Detecting the no-control state in self-paced Brain-Computer Interfaces

Mircea Stoica

Enschede, December 19, 2012

Master's Thesis

Human-Media Interaction

Faculty of Electrical Engineering, Mathematics and Computer Science

University of Twente

Graduation committee:

Dr. Mannes Poel (1st supervisor)

Dr. Hayrettin Gürkök

Dr. Boris Reuderink

Bram van de Laar, MSc

Abstract

The field of brain-computer interfaces (BCI) has skyrocketed in recent years, and the advent of increasingly powerful technology is rapidly bringing it to commercial use. Crucial to its adoption for many real-world applications is the ability to respond only to intentional commands of the user. This is rather difficult because of the constant and seemingly random activity of the brain.

Research in BCIs based on voluntary changes of brain activity has traditionally focused on distinguishing between two or more different and reproducible patterns, for the goal of communication. Additionally distinguishing them from all other possible brain activity is a considerable challenge for the signal processing and machine learning methods commonly used.

This is an exploratory study into the behavior and characteristics of different approaches to self-paced motor imagery BCIs. Subject-specific band-power features are extracted and different classifiers are applied on a publicly available dataset, consisting of four subjects performing two types of motor imagery with no cues. The usefulness of dwell times and refractory periods is also studied, and we investigate the relation between different performance metrics.

Our results suggest that linear discriminant classification between the three states is the most suitable approach, given the spatial filtering methods presently available. We find that refractory periods do not generalize well to new data but that the dwell time is highly recommended. The results show that the current event-based definition of true positive rate has an inverse relation to mutual information, which may be more related to how much a subject can maintain a desired state.

Table of Contents

List of figures **1**
List of tables **4**
List of acronyms **5**
Introduction 6
Motivation and objectives 6
Report outline 7
Chapter I Introduction to BCI **8**
1. Objective of BCI research 8
2. Structure of a BCI system 9
3. Recording techniques 10
3.1. Invasive methods 10
3.1.1. Intracortical electrodes 10
3.1.2. Electrocorticography (ECoG) 12
3.2. Non-invasive methods 13
3.2.1. Electroencephalography (EEG) 13
3.2.2. Magnetoencephalography (MEG) 14
3.2.3. Functional magnetic resonance imaging (fMRI) 15
3.2.4. Functional near-infrared spectroscopy (fNIRS) 15
3.3. Conclusion 16
4. Neurological phenomena 16
4.1. Sensorimotor rhythms 16
4.2. Evoked potentials 19
4.3. Event-related potentials 19
4.4. Slow cortical potentials 19
4.5. Conclusion 20
Chapter II The self-paced BCI **21**
1. The no control (NC) state 21
2. BCI control paradigms 22
2.1. Timing mechanisms 22
2.2. Continuous and discrete outputs 23
2.3. Event-driven and state-driven outputs 23
2.4. Conclusion 24
3. Performance metrics for self-paced BCIs 25
3.1. Sample-by-sample evaluation 25
3.2. Event-by-event evaluation 27
4. Previous work 28
Chapter III Materials and methods **31**
1. Experimental data 31
2. Methods 32
2.1. Overview 32
2.2. Common spatial patterns (CSP) 33
2.3. Feature extraction and processing 35
2.4. Feature selection 36

2.5. Classification 38
2.5.1. Classification schemes 38
2.5.2. Classification algorithms 39
2.6. Regression 42
2.7. Genetic algorithms 44
3. Experimental procedure 44
Chapter IV Experiments and results 45
1. Feature selection 45
1.1. NC/IC discrimination 45
1.2. IC discrimination 48
2. Classification 51
2.1. Cross-validation 51
2.1.1. Two-stage classification 52
2.1.2. Separate detection of each IC state 53
2.1.3. Three-state classification 55
2.1.4. Conclusion 56
2.1.5. Influence of the feature processing pipeline 57
2.1.6. Sensitivity VS hold time 59
2.1.7. Exploring additional false positive rates 61
2.1.8. Dwell and refractory periods 62
2.1.9. Selection of relevant intervals 64
2.2. Evaluation 66
3. Regression 70
3.1. Cross-validation 70
3.1.1. Influence of the feature processing pipeline 70
3.1.2. Assigning different regression targets 71
3.1.3. Influence of tap delays 73
3.1.4. Larger FP rates and refractory periods 75
3.2. Evaluation 76
4. Combining multiple outputs 80
Chapter V Discussion 82
1. Classifier designs 82
2. Parameter tuning and test results 85
3. Performance metrics and target applications 87
4. Limitations of the study 89
Chapter VI Conclusions and future prospects 90
1. Translation algorithms 90
2. Dwell time and refractory period 91
3. Performance metrics 92
4. Future work 92
4.1. Supervised spatial filtering for self-paced BCI 92
4.2. Online study 93
4.3. Performance metrics 93
Bibliography 94

List of figures

Figure 1: Functional model of a BCI system 9
Figure 2: Sequential action potentials form spike trains 11
Figure 3: a) The BrainGate sensor resting on a US penny and the percutaneous pedestal which connects the sensor to the rest of the system. b) Close-up of the 10 × 10 microelectrode array. c) Location of the sensor. d) The first participant in the BrainGate trials. Picture taken from the original article published in Nature [11] 12
Figure 4: Typical EEG power density spectrum (left) and close-up of the low-frequency range (right) depicting the 1/f nature of EEG spectra. Notice the 50 Hz power line noise and its harmonics (left) and the peaks close to 10 and 20 Hz (right) corresponding to alpha and beta rhythms, respectively 14
Figure 5: Participant in an MEG study at the National Institute of Mental Health, USA 15
Figure 6: Typical brainwaves and their associated frequency bands. Notice the inverse relation between amplitude and frequency 17
Figure 7: Homunculus model of the sensorimotor cortex, showing the layout of different body parts 18
Figure 8: a) Event-driven discrete control, based on transient neurological phenomena, such as movement-related potentials or P300; the return to the NC state is automatically done by the brain. b) State-driven discrete control: the user has the ability to initiate, maintain and release an intentional control state. Adapted from the technical report on self-paced BCI by Mason et al [8] 24
Figure 9: Event-driven BCIs allow only transitions between each IC state and NC, while state-driven BCIs allow all possible transitions 24
Figure 10: Depending on the distribution of IC and NC states in feature space, a single linear classifier might not be able to separate the two IC states from NC, if both IC states are considered one class 29
Figure 11: Electrode layout for dataset 1 of BCI Competition IV 32
Figure 12: Block diagram of the BCI architecture 33
Figure 13: The CSP algorithm finds a set of virtual channels which maximize variance differences between two classes 34
Figure 14: Processing steps of our feature vectors 35
Figure 15: Two-stage detection and classification of IC states 38
Figure 16: Separate detection of each IC state 39
Figure 17: Direct three-state classification 39
Figure 18: SVMs determine the hyperplane which maximizes the distance to the nearest training points 42
Figure 19: Finite impulse response topology of order M 43
Figure 20: The patterns of the three most relevant spatial filters for IC/NC discrimination extracted for each subject and the corresponding frequency bands 46
Figure 21: Mutual information between extracted features and the desired output for IC/NC discrimination with respect to the number of features 47
Figure 22: Mutual information between extracted features and desired output for IC/NC discrimination with respect to the window position within the trial 48
Figure 23: The patterns of the seven most relevant spatial filters extracted for IC discrimination and the corresponding frequency bands 49

Figure 24: Mutual information between extracted features and the desired output for IC discrimination with respect to the number of features 50

Figure 25: Mutual information between extracted features and desired output for IC discrimination with respect to the window position within the trial 51

Figure 26: LDA, QDA and one-class SVM detection rates in differential mode. Error bars represent the percentage of misclassified, yet detected IC states. TP_i represents the true positive rate of IC state *i*, TP is the average true positive rate of IC detection. All TP rates correspond to an FP rate of 1% or lower. 53

Figure 27: LDA, QDA and one-class SVM detection rates in parallel mode. Error bars represent the percentage of misclassified, yet detected IC states. TP_i represents the true positive rate of IC state *i*, TP is the average true positive rate of IC detection. All TP rates correspond to an FP rate of 1% or lower. 54

Figure 28: LDA and QDA detection rates for direct 3-state classification. Error bars represent the percentage of misclassified, yet detected IC states. TP_i represents the true positive rate of IC state *i*, TP is the average true positive rate of IC detection. All TP rates correspond to an FP rate of 1% or lower. 55

Figure 29: Average IC detection rates for all subjects and classifier designs, and their standard deviation between subjects. Subscripts 1C, D, P and 3C represent one class (IC/NC), differential, parallel and direct 3-state classification, respectively 56

Figure 30: Influence of different parameters on detection rate. First column: classifier trained with average sample values, log transformed; second column: classifier trained with all samples from the window, log transformed; third column: all samples, no log. A window size of one sample is equivalent to not performing the respective operation. All TP rates correspond to an FP rate of 1% or lower. 58

Figure 31: Influence of smoothing operations on hold time. 60

Figure 32: TP rates for FP rates ranging between 1% and 10% 61

Figure 33: Relative improvement in TPR due to the use of refractory periods at two FPR values 63

Figure 34: Influence of the selection intervals on TP rate. Four sizes and four offsets are tested, of 0.5, 1, 1.5 and 2 seconds. For each combination, the best dwell and refractory periods are found. On the axis they are grouped with respect to the size, hence there are four groups [($size_1$,$offset_1$), ($size_1$,$offset_2$)…], [($size_2$,$offset_1$),($size_2$,$offset_2$)…] etc. The ticks represent the size of the selection window, thus the tick following that at 1.5s corresponds to a size of 1.5s and an offset of 1s, thus a 1.5s window centered at 1.75s 65

Figure 35: Comparison between three approaches in cross-validation (CV) and evaluation (Eval). Values are averaged over all subjects and shown with the standard deviation. The default configuration uses the one second selection interval found in feature selection and no refractory period; +refractory adds the use of refractory period; +optimized selection uses both optimal training intervals and refractory periods. 66

Figure 36: Event analysis for LDA on the evaluation data, with TP rates as a function of event duration. Refractory periods and the optimal training intervals are used. The TPR above each plot is the TP rate of the corresponding IC state, averaged over all event durations. FPR is the sample FP rate, FPR_E is the event FP rate 68

Figure 37: Inter-FA periods distribution for LDA. The minimum and maximum time between false activations is given above each plot. 69

Figure 38: Influence of the feature processing pipeline in terms of detection rate, mean squared error and mutual information, averaged over all subjects. 70

Figure 39: Detection rates for the Wiener filter (left) and comparison to LDA (right). Error bars represent the percentage of misclassified, yet detected IC states. TP_i represents the true positive rate of IC state *i*, TP is the average true positive rate of IC detection. All TP rates correspond to an FP rate of 1% or lower. 71

Figure 40: Influence of tap delays on TP rate, event FP rate and response time. One tap delay is equivalent of using 100 ms of past data. The shaded area in the response time plots represents the standard deviation..... 74

Figure 41: TP rates of the Wiener filter for FP rates ranging between 1% and 10% 75

Figure 42: Regression results on the evaluation data, with the original targets (Orig), with the ones determined by the genetic algorithm (+GA), and the additional use of refractory periods (+RP). 77

Figure 43: Event analysis for the Wiener filter on the evaluation data, with TP rates as a function of event duration. The TPR above each plot is the TP rate of the corresponding IC state, averaged over all event durations. FPR is the sample FP rate, FPR_E is the event FP rate..... 78

Figure 44: Inter-FA periods distribution for the Wiener filter. The minimum and maximum time between false activations is given above each plot. 79

Figure 45: True positive rates and event false positive rates of linearly combining classifier outputs with a genetic algorithm. Results are averaged over all subjects and shown with the standard deviation. On the *x* axis, LDA represents the initial results of 3-state classification. The other four labels indicate the optimization criterion used in the genetic algorithm: average true positive rate (TPR), difference between TPR and event false positive rate (FPE), true-false difference (TF) and mutual information (MI). 80

Figure 46: If classes -1 and 1 need to be distinguished from class 0 with a single threshold (linear classifier), it is easier when the separation between them is poor. 83

List of tables

Table 1: The number of spatial filters selected for each subject, class and frequency band for NC/IC discrimination..... 45

Table 2: The number of spatial filters selected for each subject, class and frequency band for IC discrimination..... 48

Table 3: TP rates of IC detection for different classifiers at 1% FP rate. Both IC states are treated as one class 52

Table 4: Influence of dwell time and refractory period on TP rates for a maximum FP rate of 1%. The left part of the table shows the original TP rates obtained with no dwell or refractory post-processing. The right part of the table shows the TP rates obtained with the optimal dwell and refractory periods. FPR_{Orig} is the initial FP rate obtained at the corresponding TP rates, FPR_{Optim} is the FP rate after post-processing. Numerical subscripts of TP rates indicate the IC state, Avg indicates their average. 62

Table 5: Influence of dwell time and refractory period on TP rates for a maximum FP rate of 1.5%. For a detailed description refer to the caption of Table 4. 62

Table 6: Average true positive rates with the original labels and with the ones determined by LDA. The target values are expressed in terms of their mean and standard deviation 72

Table 7: Average true positive rates with the original labels and with the ones determined by the genetic algorithm 72

Table 8: Influence of dwell time and refractory period on TP rates of the Wiener filter for a maximum FP rate of 1%. The left part of the table shows the original TP rates obtained with no dwell or refractory post-processing. The right part of the table shows the TP rates obtained with the optimal dwell and refractory periods. FPR_{Orig} is the initial FP rate obtained at the corresponding TP rates, FPR_{Optim} is the FP rate after post-processing. Numerical subscripts of TP rates indicate the IC state, Avg indicates their average. 76

Table 9: Ratios of sample and event FP rates in cross-validation (CV) and on the test set, respectively. *RP* stands for refractory period; FPR subscripts *pre* and *post* (corresponding to subscripts *orig* and *optim* in Table 8) indicate the FP rates before and after applying RP; subscripts *w* and *w/o* indicate the event FP rates with and without RP, respectively. The rightmost column is the ratio of the two anterior columns. 77

List of acronyms

1-NN	1-Nearest Neighbour
ALS	Amyotrophic Lateral Sclerosis
AUC	Area Under (ROC) Curve
BCI	Brain-Computer Interface
BMI	Brain-Machine Interface
BOLD	Blood Oxygen Level Dependent
CNS	Central Nervous System
CSF	Cerebrospinal Fluid
CSP	Common Spatial Patterns
DBI	Direct-Brain Interface
DSLVQ	Distinction Sensitive Learning Vector Quantization
DWT	Discrete Wavelet Transform
ECoG	Electrocorticography
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
ERD	Event-Related Desynchronization
ERP	Event-Related Potential
ERS	Event-Related Synchronization
FLD	Fisher Linear Discriminant
FIR	Finite Impulse Response
FPR	False Positive Rate
GMM	Gaussian Mixture Model
IC	Intentional Control
ITR	Information Transfer Rate
LDA	Linear Discriminant Analysis
LF-ASD	Low-Frequency Asynchronous Switch Design
LFP	Local Field Potential
ME	Motor Execution
MEG	Magnetoencephalography
MI	Motor Imagery
MRP	Movement-Related Potential
MSE	Mean Squared Error
NC	No-Control
PSD	Power Spectral Density
QDA	Quadratic Discriminant Analysis
ROC	Receiver Operating Characteristics
SCP	Slow Cortical Potential
SMR	Sensorimotor Rhythm
SNR	Signal-to-Noise Ratio
SQUID	Superconducting Quantum Interference Device
SSVEP	Steady-State Visually Evoked Potential
SVM	Support Vector Machine
TF	True-False Difference
TPR	True Positive Rate
fMRI	functional Magnetic Resonance Imaging
fNIRS	functional Near-Infrared Spectroscopy

Introduction

Humanity has always been fascinated by the possibility of interacting with the outside world only through the power of the mind. Throughout history, the apparently unlimited potential of the human brain has often led philosophers and scientists on a path of trying to prove that there is more to the mind than what we know. A fine example are the pioneering studies of Hans Berger in the late 20's, which have led to the development of electroencephalography (EEG) and ultimately paved the way for modern neuroscience, and were motivated by his quest to find proof of telepathic abilities in humans **[1]**.

Nowadays, the blazing advances in diverse fields such as medicine, physics and engineering have allowed scientists to gain a much deeper understanding of the brain's inner workings, and the exponential progress of technology promises to bring such developments to the masses, ultimately for the goal of an increased quality of life. A promising and relatively new direction is the field of brain-computer interfaces (BCI), which seeks to give people the capability of controlling various devices through the electrical activity of the brain.

As BCI technology steadily moves out of the lab and into hospitals, homes, and even military use, one pressing issue still remains. Due to the high complexity and apparent randomness of our brain, it is not very clear how to distinguish intentional commands from the seemingly stochastic ongoing cerebral activity. The problem of self-paced operation for BCI attracted the interest of many researchers in the past decade, but a widely accepted solution has yet to be found.

Motivation and objectives

Throughout the literature, one can find several studies related to self-paced BCI, each with different subjects, methods and experimental protocols. Deciding on a specific methodology is thus rather difficult, as the results of these studies cannot be directly compared because of the large variability between them. Differences in experimental protocol, subjects' experience with BCI and reported performance metrics all hinder the process of making educated decisions on the many possibilities in designing self-paced BCIs.

What therefore seemed to be lacking was a comparative review of the different approaches to the problem, and this is exactly the gap that this thesis attempts to fill. By no means however do we claim that it is a complete overview of all possible methods and their combinations, simply because of the huge number of possibilities.

Readers familiar with the field know that there are many types of BCI, based on different neurophysiological phenomena. Throughout the first two chapters we motivate our choice of using modulations of sensorimotor rhythms as the source of control, by showing that they theoretically offer the richest possibilities for self-paced BCIs.

The first research question we address concerns the choice of an appropriate translation algorithm. Not only different machine learning methods, but also different implementation schemes are possible for classification in a self-paced context. One could opt for a two-stage design, in which the first classifier would detect intentional control (IC) commands of any kind, and the second would assign them to one of the possible classes. This seems to be the most popular choice in self-paced BCI research, but it is not the only one. A single classifier could be trained to directly distinguish between no control (NC) and the possible IC states. A third option is to train one classifier for each IC state, and to the best of our knowledge this approach has not even been tested in previous research. We hypothesize that the common approach of considering multiple IC states as one class is highly inappropriate, mainly because the neurophysiological phenomena underlying different IC states are chosen to be as different as possible.

To decrease incorrect activations of self-paced BCIs, specific tools are used, such as the refractory period or the dwell time. The second research question is how large and reliable is the performance gain obtained with such methods. To this end, appropriate evaluation criteria need to be defined. A wide variety of performance metrics exist for the evaluation of self-paced BCIs, yet no consensus exists regarding which is most informative. It is possible that a truly informative description of performance could not even be obtained by using a single metric. The third research question aims to find which evaluation criteria are most important for self-paced BCIs.

In the hope of objective, reproducible and more meaningful results, the methods under analysis are applied on a publicly available dataset, specifically designed for the evaluation of self-paced BCIs.

Report outline

The thesis is structured into six chapters. The first chapter is aimed at readers unfamiliar with BCI technology and provides an introductory overview to the field, such as functional model, recording techniques and neurophysiological phenomena. Readers already acquainted with BCI can directly proceed to chapter two, where the issue of self-paced operation is introduced and discussed. Evaluating the performance of self-paced BCIs is a hot research topic in its own right, and commonly used performance metrics are also defined and discussed. The chapter concludes with an overview of previous research into self-paced operation for BCI. The third chapter first presents the experimental data used in the study and then proceeds to a detailed description of the algorithms and methods employed. The results of the different analyses and experiments are presented in chapter four. The fifth chapter interprets and discusses these results, and also presents the author's more general views on the study and the field of self-paced BCI. The thesis concludes with chapter six, which provides the conclusions drawn from the experimental results and also suggests future directions for research.

Chapter I
Introduction to BCI

1 Objective of BCI research

Brain-computer interfaces, also called brain-machine interfaces (BMI) or direct brain interfaces (DBI), are a direct communication pathway between the brain and a computer or some other external device. BCIs convert electrophysiological signals of the central nervous system (CNS) into meaningful messages and commands that act on the outside world and accomplish the user's intent, much in the same manner as conventional neuromuscular pathways. In that sense, a BCI replaces nerves and muscles with hardware and software that measure brain activity and translate it into actions **[2]**.

Assistive technology

Presently, the first and foremost objective of BCI research is restoring mobility and/or communication to those that have lost these abilities. Several neuromuscular disorders, such as amyotrophic lateral sclerosis (ALS), stroke, brain and spinal cord injuries, muscle dystrophies and many more affect millions of people worldwide and impair neural pathways or the muscles themselves. In extreme cases, patients might lose all muscle control, including eye movements and respiration, leaving them completely locked-in to their bodies with no ability of communication **[2]**. In such cases, restoration of even basic communication abilities would increase the patients' quality of life and independence, as well as reducing social isolation and the cost of care **[3]**. The most desired outcome of assistive BCIs would be the reanimation of a paralyzed limb **[4]** but short of that brain-controlled robotic prostheses are also highly desirable.

Augmentative technology

The new control possibilities offered by BCI could also be valuable for healthy users, enriching their experience in various games or applications. At the present moment however, BCIs do not bring any benefit to healthy users, except for the possible joy of using a novel technology and a more immersive experience in some contexts. Presently, it can be argued that the only BCIs that healthy users might consider are EEG-based, as they are non-invasive and relatively cheap. However, considering the modest information transfer rates (ITR) of current EEG-based BCIs[1] (25 – 35 bits per minute **[6]**), it is hard to see a reason why present-day healthy users would choose a slow, cumbersome and error-prone control input over the traditional mouse and keyboard input. This is not to say that we do not have faith in the future success of BCI; one must admit that the relative success of a technology partly depends on its availability to the masses. Even with the spectacular progress of BCI research in the past decades, it is no doubt that its most influential breakthroughs and applications are still to come with widespread adoption of the technology by healthy users.

[1] These values apply to motor imagery-based BCIs. SSVEP-based systems can achieve higher ITRs as they typically use a much larger number of classes (ITR of 75.4 bits/min was achieved with 16 targets **[5]**)

2 Structure of a BCI system

Throughout the literature there are several functional models of BCIs presented but considering the variability of different implementations they can be reduced to five basic components: signal acquisition, signal enhancement, feature extraction, the translation algorithm and feedback. A schematic model based on these components is presented in Figure 1. For a more detailed functional model of BCI the reader is referred to the work of Birch and Mason **[7, 8]**.

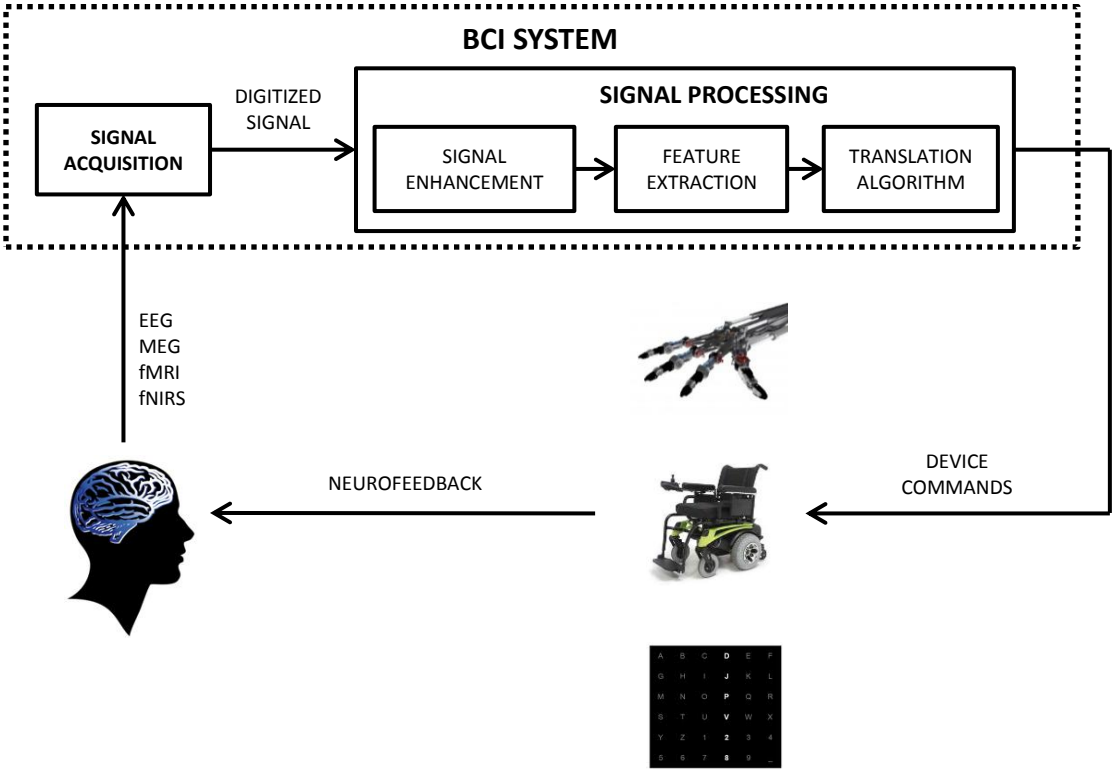

Figure 1: Functional model of a BCI system

Signal acquisition

Signal acquisition is the process of recording brain activity over time. Cortical activity can be inferred by measuring the electrical, magnetic or metabolic responses of the brain, through various techniques which will be discussed shortly.

Signal enhancement

Signal enhancement refers to techniques used to increase the signal-to-noise ratio (SNR) or reduce dimensionality. Undesired components of the signal, called artifacts, which are not a result of brain activity, are also removed in this step. These may include the 50/60 Hz power-line noise, various electromagnetic interferences and artifacts resulting from physical movements or hardware faults. Artifact detection and correction is a highly specialized and complex field in itself. It often requires additional sensors and measurements, such as recording eye movements through

electrooculography (EOG) or muscle activity through electromyography (EMG). Some applications skip this process altogether as possible artifacts might not influence the frequency bands of interest.

Feature extraction

The process of feature extraction transforms the measured cortical activity into meaningful and useful representations for predicting the intent of the user. In many cases feature extraction involves transforming the signal to a different domain and calculating relevant characteristics based on a priori knowledge of the neurological phenomena under analysis.

Translation algorithm

The translation algorithm of a BCI system converts the extracted features into a discrete or continuous control signal that is an estimation of the intent or mental state of the user (within the analyzed possibilities, of course – BCIs don't read minds). BCI spellers in which users select letters on the screen is an example of discrete output, while applications in neuroprosthetics require real-valued output for controlling each joint of the artificial limb.

Feedback

The final step is to provide the user with information on the BCI predictions. This usually consists of displaying the result on a computer screen or issuing the associated command to a device, such as an actuated wheelchair or prosthetic limb. Feedback is very important in real-world applications, as the brain is capable of reorganizing neural connections for the purpose of learning and adaptation. This phenomenon is known as neuroplasticity and, in essence, is no different than learning to walk or to speak at an early age.

3 Recording techniques

There are quite a few methods available for recording cortical activity, each with specific strengths and weaknesses. While the brain is essentially an electrical machine, neuroimaging techniques have been developed that are also sensitive to its magnetic and metabolic activity. Following is a brief description of each technique.

3.1 Invasive methods

As the name suggests, invasive neuroimaging techniques require surgery for the implant of electrodes which measure the electrical activity of single neurons or neural ensembles. Given the close proximity to the grey matter, these methods currently offer the best quality signals and consequently the most natural control possibilities of prosthetic limbs. They are divided in two categories, depending on whether electrodes are implanted inside brain tissue or on the surface of the cortex.

3.1.1 Intracortical electrodes

The most invasive type of neural implant consists of microelectrode arrays usually implanted directly into the motor cortex of paralyzed patients **[9]**. At the smallest level, even the

activity of a single neuron can be measured. The activation of a neuron produces a so called action potential, which is the most basic form of neural activity and manifests as a sharp increase in the electric potential of a neuron's membrane, typically from -70 mV to several tens of millivolts. Neurons can fire up to 100 times per second and the resulting sequence of voltage impulses is called a spike train (see Figure 2), which forms the basis of information processing within the brain and can be captured by intracortical electrodes.

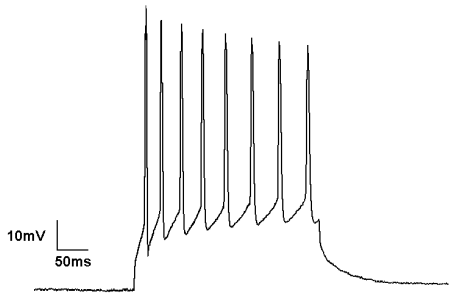

Figure 2: Sequential action potentials form spike trains

The firing patterns of multiple neurons can synchronize and the summation of their synaptic currents produces oscillations in the electric potential of local extracellular space. These oscillations are called local field potentials (LFP) and can also be captured by intracortical implants. These electrodes therefore capture either the firing rates of single neurons or local field potentials of several neurons and translate them into complex movements.

After extensive training of monkeys in a reaching and grasping task, it was shown that the firing patterns of only 32 neurons are sufficient to perform the same movement directly with a robotic limb **[10]**. In 2004, Donoghue's group presented a highly successful demonstration of an invasive BCI **[11]**, in which a tetraplegic subject could successfully check e-mails, control the TV, play a game and even draw a circle with non-muscular control of a computer cursor. Furthermore, the subject learnt to open and close a robotic hand in just a few trials while looking at the prosthetic hand and with no feedback from the cursor display, and was able to use a multi-jointed robotic limb to grasp an object from one location and transport it to the other. Figure 3 presents the BrainGate sensor developed by the biotech company Cyberkinetics for this pioneering study and its location within the brain of the tetraplegic patient.

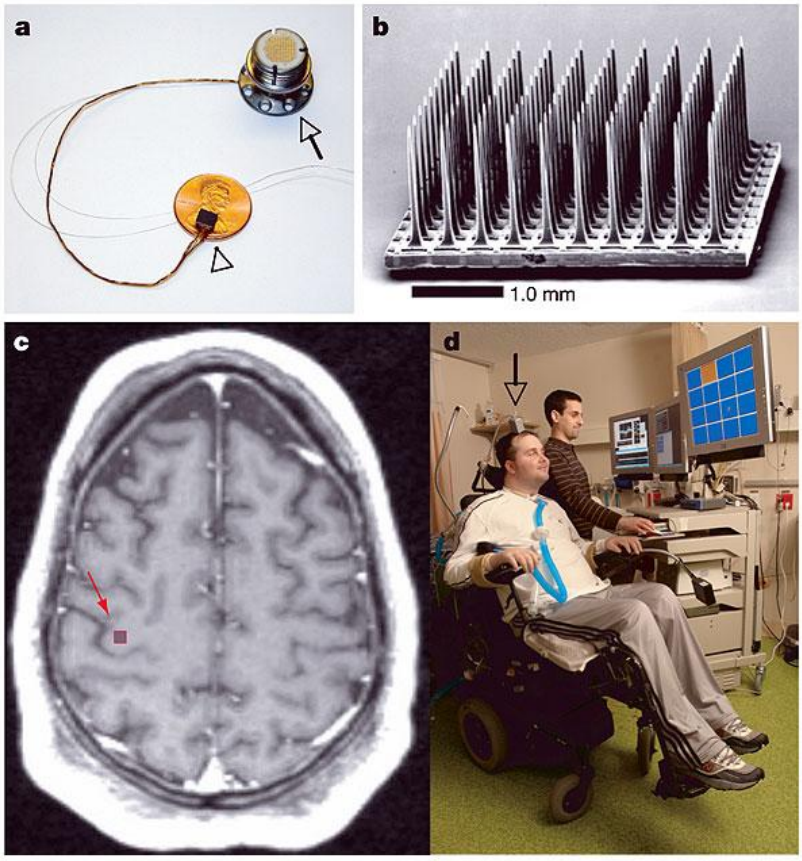

Figure 3: a) The BrainGate sensor resting on a US penny and the percutaneous pedestal which connects the sensor to the rest of the system. b) Close-up of the 10 × 10 microelectrode array. c) Location of the sensor. d) The first participant in the BrainGate trials. Picture taken from the original article published in Nature [11]

3.1.2 Electrocorticography (ECoG)

A less invasive option is electrocorticography, which consists of electrode arrays placed directly on the surface of the cortex. The signals thus recorded are a mixture of several LFPs, smeared and attenuated by the layers of brain tissue and cerebrospinal fluid (CSF). Because ECoG implants are only used by neurosurgeons for the purpose of clinical monitoring and localization of seizure foci in epileptic patients, the number of human subjects for ECoG research is rather limited **[12]**. Nevertheless, ECoG signals were successfully used for decoding two-dimensional hand movements with an accuracy comparable to what is achieved in monkeys with intracortical microelectrodes **[13]**. In an ECoG study on monkeys, Chao et al proved that ECoG signals can be used for decoding self-paced three-dimensional arm movements, again with comparable performance to that obtained by more invasive techniques. Another highly promising result of this study was that predictive performance remained stable across several months with no need for recalibration **[14]**, validating ECoG as a good candidate for long-term BCI usage.

3.2 Non-invasive methods

Invasive BCIs are clearly well suited for restoring mobility and communication to those in need, but they also pose the risks normally associated with brain surgery, such as infection and tissue damage **[15]**. Even so, it is conceivable that patients suffering from severe neuromuscular disorders such as amyotrophic lateral sclerosis, stroke or spinal cord injuries might choose the invasive option if there is no alternative. This would not be an option for healthy users however. Fortunately, the majority of BCI research focuses on non-invasive methods and continued technological advancements in both hardware and software hold promise for the future of such techniques.

