Covert speech decoding from EEG signals

Clément Dauvilliers, Emma Farina, Husam Jubran
clement.dauvilliers@epfl.ch, emma.farina@epfl.ch, husam.jubran@epfl.ch
Machine Learning CS-433 - EPFL
Project hosted by Prof. Anne-Lise Giraud, Department of Basic Neurosciences, Université de Gèneve
Supervisors: Kinkini Bhadra, Shizhe Wu

***Abstract*—Covert speech is performed everyday by most people without even noticing, Decoding covert speech directly from the neural signal could be life-changing for patients who are unable to speak. In this work, we assess the importance of several frequency bands and areas of the brain in classifying two syllables of imagined speech. We use the optimal features to improve three classifiers: a random forest, logistic regression and convolutional neural network, by reducing their overfitting.**

I. Introduction

Covert speech, also called imagined speech, is the internal pronunciation of phonemes, words, or sentences, without the movement of the phonatory apparatus or any audible output [1]. Although speech related disabilities such as in aphasia or locked-in-syndrome commonly restrict overt production of speech, even in these conditions it is possible to actively imagine speaking [2]. Brain-Computer Interfaces (BCIs) interpret brain activity into digital form that acts as a command for a computer, allowing users to control external devices by using brain signals [3]. A BCI system which is able to decode the electrical activity of the brain during covert speech and translate it into words would therefore improve the quality of life of people with disabilities [2].

Among the neuroimaging techniques currently available for BCI systems, electroencephalography (EEG) has the advantage of being cost-effective and non-invasive with high temporal resolution of less than 1 millisecond. Nevertheless, such systems present some challenges, including low signal-to-noise ratio, low spatial resolution and frequent artifacts due to eye blinking or muscular activities [2], [3]. Furthermore, even though some areas of the brain are known to be specifically dedicated to speech perception and production, there is a relevant inter-subject and intra-subject variability in the spatial features of speech related tasks [4], which makes finding a model that provides reliable decoding challenging even for a single person over several days.

A. Objectives

The study surrounding this work aims at using machine learning methods to classify two imagined syllables ("*fo*" and "*gi*") based on EEG signals. Such a system, although developed using offline data, would potentially be useful in the clinical context, if the complexity of the algorithm allows for real-time application. The objective of this work precisely is to apply the pipeline used in [5], based on Convolutional Neural Networks (CNNs), to our dataset. However, a secondary objective is to understand which areas of the brain, as well as which frequencies, give the most information regarding the imagined speech. By limiting the features to the essential electrodes and frequencies using this knowledge, we hope to help the CNN avoid overfitting.

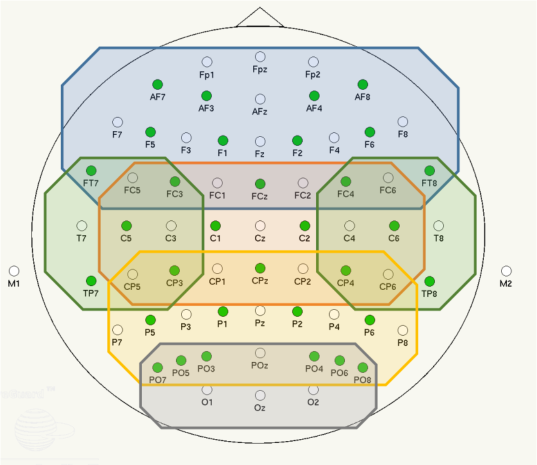

Fig. 1. Ant Neuro System channel map with electrode grouping. Purple: frontal electrodes, green: left and right temporal electrodes, red: central electrodes, orange: parietal electrodes, black: occipital electrodes.

II. Models and Methods

A. Dataset

EEG data was acquired from 1 subject trained for 5 consecutive days. Each day, the subject performed a total of 90 trials imagining different syllables (45 times "*fo*" and 45 times "*gi*").

