UCLouvain

epl

École polytechnique de Louvain

Side-channel attack on the human brain : facial recognition and preferences

Study of different experimental conditions

Author: **Kenzo Fujii**
Supervisors: **François-Xavier Standaert, André Mouraux**
Readers: **Renaud Ronsse, Benjamin Chiêm**
Academic year 2019–2020
Master [120] : ingénieur civil biomédical

Contents

List of abbreviations

ERP	**E**vent **R**elated **P**otential
EEG	**E**lectro **E**ncephalo**G**gram
ICA	**I**ndependent **C**omponent **A**nalysis
N170	**N**egative **170**
N250	**N**egative **250**
P300	**P**ositive **300**
PCA	**P**rincipal **C**omponent **A**nalysis
SNR	**S**ignal to **N**oise **R**atio
SR	**S**uccess **R**ate

Introduction

Nowadays, the development of so-called "brain-computer interfaces" is the focus of much attention. These devices are machines directly "linked" to the brain and able to decrypt our thoughts for different purposes [1]. They can be used to control the movements of an exoskeleton or as a lie detector. In this respect, studies have shown that it was possible to extract information from the human brain by analyzing EEG signals. Tamara et al. depicted methodologies and experimental protocols that are efficient to obtain precise information about the user [2]. In another study, Abhari et al. tried to use the EEG signals as a lie detector [3]. Target stimuli were presented at low frequencies among irrelevant items. We can also cite Berald et al. that observed neural response to the subject's own name [4]. Again, the target stimulus was the subject's first name and the non-target stimuli were other words, unrelated to the user.

The aim of these studies is to classify the EEG segments in different categories through statistical analysis. The term of "side-channel attacks" on the human brain, that inspired the title of this thesis is a good term to categorize these kind of studies. The term "side-channel attack" itself, in the context of computer systems describes the fact that an attacker tries to get information about the functioning of the whole system by analyzing features like timing information, power consumption or electromagnetic leaks. Applied to the human brain, this means that a researcher tries to obtain "hidden" information on a subject by analyzing his EEG signals.

In the task of EEG classification, the event called P300 is often used [5]. This event is a positive peak in the EEG signal that appears on average 300 millisecond after stimulus perception and observed by confronting a subject to low-probability (target) items mixed with high-probability/non-target items. This peak is elicited in the process of decision making.

It is possible to categorize many of these studies thanks to predefined features (concerning the users and the classes to be defined in the experiment). This concern will introduce the taxonomy that we designed and that we present in this thesis. With this taxonomy, we are able to observe that many experiments are designed in convenient conditions that ease the acquisition of significant results. It is interesting to investigate the viability of these experimental setups in comparison to real situations where brain-computer interfaces could be used.

This thesis will thus be based on this taxonomy and we will investigate to what extent we can expect to obtain probative outcomes from an experiment. It is indeed expecable that at one point, it will be very difficult to observe meaningful results by analyzing the EEG signals.

In this respect, we designed two experiments that put the users in different conditions from our taxonomy. In the first one, the task was designed to ease extraction of information. In the second experiment, we put the user in a less favourable context to determine if we could still observe significant results. For these two experiments, the focus is facial recognition: we try to classify EEG segments where the user identifies the target among the ones where he does not. Some studies already tried to conduct similar researches [6][7][8].

These studies put in evidence two more observable events present in EEG segments involved in the task of face identification. One response is known to be involved in the processing of stimuli related to a face while the second is observed when confronted to familiar or newly learned faces [9]. After this analysis, we tried to classify the same segments according to another feature, unrelated to the previous one: "Do the faces displayed inspire the user with a positive or a negative feeling?" This criterion is purely subjective and the users were unaware of this concern during the tests. We believe that these experiments can give us some insights about the question presented here above.

In the first chapter, we will describe the taxonomy mentioned in this introduction and present the protocol and experiments, the hardwares used and the subjects. Then we will explain the preprocessing of our data and how we made them usable for statistical analysis. Further on, we take on the reduction of dimensionality of the signals that allows us to manipulate data reduced to one dimension. Then we present the methodology used to build models, to analyze the data and evaluate their performances. After that we will observe and discuss the results obtained with respect to the taxonomy. Finally, we will conclude this thesis.

Background and important EEG events

In this introduction, we talked about different observable events in EEG signals. The first is the **P300**, already presented. It is a positive peak generally present around 300 milliseconds after the perception of a stimulus. This event is related to a task of classification and usually observed in the context of the "oddball paradigm". A second event is the **N170**, that many studies about face recognition cites [7][6]. This event is a negative peak that appears around 130 to 200 milliseconds after the stimulus presentation. It has been linked with the structural encoding of faces. The last peak is the **N250**, a negative peak that appears around 250 millisecond after the stimulus presentation. This last event has been related to familiar or newly learned faces [9].

These events will take an important roles in our analysis, as we will see in the description of our experiments.

Chapter 1

Taxonomy, hardwares and experiments

1.1 Taxonomy and features

Here is the taxonomy defined during this thesis. This list of features regroup parameters that can describe and categorize most of the studies mentioned in the introduction [10]:

User's related features

1. **Users' condition** : represents the level of concentration of the user
    - *Focused users* are asked to perform a single task
    - *Distracted users* are asked to perform several tasks in parallel
2. **Users' involvement** : define the level of application of the user
    - *Participative users* are asked to perform a task directly related to the classes when they are displayed during the experiment
    - *Passive users* are not asked to perform any task related to the definition of the classes during the experiment
3. **Users' knowledge**: define if the user has a hint about the classes
    - *Informed users* have some hints about the definition of classes to identify during the experiment
    - *Uninformed users* do not have any information about the classes to identify

Classes-related features

1. **Classes definition** determine how and when the classes are defined
    - Classes determined *a priori* are determined before the experiment
        - Classes defined *by the user* are based on questions asked to the subjects
        - Classes defined *independently of the user* are based on generic information
    - Classes determined *a posteriori* are determined after the experiment
        - Classes defined *by the user* are based on questions asked to the subject
        - Classes defined *by clustering* are defined by using unsupervised clustering algorithms
2. **Number of classes**
    - *Discrete classes* correspond to a discrete and specific event
    - *Continuous then discretized classes* correspond to continuous levels that are discretized by the experimenter
3. **Type of classes**
    - *Knownledge classes* is related to the user's memory of some displayed event
    - *Preference classes* correspond to user's taste (or not) for some displayed event
    - *Physiological classes* correspond to the physiological state of the user

Now that the taxonomy is established, let us try to classify some studies involving informations' extraction thanks to these parameters.

1.1.1 Classification of some experiments

Descrpition of the experiments[1]

- Tanakaet al. : Subject had to recognize a subject of reference "Joe" whose face was presented in the same set as that strangers' faces and their own [8]

[1]These are very short summaries of the studies, the references are in the bibliography if the reader wants more details

- Boehm et al. : one face as target that the subject has to identify. He had to press the button target/non-target [11]
- Jemel et al. : 65 faces of unknown and 65 faces of celebrities, the subject had to press "yes" or "no" if the image is the target or not. The images are displayed with an decreasing level of noise [6]
- Saavedra et al. : 40 known and 40 unknown faces, 2 tasks.The first one is judgment of familiarity of the faces (one key for known and one key for not familiar), for the second the task is to tell whether the face presents any emotional expression or not [7]
- Neupane et al. : distinguish phishing attacks and differentiate reals and fake websites. Eye-tracker and EEG to collect data [12]
- Tamara et al. : Treats the future rise of BCIs and the related privacy dangers that will come with them. The article presents the different ways by which a malicious third part could extrcat information about the users [2]
- Abhari et al. : Lie detection based on P300. (Reread for attentional blink, repetition blindness and habituation : sources of errors). Show pictures classed in three different classes : probes, target and irrelevant. The subject is supposed to respond to a specific type of images (targets) [3]
- Inzlicht et al. : study of the correlation between religious belief and reduced reactivity in the ACC (anterior cingulate cortex) which is involved in the mechanism of anxiety and is important for self-control. The test was a classic Stroop test : colored words are displayed with different color fonts, the subject has to press the button with the color matching the one of the font of the word displayed [13]
- Martinovic et al. : feasibility of side-attack channel on the human brain. The extraction of different type of informations was tested : PIN code (memorize a randomly generated number), bank information (try to extract the bank of the user by showing logos of banks => unsuccessful), month of birth, face recognition and geographical location [14]
- Rosenfeld et al. : Feasibility of side-channel attacks on human brain by diverse means [15]
- Berlad et al. : Study the reaction of the patient to the sound of his own name. Three different sets of data were presented. Set 1 : "oddball" paradigm where the subject has to respond to the low probability target by pressing a button, the two presented words were irrelevant to the subject. Set 2: same "oddball"

paradigm with one of the stimuli being the patient's own name. Set 3 : 2 low probability words with same length and physical characteristics (one of which is the subject's name) and a third word (60%) with similar characteristics. [4]

- Farewell et al. : lie detector based on P300. First the subjects are trained by performing a one of two spy scenarios. In this scenario, they encounter six critical items. Afterwards, the test is performed, consisting in listening to phrases of 2 words. There are 3 categories of stimuli : target, irrelevant and probe (which concerns the critical items) [16]

- Kaongoen et al. : Authentication system based on the analysis of P300 response to stimuli. The principle is that only the client would react by visioning the targets during the oddball paradigm [17]

- Frank et al. : investigating the feasibility of a subliminal side-channel attack on the human brain. The concept is to show images to the subject during a duration that would be too short for him to react consciously to it, but long enough to provoke some brain reaction (such as P300) to it [18]

Let us see how we can characterize these studies(Table 1.1.1). We can see that our taxonomy makes it possible to describe the characteristics of this type of studies quite well. We see in the table that the main tendency is to have **focused**, **informed**, **participative** users. The classes are generally defined **a priori** and **independently** from the user. This table does not contain the type of classes. In almost all experiments, classes are related to users' **knowledge**. We will thus try to design experiments that gradually increases the difficulty of retrieving information.

1.2 Experiments

1.2.1 Hardware and subjects

A cap of 64 Ag-AgCl electrode, using the international 10-10 system was used to record the signals that were sampled at a rate of 1000Hz. A total of 6 subject took part in these experiments aged from 19 to 82 years old. The users all signed a disclaimer ensuring their consent. The study was also approved by the ethics committee IPSY[2].

[2]Institut de Recherche en Sciences Psychologiques de l'Université Catholique de Louvain

	F/D	Inf/Uninf	Participative/Passive	Prior/ Post	Users/Independent/Clust
Tanaka	F	Inf	Participative	Prior	Independent
Boehm	F	Inf	Participative	Prior	Independent
Jemel	F	Inf	Participative	Prior	Independent
Saavedra	F	Inf	Participative	Prior	Independent
Neupane	F	Inf	Participative	Prior	Independent
Tamara	?	?	?	?	?
Abari	F	Uninf	Participative	Prior	Independent
Inzlicht	F	?	Participative	?	?
Martinovic1	F	Inf	Passive	Prior	User
Martinovic2	F	Inf	Passive	Post	User
Martinovic3	F	Inf	Passive	Post	User
Martinovic4	F	Inf	Passive	Post	User
Martinovic5	F	Inf	Passive	Post	User
Berlad1	F	Inf	Participative	Prior	Independent
Berlad2	F	Inf	Passive	Prior	User
Berlad3	F	Inf	Passive	Prior	User
Farewell	F	Inf	Participative	Prior	User
Kaongoen	F	Inf	?	Prior	User
Frank	F	Uninf	Passive	?	Independent
Rosenfeld	F	?	Active	Prior	Independent

Table 1.1: Classification of the experiments presented in the articles by features. F/D = subject focused / distracted, Inf/Uninf = subject informed/ uninformed, Prior/Post = classes defined a priori/ a posteriori

The stimuli are showed on a screen **Display++**[3]. The dimensions of the images displayed are 770x700 pixels. The remaining experimental conditions are presented in the experiments' descriptions.

1.2.2 Description of the experiments and protocol : face recognition

To build our models, we need to record EEG signals. To obtain them, we designed 2 experiments with different conditions. The general idea is to show to a user faces of people whom he identifies [4] and other whom he does not recognize. After the experiment, the subject is asked to complete a document where he clearly states if

[3]Specifications available at https://www.crsltd.com/tools-for-vision-science/calibrated-displays/displaypp-lcd-monitor/

[4]Celebrities like leading political figures, top sportsmen, actors,...

he does or does not recognize the faces that were presented during the experiment. We will later use this form to get the classes of each observation of our EEG signals.

It is to be noted that the concept of **recognizing** a face was stated as : "I know precisely that I recognize this person and I remember in which context." The subject does not have to remember the name of the people he sees on the screen.

However, it is difficult to state if the identification of a face is a binary random variable or a continuous one. How can we classify a "déjà-vu", or someone that the user used to know but forgot after a certain time? In this study, for simplicity we will consider that the classes are discrete, but the question remains.

