EEG-Based Brain-Computer Interfaces

Rita Guilherme Matias Carvalho Barreiros
rita.barreiros@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

September 2020

Abstract

Communicating with the mind seems something futuristic and taken from a science-fiction novel. However, machines which are able to decode brain signals and assign them to specific outputs, such as moving a robotic arm, or typing words on a screen, already exist. Such machines are called Brain-Computer Interfaces. In this work, we aim to find measurable differences between two simple forms of communication: conveying *yes* or *no*, in a direct-recognition approach, i.e., by recognizing a signature of the actual answer. 22 subjects (ages between 23 and 60) underwent a visual task where they answered to *yes* and *no* questions, as their EEG recordings were gathered. The recordings were processed using, among others, an automatic artifact removal algorithm (Artifact Subspace Reconstruction). The Common Spatial Patterns method was then employed to design spatial filters (by maximally differentiating the variance of each class) and to build the input feature vectors. These input vectors were classified using SVMs whose parameters were tuned through the gridsearch method, and the results were plotted in a 2D graphical representation which displayed the classification accuracy by frequency and time of the classified trials. Finally, the features which presented the highest classification accuracy on the training sets, were used to evaluate the performance of the classifier on new data. The classification accuracy of the test sets failed to reach the theoretical chance level (65%), even when the classification accuracy on the training sets was over 86% for all analysis. Our analysis, following this methodology, failed to find detectable differences between these two answers.

Keywords: Brain Computer Interfaces - Electroencephalogram - EEG Acquisition - Common Spatial Patterns - Binary Classification

Communication is a human basic need, of which we make an everyday use without even thinking about it. Alas, there are people who are not able to communicate, such as the ones who suffer from a condition called Complete Locked-In Syndrome. In such a state, a person is made hostage in their body, unable to talk, to move, to even move their eyelids.

Fortunately, there exist assistive systems, called Brain-Computer Interfaces (BCIs), which allow the interaction of a person with a computer by interpreting their brain signals and assigning them to a specific output. Such devices make use of different brain imaging techniques to decode a user intention and link it to a physical outcome. However, often the outcome of these devices is not related with the desired intention. For instance, there are BCIs which link the imagination of a left-arm movement with a specific word, and the right-arm movement with another. This is an indirect approach, in the sense that the task required from the BCI user does not reflect one's own will.

In this work, we are interested in a more direct approach, which facilitates two simple forms of communication: conveying *yes* and *no*. That is, we aim to understand if there is a specific distinguishable signature of a *yes* and a *no*, through the analysis of electroencephalograms (EEGs) of subjects covertly answering to *yes*/*no* questions.

To do so, we designed an EEG-based visual experiment which required simple and direct answers (*yes* or *no*) from the analyzed subjects, who were asked about specific characteristics of the cards from a common card deck.

1. Background

1.1. Electroencephalograms

EEGs measure the brain's electrical activity, generated by the simultaneous firing of specific neurons,

named pyramidal cells.

Neurons are the functioning units of the nervous system [7]. They can be divided into its soma, where the nucleus of the cell and the major organelles lie; the dendrites, which receive information from other neurons at specialized areas of contact; and the axon, which emerges from the soma and projects onto target cells. The cellular membrane chemically insulates the neuron, separating the ionic concentrations in each mean (the intracellular and extracellular means).

When the membrane is at its resting potential, a steady state on which the ionic flow is at an equilibrium, the potential outside the cells is conventionally defined as zero, whereas the potential inside is around -70 mV [8]. Upon an electrical or chemical stimulus, if the potential reaches -60 mV, a cascade of fast-opening Na^+ channels generates a current across the neuron's axon, called the action potential, which is how information flows through a neuron.

Neurons communicate either electrically through gap junctions, or chemically, through synapses. Either by chemical or electrical communication, two types of postsynaptic potentials are elicited: if a depolarization of the postsynaptic membrane happens, there is a production of a postsynaptic action potential (Excitatory Postsynaptic Potential, or EPSP); whereas if a hyperpolarization occurs, the membrane of the postsynaptic cell remains more negative than the threshold for the action potential to occur (Inhibitory Postsynaptic Potential, or IPSP).

Both the generation of EPSPs and IPSPs produce a current flow inside the neuron which, in turn, generate potentials, which are measurable by the EEG electrodes.

The individual voltage changes from each pyramidal neuron, firing simultaneously, sums up, producing a detectable surface EEG.

1.2. Selecting EEG Information

An EEG recording is, in terms of data size, a considerably big file. When developing a BCI, it is important that the acquisition and processing of the EEG, and its association with a certain outcome, is a fast process, to ensure usability in real-time scenarios. Hence, one must select relevant information (which characterize the mental task of interest) from the EEG recordings, ideally a small number of values - the extracted features.

There are two main types of analyzed processes in an EEG-based BCI: oscillatory activities, or Evet-related Potential (ERP) values, and the types of features and analysis one can make differ from one another. In general, the features used in EEG-based BCIs can be categorized in temporal, spectral and spatial features [9].

BCIs based on oscillatory processes use preferentially the spectral content of an EEG as features. Specific brain rhythms are associated with specific mental states, and their modulation can be used for the development of a BCI. These rhythms are the delta band (1-4 Hz), the theta band (4-8 Hz), the alpha band (8-12 Hz) and the gamma band (>25 Hz). It is common to compute the power, or energy, in a specific frequency band (commonly referred to as bandpower). An increase of the power is called an Event-Related Synchronization (ERS), and a decrease is called an Event-Related Desynchronization (ERD).

