, shows results better than chance for several subjects, and the weight plots (described in the previous paragraph) show only few electrodes with non-zero weights soon after speech onset, similar to the patterns seen for the articulation features (Figure 23 in Appendix B).

5.3 Predicting speech volume

In section 5.1 we have seen that both the cortex and the STN contain enough information to distinguish when a subject is speaking. The next question that we raise is whether we can predict more fine grained features of speech. In this section we address the problem of predicting the volume (or loudness) of speech.

In order to obtain a measure of the loudness, we convert the audio data into *perceived volume*. We chose a transformation commonly used for perceived volume: root mean square of the audio over small windows of time. That is, the volume at time t is: $volume(t) = \sqrt{\left(\frac{1}{2w+1}\sum_{i=-w}^{w} audio(t+i)\right)}$, where $2w+1$ is the size of the chosen time window centered at t. The effect of this transformation is shown in Figure 17.

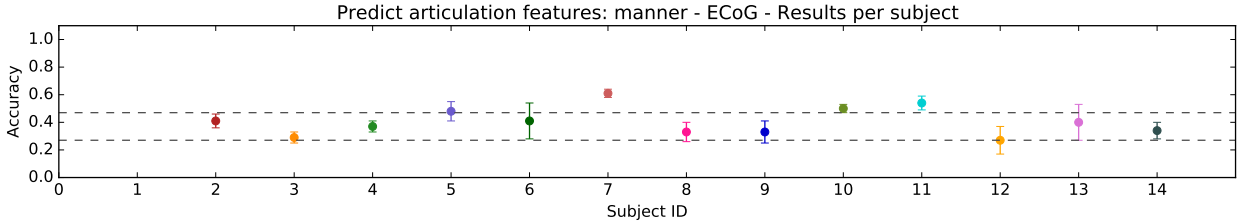


Figure 16: Accuracy of predicting **manner** features in the first consonant, using Logistic Regression with L2 regularization on ECoG data. The subjects are represented along the x axis, with a different color per subject. We mark the mean and standard deviation of the accuracy across 5 cross-validation folds as a dot with error bars. The dashed horizontal lines marks the 0.25 chance level, and 0.46 mode predictor accuracy.

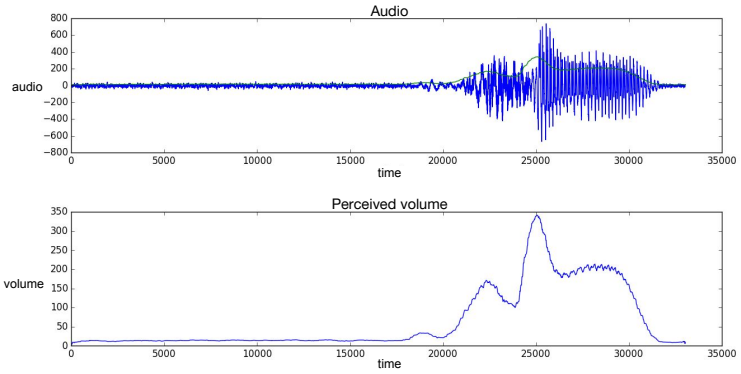
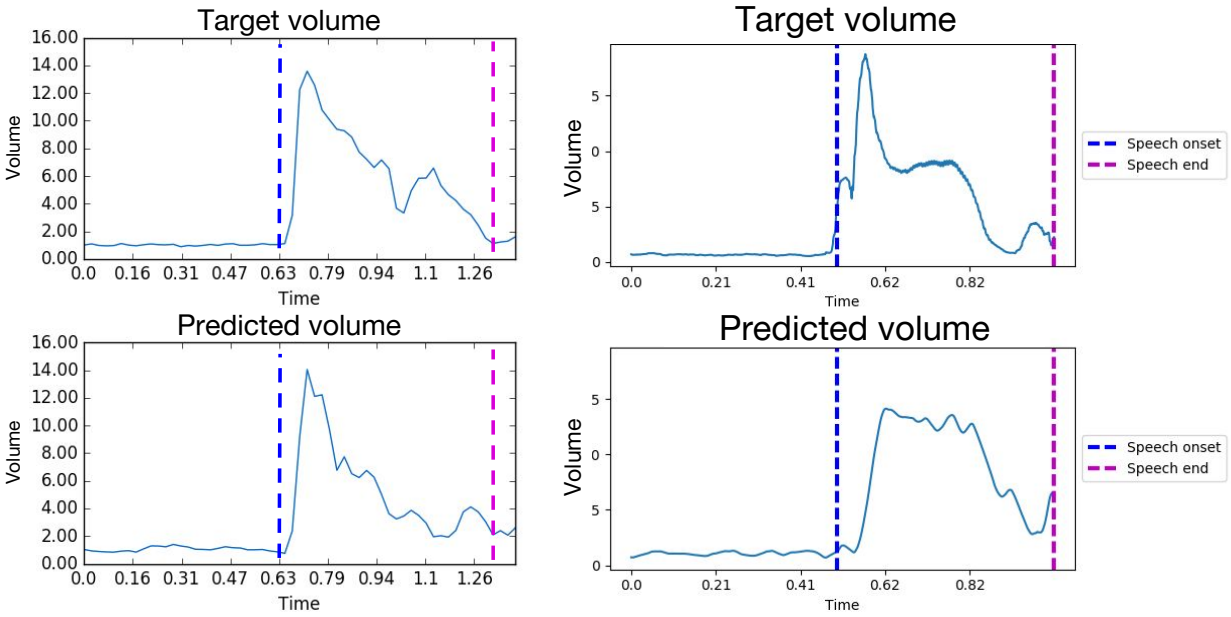