In each experiment, we show 600 faces to the subject in 8 sets of images containing 75 faces each. In each set, the subject could recognize a maximum of 50 faces (celebrities), the 25 others were random people that should be unknown to everyone. The images were shown during 1.5 second and there was a gap of 1 second between each image. There was a blue fixation cross at the center of the screen (blue and sometimes red for the third experiment) and placed at eye level of the faces shown[5]. There was a pause of 2 to 5 minutes between each set of images in order to avoid potential fatigue effects.

First experiment

During this experiment, faces of famous and anonymous persons were shown to the user following the protocol presented before. The user was then asked to count each time a face that he recognizes is displayed. After the experiment, the user is asked to fill in the document that will be used to determine the classes. For the users' related features, it is clear that the user is **focused**, **participative** and **informed** since there is only one task, related to the classes to perform and the user clearly has some hints about them. For the classes related features, classes are defined **a posteriori by the user**, the classes are **knowledge related classes** and are **discrete** since there are only 2. 3 subjects took part in this experiment.

Second experiment

During this experiment, the faces are displayed following the same protocol as before. This time the user is asked to count the number of time the fixation cross will change from blue to red. This time, the user is **focused** since there is only one task, but this time he is **passive** with respect to the classes because the task

[5]The gaze must be at this level for the task of identifying faces

has nothing to do with it. Finally, the classes are defined again **a posteriori by the user**. For the classes-related features, they are the same than in the first experiment. 3 persons also took part in this experiment. Since the user is passive, we expect the classes to be more difficult to distinguish statistically.

Since the two experiments involve face recognition and a classification of the stimuli, we expect to observe the events mentioned in the introduction. P300, related to decision making, **N170** related to the encoding of faces and **N250** related to the response of newly learned or familiar faces.

1.2.3 Subsequent study: preferences

This study was performed only using the data collected during the two experiments described here above. The document that the users had to fill in actually contained another question that they had to answer for each face. The question was about the feeling that the face inspired to the user. This time, there was 3 possible choices:

- I have a positive feeling about this person
- I have a negative feeling about this person
- I feel neutral towards this person

Here, the users are clearly **passive** and **uninformed** since they have no clue about the classes during the experiment, and the task has nothing to do with it. The type of classes becomes also **preferential**. The rest of the taxonomy remains the same as for the previous study. For this analysis, since the user is even less aware of the classes, we expect the distinction of the two classes to be even less observable. It is also to be noted that since there were 3 options, the whole 600 observations could not be used as data set. We are restricted to the data sample that the users chose as inspiring a **positive** and **negative** feeling only.

1.2.4 Important note on the P300 and the "oddball paradigm"

As explained in the introduction, the P300 is usually observed in classification tasks that involves low-probability target items, mixed to high-probabilities non-target items. This precise type of experiments is called the "oddball paradigm". In our case, since there are 50 celebrities for each set of data and since we do not known a priori which face the user will be able to recognize, we cannot ensure that the target (here the *known* faces) are displayed with low probabilities.

In many studies (see Table 1.1.1), the experiments are designed to define the

classes **a priori**, which allowed to ensure that the target item was displayed with low-probability. The fact that the target was presented in an isolated manner in regard of the other standard items possibly played an important role in the experimental design of these studies.

In the experiments presented in this thesis, since the classes are defined **a posteriori by the user**, it is not possible to ensure that the target items are isolated, even if we can verify it in our analysis. These conditions therefore seem far from ideal to collect informations. However, we can imagine that in the context of use of a brain-computer interface, the conditions would not be always ideal either.

Chapter 2

Preprocessing

2.1 Filtering signals

It is common to preprocess the EEG signals that we will work with, since they are filled with noise. Indeed, the electric activity of the brain is very small in amplitude and the signal is corrupted by many sources of noise, as other electronic devices or even muscular activity of the subject. In this section, the different processes used to filter the EEG signals are detailed. The software used is MATLAB with the extension Letswave 6 [19], in which many filters and methods are preimplemented making the treatment of data easy and intuitive.

Band-pass filter

In order to keep only the frequencies that interest us, we first apply a band-pass filter with a low-frequency cut-off of 0.5Hz and a high-frequency cut-off of 30Hz.

This allows us to neglect the most of noise sources that can corrupt our signal, as the frequencies above 30Hz often corresponds to electronic devices and the sources that emit under 0.5Hz are not likely to participate much at the response elicited by a visual stimulus.

Epoch segmentation

As the signals that we record for each set are composed of the segments corresponding to the neural response for each 75 images placed end-to-end, we want to decompose them into 75 different segments called "epochs". We will thus take for each image the window of [-0.5 s, 1s] where 0 is the time at which the image is displayed on the screen.

Since the events that we want to observe are known to occur on average between 0 and 500 milliseconds after the stimulus perception[1], this time window seems reasonable. That gives us 1500 points for each epochs.

In this thesis, each epoch, also referred to as "observation" will be written as $e_i^n(t)$, where i is the name, or number of the electrode selected ($i \in [1; 64]$), $\mathbf{n}$ is the number of the observation ($n \in [1; 600]$) and t is the time of observation in millisecond ($t \in [-500; 1000]$).

Baseline Correction

For each epoch, we substract the average of the values contained in the interval of 0.5 seconds before the stimulus. This is computed as:

$$e'_i = e_i - \text{MEAN}(e_i|_{[-500..0]}) \tag{2.1}$$

Independent Component Analysis

We will then apply an ICA to our data, in order to reject some artefact that can be caused by the movement of the patient, particularly ocular artifacts. The main task for ICA for a vector is to find a linear transformation that minimizes the statistical dependence between the components of the signal [20]. It is a common way to try isolating the signal that interests us from other sources of noise that can be added to it.

Mathematically, that means we want to compute the so-called "unmixing matrix" W defined as:

$$s = Wx \tag{2.2}$$

Where x is the signal observed (our data collected), and s represents the different independent sources. In the specific case of neuroscience experiments, we can say that an epoch e_i can be corrupted by an artefact of intensity I that correspond to a perturbation. We aim thus to compute the unmixing matrix W that will remove these artefacts by multiplying it by the observed epoch e_i.

Of course, the ICA matrix must be computed separately for each user. When it is computed, it is easy to retrieve the ocular artefacts in our signals. We can retrieve them by analyzing the waveform morphology and topological distribution of it that dissociates from the rest of the signal. We finally apply the matrix to these epochs to remove the artifacts.

[1] Recall that we are likely to observe N170, N250 and P300

2.2 Separating segments according to class

After this preprocessing, we must separate our data according to the classes following our experimental protocol. We have at our disposal 8 sets of 75 observations per user, for which we have different proportions of the two classes of interest. We thus have 600 observations per users.

For each user, we separate the 600 observations according to their class. We know the class of each observation by reading the form that the users completed. We have for a user a matrix of dimension $600 * 64 * 1500$.

For the preference (*like* vs *dislike*) analysis, all the observations are not used since there was a "neutral" option in the document. We dispose of less data for this study.

2.3 Some notations

Throughout this thesis, we will ofter use the concept of **ERP**. This acronym stands for Event Related Potentials. It is the measured brain response to a sensory, cognitive or motor stimulus. In this section and the ones following we will denote the time t and corresponding to the electrode named i for an observation n as $e_n^i(t)$.For the face recognition analysis, there is 8 sets of 75 observations, we dispose of 600 epochs per user.

Considering the context, each observation belongs to one of the two classes *known* or *unknown* (and *like* or *dislike* for the second analysis) the class of one observation will be depicted by the variable p. In the following chapters, $n \in [1; N]$ where N is the total number of observation for an user (600). Also, $t \in [1; T]$ with T the total number of time steps.

Chapter 3

First approach and Dimensionality reduction

3.1 Averaged ERP vizualisation

To have a general idea about our data, it is common to plot the averaged ERP of the different classes and observe if we can see any differences a priori. We will also join the variance to this first glance. The averaged ERP is simply computed by averaging the signal for each time t, among all observations relative to one class. We obtain an averaged ERP over time. This is mathematically equivalent to:

$$\bar{e}_p^i(t) = \frac{1}{n_p} \sum_{n=1}^{n_p} e_n^i(t) \tag{3.1}$$

Where n_p is the total number of observation for this class, i is the electrode selected and $\bar{e}^i(t)$ is the averaged value of the ERP at time t for class p.

On Fig. 3.1 and 3.2 are represented the averaged ERP and standard deviations of the ERP of 2 users that participated the first experiment. These subjects were focused on counting each time they saw a face that they did recognize. We observe that the pattern of the two classes is very similar and that it is difficult to notice clear differences between the means. When we take the variance of the data into account, it seems even harder to tell anything.

The Fig. 3.3 represents the averaged and the standard deviation of the ERP for a user who participated in the second experiment. This time, the subject was only asked to count the number of times the fixation cross was changing color (**passive user**). We can also observe very little differences between the classes. It seems however that the patterns are quite similar, with a positive and negative peak around $200 - 300$ milliseconds after the stimulus. However, we will need to perform deeper analysis to deduce something. As expected, the negative peak

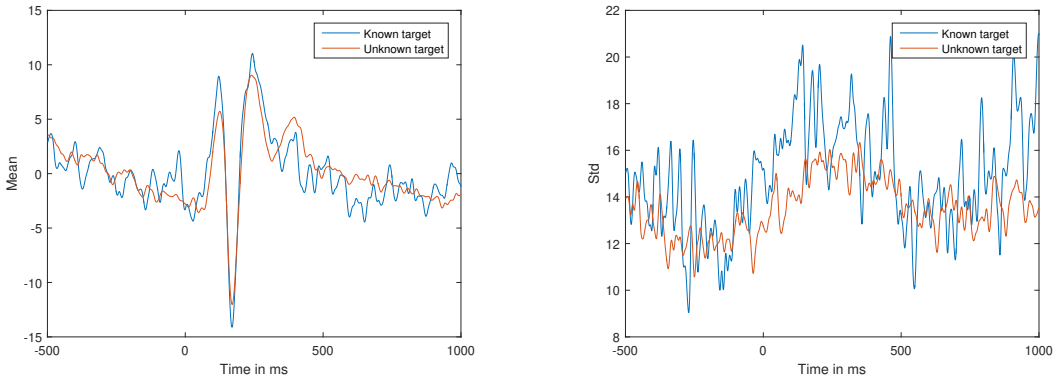

Figure 3.1: Average (left) and standard deviation (right) of ERP for User 1 electrode P8

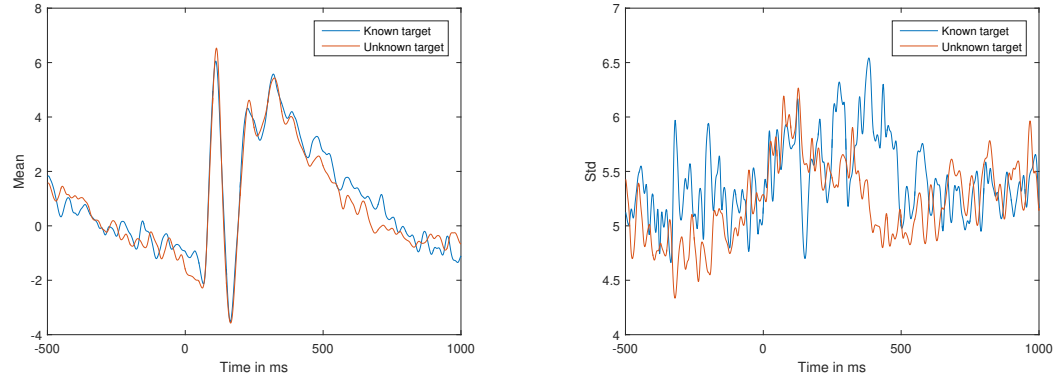

Figure 3.2: Average (left) and standard deviation (right) of ERP for User 2 electrode P8

should correspond to the **N170**[1] mentioned in the introduction and the positive to the **P300**.
The first step of the data analysis will be to **center** the data. We will subtract the mean value of the ERP over all observations at time t to the value of each observation at time t. Let us compute the averaged ERP again and look if we see any differences. Mathematically this is computed by:

$$\bar{e'}_p^i = \frac{1}{n_p} \sum_{n=1}^{n_p} e_n^i(t) - \bar{e}^i(t) \tag{3.2}$$

Where e'^i_p is the centered averaged ERP for class p and $\bar{e}^i(t)$ is the averaged ERP at time t for the 600 observations.

