Faculty of Engineering and Architecture
Departement ELIS
2015–2016

Recognize Emotion in the brain using EEG Data

by

Andreas DE LILLE

Promotors:	Prof. J. DAMBRE
	Dr. Ir. P. VAN MIERLO
Assistent:	Ir. T. VERHOEVEN

Contents

Nomenclature

ANET Affective Norms for English Text

ANEW Affective Norms for English Words

BCI Brain Computer Interface

CSEA Center for the Study of Emotion and Attention

CSP Common Spatial Patterns

DEAP Dataset for Emotion Analysis using Physiological Signals

EEG Electroencephalography

ERP Event Related Potential

Fm Frontal Midline

IADS International Affective Digital Sounds

IAPS International Affective Picture System

LDA Linear Discriminant Analysis

MEG magnetoencephalography

SAM Self-Assessment Manikin

SAM self-assessment manikins

.

1

Introduction

This chapter describes the context of the thesis, starting with brain computer interfaces(BCI), before defining some BCI basics. After that, the P300 speller and P300 paradigm are introduced. Before the need for an emotionally aware P300 speller is justified, the basic process of emotion in the brain is explained.

1.1 Brain computer interfaces

A Brain Computer Interface (BCI), creates a direct neural link from the brain to the computer[1], that tries to recognize patterns and based on the extracted information, performs actions. A BCI removes to need for physical actions, i.e. typing or moving a mouse, for the transfer of information. The neural link provided by the BCI is made of two important components. The first component is the extraction component, which extract brain signals from the brain. The second component is the computer that interprets signals and performs actions based on the outcome.

1.1.1 Electroencephalography (EEG)

Different technologies exist to analyze the brain, the most convenient method is via Electroencephalography (EEG), since it is a non-invasive method. Non-invasive methods, in contrast to invasive methods require no surgery; they simply measure electrical activity using electrodes placed on the scalp.

The electrical activity in a brain is caused when an incoming signal arrives in a neuron. This triggers some sodium ions to move inside the cell, which in turn, causes a voltage rise[2]. When this increase in voltage reaches a threshold, an action potential is triggered in the form of a wave of electrical discharge that travels to neighboring neurons. When this reaction occurs simultaneously in a lot of neurons, the change in electrical potential becomes significantly, making it visible to the EEG surface electrodes. EEG can thus only capture synchronized activity of many, many neurons.

Signals originating from the cortex, close to the skull, are most visible, while signals originating deeper in the brain cannot be observed directly. Even for signals originating close to the cortex, EEG is far from precise as the bone between the the cortex and electrodes distorts the signal.

Additionally other artifacts like eye and muscle movement add a lot a noise to the signal, noise removal techniques are therefor advised. Even though the noise is persistent and EEG data has very low spatial resolution, it still can provide significant insight into the electrical activity of the cortex while offering excellent temporal resolution[3].

Note that EEG records electrical activity, other methods like magnetoencephalography (MEG) measure brain activity using magnetic fields. Since MEG is more prone to noise from external magnetic signals, i.e. the earth's magnetic field and electromagnetic communication, a magnetic shielded room is required, making this method very expensive and not mobile.

EEG uses electrodes which are placed on the scalp to measure the electrical activity. To ensure that experiments are replicable, standards for locations of electrodes have been developed. One of these systems is the 10/20 system, an internationally recognized methods to describe location of scalp electrodes[4]. The numbers 10 and 20 refer to the distances between the electrodes, which are either 10% or 20% of the total front-back or left-right distance of the skull. Each site is identified with a letter that determines the lobe and hemisphere location.

- **F:** Frontal
- **T:** Temporal
- **C:** Central
- **P:** Parietal
- **O:** Occipital

Note that no central lobe exists; the C letter is only used for identification purposes. The letter z indicates that the electrode is placed on the central line. Even numbers are use for the right hemisphere, while odd numbers are used for the left hemisphere. A picture of a 23 channel 10/20 system is added below for clarification. Even though some experiment setups may use a different set of channels than shown in figure 1.1, they all follow the same naming convention.

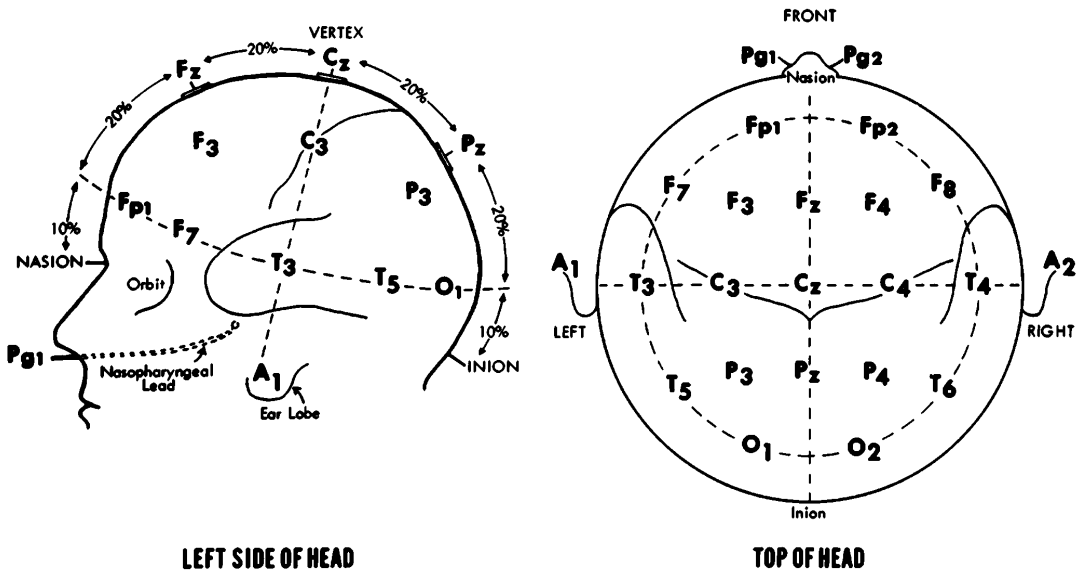

Figure 1.1: The electrode placement of a 23 channel system[5].

Two different types of EEG channels exist, monopolar and dipolar. A monopolar channel records the potential difference of a signal, compared to a neutral electrode, usually connected to an ear lobe of mastoid. A bipolar channel is obtained by subtracting two monopolar EEG signals, which improves SNR by removing shared artifacts[6].

In the frequency domain, brain waves are usually split up into different bands[7, 8], each band has a different medical interpretation. These wavebands are:

1. **Alpha:** 8-13Hz, indicate how relaxed and/or inactive the brain is.
2. **Beta:** 13-30HZ, indicate a more active and focused state of mind.
3. **Gamma:** 30-50Hz, relate to simultaneous processing of information from different brain areas.
4. **Delta:** 0-4hz, these waves are generated during dreamless sleep and meditation.
5. **theta:** 4-8Hz, occur during dreaming.

Most muscle and eye artifacts have a frequency around 1.2Hz. Artificats caused by nearby power lines, have a frequency around 50Hz[2]. To remove most of this noise, a bandpass filter is usually applied to filter out frequencies below 4Hz and above 40-45Hz.

Event related potentials

An Event related potential (ERP), is a measured brain response to an event, measured by EEG or MEG. The ERP nomenclature usually starts with a letter that indicates the polarity: the P corresponds to a positive polarity, while the N indicates a negative polarity. The number indicates the mean latency, measured between the ERP and the stimulus, which might may variate significantly between subjects.