The participant is asked to imagine the syllable "*fo*" or "*gi*" depending on color cues shown on a screen (pink and blue, respectively), and internally pronounce it multiple times guided by the fixed rhythm of an auditory cue. The duration of the imagery task is 5 seconds. EEG signals were recorded using a 64-channel Ant Neuro System with sampling rate 512 Hz. During each trial, a Random Forest model tries to predict which syllable was imagined based on the subject's brain recordings. As the days and trials go on, **the subject learned to adapt to the model**. Therefore, we expect the classification task to be easier for the trials from day-4 and 5 than day-1 and 2.

Eventually, the brain signals are extremely sensitive to many environmental conditions, making the recordings from two different days and even trials potentially very different. Hence, a crucial point of the study is that **the models are trained and evaluated for each day separately and independently**.

While we originally worked on the raw EEG signals using denoising and filtering methods such as EMD decomposition and Independent Component Analysis, we were at some point provided with a clean dataset including 5 days of trials from a single subject.

B. *Pre-processing*

The following pre-processing was performed:

1) Time window selection, to isolate the speech imagery time window in each epoch;
2) Baseline correction, by subtracting the average of the signal between -1.5 s and -1 s of each trial;
3) Electrode grouping according to the electrode position, as shown in Figure 1;
4) Frequency band division, as shown in Table I.

EEG Waveform	Frequency Range
Delta	1-4 Hz
Theta	4-8 Hz
Alpha	8-12 Hz
Beta	12-25 Hz
Low Gamma	25-40 Hz
High Gamma	40-70 Hz

TABLE I
FREQUENCY BANDS IN EEG SIGNALS.

C. *Features*

Following [5], the pre-processed channels from Section II-B are transformed into images to be given to a Convolutional Neural Network (CNN). The input of the CNN is a matrix that contains the FFT of the differences between every couple of electrodes. The resulting input is an image of a single channel, of dimensions $\frac{N_T}{2}\text{x}\frac{N_E(N_E-1)}{2}$ where N_T is the number of timesteps and N_E the number of electrodes.

Feature selection, *i.e.* choice of electrode group and frequency band, was done based on the results in Section III. For electrode group and frequency band selection, the power spectrum of the channels was given as input to the model.

D. *Models*

The following models were tried out:

- Random Forest
  We used a standard Random Forest Classifier with a maximum depth of 3 nodes. This shallow depth allows us to use numerous trees (their number being fixed at 300) and still train and evaluate rapidly.
- Logistic Regression
  We used Logistic Regression adjusted using Ridge regularization.
- Convolutional Neural Network

The random forest and logistic regressions were chosen for their short training times and simplicity. Those two models were thus used to perform the features optimization. They approach the representation of the data in two very distinct manners, which allows us to better estimate the relevance of our features for different types of machine learning techniques. The CNN was the final model to which the features optimization supposedly benefits.

III. RESULTS

A. *Features optimization*

The following section evaluates the different combinations of features, *i.e.* electrode groups and frequency bands, by optimizing random forest and logistic regression. The models were optimized on each combination from the recordings of day 4, as they are supposedly better adapted to the classification task. The best model in each case is then evaluated on the remaining days (1, 2, 3, 5) (Figures 4 and 5).

The Random Forest model, used on the clean data of day 4, produces the results shown in Table II. The best combination appears to be Gamma frequencies and Right Temporal electrodes, reaching around 0.9 accuracy. Since the test subset contains a small number of samples (15 samples, 5 cross-validation folds), a slight difference in accuracy should not necessarily be interpreted as significant.

The Logistic regression model, used on the clean data of day 4, produces the results reported in Table III. The best combination appears to be High Gamma frequencies and Right Temporal electrodes, reaching around 0.87 accuracy. The *Param* column indicates the inverse of the regularization weight (the higher the *Param*, the less the model is penalized). As expected, the logistic regression needs that the number of features be limited to perform well on unseen samples.