Figure 17: Converting audio data to perceived volume.

Experimental setting. The goal in this experiment is to predict, given as input the brain activity signal, the perceived volume of speech. The input has the same format as in the previous tasks. The output is a real valued function over time, representing perceived volume, as the example shown in Figure 17. Since this is a regression task, we experiment with the regression models described in Section 4. We evaluate our results quantitatively in terms of mean squared error (MSE) between the predicted and target volume, as well as qualitatively, by plotting the predicted volume and comparing visually with the target.

Results. Our experiments show that predictions using the STN data as input show better qualitative and quantitative results than similar models applied on cortex data. We show some sample predictions from the STN in Figure 18. To obtain the prediction in Figure 18a, all brain activity time series (i.e. all time points) was used as input to a Ridge Regression classifier. The results look well, however, our concern was that the classifier may learn the structure of the experiment (i.e. first there is a period of silence, followed by a period of speech, followed by silence again) in order to improve its predictions. To remove this effect, we used a rolling window of 500ms over the input brain activity and predicted each time point in the volume. The results are shown in Figure 18b. Indeed, the prediction no longer looks so accurate, but the model is still able to discover some level of volume. In fact, Ridge Regression no longer worked well, so the results in Figure 18b were obtained with a more complex model, a Multilayer Perceptron with 3 hidden layers of 256, 128, and 64 units, respectively. These preliminary results look very exciting, as the STN is not known to participate in speech production. More investigation is though required to describe the degree to which it is involved.



(a) Predicting volume at all time points simultaneously, from all brain activity.

(b) Predicting volume at every time point from a 500ms window of brain activity.

Figure 18: Volume prediction from STN data. The vertical dashed lines mark the speech onset and end times.

6 Conclusions

In this project, we analyzed the neural activity associated with speech production. We approached the problem from two perspectives. From a neuroscience perspective, we attempted to understand which parts of the brain and at what time points contain information related to various aspects of speech (e.g. is the person speaking, how loud are they speaking, what articulators are they using), by trying to predict these features from the neural activity, using machine learning models. Our experiments showed that it is possible to predict when speech is occurring from both the ventral primary motor and primary sensory cortex, as well as from the subthalamic nucleus (STN). Moreover, we can predict from both regions, with different degrees of accuracy, even more fine grained features, such as different articulation and linguistic features, or the speech volume. In our experiments, the data from the cortex was better at predicting articulation features than STN data, but STN data was better at predicting speech volume. These results suggest that the inspected brain regions may be involved in speech production. The cortex results confirm existing studies that have found a topographic organization of articulation features in the sensorimotor cortex. The STN results are interesting, as they challenge the current theories of speech production, which include other basal ganglia regions, but not the STN.

From a machine learning perspective, we compared the performance of several models on the tasks described above. In our experiments, simple regression models with regularization, such as Logistic Regression with L2 regularization, seem to work well consistently across subject and tasks, and are easy to train. More complex models, such as neural network models, can work even better if their parameters are tuned properly. However, they require more computational resources and more time to train.