The most important one is the P300 wave, which is usually elicited using the oddball paradigm. The oddball paradigm is the occurence of a low probability target item between high probability targets, e.g. flashing a specific symbol in a grid of different symbols. It consists of two components, the P3a with a latency of 240ms and the P3b with a latency of 350 ms[9]. The later component, P3b only occurs when the subject actively counted either the targeted or more frequent stimuli.

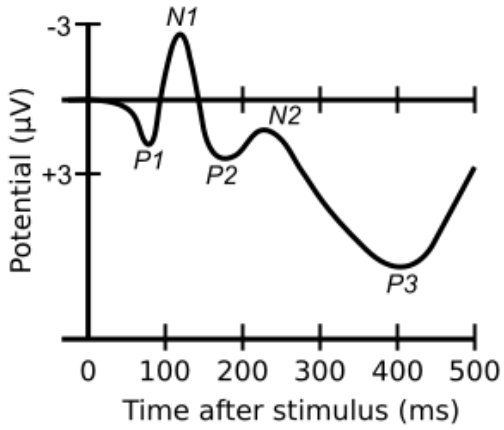

Figure 1.2: The Different ERP linked to an oddball paradigm, found at[10].

1.1.2 The P300 speller

The P300 speller is an active topic of research that uses EEG data to enable persons with the locked in syndrome to communicate[11]. The basic version uses a six by six grid of characters, each row and column is flashed in a random order while the subject silently counts the number of flashes of a certain character, as shown in figure 1.3. This procedure, where a train of stimuli with some infrequent occurring target stimuli is applied, is called the oddball paradigm[12]. It is known that this technique triggers an increase in the potential difference in the EEG around the parietal lobe. This ERP occurs +- 300 milliseconds after the stimuli is flashed, hence its name, the P300 waveform[13]. The presence or absence of the P300 waveform is used by the P300 speller to determine what character the subject was focusing on, which basically allows the subject to spell text.

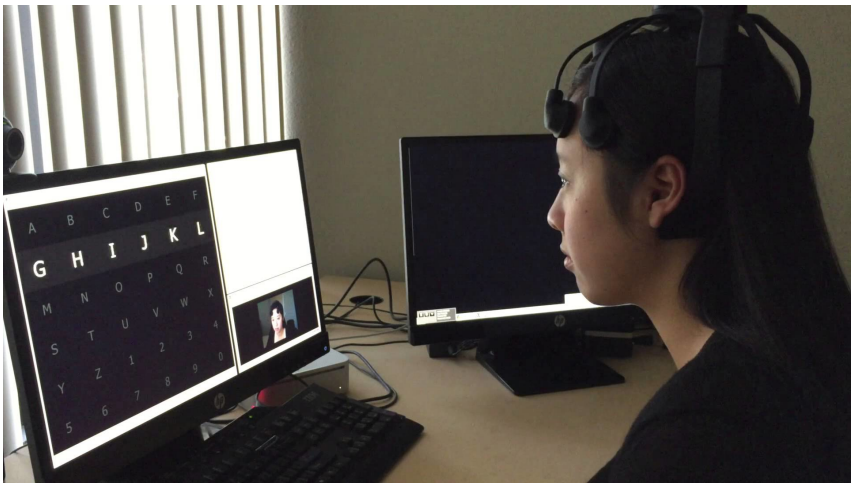

Figure 1.3: Different parts of the P300 speller, found at [14].

To improve the spelling time, many improvements and research has been done. Language models were used to predict the word based on the first characters, which enabled great speedups[1],

classifiers were compared and tested on both healthy[15] and unhealthy subjects[13]. Since many unhealthy subjects might have an impaired vision or eye movement, tactile[16] and auditory[17] spellers have been developed to circumvent this problem.

To improve accuracy, common problems such as adjacency distraction, when a subject is distracted by a neighboring flash, and double flashes, when the target row and column are flashed close after each other, were avoided using new randomized paradigms[12]. Other input layouts like the T9 interface P300 speller have also been developed[18].

To further speedup the spelling, error potentials were explored. Error potentials are triggered when the user becomes aware of an erroneous action[19], i.e. when a wrong character is selected. When an Error potential is detected, the character is usually changed to the second most probable character according to the P300 decoding[20], which is the most viable character.

The basic P300 speller needs a calibration period before it can be used, when a healthy subject makes a mistake during calibration, he can simply communicate this. This is not the case for a patient, who has no other means of communication than the P300 speller. Having wrongly labeled data during calibration can lead to severe problems. The unsupervised speller as proposed in [21] solves this problem by removing the need for a calibration procedure. The speller works with expectation maximization and has an undemanding linear classification backend. This system starts with a warm-up period where the system adapts to the given condition.

1.2 Emotion recognition

Psychology makes a clear distinction between physiological behavior and the conscious experience of an emotion, called expression[2]. The expression consists of many parts, including the facial expression, body language and voice concern. Unlike expression, the physiological aspect of an emotion, e.g. heart rate, skin conductance and pupil dilation, is much harder to control. To really know one's emotions, it seems, one has to research the physiological aspect of the emotion. One possibility for this is analysis of brain activity via Electroencephalography[22], which is the main method for this thesis.

1.2.1 Emotion in the brain

Before emotions can be recognized, a classification model is needed. A common model to classify emotions is the bipolar arousal-valence model[2, 23], that places emotions in a two dimensional space. The main advantage of using a multidimensional model, is that all emotions are modeled in its space, even when no particular discrete label can be used to define the current feeling. Figure 1.4 shows the mapping of different emotions for this model.

Even though arousal and valence describe emotion quite well, a third dimension can also be added. The new model then has three dimensions: arousal, valence and dominance. Arousal indicates how active a person is and ranges from inactive, bored to active, excited. The valence indicates if the emotion is perceived as positive or negative. The third dimension, the dominance, indicates how strong the emotional feeling was and ranges from a weak feeling to an empowered, overwhelming feeling. The dominance component can aid to filter out samples of strong feelings, since feelings with low dominance are less likely to show significant effects.

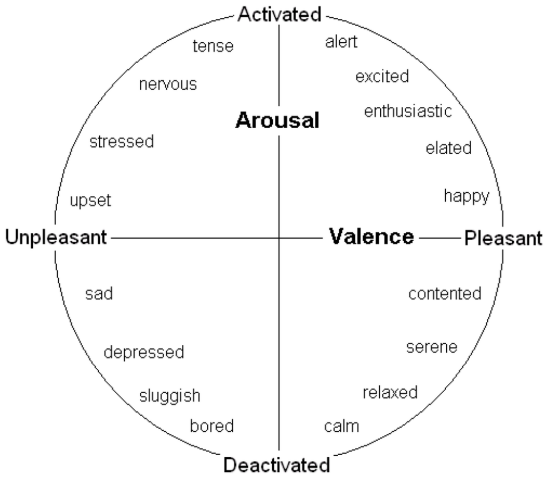

Figure 1.4: The arousal - valence model maps emotions in a two dimensional plane.

Determining valence

The most known and used feature is the frontal asymmetry of alpha power[3]. The right hemisphere is generally speaking, more active during negative emotion than the left hemisphere which is in turn more active during positive emotions[23, 22]. The asymmetry is given for L and R being the Left and Right frontal alpha powers as:

$$Asymmetry = \frac{L-R}{L+R}$$

Computing the spectral power of the alpha band is possible via e.g. the fast Fourier transform or wavelet transform.