On the other hand, ERP-based BCIs preferentially use the temporal characteristics of the EEG. An ERP study analyzes the amplitude of the EEG, locked to a specific time point in relation to the stimulus-onset. To do so, a multitude of EEG segments, locked to a stimulus, are averaged to enhance a characteristic waveform of the potential in study. In general, the more trials are averaged, the better will be the noise cancellation, which leaves the waveform more visible.

Both ERP-based and oscillatory-processes-based BCIs may benefit from the spatial information of an EEG. For instance, if the brain area where an ERD occurs when a specific task is performed (such as imagining a left-arm movement, occuring over the motor cortex of the brain) we can choose a specific set of channels to focus on. Sometimes, however, there is no knowledge about the specific brain area which will be engaged in a specific task. An alternative to *a priori* selection of channels, is making use of spatial filtering - making a linear combination of the existing channels to use only a smaller subset of (relevant) electrodes. There are two main categories of spatial filters: data-driven spatial filters, optimized by using data, or fixed spatial filters. The Laplacian and bipolar filters are examples of fixed filters, which exclude channels by geometrically subtracting them from one another. A particular data-driven spatial filtering technique, called Common Spatial Patterns (CSP) is used in this work.

1.3. Brain-Computer Interfaces

A machine-learning-based BCI system can be schematized into two major components, or phases: the calibration phase and the feedback phase [1]. In the first phase, the brain signal of a subject executing a certain mental task is acquired and the features are extracted. Then, a classifier is trained to associate the different features with a specific output. In the feedback phase, sliding windows from a continuously acquired brain signal are used as inputs for the trained classifier, which in turn conveys

the measured input into a desired output.

In this work, we are focused on the analysis of the obtained EEG signal itself, which places us at the design of the calibration phase of the BCI.

EEG presents low signal-to-noise ratio (SNR) and spatial resolution when compared to other imaging techniques (such as fNIRS or fMRI). However, it is the most common imaging technique used in non-invasive BCIs [11]. It directly measures neural activity; it is a portable technique and easy to use (even for non-experts). Besides, and maybe one of the strongest qualities of EEG given the swiftly occurring changes in the brain's electrical activity, it presents a temporal resolution in the order of the milliseconds.

1.3.1 EEG-based Yes/No Discrimination BCIs

Covert (imagined) speech has proven usability for BCI purposes [12]. Suppes et al. [14] recorded the EEGs and magnetoencephalograms of several subjects during covert articulation of seven different words and achieved above chance-level accuracy scores in the classification of such words for most participants; regarding EEG alone, the covert articulation of syllables and vowels have been used as inputs for classifiers [2] [6], also yielding above chance-level accuracy scores. These results support the fact that covert speech may be used in the development of a direct-recognition-approach, EEG-based, BCI.

In the study of Rezazadeh Sereshkeh et al. [12] the reliability of EEG in discerning different covert speech tasks was investigated. They were the first authors to use Artificial Neural Networks (ANNs) for the classification of covert speech measured by EEG over various sessions. Discrete Wavelet Transform (DWT) features were chosen to characterize the EEG trials due to its ability to analyze both time and frequency domains. Common spatial patterns and autoregressive components were also analyzed as input features, but the DWT features alone yielded the highest cross-validation classification accuracy on the training set.

Choi and Kim [5] wanted to understand the differences between the intention to answer *yes* or *no* when self-referential questions were made to subjects. Firstly, they identified in their ERP study [4] that, when asked self-referencing questions, there is an integration of semantic and autobiographical information processing that precedes the answers. The occurrence of the posterior N400 potential (at 300-500 ms) was associated with the mentioned information processing. Hence, they expected that the decision and intention to answer either *yes* or *no* would occur either simultaneously or immediately after the occurrence of the N400; they also hypothesized that the activities of the brain upon the decision to answer, in working memory, could be identified during this period of time, after the information processing and before the actual covert answer, on what they called "a period of retaining intention in mind".

In their experiment a series of self-referential questions was presented word by word, the last (or critical) word being the one that either violated the autobiographical fact (*no* answer) or did not (*yes* answer). After showing the critical word, a blank screen appeared until a cue to answer to the question ("Please respond") appeared on the screen. It was the period between the critical word and the answering cue that the authors hypothesized to be the retaining intention period, and it was this period that deserved the focus of their work.

They analyzed the single-trial EEGs in a -500 to 1300-milliseconds interval (0 being the critical word onset) using CSP [5]. This technique designs the spatial filters that optimize the discrimination of single band-passed EEG trials from different classes based on their variance. And because the variance of a band-passed EEG trial is equal to its band-power in that frequency [1], their approach aims to find the maximally different frequency bands (and time) for each response. They found two time-frequency regions that contained what they called "useful information for the yes/no" decoding, for these were the bands where the differences were most noticeable: in early theta (in the frontal region at 200-500 ms) and late alpha bands (in the right parietal regions at 800-1200 ms).

2. Methods and implementation

We adopted the feature extraction and classification methodology of Choi and Kim [5], to have a basis of comparison and to see if any similarities in the results could be found. The two most evident differences between our goals and theirs were: they focused on the intention of answering *yes* or *no* rather that the answer itself; and they used self-referencing questions, which may be crucial in their findings (e.g. the appearance of the posterior N400 upon self-referencing questions).

2.1. Participants

There were 22 participants, 16 females and 6 males, with ages between 23-60 years-old. All subjects were of Portuguese Nationality.

After subject-exclusion, approximately one third of the subject sample was discarded. The remaining sample was composed of 17 participants (ID0 to ID16), ages between 23-29 years-old (average age: 25 years-old), 11 females and 6 males. 15 partici-

pants were right-handed and the remaining participants were left-handed. All participants signed a consent form and a simple questionnaire, by which general information (age, gender, dominant hand) was collected as well as relevant information for the experiment at hand.