[1]Related to face processing

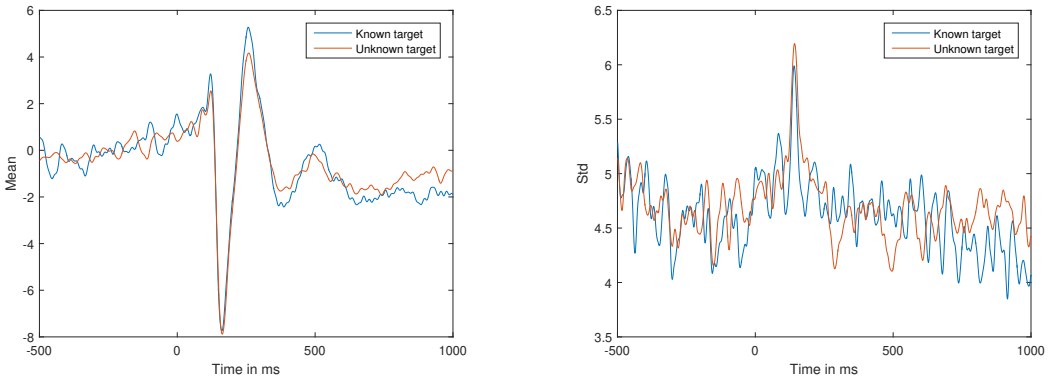

Figure 3.3: Average (left) and standard deviation (right) of ERP for User 4 electrode P8

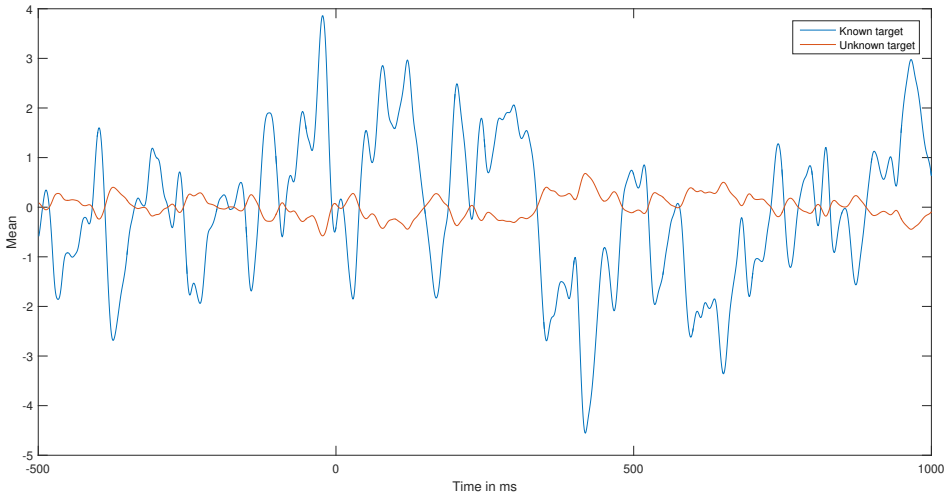

Figure 3.4: Average ERP after centering for User 1, electrode P8

On Fig. 3.5 and 3.4, after centering, we can clearly observe many more differences between the classes. Note that for user 1, there were only **78** items defined as *known* vs **522** *unknown*. This implies that the average of all observation that we subtracted was mainly defined by the *unknown* class. This explains why the mean corresponding to the unknown class is closer to zero. On Fig. 3.5, that corresponds to user 2, the classes were much more balanced (261 observations *known* vs 339 *unknown*). And we see that the differences between the two averaged classes are way less obvious. This will play a major role in further analysis but it will be discussed in the next sections.

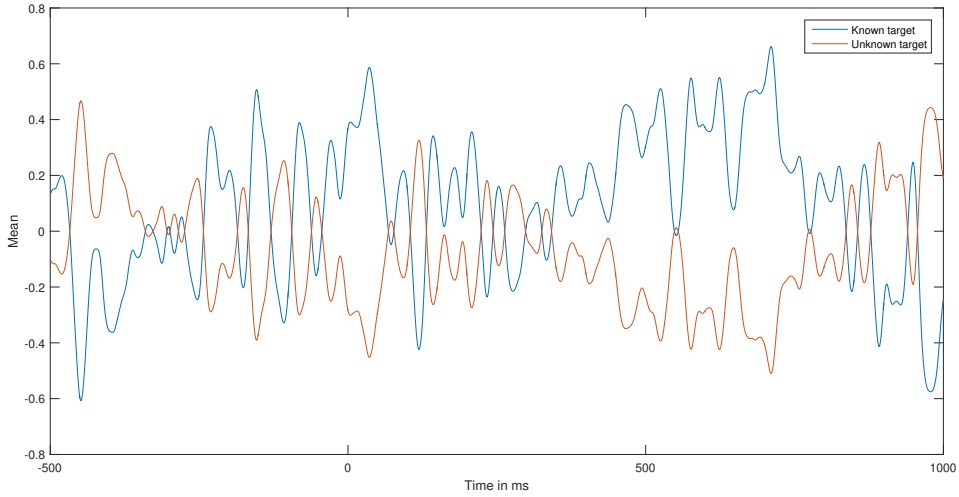

Figure 3.5: Average ERP after centering for User 2, electrode P8

The first challenge to take on is related to the size of the data. Indeed a matrix of size $(600 * 64 * 1500)$ is hard to manipulate. There are two different feature selections that we can use to reduce drastically the dimensions of our data.

3.2 Signal to noise Ratio

In the context of neuroscience, signal to noise ratio (SNR) represents a dimensionless measure of the signal that we want to quantify, relative to the fluctuations that are outside experimental control [21]. A neuroscience experiment usually involves different responses to a collection of stimuli presented to a subject.

During the experiment, different stimuli are presented to the user, each of them inducing a response that will be recorded. In this way, we have a distribution of the responses over the different observations. The noise can be represented as the variance of the signal at a time t throughout each observation. In our particular case, we want to use the SNR in order to determine in which temporal window, and on which electrode we can observe a significant response to our stimulus, that will differentiate itself from the noise.

For N stimuli presented (600 in our case), and for a value $e_n^i(t)$ of the ERP

at time t for electrode i. The SNR at time t is given by [21]:

$$SNR^i(t) = \frac{\frac{1}{N}\sum_N (e_n^i(t))^2}{\frac{1}{N}{\sigma_N^i}^2(t)} \tag{3.3}$$

Where ${\sigma_N^i}^2(t)$ is the variance of the signal throughout all observations at time t and electrode i.

From the literature, we know that the P300 reaction that we expect is mostly measured strongly by the electrodes covering the parietal lobe [5][3]. We also know that the strong reaction to a visual stimulus is recorded by the electrodes located in the occipital lobe [22]. We thus have heavy assumptions in the electrodes covering the parietal and occipital zone.

3.3 Electrode selection

As mentioned in the previous section, we have the information given by 64 electrodes at our disposal. It is more than likely that some are more interesting than others to observe, and in any case, 64 additional dimensions is too much to perform an efficient analysis with the tools we are going to use. We have also strong assumptions regarding the electrodes covering the parietal and occipital lobe, since they are recording the best the activity related to P300 and N250 events and responding to visual and face processing related stimuli.

In section 3.2, we gave a definition for the SNR defined for neuroscience experiments. Let us compute the SNR for our subjects in function of time and for each electrode, this way we should be able to determine in which time windows and which electrode are the most informative.

On the Fig. 3.6 and 3.7, we can clearly see a concentration of peaks in a window between 0 and 500 milliseconds. This confirms us the intuition we had about the P300, N170 and N250, but is also consistent with the expectation of observing a reaction to the visual stimulus. However, there are many peaks for each electrode over different periods of time.

We would like to keep only one statistic per electrode to determine which are the most informative. Two options were considered : the maximum SNR over time or the averaged SNR over time. The latter option was excluded, since the differences between the averaged SNR throughout the electrodes were very little. For each user, we selected the 3 electrodes that presented the higher SNR over time. Since we have two different experiments, it would seem logical to separate

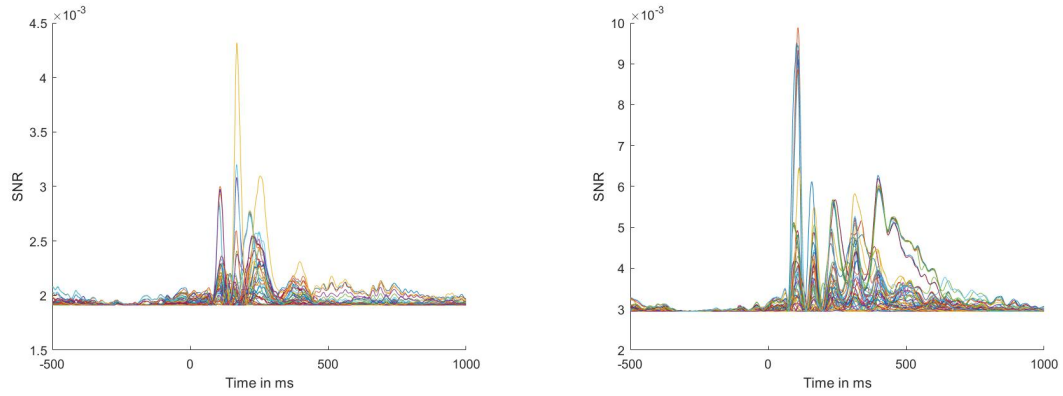

Figure 3.6: SNR over time for Users 1 (left) and 2 (right)

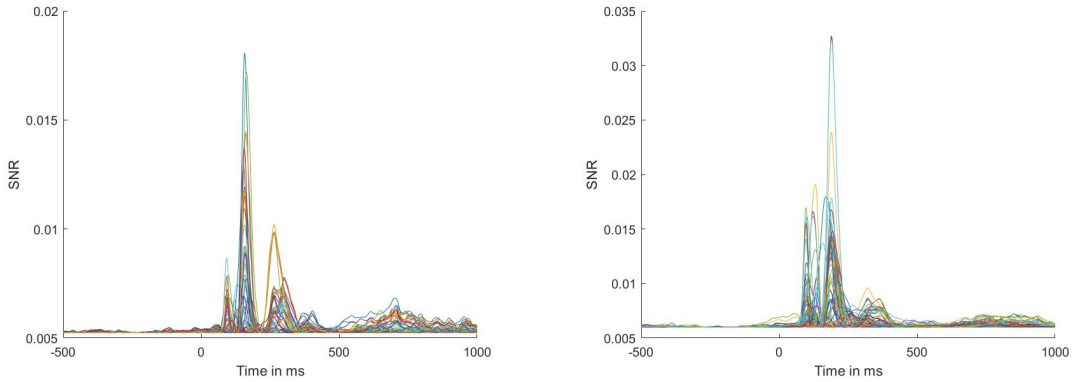

Figure 3.7: SNR over time for Users 4 (left) and 5 (right)

the selection of electrodes independently for the two conditions.

Recall that the users 1, 2 and 3 took part in the first experiment, and the users 4, 5 and 6 in the second.

On Fig. 3.8 the statistic that we used to choose our electrodes is represented. Table 3.1 contains a summary of the electrodes with the highest values of SNR per user.

As expected, the electrodes covering the parietal lobe (prefix P) and the occipital lobe (prefix O) are very well represented. We see for example that the electrode P8 is selected for 3 users. For the rest of the analysis, and since the overall section of electrodes is similar throughout the users and consistent with literature, we will study these electrodes for all users.

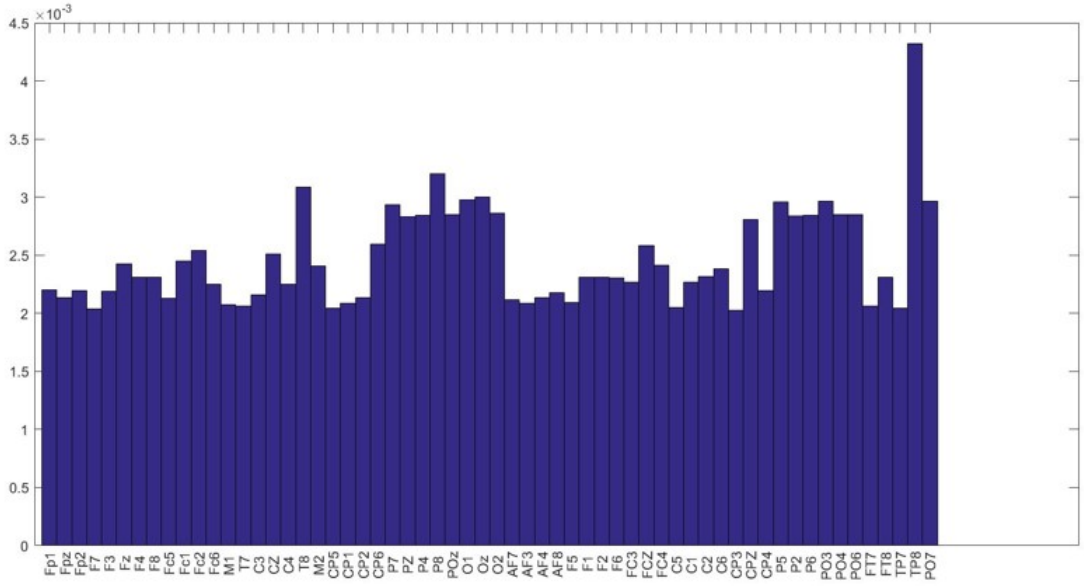

Figure 3.8: Maximum SRN per electrode for User 1

	1	2	3
User 1	TP8	P8	T8
User 2	O2	Oz	PO8
User 3	P4	P8	POz
User 4	P7	P8	PO4
User 5	O2	PO6	PO8
User 6	O2	PO6	PO8

Table 3.1: Electrodes with maximum SNR per user

Our selection is thus : TP8,P8,T8,O2,Oz,PO8,P4,P8,POz,P7,PO4,PO6. For the rest of the analysis, we will use a technique that consists of taking the signals of electrodes of the selection placed end-to-end. This way we are able to analyze all the electrodes at the same time, using the information contained in all of them at once.