It is also possible to include beta waves in the process. High alpha rates correspond with an inactive brain, while high beta waves with an active brain. Looking for an increase in beta activity and a decrease in alpha activity at one side, while the other hemisphere should show an increase in alpha waves and a decrease of beta waves as indication that is becomes less active, offers an insight to the frontal asymmetry and thus the valence[2].

Another feature is the frontal midline (Fm) theta power, that is shown to increase with increasing pleasant ratings for audio stimuli[24].

Determining arousal

Arousal can be determined in several ways. A first methods for the arousal recognition uses only alpha and beta bands. the Alpha band corresponds with a relaxed state and is often connected to brain inactivity[2, 23]. Beta waves, on the other hand, are an indication that the brain is a more active state or has a higher level of arousal. Combining these two parameters gives the beta/alpha ratio as an indication of the arousal level.

Other methods are based on the EEG coherence across the prefrontal and posterior beta oscillations, which is known to increase when high arousal images are viewed. Additionally, gamma power is said to increase with arousal after a delay of 500ms[3].

Datasets

One of the most used datasets in the context of emotion recognition is the Dataset for Emotion Analysis using Physiological Signals (DEAP) dataset[25]. This dataset contains EEG samples at 512 Hz of 32 persons each viewing 40 videos. A preprocessed version of this dataset, that is down sampled to 128Hz and has EOG removal will be used extensively during this thesis.

Furthermore, the center for the study of emotion and attention (CSEA), by the university of Florida made several visual datasets available:

- **IAPS:** International Affective Picture System
  This dataset consists of a large set op emotional stimuli in the form of color photographs.
- **IADS:** International Affective Digital Sounds
  This dataset consists of acoustic emotional stimuli designed for investigation and research of emotion and attention.
- **ANEW:** Affective Norms for English Words
  This dataset provides a set of emotional ratings for a large number of english words.
- **ANET:** Affective Norms for English Text
  This set provides normative ratings of emotion for a large set of brief texts.
- **SAM:** Self-Assessment Manikin
  A non-verbal pictorial assessment technique that measures the pleasure, arousal and dominance associated with a person's affective reaction to a wide variety of stimuli.

The stimuli from these sets are used in experimental setups to trigger emotions[2, 26, 23, 22].

1.2.2 Benefits of creating an emotionally aware P300 speller

Emotions play a major role in non-verbal communication, are quite complex and essential to understand human behavior. The ability to recognize emotions will improve the ability of computers to understand human interaction[27] and are likely to improve the P300 speller's accuracy in different ways.

To improve the detection of error potentials, emotion can be used. It is expected that when a person makes a lot of mistakes, his/her emotional state will change to a less happy, more frustrated state. Making the speller emotionally aware, could improve the detection of error potentials.

Research with visual stimuli on healthy subjects, show that emotion has an effect on the auditory P300 wave[28]. Both the P300 peak amplitude and area was highest when viewing neutral pictures and descended further, in decreasing order, for sadness, anger and pleasure. The amplitudes were significantly lower at both Fz and C3 positions than Pz and Oz. The latency

of the P300 ERP speller was shortest or neutrality and in increasing order longer for pleasure, anger and sadness. It is expected that a visually triggered P300 wave, will also be influenced by emotion. Detecting emotion can therefore improve the detection of P300 waves.

Contrary to what subjects might think, the P300 speller is unable to read the mind and know what a person is thinking about[11]. The P300 speller provides no more than a means of communication that the subject can use. Should he chose to ignore the instructions and focus his attention elsewhere, then the recordings become useless. Nevertheless, ethical question often remain unanswered. Knowing how the subject feels, can help him communicate more humane on one hand, while providing more insight for ethical issues, on the other hand, e.g. "How does the subject think about the P300 speller recording and analyzing his brain activity?". Information about the subject's emotional state can help answer some of these ethical questions.

1.3 Goal of the thesis

This thesis aims to improve the performance of the P300 speller by making the speller emotionally aware. An emotional aware speller is expected to yield better performance, since the P300 wave is affected by the emotion. Furthermore, the detection of error potentials can be combined with the emotional state, since the emotional state is expected to change with increasing errors.

More concrete, the main goal is to recognize emotions in the arousal-valence model, using the DEAP dataset. First the emotions of a single person should be recognized, since the features are known to differ from person to person. Later, the model will be adjusted so that it is capable of detecting emotion of different persons. Once the emotion recognition is able to classify the emotions with decent accuracy, it will be integrated in the P300 speller, which should give additional accuracy and aid in the error potential detection. The expected results are:

- Being able to recognize emotions of a single person
- Being able to recognize emotions across different persons
- Improved accuracy for the P300 speller
- Improved error potential recognition for the P300 speller

Additionally the gained information for emotion recognition can be used for other ethical research studies, to answer ethical questions about the usage of BCI on patients.

2

Machine Learning

2.1 What is machine learning?

Since machine learning is a very broad domain, this discussion will be limited to the application of machine learning and machine learning techniques as this is the most relevant part. Machine learning is the science of getting computers to act without being explicitly programmed. Machine learning uses pattern recognition to find patterns or structure in the data. When applied correctly, machine learning algorithms can recognize objects in images, e.g. Optical Character Recognition (OCR).

Quite vague explanation so far, let's illustrate this with a small example. Suppose one has a price list of house that are for sale combined with their total area. Now since it is easy to see that a bigger house will have a higher asking price than a smaller house, one says that the area of the house is correlated with the asking price. Suppose you want to predict how much a certain home is worth, based on their area. This is possible with machine learning, you first learn/teach or train your machine learning algorithm with the aforementioned list. Once this is done you can predict prices of new houses based on the corresponding area. The area of the house is only one property or feature for the algorithm. So far the algorithm can only look at one property and thus is not likely to achieve good results. Looking with more detail at the data, i.e. adding additional features will often improve the performance of an algorithm. For example, a house with 5 bedrooms is more expensive than a house with only 3 bedrooms.

Machine learning algorithms are responsible for the relation between features and the predicted value. One way to group this algorithms is to look at their produced output. In the asking price examples above, the output is a price, which is a continuous value. When one tries to recognize characters in a picture, the problem becomes a classification problem, as there are only 26 distinct characters.

Another way to group the algorithms is based on their training data. In the asking price examples above one gets labelled results; the asking price is given for each area, this is referred to as supervised machine learning. The other possibility is unsupervised machine learning, which often results in finding groups of similar data points (clustering). Not that the combination of supervised and unsupervised data is also possible. Support you have a dataset with 5000 webpages and you want to categorise them in 10 distinct categories, e.g. science, nature, cooking, ... , but you only have the labels for 100 of the 5000 pages. Then you could first cluster the

pages in similar groups. As soon as a group contains one labelled page, you can label all the pages in the group as they are similar. This technique is known as semi supervised learning and has the advantage that one can also use unlabelled data. Note that labelled data is usually quite rare; if you had a fast and easy way to label the data then you wouldn't be needing machine learning.