3.2.1 Electroencephalography (EEG)

EEG is the oldest technique for measuring brain activity. It was developed by Hans Berger in 1929 **[1]** and almost 90 years later it continues to be a valuable tool in both clinical and research applications. EEG measures brain activity through sensitive electrodes placed on the scalp of the subject. The number of electrodes varies depending on the application and montages of up to 512 electrodes are used. Electrolytic gel is usually applied to form a conductive bridge between skin and electrode for decreasing impedance and the recorded voltages are amplified with a factor commonly ranging between 10^3 and 10^5. EEG amplitudes are quite weak, in the order of microvolts, and the recorded activity is a spatial integration of multiple local field potentials. Thus, the spatial resolution of EEG is quite poor (~ 1 cm **[16]**) but source localization methods can be used in multi-channel recordings for improving resolution. Furthermore, EEG is susceptible to movement artifacts, electromagnetic interference, muscle activity, and the signal is severely degraded because of the meninges, cerebrospinal fluid (CSF) and the skull.

One advantage of EEG is considered to be the high temporal resolution, as sampling rates of 4 KHz or higher are quite common in such devices. However, the usefulness of such a high sampling rate is limited due to the 1/f nature of EEG spectra (see Figure 4). As the amplitude of oscillations is proportional to the number of synchronously active neurons **[17]**, slowly oscillating cell assemblies comprise more neurons than fast oscillating ones **[18]**. The signal to noise ratio of EEG is rather poor to begin with, and the severe degradation of high frequencies drastically limits the usefulness of high sampling rates.

Despite all of these drawbacks, EEG continues to be the instrument of choice for the vast majority of BCI research, owing to its low costs, non-invasive nature and the continued interest in the development of increasingly sensitive equipment.

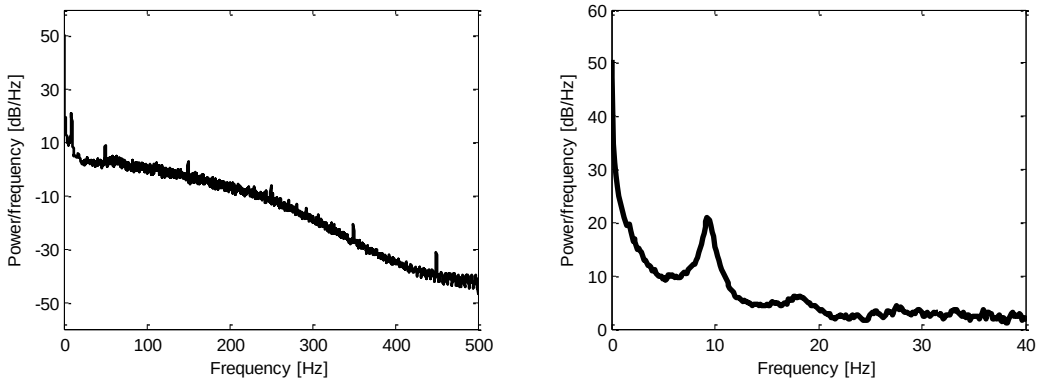

Figure 4: Typical EEG power density spectrum (left) and close-up of the low-frequency range (right) depicting the 1/f nature of EEG spectra. Notice the 50 Hz power line noise and its harmonics (left) and the peaks close to 10 and 20 Hz (right) corresponding to alpha and beta rhythms, respectively

3.2.2 Magnetoencephalography (MEG)

Electrical currents flowing through the axons of neurons induce a very weak orthogonal magnetic field, which can be measured outside the skull by means of magnetoencephalography. Unlike the electric field measured by EEG, the magnetic field is barely influenced by the surrounding brain tissue, CSF and skull. Furthermore, MEG has a spatial resolution of approximately 5 mm **[16]**, superior to that of EEG **[19]**, and also has a wider frequency range **[20]**. BCI experiments with MEG in two-dimensional control provided satisfactory results, achieving 69% accuracy in a four-class scenario **[21]**.

MEG-based BCIs are still far away from widespread use. The magnetic field produced by the brain is very weak, roughly 10 fT (10^{-14} T), many orders of magnitude weaker than the Earth's magnetic field, which is about 0.5 mT. MEG therefore requires magnetically-shielded rooms. MEG devices are also bulky and expensive, as they consist of arrays of extremely sensitive SQUID (superconducting quantum interference device) detectors which need to be cooled in liquid helium (see Figure 5). Furthermore, while EEG recordings permit some level of head movements, MEG requires complete stillness.

Figure 5: Participant in an MEG study at the National Institute of Mental Health, USA

3.2.3 Functional magnetic resonance imaging (fMRI)

Up to this point, we have discussed neuroimaging techniques that record brain activity based directly on electromagnetic measurements. A different approach is to indirectly infer cortical activity by measuring the metabolic response of specific brain regions, particularly the blood oxygen level dependent (BOLD) response. Active neurons produce an increase in oxygen-rich blood flow in surrounding tissue, and this change can be captured by functional magnetic resonance imaging. The prefix "functional" refers to the usage of standard MRI techniques for the imaging of blood flow instead of structural tissue.

fMRI provides good spatial resolution (1 mm **[16]**) and was successfully used for online two-dimensional control of a robotic arm by a human subject **[22]**. Its practical use in BCI is however limited by the large, expensive equipment and the rather poor temporal resolution of BOLD signals (up to several seconds), as the hemodynamic response is slower than the underlying neural activity.

3.2.4 Functional near-infrared spectroscopy (fNIRS)

Much like fMRI, functional near-infrared spectroscopy also measures the BOLD response, this time by projecting light in the near infrared range (700 – 900 nm) onto the scalp and monitoring the relative amounts of oxygenated and deoxygenated hemoglobin (HbO and HbR). Within these limits of the electromagnetic spectrum, bone and tissue are almost transparent, while blood is a stronger absorber of light. NIRS devices are relatively cheap and portable, but their

performance in BCI applications is lower than what can be achieved by EEG **[23]** and they also suffer from the relatively high latency of the hemodynamic response.

3.3 Conclusion

While in clinical and neuroscience research applications expensive and non-portable equipment have their use, for practical, everyday BCIs, portability is a must. Leaving invasive methods aside for similar reasons, we are thus left with a choice between EEG and fNIRS.

Combining the two techniques might prove to be most beneficial. A recent study by Fazli et al found a 5% increase in classification accuracy in a two-class motor imagery experiment by complementing EEG with fNIRS. The increase is modest though, and EEG surpassed fNIRS in accuracy for both executed and imaginary movements. An interesting detail is that while EEG performance dropped 13% (from 90.8% to 78.2%) in the case of motor imagery (MI) compared to motor execution (ME), fNIRS performance remained quite stable, and quite surprisingly the oxygenated hemoglobin feature even had a slightly higher classification accuracy (71.7% for MI compared to 71.1% for ME) **[24]**.

Simultaneous recordings of EEG and fNIRS will most likely be a hot BCI research topic in the years to come, but the slow BOLD response will still hamper the information transfer rates achievable by fNIRS systems. Berger would be proud to know that even though we landed a man on the Moon and have supercomputers in our pockets, his invention still holds the most promise for the future of BCI. The rest of this paper will be therefore focused on EEG-based BCIs.

4 Neurological phenomena

A common myth surrounding BCIs is that they read minds. While we hope that the above review of neuroimaging techniques shed some light on the difficulty of capturing brain activity, we must stress that BCIs (especially non-invasive variants) simply rely on the detection of specific cerebral patterns and associate them with commands that can be outputted to external devices. These patterns can be either a result of the subject performing some mental task or a natural response of the brain to some stimulus. Neurological phenomena generated as the result of cognitive processes are called *endogenous*, while the ones evoked by an external stimulus are called *exogenous*.

To gain a better understanding of the inner workings of BCI systems, in this section we will outline the most important neurological phenomena that are presently used in EEG-based BCIs.

4.1 Sensorimotor rhythms

The neural oscillations recorded by EEG are generally referred to as brain waves or rhythms. They have distinct topographical and spectral distributions and are linked to specific brain activities. The brain waves associated with movement tasks (among others) are called sensorimotor rhythms (SMRs) and are predominantly found over the sensorimotor cortex. Figure 6 illustrates the major rhythms recorded from a healthy human subject through EEG.

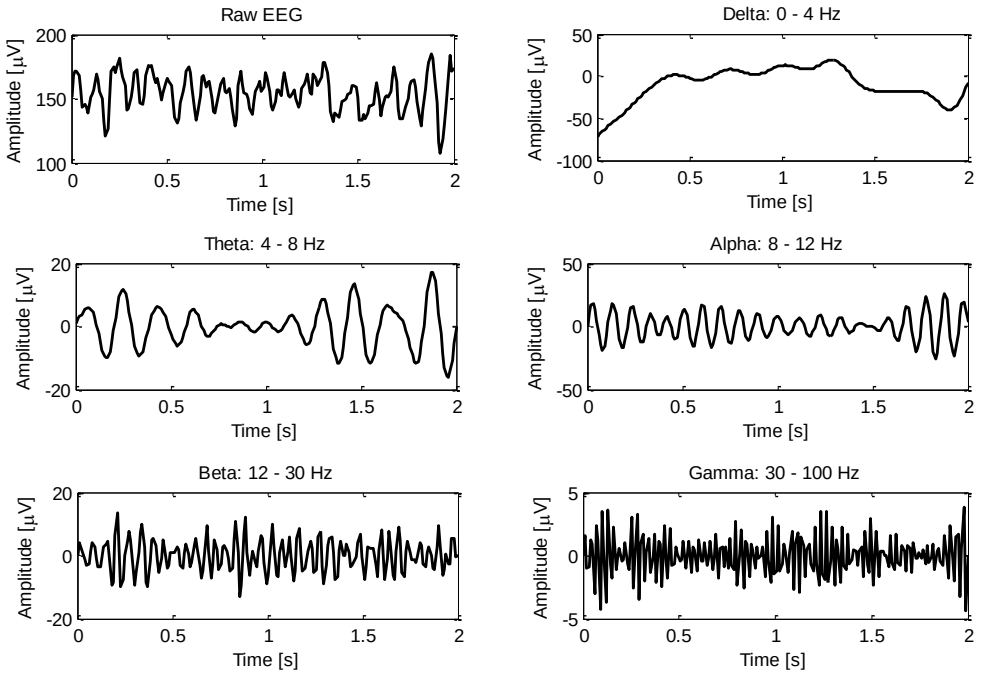

Figure 6: Typical brainwaves and their associated frequency bands. Notice the inverse relation between amplitude and frequency

In addition to the rhythms illustrated above, a particularly important brainwave for BCI is the mu rhythm, which commonly occupies a similar frequency band as alpha (8 – 13 Hz), but has a distinct arc shape and is predominantly found over the sensorimotor cortex. The mu and beta rhythms are the most important to BCI as their amplitudes are modulated by real or imaginary movements. These modulations have distinct topographical distributions for different body parts.

All muscles and organs are innervated by neurons, the wiring of our body responsible for carrying information. Muscular commands are downloaded from the brain and sensory information is uploaded back. The cortical areas devoted to controlling or feeling different body parts have distinct locations on the sensorimotor cortex. Neuronal networks responsible for adjacent parts of the body are also located next to one another, but their area is not a reflection of the physical size of the body region they control, but rather of the number of neurons that innervate the respective region. The representation of the body within the brain is often called the cortical homunculus and is illustrated in Figure 7.

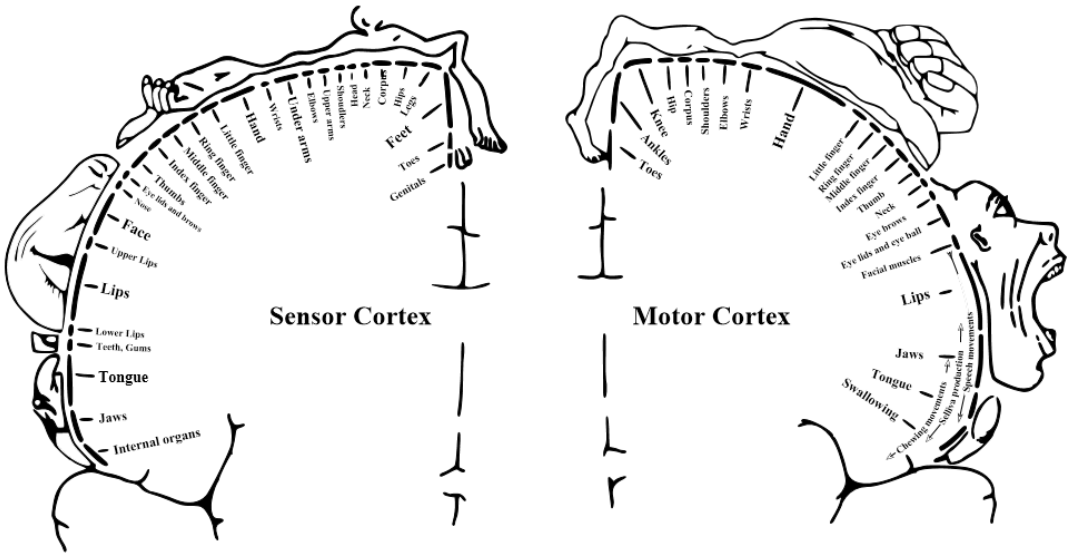

Figure 7: Homunculus model of the sensorimotor cortex, showing the layout of different body parts

Certain events can synchronize or desynchronize distinct neuronal networks, resulting in frequency and spatially specific increases or decreases in amplitude. Similar to ERPs, these changes are time-locked to the event, but are not phase-locked, thus cannot be extracted by time-domain averaging, but rather by spectral analysis **[25]**. An increase in band power is called event-related synchronization (ERS) and correspondingly, a decrease is referred to as event-related desynchronization (ERD).

Idling periods are characterized by synchronous neural firing and generally large power in the mu and beta bands. Executed, imagined, or attempted movements (in the case of paralyzed patients) desynchronize neural activity, producing ERD of the mu and beta rhythms. Upon completion of the movement, an increase in power called post-movement ERS is usually found, especially in the beta band **[25]**. The distinct topographical mapping means that movements of different body parts will elicit power modulations in specific brain regions, thus forming the basis of sensorimotor-based BCIs.

These BCIs have many advantages. Paradigms requiring the user to change gaze direction for selecting a letter do not rely solely on brain activity, as they also require ocular movements. They are therefore called *dependent* BCIs **[2]**. SMR-based BCIs do not rely on any stimulus or muscular output path, but rather on internally generated events, such as an imaginary or attempted movement, thus being *independent* BCIs which could be used even if the senses are impaired. This is also true for SCP-based BCIs, but the SMR paradigm is less affected by EOG artifacts, thus potentially requiring fewer sensors and processing power invested in artifact correction.

Furthermore, SMR systems are a valuable tool in rehabilitation **[26]** and in potential neuroprosthetic applications due to the movement-related paradigm. In what can be considered state-of-the-art in EEG-BCI, Wolpaw et al have demonstrated for the first time 3D cursor control, potentially validating EEG as a source of control for artificial limbs **[27]**. The study shows that the three dimensions were controlled simultaneously rather than sequentially. Users reported that 3D

control became easier with training as the precise nature of the mental task became less important and control became more automatic. This is similar to the training of regular motor skills, such as walking. The authors also compare the 2D target acquisition time for the intracortical BrainGate implant **[11]** with a previous motor imagery EEG study **[28]** and show that surprisingly, EEG control had the same success rate and latency as intracortical implants.

4.2 Evoked potentials

An evoked potential is the brain's response to an external sensory stimulus. In EEG recordings, evoked potentials manifest as slow changes in voltage. The most commonly used evoked potentials in BCI are visually evoked potentials. Flickering light at a given frequency elicits cortical activity at the same frequency in the occipital lobe, which governs visual information processing. This particular type of evoked potential is called the steady-state visually evoked potential (SSVEP). Thus, when two or more targets are used, each with a different frequency, the user's gaze direction can be inferred. This paradigm can be used for communication and control in BCI systems. As a matter of fact, the first BCI developed by Vidal in 1973, used SSVEP for navigating a maze **[29]**. The pioneering work of Jacques Vidal was the first proof on the feasibility of EEG-based communication and also introduced the term "brain-computer interface".

4.3 Event-related potentials

A particular type of evoked potentials is the event-related potentials (ERP), which are the result of higher-level processes, involving attention, expectation or memory, among others. The most common ERP employed in BCI research is the P300 response, which manifests as a positive voltage deflection appearing roughly 300 ms after the presentation of a particularly important visual, auditory or somatosensory stimulus, when interspersed with other frequent or routine stimuli **[2]**. The most common application of P300 is in speller devices, where the user is presented with a 6 × 6 matrix of letters, numbers and/or commands. Each column and row is flashed multiple times for averaging EEG noise, and the P300 is detected only for the column or row which contains the desired symbol. Thus, the user's intended choice can be inferred **[30]**. As with SSVEP, one advantage of P300 is that it is a natural response of the brain, thus requiring no user training.

4.4 Slow cortical potentials

Unlike evoked potentials, slow cortical potentials (SCP) are not dependent on external stimuli, but rather reflect an increase or decrease in cortical activity. They are among the lowest frequency features detectable by EEG, and manifest as negative or positive voltage deflections over the course of 0.5 – 10 seconds. Negative SCPs are associated with movement, while positive ones are commonly associated with reduced cortical activation **[31]**. With extensive user feedback training, up to several months, people can learn to control the polarity of SCPs. This self-regulation can be used for binary selections in BCIs, and formed the basis of the "Thought Translation Device" BCI, used to restore basic communication abilities in ALS patients **[32]**. Besides the obvious drawback of requiring prolonged user training, SCP-based BCIs provide relatively slow communication speeds, at a rate of 0.15 – 3 letters per minute **[2]**, and due to their low frequency nature, are highly susceptible to artifacts such as eye movements.

4.5 Conclusion

The obvious advantage of BCIs that use exogenous neurological phenomena is that they require no user training. The downside of such systems is the constant commitment of a sensory modality such as vision **[33]** to potentially repetitive stimuli that might cause user fatigue and frustration. Endogenous neurological phenomena indeed require some level of user training for generating stable and consistent cortical patterns, but once the user familiarizes with the control method, it is expected that control will gradually become easier and more natural. Furthermore, because such independent BCIs do not rely on any sensory pathways, they are highly valuable for patients with impaired senses. It comes to no surprise that more than 80% of BCI research is focused on endogenous neurological phenomena **[34]**.

Chapter II
The self-paced BCI

We have seen that, when presented with specific stimuli or performing certain mental tasks, characteristic changes affect cerebral activity. These cortical patterns can be used for communication and control in brain-computer interfaces. Most recording techniques are noisy and spontaneous EEG activity in particular resembles a stochastic process with the relevant features for control buried in noise. This means that the characteristic patterns associated with possible output commands vary quite a lot and often overlap. Unlike traditional, deterministic communication systems in which noise mostly affects the transmitted message, in EEG-based communication noise is present all the time, whether the user desires to output a command or not. Until some complete model describing all possible EEG activity is established, we are forced to consider as noise all EEG activity except the characteristic neurophysiological phenomena used for control, depending on the paradigm. This can be troublesome, as the relevant EEG features are often smaller in amplitude than the noise itself; hence there is a considerable probability that spontaneous activity might emulate the control features without the user having intended to output a control signal.

This brings up an important topic: while in a discriminative scenario, such as distinguishing imaginary movements of the left and right hand, it is fairly easy to attain good performance (>90%) **[35]**, how do we go about distinguishing the active tasks from all other possible EEG activity? To gain a better understanding of the problem, let us first put it better in context.

1 The no control (NC) state

In most communication systems and interfaces, self-paced control is not an issue. We pick up the telephone and call somebody if and when we want to, use computers when we desire and turn the steering wheel of a car only when needed. That is because such systems rely on physical interaction, and periods in which the user has no control (NC) intention are characterized by no input to the system. Such a clear delimitation is presently impossible to achieve with EEG-based BCIs, due to the spontaneous cortical activity that often resembles the one associated with periods of intentional control (IC).

NC support is necessary in all applications where frequent periods of intentional control are interspersed with periods of inaction. It is noteworthy to mention that the NC state is not the same as an idle or relaxed state in which the user tries to think of nothing, but rather entails these as sub-states. As discussed in section 4.1, these periods have distinct characteristics, such as large power in the mu band of sensorimotor rhythms, and represent only one possible type of NC. The NC state is therefore characterized by all possible cortical activities other than the ones used for intentional control. This tremendously complicates designs, as BCIs with NC support must handle the variety of additional tasks the user might be doing between IC periods, whether daydreaming or staring out the window for a few minutes, watching a movie for a few hours or performing any action other than trying to control the BCI. Arguably, it is unlikely that one can model all possible NC sub-states **[36]**.

One possibility would be to simply turn the BCI off when not needed. This could be performed manually by the user or through some automatic mechanism that would put the BCI into a low-power state, similar to the "sleep" mode of laptops **[8]**. Implementing an automatic "sleep" mode is not a solution though, as it implies that the system is already able to recognize long periods of inactivity, therefore entailing the availability of NC support. A solution for turning off the BCI is the so-called "brain switch", a specific mental task that the user needs to perform in order to turn the BCI on or off. This has been experimented with mostly in hybrid BCIs, which use SSVEP for control and the post-movement ERS of brisk imagined foot movements as the brain switch. In such an experiment, the brain switch decreased false activations in NC periods from 5.4 to 1.4 per minute **[37]**. However, the use of a brain switch does not resolve the intrinsic problem of NC support for the primary BCI, and might not be convenient in many applications. For long periods of inactivity, turning the system off might indeed be beneficial and even necessary, but it is not an appropriate solution for the majority of everyday interactions which require short and frequent pauses, such as conversations or the navigation of a motorized wheelchair through the environment.

2 BCI control paradigms

For a better clarification of the problem, in the following we will review the various timing mechanisms used in BCI research and the different types of output. We adopt the classification proposed by Mason et al in a technical report dedicated to self-paced BCIs **[8]**.

2.1 Timing mechanisms

Synchronized control

In synchronized control, the BCI is periodically available to the user when it is on/awake and it does not support NC. The user is prompted via a cue when a control period (trial) starts and does not have the possibility of not issuing any command to the system. Especially for BCIs based on the modulation of sensorimotor rhythms, this is the predominant form of control and is also the most simple.

Constantly engaged

These BCIs are constantly available to the user but do not support NC. This is not a practical mode of control, as the user must continuously control the BCI and correct false activations during periods of inactivity. An example of such a system is the virtual keyboard speller developed by Scherer et al, a 3-class BCI in which users select letters with imaginary movements of the left hand, right hand and foot **[38]**.

System-paced

With system-paced timing, the BCI is periodically available to the user when it is on/awake and it supports NC. Thus, the system is still controlled on a single-trial basis, but the user has the option of not issuing any command during control periods.

Self-paced

Self-paced control means that the BCI is constantly available to the user when it is on/awake and it supports NC. The user thus has full control over the BCI and uses it at his or her discretion. This is arguably the most natural form of control for any interface.

2.2 Continuous and discrete outputs

As discussed, BCIs can have either continuous or discrete outputs. The vast majority of classification procedures used in BCI gives a continuous output. An unknown sample goes through the same feature extraction and processing pipeline used in training and receives a score which represents its similarity to the known classes. This score can be either a distance measure (how close is the sample to each of the known classes in feature space) or a probabilistic one (what is the probability that the sample belongs to the known classes). Irrelevant of its nature, a discrete output can be obtained by simply applying a threshold on the continuous score.

Most BCIs which rely on exogenous neurological phenomena have discrete outputs, as to express the presence or absence of a certain stimulus. In general, presenting a measure of the similarity score is useful in endogenous BCIs for providing feedback, which in turn allows the user to learn and adapt. However, the desired type of output depends on the target application. For example, in the case of the virtual keyboard developed by Scherer et al **[38]**, the presented feedback is continuous but the output is discrete, as the user chooses from a finite set of possible symbols. On the other hand, applications in neuroprosthetics might require continuous output for the precise control of each joint.

In general, self-paced applications benefit from continuous outputs, as the thresholding of similarity scores allows fine-tuning the balance between BCI sensitivity and false activations.

2.3 Event-driven and state-driven outputs

Mason et al describe two possible implementations of self-paced BCIs with discrete outputs: event-driven and state-driven designs **[8]**.

In event-driven self-paced BCIs, the user has the ability to initiate a state change whenever desired, but cannot maintain this state because of the transient nature of the underlying neurological phenomena. That is to say, the return to the NC state is automatically performed by the brain. Because of the automatic return to NC, event-driven designs lack the possibility of directly transitioning from one IC state to another IC state. BCIs based on movement-related potentials or P300 are common examples of event-driven designs.

In state-driven self-paced control, the user can initiate a state change when desired and additionally has the ability to maintain (hold) the given state. The duration of IC periods is therefore also at the user's discretion. Compared to event-driven control, the user also has the possibility of transitioning from one IC state to another without the need of going through the NC state. Arguably, state-driven implementations provide the richest form of control. These designs are only possible in BCIs which use thresholded neurological phenomena, such as the modulation of sensorimotor rhythms or SSVEP. A graphical representation of the differences between the two control methods is depicted in Figure 8.

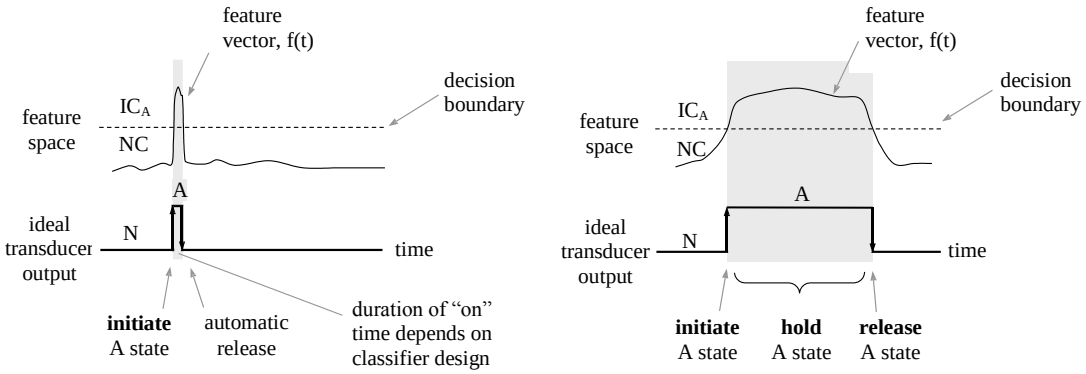

Figure 8: a) Event-driven discrete control, based on transient neurological phenomena, such as movement-related potentials or P300; the return to the NC state is automatically done by the brain. b) State-driven discrete control: the user has the ability to initiate, maintain and release an intentional control state. Adapted from the technical report on self-paced BCI by Mason et al [8]

While it may not become immediately apparent from Figure 8, the possibilities offered by state-driven BCIs are substantially larger, as the user has the option of maintaining an IC state or directly switching to another IC state. This is represented by the fully connected graph in Figure 9.

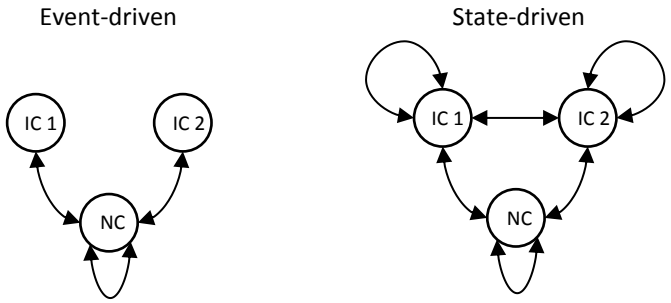

Figure 9: Event-driven BCIs allow only transitions between each IC state and NC, while state-driven BCIs allow all possible transitions

2.4 Conclusion

There is an important point to be emphasized here: self-paced control is an issue only for BCIs relying on endogenous neurological phenomena. Exogenous paradigms such as P300 or SSVEP are inherently self-paced, as the user pays attention to the stimulus only when desired. While false activations are still possible, they are less likely. So then why go through all the trouble of enabling self-paced operation for endogenous BCIs? Aside from the drawbacks already discussed in section 4.5, the requirement of an external stimulus makes P300 and SSVEP solutions rather unusable in one of the most sought-after BCI applications: the control of artificial limbs. In this context, endogenous BCIs are the only option, as they present another crucial advantage: they give the user the possibility to adapt and learn and ultimately better control the BCI.

Unsurprisingly, it becomes apparent that the very existence of such applications is conditioned on the availability of NC support.

Thus, from now on our discussion will be focused on endogenous neurological phenomena. State-driven BCIs based on sensorimotor rhythms seem the best candidate for self-paced operation, simply because they give the user full control over the transitions and durations of IC states.

3 Performance metrics for self-paced BCIs

Evaluating synchronized BCI experiments is not a challenge. For each trial, the true and predicted labels are known, thus a myriad of performance metrics can be derived. For a review, interested readers are referred to the report of Schlögl et al **[39]**. One could argue that the same holds true for self-paced BCIs as well. After all, we know the true label of each sample, thus we can compare it to the predicted label and apply the same evaluation criteria. The problem with this approach is that more often than not, self-paced data is unbalanced, i.e. there are many more samples from the NC state than from IC states. This means that many of the well-established criteria such as accuracy are no longer applicable. Consider a scenario where there are 9 times more NC samples than IC, and the system has 90% accuracy. Normally, such a performance would be considered very good, but it might very well be the case that the system classified all samples as NC, therefore achieving an accuracy of 90%. What is therefore needed are performance metrics that measure the detection rate of IC states (call them positives) and the rate of false detections during NC (call them negatives). Ideally, all IC states would be correctly detected and no false activations would occur during NC.

But this is only part of the story. While we can come up with adequate performance metrics for unbalanced data, we need to ask ourselves whether such "sample-by-sample" evaluations are actually meaningful. Consider the following scenario: there are 10 IC-related activations (call them events), each having 10 samples and we have two algorithms to evaluate. One algorithm correctly detects all 10 samples of only one out of the 10 events, while the other detects only one sample for each of the ten events. While neither of the two possibilities is exactly optimal, generally the second option is preferable. This type of evaluation is the most appropriate for self-paced BCIs and is called "event-by-event" evaluation.

3.1 Sample-by-sample evaluation

In sample-by-sample evaluation, each sample of the BCI output is compared to the label of the intended output for that sample. Different metrics can be used for this evaluation, such as the mean squared error, the mutual information between true and predicted labels, or the area under the ROC curve.

Mean squared error (MSE)

The mean squared error of the predicted output with respect to the true labels is not exactly a common metric in self-paced BCI evaluations, but it is frequently used in the case of continuous outputs. The main reason of including it in this list is because it was the performance metric for

self-paced BCIs in the fourth international BCI competition[2]. What was most surprising about this decision is that previously (in the third competition) the evaluation criterion for self-paced BCI was mutual information, which is generally considered to be a much better alternative.

In general though, the mean squared error is not a good indicator of performance for self-paced evaluations. Let us consider a three-state self-paced BCI with two IC states. Assuming the label of the NC state is 0 and the two IC states are represented by -1 and 1, misclassification errors between the two IC states will be larger than false activations, i.e. errors between 0 and ±1. While this might seem intuitive, we need to keep in mind that in most of the time the different IC states can be distinguished with satisfactory accuracy. The difficult part is distinguishing them from the NC state, and exactly this aspect is not emphasized by the mean squared error. The MSE is a highly informative metric in cursor control applications with continuous output, such as neuroprosthetics **[40]**, where the preciseness of control is to be evaluated.

Mutual information

The mutual information between two processes *X* and *Y* is a measure of the decrease in the uncertainty of *X* when knowing *Y*, or vice versa. It is a symmetric function of their joint probability distribution and is defined as:

$$I(X,Y) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy \qquad (1)$$

This is a highly desirable property of a BCI, as it is a direct measure of the user's intent **[39]**. For both discrete and continuous outputs, as well as different number of states, mutual information is considered a relevant and informative performance metric. However, it is difficult to compare the results of different studies based on the mutual information, because this metric is also sensitive to the sample size of each state, which generally varies between studies.

Receiver operating characteristics (ROC)

The receiver operating characteristics (ROC) curve is a popular and insightful evaluation criterion for signal detection systems. It is created by varying the decision threshold of a binary classifier with continuous output and calculating for each step the true positive rate (TPR) and the false positive rate (FPR). The former is a measure of sensitivity, while the latter is a measure of specificity.

Considering that *TP* is the number of true positives (correctly classified positive samples), *P* is the total number of positives, *FP* is the number of false positives (incorrectly classified negative samples) and *N* is the total number of negatives, the true and false positive rates are given by

$$TPR = \frac{TP}{P} \qquad FPR = \frac{FP}{N} \qquad (2)$$

[2] http://bbci.de/competition/iv/

These performance metrics are advantageous as they provide a natural solution to the issue of unbalanced data. The true and false positive rates can be combined in a single performance metric, the area under the ROC curve (AUC). The area under the ROC curve is equal to the probability that a random positive sample will be ranked higher than a random negative sample. An AUC of 0.5 represents random performance and an AUC of 0 or 1 represents perfect classification. While the AUC is only applicable in binary classification, multi-class extensions have been proposed **[41]**.

3.2 Event-by-event evaluation

The common practice in self-paced BCI evaluation is to report event-based TPR and sample-based FPR **[8, 36, 42]**. The event-based TPR is defined as the number of successful IC-related activations relative to the number of attempted IC states **[8]**. Therefore, it becomes necessary to count multiple detections within a single event as a single true positive, otherwise multiple correct activations within an IC state can create the illusion of many successful detections, when in fact only one event occurred **[43]**. A drawback of this approach for self-paced BCIs with more than one IC state is the lack of a formal definition of classification errors between IC states. It is possible that more than one IC state is detected during an event, including the correct one. In this paper, such an event is counted as a true positive, but is considered a misclassification error between IC states.