We hope these results will be of use to researchers using machine learning to studying the brain. In particular, we hope these results bring us a step closer to understanding the neural basis of speech production, which has major implications for science and medicine.

7 Acknowledgements

I would like to thank my advisers, Dr. Barnabàs Pòczos and Prof. Tom Mitchell, who provided great insights for this work. Many thanks to Dr. Mark Richardson and Dr. Witold Lipski for sharing their neuroscience expertise and collecting the data used in our experiments, as well as Dr. Anna Chrabaszcz for sharing her knowledge in linguistics. This research was sponsored in part by the National Institutes of Health (NIH) grant U01NS098969, and the Army Research Laboratory accomplished under Cooperative Agreement Number W911NF-10-2-0022. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government.

References

[1] K. E. Bouchard and E. F. Chang. Neural decoding of spoken vowels from human sensory-motor cortex with high-density electrocorticography. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 6782–6785. IEEE, 2014.

[2] D. De Gaspari, C. Siri, M. Di Gioia, A. Antonini, V. Isella, A. Pizzolato, A. Landi, F. Vergani, S. Gaini, I. Appollonio, et al. Clinical correlates and cognitive underpinnings of verbal fluency impairment after chronic subthalamic stimulation in parkinson's disease. *Parkinsonism & related disorders*, 12(5):289–295, 2006.

[3] D.-B. S. for Parkinson's Disease Study Group et al. Deep-brain stimulation of the subthalamic nucleus or the pars interna of the globus pallidus in parkinson's disease. *N Engl J Med*, 2001(345):956–963, 2001.

[4] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[5] S. L. Kowal, T. M. Dall, R. Chakrabarti, M. V. Storm, and A. Jain. The current and projected economic burden of parkinson's disease in the united states. *Movement Disorders*, 28(3):311–318, 2013.

[6] T. D. Parsons, S. A. Rogers, A. J. Braaten, S. P. Woods, and A. I. Tröster. Cognitive sequelae of subthalamic nucleus deep brain stimulation in parkinson's disease: a meta-analysis. *The Lancet Neurology*, 5(7):578–588, 2006.

[7] S. Pinto, C. Ozsancak, E. Tripoliti, S. Thobois, P. Limousin-Dowsey, and P. Auzou. Treatments for dysarthria in parkinson's disease. *The Lancet Neurology*, 3(9):547–556, 2004.

[8] N. Ramsey, E. Salari, E. Aarnoutse, M. Vansteensel, M. Bleichner, and Z. Freudenburg. Decoding spoken phonemes from sensorimotor cortex with high-density ecog grids. *NeuroImage*, 2017.

[9] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

Appendix A

Table 2: Results for predicting speech versus silence using ECoG – all classifiers, a single subject, 4 recording sessions.