2.2 over and underfitting and the relation to high bias and high variance

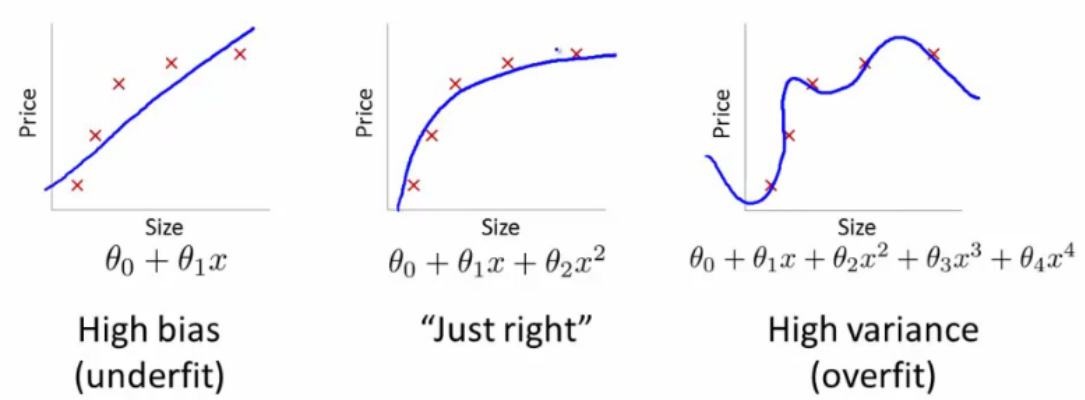

Figure 2.1: Overfitting versus underfitting[**?**].

Suppose the example in Figure 2.1, where one tries to find a good function to fit the given data points. Looking at the three proposed functions, one can easily see that the middle figure is the most likely generator function of the red points.

The figure on the left corresponds to an underfit, where the proposed function is not able to capture sufficient detail of the points. The function is not complex enough to approach the generator function, which is known as high bias. A high bias problem has a high training error, as the function is not able to fit the points correctly, this is visible in Figure 2.2

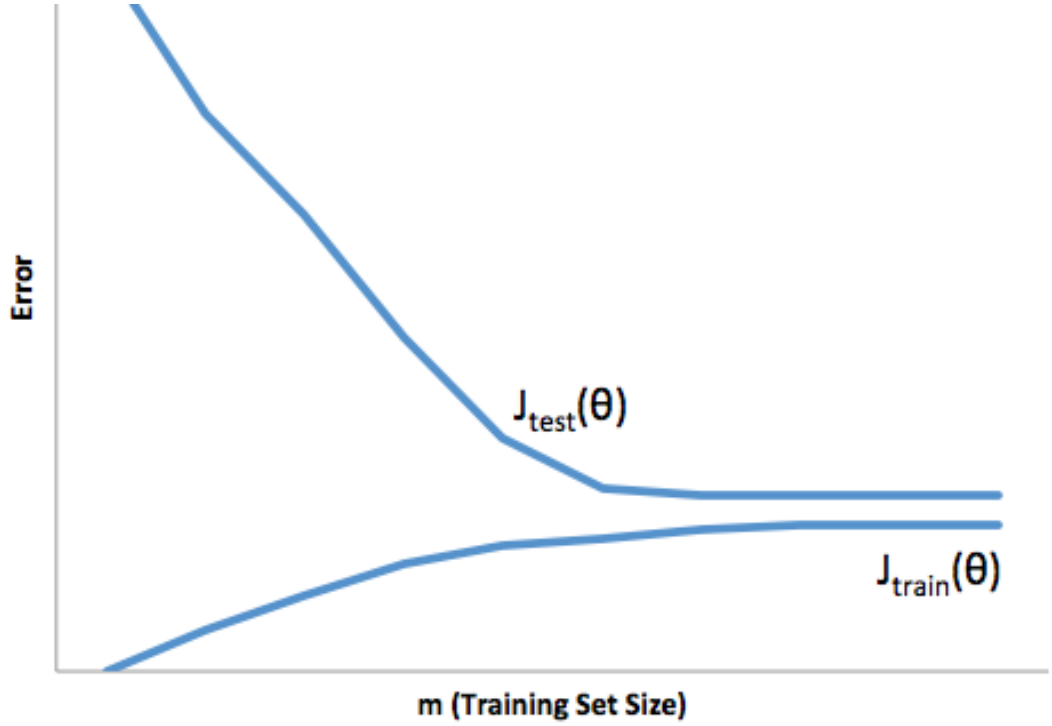

Figure 2.2: A high bias function is not complex enough to approach the generator function closely.

The function on the rights The function on the right corresponds to an overfit; the function 'goes through' each point exactly, but one can see that in between data points the behaviour of the hypothesis function is not logical. This problem is known as a high variance problem, where the train error is close to zero, but the test error is quite dramatic.

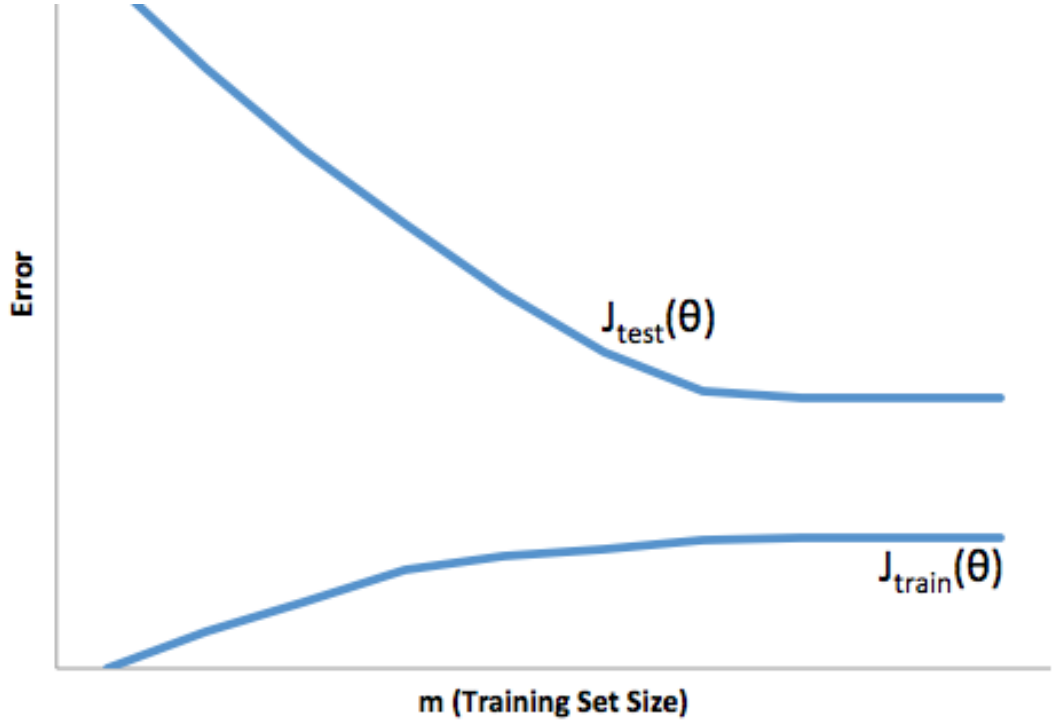

Figure 2.3: A High variance function is too complex and fits the data point too closely.

Another way to explain the bias variance tradeoff is by an example. Suppose you have a dart board, as shown in Figure 2.4. Suppose the situation on the top left corner, this corresponds to a world class player that has perfect aim, and very little variation on his precision. The situation on the left bottom corresponds to a player that has very little variation on his precision, but that is consistently aiming too high, he is biased to hit higher than needed. The pictures on the right side are different, there the person may or may not have a biased aim, but it is clear that he has a lot of variation in the precision of his aim.

In the context of machine learning, the low bias corresponds to having a hypothesis set that is close to the generator function, which allows you to get quite close. However you still have to pick the right function from that set, which is hard to do if you don't have enough data. If you are not able to take the best solution from the hypothesis set, you have high variance.