Using both models, we compared the performances using different groups of electrodes (Figure 2) and different frequency bands (Figure 3). We can conclude that the right temporal lobe area gives the most valuable information among all brain areas considered. The central area gives similar information, while the others drop significantly. The occipital area does not seem to help the classification task. Regarding the frequency ranges, the Gamma band is essential - especially the high gamma frequencies. According to Figure 3, using only the high gamma frequencies instead of both low and high gamma or all frequencies allows to gather almost all of the important information while reducing the number of features.

B. *Convolutional Neural Network*

Two CNN models were trained and evaluated. The Baseline CNN is trained on the *FFT of electrodes difference* input using all electrodes and all frequencies. The dimensions of those images are 1x1281x1891. The Optimized features CNN is trained on the FFT of electrodes difference input but the features are filtered accordingly to our findings during the features optimization phase:

- Right and left temporal electrodes only;
- Applying a bandpass filter of critical frequencies (40Hz, 70Hz) corresponding to the High gamma frequencies;

Frequencies	Electrodes	Train acc	Train std	Acc	Std	Param	n_features
hgamma, lgamma	right temporal	0.9968	0.0063	0.9125	0.0935	2	4509
hgamma, lgamma	right temporal	1	0	0.9125	0.0637	3	4509
all	central	1	0	0.8991	0.0632	3	9519
all	right temporal	1	0	0.8983	0.0853	2	4509
hgamma, lgamma	right temporal	0.9494	0.0116	0.8875	0.1000	1	4509

TABLE II
BEST COMBINATIONS WITH RANDOM FOREST.

Frequencies	Electrodes	Train acc	Train std	Acc	Std	Param	n_features
hgamma	right temporal	1	0	0.8733	0.0685	10	4509
hgamma	right temporal	1	0	0.8733	0.0685	100	4509
hgamma	central	1	0	0.8616	0.0911	10	9519
hgamma	right temporal	0.9810	0.0063	0.8608	0.0916	1	4509
hgamma	central	1	0	0.8491	0.0927	100	9519

TABLE III
BEST COMBINATIONS WITH LOGISTIC REGRESSION.

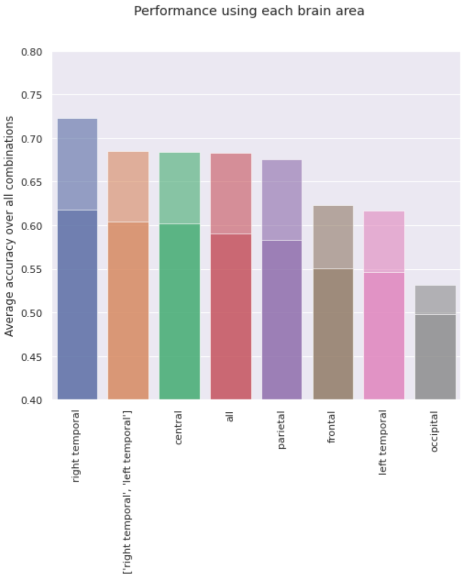

Fig. 2. Average accuracy over all combinations of frequencies and model hyperparameters, for different selections of brain areas. The random forest yields the upper accuracy for every area, while the lower is logistic regression.

- Keeping only the coefficients of the spectrum corresponding to frequencies 25Hz to 70Hz after the FFT has been computed. While frequencies 25Hz to 40Hz have been filtered, keeping them keeps the images dimensions closer to a square image.

The dimensions of the images for the optimized features CNN are 1x225x153.