The recordings of the 17 subjects were acquired in 4 different days (recording sessions). In the first recording session participated 3 subjects; in the second 5 did; in the third session, 4 subjects participated; and in the forth session, 5 subjects did.

2.2. Experiment Design

In the experiment of this work, we made use of the images of the 36 numbered cards of a common card deck.

The experiment was composed of three consecutive runs, with a small break between them. Subjects were sitting in front of a laptop (at approximately 80 cm from the screen), where the images and questions of the experiment were shown, while their EEG was being recorded.

On the first run, subjects were orally instructed, at the beginning of the experiment, to answer the question "Is the card on the screen red?". The cards were randomly displayed at constant intervals, preserving an even number of red cards (*yes* answers) and black cards (*no* answers). On the second and third runs, subjects were instructed to answer each question that would appear written on the screen. These were questions regarding the number and suit of the proceeding card (e.g. "Is the card a 4 of diamonds?").

On the second run, to reduce any biases, we made sure that whenever the card shown violated the preceding question (expected *no* answer), the number and suit of the card would be different from the one on the question (e.g. if the subject was shown the question "Is the card a 4 of diamonds?", the shown card must not have been a number 4 nor from the suits of diamonds). Run three of the experiment was very similar to run two except that this time, when a *no* answer was expected, the shown card would have the same number, but a different suit from the card mentioned at the preceding question (e.g. if the question was "Is the card a 4 of diamonds", the proceeding card would be a 4 from a different suit).

All the questions were written (or orally instructed in the case of Run 1) in Portuguese. Each run would begin with the screen with a dark grey background for 4-seconds, after which either the card (if on Run 1) or the question (if on Run 2 or 3) was shown. On the first run, each card was separated by the dark grey background, displayed for 4-seconds. On the second and third runs, the 4-second interval occurred between questions and cards. Both questions and cards were displayed for 2-seconds.

At the end of each run of the experiment, each subject had answered to 18 *yes* questions, and 18 *no* questions.

2.3. Software and Hardware

For the EEG recordings we used the openBCI© Cyton + Daisy Biosensing Boards, which allows sampling 16-channel-EEG data with a sampling frequency of 125 Hz. Each input channel was connected to a dry electrode which was attached to an elastic cap, following the international 10-20 system. Two other electrodes were used for reference, and were attached to the ear lobes of the subjects with tape and a conductive gel, Ten20©. In total, 18 electrodes were used. A laptop was used to run the openBCI© software and connect the hardware. The connection of the boards with the laptop was ensured by a Bluetooth dongle, from the same provider.

The openBCI© software lets any user develop their own *widget* - an interface that allows the interaction of the user with the recordings. In this work, we developed our own custom *widget*, specifically designed to show the images of the cards and the questions, while recording the EEG of the subjects. The development of the *widget* was made in `Processing` language.

2.4. EEG Preprocessing

Preprocessing an EEG signal involves a series of important steps that reduce the noise of the data, increasing the SNR of the EEG, and preparing it for further analysis.

Fixed-frequency artifacts can be detected by computing the FFT of the EEG recording. When doing so, we detected a peak at 50 Hz (the powerline noise frequency in Europe), as well as two other peaks: one at 25 Hz and another at 32.15 Hz, across the majority of the recordings. A notch filter was used twice to remove the above-mentioned frequencies of the EEG signals of all recordings. A 2^{nd} order butterworth high-pass filter, with a cutoff frequency of 0.5 Hz was used to remove low frequency artifacts, such as the ones caused from sweat. Flat and noisy channels were also removed.

Movement-induced artifacts (eye blinking, jaw clenching, heart rate, muscular activity) are usually harder to handle, for their frequency varies, and are inconsistent over time (due to their nature). We used an automatic approach to remove such artifacts, called Artifact Subspace Reconstruction (ASR). This method automatically identifies clean portions of the data, which are used as a reference for the remaining analysis, and rejects large-variance components. The cleaned data is recon-

structed from remaining components. ASR is available as a plugin from EEGLAB, a `Matlab` toolbox used for neuroimaging research, which we used in this work.

The most important parameter of ASR is the "cutoff" parameter k, which dictates how aggressively the data is cleaned. Small values of k result in a strict cleaning, meaning that some of the eliminated components might correspond to brain-sources signal, whereas very high values of k results in a less aggressive cleaning. To understand how much the values of k affected the data, we evaluated how much data was modified and how much data variance was reduced, for different values of k [3]. For $k \in [20, 40]$ the average percentage of modified data lie in the interval $[20, 40]$. The value of k was set 30.

After cleaning the data, all files were normalized, leaving the measured voltages comprised in the interval $[-1, 1]$.

The EEG recordings were then segmented into 5-second long pieces of the EEG recording, from 1-second before the stimulus-onset (the appearance of a card on the screen) to 4-second after the stimulus.

2.5. Subject Exclusion

The EEG data file generated by the openBCI© software includes a time-stamp and a sample index, for each sample. The former was used to synchronize the data with the stimuli. However, as the experiments took place and the recordings were analyzed, we found some irregularities regarding this times-stamp: several recordings yield the same time-stamp for several samples, which would not allow us to time-lock the stimuli. The recordings of 6 subjects were discarded because they exhibited very irregular sample rates (either several samples being recorded at the same instant, or more than 200 ms separating consecutive samples).

The sample index is an integer which consecutively accompanies each sample (cyclically repeating from 0 to 127). Hence, if any sample is loss, a non-consecutive line of integers is found. The first run of ID13 was discarded because 30% of its data had been lost. Some EEG segments were also discarded, whenever any sample within that segment had been lost.