Figure 2.4: The bias variance explained using the dartboard example found at [**?**]

2.3 Support Vector Machines (SVM)

2.4 Linear Discriminant analysis (LDA)

2.5 Common Spatial Patterns (CSP)

2.6 Random Forests (RF)

2.7 Pearson correlation

3

A first look at the data

3.1 The DEAP dataset

This thesis uses the DEAP dataset[25], a dataset for emotion analysis that is publicly available for academic research. This dataset contains EEG recordings of 32 participants, each watching 40 one minute excerpts of music videos. Each video was rated individually by each person on valence, arousal, dominance and liking. The first three ratings correspond to the valence, arousal and dominance space of an emotion 1.2.1. The liking component indicates how much the person liked the video excerpt and should not be confused with the valence component; it inquires information about the participants' tastes, not their feelings, i.e. a person can like a video that triggers angry or sad emotions. The liking rates are neglected, since they are not part of the emotion space.

For assessment of these scales, the self-assessment manikins (SAM), were used[25]. SAM visualizes the valence, arousal and dominance scale with pictures, each picture corresponds to a discrete value. The user can click anywhere in between the different figures, which makes the scales continuous. All dimension are given by a float between 1 and 9, but for the context of this thesis, a preprocessing step scaled and translated these values to ensure they range between 0 and 1.

The used SAM figures are shown in Figure 3.1. The first row gives the valence scale, ranging from sad to happy. The second row shows the arousal scale, ranging from bored to excited. The last row represents the different dominance levels. The left figure represents a submissive emotion, while the right figure corresponds with a dominant feeling.

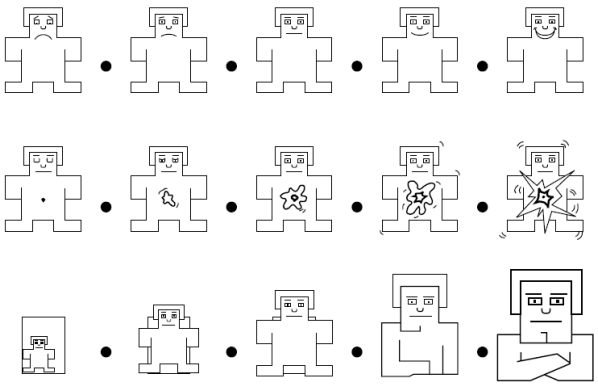

Figure 3.1: The images used for the SAM[25].

To further inspect the distribution of the user ratings and whether or not the data is balanced, the average for each emotion dimension (valence, arousal and dominance), was determined using all videos of all persons. These can be seen in Table 3.1. A uniform distribution, which is the ideal case for machine learning, should give a value of 0.5. The averages of the different dimensions are a just a little above 0.5, which gives a first indication that overall, the data is only slightly unbalanced.

	Valence	**Arousal**	**Dominance**
value	0.532	0.520	0.548

Table 3.1: The average value of each component.

3.2 A first model to classify the valence

The first model in this thesis recognizes the valence of a single person. This was done with a linear SVM classifier in a first attempt, since SVMs were used in a lot of the covered literature. SVMs have the advantage that they are able to handle limited datasets well, because they only look at points close to the separation boundary. In this attempt 2 features were compared, the usage of the frontal alpha asymmetry and the frontal theta power. Both feature are in literature frequently reported as being correlated to the valence. A third feature set used a combination of both these features.

The first classifier labels the valence values into two classes: low valence (unpleasant feelings) and high valence (pleasant feelings). This also meant that labels needed to be assigned to the dataset. This can be done in different ways.

The first and most straightforward method simply splits the valence range in half, all values positioned in the first and second half of the valence region were assigned to the low and high valences classes respectively. Note that even though the data set is overall quite balanced, the situation becomes quite different in case of a person specific classifier. Looking at Figure 3.2, it is clear that for some persons the data is quite unbalanced, for examples the unbalance is quite high for person 6. There are only 10 examples assigned to the low valance class, while 30 examples are assigned to the high valence class. Having an unbalance in training examples leads to less accurate classifiers.

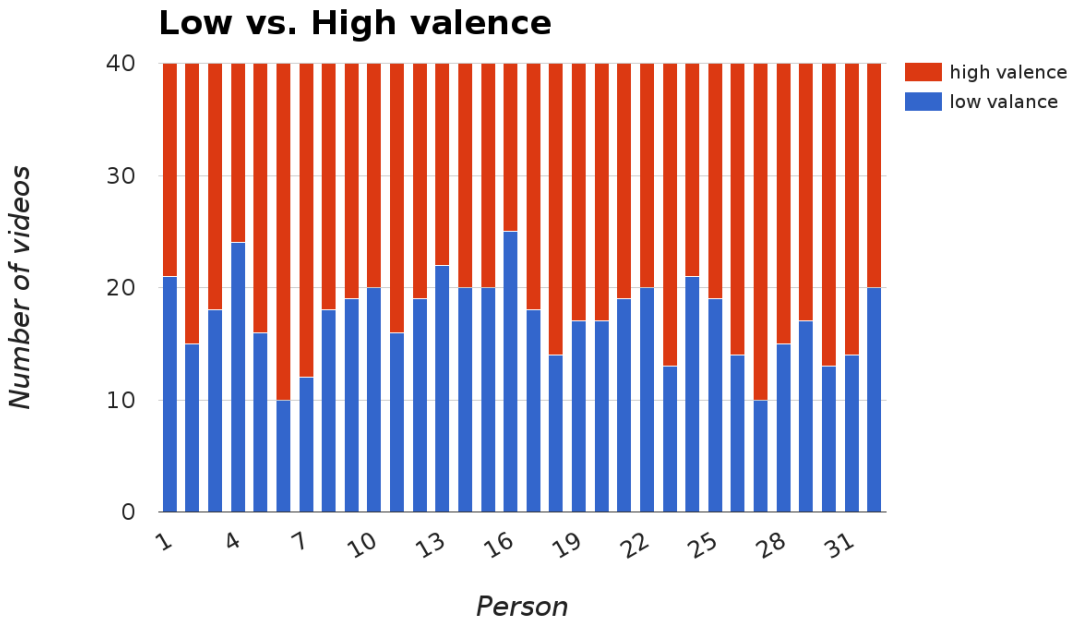

Figure 3.2: This graph show the assignment of the data set in to two classes: high and low valence. Each bar on the X axis represents a person, while the Y axis represents the number of low and high valence values.

The unbalance and especially the difference between unbalances for different persons is actually quite remarkable, given that each persons watched the same set of videos. Even taking into account personal differences, the different is high. One explanation for this might be that each persons rated movies differently; some persons are prone to giving higher values than other which results in an higher average valence value for some persons. To solve this problem one could simply order all the rating in ascending or descending order and assign the lower and higher half to the lower and higher valence class respectively. Using the median of the ratings as a second assignment method will do just that and as a result the classes will always be balanced. Both assignment methods are visually compared in Figure 3.3 and Table 3.2 below.

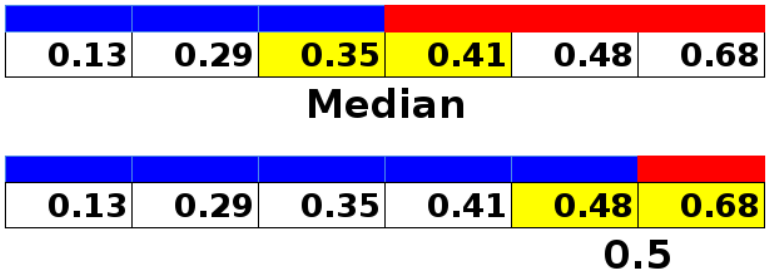

Figure 3.3: Assignments of different values to the two classes using two methods.