After this operation, since the signal of each electrode is of size 1500, we are left with a signal of size $11 * 1500 = 16500$. Note that since the EEG caps used to perform the experiments were sometimes defective, some electrodes did not record any signals during the experiment. In consequence of that, for the user 1 we do

not have a signal for electrode PO8, which left us with a segment of 15000 points per observation.

Now that the electrodes are selected, the value of the ERP at time t will be refered as simply $e_n(t)$, since the electrodes are selected at once. Note that now $t \in [1; 16500]$.

3.4 Principal Component Analysis

Now that the signals are preprocessed, we want to reduce the dimension of our data. A common way to do this in EEG signal analysis is the **Pincipal Component Analysis** [23]. This analysis allows redefining our signals by projecting them onto new components that maximizes the variance. Computing the PCA matrix gives us the eigen vectors of the variance-covariance matrix of our initial data matrix. There are different steps necessary to obtain these eigen vectors.

The first step is to standardize the data. Our aim is to obtain data with a mean of 0 and a standard deviation of 1. Given $\mathbf{X}$, the initial data matrix (of dimensions $N * T$[2]), we compute $\mathbf{X}_{cen}$, the centered and standardized data matrix as:

$$X_{cen} = \frac{(e_n(t) - \bar{e}(t))}{s(t)} \tag{3.4}$$

Where $e_n(t)$ described in the next section is the value of ERP for observation n at time t, for our selection of electrode and $\bar{e}(t)$ and $s(t)$ are respectively the mean and the standard deviation for $N = 600$ observations at time t. With $i \in [1, N]$ and $j \in [1, T]$. The next step is to compute the variance-covariance matrix as:

$$C = \frac{1}{N} X'_{cen} X_{cen} \tag{3.5}$$

Finally, the matrix containing the eigen vectors of C, $(u_1 ... u_t)$ is such that:

$$C u_\alpha = \lambda_\alpha u_\alpha \tag{3.6}$$

With $\lambda_1, ..., \lambda_t$, the eigen values of C.

In the context of our EEG signal analysis we have two different choices. Either we choose to apply the PCA considering all the observations at once, either we apply the PCA on the averaged signal of each class, considering that it contains

[2]Where N is the number of observations and T the times steps

most of the information needed. In this thesis, we will consider the latter option. This implies that matrix C contains only two observations, the means of each class:

$$C = \begin{pmatrix} \bar{e}_{p_0} \\ \bar{e}_{p_1} \end{pmatrix}$$

(3.7)Where $\bar{e}_{p_0}$ and $\bar{e}_{p_1}$ are the mean of the ERP for each class. At the end, we obtain one eigen vector u_1 on which we can project our data to reduce the dimensions.

3.5 Univariate reduction with PCA

Performing the PCA on the averaged signals for both classes leaves us with one coefficient vector of size 16500 (or 15000 for user 4). Then we project both the training and test set on this eigen vector $((600 * 16500) * (16500 * 1))$. In the end, that leaves us with only one dimension per observation.

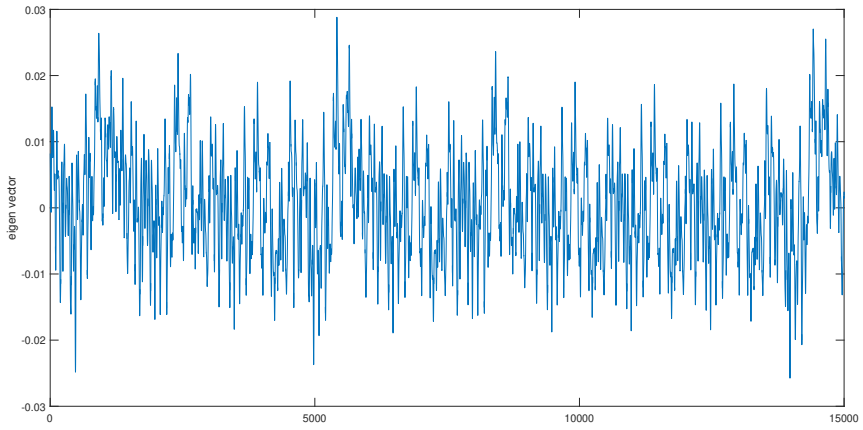

Figure 3.9: Eigen vector obtained from the means of 540 obs., User 1

On Fig. 3.9 and 3.10, we can see the eigen vectors used for the projection of the training and test set for 540 observations within the training set for users 1 (1st experiment) and 5 (2nd experiment). It is thus the PCA computed using the maximum of the data to compute the average of the two classes. By looking at the eigen vectors, we can see that it is very noisy, making the interpretation of the principal components quite hard. The vector of user 5 seems to be

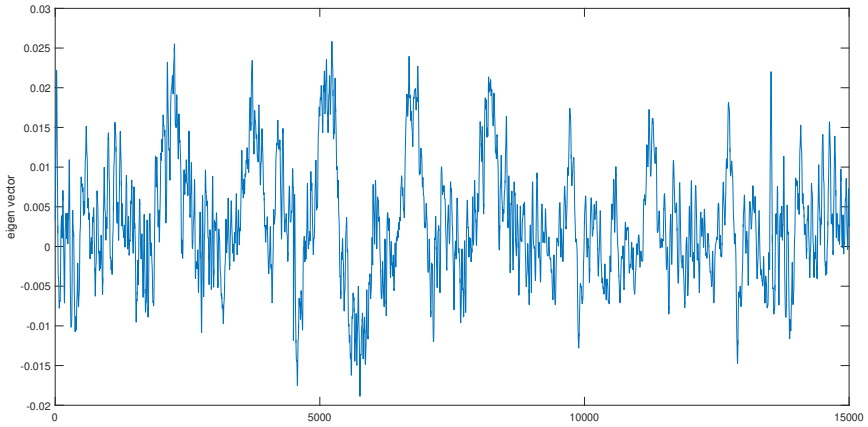

Figure 3.10: Eigen vector obtained from the means of 540 obs., User 5

composed of some peaks but the intervals between them are not periodic. Note that this vector is the combination of the vector of each electrode placed end-to-end.

Let us try to superpose the parts of the eigen vector corresponding to each electrode to look if we do not see periodic behaviors more distinctly (Fig. 3.11).

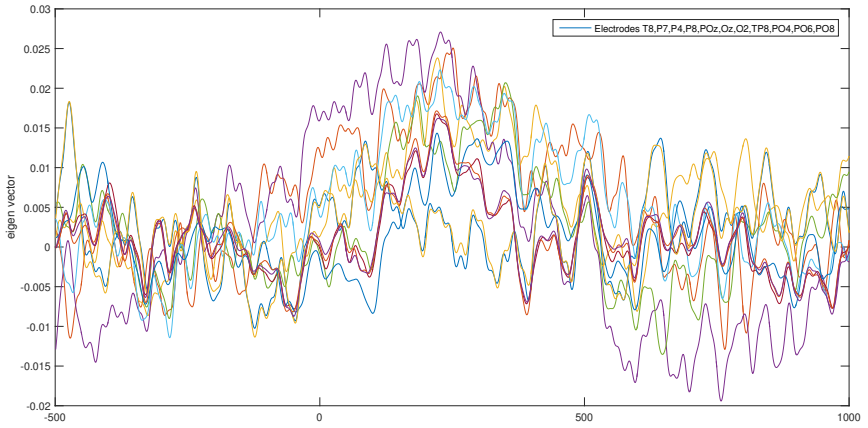

Figure 3.11: Segments of the eigen vector for each electrode (*face recognition* analysis) , User 5

Here we can clearly observe that the highest values present in the vector for each electrode are in a window $t \in [0; 500]$ which again confirms our expectation about N170 and P300 events.

Now we observe the eigen vectors computed for the *preference* analysis. On Fig. 3.12, we can see that the vectors do not present the same peaks than in the previous study. The values between 0 and 500 milliseconds seems however a bit higher than the rest but it is less visible than on the previous figure. This is coherent with the fact that here, the face recognition process and the knowledge of the user are not involved. Indeed, the classes are this time related to the *preference* of the users.

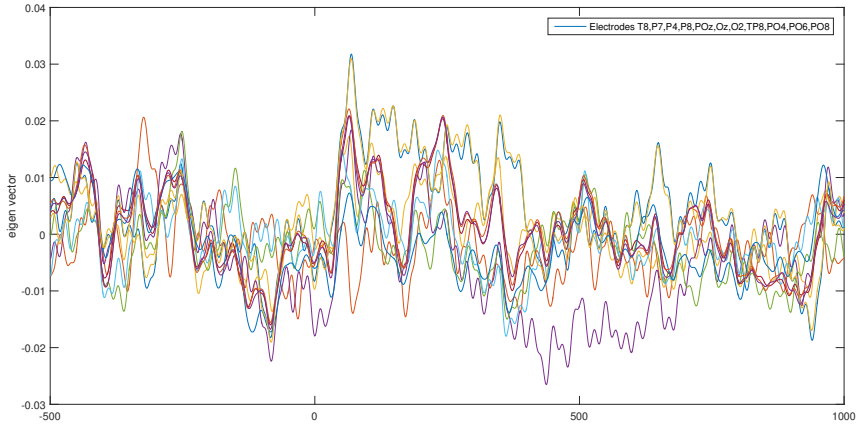

Figure 3.12: Segments of the eigen vector for each electrode (*preference* analysis) , User 5

One question remains: when do we apply the PCA during our methodology ? This will be explained in chapter 4.

For now on, we will refer as each ERP value for observation n as e_n since the time dimensionality disappeared.

Chapter 4

Evaluation Metric and Methodology

In this chapter we detail the methodology used to analyze our data. We first present the mathematical concepts used (some were introduced in previous section, as PCA analysis). Since we want to evaluate the performances of our model, we build our model over a training set and then evaluate it over a test set.

4.1 Cross-validation

Let us explain how we can assess the performances of the model thanks to a training and a test set. The problem is that if we use a certain portion of the data set to train our model and then test it on the part that was left, we lose some information building our model. In addition, we test this model only over a restricted portion of the data. This can lead to severe overfitting problems. A very common way to deal with this issue in data process is the cross-validation.

The idea to divide our data set in k different folds (here we chose $k = 10$) of equal size, use one of this fold as test set and the rest as training set and finally iterate m times. This way, we use m (we chose $m = 50$) different partitions of the data as training and test set, reducing the overfitting. This process differs a bit from the famous K-fold cross-validation(Fig 4.1), often used in data analysis, since the training set is taken 50 times at random. As explained the test set is built with what is left of the data. We chose 50 iterations in order to smooth the curves of results, since with only 10 iterations, the variance was too high to get a representative idea of the results.

When the training set is computed, we obtain some observation of both classes that

compose it. We apply the PCA on these means and we project both the training and test set on the eigenvector obtained to get unidimensional data.

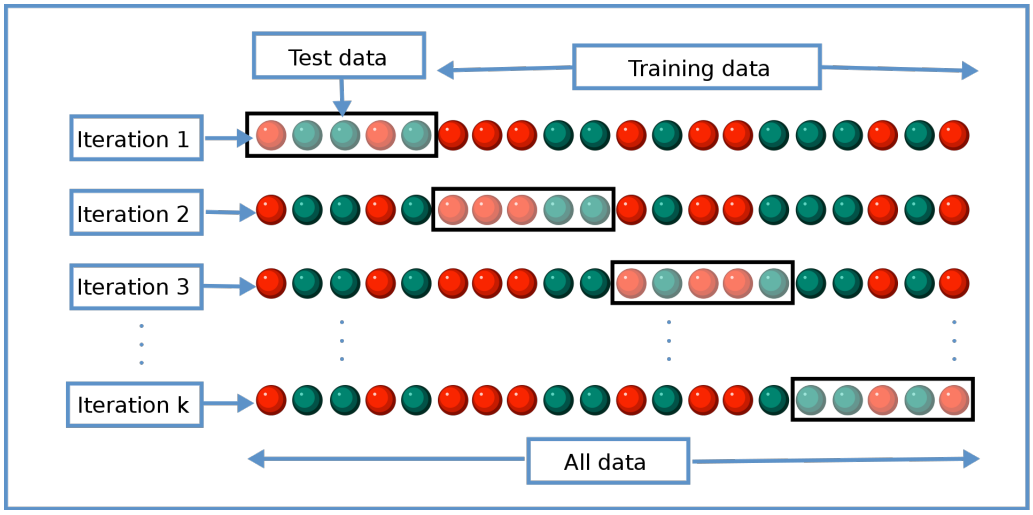

Figure 4.1: K-fold cross validation (source Wikipédia.org)

Since for each user, there is a different proportion of sample *known* and *unknown*, this has an impact on the model built for each class. We thus ensure to respect this proportion when we define our training and our test set for each user. For each proportion, we compose our test set by taking 1/10 of the total proportion of observation taken, and 9/10 as training set.