2.6. Common Spatial Patterns

CSP is a supervised data-driven spatial filtering technique. To apply a spatial filter to an EEG signal is to use a smaller subset of the channels, defined as a linear combination of the original complete set [9]. For a given EEG trial:

$$x^s = \sum_i w_i x_i = wX \tag{1}$$

where x^s is the spatially filtered signal, x_i is the original EEG trial and the w_i are the spatial filters.

Considering a band-passed EEG single trial ($X_i \in \mathbb{R}^{C \times N}$, C being the number of channels and N the number of data samples) - in this case, the set of samples which corresponds to one expected *yes* or *no* answer - the goal of CSP is to find a set of spatial filters (W) that, when used to project the EEG trial, yield maximum variance for the EEG trials in one condition (class 1, *yes*) and minimum variance for the other (class 2, *no*). The way CSP does it, is by simultaneously diagonalizing the covariance matrices for class 1 ($\Sigma^{(1)} \in \mathbb{R}^{C \times C}$) and class 2 ($\Sigma^{(2)} \in \mathbb{R}^{C \times C}$):

$$\begin{aligned} W^T \Sigma^{(1)} W &= \Lambda^{(1)} \\ W^T \Sigma^{(2)} W &= \Lambda^{(2)} \end{aligned} \tag{2}$$

subject to the restriction $\Lambda^{(1)} + \Lambda^{(2)} = I$. The covariance matrices for each class are computed as follows [17]:

$$\Sigma^{(k)} = \frac{1}{|\phi^{(k)}|} \sum_{i \in \phi^{(k)}} X_i {X_i}^T \tag{3}$$

ϕ^k is the set of trials that belongs to each class, and $|\phi^{(k)}|$ denotes the size of each set, i.e., the number of EEG single-trials in each class.

Equation (2) can be put into the form of a generalized eigenvalue problem:

$$\Sigma^{(1)} w_j = \lambda_j \Sigma^{(2)} w_j \tag{4}$$

The w_j ($j = 1, ..., C$) in equation (4) are the generalized eigenvectors, corresponding to the columns of W in equation (2). $\lambda_{k,j}$ are the generalized eigenvalues and the diagonal elements of Λ^k (2), defined as $w_j^T \Sigma^{(k)} w_j$. In equation (4), λ_j is equal to $\frac{\lambda_j^{(1)}}{\lambda_j^{(2)}}$. The eigenvalues are constricted between 0 and 1 and reflect the variance of each class. That is, a large (close to 1) $\lambda_j^{(1)}$ tells us that the EEG trial, projected by the spatial filter w_j shows a high variance if from class 1 and a low variance if from class 2. It is the difference between the variances of each class in the projected space that allows their discrimination.

In `Matlab`, the eigenvalue problem of equation (4) can be simply solved by the command `[W,D] = eig(`$\Sigma^{(1)}$`,`$\Sigma^{(1)} + \Sigma^{(2)}$`)`. W is the matrix whose columns are the spatial filters, sorted by ascending order of the respecting eigenvalues. That is, the first column of W is the spatial filter that yields the highest variance for class 1 and lowest variance for class 2. D is the matrix whose diagonal is composed by the eigenvalues associated with the eigenvectors that form W. Due to the sorting order, the first values of diag(D) will be the ones that are closest

to one, and the last values of diag(D) will be the ones closest to 0, which means that the first and last eigenvectors, i.e. first and last columns of W, will be the ones which yield the most discriminating power, for they are the ones that maximize the difference of the variance between classes.

We can then obtain the projected signal, $Z_i \in \mathbb{R}^{C\times N}$ by multiplying the original EEG single trial by the spatial filters:

$$\begin{aligned} Z_i^{(1)} &= W^T X_i^{(1)} \\ Z_i^{(2)} &= W^T X_i^{(2)} \end{aligned} \tag{5}$$

These signals are now projected into the space that yields the highest variance for class 1 and lowest variance for class 2.

2.6.1 Feature Computation

The features that will be used to characterize the EEG data, will be the normalized variance of the rows of the projected signal, since these will be the features that, by the construction of the method, best discriminate the two classes:

$$f_p = log\left(\frac{var(Z_p)}{\sum_{i=1}^{2m} var(Z_i)}\right), \quad p = 1, ..., 6 \tag{6}$$

The *log* transform is used to normalize the distribution of the data. We used 6 spatial filters, which means that each EEG trial will lead to a 6-dimension feature vector ($p = 1, ..6$) to be used as input for the classifier.

2.7. Classification

Similarly to Choi and Kim [5], the CSP features were computed for time-frequency (TF) subwindows of $200ms \times 2Hz$ of the 5-second long EEG segments, in order to find which subwindow would yield the highest discrimination power between two classes (*yes* and *no*). The procedure for the subwindow division is described below.

A 2$^{\text{nd}}$ order bandpass filter of 2 Hz bandwidth was used to consecutively obtain the 2 Hz frequency subwindows (from 0.5 Hz to 60.5 Hz), resulting in 30 band-passed EEG recordings. After segmentation, each 5-second bandpass-filtered EEG recording was subdivided in 200-millisecond time intervals, resulting in 25 time-based EEG subsegments for each band-passed EEG recording (5 s / 200 ms = 25). The CSP features were computed for each of the 750 (30 × 25) time-frequency (TF) subwindows.

Each EEG segment then resulted in a six-dimensional feature vector for each subwindow. These feature vectors will be the input of an SVM model, built using the `scikit-learn` libraries, whose output will be either *yes* or *no*.