Modality	Model	Accuracy mean | std		Precision-Recall AUC mean | std	
		train	**test**	**train**	**test**
/	Random	0.49 | 0.01	0.51 | 0.03	0.49 | 0.01	0.63 | 0.0
/	Random with counts	0.50 | 0.01	0.53 | 0.03	0.63 | 0.01	0.65 | 0.02
/	Mode	0.50 | 0.00	0.50 | 0.00	0.75 | 0.00	0.75 | 0.00
ECoG time domain	Naive Bayes	0.59 | 0.02	0.56 | 0.05	0.70 | 0.02	0.69 | 0.03
	Logistic Regression	1.00 | 0.00	**0.60 | 0.16**	1.00 | 0.00	**0.71 | 0.12**
	K Nearest Neighbor k = 5	1.00 | 0.00	0.53 | 0.02	1.00 | 0.00	0.67 | 0.01
	Boosted Random Forest max_depth = 3, subsample = 0.9	1.00 | 0.00	0.54 | 0.02	1.00 | 0.00	0.61 | 0.04
	Multilayer Perceptron hidden layer size = 32	0.96 | 0.0	0.54 | 0.03	0.98 | 0.02	0.62 | 0.04
	Multilayer Perceptron + Autoencoder autoenc_hidden = 128, mlp_loss_weight=0.01, encoder activation=none, decoder activation=none, dropout=0.2, mlp_activation=ReLU	0.63 | 0.01	0.59 | 0.04	0.73 | 0.01	0.69 | 0.04
	Recurrent Neural Network (LSTM) hidden layer size =32	1.00 | 0.00	0.51 | 0.06	1.00 | 0.00	0.58 | 0.07
ECoG spectrogram frequencies 0-200Hz	Naive Bayes	0.61 | 0.08	0.59 | 0.08	0.77 | 0.02	0.76 | 0.03
	Logistic Regression L1 regularization, weight=1	1.00 | 0.00	0.63 | 0.01	1.00 | 0.00	0.71 | 0.01
	K Nearest Neighbor k = 5	1.00 | 0.00	0.52 | 0.03	1.00 | 0.00	0.57 | 0.07
	Boosted Random Forest max_depth = 3, subsample = 0.9	1.00 | 0.00	0.60 | 0.02	1.00 | 0.00	0.71 | 0.03
	Multilayer Perceptron hidden layer size = 32	1.00 | 0.00	0.55 | 0.06	1.00 | 0.00	0.65 | 0.05
	Multilayer Perceptron + Autoencoder autoenc_hidden = 128, mlp_loss_weight=100, encoder activation=none, decoder activaion=none, dropout=0.2, mlp_activation=ReLU	0.99 | 0.01	**0.73 | 0.09**	0.99 | 0.01	**0.81 | 0.09**
	Recurrent Neural Network (LSTM) hidden layer size = 32	1.00 | 0.00	0.61 | 0.09	1.00 | 0.00	0.70 | 0.10
ECoG average power gamma frequencies 85-175Hz	Naive Bayes	0.70 | 0.05	0.69 | 0.08	0.80 | 0.02	0.78 | 0.07
	Logistic Regression L2 regularization, weight=1	0.98 | 0.01	**0.78 | 0.04**	0.98 | 0.00	**0.85 | 0.02**
	K Nearest Neighbor k = 5	1.00 | 0.00	0.66 | 0.06	1.00 | 0.00	0.75 | 0.04
	Boosted Random Forest max_depth = 3, subsample = 0.9	1.00 | 0.00	0.69 | 0.04	1.00 | 0.00	0.78 | 0.05
	Multilayer Perceptron hidden layer size = 32	1.00 | 0.00	0.64 | 0.06	1.00 | 0.00	0.72 | 0.07
	Multilayer Perceptron + Autoencoder autoenc_hidden = 128, mlp_loss_weight=1, encoder activation=none, decoder activation=none, denoising_dropout=0.2	0.84 | 0.02	0.67 | 0.03	0.88 | 0.01	0.75 | 0.03
	Recurrent Neural Network (LSTM) hidden layer size = 32	1.00 | 0.00	0.75 | 0.10	1.00 | 0.00	0.81 | 0.08

Table 3: Results for predicting speech versus silence using LFP – all classifiers, a single subject, 4 recording sessions.