Success rate[1] (defined in the next section) is finally computed taking the average of the SR calculated for the 50 folds.

An additional angle that we will investigate is **the number of observations** used to build the model. We investigate if the SR changes and converges as the number of data grows. For this purpose, we use at first 60 of the 600 data, then we increase the size of the dataset of 60 at each step, until reaching the full capacity of the data. So we have 10 iterations of the whole process. For the **preferences** study, the process is the same. We iterate over 10 different proportions of the data, but adapted to the size of the whole data set, of course. Recall that we do not dispose of 600 observations per user in this case.

[1]The value that interests us

4.2 Success Rate

The evaluation metric that will interest us in our analysis is the **Success Rate** (SR). The SR is easy to explain: it is an estimation of the rate at which our model will correctly classify any new observation into one of the two classes. Let us now detail how we compute this criterion, that will only be computed on the test set.

The Success Rate is simply defined as the percentage of success of our model. For each observation e_n of our test set, we take the probability that it is assigned to its true class (i.e if p=1 we take $Pr[P = 1|E = e_n]$) which is obtained from the model computed by the training set (see section 4.3). We add 1 to the success rate each time this probability is above 0.5 (that means that this observation is well classified), then we divide it by the size of the test set. The SR is finally computed by:

$$SR = \frac{\sum_{e_n \in test} 1_{|Pr[P=1|E=e_n]>0.5}}{S_{test}} \tag{4.1}$$

Where the numerator is the number of correctly classified observations from the test set and S_{test} is the number of observation of the test set.

Since the SR will vary in function of the training and test set, we have to compute a confidence interval to estimate precisely the range of the SR for each user. This will be explained in section 4.4.

4.3 Gaussian densities approximation

Let us now explain how we will build our model thanks to the test set and how we compute $Pr[P = 1|E = e_n]$, the probability that interests us. Assuming that each observation belongs to class $P = 1$ if the user knew the face and to class $P = 0$ if he does not know it. Since we want to observe the success rate of the model, we want for each of our observation to obtain $\hat{Pr}_{model}[E = e_n|P = p]$, which correspond to the probability for a certain observation E to take the value e_n, knowing that its true class is p.

To estimate this probability, we can approximate the true distribution of our data by an existing one that will fit it decently. A common way to perform this in EEG classification is approximating the distribution of our observations by the normal law [24].

In most cases, the distribution of the data for each electrode does not perfectly fit a normal distribution[2]. However, the normality will be assumed for its conve-

[2]We can verify this asset by performing an Anderson Darling test over the dataset

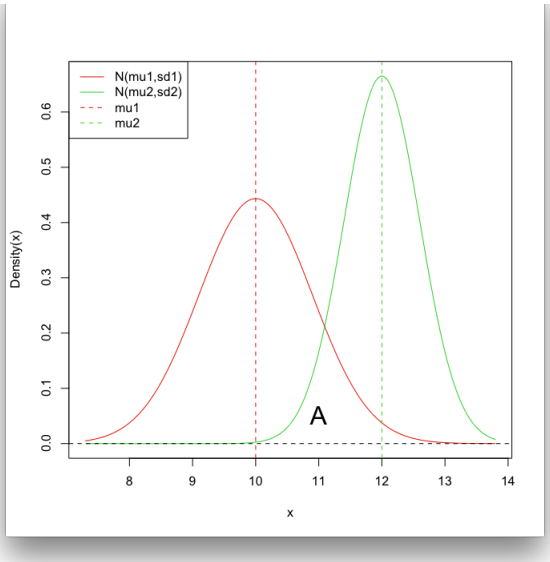

Figure 4.2: Two Gaussian densities

nience in this application.

The idea is thus to compute the mean and the standard deviation of both classes (known and unknown) by using all the observations contained in the training set to obtain the two corresponding probability density functions. These functions will take the role of the *model* mentioned above. The normal probability density function, expressed in terms of the mean μ and the standard deviation σ is :

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu)^2}{2\sigma^2} \tag{4.2}$$

Once we have our two density functions (one for each class) we can obtain the probability for each of our observations[3] to take their actual value e_n, knowing their belonging to the *known* or *unknown* class.
Observing the Fig. 4.2, if each Gaussian curve represents the distribution of one of our classes, then the probability of an observation to take a value e_n is the height of the corresponding curve for this value on the x axis. Note that if e_n is very distant from the mean of the two normal curves, the associated probability will be very small for both classes. This effect should be taken into consideration during our analysis.

[3]Remember that after the PCA projection, we are left with only one value per epoch

Now that we got $\hat{Pr}_{model}[E = e_n|P = p]$, we can easily compute the probability that interests us the most : $\hat{Pr}_{model}[P = p|E = e_n]$. This is the probability for an observation e_n to belong to the class p. This quantity can be calculated by using Bayes' rule:

$$\hat{Pr}_{model}[P = p|E = e_n] = \frac{\hat{Pr}_{model}[E = e_n|P = p] * Pr[P = p]}{\sum_{p'} \hat{Pr}_{model}[E = e_n|P = p'] * Pr[P = p']} \tag{4.3}$$

Where $Pr[P = p]$ are the a priori probabilities of each class. Since this information is not available in real situation, we set it to 0,5.

4.4 Confidence interval

To better assess the results given by the computation of the Success Rate, we will calculate confidence intervals for each of them. During data analysis, as mentioned we estimate the SR for 50 different folds for a certain proportion of our data. We will use the theory of bootstrapping to compute our intervals [25]. Recall that in the cross validation, we computed the probabilities that are used to compute the SR 50 times for a given proportion. The proportion that interests us here is the total number of observations (600 for the face recognition analysis). In this case the size of the test set is 60 observations. We obtain thus a total of 60*50 probabilities. We perform 100 resampling of size 60 on these 3000 probabilities and we compute the SR on these samples. We obtain 100 different success rates. Now we just take the 2.5 and 97.5 percentiles and we obtain 95% confidence interval. Note that this process is quite different that the one used to compute the average SR over the 50 folds. In some rare cases, the average SR computed can thus be outside the bounds (but still very close).

Summary of the methodology

Fig. 4.3 represents the methodology step by step. First we pick a proportion of the data set, then we separate it into a test (1/10) and a training set (9/10). We compute the PCA based on the training set and we project both training and test set on the eigenvector to get our unidimensional data. The projected training set is used to compute our two gaussian densities, and finally the test set is evaluated over it to obtain our success rate. This process is repeated 50 times with the same proportion of the whole data. The success rate for each proportion is taken as the average SR over the 50 iterations.

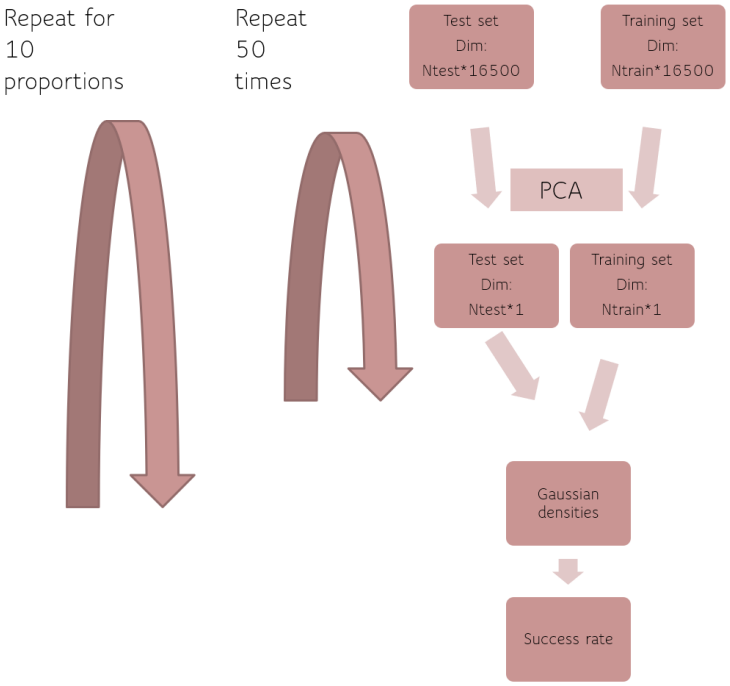

Figure 4.3: Methodology used to evaluate the SR

Chapter 5

Data analysis

Let us analyze the data obtained by our model.

5.1 Face recognition

Model computation

First, we will analyze the gaussian densities that serve as model in our study. We will thus observe the densities output for different proportions of the whole set of observations. For a given proportion *prop*, recall that the training set is composed of $9/10 * prop$ observations. This training set is used to compute the PCA, then both sets are projected on the eigenvector obtained. Now that we have unidimensional data, we train our model by computing the mean and standard deviation and compute the gaussian densities associated to them. Let us take a look at the densities obtained. We can compute the first probability that interests us : $\hat{Pr}_{model}[E = e_n|P = p]$. You may recall that this value is the probability that the ERP (now in one dimension) takes the value e_n, knowing its belonging to class p.

On Fig. 6.1 and Fig. 5.2 we can see the distributions of our classes, fitted to normal distribution for respectively 60 and 600 observations. The density corresponding to the *unknown* class is wider in both case because there are more observations belonging to this class : 78 *known* vs 522 *unknown* for User 1.

We known that the closer the densities are, the more difficult it will be for us to guess the class of one observation of the test set. On Fig.6.1 we see that the densities that are somewhat distinct for 60 observations tends to merge when reaching the 600 observations. This means that the observations of the test set will be even more difficult to distinguish.

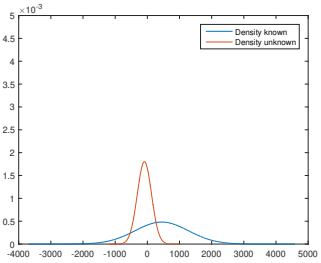

Figure 5.1: Densities for 60 observations, User 1

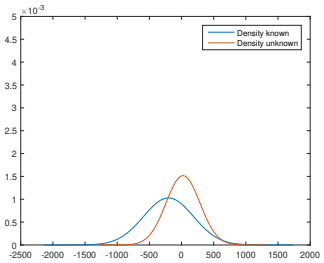

Figure 5.2: Densities for 600 observations, User 1

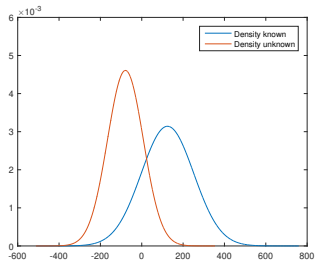

Figure 5.3: Densities for 60 observations, User 2

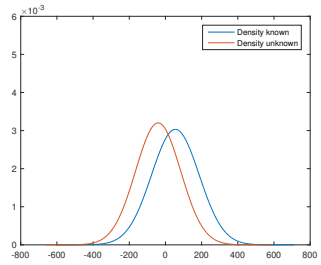

Figure 5.4: Densities for 600 observations, User 2

This effect is amplified if the probabilities are close to each other. Indeed, since we use Bayes' rule to compute the final probability of interest ($\hat{Pr}_{model}[P = p|E = e_n]$, see Eq. 4.3) the closer $\hat{Pr}_{model}[E = e_n|P = 1]$ and $\hat{Pr}_{model}[E = e_n|P = 0]$ will be, the more this probability will be close to 0.5 which is of course not desirable. User 2 was here chosen on purpose because the number of observations in each class is pretty similar (261 *known* vs 339 *unknown*) so it was expected that the two densities will be close to each other.

Success rate

At this point, we just need to compute the final probability thanks to the two others obtained with the model and Bayes' formula. We finally obtain for each observation the probability of belonging to its true class. At that stage, the success rate is easy to obtain. Here again, we chose to observe the evolution of the success rate in function of the proportion of the data set used to build the model.

Recall that the success rate computed is for the whole section of electrodes ob-

tained in section 3.2. We will first compute the curves for the 3 users of the firs experiment, then for the 3 users of the second separately, with their confidence interval computed for 600 observations. Please note that on the following figures, the SR was calculated as the averaged SR over the 50 folds.

An important quantity to join to our study is the proportion of *known* and *unknown* observations per user. Recall that these proportions are respected when building the training and test set, for more consistency. These proportions can be seen in Table 5.1.

Experiment 1

For this experiment, the users were asked to count each time they recognize a face on the screen. We expect that the action of counting induces a neural response significantly different that could be distinguished from the *unknown* observations. The users are **participative** with respect to the tasks.