The SVM's hyperparameters were optimized using gridsearch, in which the C and γ parameters of the SVM were tested for each time-frequency subwindow, as well as two kernels (linear and radial basis function). Thus, a specialized model was developed for each TF subwindow. The tested values of C parameter were 10^{-3}, 10^{-2}, 10^{-1}, 10, 10^{1}, 10^{2}, and for of γ were 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1.

A training set was used to design the spatial filters; a validation set was used for the hyperparameter optimization of the SVM model; a test set was used to evaluate the performance of the developed SVM model. The training:test:validation ratio of for each complete set of trials (each run of each subject) was 24:6:6, that is, 24 trials were used for training, 6 for validation and 6 for test. For the subjects whose EEG trials were discarded the following ratio was applied: 24:6:5 (when only one trial was discarded); 22:5:6 (when 3 trials were discarded).

In all sets, the balance of classes (number of *yes* and *no* trials) was taken into account, that is, we used, when possible, the same number of trials from each class on each set. When not possible (due to segment discarding), the difference between the number of trials from each class was never larger than one, and when it occurred it was on the test set.

2.8. Performance Evaluation

Considering that the two classes were balanced, we chose to use the classification accuracy (the number of correct predictions over the total number of predictions) of the SVM as the metric for the performance of the model.

For a 2-class BCI with 20 trials per class, the chance-level accuracy (for a confidence level of 95%) was 65% [10]. This is a reference for the discussion of our results, bearing in mind that in our experiment there were slightly fewer trials than the ones for which the chance-level accuracy was determined (18 per class, instead of 20).

2.9. Classification Tasks

Once the hyperparameters of the classifier were tuned for each TF interval, we computed the classification accuracy of the training and test sets for each subwindow. The results were plotted in a 2D colored map, the y-axis representing the frequency and the x-axis the time, in relation to the stimulus. Colors represent the accuracy scores: values of classification accuracy equal to and beneath 50% are represented by a dark blue tone; increasing yet still low scores are represented by light blue and green; warmer tones (yellow, orange and red) represent higher values of classification accuracy (from around 82% up to 100%).

The main line of analysis was, as follows: visu-

alization of the achieved accuracy for each time-frequency subwindow of the training set; selection of the most relevant subwindows on the training set; visualization of the accuracy on the test set when using the selected subwindows. We aimed to find whether the reported TF subwindows identified in Choi and Kim [5] would be depicted, or not. Moreover, we analyzed if the subwindows which yielded higher classification accuracy scores in the training sets would also yield high classification accuracy scores in the test set.

The analysis was applied to four different groups of EEG results: **individual**, where each run of each subject was contemplated; **intersubject**, where the average classification accuracy across subjects, of each run, was contemplated; **intersession**, where the classification accuracy results were averaged across recording sessions; **interrun**, where the classification accuracy results of each subject were averaged across runs.

3. Results and discussion

3.1. Colorplots and inspection of relevant features

As Choi and Kim [5] the selection of most relevant subwindows were defined, as follows: the subwindows for which the classification accuracy was both above the theoretical chance-level classification accuracy (65% in the case of this work), and above the mean for all the TF scores, plus 2 standard deviation, were selected as the most relevant. There were cases where the second criterion lead to a threshold above 100%. In those cases, this criterion was updated to choosing the TF subwindows for which the classification accuracy was equal to 100%. After the selection of the important TF subwindows, the scores of the test sets for those subwindows were analyzed and will be further discussed.

The shown colorplots are representative of the entire sample of subjects, since the differences between them was not significant. Moreover, we shall only present the colorplot of Run 1, for the same reason.

3.1.1 Individual EEG results

Figure 1 shows the individual colorplot produced for the training sets of Run 1 of subject ID0.

Only TF subwindows with light blue or warmer coloring lie above the theoretical chance-level accuracy score of 65%.

There is no particularly significant TF for which the discrimination between *yes* and *no* is better or worse than any other one, except for sparsely distributed subwindows for which the classification accuracy is lower.

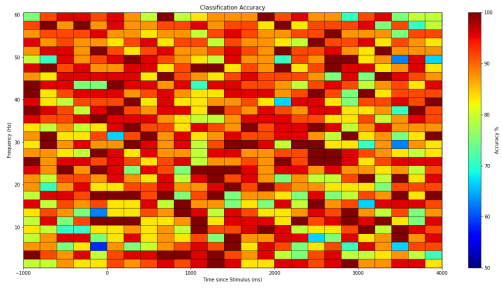

Figure 1: Colormaps of the classification accuracy of the training set of Run 1 of subject I*D*0.

Indeed, the mean classification accuracy for each run (1,2 and 3), averaged across all TF subwindows, is 87.77% ± 1.39%, 87.43% ± 1.50% and 88.32% ± 1.81%, respectively.

Figure 2 shows the produced colorplot for the same run of the same subject, this time for the test set.

The overall classification accuracy is considerably lower for most TF subwindows when comparing with the results for the training set: 50.53% ± 1.56%, 50.33% ± 1.50% and 50.76% ± 2.58% for each run of the experiment, below the 65% chance level. Results of this kind suggest that the model might have overfit the training data, since it shows such different (much higher) classification scores than the test set does. Hence, it fails to predict the correct labels (*yes* or *no*) of unseen, new data.

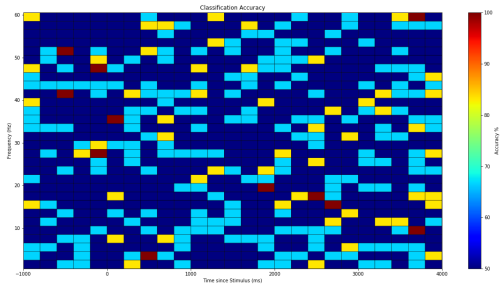

Figure 2: Colormap of the classification accuracy of the test set of Run 1 of subject ID0.