Both models use the same type of architectures and are built upon the same convolutional blocks. However, the baseline CNN includes more of these blocks as its input images have larger dimensions, requiring more downsampling before the linear layers that constitute its classification head. The difference in architecture did not play any role in the difference in performances between the models, as the baseline CNN could not yield better accuracies even with fewer weights and filters. We compare the models in the following manner:

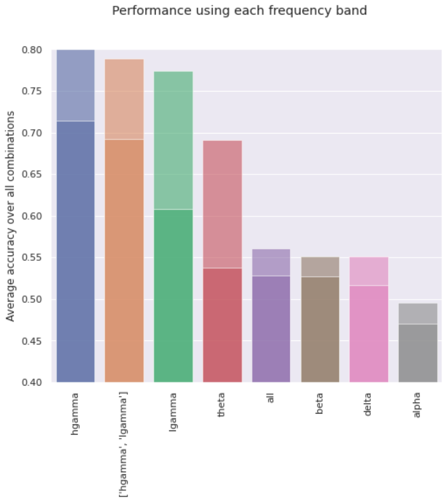

Fig. 3. Average accuracy over all combinations of electrodes and model hyperparameters, for different frequency bands. Once again, the Random Forest classifier is the upper accuracies and logistic regression is the lower ones.

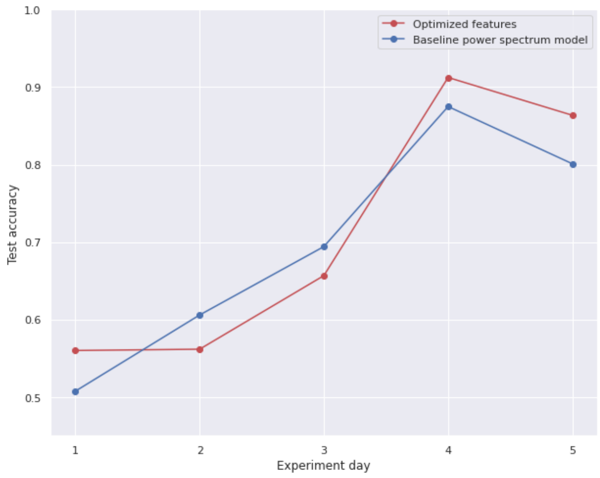

Fig. 4. Evaluation of a random forest with optimal features and maximum depth on each day. The optimized features do not significantly change the results, which is result of the random forest's ability to avoid overfitting.

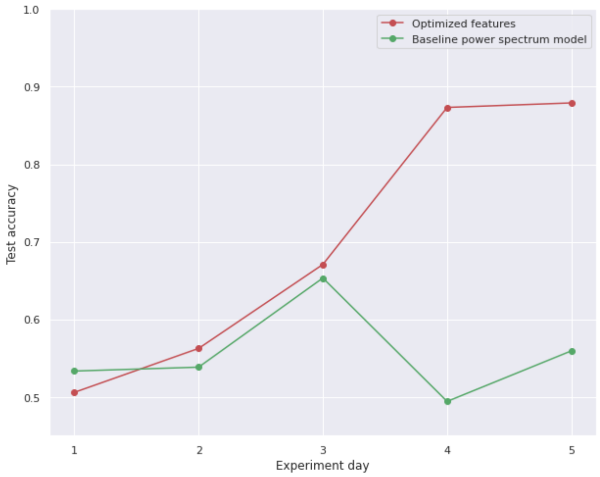

Fig. 5. Evaluation of a logistic regression with optimal features and regularization parameter on each day. Contrary to the random forest, the logistic regression is unable to correctly predict unseen samples if all frequencies and electrodes are used. The optimized features succeed in bringing enough valuable information to almost catch up to the best random forests' accuracies while considerably reducing the amount of features.

- Each CNN is trained and tested on each day independently using 4-fold cross-validation;
- The Binary Cross-entropy is used as cost function.
- The training phase lasts for 20 training epochs, which was found to be sufficient to reach optimal test results.
- For each day, we gather the mean training loss, test loss, accuracy as well as their standard deviations.
- No dropout is used in either model.

An NVidia RTX 3060 Laptop was used to train the models. The Optimized features CNN is significantly quicker to train as it includes less parameters and takes smaller inputs, requiring only a few seconds to complete 20 epochs. The results of the comparison are given in Figures 6 and 7. The Optimized features CNN performed significantly better overall, reaching accuracies superior to those of the best random forests on days 2 and 5. It still performed worse than the random forests on the other days, even though those are much simpler models. Nonetheless it is likely, given those results, that a CNN can outperform the simpler algorithms by fine-tuning its number of weights and using specific layers such as dropout.