	# known	# unknown
User 1	78	522
User 2	261	339
User 3	234	366

Table 5.1: Proportions of known vs unknown observations per user

	CI 95% for SR	Average SR
User 1	[0.5167;0.7500]	0.768
User 2	[0.3333;0.6167]	0.5661
User 3	[0.3500;0.6029]	0.6004

Table 5.2: 95% Confidence interval for the success rate per user

We can see that for all users, except user 1, the success rate seems to converge with the number of observations used. In this case, it can be interesting to check the probabilities computed by our model: this will be studied in chapter 6.

By looking at Fig. 5.5 and Table 5.2, we obtain a significantly higher success rate for the first user compared to the two others. This could be related to the fact that there is a big difference between the proportion of *known* and *unknown* observations (see Table 5.1). Note also that the intervals are wide, and contain

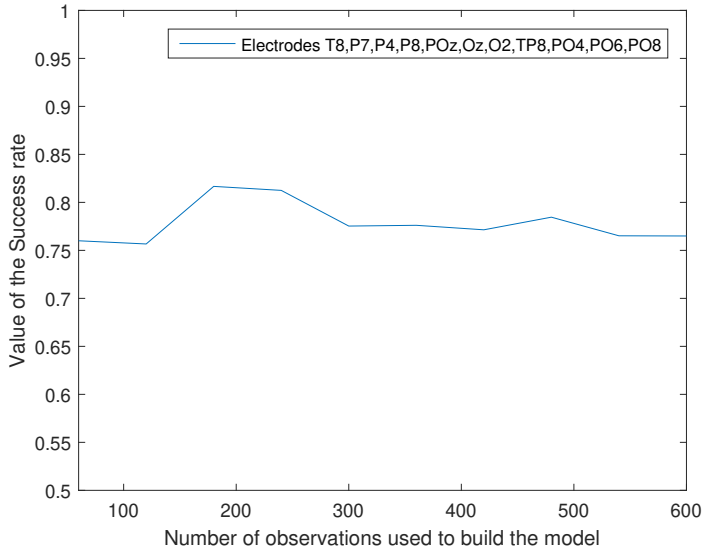

Figure 5.5: Evolution of the success rate in function of the size of data set used, User 1

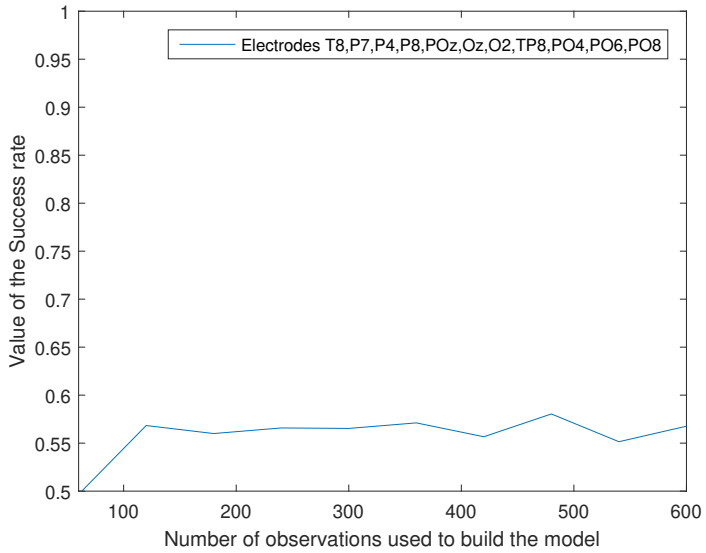

Figure 5.6: Evolution of the success rate in function of the size of data set used, User 2

0.5 for 2 on 3 users, which assesses that in some cases, the model presents bad performances. This will be further investigated in chapter 6.

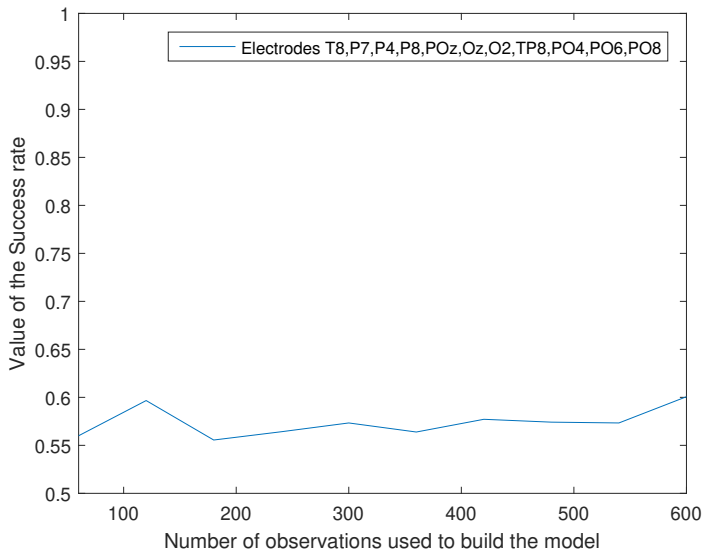

Figure 5.7: Evolution of the success rate in function of the size of data set used, User 3

Experiment 2

During this second experiment, the users were asked to count the number of time the fixation cross changes color, from blue to red. Since the users are **passive** in this case (no task related to the classes whatsoever), we expect a lower success rate.

	# known	# unknown
User 4	151	374
User 5	99	501
User 6	193	407

Table 5.3: Proportions of known vs unknown observations per user

In this case, compared to the previous experiment, we can see that the success rate curves tend to decrease in function of the size of data set used to compute the model. This can be rather counter intuitive but this effect will be examined more in detail in chapter 6. However, the values of the success rate does not seem to be so different from the previous experiment (Table 5.1). The CI for all users again contains the value 0.5. The fact that the users did perform a task unrelated to the definition of the classes does not seem to change that much in terms of results.

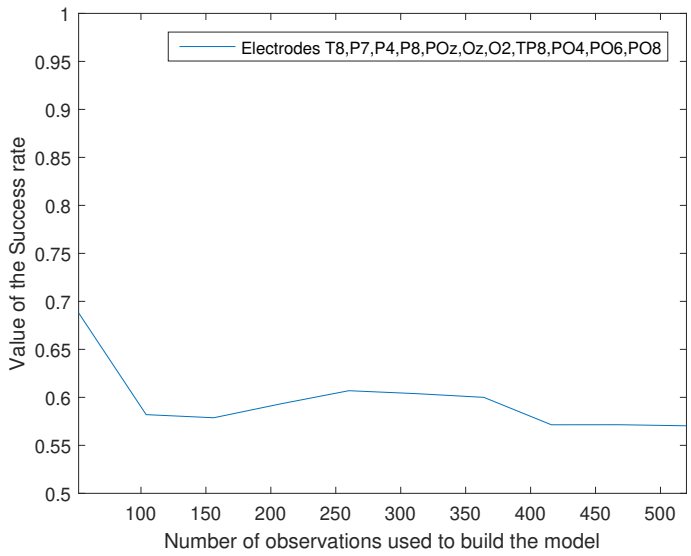

Figure 5.8: Evolution of the success rate in function of the size of data set used, User 4

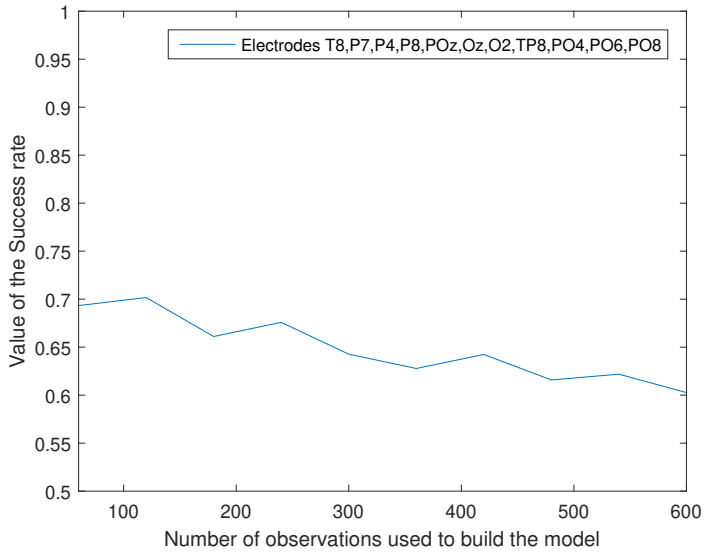

Figure 5.9: Evolution of the success rate in function of the size of data set used, User 5

However the model from user 1 computed in the first experiment seems to beat the other users from both experiments in terms of efficiency.

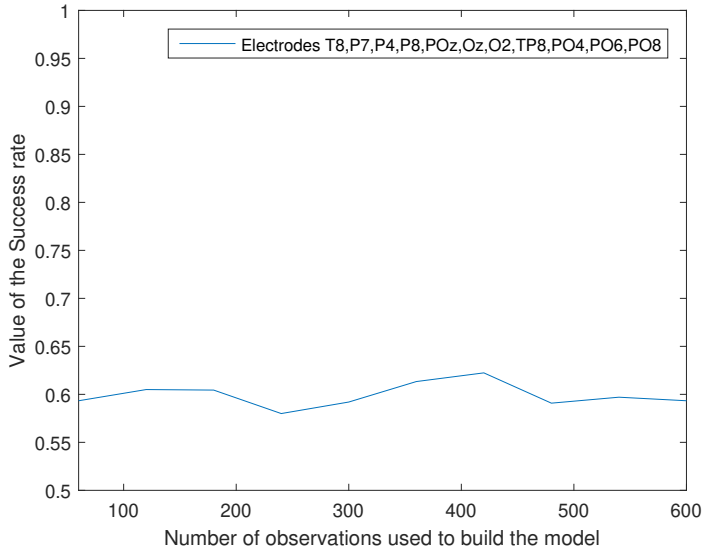

Figure 5.10: Evolution of the success rate in function of the size of data set used, User 6

	CI 95% for SR	Average SR
User 4	[0.4231;0.6923]	0.5678
User 5	[0.4000;0.6333]	0.6098
User 6	[0.3500;0.6335]	0.5974

Table 5.4: 95% Confidence interval for the success rate per user

5.2 Preferences

Let us now take a look at the subsequent analysis, performed with the same data as the previous experiments. As mentioned, we do not dispose of 600 observations to perform the analysis, since there was 3 check boxes in the document:

- I have a positive feeling about this person
- I have a negative feeling about this person
- I feel neutral towards this person

The items categorized as neutral are not used for this analysis. Note also that in this case, we are not able to determine what the target items are, since we are not able to predict which class of items will be presented with a lower frequency. However, since the users are not informed about the definitions of the classes in this case, this may not have a big influence on the results.

	Proportion of like vs dislike
User 1	102 vs 18 (120 total)
User 2	234 vs 99 (333 total)
User 3	149 vs 52 (201 total)
User 4	85 vs 96 (181 total)
User 5	224 vs 122 (346 total)
User 6	150 vs 54 (204 total)

Table 5.5: Proportions of observations per user for like and dislike class

In Table 5.5 we can observe the total number of observations used and the proportion of item from both classes.

Model computation

We will first observe the densities computed with the training set.

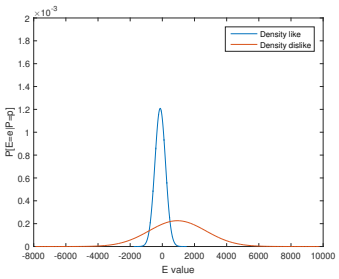

Figure 5.11: Densities for 24 observations, User 1

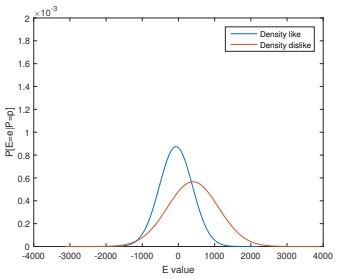

Figure 5.12: Densities for 120 observations, User 1

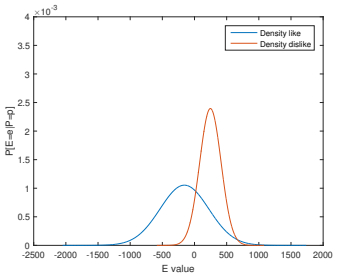

Figure 5.13: Densities for 30 observations, User 2

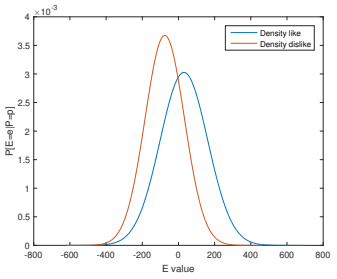

Figure 5.14: Densities for 330 observations, User 2

As we can see on Fig. 5.12, on the left part we start with 2 gaussians that are relatively close to each other. This figure corresponds to 24 observations

only, used to compute the two densities In addition, we see that the standard deviation of the dislike class is huge compared to the other class. This implies that $P[E = e_n|P = 1]$ and $P[E = e_n|P = 0]$ will be pretty different for a new point that would be somewhere between the means of the two densities. Of course, since there are only 24 observations used to build the classes this result can vary, but this is the general tendency. For the total number of observations, on the other hand, the standard deviation and the means of the classes are much closer to one another. This implies that the probabilities $P[E = e_n|P = 1]$ and $P[E = e_n|P = 0]$ will be closer for a point situated between the two means and that the probability to belong to one class will be closer to 0.5.