The results of different runs are obtained with leave-one-out validation where each of the 40 samples is predicted once, using the remaining 39 samples as training data. The 40 predictions' accuracies are then averaged which gave the following results:

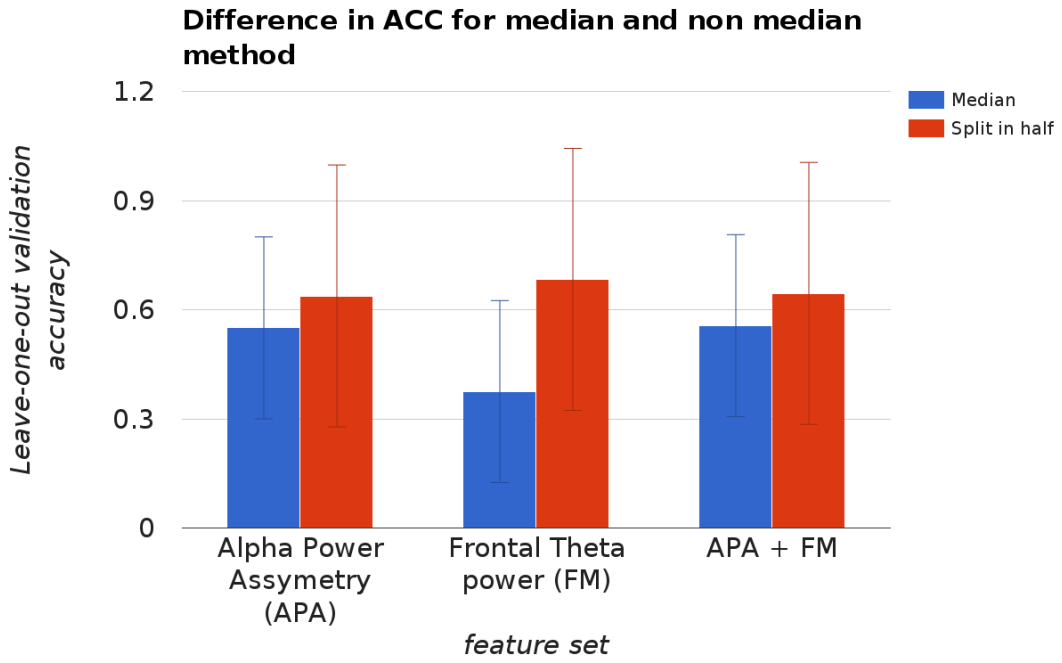

Figure 3.4: A comparison of the first results.

Features	Median Method	split in half Method
Alpha	0.55 ± 0.26	0.64 ± 0.37
FM	0.38 ± 0.25	0.68 ± 0.41
Alpha + FM	0.56 ± 0.28	0.64 ± 0.35

Table 3.2: Different accuracy values for the features sets and assignment methods (avg accuracy $\pm$ standard deviation).

Looking at the results it becomes clear that the average accuracy of the split in half method is higher than the median method. This is a result of assigning the classes with the median. For example with persons 6, some of the examples might end up in the low valence class even though person 6 might be feeling quite pleasant feelings during these EEG recordings, which confuses the model. Note that the standard deviation for the median method is lower than the split in half method, so in this perspective it is still possible that the results for the median method might be better.

3.3 CSP + LDA

Overall the results of the SVM explained above are reasonable for a first model, but higher accuracy is desired. Even though SVMs are capable of dealing with small datasets, it does not always provide the best results[29]. A frequently used model in the context of Brain Computer Interfaces is a combination of Common Spatial Patterns (CSP) and Linear Discriminant analysis (LDA) [19, 29, 30].

CSP is a spatial filtering technique that was originally designed to optimize the performance of motor imagery BCIs. CSP does so by creating linear combinations of the original EEG channels that maximize the variance for one class while simultaneously minimizing the variance of the other class [19]. One disadvantage of using CSP is that the default version can only distinguish between 2 classes, though one can easily aggregate multiple CSP models to create one-vs-one and one-vs-all models, similarly to the one-vs-one and one-vs-all SVMs.

LDA, on the other hand tries to find a projection of the data where the data is linearly separable.

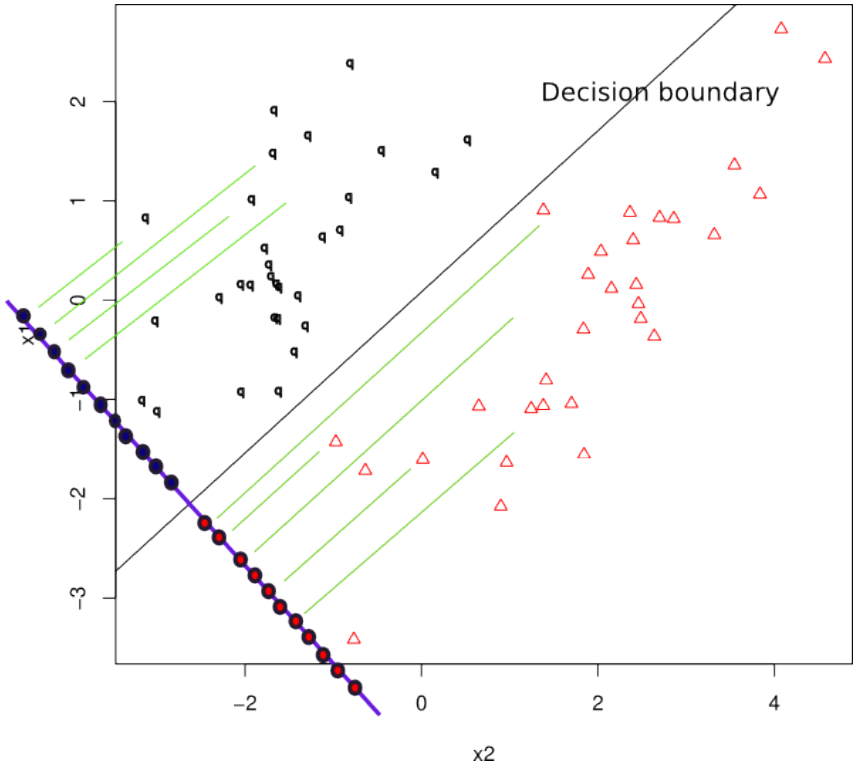

Figure 3.5: LDA finds a projection of the data where the separation of the data is clear.

To evaluate the model the dataset is first split in a train and testset (30/10), the testset is only used in the end to evaluate how well the model generalizes. During loading, the 32 original EEG channels are transformed to 16 CSP channels pairs, then the optimal number of channels pairs is determined used leave-one-out validation. This procedure thus starts with the outer channel pairs, performs leave one out validation and then uses the two outer channel pairs, etc. The set of channelpairs corresponding to the highest achieved validation accuracy is then used to predict the testset. Those results are shown in Figure 3.6.