Modality	Model	Accuracy mean | std		Precision-Recall AUC mean | std	
		train	**test**	**train**	**test**
/	Random	0.49 | 0.01	0.51 | 0.03	0.49 | 0.01	0.63 | 0.0
/	Random with counts	0.50 | 0.01	0.53 | 0.03	0.63 | 0.01	0.65 | 0.02
/	Mode	0.50 | 0.00	0.50 | 0.00		0.75 | 0.00
LFP time domain	Naive Bayes	0.63 | 0.01	0.60 | 0.05	0.73 | 0.01	0.70 | 0.04
	Logistic Regression L1 regularization, weight=1	0.99 | 0.01	0.57 | 0.06	0.99 | 0.01	0.65 | 0.07
	K Nearest Neighbor k = 5	1.00 | 0.00	0.56 | 0.06	1.00 | 0.00	0.68 | 0.05
	Boosted Random Forest max_depth = 3, subsample = 0.9	1.00 | 0.00	0.57 | 0.0	1.00 | 0.00	0.64 | 0.12
	Multilayer Perceptron hidden layer size = 32	1.00 | 0.00	**0.66 | 0.04**	1.00 | 0.00	**0.76 | 0.06**
	Multilayer Perceptron + Autoencoder autoenc_hidden = 128, mlp_hidden=32 encoder activation=none, decoder activation=none, dropout=0.2, mlp_activation=ReLU	0.90 | 0.01	0.64 | 0.05	0.92 | 0.01	0.74 | 0.04
	Recurrent Neural Network (LSTM) hidden layer size = 64	0.82 | 0.08	0.57 | 0.07	0.86 | 0.06	0.66 | 0.08
LFP spectrogram frequencies 0-200Hz	Naive Bayes	0.72 | 0.01	0.54 | 0.03	0.79 | 0.01	**0.63 | 0.03**
	Logistic Regression L1 regularization, weight=1	1.00 | 0.00	0.52 | 0.01	1.00 | 0.00	0.61 | 0.0
	K Nearest Neighbor k = 10	1.00 | 0.00	0.52 | 0.03	1.00 | 0.00	**0.63 | 0.03**
	Boosted Random Forest max_depth = 3, subsample = 0.9	1.00 | 0.00	0.51 | 0.03	1.00 | 0.00	0.56 | 0.15
	Multilayer Perceptron hidden layer size = 32	1.00 | 0.00	0.52 | 0.05	1.00 | 0.00	0.61 | 0.05
	Multilayer Perceptron + Autoencoder autoenc_hidden = 128, mlp_loss_weight=1, encoder activation=none, decoder activation=none, dropout=0.2, mlp_activation=ReLU	0.98 | 0.01	**0.55 | 0.04**	0.98 | 0.01	**0.63 | 0.05**
	Recurrent Neural Network (LSTM) hidden layer size = 64	1.00 | 0.00	**0.55 | 0.05**	1.00 | 0.00	**0.63 | 0.08**
LFP average power gamma frequencies 85-175Hz	Naive Bayes	0.58 | 0.01	**0.54 | 0.06**	0.69 | 0.02	**0.65 | 0.05**
	Logistic Regression L2 regularization, weight=1	0.62 | 0.01	0.50 | 0.06	0.71 | 0.01	0.61 | 0.05
	K Nearest Neighbor k = 5	1.00 | 0.00	0.47 | 0.07	1.00 | 0.00	0.61 | 0.05
	Boosted Random Forest max_depth = 3, subsample = 0.9	1.00 | 0.00	0.46 | 0.0	1.00 | 0.00	0.47 | 0.07
	Multilayer Perceptron hidden layer size = 32	1.00 | 0.00	0.46 | 0.06	1.00 | 0.00	0.52 | 0.07
	Multilayer Perceptron + Autoencoder autoenc_hidden = 128, mlp_loss_weight=1, encoder activation=none, decoder activation=none, dropout=0.2, mlp_activation=ReLU	0.60 | 0.01	0.52 | 0.06	0.70 | 0.01	0.64 | 0.05
	Recurrent Neural Network (LSTM) hidden layer size = 32	1.00 | 0.00	0.51 | 0.04	1.00 | 0.00	0.57 | 0.06

Appendix B

Table 4: Consonant articulation and linguistic features.

Consonant	Articulator			Voicing	Place					Manner					Aquisition	
	tongue	lips	teeth	voice (larynx)	alveolar	bilabial	dental	labiodental	palato-alveolar	fricative	lateral	nasal	plosive	trill	early	late
s	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1
z	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1
l	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0	1
n	1	0	0	1	1	0	0	0	0	0	0	1	0	0	1	0
t	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0
d	1	0	0	1	1	0	0	0	0	0	0	0	1	0	1	0
r	1	0	0	1	1	0	0	0	0	0	0	0	0	1	0	1
m	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1	0
p	0	1	0	0	0	1	0	0	0	0	0	0	1	0	1	0
θ	1	0	1	0	0	0	1	0	0	1	0	0	0	0	0	1
ð	1	0	1	1	0	0	1	0	0	1	0	0	0	0	0	1
f	0	1	1	0	0	0	0	1	0	1	0	0	0	0	1	0
v	0	1	1	1	0	0	0	1	0	1	0	0	0	0	1	0
ʃ	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1

Figure 19: Accuracy of predicting when **lips** are used in the first consonant from ECoG data, using Logistic Regression with L2 regularization. The subjects are represented along the x axis, with a different color per subject. We mark the mean and standard deviation of the accuracy across 5 cross-validation folds as a dot with error bars. The dashed horizontal line marks the 0.5 chance level.