Success rate

We can now finally observe the curves representing the success rate in function of the observations used for our model. As mentioned all 6 users will be presented as a single group. The proportions of both classes can be seen in Table 5.5.

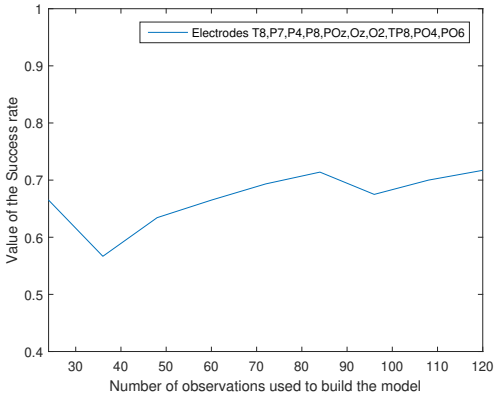

Figure 5.15: Evolution of the Success Rate in function of the size of the data set, User 1 (preference analysis)

In Table 5.6, we can observe the SR intervals and average for all our users. We can see that the Success rate is lower than in the previous study for all users, except user 6.

Recall that in this study, the users were all passive and had no clue about the classes, so we expected lesser SR. Note however that the Success rate may not converge completely, since we have even less data available than in the previous study.

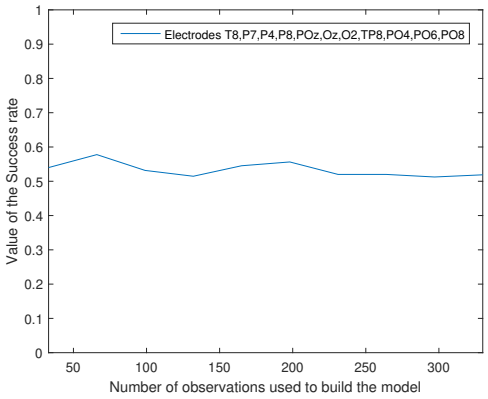

Figure 5.16: Evolution of the Success Rate in function of the size of the data set, User 2 (preference analysis)

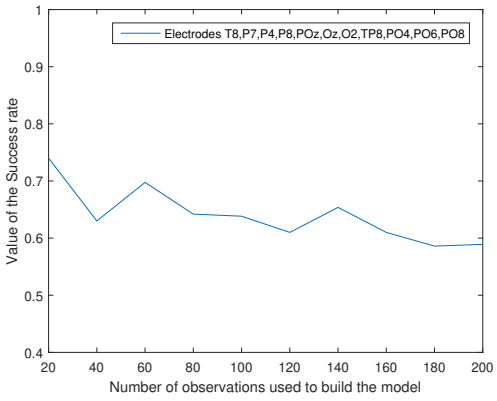

Figure 5.17: Evolution of the Success Rate in function of the size of the data set, User 3 (preference analysis)

	CI 95% for SR	Average SR
User 1	[0.5129;0.7215]	0.7126
User 2	[0.3014;0.5587]	0.5113
User 3	[0.3148;0.6168	0.5927
User 4	[0.4001;0.5439]	0.5005
User 5	[0.3561;0.5772]	0.5543
User 6	[0.3129;0.6356]	0.6125

Table 5.6: 95% CI for the Success Rate for all users, preference analysis

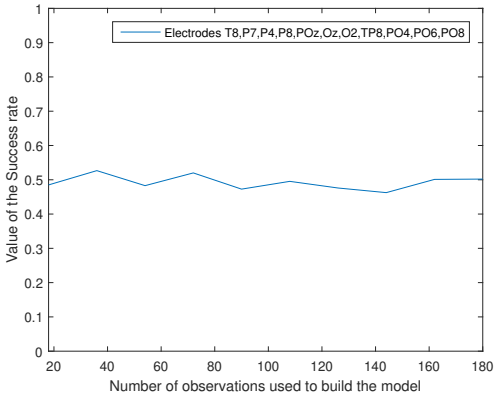

Figure 5.18: Evolution of the Success Rate in function of the size of the data set, User 4 (preference analysis)

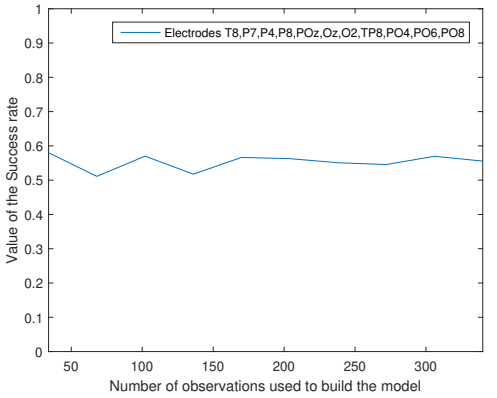

Figure 5.19: Evolution of the Success Rate in function of the size of the data set, User 5 (preference analysis)

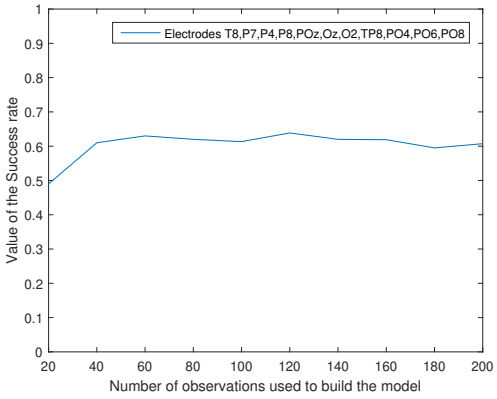

Figure 5.20: Evolution of the Success Rate in function of the size of the data set, User 6 (preference analysis)

Chapter 6

Discussion

In this section, we will discuss our results and try to explain why we reach them. First we will discuss the results of the face recognition analysis and the differences between the two experiments, then we will take on the preference study.

6.1 Experiment 1 vs experiment 2 : face recognition

Increasing and decreasing Success Rate

The first comment that we can make is that, unexpectedly, there seem to be no real differences between the results of the two experimental conditions. This is surprising because as mentioned, the first experiment included a counting task related to the classes. We expected thus to induce a stronger neural response in this case that in the other experiment, where the user was *passive* regarding the classes.

Another counter intuitive result is the fact that for some users, the Success Rate was decreasing inversely proportional to the size of the data set used to build the model (Fig. 5.8 and 5.9). By looking closely at the probabilities and densities computed, we can partially explain this effect. For users 4 and 5, when building the model after the PCA projection with 60 observations (Fig. 6.1 on the left), we obtain two normal densities with always one with positive and one with negative mean. The standard deviation is quite low and that means that the two classes are rather distinct.

So when one observation (let us say belonging to the *known* which has a density with negative mean) is projected on the eigen vectors, if the PCA was efficient, the value of the projected point will be negative aswell. Furthermore the probabilities

Figure 6.1: Densities of both classes for 60 and 600 observations, User 5

$P[E = e|P = p]$[1] for $p = 0$ and $p = 1$ will be unbalanced. One will be of the order of magnitude $10e - 3$ (for its actual class) against $10e - 7$ for the other or even less. This involves that the final probabilities $P[P = p|E = e]$, when Bayes formula (see Eq. 4.3) is applied will be far from 0.5 (close from 0 or close from 1). Since on average, the observations are projected efficiently following their classes (negative value for a*known* and positive for an *unknown* observation), the success rate value is rather high (average SR for 60 observations of 70%). In this case, the test set is composed of 6 observations.

Now looking at the densities for 600 observations (Fig. 6.1 right), we can see directly that the centers of the gaussian are very close and the standard deviation are of the same order. In this case $P[E = e_n|P = p]$ for $p = 0$ and $p = 1$ will be similar if the observation is projected on a value situated between the two means. This implies that $P[P = p|E = e_n]$ will be close to 0.5 in most of these cases. This explains why the success rate decreases for these 2 users. Recall that, for then case in which 60 observations are used, since there is 50 iterations for the computing of the densities and probabilities, the above explanation translates the general tendency throughout these iterations. There are always outliers and particular cases where the projection of the test set is not accurate, the densities are not representative... This variability is of course less present when the whole set is used to compute the training and test set.

Let us now take on the case where the success rate increases (user 1 for example). In this case, when 60 observations are used to compute the model, we can see (Fig. 6.2 on the left) that the *known* density has a high value of standard deviation compared to the other class. This means that there is a small window for which $P[E = e_n|P = p]$ will be greater if $p = 0$. For this particular user, there is only

[1]Recall that this quantity represents the probability for this observation to take value e knowing that it belongs to class p

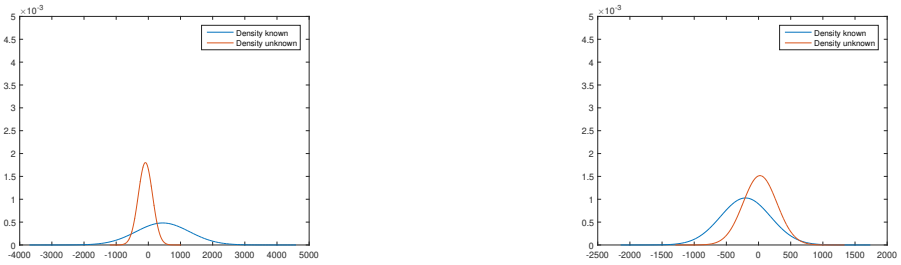

Figure 6.2: Densities of both classes for 60 and 600 observations, User 5

78 observations belonging to the *known* class ($p = 1$) against 522 of the *unknown* class. So the model has higher risks of making mistakes during the classification of the test set. When 600 observations are used, the densities computed are close to each other (Fig. 6.2 on the right) but the standard deviation of the *unknown* class as well as its peak of probability. So the probabilities that $P[P = p|E = e_n]$ (where $p = 0$) will be above 0.5 in most cases are much higher. Since there are more than 5 times more *unknown* observations for this user, the model has few risks of committing mistakes when evaluating the test set, since the proportions are respected.

Let us recall also that we made the hypothesis of **normality** for the distribution of our data, which was not exactly the case. This also has an impact on the performance of our model.

To put it in other words, the quality of the model will tend to increase if the final densities represent well the differences between the two classes, and will tend to decrease if the model has hard times to differentiate them.

The oddball paradigm in classification tasks

Let us investigate why the differentiation of the classes can be difficult for most of the users. Another surprising result is that the differences between the Success Rate of the two experiments were not so obvious. Let us think a bit outside the numbers and statistics and look at our experimental design. As mentioned in section 1.2.4, the P300 event is often used during classification tasks as this one, but usually in the context of use of the "oddball paradigm".

The idea is to present to the subject a target item mixed with non-target ones. An important feature of the paradigm is that the target item is shown with a low

probability compared to the standard items [15]. In our experimental conditions, the problem is that we cannot ensure that the target (in this context the targets are the *known* observations) will be presented with a low probability. As we saw in section 5, the proportion of *known* and *unknown* observations varies a lot between the users. For user 1 (78 *known* vs 522 *unknown*), we can say that the target was presented with a low probability, but this is not the case for all users.

If a user sees 3 faces that he recognizes in a row, the neural response may decrease by the phenomenon of habituation, making the differences between the two classes harder to observe [26]. That could explain the fact that we do not see huge differences between the results of the two experiments at first glance by looking at the average success rates. A result that could support this hypothesis is the fact that the user for which the Success Rate was the highest was user 1 (see Fig. 5.5), for whom we saw a clear differences in the proportions of the two classes of observations. As reminder, the higher bound of the CI for the SR for this user was 0.7500 and for all the other users this higher bound was around 60%. Also, user 1 is the only user for which the CI is entirely above 0.5.

We did not observe any user who presented such a high success rate in the second experiment (even for user 5 for which the classes were very unbalanced : 99 *known* against 501 *unknown*). We could thus think that the experimental conditions are the reasons that made the performances worse.

6.2 Preference analysis

For this experiment, we observe that the CIs for the Success rate seems lower than in the previous study, but we cannot conclude that it is significant at this stage. Let us recall that since we disposed of few data for each user, we cannot conclude that the SR converged at this point. We still observe a Success Rate higher for user 1, the only user for which the CI is again above 0.5. The fact that the CI are wide indicates also a rather high variance in terms of the model computed and test sets. Note that in this case, the type of classes is the *preference* of the user. The knowledge and thus, the P300 is not expected to be involved in the differences that we can observe between the two classes of observations.

Still as mentioned, the success rate are lower than in the previous study is lower for each user, except user 6. Indeed, the average success rate is close to 0.5, which clearly means that the model is inefficient. The average success rate, however is a poor indicator to compare the statistical difference between the success rates of the different experimental conditions. We will need to perform a hypothesis test to make our conclusions.