3.2. Intersubject EEG results

In an attempt to find a more restricted set of relevant subwindows, as well as to see if there is a particular set of subwindows which yields good discrimination power among all subjects, we averaged the classification accuracy scores of all the TF subwindows for each run of every subject. **Figure 3** shows the averaged colorplot of Run 1 for the training sets of all subjects.

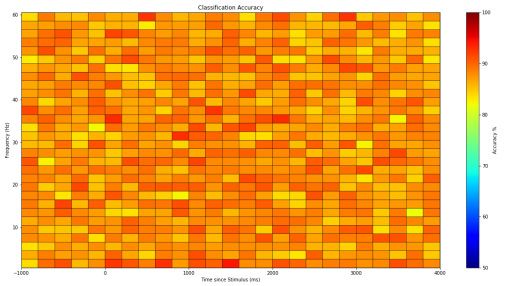

Figure 3: Colormaps of the classification accuracy of the training set of Run 1, averaged across all subjects.

Compared with the individual colorplots, there are no TF subwindows which yield scores under 80%. At the same time, there are no subwindows which yield higher scores like those of the individual colorplots. This is, of course, expected, since averaging led to a reduction of the maximums and minimums.

Similarly to the methodology applied to the individual results, the criteria for features selection was applied: only the subwindows which presented a classification accuracy of more than the mean of the TF subwindows scores for the average across subjects, plus two standard deviation, were selected (91.98% for Run 1, 91.74% for Run 2 and 90.75% for Run 3). **Figure 4** shows the colorplot of the same average, relative to the test sets.

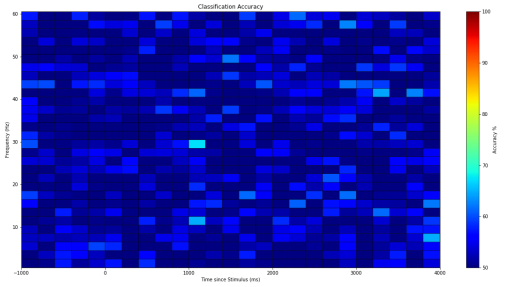

Figure 4: Colormap of the classification accuracy of the test set of Run 1, averaged across all subjects.

Compared with the individual results, very few subwindows that met the criteria were found: only 9 for the first run, 11 for the second run, and 14 for the third run. This is not something which is disappointing. Rather, if we could find a small set of subwindows which yield high decoding power, it would more closely resemble the results of Choi and Kim [5] who found only 3 subwindows that met their criteria (even if their focus was different).

As for the produced colorplots, the average classification accuracy in the test set was no higher than 70%, and the features that yield this classification were a minority (only 2, from the existing 750). Which leads us to say that we could not find, by this methodology, a set of subwindows that yield a good discrimination power across all subjects. Instead, if such can be said, the intersubject (and possibly intersession) variability might be too great to allow for the finding of similarities between the mental answers of different people.

The average classification accuracy for the test set, for the selected subwindows, was very much below the theoretical chance-level classification accuracy of 65%: 49.98% ± 4.51% for Run 1, 49.51% ± 2.83% for Run 2 and 49.61% ± 2.94% for Run 3.

3.3. Intersession EEG results

EEG is known for its intersession variability, as mentioned in the above paragraphs. In an attempt to understand whether the low scores of the intersubject results were due to this variability, we grouped subjects per recording session. The results of the first run for the 4 different recording sessions of the training sets, averaged across the subjects in each session are shown below **Figure 5**. We decided to include in this discussion the results of only one session because the results were visually similar between sessions.

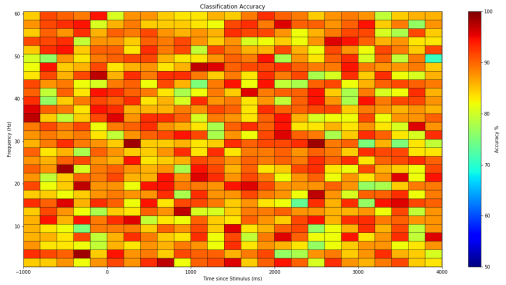

Figure 5: Colormap of the classification accuracy of the training set of Run 1, averaged across the subjects in the first recording session.

Comparing with the intersubject results, there are more subwindows which yield high classification accuracy scores, as well as the ones that yield low scores. This is true for all sessions, which include different numbers of subjects.

Following the line of the previous analysis, we computed the classification accuracy of the selected subwindows for the test sets averaged across the subjects in each session. In this article, we will only show the average classification accuracy across runs for each session: 87.63% for session 1, 86.52% for session 2, 87.09% for session 3 and 87.71% for

session 4.

The colorplots for the test sets are presented in **Figure 6**.

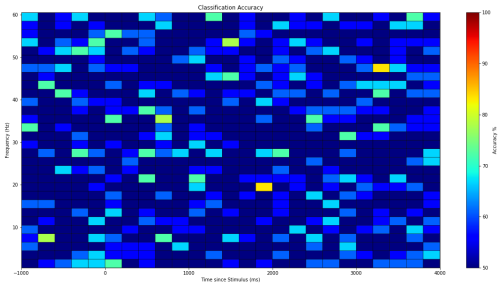

Figure 6: Colormap of the classification accuracy of the test set of Run 1, averaged across the subjects in the first recording session.