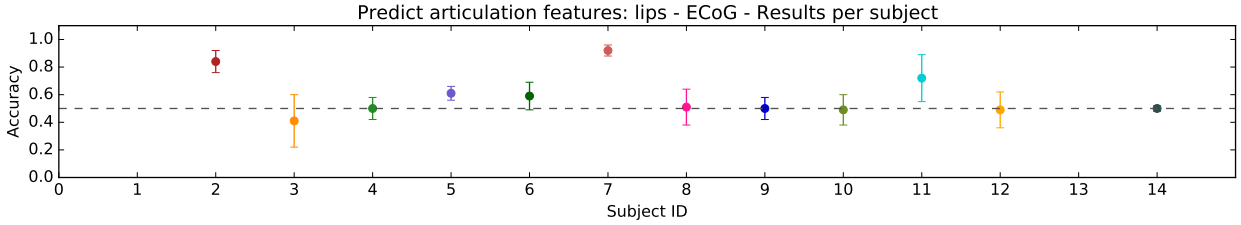


Figure 20: Accuracy of predicting when **teeth** are used in the first consonant from ECoG data, using Logistic Regression with L2 regularization. The subjects are represented along the x axis, with a different color per subject. We mark the mean and standard deviation of the accuracy across 5 cross-validation folds as a dot with error bars. The dashed horizontal line marks the 0.5 chance level.

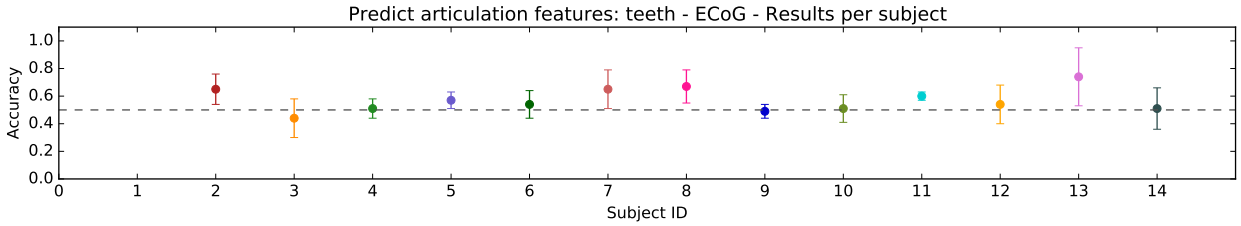


Figure 21: Accuracy of predicting when **voicing (larynx)** are used in the first consonant from ECoG data, using Logistic Regression with L2 regularization. The subjects are represented along the x axis, with a different color per subject. We mark the mean and standard deviation of the accuracy across 5 cross-validation folds as a dot with error bars. The dashed horizontal line marks the 0.5 chance level.

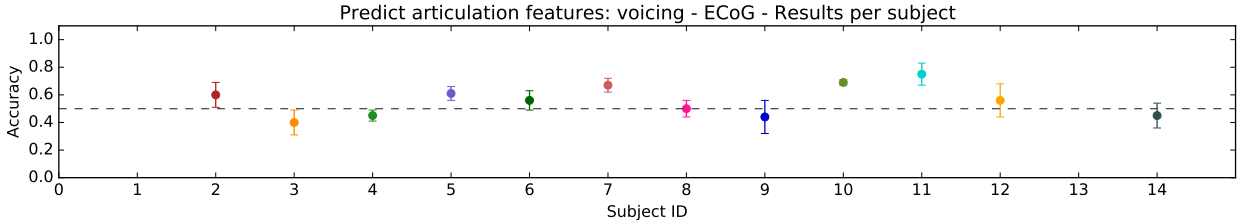


Figure 22: Model weights trained with Logistic Regression (LR) with L1 regularization when predicting whether the first consonant is **fricative** from **LFP** data from the STN. The vertical dashed line marks the speech onset time.