Event-based false positive rates can be calculated as well, but this may lead to an overly pessimistic view on performance, as any false activation during an arbitrarily long NC period would label the whole non-event as a false positive. Because of the variable durations of NC states, sample-based FPRs are generally preferred. The standard practice is to report event-based TPR corresponding to a sample-based FPR of 1% **[36]**. Fixing the FPR to a predefined value leaves only one metric, the event-based TPR, which is more easily comparable between studies.

Another possibility of combining event TPR and sample FPR in a single metric is the true-false difference (TF) introduced by Townsend et al **[43]** and defined as

$$TF = \left(\frac{TP}{E} - \frac{FP}{E + FP}\right) \times 100 \tag{3}$$

where E represents the number of IC events. In this context, all detections during a non-event are counted as false positives, and a true positive is defined to be one or any number of detections during an event.

To provide a richer description of performance, these metrics should also be accompanied by others describing the temporal characteristics of the BCI. An important factor is the response time, i.e. the delay between user intention and BCI response, which can be given either as a histogram or in terms of mean and standard deviation. Because in event-based evaluations a single correctly detected sample is sufficient to label the whole event as a true positive, regardless of its duration, it becomes necessary to also report a measure of the BCI's ability to maintain a certain IC state. The hold time can be evaluated and presented similarly to the response time, either as a histogram or through statistical parameters. The distribution of inter-false activation periods is also highly relevant. If this distribution is biased towards 0, it means that false activations tend to appear in patches; if it is uniform, one can assume that false activations tend to appear randomly. Finally, a characterization of NC period durations must also be given, similarly in the form of a histogram or

statistical parameters. This helps in determining how often a subject attempted control and consequently if the results are applicable to the target application.

One problem with the TPR and FPR metrics is that they are independent of the decision rate. A BCI with a decision rate of 10 Hz is expected to produce one false activation every 10 seconds for an FPR of 1%, while a decision rate of 1 Hz would produce one false activation every 100 seconds, which is potentially more useful in many applications. For all evaluations based on TP and FP rates, the decision rate must also be mentioned.

4 Previous work

There is no doubt that many potential applications of BCI are currently limited by the lack of NC support. The variety of spontaneous EEG activity makes modeling the NC state quite difficult. From the signal processing point of view, self-paced operation poses considerable challenges on the techniques commonly used in single-trial classification, and there is currently no golden standard for such systems.

Research into self-paced endogenous BCIs kicked off in 2000, when Mason and Birch proposed the Low-Frequency Asynchronous Switch Design (LF-ASD) **[42]**. The LF-ASD was designed to detect the movement-related potentials (MRP) of a finger flexion task. It used features in the 0.1-4 Hz band of six bipolar EEG channels recorded from the locations F_1-FC_1, F_Z-FC_Z, F_2-FC_2, FC_1-C_1, FC_Z-C_Z and FC_2-C_2. The features were extracted by a simplified version of the discrete wavelet transform (DWT) and classified with a 1-nearest neighbor (1-NN) classifier. EEG data was recorded from five subjects and the initial offline results were highly promising, with true positive rates in the range of 38%–81% and corresponding false positive rates in the range of 0.3%–11.6%. Subsequent online evaluations showed the system's ability to operate with imagined finger flexions as well; two subjects with high-level spinal-cord injuries achieved true positive rates ranging between 45-48% and false positive rates below 1% **[44]**. Over the years, the LF-ASD provided the general framework of many studies which brought improvements to the original design, such as the addition of an energy normalization transform **[45]**, the addition of a debounce window in post-processing for reducing FP rates **[46]** or the use of genetic algorithms for improved model selection **[47]**.

Despite the promising results, there are several drawbacks that limit the functionality of the LF-ASD and many of its variants. First of all, this is a two-state BCI which can only distinguish between NC and IC periods of a single mental task. This approach is different from what is normally pursued in BCI research, which started with the discrimination of two or more mental tasks and additionally seeks to add NC support. It is not clear how a binary output with only one IC state would be useful in common BCI applications. The most obvious application of the LF-ASD would be the push of a button, such as turning the TV on or off. However, the lack of an additional IC state would make it impossible for the user to directly switch between devices or between different controls of a single device. One possibility would be to assign two or more functionalities to the unique IC state, depending on its duration. Unfortunately, this is not feasible either, as the LF-ASD is an event-driven BCI, owing to its use of movement-related potentials (see section 2.3). This is not to say that using a brain switch for more than one command is impossible; one could simply let the BCI cycle through possible choices and stop it on the desired command. However, there is no doubt that having more than one IC state is advantageous.

A crucial improvement to the LF-ASD was proposed by Bashashati, who extended the design by adding a second IC state **[36]**. The resulting 3-state BCI detects executed left and right hand extensions from ongoing EEG activity, and consists of two detectors. The first distinguishes IC commands from NC periods, effectively considering both IC states as one class, and the second detector discriminates between the two IC states. In both cases, bipolar montages were found to give better results than monopolar channels. Offline analysis of the EEG data recorded from four able-bodied subjects resulted in average TP rates of 37.5% and 42.8% (at the FP rate of 1%) in detecting right and left hand extensions, respectively.

Bashashati evaluated two different designs of the IC/NC detector, one based on the LF-ASD and the other based on power spectral density (PSD) estimation combined with linear discriminant analysis (LDA). For three out of the four subjects, the detector based on LF-ASD outperformed the one based on PSD-LDA. Apparently, this result suggests that compared to sensorimotor rhythms, low-frequency features such as MRPs might provide better discrimination ability between NC and IC states. However, a closer look shows that Bashashati's SMR detector has several drawbacks. First, the bipolar montage is not the best choice for capturing SMR modulations. The common spatial patterns (CSP) method linearly transforms multi-channel EEG data such that band-power differences between two classes are maximized and was shown to be superior to bipolar montages **[48]**. Second, if the two IC states are separated by the NC state in feature space, it might be impossible for a single linear classifier to effectively separate the distributions. This situation is illustrated in Figure 10. Furthermore, the subject for which the PSD-LDA detector gave better performance was the only one with prior experience in using BCI. These insights suggest that sensorimotor rhythms might be a viable, if not superior alternative to low-frequency features such as MRPs.

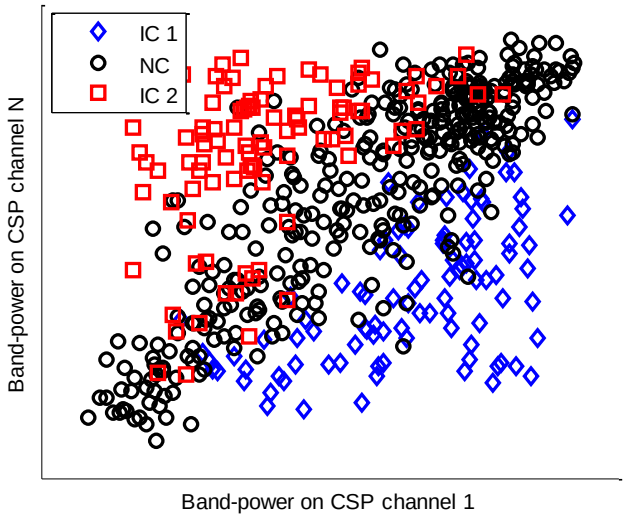

Figure 10: Depending on the distribution of IC and NC states in feature space, a single linear classifier might not be able to separate the two IC states from NC, if both IC states are considered one class

The Graz group developed a self-paced BCI based on SMR modulations which uses three bipolar electrodes and detects imaginary movements of the left hand, right hand and foot from the ongoing EEG **[49]**. The system also employs a two-stage classification procedure that first detects the presence of an IC command and then classifies it into one of the three IC classes. Similar to

Bashashati's design, the IC/NC detector also consists of a single LDA, while IC commands are classified by majority voting of three LDA classifiers. The distinction sensitive learning vector quantization (DSLVQ) was used to select subject-specific frequency bands **[50]**. Three healthy subjects participated in the study. Prior to using the self-paced BCI, subjects had approximately 4 hours of cue-based feedback training in which they learned to generate discriminative motor imagery patterns. The self-paced interface was used in two applications. The first consisted of navigating through a virtual environment and collect coins. Two out of the three subjects successfully collected all three coins. The second application was a simplified and intuitive interface between the BCI and Google Earth, called Brainloop. In a 40-minute media performance, one subject successfully used the interface and reported that "most of the time the BCI was correctly detecting the motor imagery patterns as well as the non-control state". Further evaluations of self-paced operation were not provided.

Similar approaches based on LDA and PSD estimation were used for controlling a bipedal robot **[51]** or playing a Hangman game **[52]**. While details of the implementation vary between studies, one can observe certain trends in self-paced BCI research, such as the two-stage classification of SMR modulations, the selection of subject-specific frequencies and the use of linear classifiers.

A completely different framework was introduced by Schalk et al, based on detection instead of classification **[53]**. Instead of explicitly training classifiers with examples of different motor imagery tasks, the authors propose an unsupervised Gaussian Mixture Model (GMM) trained only with examples of the NC state. Features were extracted by PSD estimation but no subsequent feature selection procedure was applied. Examples of IC states from the training data are used to calculate optimum log-likelihood thresholds that maximally separate different IC commands. This new approach proved superior to linear regression in discriminating between different motor tasks, but no analysis of NC periods was performed. However, the proposed detection approach is potentially unsuited for self-paced operation. Distinction between different IC commands is possible only provided that the IC states represent progressive modulations of signal features. It is expected that in situations such as the one illustrated in Figure 10, where IC states are somewhat equally far from NC, this method would be unable to distinguish between the two IC classes. Furthermore, a reliable self-paced BCI is expected to react only to specific commands and not to some measure of perturbation. Therefore, an intuitive approach would be to model the IC states as accurately as possible and consider everything else NC. The authors adopt the opposite approach, where the "rest" state is modeled (rather than a more general NC state) and anything else is considered IC. This is arguably not a wise choice for reliable self-paced operation, although admittedly, this was not the objective of the study.

One problem that arises is the difficulty of directly comparing the results of different studies. Differences in recording equipment, processing and classification pipelines are expected and even informative, but the most pressing issues are the different amounts of subject training and differences in the experimental protocol, especially differences in the time intervals during which false activations might occur. For the sake of objective and informative comparisons, new algorithms and methods could first be tested on publicly available datasets, rather than in application-specific scenarios.

Chapter III
Materials and methods

In the first part of this thesis, we have highlighted the advantages of using endogenous neurological phenomena for real-life, self-paced BCI applications. We have also seen that among the possible choices, BCIs based on sensorimotor rhythms could theoretically provide the richest level of control due to their state-driven design. These self-paced implementations provide the user not only with the option of initiating a change of mental state at arbitrary moments, but also with the ability to manipulate the duration of intentional control states.

Having decided on the appropriate neurological phenomenon, the next step is to consider the existing possibilities for the signal processing block: feature extraction and the translation algorithm. Regarding the extracted feature, we have seen that most implementations are based on capturing power modulations in specific frequency bands, as it directly correlates with the underlying neurological phenomenon (see section 4.1 of the first chapter): due to the distinct mapping of different body parts on the sensorimotor cortex, executed, imagined or attempted movements produce spatially localized de-synchronizations of sensorimotor rhythms, forming the basis of motor imagery BCIs.

In general, the mapping from the N-dimensional EEG feature space to the desired device commands can be performed by three broad types of methods: classification, regression and generative models. While all three have been used in the field of BCI, it is not clear whether any clear winner exists when applied in the context of self-paced operation. In this work, we focus on classification and regression. Our objective is to evaluate the two approaches and also investigate any possible advantages in combining them.

1 Experimental data

In the interest of an objective evaluation of our methodology, we opted for a publicly available dataset, specifically designed for the challenge of self-paced BCIs. Dataset I of the BCI Competition IV, originally recorded in **[6]**, consists of EEG recordings from four healthy human subjects (subjects *a, b, f* and *g*) and three artificially generated datasets (subjects *c, d* and *e*); in this work, we discard the artificial datasets. Signals from 59 EEG positions were measured that were most densely distributed over sensorimotor areas (see Figure 11 for layout). Signals were band-pass filtered between 0.05 and 200 Hz and then digitized at 1000 Hz with 16 bit (0.1 μV) accuracy. There is a 100 Hz downsampled version available as well, which was used in the present study. Downsampling was performed by first low-pass filtering the original data (Chebyshev Type II filter of order 10 with stopband ripple 50dB down and stopband edge frequency 49Hz) and then calculating the mean of blocks of 10 samples.

In the whole session motor imagery was performed without feedback. For each subject two classes of motor imagery were selected from the three classes *left hand*, *right hand*, and *foot* (side chosen by the subject; optionally also both feet).

- **Calibration data:** In the first two runs, arrows pointing left, right, or down were presented as visual cues on a computer screen. Cues were displayed for a period of 4s during which the subject was instructed to perform the cued motor imagery task. These periods were interleaved with 2s of blank screen and 2s with a fixation cross shown in the center of the screen. The fixation cross was superimposed on the cues, i.e. it was shown for 6s.
- **Evaluation data:** Then 4 runs followed which are used for evaluating the submissions to the competitions. Here, the motor imagery tasks were cued by soft acoustic stimuli (words *left*, *right*, and *foot*) for periods of varying length between 1.5 and 8 seconds. The end of the motor imagery period was indicated by the word *stop*. Intermitting periods also had a varying duration of 1.5 to 8s. In the evaluation data there are not necessarily equally many trials from each condition.

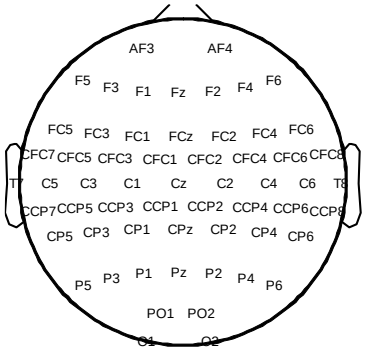

Figure 11: Electrode layout for dataset 1 of BCI Competition IV

2 Methods

2.1 Overview

A schematic overview of the pipeline is illustrated in Figure 12. The common spatial patterns algorithm is used to extract discriminative spatial filters in 26 frequency bands, with central frequencies from 7 to 32 Hz and a constant bandwidth of 2 Hz. All band-pass filters are 6^{th} order Butterworth filters. Band-power features are extracted by squaring the EEG amplitudes in non-overlapped short time windows of 100ms, which are then processed by the combined use of a median filter, log-transform and moving average. An information-theoretic feature selection based on the maximum relevance minimum redundancy criterion **[54]** is used to select the features that maximize the mutual information between class labels and features. Finally, classification and regression methods are used to map the input features to the output commands.

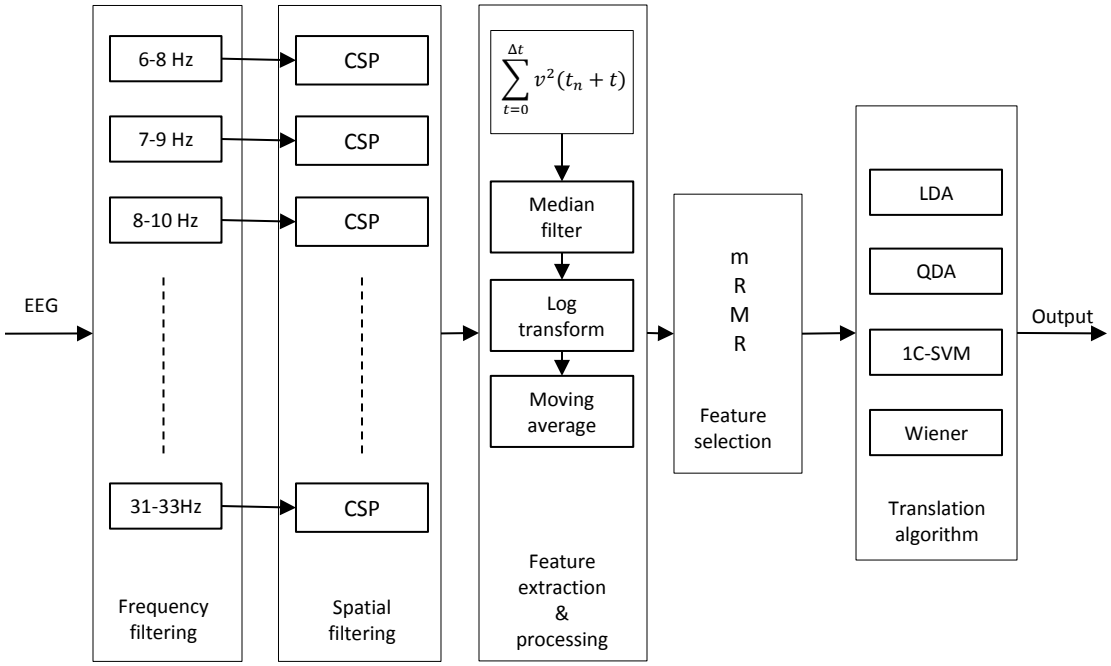

Figure 12: Block diagram of the BCI architecture

In the next subsections each of these steps will be presented in detail.

2.2 Common spatial patterns (CSP)

EEG data is very noisy and electrode signals are highly correlated due to volume conduction **[55]**. The common spatial patterns algorithm **[56]** linearly combines multichannel data for increasing spatial resolution and was proven to also increase signal-to-noise ratio **[57]**. It works by finding a set of spatial filters that maximize variance differences between two classes in a least squares sense **[48]**.

Let us denote a trial of EEG data by a matrix $X_i \in \mathbb{R}^{M \times N}$, where M is the number of electrodes and N is the number of samples. The spatial covariance estimate for each of the two conditions is given by

$$\Sigma^{(k)} = \frac{1}{N_k} \sum_{i \in C_k} \frac{X_i {X_i}^T}{\text{trace}\left(X_i {X_i}^T\right)} \qquad k \in \{+, -\} \tag{4}$$

where trace(x) is the sum of the diagonal elements of X, T is the transpose operator, k is the class label and N_k is the number of trials in class C_k. Normalizing the covariance estimates was common practice in the early days of CSP and was done to eliminate trial-to-trial variations in the absolute values of the second-order moments **[58]**. This idea became less and less popular in BCI research and was gradually abandoned for unknown reasons. Our preliminary experiments with single-trial classification resulted in better performance when using the normalized covariance estimate, thus we choose to do the normalization.

The composite spatial covariance is subject to eigenvalue decomposition

$$\bar{\Sigma} = \Sigma^{(+)} + \Sigma^{(-)} = U_c \lambda_c U_c^T \tag{5}$$

We sort the eigenvalues λ_c in descending order and compute the whitening transform P

$$P = \lambda_c^{-\frac{1}{2}} U_c^T \tag{6}$$

which equalizes the variances in the space of U_c and decorrelates channels such that $P\bar{\Sigma}P^T = I$, where I is the identity matrix. The whitened within-class covariance estimates then share common eigenvectors

$$S^{(k)} = P\Sigma^{(k)}P^T = B\lambda^{(k)}B^T \tag{7}$$

with $\lambda^{(+)} + \lambda^{(-)} = I$ and the eigenvalues in $\lambda^{(k)}$ also sorted in descending order. Therefore, an eigenvalue $\lambda_j^{(k)}$ close to one indicates that the associated spatial filter yields large variance for class k and small variance for the other class. The CSP projection W is then given by

$$W_{CSP} = (B^T P)^T \tag{8}$$

The columns of W are the spatial filters, and the columns of W^{-1} are the responses, or spatial patterns of the filters. Considering that the variance of a band-pass filtered signal (i.e. zero-mean) represents the average power in that band, the projections thus obtained are very convenient for capturing the specific topographies of power modulations during motor imagery. Moreover, because the eigenvalues are sorted in descending order, most of the discriminative information lies in the first and last m spatial filters, where usually m = 3. A visual representation of this situation is presented in Figure 13.

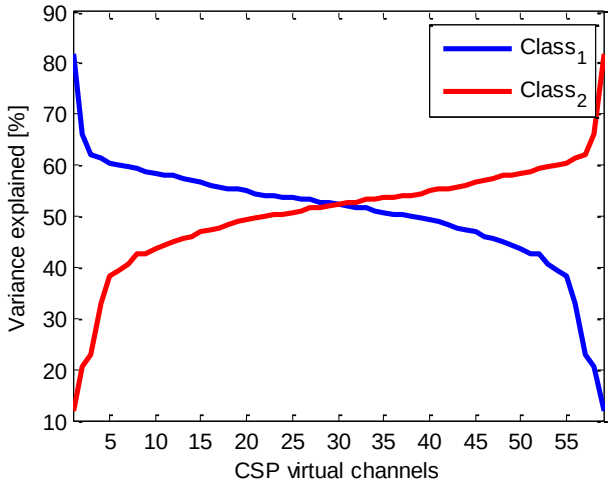

Figure 13: The CSP algorithm finds a set of virtual channels which maximize variance differences between two classes

As mentioned previously in the overview, the CSP algorithm is applied on a total of 26 frequency bands, with central frequencies equally spaced between 7 and 32 Hz and constant bandwidth of 2 Hz.

2.3 Feature extraction and processing

In order to take advantage of the discriminative projections obtained by CSP, we extract amplitude modulation features by calculating EEG signal power in short time windows:

$$x(t_n) = \sum_{t=0}^{\Delta t} v^2(t_n + t) \tag{9}$$

where $v(t)$ represents the amplitude of spatially filtered EEG at time t and $\Delta t = t_{n+1} - t_n$ is the duration of the window. Because windows are not overlapped, the window duration gives the maximum output rate of the BCI and we let Δt = 100ms for a decision rate of maximum 10 Hz.

Given the poor signal-to-noise ratio of EEG signals, squaring their amplitudes will drastically amplify noise and undesired artifacts such as power spikes. Noise reduction is therefore necessary and our approach is to handle Gaussian and impulse (non-Gaussian) noise separately with appropriate techniques. The noise-reduction pipeline employed here consists of a median filter for removing spikes and a moving average window on log-transformed data for attenuating normally distributed noise. These processing steps are illustrated in Figure 14.

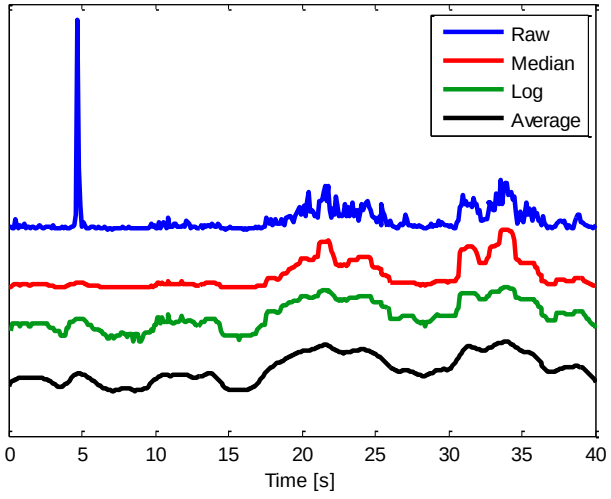

Figure 14: Processing steps of our feature vectors

The median filter is a non-linear order statistic filter, in which the center sample of the moving window is replaced by the median value of that window. The robustness of the median estimator in the presence of outliers makes it particularly useful in removing impulse noise (see Figure 14), while preserving the underlying structure of the data **[59]**. These properties make it a popular tool in image processing for removing "salt and pepper" noise, and we believe that the

same properties are also useful in EEG noise reduction. Unlike the moving average, the moving median does not perform any calculations on the window, making it invariant to the log operation. That is to say, $\mathrm{med}(\log(x)) = \log(\mathrm{med}(x))$.

The moving average is effective at removing Gaussian noise so we choose to log-transform the data prior to its application, in order to better approximate a normal distribution. The logarithm additionally has the advantage of facilitating subsequent processing by reducing dynamic range and also attenuates any remaining impulse artifacts **[60]**.

As with most motor imagery BCIs, the feature we extract is a measure of band-power. The key difference compared to other implementations is the short window duration. For example, the authors of the winning algorithm of the BCI Competition IV for this dataset used a window duration of 2.5 seconds **[61]**. The longer the size of the window, the less influence noise and artifacts will have, at the expense of a slower reaction time. Even when using long window durations, the variance estimator is highly susceptible to outliers, especially to high-powered artifacts such as impulse noise. Moreover, the influence of these artifacts is spread across multiple windows because of the overlap. Separately dealing with impulse and Gaussian noise therefore seems a better approach.

2.4 Feature selection

Research has shown that a key ingredient in increasing the performance of motor imagery BCI is the selection of subject-specific frequency bands **[35]**. Several approaches have been used for this purpose, such as distinction sensitive learning vector quantization (DSLVQ) **[49]**, particle swarm optimization **[62]** and, more recently, filter bank CSP methods based on the maximization of mutual information between features and class labels **[63, 64]**. The latter approach is adopted here, as it was proven to be very effective in extracting discriminative spatio-spectral features, and was also part of the winning algorithm of the BCI Competition on this dataset **[61]**.

The main idea is to apply CSP on several frequency bands and select the most relevant features based on the maximum mutual information criterion. There are, however, several open challenges which need to be addressed. The first issue is finding the optimum combination of spatial filters for each frequency band. An exhaustive search is not feasible, as even when only the first and last three filters are kept, there are already 63 possible combinations that can be evaluated. The problem gets even worse when searching for an optimum combination of frequency bands. For example, Zhang et al **[64]** propose keeping only the first and last two filters for each band and determining the optimal pair, thus reducing the search space to six combinations per frequency band. Only the band associated with the most informative subset is kept. Clearly, such a procedure is not optimal. Because CSP is highly sensitive to noise and artifacts, it is possible for the relevant filters to lie more towards the center of the projection matrix, thus not being part of the small subset considered. For each band the best subset might be formed by different number of filters. Most importantly, considering more than one band is intuitively advantageous, but the best configuration is not necessarily given by the most informative bands, both because of correlations and of the fact that less informative bands might give complementary information.

To overcome the issues described above, namely selecting the optimum configuration of features from a large set of possibilities, we adopt the maximum relevance minimum redundancy (mRMR) criterion proposed by Peng et al **[54]**. The mRMR algorithm maximizes the mutual information between the selected features and the desired output (max relevance), while minimizing the mutual information between the selected features (min redundancy).

The purpose of feature selection methods based on maximizing mutual information is to find a feature subset S with m features $\{x_i\}$, which jointly have the largest dependency on the target class c. This scheme, called max-dependency, takes the following form:

$$\max D(S,c), D = I(\{x_i, i = 1, \dots m\}; c) \tag{10}$$

The max-dependency criterion is hard to implement because of the difficulty of evaluating a very large number of possible subsets, theoretically $\sum_{k=1}^{M} \binom{M}{k}$, where M is the total number of features. An alternative is to find a subset based on the maximal relevance criterion, which approximates $D(S,c)$ with the mean value of all mutual information values between individual feature x_i and class c:

$$\max D(S,c), D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c) \tag{11}$$

It is likely that the features selected according to the max relevance criterion have rich redundancy, thus the minimal redundancy condition can be added to select mutually exclusive features:

$$\min R(S), R = \frac{1}{|S^2|} \sum_{x_i, x_j \in S} I(x_i; x_j) \tag{12}$$

The criterion which combines the above two constraints is called minimal redundancy maximal relevance. The operator $\Phi(D,R)$ is defined to simultaneously optimize D and R:

$$\max \Phi(D,R), \Phi = D - R \tag{13}$$

The authors also provide a Matlab-ready implementation[3] of the mRMR algorithm, which is used in this work. The mRMR feature selection method is applied for determining both the optimal combination of spatial filters for each frequency band and the optimum subset of frequency bands. One complication is that mRMR requires specifying the desired number of features. Thus, we combine mRMR with the maximal dependency criterion by determining subsets of different number of features and selecting the one which maximizes the mutual information with the desired output. The mutual information is estimated with the Information Theoretical Estimator Matlab toolbox developed by Zoltán Szabó **[65, 66]**. The estimation of multivariate mutual information is computationally demanding, therefore the number of samples is reduced by averaging feature vectors in overlapped windows with a duration of one second. Because of the prevalence of the NC state in the data, we use different overlaps for IC trials and NC segments, in the attempt to balance the data. The overlap of IC trials is of 0.5 seconds, and 0.2 seconds in the case of NC segments.

In total, three sets of features are extracted for each subject: two sets which optimally separate each of the two motor tasks from the NC state and one set which is optimal in discriminating between the two IC states. In all situations, the first and last five spatial filters are

[3] http://www.mathworks.com/matlabcentral/fileexchange/14608

kept for each frequency band. For IC/NC discrimination, we determine the optimal subset with at most three filters and the configuration which maximizes mutual information is kept for each band. For discrimination between the two IC states, we analyze subsets with at most six spatial filters for each band. For each frequency band, CSP is applied on multiple window configurations with durations of 1.5, 2, 2.5, 3 and 3.5 seconds and variable positions within the trial. The motivation behind this is that the temporal dynamics of the two motor imagery tasks might be different and ERD might occur at different moments. In total, 22 window configurations are tested for each frequency band and for each window position the optimal subset according to the mRMR scheme is kept. For each band, the most informative out of the 22 resulting subsets is selected. Finally, a maximum of 20 filters are selected from all the resulting subsets for determining the optimal configuration of frequency bands.

The fact that the feature vectors are averaged in overlapping sliding windows gives rise to the possibility of determining the most discriminative window position within a trial. Therefore, after selecting the optimal subset of features, we determine the most informative window position for discriminating between the NC state and each of the two motor tasks, as well as the optimal position for discriminating between the two IC states.

2.5 Classification

2.5.1 Classification schemes

For classification, there are three designs which can be used for self-paced BCI. The most common approach consists of two detectors: the first detects the presence of motor imagery and the second assigns the detected movement to one of the possible IC states. This method is illustrated in Figure 15. The advantage of this design is that different features and/or algorithms can be used for the two detectors, potentially improving performance. One possible drawback is that all IC states are considered as one class, which may not be appropriate for some classifiers because of the different distributions of IC states in feature space (see Figure 10, page 29).

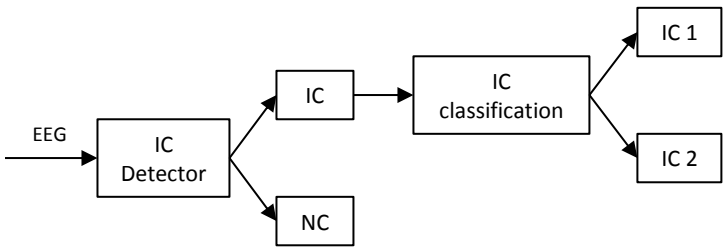

Figure 15: Two-stage detection and classification of IC states

The second possibility is to directly classify each sample as belonging either to NC or one of the possible IC states. There are two possibilities for this approach. The first is to train two classifiers for the detection of each IC state. This scheme, illustrated in Figure 16, might prove to be advantageous as different and optimally discriminative features can be used for the detection of each IC state.

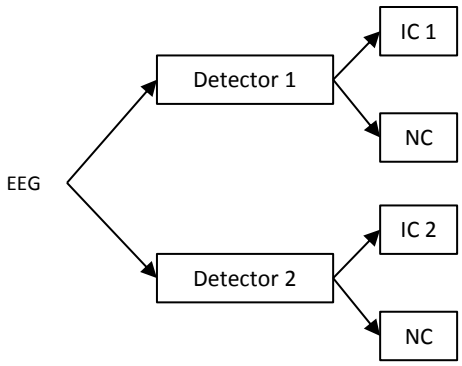

Figure 16: Separate detection of each IC state

The final possibility is to perform direct three-state classification. Modeling each class separately can lead to a better discriminative ability but the features and algorithms suited for NC/IC discrimination might not be as appropriate for the classification of different IC states.

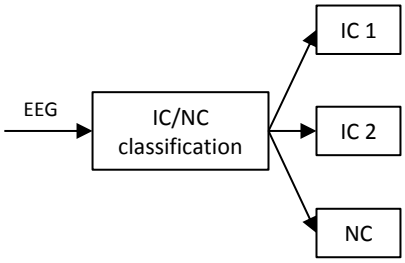

Figure 17: Direct three-state classification

For each of these designs, linear and non-linear classifiers will be employed. Linear discriminant analysis forms the basis of many synchronized BCIs and was successfully used for self-paced variants as well **[36, 49, 51, 52]**. For non-linear classification, quadratic discriminant analysis (QDA) and support vector machines with Gaussian kernels will be used.

2.5.2 Classification algorithms

Discriminant analysis

Discriminant analysis is based on the Bayes criterion of assigning to an unknown sample the class label which is most likely given the posterior and prior probabilities $p(x|C_i)$ and $P(C_i)$, where C_i is the i^{th} class. A so-called discriminant function is calculated for each class, and takes the form **[67]**

$$g_i(x) = \log p(x|C_i) + \log P(C_i) \quad (14)$$

For classification, the discriminant function of each class is calculated for an unknown sample, and the class with the highest value of the discriminant is assigned to the unknown sample. Assuming a normal distribution of the posterior probabilities, after some manipulation $g_i(x)$ can be written as

$$g_i(x) = -\frac{1}{2}\log|S_i| - \frac{1}{2}(x - m_i)^T S_i^{-1}(x - m_i) + \log \hat{P}(C_i) \tag{15}$$

where m_i, S_i and $\hat{P}(C_i)$ are the maximum-likelihood estimators of the mean, covariance and prior probability of class *i*, respectively. Note that (15) can also be written in quadratic form

$$g_i(x) = x^T W_i x + w_i^T x + w_{i0} \tag{16}$$

where

$$W_i = -\frac{1}{2}S_i^{-1}$$

$$w_i = S_i^{-1} m_i$$

$$w_{i0} = -\frac{1}{2}m_i^T S_i^{-1} m_i - \frac{1}{2}\log|S_i| + \log \hat{P}(C_i)$$

Equation (16) defines a quadratic discriminant which forms the basis of quadratic discriminant analysis (QDA). The separating hyperplanes thus obtained are conic sections, i.e. lines, circles, ellipses, parabolas or hyperbolas. Note that in the case of equal prior probabilities, $\log \hat{P}(C_i)$ can be dropped as it would be common to all classes.