IV. Discussion

Through this work we were able to identify the important features for decoding covert speech in terms of frequency bands and electrode locations. We showed that simple Random Forest models are able to reach accuracies of over 90% in the precise conditions of the experiments, with subject training over several days. Eventually, we showcased the beneficial effect of using the best specific frequencies and electrodes to help logistic regressions but also a convolutional neural network to avoid overfitting. Nevertheless, the conditions of the study are extremely precise and testing the same method on another subject is critical to confirm the efficiency of the method. Eventually, the possibility of using such models in practical application remains a question, which requires online testing to be answered.

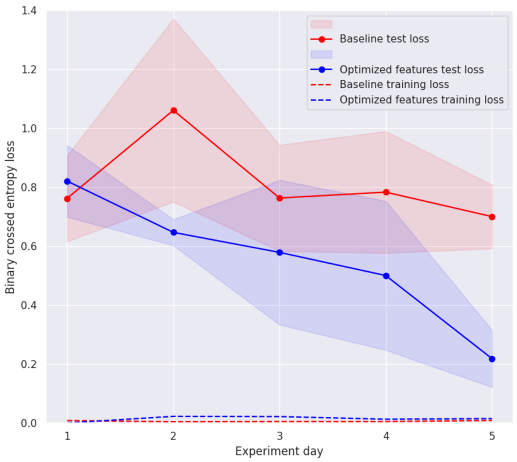

Fig. 6. Comparison of the two CNN models in terms of loss.

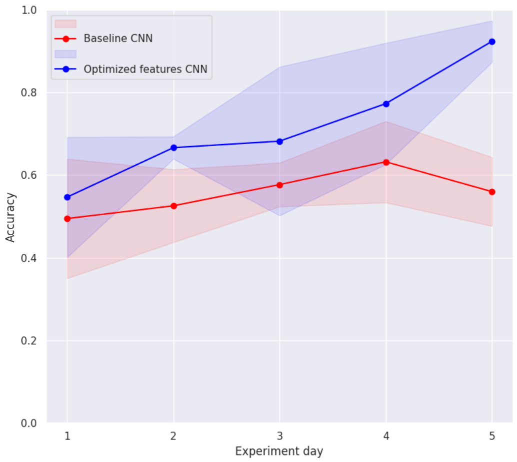

Fig. 7. Comparison of the two CNN models in terms of prediction accuracy.

Acknowledgements

We would like to thank our supervisors Kinkini Bhadra and Shizhe Wu for their precious help, as well as Prof. Anne-Lise Giraud for giving us the opportunity to work on an interesting research application of Machine Learning.

References

[1] C. Cooney, R. Folli, and D. Coyle, "Neurolinguistics research advancing development of a direct-speech brain-computer interface," *iScience*, vol. 8, pp. 103–125, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2589004218301512

[2] J. T. Panachakel and A. G. Ramakrishnan, "Decoding covert speech from eeg-a comprehensive review," *Frontiers in Neuroscience*, vol. 15, p. 392, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2021.642251

[3] S. Vaid, P. Singh, and C. Kaur, “Eeg signal analysis for bci interface: A review,” in *2015 Fifth International Conference on Advanced Computing Communication Technologies*, 2015, pp. 143–147.

[4] G. A. Ojemann, “Brain organization for language from the perspective of electrical stimulation mapping,” *Behavioral and Brain Sciences*, vol. 6, no. 2, p. 189–206, 1983.

[5] L. C. Sarmiento, S. Villamizar, O. López, A. C. Collazos, J. Sarmiento, and J. B. Rodríguez, “Recognition of eeg signals from imagined vowels using deep learning methods,” *Sensors*, vol. 21, no. 19, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/19/6503