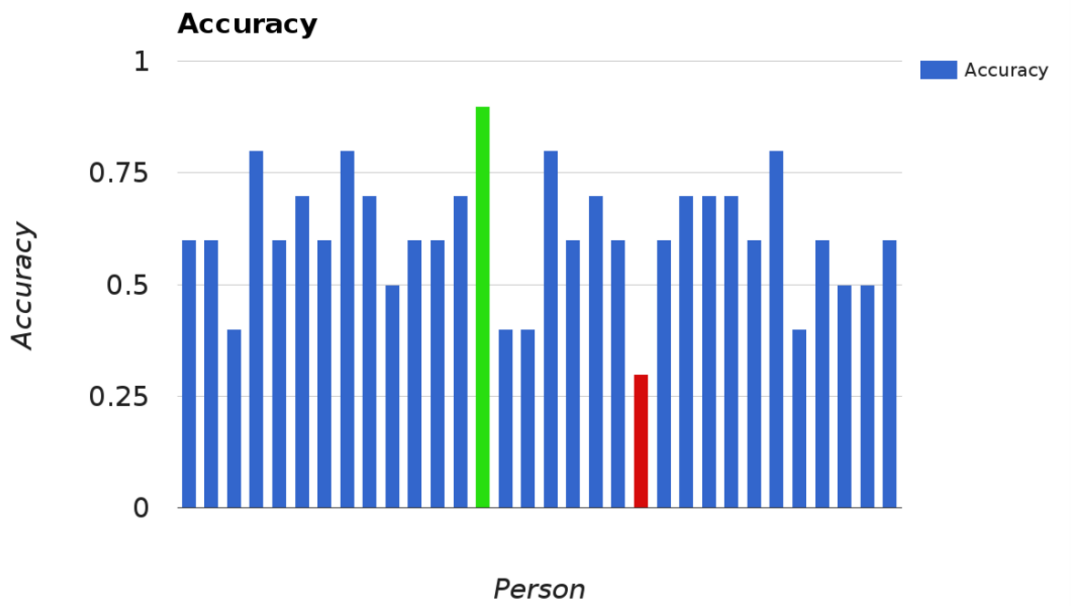

Figure 3.6: Accuracies for different persons, the highest (green) and lowest (red) accuracy are obtained by person 14 (90%) and 21 (30%) respectively.

From the results it is clear that this model actually works quite well for some persons while achieving lower accuracies for other persons. This variation in performance is typical for EEG and BCI related problems as EEG data is very person specific. Additionally one could plot the distribution of the original valence values, ranging between 0 and 1. Remember that these values were mapped to a label, where all valence values > 1 correspond to the high valence class while all valence values < 1 correspond to the lower valence class. Dividing these value in 8 bins and plot how many samples are present of each samples shows that for person 14 (highest accuracy) the greatest amount of samples is either very low or very high, and not a lot of mass is present near the classification boundary. For person 21 however, most of the samples are either just smaller of just greater than the classification boundary. With the absence of more extreme samples, it becomes hard for the classifier to achieve good performance.

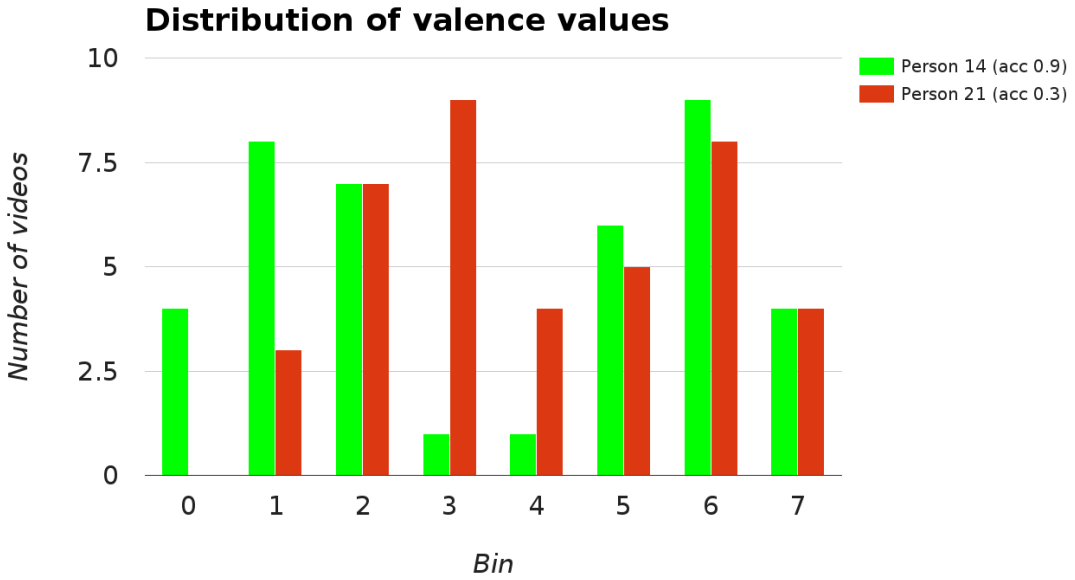

Figure 3.7: Distribution of valence values for the lowest and highest achieved test scores of the model.

Since CSP is applied it is also possible to look at the generated filter pairs to improve the understanding of the CSP algorithm. Figure 3.8 shows generated plots of the CSP values. The black dots represent given weights at the electrode locations the plot is then generated by interpolating the absolute weights of these points for the whole image. Blue colors correspond too less importance and red colors are the most important. Looking at the generated plots it become quite clear that the generated CSP filter is very person specific, CSP filter pairs of different persons look quite different.

Generally speaking, you would expect something similar to person 2 where the algorithm measures power components in the right and left parts of the scalp. This is somewhat the case for person 21, where the CSP gives high weights to different electrode locations. For person 14 however, the result is somewhat less desirable. Even though person 14 had the highest accuracy of 90%, the entire classification is based on a single electrode, which is an indication that the CSP is overfitting. This means that the performance might be great for the given test set, but its performance is likely to diminish quickly if one were to redo the experiment with the same classifier and a different set of samples.

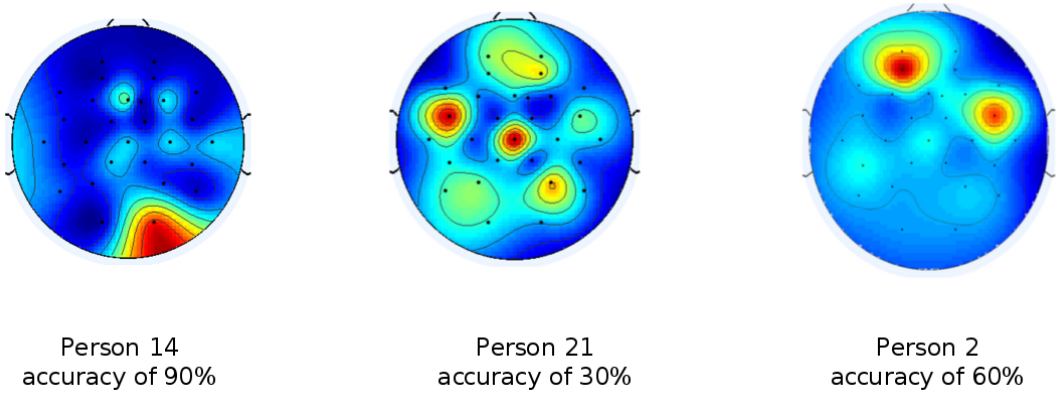

Figure 3.8: CSP filters for different persons

4

Feature Selection methods

This section will first explain what feature selection means and why it is used, before giving a detailed overview of the different methods.

4.1 The need for feature selection

The need for feature selection is twofold: first by reducing the number of features, you can improve the accuracy, certainly when the training dataset is limited and there is a severe risk of overfitting. Second, reducing the number of features can increase performance. Additionally, in the context of research and this thesis, looking at which features are importance gives insight in the problem. For thesis it is important to find out which features are useful in the prediction of emotion.

4.2 Different methods

There are several methods for the feature selection available, the following section will give a list of the most common used methods and their advantages / disadvantages.