A simplification of (16) is to assume a common covariance matrix for all classes, thus eliminating the quadratic term. The decision boundaries are hence linear and $g_i(x)$ now defines a linear discriminant classifier

$$g_i(x) = w_i^T x + w_{i0} \tag{17}$$

with

$$w_i = S^{-1} m_i$$

$$w_{i0} = -\frac{1}{2}m_i^T S^{-1} m_i + \log \hat{P}(C_i)$$

Equation (17) is the basis of linear discriminant analysis (LDA), which along with CSP forms the gold standard in classification for synchronized motor imagery-based BCIs. Because CSP is only applicable for two classes, and because traditional motor imagery experiments involved the discrimination of only two motor tasks, a common practice in BCI is to subtract the discriminant functions of the two classes and assume equal priors, resulting in a single discriminant function

$$f(x) = w^T x + w_0 \tag{18}$$

with

$$w = S^{-1}(m_1 - m_2)$$

$$w_0 = -\frac{1}{2}(m_1 - m_2)^T S^{-1}(m_1 + m_2)$$

The common practice in binary classification is to assign class labels according to the sign of $f(x)$, which indicates on which side of the hyperplane the unknown sample is. This approach is often misleading though, as it can give the false impression that LDA is strictly a binary classifier and it has to be "extended" for multi-class scenarios through one-VS-one or one-VS-rest techniques. This is unfortunately the case even in well-respected and peer-reviewed papers **[68]**.

LDA is often preferred in BCI research due to its simplicity, which also makes it a popular choice for either supervised or unsupervised adaptation **[69]**. However, LDA classification typically suffers from several drawbacks. The normality assumption may not hold for certain conditions and the need for estimating covariance matrices and inverting them means that for N-dimensional feature space at least N+1 samples are needed for the covariance matrix to have full rank. Also, in many cases the assumption of a shared covariance matrix may not hold either.

Support vector machine (SVM)

Similar to discriminant analysis, support vector machine classifiers also determine an optimal separating hyperplane. However, in the case of SVM, the obtained hyperplane is the one that maximizes the margins, i.e. the distance to the nearest training points (see Figure 18). Maximizing the margins is known to increase generalization capability **[70]**. To accommodate for outliers and some level of misclassification on the training set, a regularization parameter $C > 0$ is introduced in the optimization function of SVM.

Non-linear classification can be performed by first mapping the data into a higher (maybe infinite) dimensional space by a kernel function and determining a linear separating hyperplane in this new space. The most common kernel is the Gaussian or radial basis function (RBF) kernel:

$$K(x, y) = \exp\left(\frac{-||x - y||^2}{2\sigma^2}\right) \tag{19}$$

Thus, for a Gaussian SVM there are two parameters that need to be specified, namely the regularization parameter C and the width of the Gaussian kernel, σ. We determine the two parameters through a grid search in cross-validation with exponentially growing sequences of C and σ. More specifically, $C \in \{2^{-1}, 2^0 \dots 2^3\}$ and $\sigma \in \{2^{-4}, 2^{-3} \dots 2^1\}$. Because of the unbalanced nature of the data, we do not choose the parameters which maximize cross-validation accuracy, as is normally the case, but rather by the ones that maximize the difference between event-based TPR and sample-based FPR. For the definition of these metrics, the reader is referred to Chapter II, section 3.

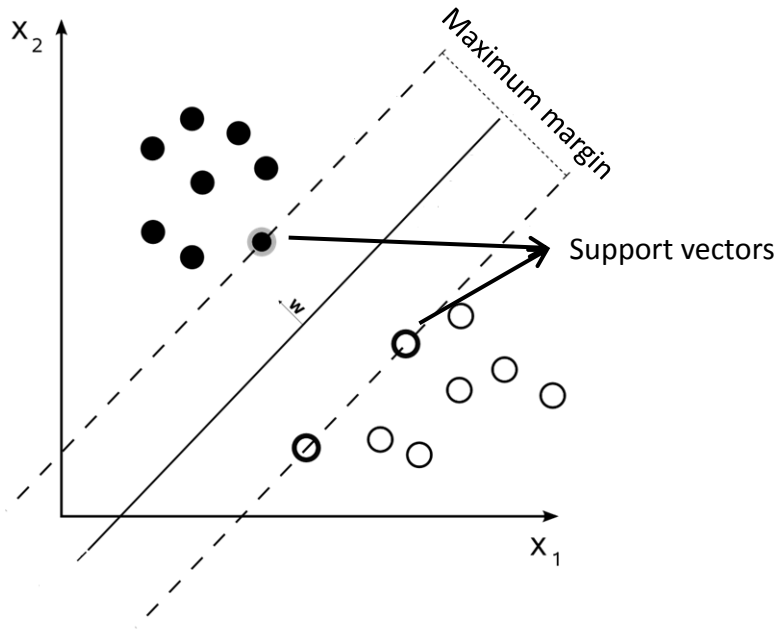

Figure 18: SVMs determine the hyperplane which maximizes the distance to the nearest training points

Linear kernels were not used here, as LDA is already known to offer good performance. Linear SVMs truly shine compared to LDA in the case of small sample size when the covariance estimator would not be trustworthy, but this was not the case here. Therefore, only Gaussian SVMs are used.

SVMs such as the ones described above and illustrated in Figure 18 are called soft-margin SVMs. Another, less common type is the one-class SVM. These are trained only with positive instances and they work by fitting a tight hyper-sphere to include most, but not all of the training examples, in order to avoid overfitting. Therefore, instead of fine-tuning the misclassification cost, one needs to determine ν, the regularization term for errors, which can take values between 0 and 1. The selection of ν is performed similarly to soft-margin SVMs, with $\nu \in \{0.1, 0.2 \ldots 0.9\}$.

One-class SVMs were tentatively employed here for two reasons. First, they do not need specific examples of the negative class, which makes them an ideal candidate for detection and might circumvent the issue of modeling the NC state. And second, the smaller number of samples makes for a much faster training and model selection.

The LibSVM toolbox **[71]** is used for the practical implementation of all SVMs.

2.6 Regression

A conceptually different approach for the translation algorithm would be to treat self-paced operation as a regression problem instead of a classification one. In essence, we consider the whole training data as a continuous N-dimensional input signal which needs to be mapped to a one-dimensional control signal, representing the desired output. The targets for regression are therefore the labels of each sample.

Wiener filter

The mapping of the multivariate input to the one-dimensional control signal can be done by either linear or non-linear functions. We choose a linear transformation as linear methods proved to be very effective in BCI **[72, 73]**.

Consider a *d*-dimensional input signal $x \in \mathbb{R}^{d \times N}$ and the desired output signal $y \in \mathbb{R}^{1 \times N}$. We are looking for a transform *W* that minimizes some error criterion between the predictions $\hat{y}$ and the true *y*:

$$\hat{y} = W^T x \tag{20}$$

The projection which minimizes the mean squared error is given by the Wiener solution **[74, 75]** and is based on estimates of the autocorrelation of the data and the cross-correlation between data and desired output:

$$W = E[x^T x]^{-1} E[x^T y] \tag{21}$$

Past samples can be used for the current prediction by extending the Wiener filter to a finite impulse response (FIR) model. Such a topology is illustrated in Figure 19, where the current sample *x(n)* and the preceding *M* samples are used to calculate $\hat{y}(n)$, our estimate of the true system output at time *n*.

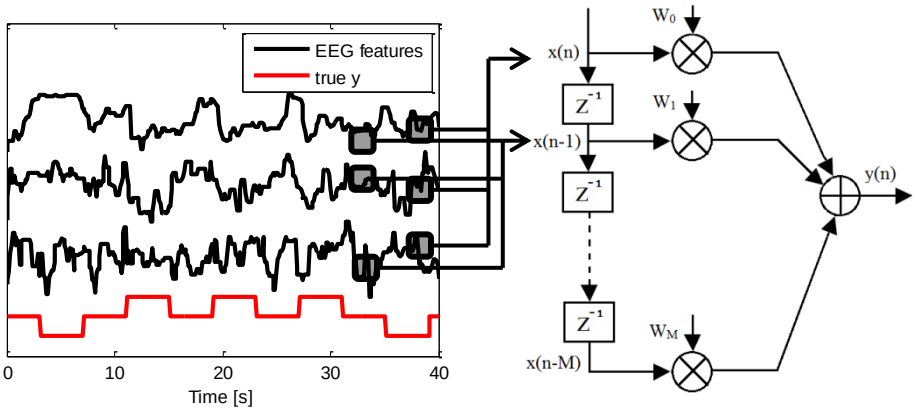

Figure 19: Finite impulse response topology of order M

2.7 Genetic algorithms

Genetic algorithms are optimization methods which mimic the process of natural evolution and are commonly used in situations where the search space is very large. In this work, they are used for linearly combining the outputs of several classifiers and for determining new regression targets.

The algorithm starts with a set of randomly generated candidate solutions to the optimization problem. A fitness function measures the fitness of candidate solutions, also called individuals. The best individuals are recombined such that their best traits are inherited by their "children", which will form the new generation. The whole procedure is then repeated until some criterion is satisfied. To reduce the risk of finding a local minimum, random mutations are also possible.

3 Experimental procedure

For a proper evaluation of performance, an important step is to determine the influence of the various parameters, such as the size of the median and moving average windows. For classification in particular, the number of parameters which can be fine-tuned is even larger, as a choice needs to be made regarding the positions within trials with which to train the classifier. Due to the large number of design choices, our approach is to set default values, determine a promising model, and only then fine-tune these parameters. For the two moving windows a size of 11 samples was chosen (or 1.1 seconds as band power is calculated in windows of 100ms) and classifiers are initially trained with the features within the most informative one-second window for each class, as determined by the feature selection procedure. This also implies that not all samples are used directly for classification, but are averaged in sliding windows of one second. The step of the sliding window is now of only one sample though, otherwise we risk skipping the portions on which the classifier was actually trained. More details will be given for each specific implementation, where needed.

Chapter IV
Experiments and results

1 Feature selection

We start by presenting the results of the feature selection procedure based on the combination of mRMR and the maximal dependency criterion. As explained in the previous chapter, optimal subsets of features are determined for discriminating both between the two IC states, as well as between each motor task and the NC state.

1.1 NC/IC discrimination

For each mental task, the feature selection procedure determines an optimal subset which best separates it from the NC state. The number of selected features is thus not constrained to be the same for both classes or for each frequency band, but it is limited to a maximum number of 20 filters. The results of the feature selection procedure are summarized in Table 1.

Table 1: The number of spatial filters selected for each subject, class and frequency band for NC/IC discrimination

Subject	Class	Central frequency [Hz] / Number of spatial filters												Total
		8	9	10	11	12	16	19	20	21	22	23	30	
a	Left				1	1				1	1	1		**5**
	Foot	1					2							**3**
b	Left	1				1								**2**
	Right	1				1								**2**
f	Left		2	1	2	1								**6**
	Foot			1	2									**3**
g	Left			1				1	1					**3**
	Right		1	1				1	1				1	**5**

Most of the selected features belong to the mu band, between 8 and 12 Hz. With the notable exceptions of the *foot* class of subject *a* and the *right* class of subject *g*, the rest of the features are from the 19-23 Hz band, presumably reflecting modulations of the beta rhythm. Whether they truly belong to the beta band or are simply harmonics of mu is still generally an open question in motor imagery BCI.

In Figure 20 the normalized patterns of the most discriminative spatial filters are presented, along with the corresponding frequency bands, as determined by the feature selection procedure. For reasons of space, only the first three filters are shown.

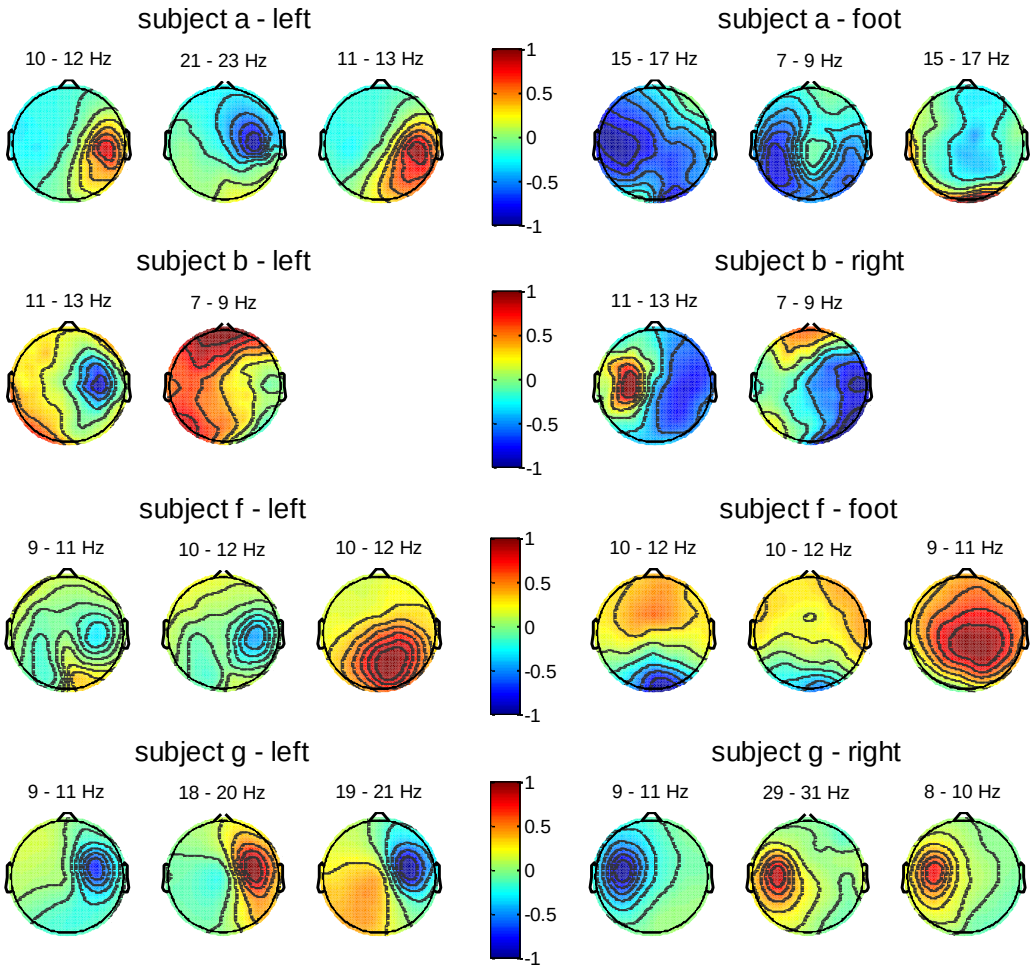

Figure 20: The patterns of the three most relevant spatial filters for IC/NC discrimination extracted for each subject and the corresponding frequency bands

Most contributions indeed seem to originate from the motor cortex, but also from parietal and occipital regions in some cases. Some of the patterns do not appear neurophysiologically plausible, especially some of the ones associated with the *foot* class. One encouraging result of the feature selection procedure is that a relatively small number of filters were selected from the maximum of 20, suggesting that correlations between features were indeed minimized. In Figure 21 the mutual information between features and class labels is plotted as a function of the number of features. The maximum limit of 20 features seems a good choice, as generally the mutual information monotonically decreases after about 15 features.

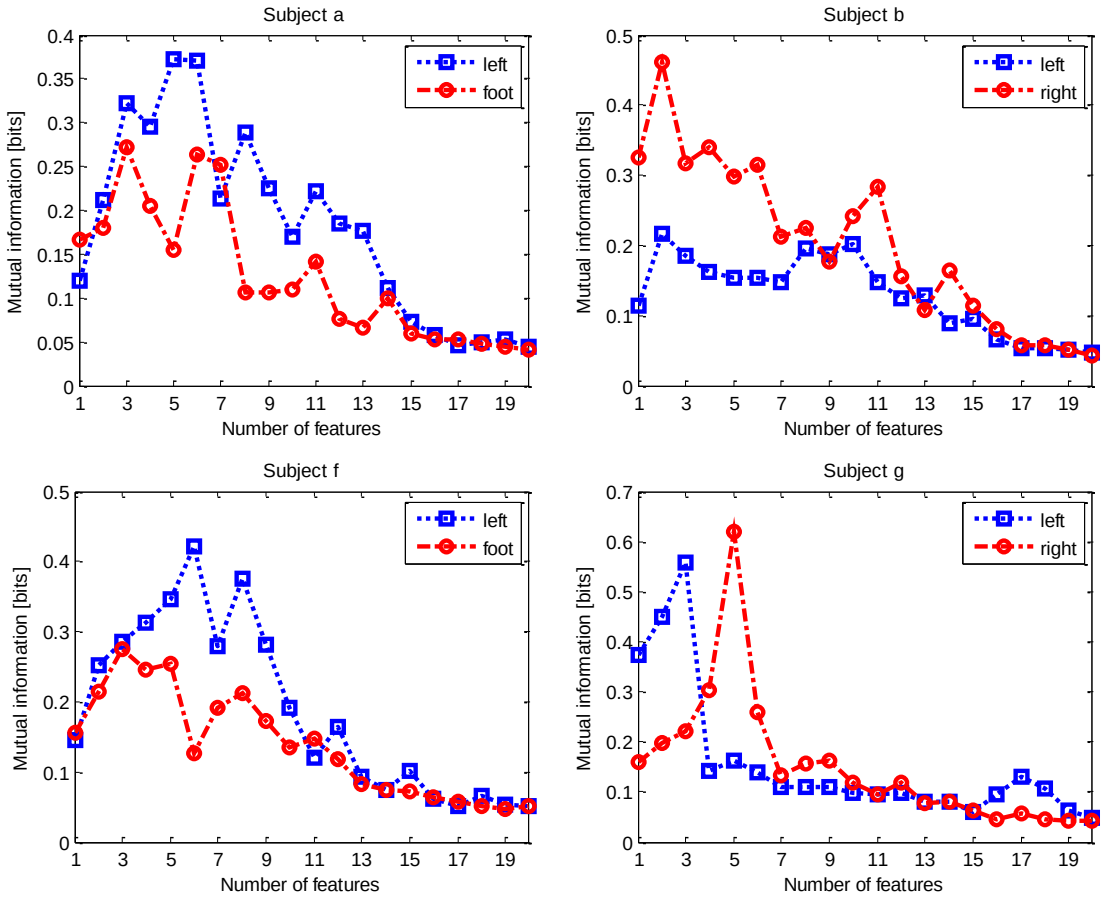

Figure 21: Mutual information between extracted features and the desired output for IC/NC discrimination with respect to the number of features

As discussed in section 2.4 of the previous chapter, the feature vectors are averaged in overlapped sliding windows with durations of one second. This opens up the possibility of evaluating the separability of IC/NC states as a function of time, i.e. the central position of the one second window within the trial. These results are plotted in Figure 22. In general the separability of IC/NC states reaches a maximum around two seconds, half of the duration of training trials. The keen reader might observe a seemingly strange phenomenon taking place: the ratios of separability with respect to the mutual information do not seem to hold between motor tasks or even subjects when compared to the results of Figure 21. For instance, subject *g* no longer presents the largest mutual information, and for subject *f* the *foot* class now seems better separated from NC than the *left* class. The explanation is that in Figure 21 all seven window positions are evaluated, whereas in Figure 22 the separability of a single position is tested.

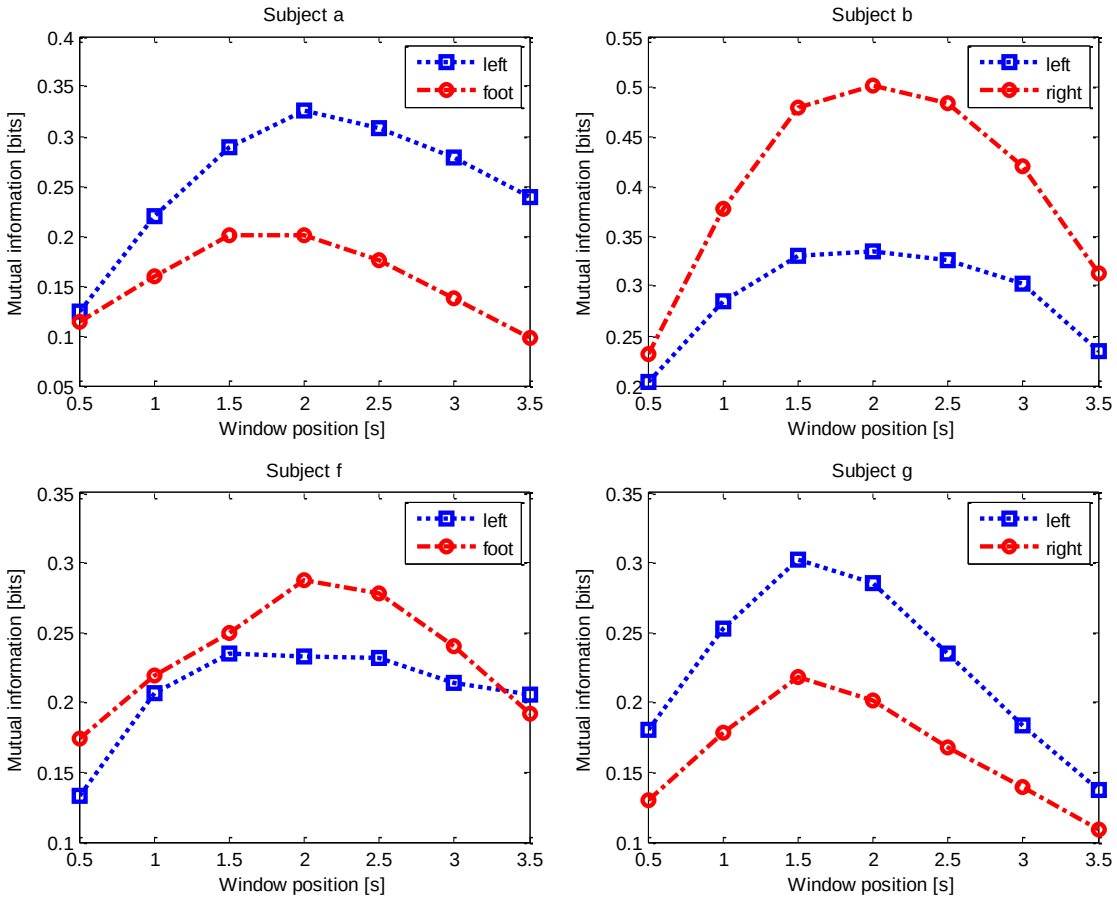

Figure 22: Mutual information between extracted features and desired output for IC/NC discrimination with respect to the window position within the trial

1.2 IC discrimination

We now proceed to the results of feature selection for discrimination of the two IC states. Similarly to the previous section, the number of spatial filters selected for each subject and frequency band is summarized in Table 2.

Table 2: The number of spatial filters selected for each subject, class and frequency band for IC discrimination

Subject	Central frequency [Hz] / Number of spatial filters															Total
	9	10	11	12	13	14	18	19	20	22	23	24	27	29	30	
a		2	2	2						1	1	1				**9**
b		1	2	2	1								1			**7**
f	2	2	2	1	1	1										**9**
g	1	1					1	1	1					1	1	**7**

A considerably larger number of spatial filters are selected for distinction of the two IC states compared to IC/NC discrimination. This might be related to numerical issues in computation

of the mutual information due to the much larger number of samples considered in IC/NC discrimination.

The normalized patterns of the seven most influential spatial filters for discriminating between IC states are shown in Figure 23. As was the case with IC/NC discrimination, some are not neurophysiologically plausible and again in the case of subject *f*, considerable contribution originates from the visual cortex. A rather unexpected result is that in the case of subject *g*, no spatial filters associated with class *right* were selected. This is probably due to between-class differences in the magnitude of band power modulations. In binary classification, if one motor task produces considerably stronger EEG amplitude modulations than the other, it could be that better discrimination can be obtained by measuring EEG power solely over the area corresponding to the dominant class. In this case, ERD over the right hemisphere would indicate movements of the left hand and the absence of ERD would indicate movements of the right hand.

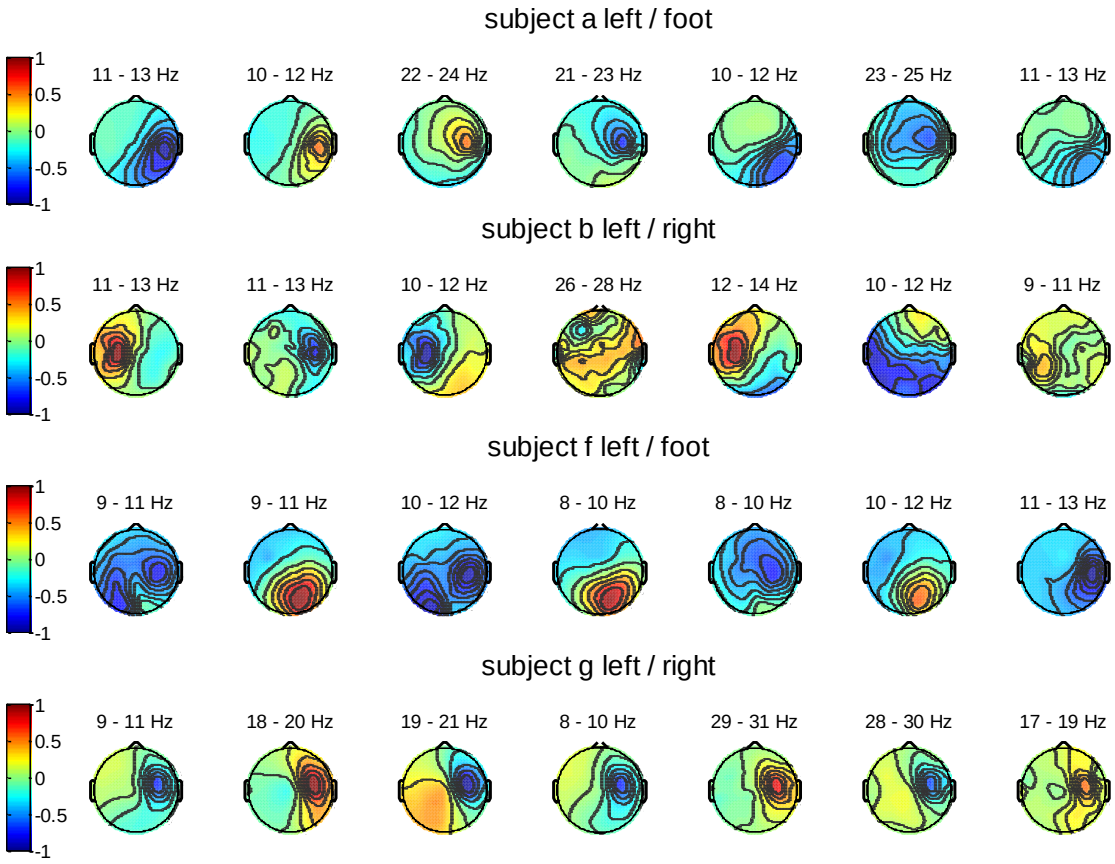

Figure 23: The patterns of the seven most relevant spatial filters extracted for IC discrimination and the corresponding frequency bands

The separation between IC states in terms of mutual information is plotted in Figure 24 with respect to the number of features. While the inverse relation between mutual information and

the size of the selected feature subset is not as pronounced compared to IC/NC discrimination, a considerable decrease can still be observed for subsets of more than 11 features.

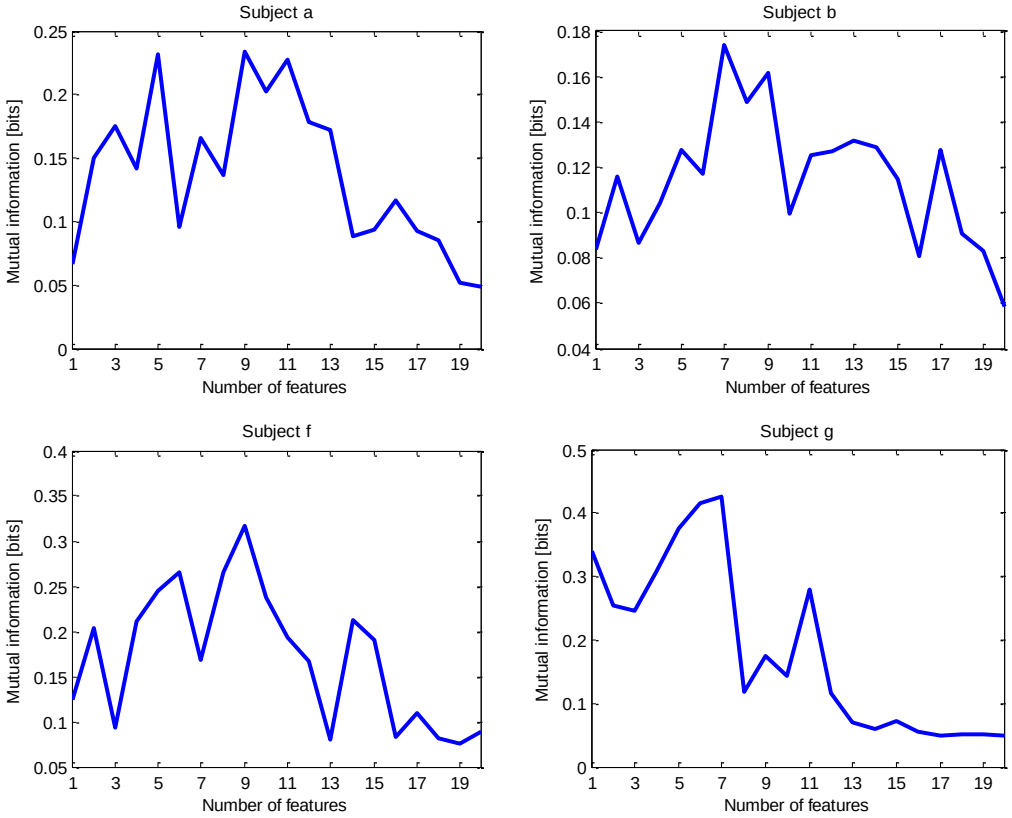

Figure 24: Mutual information between extracted features and the desired output for IC discrimination with respect to the number of features

The separation between the two IC states across the duration of a trial is plotted in Figure 25. The positions which offer maximum separation are similar to those determined for IC/NC discrimination in the cases of subjects *a* and *b*. For subject *f* the optimal position shifted towards the beginning of the trial, while for subject *g* it shifted towards the middle. Again it becomes apparent that there are considerable differences in the maximum values of mutual information between Figure 24 and Figure 25, which might be an indication of large variability of band-power levels across the duration of a trial.

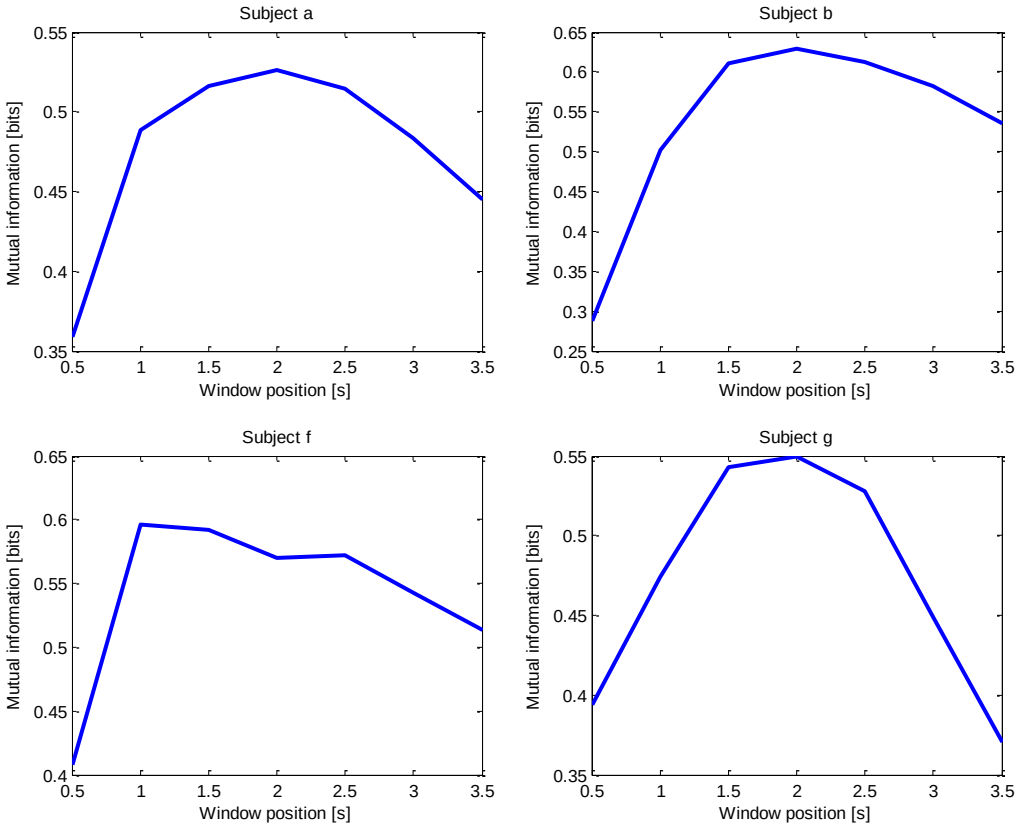

Figure 25: Mutual information between extracted features and desired output for IC discrimination with respect to the window position within the trial

2 Classification

As per the experimental procedure explained in the previous chapter, all classifiers and configurations thereof are first tested with default parameters in cross-validation. The influence of these parameters will be investigated on the best performing model, and the best configuration will then be applied on the evaluation data. The performance in cross-validation is measured in terms of the event-based TPR, while keeping the sample-based FPR as close to 1% as possible. For simplicity of notation, in the following we will omit the distinctions "sample-" and "event-based", and refer to the two metrics as TP and FP rates, or simply TPR and FPR.