Similarly to the colorplots of the training sets, test set colorplots are less smooth than the intersubject colorplots. There are many more TF subwindows whose classification accuracy lies above the 65% chance level score, although only a very small subset of features for each recording session show scores above 75%. It seems that the intersession variability does not have a great influence (at least, not one that can be detected by this analysis), and the slightly better classification scores for some TF subwindows might be caused by the smaller number of subjects for whom the average is computed (averaging across more subjects leads to "smoother" colorplots, such as the ones for the average across all subjects).

Classification scores for the test set of each session were, on average across runs: 49.82% for session 1, 52.26% for session 2, 51.33% for session 3 and 50.36% for session 4.

The number of features is of the same order, as well as the average classification accuracy than for the intersubject results. The impact of intersession variability is not conclusive from this analysis.

3.4. Interrun EEG results

The runs, despite requiring the same answer (*yes* or *no*), were different from one another. We were then expecting to find some differences in the results of each run, possibly due to different cognitive loads required (which can be reflected in different values of bandpower). Bearing that in mind, it might also happen that a *yes* and *no* answer would be independent of the context of each run.

Below are the results for the training sets of subject ID0, average across runs (**Figure 7**).

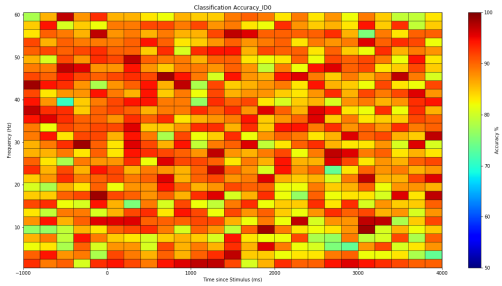

Figure 7: Colormap of the classification accuracy of the training set of subject ID0, averaged across all three runs of the experiment.

This colorplot has a very similar appearance to the individual colorplot for the same subject (1). The values of the classification accuracy of the training set of subject ID0 do not vary much between individual and interrun results (an average classification accuracy of 87.10% across subjects).

Figure 8 shows the colorplots for the test set of ID0 for the Interrun analysis.

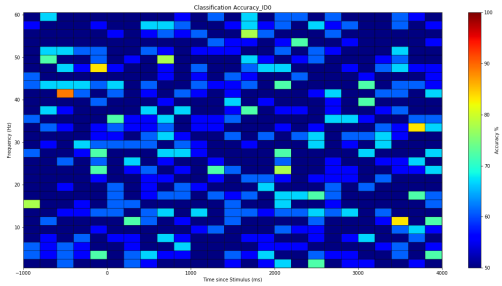

Figure 8: Colormap of the classification accuracy of the test set of subject ID0, averaged across all three runs of the experiment.

And as well as for the colorplot of the training set, the distribution of scores for the test set in this analysis are similar to the ones for the individual results of the same subject. The classification accuracy for the selected features is once again below the chance-level results (an average of 50.34% across subjects). This analysis does not allow us to conclude whether answering *yes* or *no* is dependent on the context or variables involved in the questions asked.

4. Global Discussion

The different analyses presented in the previous section showed that the proposed task was not fulfilled. The developed classifiers seem to overfit the data, for they fail to predict unseen data, i.e. the classification accuracy drops below the chance-level

threshold of 65% when predicting the labels of the test data. Even when one looks at the results of the training sets (whether individual, intersubject, intersession or interrun), no obvious pattern was found, that is, the selected relevant subwindows which yield the most discrimination power (whose criteria are based in the achievement of high classification accuracy) are different from subject to subject, and when individual results are observed, too many subwindows fulfill the criteria, inconsistently between subjects. One should mention, however, that only an analysis of averages might fail to find more complex connections between the data.

In an attempt to avoid "data leakage" we partitioned the data sets into a training, validation, and test set, the second set only to be used to tune the hyperparameters of the SVM. This approach has the downside of reducing the (already small) dataset, which makes the classifier prone to overfit the data, the opposite of what we aimed for by using this approach. A possible alternative would be to use cross-validation to a concatenated training and validation set, when tuning the classifier, so as not to waste any samples.

Also to avoid overfitting, fewer spatial filters can be selected, to characterize the data, which would lead to lower-dimensional feature vectors [13]. A feature selection algorithm could be used instead of setting a fixed number *a priori*, which could aid in the setting of the optimum number of filters to use.

Moreover, CSP is sensitive to noise, which could explain why no congruent results were found, given the probability of noise being captured instead of real brain signals. There are some alternative implementations of CSP that circumvent this issue [9],[1], which, if more time were available, could have made the employed approach more robust.

The number of subjects, and EEG trials of each run of the experiment, were very small when compared to the literature [5], [15], [16] which might not be enough for the classifier to accurately learn the data. The choice of the number of trials was made to avoid the subject's fatigue. Several subjects reported, even with this number of trials per experiment (36 in each run which led, approximately, to a 5-minute long first run and a 10-minute long second and third runs), that they felt unfocused once they had learned how to respond to the very simple tasks. In particular, they mentioned that the intervals between cards and questions were too long. Future tuning of these time intervals might be important to avoid this kind of fatigue, since such a mental state affects the bandpower of the signal.

Another very important limitation was the fact that we were unsure of the meaning of the timestamps used for synchronization. We assumed, so as to be able to conduct any kind of analysis, that the recorded time-stamps were accurately being acquired. However, it is possible that there was some desynchronization which would mean that we might be capturing different moments of the mental answer of the subjects.

5. Conclusions

The task proposed by this work was to find a signature for a *yes* and a *no* when recorded from an electroencephalogram. We aimed to use an approach which would allow us to study the EEG in a large scope, similarly to what we found in the approach followed by Choi and Kim [5].