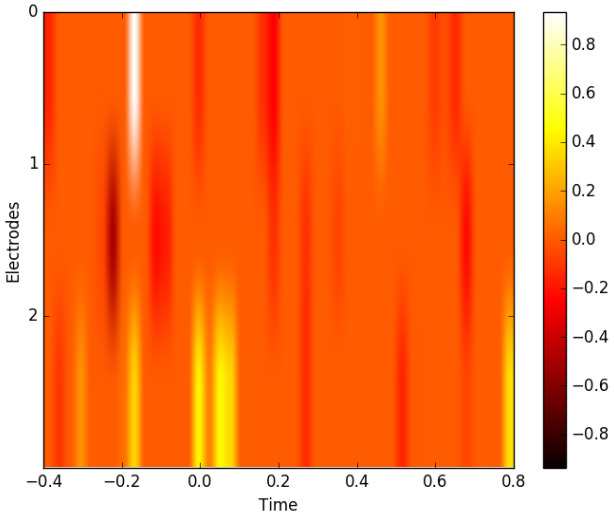
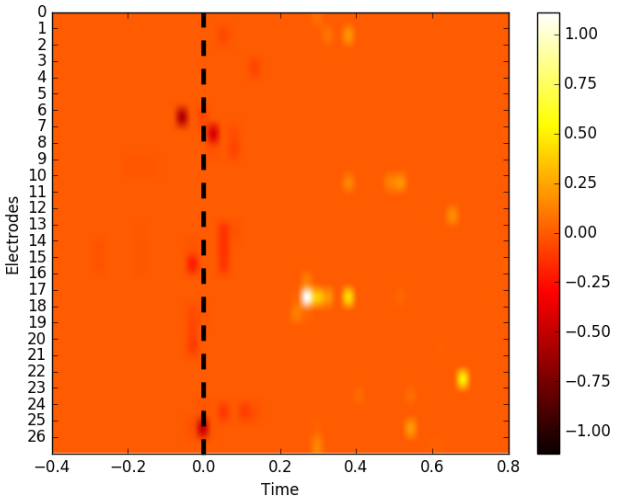


Figure 23: Model weights trained with Logistic Regression (LR) with L1 regularization when predicting whether the first consonant is **fricative** from **ECoG** data from the cortex. The vertical dashed line marks the speech onset time.



Appendix C

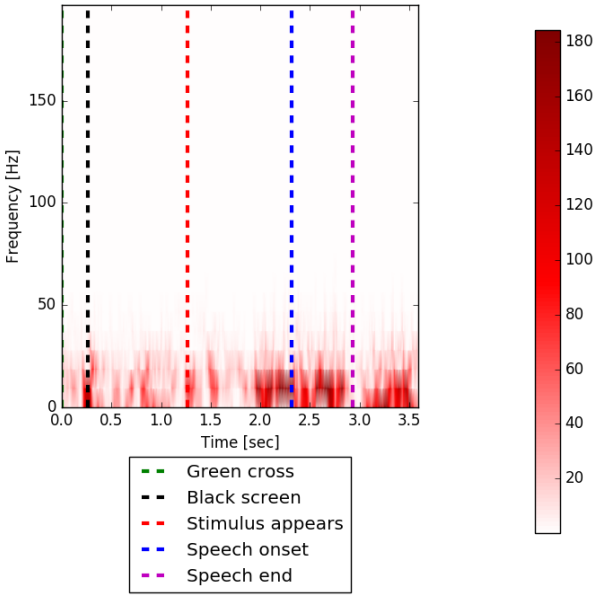


Figure 24: ECoG spectrogram for a single trial and a single electrode - no normalization

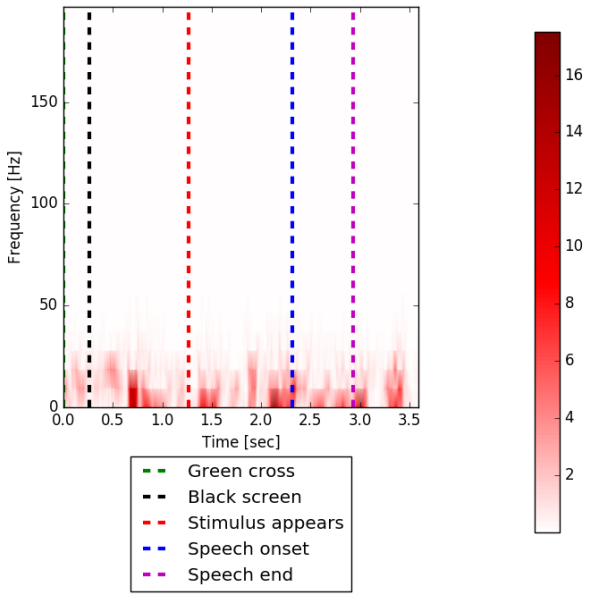


Figure 25: LFP spectrogram for a single trial and a single electrode - no normalization