2.1 Cross-validation

Judging from the timing of the training labels, it seems that the calibration data has been recorded over 14 runs with breaks of approximately 24 seconds between them. We therefore divide the training set in 14 subsets and perform 14-fold cross-validation, by training on 13 sets and evaluating the model on the remaining one. The above mentioned breaks between runs are considered part of the NC state.

2.1.1 Two-stage classification

For two-stage classification, what we are mostly interested in is evaluating the performance of the IC/NC detector (DET1), responsible for distinguishing IC states of any kind from the ongoing EEG. The feature vectors used for training DET1 are formed by concatenating the IC/NC discriminative feature sets of both classes.

The TP rates of IC detection for the four classifiers are given in Table 3, at the corresponding FP rate of 1%. A rather unexpected result is the incredibly low performance of soft-margin SVM. Initially we considered this to be an effect of the unbalanced data. As SVMs minimize the total misclassification cost, they act like majority classifiers for unbalanced training examples. One solution is to assign different misclassification costs for the two classes based on the number of training instances **[76]**, such that:

$$\frac{C_+}{C_-} = \frac{n_-}{n_+} \tag{22}$$

where C is the misclassification cost, n the number of samples and +/– represent the positive (IC) and negative (NC) classes, respectively.

Table 3: TP rates of IC detection for different classifiers at 1% FP rate. Both IC states are treated as one class

Classifier / Subject	LDA	QDA	SVM	1C-SVM
a	10.5	30	3.5	14.5
b	45	37	4	36
f	11	24	1	17.5
g	4	22.5	4.5	5.5
Average	**17.63**	**28.38**	**3.25**	**18.38**

The SVM evaluation was repeated with the different costs but results were even worse, as FP rates were actually higher than TP rates for all subjects. Because of unsatisfactory performance, the need to perform model selection and a very long training time, SVMs (at least the standard soft-margin variant) will not be considered in further evaluations.

One-class SVMs fared considerably better than their soft-margin counterparts, but the clear winner is the quadratic discriminant. Except for the soft-margin SVM, the linear discriminant had the lowest overall performance, most likely due to the assumption of identical covariance matrices and the consideration of both IC states as one class. It is interesting to note that while LDA has low detection rates for three out of four subjects, for subject *b* it presents very good performance. Not only is the TPR for subject *b* four times larger than that of the next best subject, but it is considerably larger than what is achieved by the non-linear classifiers for the same subject.

2.1.2 Separate detection of each IC state

Because we have two classifiers, there are two possibilities for interpreting their output. The first one is to subtract their scores, thus resulting in a one-dimensional control signal, which can then be thresholded for both IC states. We call this "differential mode". This is expected to give relatively low errors in classifying between the two motor tasks, but might reduce the overall detection rate. The second possibility is to separately threshold the scores of the classifiers, using them in "parallel mode". While FP rates are still expected to be low, false classifications between IC states are more likely.

Differential mode

The results for the differential mode are plotted in Figure 26.

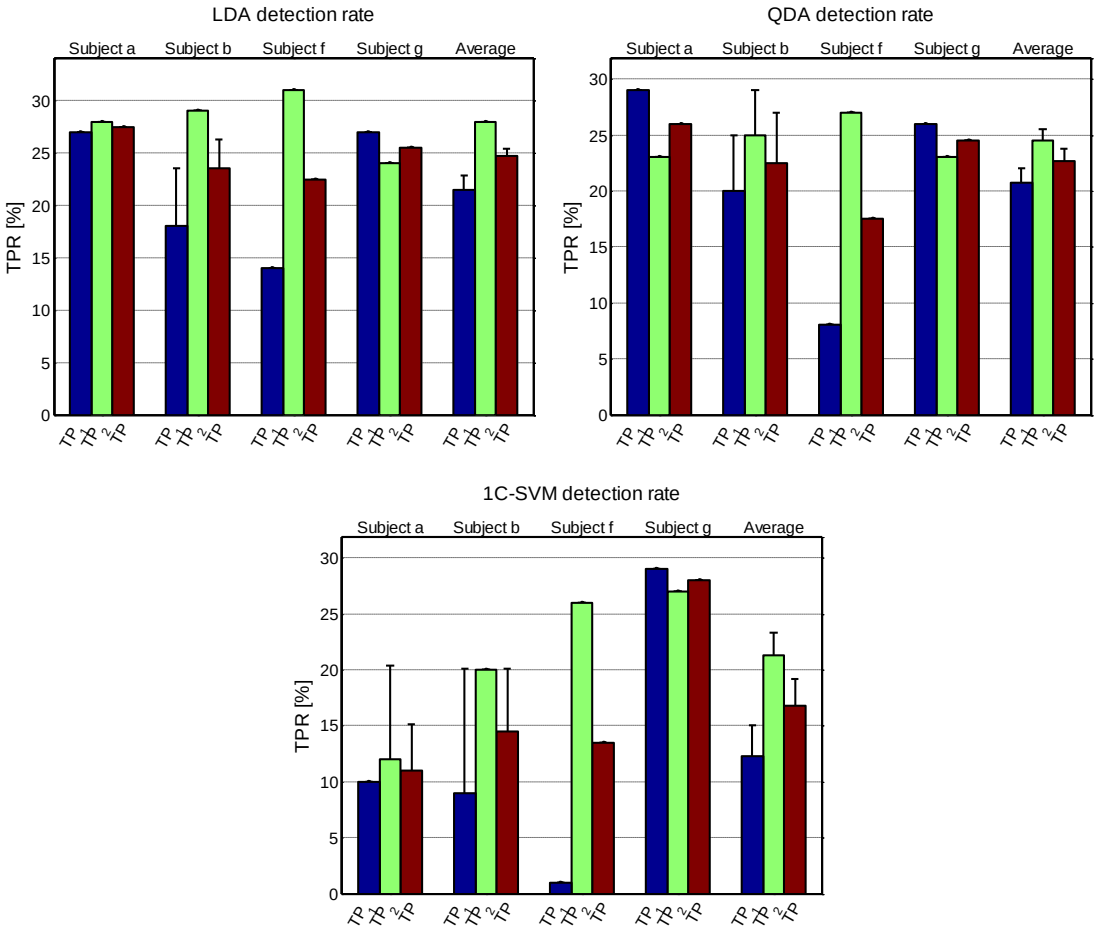

Figure 26: LDA, QDA and one-class SVM detection rates in differential mode. Error bars represent the percentage of misclassified, yet detected IC states. TP_i represents the true positive rate of IC state *i*, TP is the average true positive rate of IC detection. All TP rates correspond to an FP rate of 1% or lower.

With the exception of subject *g*, where one-class SVM performed the best, LDA presents the highest overall detection rate, although not substantially different than that of QDA. It is possible that the linear separating hyperplane provides a better generalization capability. One surprising result is the low detection rate of the *left* class for subject *f*, with which all classifiers struggled, especially the SVM, which has a TP rate of only 1%. Another interesting observation is that while the discriminant classifiers provide the best results for subject *a*, in the case of one-class SVM this subject has the lowest overall detection rate. The misclassification error between IC states is indeed low, of maximum 1 misclassified event. However, for low detection rates such as those of the one-class SVM, this can still translate in errors as large as 11%.

Parallel mode

The detection rates obtained in parallel mode are plotted in Figure 27.

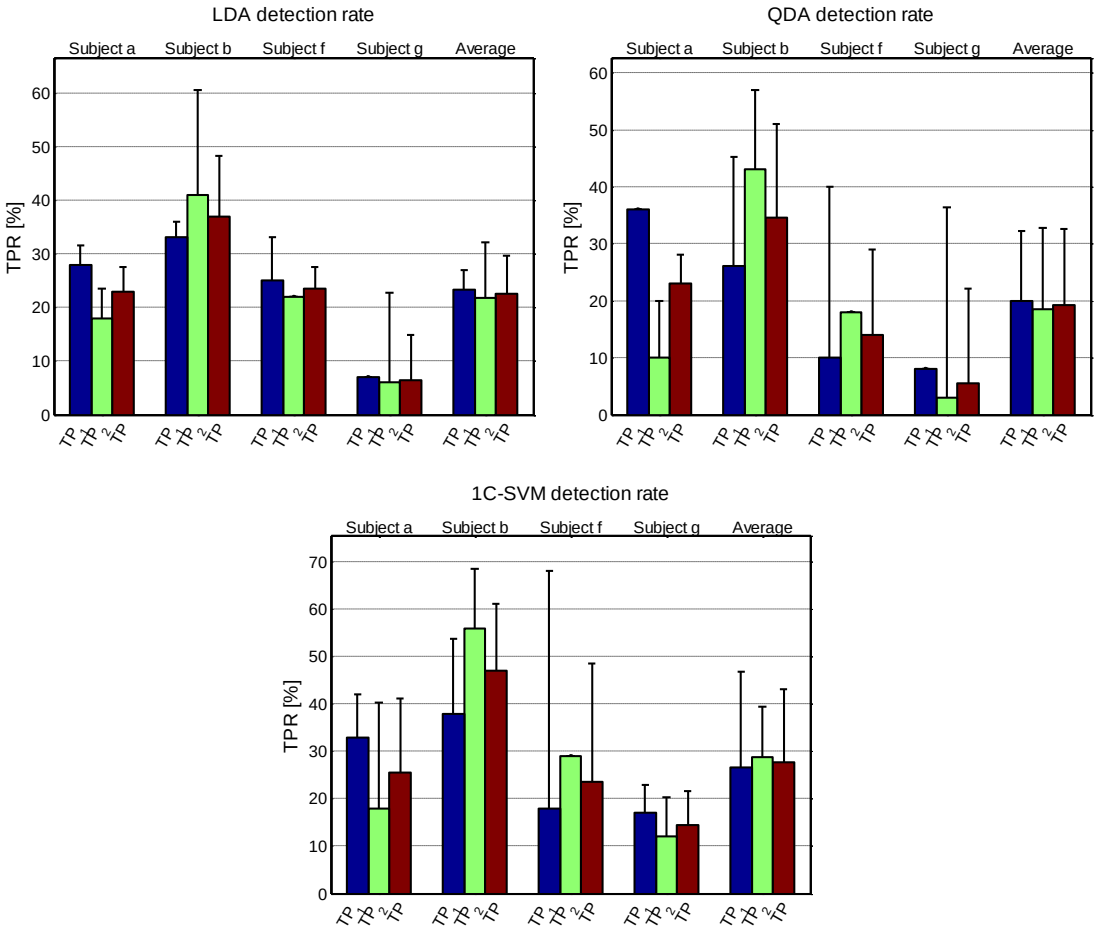

Figure 27: LDA, QDA and one-class SVM detection rates in parallel mode. Error bars represent the percentage of misclassified, yet detected IC states. TP_i represents the true positive rate of IC state *i*, TP is the average true positive rate of IC detection. All TP rates correspond to an FP rate of 1% or lower.

It becomes immediately apparent that misclassification errors between IC states are considerably larger than in differential mode, ranging between 0 and 9 wrongly classified events, and are especially prominent for the one-class SVM. In the case of class *left* of subject *f* this translates into chance-level accuracy. Nevertheless, we must emphasize that this is not necessarily an issue, as a second classifier can be used to discriminate between IC states once an event has been detected.

Interestingly, while all three classifiers provided good detection rates for subject *g* when used differentially, they all perform the worst for this subject in parallel mode. Only the one-class SVM provides TP rates higher than 10% for this subject. A large performance decrease is also observed for subject *a*, even though it is not as drastic. It seems that these differences are subject-specific and should not be generalized. In the case of subject *b*, all classifiers provide much better detection rates. The discriminant classifiers gain an improvement of roughly 50%, as TP rates increase from around 24% to 37% for LDA and 34.5% for QDA. Percentage-wise, most impressive is the 335% performance gain of the one-class SVM for subject *b*, from only 14% in differential mode to a respectable 47% in parallel. While for the other subjects the differences between the discriminant classifiers and the SVM are not as substantial, the one-class SVM seems to cope better with parallel mode compared to LDA and QDA.

2.1.3 Three-state classification

The final possibility is to directly perform three-state classification. Only the discriminant classifiers can be used for this approach and the results are plotted in Figure 28. Detection rates are now more balanced between subjects and IC states. It is encouraging that misclassification errors are low, of maximum 2 incorrect classifications for LDA and 3 for QDA, both in the case of subject *b*. Again, the linear discriminant outperforms the quadratic variant, even though not by a wide margin. Both classifiers provide the best performance in this mode compared to all other configurations, except for QDA, for which the maximum detection rate of 28% was obtained when both IC states were considered as one class.

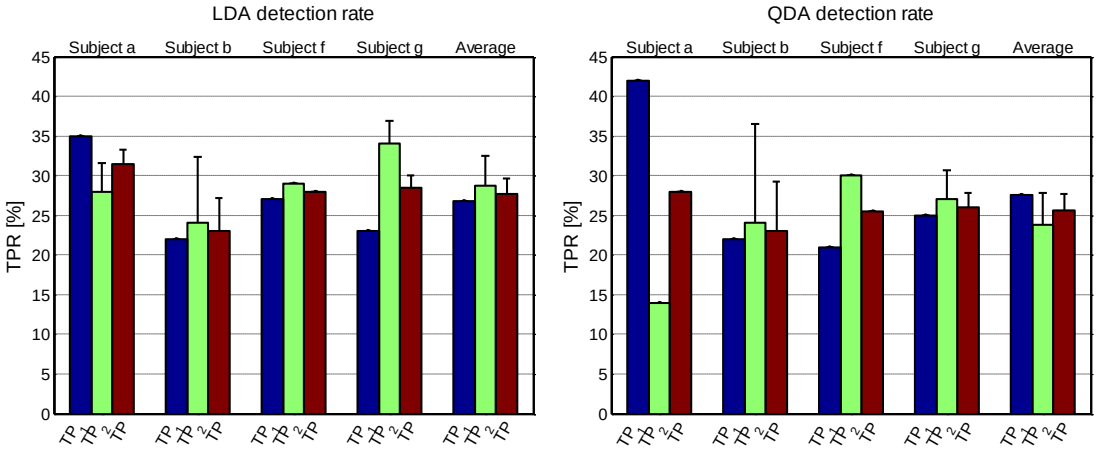

Figure 28: LDA and QDA detection rates for direct 3-state classification. Error bars represent the percentage of misclassified, yet detected IC states. TP_i represents the true positive rate of IC state *i*, TP is the average true positive rate of IC detection. All TP rates correspond to an FP rate of 1% or lower.

It should be noted that when three-state classification is performed with discriminant classifiers, three scores are obtained, one for each class. Theoretically, this also allows for the differential and parallel modes of interpreting the outputs. In this case, however, both scores are generated by the same classifier, with the same features, thus the results of the two modes are almost identical. In Figure 28 the differential score is presented.

2.1.4 Conclusion

It is now time to conclude these tests and choose an appropriate classifier design. In Figure 29 we present the average detection rates over all subjects for LDA, QDA and one-class SVM. Except for the SVM which cannot be employed for direct three-state classification, the results for all four modes of operation are presented: IC detection where both motor tasks are considered as one class, separate classification for each IC state in both differential and parallel modes and direct three-state classification for the discriminant classifiers.

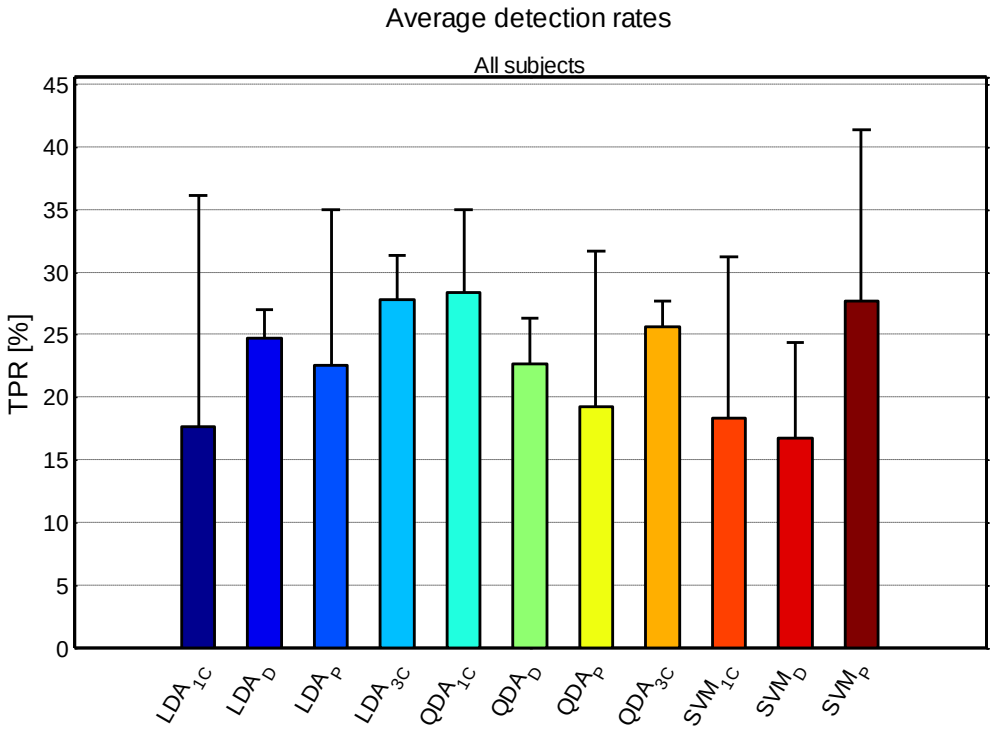

Figure 29: Average IC detection rates for all subjects and classifier designs, and their standard deviation between subjects. Subscripts 1C, D, P and 3C represent one class (IC/NC), differential, parallel and direct 3-state classification, respectively

For IC detection LDA performs the worst and also presents the largest standard deviation between subjects, due to the unexpectedly good performance obtained in the case of subject *b*, while QDA has the highest overall detection rate with relatively consistent performance across subjects. For separate detection of the two IC states, LDA offers the highest performance, followed by QDA and the SVM. Interestingly, both parametric classifiers perform better in differential mode than in parallel, and also have much more stable performance between subjects. The opposite is true

for the one-class SVM, which provides very good detection rates in parallel mode, but also large inter-subject variability. Regarding direct three-state classification, LDA outperforms QDA and, in fact, offers almost the same IC detection rate as one-class QDA and with more consistent results.

Based on these results, it seems that the best approach is to directly perform three-state classification. Detection rates are on par with the best results obtained by QDA in IC/NC discrimination, are more consistent between subjects, and misclassification errors between IC states are low. In fact, subject *b* is the only case where other configurations are advantageous. For this subject, the LDA-based IC/NC detector of two-stage classification gave an impressive average TP rate of 45%, and the one-class SVM in parallel mode gave an even larger average TP rate, of 47%. For the rest of the subjects direct three-state classification performs the best across all classifier designs. We therefore conclude that the most suitable approach is direct three-state classification with a linear discriminant classifier, and proceed with further investigations.

2.1.5 Influence of the feature processing pipeline

The next step is to determine how different parameters affect the performance of the three-state LDA classifier. We will start by analyzing the influence of the feature processing pipeline, which consists of the median filter, the log transform and the moving average window. In order to investigate the influence of these processing steps, a grid search over different sizes of the moving windows was performed, with and without the log-transform, and TP rates were calculated in cross-validation. The window durations under consideration are all odd and range from one sample (equivalent to not performing any moving average/median) to 19 samples (or 1.9 seconds for Δt = 100ms).

Recall that the classifier is trained with the average value of the samples within the most informative one-second window. This corresponds to an additional smoothing step, and this evaluation provides a good opportunity to check its influence and determine whether training the classifier with single samples would be better. Thus, the same grid search procedure is also applied when training the classifier directly with the samples within the window, instead of their average value. The results of this procedure are plotted in Figure 30 for all four subjects.

The most striking result from Figure 30 is the substantial performance benefit of directly training the classifier with individual samples rather than their average values. This increases true positive rates by 24%, 28%, 40% and 45 % for subjects *a*, *b*, *f*, and *g*, respectively. TP rate seems to be inversely proportional to the size of the moving windows, and this relation is more pronounced when training the classifier with individual samples. Not taking the logarithm decreases performance significantly. This was expected to be the case because log transforming the non-negative EEG band power values approximates a normal distribution. With no logarithm to attenuate impulse noise, the median window seems to be the important factor. This can best be seen in the data of subjects *b* and *f*, in which cases TP rate starts increasing only after the median window exceeds a certain threshold. For the other two subjects the median and moving average windows have more balanced influences. Nevertheless, data needs to be log-transformed in order to get good TP rates.

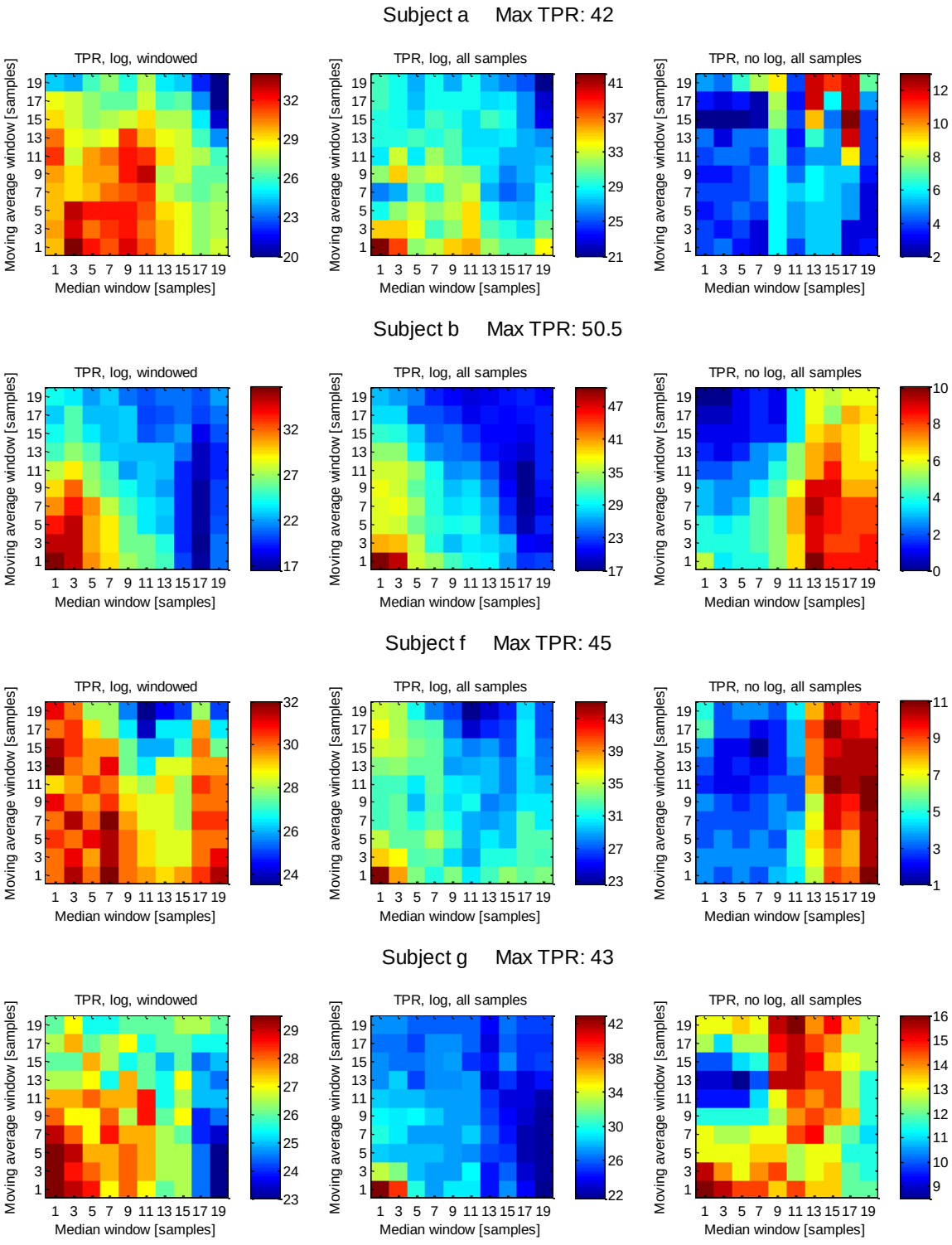

Figure 30: Influence of different parameters on detection rate. First column: classifier trained with average sample values, log transformed; second column: classifier trained with all samples from the window, log transformed; third column: all samples, no log. A window size of one sample is equivalent to not performing the respective operation. All TP rates correspond to an FP rate of 1% or lower.

For all subjects, the best results are found with no noise reduction applied, and when directly training the classifier with individual samples rather than their averaged values. This configuration will thus be used for further analysis.

2.1.6 Sensitivity VS hold time

The moving windows which were normally used are apparently not beneficial for the sensitivity of the classifier, evaluated on an event-by-event basis. However, it is very likely that reducing the amount of smoothing has a negative impact on hold time. Before proceeding with other attempts to increase true positive rates, we will first investigate how the hold time is affected by the new parameters.

In Figure 31 we analyze three configurations: the initial one, where the median and moving average windows have a length of 1.1 seconds and the classifier is trained with the average value of the samples (denoted by "sliding window"); one in which the same noise reduction is applied but the classifier is trained with individual samples; and the one which offers the largest TP rates, in which no noise reduction is applied and the classifier is also trained with individual samples. In general, the hold time increases with the extent of the smoothing operations, although there are no significant differences between applying the additional sliding window and directly using individual samples. For all subjects with the exception of subject *b*, roughly 50% of the detected events have a hold time of one second or longer with noise reduction applied. This is not stellar performance, but if we were to consider the first second of each trial as a transitory period, this would correspond to maintaining the desired IC state for about 33% of its duration. This is not very bad, especially compared to what is obtained when noise reduction is not performed, which leads to a hold time of only 0.2 seconds for all subjects in the vast majority of situations.

It seems that sensitivity and hold time are diverging objectives, and we are thus left with a choice between them. In practice, this might depend on the target application of the BCI. For the purpose of this study though, we choose to increase true positive rates as much as possible, even at the expense of decreased hold times. Based on these results, the BCI we are left with is more similar to event-driven than to state-driven BCIs.

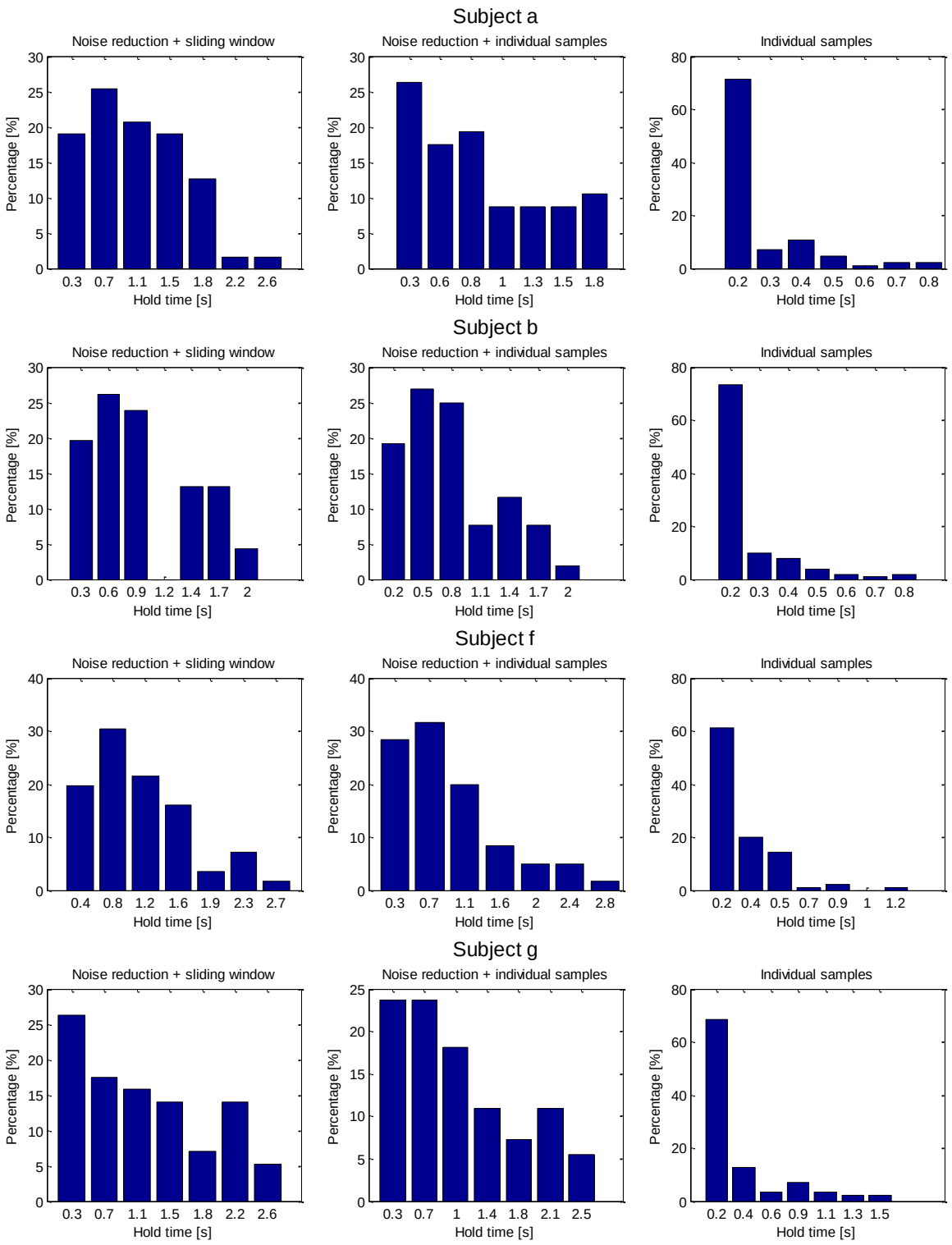

Figure 31: Influence of smoothing operations on hold time.

2.1.7 Exploring additional false positive rates

The procedure adopted until now was to determine a threshold on the cross-validation score such that the false positive rate is of maximum 1%, and calculate the resulting true positive rate. This is useful for comparing different approaches, but it would be useful to evaluate the performance at different false positive rates as well. True positive rates were calculated for false positive rates ranging between 1% and 10%, and the results are plotted in Figure 32.

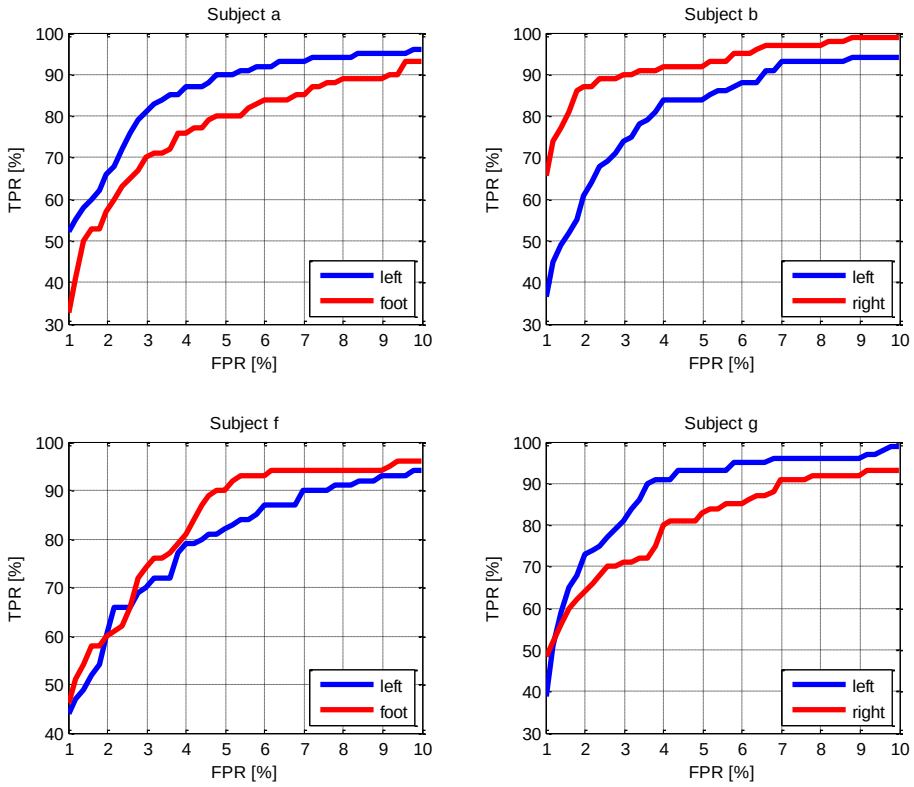

Figure 32: TP rates for FP rates ranging between 1% and 10%

Detection rates improve significantly when larger FP rates are allowed. For all subjects, the average true positive rate is above 90% when the FP rate is 8% or larger. For subject *f* there is a small range of FP rates, between 2% and 2.7%, in which the dominant class changes. This change in the dominant class is also found for subject *g*, in which case it appears that IC state *right* is more easily detectable only for the 1% FP rate initially considered. This balance between IC states is in accordance to the results of the feature selection procedure from Figure 22. The very good performance that can be achieved with more relaxed thresholds motivates us to search for methods which could decrease the false positive rate while maintaining large true positive rates.