4.2.1 Filter methods

These methods simple use a statistical test and filter out features that fail these test.

Removing features with low variance

This method simple removes all feature whose variance doesn't meet some threshold. Feature with have a low variance are less likely to be relevant, as there is very little variation in their value.

Univariate Feature Selection

The basic idea is to perform a statistical test to each feature individually. This is very simple, but will fail to recognise 'pairs' of important features; one feature may not be important on

label	feature A	feature B
Happy	+	+
Happy	-	-
Sad	-	+
Sad	+	-

Table 4.1: Some feature are not significant on its own, but a might be part of a combination of features.

its own, but might be a very good feature when combined with other features. Suppose the following example in Table **??**:

It is clear that feature A and B are very important when it comes to predicting whether or not a person is happy or sad. When both features have the same sign, the person is happy, otherwise he is not. This problem occurs in many simple selection methods.

4.2.2 Wrapper based methods

These methods use a training algorithm that guides them at finding the optimal subset of features. This methods is more recommended.

Recursive feature elimination

In this method you start by training a classifier with all the features and measuring the performance on a separate validation set. Once trained you look at the different coefficients that are assigned to each feature. Features with low coefficients are less important than features with high coefficient, so they can be removed. This process is repeated multiple times and every time a certain percentage of the lowest features is dropped until the validation score decreases significantly. This is done in L1 regression, you also have a less aggressive method , the L2 regression that lowers penalises the influence of features, but doesn't remove them, which strengthens its resistance against highly correlated features.

Randomized Sparse models

The main limitation of the above methods is that they will only select one feature out of a group of correlated features. A solution to solve this problem is to use randomization techniques, which randomly select features and look at the performance. combinations of features that give a high performance are then selected as being relevant. However this only works when the ground solution is sparse, meaning that only a small fraction of the features is relevant.

A sub group of randomized sparse models are the tree based estimators. For example, in a random forest features are randomly combined multiple times. The result is that each feature has an importance value, that indicates how important a certain feature is. Using the importance values, it is easy to remove certain features.

Coordinate Research

On coordinate descent the feature selection is formulated as a optimization problem and gradient ascent/descent is applied.

Bibliography

[1] H. Verschore, "A brain-computer interface combined with a language model: the requirements and benefits of a p300 speller," afstudeerwerk, Ghent University, June 2012.

[2] D. O. Bos, "Eeg-based emotion recognition," 2007.

[3] M.-K. Kim, M. Kim, E. Oh, and S.-P. Kim, "A review on the computational methods for emotional state estimation from the human eeg," *Computational and Mathematical Methods in Medicine*, vol. 2013, no. 573734, p. 13, 2013.

[4] T. C. Technologies, *10/20 System Positioning manual*. Fortis Tower, 2012.

[5] unknown, "Electrode placement," 2015.

[6] Y. Yang, S. Chevallier, J. Wiart, and I. Bloch, "Time-frequency optimization for discrimination between imagination of right and left hand movements based on two bipolar electroencephalography channels," *EURASIP journal on Advances in Signal Processing*, vol. 2014, no. 38, 2014.

[7] K.-E. Ko, Hyun-Chang, and K.-B. Sim, "Emotion recognition using eeg signals with relative power values and bayesian network," *International Journal of Control, Automation, and Systems*, 2009.

[8] Brainworks, "What are brainwaves?," 2015.

[9] N. K. Squires, K. C. Squires, Steven, and A. Hillyard, "Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man," *Electroencephalography and Clinical Neurophysiology*, pp. 387–401, 1975.

[10] Wikipedia, "Event-related potentials," 2015.

[11] L. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalography and clinical Neurophysiology*, vol. 70, no. 70, pp. 510–523, 1988.

[12] T. Verhoeven, "Brain-computer interfaces with machine learning: an improved paradigm for the p300 speller," afstudeerwerk, Ghent University, June 2013.

[13] N. V. Manyakov, N. Chumerin, A. Combaz, and M. M. V. Hulle, "Comparison of classification methods for p300 brain computer interface on diabled subjects," *Computational intelligence and neuroscience*, 2011.

[14] Cognionics, "Cognionics dry eeg p300 speller demo," 2015.

[15] D. J. Krusienski, E. W. Sellers, F. Cabestaing, S. Bayoudh, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "A comparison of classification techniques for the p300 speller," *Journal of Neural Engineering*, vol. 3, no. 4, p. 299, 2006.

[16] A.-M. Brouwer and J. B. van Erp, "A tactile p300 brain-computer interface," *Frontiers in Neuroscience*, vol. 4, May 2010. doi:10.3389/fnins.2010.00019.

[17] J. Höhne, M. Schreuder, B. Blankertz, and M. Tangermann, "Tow-dimensional auditory p300 speller with predictive text system," *IEEE EMBS*, 2010.

[18] F. Akram, H.-S. Han, H. J. Jeon, K. Park, S.-H. Park, J. Cho, and T.-S. Kim, "An efficient words typing p300-bci system using a modified t9 interface and random forest classifier," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pp. 2251–2254, July 2013.

[19] A. Coone, "A study on different preprocessing and machine learning techniques for the detection of error-potentials in brain-computer interfaces," afstudeerwerk, Ghent university, June 2011.

[20] R. Chavarriaga, A. Sobolewski, and J. d. R. Millán, "Errare machinale est: The use of error-related potentials in brain-machine interfaces," *Frontiers in Neuroscience*, vol. 8, no. 208, 2014.

[21] A. B. model for exploiting application constraints to enable unsupervised training of a P300-based BCI, "Pieter-jan kindermans and david verstraeten and benjamin schrauwen," *Plos ONE*, vol. 7, April 2012.

[22] Y. Lio and O. Sourina, "Eeg databases for emotion recognition," *International Conference on Cyberworlds*, 2013.

[23] Y. Lio, O. Sourina, and M. K. Nguyen, "Real-time eeg based human emotion recognition and visualization," 2010.

[24] D. Sammler, M. Grigutsch, T. Fritz, and S. Koelsch, "Music and emotion: Electrophysiological correlates of the processing of pleasant and unpleasant music," *Psychophysiology*, vol. 44, pp. 209–304, 2007. DOI: 10.1111/j.1469-8986.2007.00497.x.

[25] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals," *Affective Computing, IEEE Transactions on*, vol. 3, pp. 18–31, Jan 2012.

[26] M. Technologies, "Emotions by mensia," 2015.

[27] R. W. Picard and J. Klein, "Computers that recognise and respond to user emotion: theoretical and practical applications," vol. Interacting with computers, no. 14, pp. 141–169, 2002.

[28] Y. Morita, K. Morita, M. Yamamoto, Y. Waseda, and H. Maeda, "Effects of facial affect recognition on the auditory {P300} in healthy subjects," *Neuroscience Research*, vol. 41, no. 1, pp. 89 – 95, 2001.

[29] F. Lee, R. Scherer, R. Leeb, C. Neuper, H. Bischof, and G. Pfurtscheller, "A comparative analysis of multi-class eeg classification for brain computer interface," in *Proceedings of the 10th Computer Vision Winter Workshop*, pp. 195–204, 2005.

[30] G. Pfurtscheller, C. Neuper, C. Guger, W. Harkam, H. Ramoser, A. Schlögl, B. Obermaier, and M. Pregenzer, "Current trends in graz brain-computer interface (bci) research," *IEEE Transactions on rehabilitation Engineering*, vol. 8, pp. 216–219, JUNE 2000.