Through CSP features, we tried to analyze which particular frequency and time held maximally different bandpowers between classes, in order to verify if these were different between a *yes* and a *no*.

We analyzed both individual and grouped accuracy scores (per total number of subjects, per recording session and per run of the experiment). The achieved classification accuracy for the training sets were over 86%. However, for the test sets the achieved classification accuracy was under 59%, not reaching the theoretical chance-level reference of 65%, irrespective of attended analysis.

The performed analysis were not enough to find measurable differences between a *yes* and a *no*, although the failure of such a task was influenced by several limitations, as mentioned in the above section.

In the future, an interesting approach would be to investigate the performance of other classifiers and features. As mentioned throughout this work, we chose CSP features as this is a common technique used in several BCI studies, which provides a discrimination based on bandpower features. However, there are other features which account for both time and frequency domains of an EEG, such as DWT, which were used in the work of Rezazadeh Sereshkeh et al. [12]. They successfully verified the usability of ANNs in the discrimination of covert speech across multiple sessions. They did not analyze single trial EEGs, rather they ran through the ANNs, the DWT features collected from the covert repetition of *yes*, *no* and *rest*. Hence, it would be an interesting approach to understand the performance of ANNs with single trial EEGs, such as the ones acquired in our work.

Acknowledgements

This document was written and made publically available as an institutional academic requirement and as a part of the evaluation of the MSc thesis in Biomedical Engineering of the author at Instituto Superior Técnico. The work described herein was performed at INESC-ID's Spoken Language Sys-

tems Laboratory, Instituto Superior Técnico, Universidade de Lisboa, between September 2019 and September 2020. The work was carried out under the supervision of Prof. David Manuel Martins de Matos and Prof. Isabel Maria Martins Trancoso.

References

[**1**]: Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M. and Muller, K. [2008], 'Optimizing spatial filters for robust eeg single-trial analysis', *IEEE Signal Processing Magazine* **25**(1), 41–56.

[**2**]: Brigham, K. and Kumar, B. V. K. V. [2010], Imagined speech classification with eeg signals for silent communication: A preliminary investigation into synthetic telepathy, *in* '2010 4th International Conference on Bioinformatics and Biomedical Engineering', pp. 1–4.

[**3**]: Chang, C., Hsu, S., Pion-Tonachini, L. and Jung, T. [2020], 'Evaluation of artifact subspace reconstruction for automatic artifact components removal in multi-channel eeg recordings', *IEEE Transactions on Biomedical Engineering* **67**(4), 1114–1121.

[**4**]: Choi, J., Cha, K., Jung, K. and Kim, K. [2017], 'Gamma-band neural synchrony due to autobiographical fact violation in a self-referential question', *Brain Research* **1662**, 39–45.

[**5**]: Choi, J. W. and Kim, K. H. [2019], 'Covert intention to answer "yes" or "no" can be decoded from single-trial electroencephalograms (eegs)', *Computational Intelligence and Neuroscience* **2019**, 4259369.

[**6**]: DaSalla, C. S., Kambara, H., Sato, M. and Koike, Y. [2009], 'Single-trial classification of vowel speech imagery using common spatial patterns', *Neural Networks* **22**(9), 1334 – 1339. Brain-Machine Interface.

[**7**]: Hall, J. E. [2011], *Guyton and Hall Textbook of Medical Physiology - 12th-Ed*, Saunders - Elsevier.

[**8**]: Holmes, G. L. and Khazipov, R. [2007], *Basic Neurophysiology and the Cortical Basis of EEG*, Humana Press, Totowa, NJ, pp. 19–33.

[**9**]: Lotte, F. [2014], *A Tutorial on EEG Signal-processing Techniques for Mental-state Recognition in Brain–Computer Interfaces*, Springer London, London, pp. 133–161.

[**10**]: Müller-Putz, G., Scherer, R., Brunner, C., Leeb, R. and Pfurtscheller, G. [2008], 'Better than random? a closer look on bci results', *International journal of bioelectromagnetism* **10**(1), 52–55.

[**11**]: Nicolas-Alonso, L. F. and Gomez-Gil, J. [2012], 'Brain computer interfaces, a review', *Sensors* **12**(2), 1211–1279. **URL:** *http://dx.doi.org/10.3390/s120201211*

[**12**]: Rezazadeh Sereshkeh, A., Trott, R., Bricout, A. and Chau, T. [2017], 'Eeg classification of covert speech using regularized neural networks', *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **25**(12), 2292–2300.

[**13**]: Russell, S. and Norvig, P. [2009], *Artificial Intelligence: A Modern Approach*, 3rd edn, Prentice Hall Press, USA.

[**14**]: Suppes, P., Lu, Z.-L. and Han, B. [1997], 'Brain wave recognition of words', *Proceedings of the National Academy of Sciences* **94**(26), 14965–14969. **URL:** *https://www.pnas.org/content/94/26/14965*

[**15**]: Wang, Y., Gao, X. and Gao, X. [2005], 'Common spatial pattern method for channel selelction in motor imagery based brain-computer interface', *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* **5**, 5392–5.

[**16**]: Xygonakis, I., Athanasiou, A., Pandria, N., Kugiumtzis, D. and Bamidis, P. D. [2018], 'Decoding motor imagery through common spatial pattern filters at the eeg source space', *Computational intelligence and neuroscience* .

[**17**]: Yger, F., Lotte, F. and Sugiyama, M. [2015], 'Averaging covariance matrices for eeg signal classification based on the csp: An empirical study', *2015 23rd European Signal Processing Conference (EUSIPCO)* pp. 2721–2725.