2.1.8 Dwell and refractory periods

It may be the case that larger FP rates could be reduced by using dwell and refractory periods, while still keeping the advantage of increased TP rates. The dwell time is the amount of time the signal has to cross the threshold to be considered a valid detection. Once the dwell time has been met, the refractory period is the amount of time during which the signal is simply ignored, regardless of its value. These tools are specific to self-paced operation and are known to decrease false positive rates **[43]**.

The analysis on the influence of smoothing operations revealed the hold time to be around 0.2 seconds for all subjects. Based on this, the maximum value considered for the dwell time is 0.3 seconds. For the refractory period, we consider values between 0.5 and 1.5 seconds. The upper limit of the refractory period is chosen based on the evaluation data, where the minimum duration between events is also 1.5 seconds, but also on the more practical consideration of not reducing the speed of operation too much.

The approach we adopt is to determine the optimal dwell time and refractory period for each step of the FPR range in Figure 32. A choice must then be made regarding the maximum allowable FP rate. To better illustrate the performance benefits, two upper limits of the FP rate will be considered, of 1% and 1.5%. The results thus obtained are given in Table 4 and Table 5.

Table 4: Influence of dwell time and refractory period on TP rates for a maximum FP rate of 1%. The left part of the table shows the original TP rates obtained with no dwell or refractory post-processing. The right part of the table shows the TP rates obtained with the optimal dwell and refractory periods. FPR_{Orig} is the initial FP rate obtained at the corresponding TP rates, FPR_{Optim} is the FP rate after post-processing. Numerical subscripts of TP rates indicate the IC state, Avg indicates their average.

Subject	Original			Optimized max(FPR) = 1%						
	TPR_1	TPR_2	TPR_{Avg}	TPR_1	TPR_2	TPR_{Avg}	FPR_{orig}	FPR_{optim}	Dwell	Refract
a	52	32	42	62	53	57.5	1.78	0.99	0	0.8
b	36	65	50.5	51	81	66	1.78	0.99	0	1
f	44	46	45	53	57	55	1.79	0.98	0	0.8
g	38	48	43	89	71	80	3.59	0.99	0	1.4
Average	**42.5**	**47.75**	**45.13**	**63.75**	**65.5**	**64.63**	-	-	-	-

Table 5: Influence of dwell time and refractory period on TP rates for a maximum FP rate of 1.5%. For a detailed description refer to the caption of Table 4.

Subject	Original			Optimized max(FPR) = 1.5%						
	TPR_1	TPR_2	TPR_{Avg}	TPR_1	TPR_2	TPR_{Avg}	FPR_{orig}	FPR_{optim}	Dwell	Refract
a	58	51	54.5	81	69	75	2.97	1.49	0	0.8
b	50	80	65	65	88	76.5	2.37	1.42	0	0.5
f	50	56	53	69	74	71.5	3.18	1.45	0	1.3
g	61	58	59.5	92	83	87.5	5.17	1.47	0	1.2
Average	**54.75**	**61.25**	**58**	**76.75**	**78.5**	**77.63**	-	-	-	-

Presumably because of the typically short hold time, using dwell times decreased true positive rates considerably and thus the best TP rates were found when no dwell time was used. On the other hand, the introduction of the refractory period brings substantial increases in performance. For all subjects, the TP rates obtained when refractory periods are used at the FP rate of 1% are larger than the TP rates obtained with no refractory period at the FP rate of 1.5%. The refractory period is thus a highly useful tool, as it effectively delivers TP rates which could normally be achieved only by accepting increases in the false positive rate of more than 50%.

The relative improvement for each subject correlates with the original false positive rate at which the best results were found. For large false positive rates, we can expect that true positive rates will also be large, as Figure 32 shows. The relative performance increases brought by the use of refractory periods are shown in Figure 33 for all subjects and for both upper limits of the FPR. Subject *g* presents the most impressive improvement, of 86%, at an FPR of 1%. Even though not as large, this subject has the biggest improvement for a 1.5% FPR as well. This can be traced back to the larger initial FP rates in the case of this subject. In other words, the operating point for subject *g* is more to the right in the plots of Figure 32, compared to other subjects. For subject *f* we can see an interesting development as well: at an FPR limit of 1%, the initial FPR was the same as that of subjects *a* and *b*, and the improvements brought by the refractory period were of only 22%, the smallest across all subjects. However, at the FPR limit of 1.5%, the initial FPR was considerably larger, hence the performance benefit was also larger.

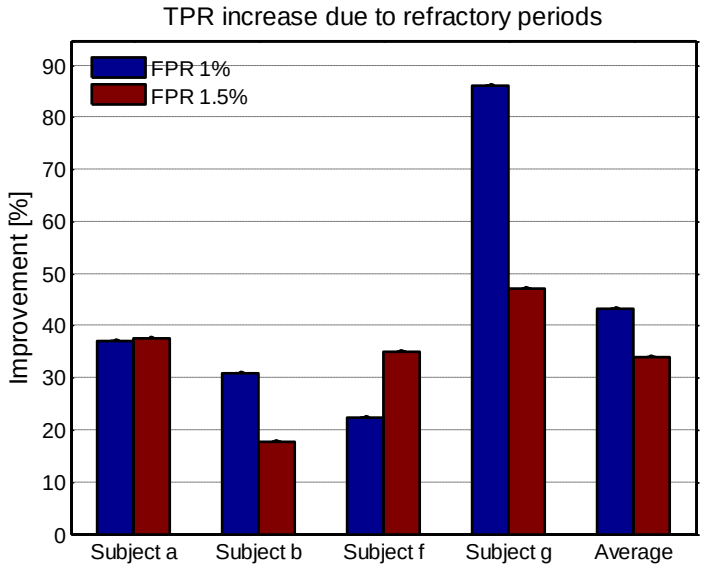

Figure 33: Relative improvement in TPR due to the use of refractory periods at two FPR values

A choice needs to be made regarding the maximum allowable FP rate. Even though an FPR of 1.5% gives very impressive TP rates (> 70% for all subjects), the relative improvements from Figure 33 suggest that the limit of 1% is better overall. For two subjects an FPR of 1% allows bigger improvements compared to the 1.5% limit, while the differences are insignificant for another subject. Moreover, maintaining the FPR at 1% is more consistent both with the results obtained until now, and with the ones found throughout the literature.

2.1.9 Selection of relevant intervals

Another key factor which can be fine-tuned is the time interval from which samples are extracted for training the classifier. The feature selection procedure already provides some information, specifically the most informative position of a one second sliding window. This is however based on the average value of the samples within that window rather than their individual values. Moreover, it may be that the best performance is achieved with different durations of this interval for the two classes. Hence, a more detailed analysis on this matter is warranted.

We have tested the cross-validation performance with four sizes of the selection window and four offsets. The sizes and offsets under consideration are 0.5, 1, 1.5 and 2 seconds. As there is no guarantee that it would be beneficial to use the same size and/or offset for both IC states, all combinations must be tested. For each configuration, the best dwell and refractory periods are determined. The results, though a bit difficult to visualize, are shown in Figure 34, where we have structured all sets of sizes and associated offsets into four groups, as per the caption.

Indeed, the best performing parameters are not necessarily identical for both classes. In fact, only in the case of subject *f* are the sizes of the selection intervals the same for the two IC states, while their offsets are different. It is noteworthy to mention that true positive rates have increased for all subjects, compared to the initial values obtained with the one second selection window. TP rates increase 7%, 5.3%, 25.5%, and 5.6% for subjects *a*, *b*, *f*, and *g*, respectively. The newly obtained TP rates are now larger than 60% for all subjects. Yet again, no dwell time was found to bring improvements.

In Figure 34, the true positive rates of each IC state are also plotted, which allows certain observations to be made. To exemplify the kind of information that can be extracted from such plots, consider the first IC state of subject *f*, which corresponds to imaginary movements of the left hand. The first thing to notice is that the pattern is made up of stripes. As this is the first IC state, corresponding to the *x* axis, it means that its detection rate is independent of the selection interval of the second IC state. The stripes become progressively darker both with increased duration and offset of the selection interval, meaning that the imaginary movement becomes easier to detect towards the middle of the trial. Analogously, for the second IC state (*foot*), we would expect horizontal lines in the pattern. This is not the case though, as the detection rates for the *foot* class seem more dependent on the selection interval of the *left* class. Therefore, in three-state classification, the selection interval of one IC state can influence the true positive rate obtained for the other IC state. Similar observations can be made for other subjects as well. The second IC state of subject *b* is apparently easily detectable even for short windows starting at only 0.5 seconds post stimulus, potentially indicating that the band power modulations of this imaginary movement are rather consistent across the duration of a trial.

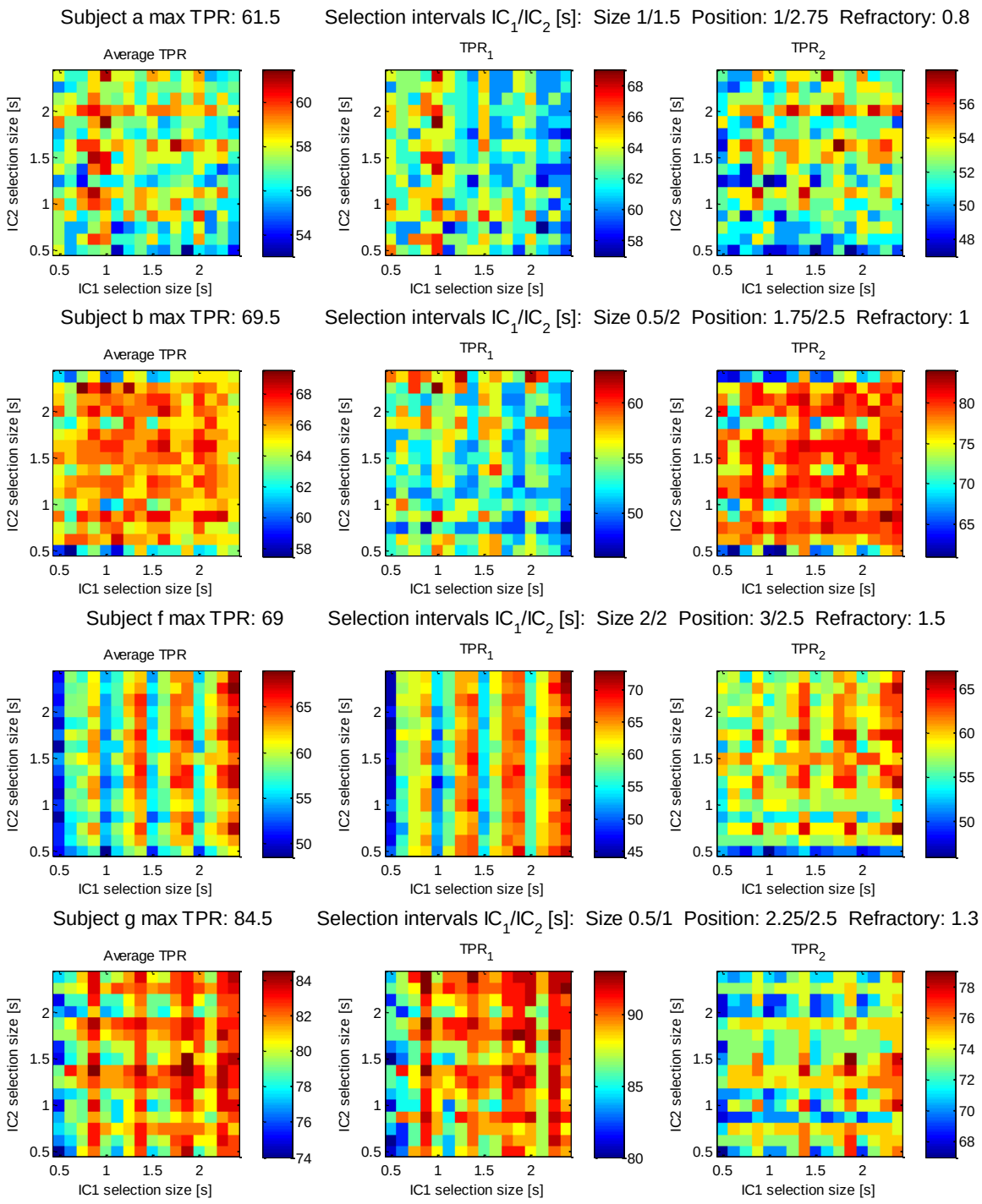

Figure 34: Influence of the selection intervals on TP rate. Four sizes and four offsets are tested, of 0.5, 1, 1.5 and 2 seconds. For each combination, the best dwell and refractory periods are found. On the axis they are grouped with respect to the size, hence there are four groups [(size$_1$,offset$_1$), (size$_1$,offset$_2$)…], [(size$_2$,offset$_1$),(size$_2$,offset$_2$)…] etc. The ticks represent the size of the selection window, thus the tick following that at 1.5s corresponds to a size of 1.5s and an offset of 1s, thus a 1.5s window centered at 1.75s

2.2 Evaluation

As explained in the section dedicated to the description of the dataset, the evaluation data consists of motor imagery trials of variable length, between 1.5 and 8 seconds, and "rest" trials with durations in the same range, which we call NC events. Therefore, two measures of the false positive rate can be calculated: sample FP rate is measured with respect to all samples belonging to the NC state, which comprises inter-trial-intervals and rest trials, while event FP rate is measured as the percentage of rest trials incorrectly detected as IC states. Therefore, a single false activation during a rest trial results in the detector being declared incorrect for the whole duration.

One of the key issues we want to investigate is whether the improvements brought by the use of refractory periods and the selection of specific training intervals hold for the evaluation data as well. Three configurations are therefore tested. The first is the "default", with no refractory period and in which training samples are extracted from the one-second intervals found by the feature selection procedure. The second configuration uses the same training intervals but adds the refractory period, and the third configuration uses both the refractory period and the optimized training intervals found in Figure 34. The three approaches are compared in Figure 35 in terms of true positive rate, misclassification errors between IC states, sample FPR and event FPR.

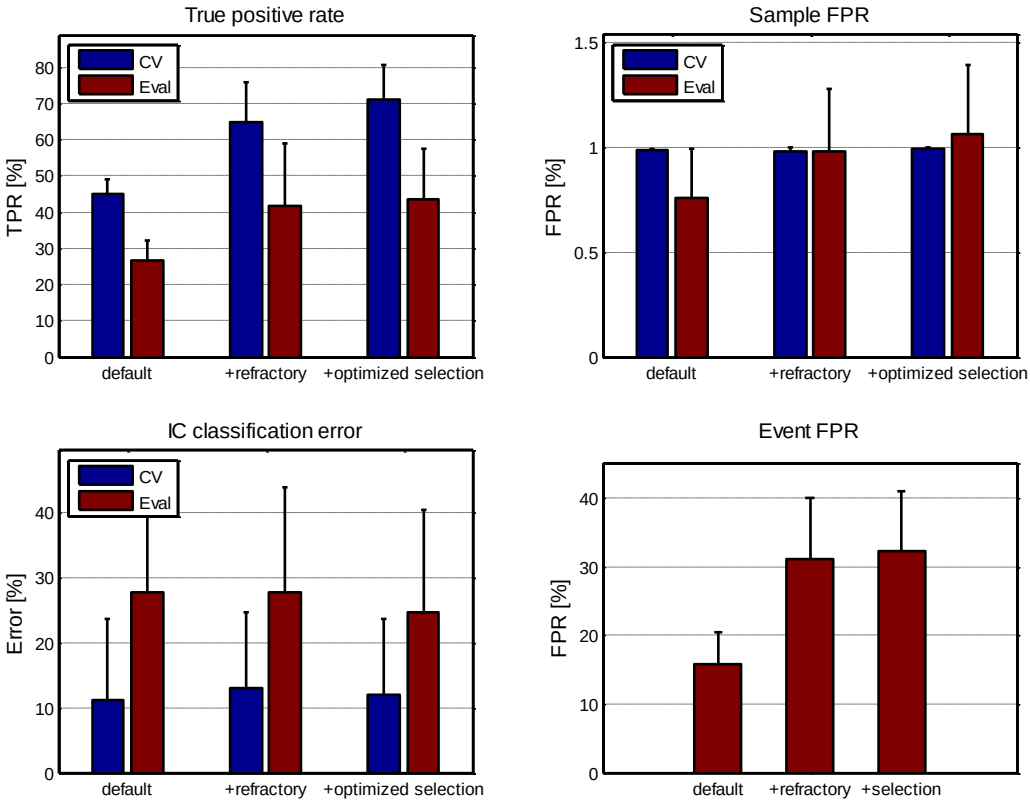

Figure 35: Comparison between three approaches in cross-validation (CV) and evaluation (Eval). Values are averaged over all subjects and shown with the standard deviation. The default configuration uses the one second selection interval found in feature selection and no refractory period; +refractory adds the use of refractory period; +optimized selection uses both optimal training intervals and refractory periods.

The improvements brought by the use of refractory windows hold quite well for the independent test set and are actually larger. Whereas TP rates increased on average by 43% in cross-validation, the average increase on the test set is of 57%. The average TPR on the test set when using refractory periods is comparable to the average cross-validation TPR when not using refractory periods, although it varies considerably more between subjects. This is not the case for the optimization of training intervals, which brought an additional increase of 10% in cross-validation, but of only 4% on the test set. However, it is interesting to note that even though this increase is small, the standard deviation between subjects is also smaller. These improvements apparently come at the cost of increased false positive rates. In cross-validation, the FP rates were fixed at 1% for all three methods. While for the default method sample-based FPR actually decreases on the test set, it increases for the other two approaches, although it is kept below 1.5% for all subjects. The small increases in false positive rate might not seem significant, but it is interesting to correlate them with the event false positive rate. Comparing the "default" and "refractory" approaches, we can see that an increase in sample FPR of only 30%, from 0.76% to 0.98%, translates into an increase of almost 100% in event FPR. One benefit of using the optimized training intervals seems to be the lower misclassification error between IC states. While in cross-validation there are no significant differences between the three methods, the situation changes on the test set, where using the optimized training intervals was found to provide consistently lower misclassification errors for all subjects.

A more detailed, subject-specific analysis of LDA with refractory periods and optimized training intervals is given in Figure 36, where the true positive rate is also presented as a function of the IC events duration. TP rates generally increase with longer event durations, although not monotonically. Most events are satisfactorily detected for all subjects, even short lasting ones between 1.5 and 2.5 seconds. Two notable exceptions are subjects *b* and *f*. For the former subject, the BCI failed to detect the shortest lasting events of IC state *right*, whereas for the latter subject the same is true for events of the *foot* class with durations between 3.5 and 4.5 seconds. Note that these are not errors due to a small number of events of the corresponding durations: for subject *b* there were 20 events of class *right* lasting between 1.5 and 2.5 seconds, and for subject *f* there were 15 events of class *foot* lasting between 3.5 and 4.5 seconds.

When analyzing the influence of the refractory period (Table 4 and Table 5) we have found that more "relaxed" thresholds could be used for the data of subject *g*, effectively allowing higher TP rates at the same FP rate. The refractory period was efficient in decreasing FP rates considerably in cross-validation, but this does not seem to hold on the test set. The large FP rates obtained with no refractory period reflect in large FP rates on the test set, both sample- and event-based.

In general, it seems that sample FPR is indeed proportional to event FPR. For instance, subject *b* presents the lowest sample FPR, of 0.68%, and consequently also has the lowest event FPR, of 19.49%. The relation between the two metrics is not linear though: subject *a* has a sample FPR 40% larger compared to that of subject *f*, but the event FPR is only 8% larger.

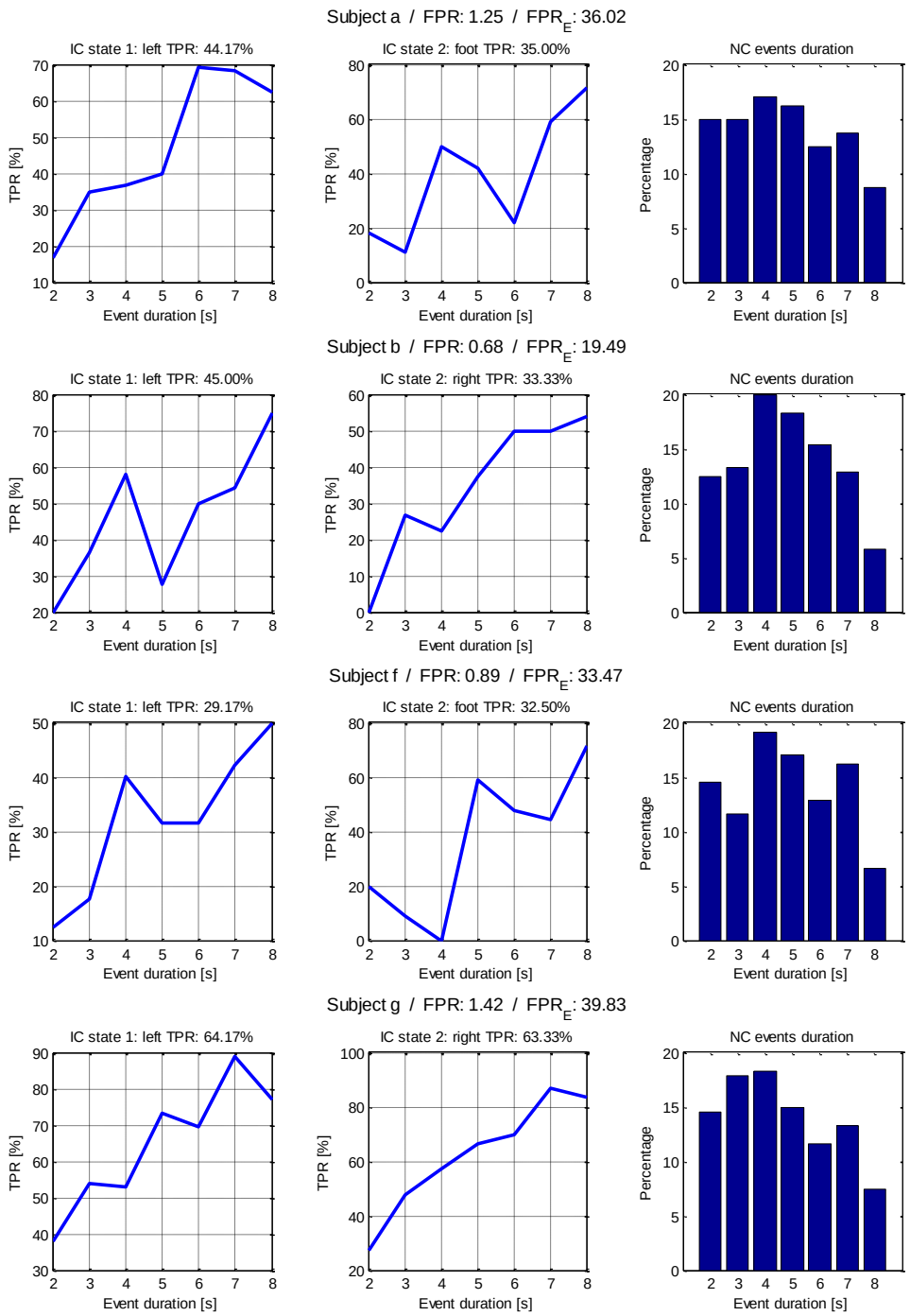

Figure 36: Event analysis for LDA on the evaluation data, with TP rates as a function of event duration. Refractory periods and the optimal training intervals are used. The TPR above each plot is the TP rate of the corresponding IC state, averaged over all event durations. FPR is the sample FP rate, FPR_E is the event FP rate

To gain a better understanding of the behavior of the BCI, it is important to also analyze how false activations are distributed, whether they are random or appear in patches (inter-FA periods, see section 3.2 of chapter II). The inter-FA periods distribution for the four subjects is presented in Figure 37 in the form of histograms. Because the time between false activations can be very long, only the most significant interval is plotted for each subject. To get a more detailed impression, the minimum and maximum time between false activations is also given above each plot.

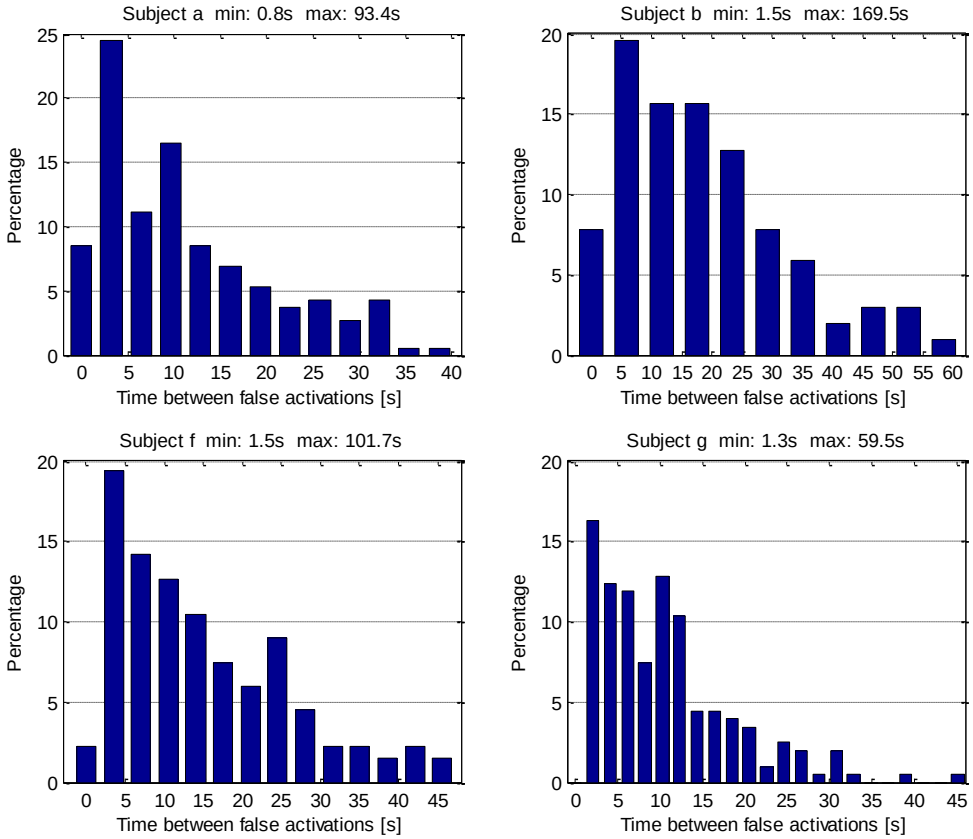

Figure 37: Inter-FA periods distribution for LDA. The minimum and maximum time between false activations is given above each plot.

The distribution is skewed towards intervals of less than five seconds for all subjects, thus false activations are not completely random. Because of the refractory window, there are no consecutive false activations.

3 Regression

3.1 Cross-validation

3.1.1 Influence of the feature processing pipeline

The first step is to determine the influence of various parameters on performance and we will start with the feature processing pipeline. Similar to classification, a grid search was performed on window durations ranging between 1 and 19 samples, with and without the log transform. Performance is now evaluated not only in terms of TP rate, but also mean squared error and mutual information with the true labels. The results, averaged over all subjects, are plotted in Figure 38. The three performance metrics are very differently affected by the size of the moving windows. The mean squared error and the mutual information both profit from increased window durations, whereas the best detection rate is clearly found when no smoothing is performed.

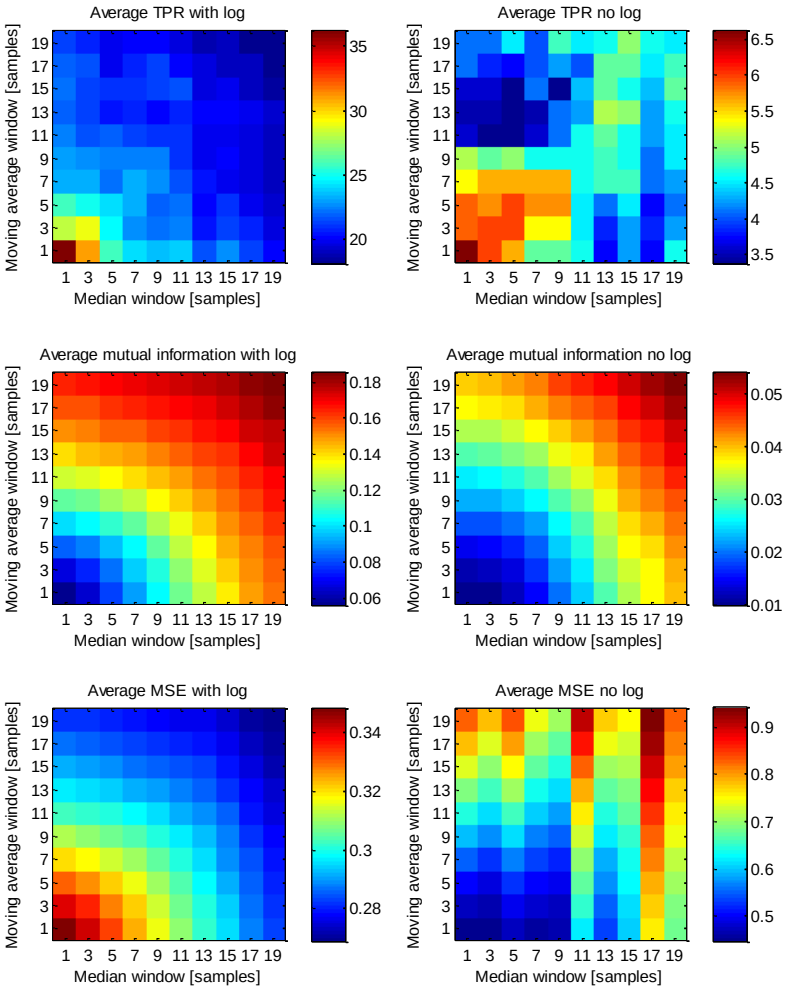

Figure 38: Influence of the feature processing pipeline in terms of detection rate, mean squared error and mutual information, averaged over all subjects.

In Figure 39 we present the cross-validation detection rates of the Wiener filter with no smoothing applied, as well as a comparison to direct three-state classification with LDA. On average, the true positive rates achieved by regression are lower than those of LDA classification. The best performing case is that of subject *g*, where both methods achieved an average TP rate of 43%. The true positive rates of the two IC states are 36% and 50% for regression, while LDA offered similar rates, of 38% and 48%, respectively.

Interestingly, the second IC state offers better detection rates for all subjects in the case of regression, whether it is the right hand or the foot class. For three subjects this was also the case of LDA, with the exception of subject *a*, in which case the TPR of the first IC state was considerably larger than that of the second IC state. This is in accordance with the findings of the feature selection procedure (see Figure 22, page 48). As with three-state classification, misclassification errors between IC states are rather low, of maximum 8.33% for IC state *left* of subject *f*.

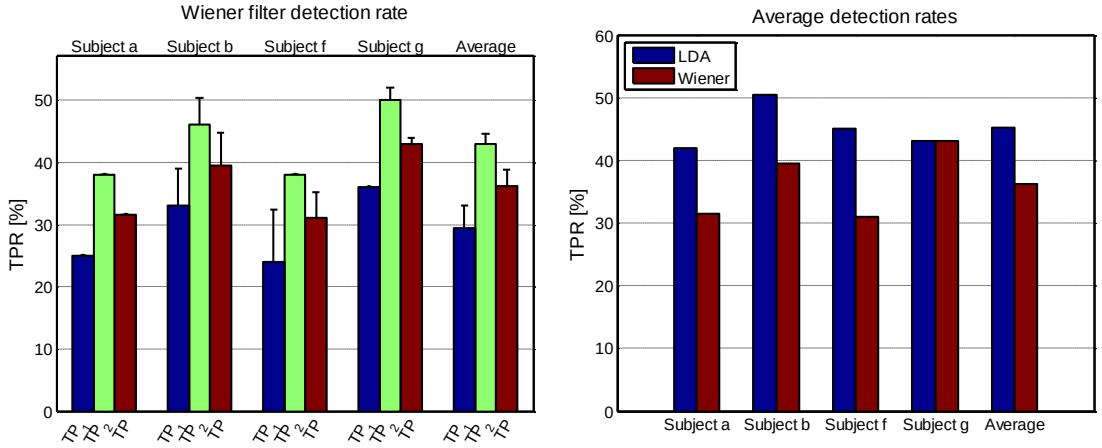

Figure 39: Detection rates for the Wiener filter (left) and comparison to LDA (right). Error bars represent the percentage of misclassified, yet detected IC states. TP_i represents the true positive rate of IC state *i*, TP is the average true positive rate of IC detection. All TP rates correspond to an FP rate of 1% or lower.

3.1.2 Assigning different regression targets

Whereas the specific values of class labels are irrelevant to classification, they are a key factor in regression. As such, one possibility of increasing performance would be to assign different targets, instead of the default labels of {-1, 0, 1}. We tentatively employ LDA for dimensionality reduction and determine the one-dimensional representation of the data which maximally separates the three classes with respect to the Fisher ratio. The new targets for regression can be either constant for all samples belonging to a class or simply their one-dimensional projections. In the former case, one could assign to all samples belonging to one class the average value of their one-dimensional projections. The latter approach offered comparatively better results however, which are presented in Table 6.

Table 6: Average true positive rates with the original labels and with the ones determined by LDA. The target values are expressed in terms of their mean and standard deviation

Subject	Target values, mean (std)			Average TPR [%]		Improvement
	IC_1	NC	IC_2	Original	Post-LDA	
a	-0.96 (0.85)	0.07 (1.02)	0.46 (1.00)	31.5	35.5	12.7 %
b	-0.58 (0.89)	0.19 (1.03)	-0.99 (0.87)	39.5	35	-11.4 %
f	-0.72 (0.92)	0 (1.02)	0.82 (0.93)	31	30	-3.23 %
g	1.12 (1.11)	0 (0.99)	-0.95 (1.14)	43	40.5	-5.81 %

The only case where the targets found by LDA improved performance is subject *a*. For the other three subjects, the average true positive rates actually decreased. Putting performance aside though, it is interesting to examine the target values found by LDA. For all subjects, the average value of the NC state is close to zero, especially for subjects *f* and *g*, where it deviates from zero only starting with the third decimal. The most interesting cases are subjects *b* and *g*, coincidentally the subjects which chose imaginary movements of the right hand instead of the feet. For subject *b*, the ordering of the classes is completely scrambled. The second IC state now has the lowest average value, which is followed not by NC, but rather by the first IC state, which also has a negative average target value. The two IC states are fairly close together and have considerable overlap. A very interesting case is also that of subject *g*, where even if the NC state remained in the central position, the two IC states have switched, with the first IC state having a positive value on average, and the second one having a negative value.