6.3 Hypothesis testing between the the different experimental conditions

To better assess the differences between the Success Rates obtained between the different experimental conditions, we will perform a Wilcoxon rank sum test on our data. This test is a non-parametric alternative to the classic hypothesis testing[2] considering the order in which the observations of the sample falls. To develop a little more, let us take two samples A and B that we want to compare with respectively n_A and n_B observations. All these observations are placed in a single vector following their *rank* : the smallest observation has rank 1, the second rank 2 and so on. The test will thus compare the sum of the ranks of the observations from each sample (A and B) and compare them [27]. The Wilcoxon rank sum test does not make the hypothesis of normality of the data, which is interesting in our situation. Indeed, we cannot ensure that the SR follows a normal distribution. The null hypothesis of this test is that the median of the two samples are the same, against the alternative hypothesis that they are not equivalent.

The conditions tested are the following :

1. Is there a difference between the success rates of the **focused** users and the **passive** users?

2. Is there a difference between the the success rates in the **face recognition** and in the **preference** analysis?

Due to our lack of data, we use as samples the SR computed in the 50 folds of our cross validation for each user.

For the first test, we have 50 values of success rates for each one of the 3 **participative** users[3] and 50 more for each one of the 3 **passive** users[4].

For the second test we have 50 values for each of the 6 users for the **face recognition** study and 50 values for the same 6 users for the **preference** study.

The null hypothesis for the two tests are :

1. $H0$: The distributions of the Success rate for **participative** and **passive** users are equivalent

2. $H0$: The distributions of the Success rate for the **face recognition** and **preference** study are equivalent

The alternative hypothesis for the two tests are:

[2] T-test or z-test
[3] Users 1, 2 and 3
[4] Users 4, 5 and 6

1. $H1$: The median of the distribution of the Success rate for **participative** users is greater than the median of the distribution of the SR for **passive** users.

2. $H1$: The median of the distribution of the Success rate for the **face recognition** study is greater than the median of the distribution of the SR for the **preference** study.

Since we want to assess if one distribution is higher than the other, we apply a **right tail** Wilcoxon rank sum test.

	P-values
Focused vs Passive users	0.0086
Face recognition vs Preference	0.0026

Table 6.1: P-values obtained by computing the Wilcoxon sum rank test on the data from the different experimental conditions

Thanks to this test, we can see that both tests reject the null hypothesis, assessing that the Success rate for the **participative** users is higher than for the **passive** users. The same goes for the **face recognition** and the **preference** study, with a p-value of 0.0026, which is pretty encouraging. These results should be taken carefully. As explained, the tests were performed through a little amount of users.

6.4 Summary of the studies

To summarize, we can say that the oddball paradigm, added to the fact that the classes are defined **a priori** by the investigator seems to be optimistic conditions to observe significant results. The conditions chosen in our two experiments and the results obtained tend to demonstrate that in a context where we have no control about the classes themselves, an efficient model is much harder to compute. In the *face recognition* study, the classes were defined **a posteriori** by the user and he has a hint about the definition of the classes. The class for each observation was thus different for each user. In the *preference study*, the difference was that the user had no clue about the classes whatsoever. Our CIs are pretty wide, which indicates a high variance in the effectiveness of our models. This means that they are very dependent of the training and test set chosen thus it will be hard to generalize these models to new data. However, the results of the Wilcoxon rank sum test tends to show that the different experimental conditions leads to significant differences between the Success rates obtained.

Let us now try to investigate what we could modify in our experimental protocol in order to make the classification more effective.

As mentioned, the element that could be the key to increase the differences in the neural response of the user is the number of target stimuli and their frequency of appearance, but also the order of their appearance in the set of images.

Indeed, we could limit the number of potential targets that the user could identify as *known* to a smaller proportion of the set. Recall that in our case, there were 50 faces of famous people that the users could potentially recognize, against 25 only that were of random people that the user could not recognize. If we apply this modification, the user will be presented many *unknown* items in a row, then suddenly will be shown a target item. This method could reduce the risk of the habituation phenomenon to a target item and increase the observable differences between the ERP of both classes. However, the question is left open.

These insights are also leading to another angle that could be used to analyze our data. Since we expect that the position of the target items in the set of stimuli could influence the Success rate, we could compute it in a different way. We could compute different Success rates following the position of target *known* items with respect to the position of the last target presented. In other words there will be a fixed number of categories defined as such : the last *known* observation before a *known* observation of index i was at position $i-1$, $i-2$, $i-3$,... until $i-n$. With n the maximum of indexes difference between observation at index i and another *known* observation. This will give us n different categories in which we will put all our *known* observations and an additional class where the *unknown* observations will be put.

Then we can compute the Success rate independently for each class and add them to obtain the general Success rate for this user.

Another investigation that would have been interesting is the test of different models to compute the Success rate and calculate the different probabilities. The Gaussian processes seems pretty accurate in this regard, since it allows to use the data of all electrodes at once for the classification [28]. We could also try to approximate the densities of the projected ERP after PCA computations by different densities than gaussian densities, such as Gaussian kernel approximations [29]. Again, the question is left open.

Chapter 7

Conclusion

Throughout this thesis, we tried to outline the difficulty to compute an efficient classification of EEG segments according to different experimental conditions.

We chose experimental conditions that were far from the ideal conditions of the articles presented in the introduction of this thesis. In the *face recognition* study the classes were determined **a posteriori** by the users and the users were *passive* according to the definition of classes in the second experiment. In the *preference* study, the users performed a task that had nothing to do with the definition of classes and were **uninformed** about them. The fact that we obtained acceptable results only for one user In the *face recognition* study in terms of efficiency of the model tells us that the task is harder than it seems at first glance. We believe that the fact that we observe a higher Success rate is due to a partition of the classes similar to the "oddball paradigm". For this user, the target items (*known* faces) were presented at low frequencies compared to the standard items.

The variance[1] of the Success rate for the other users makes the efficiency of the model very dependent of the training and test set selected. This also means that generalizing a model to multiple users would be an even harder task. Also the oddball paradigm seems to be a key component to obtain an efficient classification.

The results of the hypothesis testing however, seems to demonstrate that, as expected, the expected Sucess rate of the **participative** users was significantly greater than for the **passive** users . Also, the Success rates obtained in the **face recognition** were also significantly greater than in the **preference**. That would mean that, the **passive** of a user according to the definition of classes makes the classification more difficult, as expected. Also, the fact that we try to differentiate

[1]The variance is characterized by a wide CI for the Success rate

classes related to the preference of a user who is **uninformed** about the definition of classes makes the task even harder. As mentioned, since the amount of data used to perform these tests was small, these results should be taken carefully.

We also summarized the enhancements that could be applied to our experiment to obtain more efficient models (in term of Success rate) : lower the frequency of appearance of the *known* target items mixed with the non target items should increase the differences in the ERP between the two classes. Still, we have to keep in mind that this condition is hard to set up in a non-experimental context.

We also mentioned a new angle of analysis, taking into consideration the position of the last *known* observation with regards to the current one. The usage of different models to classify the data is also a lead that could be investigated.

Overall, we can say that in the context of utilization of brain-computer interfaces, the analysis of the EEG signals of the user can be very difficult in real life applications. The parameters and conditions in which BCIs are used would need to be carefully studied if one wants to extract information from a user. The taxonomy defined in this thesis could be extended by other studies and by adding more and more parameters. Several other types of model should also be evaluated for different parameters of this taxonomy. At one point we could be able to determine what model, what parameters and hyper parameters are optimal to extract information in a precise task. For example if we want to obtain hints about the **preferences** of a user who is **participative** with respect to the task and is **uninformed** about the definitions of the classes, the BCI would know exactly which model and which parameters to use. Of course, there is still a lot of investigation to do to broaden the taxonomy and the models themselves. At this point, we are only able to obtain efficient models in very specific conditions.

Bibliography

[1] Yasmin Afina. Human control is essential to the responsible use of military neurotechnology. https://www.chathamhouse.org/expert/comment/human-control-essential-responsible-use-military-neurotechnology?gclid=CjwKCAiAjMHwBRAVEiwAzdLWGBk7grSF2BQm6Rs_8exCzni00ofGU_CWO

[2] Tamara Bonaci, Ryan Calo, and Howard Chizeck. App stores for the brain: Privacy security in brain-computer interfaces. volume 34, 05 2014.

[3] Reza Fazel-Rezai Kamyar Abhari and Amir Meghdadi. Lie detection using brain p300 signal: Preliminary results. *CMBES*, 30(1):4, 2007.

[4] H. Pratt I. Berlad. P300 in response to the subject's own name. *Electroencephalography and clinical Neurophysiology*, 96(1):472–474, 1995.

[5] Wikipédia. P300 (neuroscience). https://en.wikipedia.org/wiki/P300_(neuroscience).

[6] Anne-Marie Schuller Boutheina Jemel and Valérie Goffaux. Characterizing the spatio-temporal dynamics of the neural events occurring prior to and up to overt recognition of famous faces. *Journal of cognitive Neuroscience*, (1):2289–2305, 2010.

[7] J. Iglesias C. Saavedra and E. I. Olivares. Event-related potentials elicited by the explicit and implicit processing of familiarity in faces. *Clinical EEG and neuroscience*, 41(1):24–31, 2010.

[8] James Tanaka, Tim Curran, Albert Porterfield, and Daniel Collins. Activation of preexisting and acquired face representations: The n250 event-related potential as an index of face familiarity. *Journal of cognitive neuroscience*, 18:1488–97, 10 2006.

[9] Sophie Boddington Danielle Droucker Tim Curran Lara J. Pierce, Lisa S. Scott and James W. Tanaka5. The n250 brain potential to personally familiar and newly learned faces and objects. *Front Human Neurosci.*, 5:111, 10 2011.

[10] Clément Massart François-Xavier Standaert Kenzo Fujii, André Mouraux. Taxonomy for eeg-based privacy experiments. October 2018.

[11] Stephan Boehm and Werner Sommer. Neural correlates of intentional and incidental recognition of famous faces. *Brain research. Cognitive brain research*, 23:153–63, 06 2005.

[12] Ajaya Neupane. A multi-modal neuro-physiological study of phishing detection and malware warnings. 10 2015.

[13] Michael Inzlicht, Ian Mcgregor, Jacob Hirsh, and Kyle Nash. Neural markers of religious conviction. *Psychological science*, 20:385–92, 04 2009.

[14] Ivan Martinovic, Doug Davies, Mario Frank, Daniele Perito, Tomas Ros, and Dawn Song. On the feasibility of side-channel attacks with brain-computer interfaces. pages 34–34, 08 2012.

[15] J. Peter Rosenfeld Bruno Verschuere. *Memory Detection, Theory and Application of the concealed information test.* Cambridge University Press, 2011.

[16] Lawrence Farwell and Emanuel Donchin. The truth will out: Interrogative polygraphy ("lie detection") with event-related brain potentials. *Psychophysiology*, 28:531–47, 10 1991.

[17] Netiwit Kaongoen, Moonwon Yu, and Sungho Jo. Two-factor authentication system using p300 response to a sequence of human photographs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, PP:1–8, 10 2017.

[18] Mario Frank, Tiffany Hwu, Sakshi Jain, Robert Knight, Ivan Martinovic, Prateek Mittal, Daniele Perito, Ivo Sluganovic, and Dawn Song. Using eeg-based bci devices to subliminally probe for private information. pages 133–136, 10 2017.

[19] UCL Institute of Neuroscience(IONS). Letswave 5 software. http://www.nocions.org/letswave6/.

[20] R. Strungaru V. Lazarescu M. Ungureanu, C. Bigan. Independent component analysis applied in biomedical signal processing. *MEASUREMENT SCIENCE REVIEW*, 4:1–8, 11 2004.

[21] Scholarpedia Simon R. Schultz (2007). Signal-to-noise ratio in neuroscience. http://www.scholarpedia.org/article/Signal-to-noise ratio_in _neuroscience.

[22] Plasticity Brain Center. The role of the occipital lobe. https://www.plasticitybraincenters.com/media/the-role-of-the-occipital-lobe/.

[23] Wikipedia. Principal component analysis. https://en.wikipedia.org/wiki/Principal$_c omponent_a nalysis$.

[24] Die Hu Ying Wen Meng Wan Jun Long Lianghua He, Bin Liu. Motor imagery eeg signals analysis based on bayesian network with gaussian distribution. *Neurocomputing*, 188(1):217–224, 2016.

[25] Joses Ho. Bootstrap confidence intervals. https://cran.r-project.org/web/packages/dabestr/vignettes/bootstrap-confidence-intervals.html.

[26] E. V. Megalou W. N. Frost. Learning and memory in invertebrate models: Tritonia. *ScienceDirect*, 1(1):401–404, 2009.

[27] Chris Wild. The wilcoxon rank-sum test. https://www.stat.auckland.ac.nz/ wild/ChanceEnc/Ch10.wilcoxon.pdf.

[28] H. Nickisch and CE. Rasmussen. Approximations for binary gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, October 2008.

[29] Andrew Cotter, Joseph Keshet, and Nathan Srebro. Explicit approximations of the gaussian kernel, 09 2011.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**