As the LDA approach was not tremendously successful, we decided to try more of a brute force method. A genetic algorithm was employed to determine constant and unconstrained target values for each class. The fitness function therefore had three variables and the criterion to optimize was simply the average cross-validation TP rate. The best results out of five runs of the genetic algorithm are given in Table 7.

Table 7: Average true positive rates with the original labels and with the ones determined by the genetic algorithm

Subject	Target values			Average TPR [%]		Improvement
	IC_1	NC	IC_2	Original	Post-GA	
a	-1.67	1.04	9.34	31.5	39	24 %
b	-1.34	0.19	0.7	39.5	41.5	5 %
f	-3.88	2.85	1.52	31	39	26 %
g	-1.65	2.87	1.46	43	46	7 %

The target values determined by the genetic algorithm successfully improve detection rates for all subjects, in particular subjects *a* and *f*, which originally presented the lowest TP rates. The small improvement of 7% in the case of subject *g* is sufficient to make this the only situation where regression performs better than classification.

The new targets are quite different than those found by LDA. The label for the NC state is no longer close to 0, except for subject *b*, where coincidentally it received the same target value as the one determined by LDA. The ordering of target values is also different, with the NC state receiving the highest target value for subjects *f* and *g*.

3.1.3 Influence of tap delays

The Wiener filter can be extended to a finite impulse response topology by using tap delays, thus making use of past samples. The motivation behind this approach is the possibility of reducing the event FP rate, as the classification results on the test set showed this to be a significant problem. The effects of tap delays on the true positive rate, the event false positive rate and the average response time are shown in Figure 40.

The results indicate that while the event FPR indeed decreases when taking past samples into consideration, so does the ability of the BCI to detect IC commands. The TP rate and the event FP rate both decrease at roughly the same pace with increased number of tap delays and seem to have an almost linear relation. The response time does not seem to be very influenced by the extended feature space and only slightly deviates from its average of 2.2, 2.5, 1.7, and 2 seconds for subjects *a*, *b*, *f*, and *g*, respectively. For subject *f* the BCI consistently provides the shortest response time, of 1.7 seconds on average.

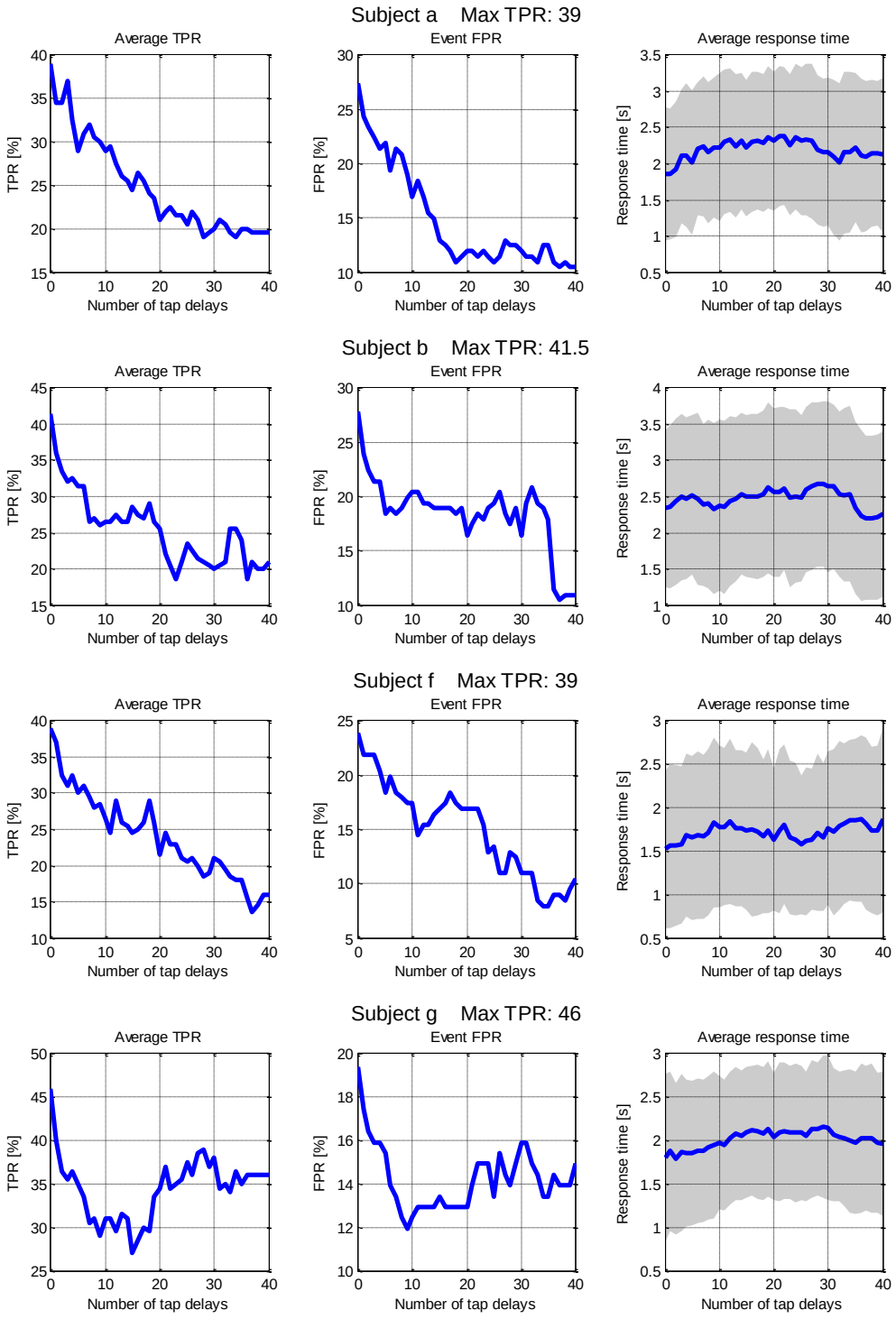

Figure 40: Influence of tap delays on TP rate, event FP rate and response time. One tap delay is equivalent of using 100 ms of past data. The shaded area in the response time plots represents the standard deviation

3.1.4 Larger FP rates and refractory periods

Similar to classification, we are also interested in the TP rates obtained at FP rates higher than 1%, as this gives a more complete view on performance. The true positive rates achieved by the Wiener filter for false positive rates ranging between 1% and 10% are shown in Figure 41.

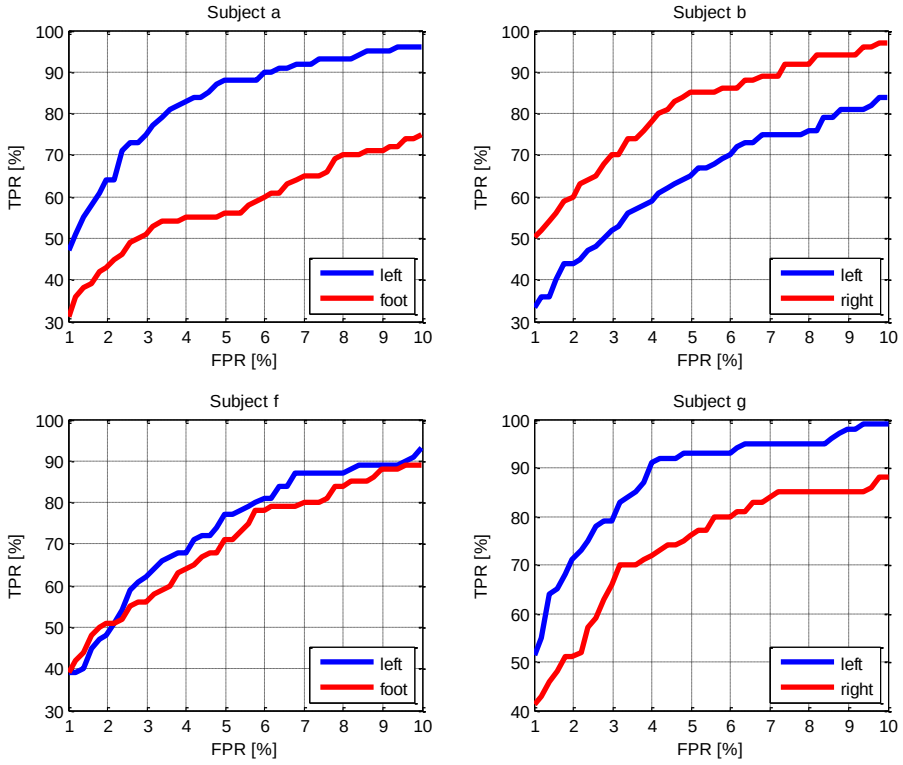

Figure 41: TP rates of the Wiener filter for FP rates ranging between 1% and 10%

Although TP rates improve significantly at higher FP rates, the TPR increase is not as pronounced or as steep compared to classification (see Figure 32). For subjects *b* and *f*, the increase is approximately linear, whereas it is approximately logarithmic in the case of LDA. Nevertheless, we have also investigated the possible benefits of using refractory periods and the results are presented in Table 8.

Table 8: Influence of dwell time and refractory period on TP rates of the Wiener filter for a maximum FP rate of 1%. The left part of the table shows the original TP rates obtained with no dwell or refractory post-processing. The right part of the table shows the TP rates obtained with the optimal dwell and refractory periods. FPR_{Orig} is the initial FP rate obtained at the corresponding TP rates, FPR_{Optim} is the FP rate after post-processing. Numerical subscripts of TP rates indicate the IC state, Avg indicates their average.

Subject	Original			Optimized max(FPR) = 1%						
	TPR_1	TPR_2	TPR_{Avg}	TPR_1	TPR_2	TPR_{Avg}	FPR_{orig}	FPR_{optim}	Dwell	Refract
a	47	31	39	70	43	56.5	2.37	0.99	0	1.4
b	33	50	41.5	43	59	51	1.77	0.93	0	1.3
f	39	39	39	47	49	48	1.78	0.98	0	0.9
g	51	41	46	82	68	75	3.17	0.97	0	1.4
Average	**42.5**	**40.25**	**41.38**	**60.5**	**54.75**	**57.63**	-	-	-	-

As the event-based ROC curves from Figure 41 suggest, the Wiener filter is less efficient than LDA for most of the FPR range considered. It comes then as no surprise that although refractory periods are useful in increasing TPR, the performance benefits are lower compared to classification. The closest results are found for subject *a*, in which case LDA offers an average TPR of 57.5%, only slightly higher than the average TPR of 56.5% offered by regression. This can be explained by the larger initial FPR of 2.37% in the case of regression, compared to only 1.78% for classification.

3.2 Evaluation

Before proceeding to more detailed analyses, we will first test whether the improvements brought in cross-validation by the new targets and the refractory period also hold on the independent test set. Three approaches are thus tested, each adding an extra processing step: regression with the original target values, with the targets determined by the genetic algorithm, and with the additional use of refractory periods. They are compared in terms of true positive rate and event false positive rate in Figure 42. On average, the target values determined by the genetic algorithm give higher TP rates, but also higher event FP rates. With the exception of subject *f*, the two metrics correlate quite well: when TPR increases, so does event FPR. In the case of subject *f*, both TP rates are the same, but the "optimized" targets give a higher event FP rate. A substantial difference can be observed for subject *g*, in which case the new targets brought a 9% increase in TPR but a much larger increase in FPR, of 65%.

The same situation holds for the refractory period as well, which although is found to increase true positive rates considerably, it increases event FPR even more. For subjects *a* and *g* the event FPR reaches almost 40%.

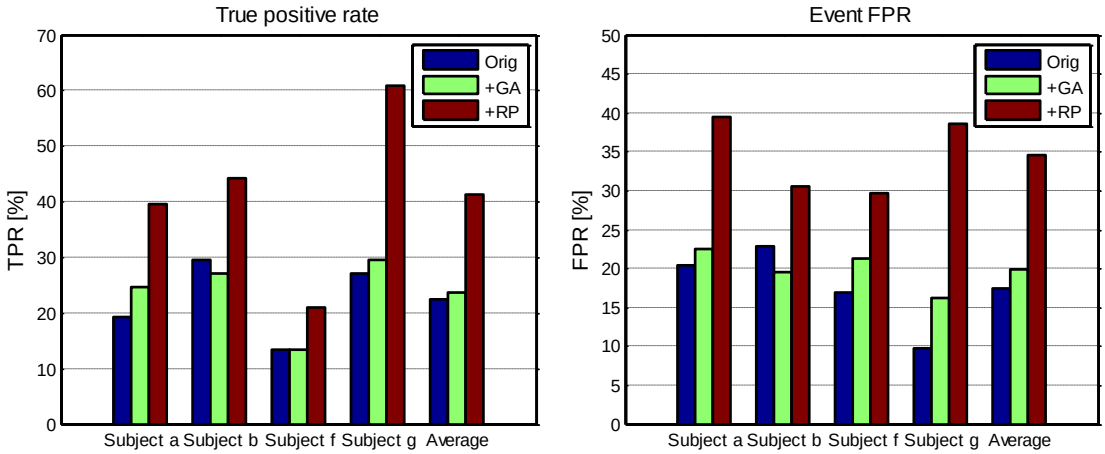

Figure 42: Regression results on the evaluation data, with the original targets (Orig), with the ones determined by the genetic algorithm (+GA), and the additional use of refractory periods (+RP).

It is interesting to compare the event FP rates with the original sample FP rates from Table 8, prior to the application of the refractory period. One can notice that for all subjects, the ratios of sample FP rates in cross-validation before and after the refractory period are proportional to the ratios of event FP rates on the test set with and without the refractory period, respectively. In other words, the reduction in sample FPR in cross-validation can give an indication on the increase of event FPR on the test set, and this seems consistent between subjects. This observation is presented in Table 9, based on the data from Table 8 and Figure 42.

Table 9: Ratios of sample and event FP rates in cross-validation (CV) and on the test set, respectively. *RP* stands for refractory period; FPR subscripts *pre* and *post* (corresponding to subscripts *orig* and *optim* in Table 8) indicate the FP rates before and after applying RP; subscripts *w* and *w/o* indicate the event FP rates with and without RP, respectively. The rightmost column is the ratio of the two anterior columns.

Subject	**CV (sample)** $\frac{FPR_{pre\_RP}}{FPR_{post\_RP}}$	**Test (event)** $\frac{FPR_{w\_RP}}{FPR_{w/o\_RP}}$	$\frac{Ratio_{CV}}{Ratio_{Test}}$
a	2.39	1.75	1.37
b	1.9	1.56	1.22
f	1.81	1.4	1.29
g	3.27	2.4	1.36

It is also interesting to compare these results with those obtained by classification with no refractory period (see Figure 35). Regardless of which target values were used for regression, classification is superior, with an average TP rate of 28% and a lower event FP rate, of 16%. The refractory period increases the TP rate of regression considerably and brings it on par with classification, although the event FPR is still larger.

In Figure 43 a subject-specific performance analysis is presented for the optimized target values with refractory periods.

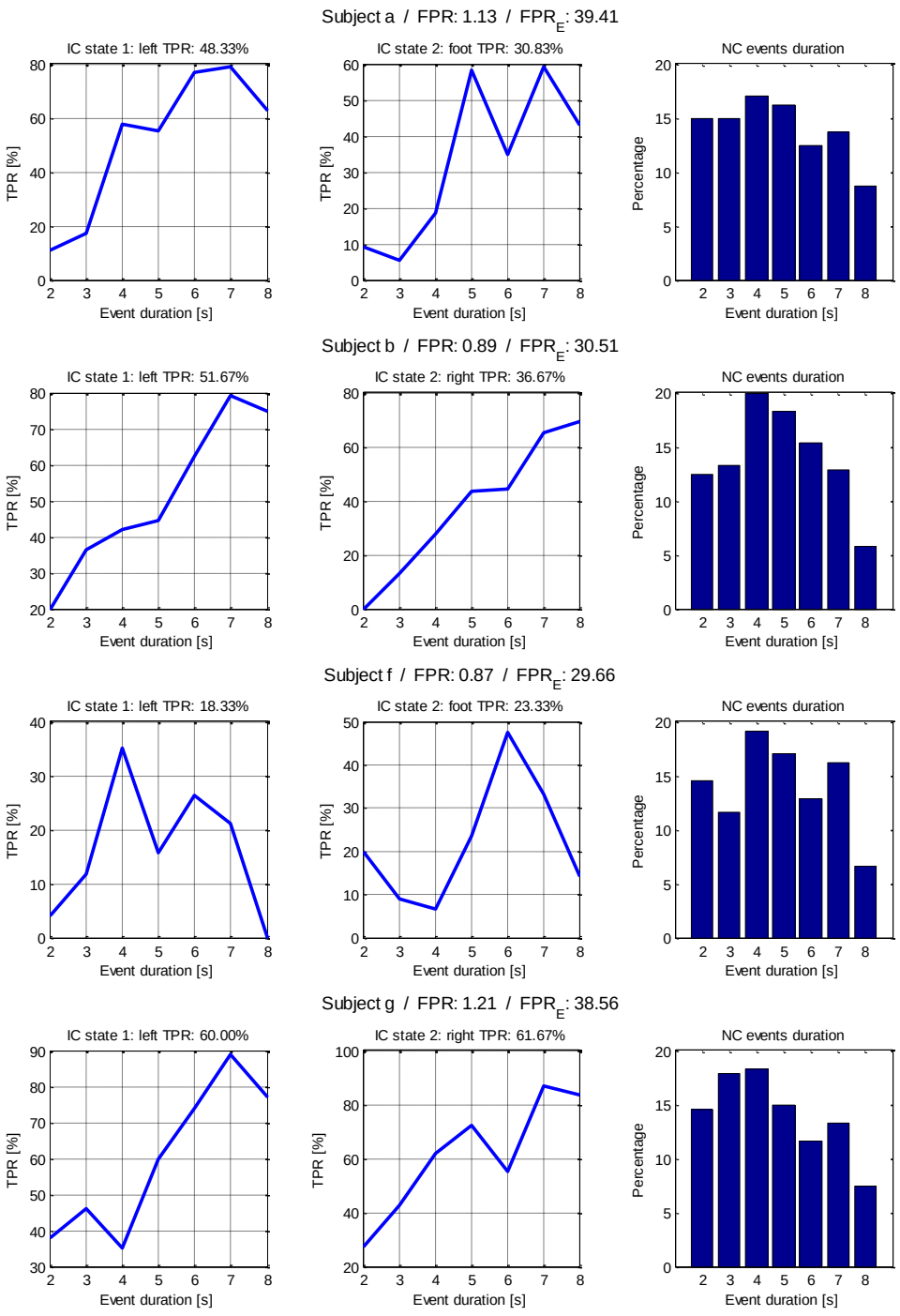

Figure 43: Event analysis for the Wiener filter on the evaluation data, with TP rates as a function of event duration. The TPR above each plot is the TP rate of the corresponding IC state, averaged over all event durations. FPR is the sample FP rate, FPR_E is the event FP rate

As was also the case with the classification results, TP rates increase with longer event durations. The exception is subject *f*, for which the detection rates of both IC states decrease considerably for events longer than six seconds. This subject also presents the lowest average TP rate, both for classification and regression. Subject *f* appears to be a special case, as the improvements brought by the refractory window (see Figure 42) are modest, although the increase in event FPR is similar to what is found for other subjects. Compared to classification, the overall results are similar: when TPR increases, so does event FPR. Sample FPR is proportional to event FPR, and the two metrics are generally also proportional between classification and regression. That is, if sample FPR is lower for classification compared to regression, event FPR will also be lower. One exception is subject *a*, in which case sample FPR and event FPR are 1.25% and 36% for classification, and 1.13% and 39.4% for regression.

The inter-FA periods distribution is plotted in Figure 44 for each subject, along with the shortest and longest durations between false activations. The distribution is not very different from that of classification, and is skewed in a similar fashion towards shorter durations. There are some differences nonetheless, such as the maximum time between false activations, which is shorter for regression.

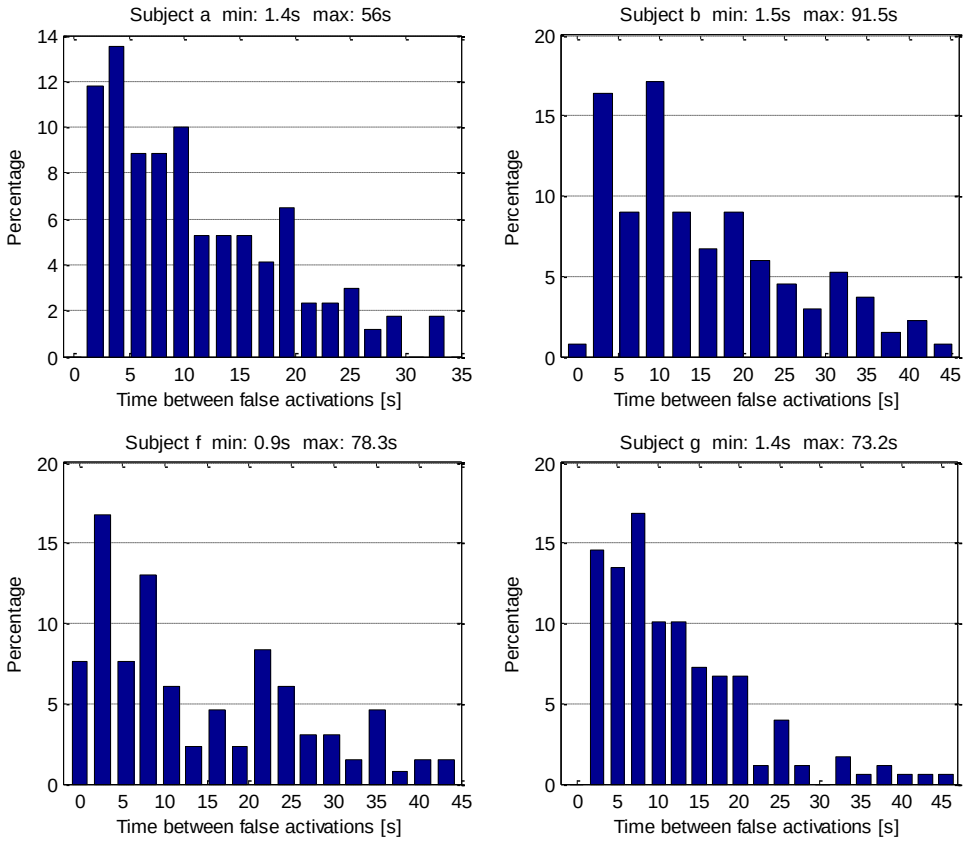

Figure 44: Inter-FA periods distribution for the Wiener filter. The minimum and maximum time between false activations is given above each plot.

4 Combining multiple outputs

Rather than trying to find the best performing algorithm and meticulously fine-tune its parameters, a more sensible idea might be to combine the output of several methods in the attempt to increase performance.

We investigate the possibility of linearly combining multiple methods by using a genetic algorithm. The translation algorithms we consider are three-state LDA and QDA classification, LDA, QDA and one-class SVM in the two classifier configuration for separate detection of each IC state, and the Wiener filter. Thus, there are nine outputs in total. A genetic algorithm determines a weight vector $W = [w_1, w_2 \dots w_9]$ with $-1 \leq w_i \leq 1, \forall i$. The fitness function simply takes the weight vector and the cross-validation scores as input and, depending on the performance metric which is to be optimized, returns either the average TPR, the difference between the average TPR and event FPR, the true-false difference or the mutual information of the resulting score with the true labels. The approach is computationally feasible because all the scores are pre-calculated. To make sure that the solution determined by the genetic algorithm is at least as good as any of the methods under consideration taken separately, we construct an initial population of six individuals representing the six methods under consideration. Thus, for three-state approaches (LDA, QDA and regression), the corresponding weight is set to 1 and all other weights to 0. For the two-classifier configurations the procedure is similar except that two weights are non-zero: the weight of the first classifier is set to -1 and the weight of the second classifier is set to 1, equivalent thus to the differential mode of operation. Refractory periods are not considered here and as usual, the sample FPR in cross-validation is kept at a maximum of 1%. The results obtained, averaged over all subjects, are shown in Figure 45, along with the original results of 3-state LDA. One might notice differences between the LDA results from Figure 45 and others presented throughout the paper, such as those from Table 4. These differences stem from the fact that different selection intervals are used for training the classifier: the grid search for determining optimal training intervals also provided the best performing ones for no refractory periods, and these are the ones that are used here.

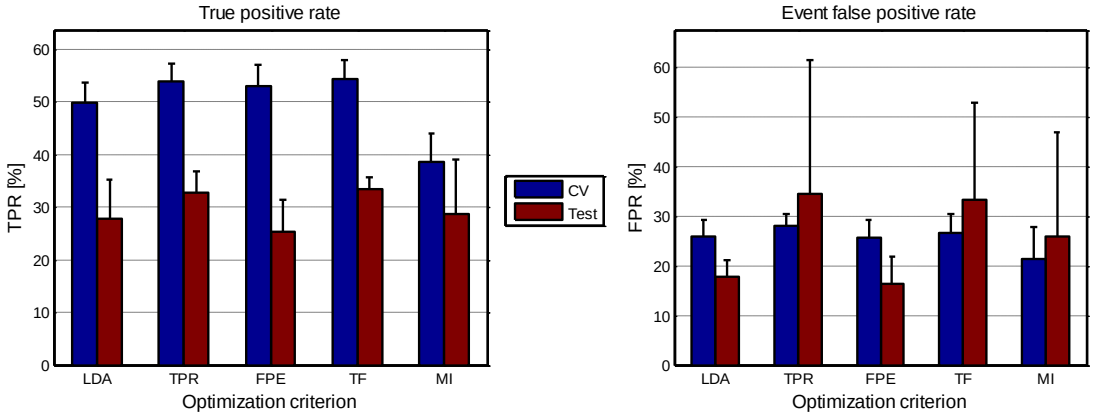

Figure 45: True positive rates and event false positive rates of linearly combining classifier outputs with a genetic algorithm. Results are averaged over all subjects and shown with the standard deviation. On the *x* axis, LDA represents the initial results of 3-state classification. The other four labels indicate the optimization criterion used in the genetic algorithm: average true positive rate (TPR), difference between TPR and event false positive rate (FPE), true-false difference (TF) and mutual information (MI).

With the exception of mutual information, the other three optimization criteria increase true positive rates in cross-validation, and the TPR and TF criteria also improve TP rates on the test set. However, this comes at the expense of higher event false positive rate on the test set, as was actually the case with all potential improvements analyzed throughout this paper. Maximizing the difference between TPR and event FPR managed to reduce the latter metric on the test set, although the former was also decreased. Note that only the original LDA results and the FPE criterion have lower event FPR on the test set compared to cross-validation. Maximizing the true-false difference provides approximately the same results as maximizing the true positive rate directly, although in the former case the standard deviation of event FPR on the test set is lower. The mutual information is not an event-by-event measurement and the regression results from Figure 38 suggest that it does not even correlate with such measurements. This seems again to be the case judging by the cross-validation TP rates, but quite interestingly, the performance degradation on the test set is very low compared to all other approaches, including LDA. It seems that regardless of the performance metric we choose to optimize, there are no clear benefits in linearly combining the outputs of multiple translation algorithms, as any improvement in terms of TPR comes at the cost of increased event FPR.

Chapter V
Discussion

There were many tests and investigations performed in the previous chapter, but also many which were not. There are both trivial and surprising results, and all the information gained until now needs to be put in perspective. This is what we will attempt to do in this chapter.

1 Classifier designs

The first classification scheme tested was binary classification between NC and both IC states. Considering both IC states as one class did not seem appropriate in theory and it did not perform well in practice either, especially for LDA. One notable exception is subject *b*, in which case an overall detection of 45% was obtained. This has a simple, yet subtle explanation, related to the fact that this subject had the highest misclassification errors between IC states, especially for discriminant classifiers. As the overlap between two normal distributions increases, so does the likelihood that samples drawn randomly from both of them could have come from one single Gaussian distribution. Assuming for the argument's sake equal covariances, when the overlap is 100% the two distributions can equally well be described by only one mean and one covariance matrix. It would not be possible for any classifier to distinguish the two distributions. But of course, when there is no overlap the distinction would be very clear. Therefore, clustering poorly separated IC states as one class would give a higher likelihood than clustering well separated ones in the same manner. The assumptions of discriminant classifiers would thus hold better, potentially improving performance. Indeed, subject *b* presents the highest misclassification errors between IC states, thus the largest overlap between them, but also the highest detection rates in binary NC/IC classification.

Still, good separation between IC states does not necessarily imply an inability to jointly distinguish them from the NC state. If the average of their means is sufficiently far from the mean of NC they could still form a well separated cluster in feature space. But in the case of motor imagery, and especially movements of the left and right hand, there is a neurophysiological symmetry of these movements with respect to the NC state due to the spatial organization of the motor cortex. A priori, we have no reason to assume significant differences in band power modulations of the two hemispheres. And with no a priori information it is also reasonable to assume that on average, modulations over the two hemispheres would be similar during NC. This symmetry with respect to NC would be especially problematic for linear classifiers. Consider the one dimensional toy example in Figure 46 where the distributions labeled -1 and 1 are to be separated from the distribution labeled 0. With good separation between -1 and 1, it would not be possible to determine a single threshold which would separate them both from 0: one of them will be left out. When they are overlapped though, it is not only easy to find an optimum threshold, but the addition of a second one, which would be equivalent to non-linear classification, would be redundant.

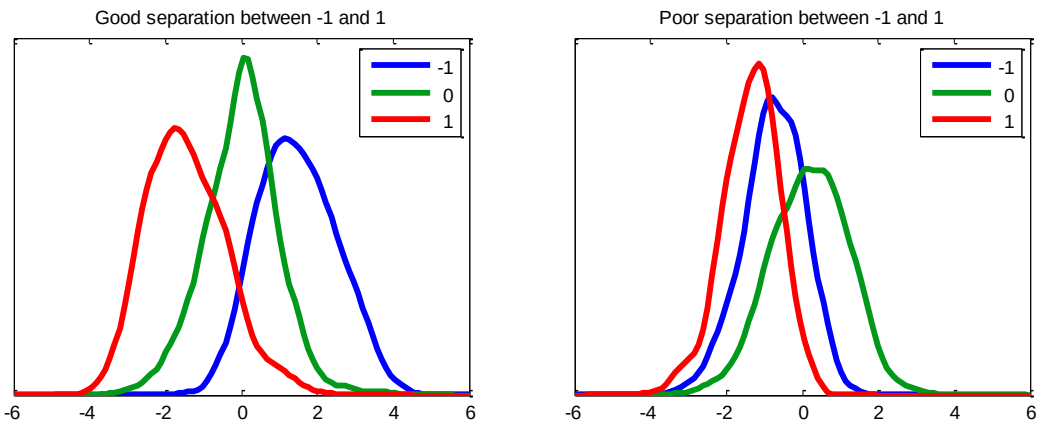

Figure 46: If classes -1 and 1 need to be distinguished from class 0 with a single threshold (linear classifier), it is easier when the separation between them is poor.

For any number of dimensions, if two classes with similar covariances are symmetrical with respect to a third one, a linear classifier would be unable to determine a threshold to separate the former two classes from the latter. Non-linear classifiers would thus be more appropriate, and this would explain the much better overall performance obtained with QDA and the one-class SVM. But is this symmetry of imaginary movements actually found in the data?

In fact, the distributions from Figure 46 are not toy examples, but rather the one-dimensional representation of the data of subjects *g* and *b* (left and right side of Figure 46, respectively), determined by LDA in our attempt to assign new targets for regression. These are simply the distributions described in Table 6, page 72, shown here in the form of probability densities rather than statistical parameters. This simple representation of the data explains the frequent misclassification errors between IC states found for subject *b*, and also nicely correlates with the spatial filters selected for each subject (see the spatial patterns from Figure 20, page 46). For subject *g*, the responses (patterns) of the filters are smooth, symmetrical and neurophysiologically plausible. For subject *b* on the other hand, the patterns associated with the second spatial filters selected for each class are very similar. In fact they must reflect a common activity of the two movements, with differences only in magnitude and not spatial distribution, as the difference between the patterns is constant. If this "common" filter ranked second most important for both IC states, it must have been among the first or last five filters of the initial CSP projection matrix, and must also represent some form of movement-related activity, as it is apparently relevant for separating the IC states from NC. Despite its neurophysiological implausibility, it is probably the reason why for subject *b* such high detection rates were achieved in binary NC/IC classification. This suggests that determining a common, movement-related activity of motor tasks goes a long way in improving detection rates for such two-stage classification schemes.

Thus, poor results of two-stage classification were very much expected, and the good results found for subject *b* were actually surprising. So then how is it that LDA worked in other experiments presented through the literature? In fact, this specific combination of one binary LDA and CSP for detecting movements of any kind is used nowhere in the literature. This is because CSP maximizes the differences between the movements, thus minimizing similarities. But the case of subject *b* shows that exactly such similarities are useful for binary IC/NC classification, therefore minimizing them leaves the classifier nothing to work with. It must be emphasized that here we

only discuss about similarities between movements which at the same time constitute differences with respect to NC; a broken electrode emitting a constant signal would of course not be useful. Thus, for two-stage designs, CSP is highly inappropriate. Surely this is the reason why there are no BCIs with binary IC detectors which use CSP; they mostly use bipolar montages, and often with a rather small number of electrodes. Decreasing the number of electrodes arguably has a myriad of real-life benefits of course, and potentially provides more robust covariance estimators for statistical classifiers, as the dimensionality of the data is lower.

Thus, we can easily argue that two-stage classifiers were not even tested in this paper, as any potential IC/NC detector was doomed from the moment it was trained with CSP projected data. What should have been done would have been to apply unsupervised spatial filtering such as CAR, Laplacian or whitening, as is common practice in other self-paced experiments, and allow the classifier to determine some common features of the two movements which would also discriminate them from NC. Considering the rather large number of channels and frequency bands, this would have complicated the feature selection procedure considerably though. Perhaps experiments with the broad-band 8-30Hz and comparison with three-state classification using CSP on the same band would have been useful and informative. But even so, it is important to understand that CSP is simply an exacerbation of the real story. The symmetry of imaginary movements of symmetric body parts with respect to NC is still there, even with no spatial filtering applied. This shows that in this context, any two-stage BCI design which uses linear classifiers is inherently sub-optimal, simply because of the underlying neurophysiological phenomena.

The very poor results of soft-margin SVMs were unexpected. It is important to note that discriminant classifiers were not affected by the unbalanced data, because the prior probabilities of each state were not estimated sample-by-sample but rather event-by-event. For SVMs however, we can only assign different misclassification costs for each class, and not prior probabilities. Misclassification costs were adjusted as suggested in **[76]** but did not help, and in fact made the situation worse. One possibility would have been to use a balanced training set, but this would have raised the issue of which NC samples are most representative. Linear kernels were also not tested in this paper. However, it is important to note that many of the tests throughout the paper involved the order of N^2 repetitions, and this would not have been feasible with the SVMs, which have computationally intensive optimization functions and also require model selection.

Using two classifiers, one for the detection of each IC state, is a much less popular choice in motor imagery BCI, although it seems promising. Having two distinct scores opens up many possibilities, out of which we have tested the differential mode, where one score is subtracted from the other, and the parallel mode, where a threshold is separately found for each of them. These are not the only options though: rather than simply subtracting them, one could additionally determine individual weights or even train a meta-classifier on the two scores. Although the former possibility was not specifically investigated, the unsatisfactory results of linearly combining multiple scores with a genetic algorithm suggest that most likely significant improvements would have not been found.

The two-classifier approach provided some surprising results. For instance, the largest average detection rate from the initial tests with default parameters was found for subject *b* when one-class SVMs were used in parallel mode. The average TP rate of 48% thus obtained is almost the same as what three-state LDA managed only after several tweaks which more than doubled its performance for this subject. For subject *g* on the other hand, the one-class SVM in parallel mode gave the lowest performance across all tests and subjects. Moreover, this is not some quirk of SVM, as both discriminant classifiers also performed poorly in parallel mode for this subject. This is perhaps the most surprising result of all, simply because we cannot find any reasonable explanation

why three different classifiers trained with data of two well separated classes (each IC state versus NC) would fare so poorly in distinguishing them. Even more, in differential mode all classifiers attained very good performance for this subject, most notably the one-class SVM which offered almost double the performance of the next best subject. Thus, for the one-class SVM subjects *b* and *g* can have both the best and worst performance among all subjects, depending on whether the differential or parallel mode is used. These ratios do not seem to hold across different algorithms though, as the best performance of the discriminant classifiers in differential mode was found for subject *a*.

Overall, three-state classification gave the highest and most balanced detection rates, and this can be traced back to the use of CSP and the symmetry of imaginary movements with respect to NC. With CSP, it is expected that the three classes would have better separation than without it, thus making it a useful tool in three-state motor imagery classification. LDA was consistently found to outperform its quadratic variant, not only for three-state classification but also for the differential and parallel modes of the two-classifier approach. This seems a bit counter-intuitive, as one could argue that if a linear separating hyperplane is optimal, this is well within the reach of QDA; but if a non-linear hyperplane would be more appropriate, the linear discriminant would clearly be sub-optimal. Moreover, LDA assumes shared covariance matrices, which is clearly not the case with NC and any of the IC states. One explanation could be that non-linear classifiers are more susceptible to overfitting and outliers due to the non-linear separating hyperplanes. This problem is well known for SVMs, where a very small width of the Gaussian kernel can lead to such a high overfitting that the decision boundary would consist of tight hyperspheres around each training instance. The hyperplanes determined by QDA would never have this exaggerated flexibility though, but it might suffer from the same problem on a smaller scale. Another possible reason why LDA gave superior results could ironically stem from its false assumption of shared covariance between classes. While QDA estimates a separate covariance matrix for each class, LDA only estimates the pooled covariance of the data. This means that more samples are available, potentially improving the robustness of the covariance estimator in the presence of outliers.

2 Parameter tuning and test results

For both classification and regression, the first step was to determine how the feature processing pipeline affects performance. It turned out that in both cases much higher detection rates can be obtained with no smoothing applied, regardless of whether it is a median filter, a moving average window or the additional sliding window used in the initial classifier tests. While less smoothing is beneficial for higher true positive rates, this comes at the expense of much lower hold times, which effectively turns the state-driven BCI into an event-driven one. And because of the very short hold time, the dwell time was unsuccessful in improving performance. This is also due to the non-overlapped short time windows from which band power features are extracted. In other works, the duration of the window is typically longer, generally up to 2.5 seconds, and the windows are overlapped. This is a form of smoothing in itself and potentially allows the dwell time to be effective.

Unlike the dwell time, refractory periods were found to significantly increase detection rates, although any improvement came at the cost of significantly more false positives on the test set. It seems that the larger the reduction of sample FPR is in cross-validation, the larger the event FPR will be on the test set. This observation is presented in Table 9, page 77, where we show that the relative decrease of sample FPR in cross-validation is indeed proportional to the relative increase of event FPR on the test set. Crucially, this relation is similar for all subjects, which seems

to indicate that in this context, refractory periods have almost no generalization ability on new data. This is because of the large differences in experimental protocol between the training set and the test set. While in the former all imaginary movements last four seconds, with four seconds breaks between them, the duration of both IC and NC periods is variable in the independent test set. The optimal refractory period is thus inherently adapted to the temporal structure of the data, as it crucially depends on the average duration of NC periods. Thus, the choice of an appropriate duration of the refractory period should not be based on the performance obtained on data with a fixed temporal structure, as is the case of our calibration set; otherwise, the improvements will not generalize to new data with different characteristics.

Both the dwell time and the refractory period are useful in decreasing false positive rate. Hence, it turns out that there are two possibilities to use these tools. The first possibility is to use a fixed threshold and determine the configuration for which the difference between true and false positive rate is maximized. This is the approach of Townsend et al **[43]**, who first introduced dwell and refractory periods for BCI. The second possibility is to fix the false positive rate, rather than the threshold, and then similarly find the dwell and refractory periods which maximize the difference between true and false positive rates. This is the approach used in this study. The former possibility might not improve detection rates, as the threshold of the classifier is still the same, but will most likely reduce false activations; the latter possibility is guaranteed to improve the true positive rate (provided that the FP rate can be decreased), but the false positive rate remains the same. Which approach generalizes better on an independent test set remains a mystery, because only the second possibility was investigated in this study. But indeed, our results show that the improvements in true positive rate not only hold quite well on the test set, but are comparatively larger than those found in cross-validation.

For classification, we have also investigated the influence of the selection of training intervals. While the moderate improvements found in cross-validation were insignificant on the test set, the results offer a valuable insight: for three-state classification in self-paced motor imagery BCI, the selection of training intervals for one IC state influences the detection rates obtained for the other one. If the same interval would have been optimal for both IC states, the largest values of the average detection rates from Figure 34 (page 65) would have been found on the diagonal, yet this was not the case. This conclusion should be taken into account for future research, considering that in most BCI experiments the same intervals are considered for both classes. One possible reason why this is not commonly investigated is the computational inefficiency of cross-validating the classifier for all possible configurations. While this is still feasible with simple models such as discriminant classifiers, it would not be appropriate for more complex ones such as support vector machines or hidden Markov models. An alternative could be a filter approach, where instead of cross-validating a classifier for each set of parameters, statistical measures of the data would be calculated and optimized.

Overall, linear regression was found to offer lower performance than classification. Unlike classification, which tries to separate the data belonging to different classes according to some criterion, the least squares regression used here is crucially dependent on the target values. Minimizing the mean squared error, the optimization criterion of the Wiener filter, is also not a wise choice given the vague relation between mean squared error and self-paced performance. It was shown that intuitive but improper target values can degrade performance significantly **[60]**. However, taking the one-dimensional projection found by LDA as the targets for the Wiener filter did not improve performance but actually decreased it. The genetic algorithm managed to improve cross-validation performance, but as usual, this came at the expense of higher event false positive rate on the test set. This brute force approach to optimization was possible because of the simple and computationally inexpensive regression model.

The final attempt to increase performance was by linearly combining the output of several classifiers and that of the Wiener filter with a genetic algorithm. Although four different optimization criteria were tested, results were unsatisfactory and any improvement in true positive rate also increased the event false positive rate on the test set. This might be caused by the fact that all classifiers were trained with the same features, extracted from the same intervals within the trial, and the same processing steps were applied for each of them. These parameters were found optimal for the three-state LDA, but there is no guarantee that they also work best for the other classifiers. Therefore, there is the possibility that performance could have been improved by using different pipelines, individually optimal for each classifier. Also, one could argue that instead of genetic algorithms, meta-classifiers could have been used, with the additional benefit of potentially determining non-linear relations between the scores of the base classifiers. We did not properly investigate this possibility, but some preliminary tests were performed with LDA, QDA and decision trees used as meta-classifiers. The results were not consistent though, and additional difficulties arose, such as which base classifiers should be used for training the meta-classifier. The genetic algorithm approach was thus preferred, partly because it offered the choice of specifically optimizing different performance metrics.

3 Performance metrics and target applications

The results on the independent test set suggest that the most difficult problem to deal with is reducing the event false positive rate. This begs the question of whether it would have been a better idea to optimize everything based on event, rather than sample FPR. However, the results also show that that sample and event FPR are proportional, and this holds quite well both across different subjects and different methods, suggesting that the approach of minimizing sample FPR is valid. Another possibility would have been to maximize the true-false difference. However, the results of linearly combining different classifiers with a genetic algorithm showed that optimization of the true-false difference gives roughly the same results in terms of event FPR as directly maximizing the true positive rate at a fixed sample FPR.

The most intriguing observations regarding performance metrics stem from the regression results from Figure 38, page 70. An inverse relation was found between true positive rates and both mean squared error and mutual information. Considering how similar MSE and mutual information appear to be in Figure 38, this apparently suggests that the two metrics are closely related and offer the same information. Whereas MSE was never perceived as an appropriate metric in this context, mutual information is generally regarded as a highly informative measure. This might be simply a coincidence for this particular dataset, but it does suggest that more detailed investigations should be performed. Then there is the inverse relation of mutual information with true positive rate, which means that one of the two metrics is not an appropriate indicator of performance. This is most disturbing, because while mutual information is based on the sound mathematical framework of information theory, event true positive rate is measured in an artificial way, as merely one correctly detected sample during an event is enough to declare the whole event as a true positive. We must stress the fact that this is not some personal interpretation of event TPR, but rather its formal definition in all self-paced BCI research **[8, 43]**. Its inverse relation to mutual information suggests however that this definition should be changed, or at least complemented with additional constraints, such as a minimum hold time or the requirement that no other IC state is detected during the event. But in any case, this performance measure weighs samples preferentially, with no clear relation to the underlying features of the data, which explains the inversely proportional relation with the two other metrics, mean squared error and the mutual information.

So does this mean that the TPR metric is useless in measuring self-paced performance? We believe this is generally not the case, but it really depends on the particular type of BCI and the target application. In the case of event-driven BCIs, the duration of the neurophysiological phenomenon that we want to detect is known a priori, and is typically much shorter than the four second trials considered here. The expected response window, during which we expect an IC-related activation, is typically only one second long **[42]**. Thus, it makes sense to declare the event as a true positive if the BCI had at least one correct activation (or more generally, if the dwell time was met) during the expected response window. Because of the short durations, the average change of EEG features during the expected response window is generally lower than in a four second motor imagery trial, for instance. To summarize, we argue that for event-driven BCIs, the TPR metric in its present definition makes much more sense and quite possible there is also a stronger relation between EEG features and its labels. Thus, in the case of event-driven BCIs, we believe that the true positive rate and the mutual information would correlate as expected, and not be inversely proportional as is the case here.

It thus becomes apparent that mutual information is more related to the hold time, rather than the true positive rate. This is based not only on thought experiments, but is backed by the combined data from Figure 31 and Figure 38 (pages 60 and 70, respectively). We have shown that not performing any smoothing operations increases true positive rates considerably, but also substantially decreases both the average hold time and the mutual information.

If so, what is the relation between a subject's perception of self-paced BCI performance, true positive rate, and hold time in the case of state-driven designs? The answer depends crucially on whether the hold time is important for the target application. Of course, one could argue that in almost any possible application we have the option of keeping a button pressed, rather than only clicking it, but again, whether this functionality is required or useful depends on the application. To illustrate this with a non-BCI example, consider the computer mouse. Provided no movements of the mouse, keeping a button pressed in the desktop environment is of no use after clicking it. But in a first person shooter, this functionality becomes important for obvious reasons.

Therefore consider a virtual keyboard, where users select letters on the screen through motor imagery. In this case, the hold time is of practically no importance. The user would start motor imagery and just keep trying until the BCI reacts and the selection changes. With no intention of keeping a button pressed, the user would only be interested in a good detection rate and perhaps a quick response time. This is a situation in which a BCI such as the one presented in this paper would be appropriate. Let us now consider a motor imagery-controlled racing game and for the sake of simplicity, let us assume that the computer accelerates and brakes automatically. This leaves us with a three-state self-paced BCI. While detection rate is of course still important, the hold time now plays a key role as it dictates the amount of left or right steering. This means that the hereby BCI would not be appropriate for the task, and it also shows that for this kind of fast-paced application, dwell and refractory periods would also be inappropriate.

Let us discuss the usefulness of the hereby BCI in practice. Even on the test set, sample FPR was still kept low, at an average of 1%. At a 10 Hz decision rate, this would correspond to about six false activations per minute. Whether this is acceptable or not again depends on the target application. Consider a BCI-controlled motorized wheelchair, which the user navigates through a normal city environment. Imagine that the person controlling the wheelchair wants to cross a busy intersection and patiently waits for the green light. Even one false activation in this scenario could be catastrophic, let alone six of them per minute. This is an extreme example requiring an almost 0% false positive rate, and no current BCI is safe enough to be used in high-risk, real-world applications. A better example would be to use a three-state BCI for changing TV channels. This is

a low-risk, real-life application, which also demands that the BCI is able to handle long periods of NC. A false activation every ten seconds would surely not be convenient for the user. In such a context though, the response time is not a crucial factor, thus it is important to consider that the 10 Hz decision rate would probably not bring any benefits. Therefore, a lower decision rate could be used, for example of 1 Hz, theoretically corresponding to one false positive every one and a half minutes, which could be reasonable. Simply downsampling the BCI output would not be appropriate because of the low hold time, and hence would decrease detection rates considerably. A more sensible approach which would theoretically offer exactly the same performance would be to output an IC command if at least one sample was classified as IC during the past second. Decreasing the decision rate would not be appropriate for all applications however: putting aside the issue of low hold times for the argument's sake, in a fast-paced racing game as described previously, the response time would be more important, and a false activation every ten seconds might even go unnoticed by the user.

4 Limitations of the study

The results from a dataset with only four subjects are difficult to generalize and used to draw performance-oriented conclusions. However, it is noteworthy to mention that although the number of subjects is small by any objective measure, this is also the case with almost all self-paced BCI research. Most studies are based on the data of two **[51, 77]**, three **[43, 49]**, four **[78, 79]**, or five **[42, 52]** subjects. Moreover, the smaller dataset could have had its merits, as some of the unexpected results and observations might not have surfaced from a large dataset with many subjects.

To validate the results obtained here and draw truly significant conclusions, experiments should also be performed online, with a user in the loop. Furthermore, experimental protocols with true self-paced environments would be recommended, as opposed to the acoustic cues used for this dataset. This method of simulating self-paced operation is useful for knowing the precise timing of movements but it might not be a valid approach, as previous research has found differences in the readiness potentials of self-paced and cued movements **[80]**. Moreover, the subjects did not receive any feedback of their actions, which is normally a very important link in BCI. In the absence of any objective to motivate them, such as effectively controlling something, users might lose interest and focus, and boredom is more likely to kick in.

While this study was not meant to investigate the use of several neurophysiological phenomena, it is important to note that combining sensorimotor modulations with movement-related potentials was shown to improve performance considerably **[81]**. For the specific purpose of self-paced operation, this approach has only been tested for two-state BCIs, in which case it was shown to significantly decrease false positive rates **[79]**.

The two other major directions which were not explored are the use of different features related to SMR modulations, such as phase synchronization **[82]**, and the use of unsupervised methods such as hidden Markov models.

Chapter VI
Conclusions and future prospects

1 Translation algorithms

The most fundamental research question which we set out to answer concerned the choice of an appropriate translation algorithm. We have tested linear and quadratic discriminant classifiers, one-class support vector machines and least squares regression. Results indicate the discriminant classifiers to be the better choice among the options considered here, as they present higher average detection rates and lower inter-subject variability. We must emphasize however that this is not a performance-oriented study aiming to declare a specific algorithm as the overall winner, both because of the rather small number of subjects and because of the small number of possibilities considered here. There are several methods which were not tested, most notable being unsupervised learning with Gaussian mixture models, hidden Markov models or conditional random fields. Nevertheless, based on our results we argue in the favor of discriminant classifiers, mostly because of their simplicity. This makes numerous computationally-inefficient optimizations possible, and it is likely that the improvements brought by these optimizations more than make up for the benefits brought by more sophisticated methods. Put simply, discriminant classifiers can be feasibly embedded in wrapper approaches, whereas more complex algorithms such as SVM cannot. Another benefit of these simple models is that corrections can easily be applied for handling unbalanced data, which is a typical situation for self-paced BCIs. Thus, among supervised learning algorithms, we recommend discriminant classifiers as most appropriate for self-paced motor imagery BCIs.

Whether the linear or the quadratic variant is more suitable depends on the classifier configuration. In fact, the key issue investigated here was not necessarily related to the characteristics of specific classifiers; the differences between them are subtle when compared to the differences found between their various configurations. The quadratic discriminant provided exceptionally good results for the IC detector of two-stage classification, especially considering the improper features on which it was trained, due to the use of CSP (see discussion). While it is regrettable that two-stage classification was not tested with unsupervised spatial filtering, the insights from the previous chapter show this approach to be inherently sub-optimal due to the topographical characteristics of motor imagery. Another drawback is the higher complexity of such an approach. In its current form with unsupervised spatial filtering, we believe two-stage classification is not a promising direction for future research, despite its popularity in self-paced motor imagery studies.

The two other possibilities investigated were two-classifier configurations and direct three-state classification. The former approach presents some appealing properties in theory, which did not hold in practice however. Thus, our results indicate that the best approach, at least when CSP is used, is to directly perform three-state classification. This provided the lowest inter-subject variability and the highest overall performance. For both the two-classifier configuration and direct three-state classification, the linear discriminant consistently outperformed its quadratic variant. Whether this is due to higher generalization ability of the linear hyperplane or to more robust estimations of covariances is not entirely clear and can only be speculated upon. However, our

conclusion is that for CSP band power features a multi-class linear discriminant classifier is highly appropriate.

But our results also indicate that apparently there is no such thing as a globally optimal approach. We have shown that depending on which methods are used, the same subject can have both the best and the worst performance among all subjects. We have also seen that IC detection rate is not proportional to IC classification accuracy. It seems that self-paced performance is more likely a measure of the adequacy of the methods we employ rather than of some intrinsic measure of individual subject performance. We propose that in addition to subject-specific feature extraction, BCI research should also incorporate subject-specific translation algorithms. And by this, we refer not only to specific classifiers but also different configurations. This would not be a troublesome addition either, as it would not add much to the complexity, especially considering the information-theoretic feature extraction procedures which are normally applied. Thus, we consider that a promising direction would be to personalize and emulate BCIs on the data of each subject, instead of trying to come up with an ultimate do-it-all algorithm. Different classifiers and features can be used where appropriate, either for each subject or even for each mental state. Nowadays when we have quad-core smartphones in our pockets, the old argument of achieving maximum efficiency does not seem to hold. This is not to be misinterpreted that particularization is preferred over generalization. Surely, everyone would prefer an ultimate method for BCI that would prove optimal for all conditions. But in the absence of this, perhaps it would be better to also turn our attention to what really matters, which is delivering the best possible performance and satisfying user and application requirements as much as possible.

2 Dwell time and refractory period

The second research question concerned the possibility of increasing the performance and reliability of self-paced BCIs through specific tools, such as the dwell time and the refractory period. The dwell time was not properly investigated because of the choice we made at a certain point, which was a preference to high detection rates at the expense of a short hold time. Regarding the refractory period however, our results show that it lacks generalization ability on new data. This is because the optimal refractory period is crucially dependent on the average duration of NC periods. Thus, if the temporal structure of the calibration data does not match closely with the temporal characteristics and requirements of the target application, the refractory period will be of no use. We have shown that the decrease of sample FPR in the calibration data is proportional to the increase of event FPR on the independent test set, which had very different temporal characteristics. Using refractory periods and taking comfort in their apparent benefits is simply risky and we do not recommend it for real-life applications. Moreover, refractory periods decrease the maximum achievable information transfer rate by imposing a limit on the minimum duration of NC periods and are completely inappropriate for certain applications, as discussed in the previous chapter.

Despite the lack of a proper analysis, the drawbacks of refractory periods should not be automatically generalized to the dwell time as well. While the refractory period reduces the speed of operation, the dwell time reduces the response time of the BCI. However, except for certain cases in which very short IC commands are required, the dwell time seems a very useful tool for self-paced BCIs, as it does not depend on the duration of NC periods. As the refractory period imposes a minimum duration of NC periods, the dwell time imposes a minimum duration of IC commands. This is a far more useful approach, as the characteristics of IC periods are known beforehand and do not change as much in different settings. While it may also not be applicable for

applications requiring a very short response time, in most situations we strongly recommend the use of dwell times.

3 Performance metrics

Any situation which requires an educated choice also requires an adequate performance metric to be defined. This was the topic of our third and final research question and the discussions from the previous chapter show that the answer depends on the application. If the user does not need to maintain an IC state for variable durations, then the true positive rate is an adequate and informative measure. However, the definition of a correctly detected event should be better formalized, so that situations such as the detection of multiple IC states during an event are better dealt with. Results show that for motor imagery events, the current form of the TPR metric is inversely proportional to the mutual information. Interestingly, the mutual information seems to be directly proportional to the hold time. Thus, our suggestion is to augment the definition of event TPR with a minimum hold time and the requirement that no other IC state is detected during that event.

While this offline study cannot directly establish a link between the subject's perception of BCI performance and various metrics, we believe that real-world applications of motor imagery BCIs will undoubtedly benefit from precise control over the hold time. Whether for steering a BCI-controlled vehicle or for assigning different functions to an IC command, the hold time will most likely be the crucial metric given a satisfactory detection rate. This is also backed by the strong relation we found between the hold time and mutual information, which is a well-established and informative performance metric. Therefore, we think that the most informative performance metric for a state-driven BCI would include both the true positive rate and the hold time in a single metric. Of course, this assumption can only be validated by online studies which would correlate the subject's personal view on BCI performance with different performance metrics.

4 Future work

4.1 Supervised spatial filtering for self-paced BCI

The results of two-stage classification provided valuable insights regarding the limitations of such designs. The only spatial filtering methods currently available for such a purpose are unsupervised. Thus, it is reasonable to assume that supervised spatial filtering would increase performance. CSP is clearly inappropriate in this context, as maximizing the differences between IC states is contrary to the assumption that they both form one class. With CSP, the two IC states are pulled apart as much as possible in the hope that increasing their separation will also separate them from NC. However, we cannot know whether this is the case. What is therefore needed is a way of finding the similarities between IC states which jointly make them as separable as possible from NC. Having two projections, one that would separate the two IC states the most (CSP) and one that would make them as identical as possible while separating them from NC could provide complementary and useful information in the context of self-paced IC detection.

4.2 Online study

To validate our results and assumptions, an online self-paced BCI should be implemented, and feedback should be presented to the user. We propose two applications, which were both discussed in the previous chapter.

The first is a racing game controlled through motor imagery. The crucial improvement we propose here is in user training. Instead of the standard approach of cued motor imagery training, we think that the user would be more entertained by watching an AI-controlled car driving around the track. The user would be instructed to try and imagine him/herself driving the car. The main advantage of this approach is that it would be a truly self-paced environment in which the true labels are known, because the computer follows a known path. User adaptation would not be an issue, because the driving skills of the computer can be adjusted in most games and because different cars would have different characteristics and would result in different paths which the computer would take. Moreover, this could be implemented with either two or three IC states: a three-state BCI would allow the user to steer the car while the computer deals with braking and acceleration, and a four-state BCI would additionally give the user control over the brakes. Different levels of difficulty could be set by limiting the cars to a certain top speed. The feedback would thus be natural and entertaining to the user, and the desire to perform as good as possible could be ensured by having a list of the best BCI drivers for a given track. Furthermore, the possibility of multiplayer BCI racing games is exhilarating.

The second application is not as adrenaline-packed, but we believe it would be highly informative nonetheless. In the racing game described above, frequent activations are expected, thus it would not provide a good testing environment for long periods of NC. Of course, a simulator could be used instead, where the user would drive for tens of kilometers, but it could become boring. In our view at least, one of the most desired applications of BCI would be the control of a motorized wheelchair. The problem in this context has been discussed in the previous chapter: even a single false activation could be catastrophic if the user wants to cross a busy intersection and is waiting for the green light. It thus becomes apparent that the BCI needs to be evaluated in situations with long NC periods. Experimental protocols of navigating a wheelchair in a small room with obstacles or in virtual environments do not offer these possibilities. We therefore propose to evaluate the BCI in the context of changing TV channels. The frequency of activations is irregular and long periods of NC are common if the user finds a nice movie or some other long TV program.

4.3 Performance metrics

Which is the most appropriate performance metric for self-paced BCIs is still an open question. We propose that the most informative measure for state-driven designs would combine both the true positive rate and the hold time, although at the moment we do not know how exactly this would be achieved. Simple linear combinations could be experimented with, but most importantly, they would need to be correlated with the user's personal perception on BCI performance. The two online BCIs above could provide a good testing environment.

Bibliography

[1] H. Berger, *Über das Elektroenzephalogramm des Menschen.* Archiv Psychiatrischer Nervenkrankheiten, 1929. **87**: p. 527-580.

[2] J.R. Wolpaw, et al., *Brain-computer interfaces for communication and control.* Clinical neurophysiology, 2002. **113**(6): p. 767-791.

[3] A. Kubler, et al., *BCI meeting 2005-workshop on clinical issues and applications.* Neural Systems and Rehabilitation Engineering, IEEE Transactions on, 2006. **14**(2): p. 131-134.

[4] R.B. Stein and V. Mushahwar, *Reanimating limbs after injury or disease.* Trends in neurosciences, 2005. **28**(10): p. 518-524.

[5] Y. Wang, Y.T. Wang, and T.P. Jung, *Visual stimulus design for high-rate SSVEP BCI.* Electronics letters, 2010. **46**(15): p. 1057-1058.

[6] B. Blankertz, et al., *The non-invasive Berlin Brain-Computer Interface: Fast acquisition of effective performance in untrained subjects.* NeuroImage, 2007. **37**(2): p. 539-550.

[7] S.G. Mason and G.E. Birch, *A general framework for brain-computer interface design.* Neural Systems and Rehabilitation Engineering, IEEE Transactions on, 2003. **11**(1): p. 70-85.

[8] S.G. Mason, et al., *Evaluating the performance of self-paced brain computer interface technology.* Neil Squire Soc., Vancouver, BC, Canada, Tech. Rep, 2006.

[9] J.P. Donoghue, *Connecting cortex to machines: recent advances in brain interfaces.* Nature Neuroscience, 2002. **5**: p. 1085-1088.

[10] J.M. Carmena, et al., *Learning to control a brain–machine interface for reaching and grasping by primates.* PLoS biology, 2003. **1**(2): p. e42.

[11] L.R. Hochberg, et al., *Neuronal ensemble control of prosthetic devices by a human with tetraplegia.* Nature, 2006. **442**(7099): p. 164-171.

[12] E.C. Leuthardt, et al., *The emerging world of motor neuroprosthetics: a neurosurgical perspective.* Neurosurgery, 2006. **59**(1): p. 1-14.

[13] G. Schalk, et al., *Decoding two-dimensional movement trajectories using electrocorticographic signals in humans.* Journal of Neural Engineering, 2007. **4**(3): p. 264.

[14] Z.C. Chao, Y. Nagasaka, and N. Fujii, *Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkeys.* Frontiers in neuroengineering, 2010. **3**.

[15] J.R. Wolpaw, et al., *BCI meeting 2005-workshop on signals and recording methods.* Neural Systems and Rehabilitation Engineering, IEEE Transactions on, 2006. **14**(2): p. 138-141.

[16] L.F. Nicolas-Alonso and J. Gomez-Gil, *Brain Computer Interfaces, a Review.* Sensors, 2012. **12**(2): p. 1211-1279.

[17] R. Elul, *The genesis of the EEG.* Int Rev Neurobiol, 1972. **15**: p. 227-272.

[18] W. Singer, *Synchronization of cortical activity and its putative role in information processing and learning.* Annual Review of Physiology, 1993. **55**(1): p. 349-374.

[19] L.A. Bradshaw, R.S. Wijesinghe, and J.P. Wikswo, *Spatial filter approach for comparison of the forward and inverse problems of electroencephalography and magnetoencephalography.* Annals of Biomedical Engineering, 2001. **29**(3): p. 214-226.

[20] J. Kaiser, et al., *Cortical oscillatory activity during spatial echoic memory.* European Journal of Neuroscience, 2005. **21**(2): p. 587-590.

[21] M. van Gerven and O. Jensen, *Attention modulations of posterior alpha as a control signal for two-dimensional brain–computer interfaces.* Journal of neuroscience methods, 2009. **179**(1): p. 78-84.

[22] J.H. Lee, et al., *Brain–machine interface via real-time fMRI: preliminary study on thought-controlled robotic arm.* Neuroscience Letters, 2009. **450**(1): p. 1-6.

[23] S.M. Coyle, T.E. Ward, and C.M. Markham, *Brain–computer interface using a simplified functional near-infrared spectroscopy system.* Journal of Neural Engineering, 2007. **4**(3): p. 219.

[24] S. Fazli, et al., *Enhanced performance by a hybrid NIRS–EEG brain computer interface.* NeuroImage, 2012. **59**(1): p. 519-529.

[25] G. Pfurtscheller and FH Lopes da Silva, *Event-related EEG/MEG synchronization and desynchronization: basic principles.* Clinical neurophysiology, 1999. **110**(11): p. 1842-1857.

[26] J.N. Mak and J.R. Wolpaw, *Clinical applications of brain-computer interfaces: current state and future prospects.* Biomedical Engineering, IEEE Reviews in, 2009. **2**: p. 187-199.

[27] D.J. McFarland, W.A. Sarnacki, and J.R. Wolpaw, *Electroencephalographic (EEG) control of three-dimensional movement.* Journal of Neural Engineering, 2010. **7**(3): p. 036007.

[28] J.R. Wolpaw and D.J. McFarland, *Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(51): p. 17849-17854.

[29] J.J. Vidal, *Toward direct brain-computer communication.* Annual Review of Biophysics and Bioengineering, 1973. **2**(1): p. 157-180.

[30] L.A. Farwell and E. Donchin, *Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials.* Electroencephalography and Clinical Neurophysiology, 1988. **70**(6): p. 510-523.

[31] N. Birbaumer, *Slow cortical potentials: Plasticity, operant control, and behavioral effects.* The Neuroscientist, 1999. **5**(2): p. 74-78.

[32] N. Birbaumer, et al., *The thought translation device (TTD) for completely paralyzed patients.* Rehabilitation Engineering, IEEE Transactions on, 2000. **8**(2): p. 190-193.

[33] T.M. Vaughan, J.R. Wolpaw, and E. Donchin, *EEG-based communication: prospects and problems.* Rehabilitation Engineering, IEEE Transactions on, 1996. **4**(4): p. 425-430.

[34] S.G. Mason, et al., *A comprehensive survey of brain interface technology designs.* Annals of Biomedical Engineering, 2007. **35**(2): p. 137-169.

[35] B. Blankertz, et al., *The Berlin Brain--Computer Interface: accurate performance from first-session in BCI-naive subjects.* Biomedical Engineering, IEEE Transactions on, 2008. **55**(10): p. 2452-2462.

[36] A. Bashashati, *Towards development of a 3-state self-paced Brain Computer Interface system*, 2007.

[37] G. Pfurtscheller, et al., *The hybrid BCI.* Frontiers in neuroscience, 2010. **4**.

[38] R. Scherer, et al., *An asynchronously controlled EEG-based virtual keyboard: improvement of the spelling rate.* Biomedical Engineering, IEEE Transactions on, 2004. **51**(6): p. 979-984.

[39] A. Schlogl, et al., *19 Evaluation Criteria for BCI Research.* Toward brain-computer interfacing, 2007: p. 327.

[40] W. Wu, et al., *Bayesian population decoding of motor cortical activity using a Kalman filter.* Neural Computation, 2006. **18**(1): p. 80-118.

[41] D.J. Hand and R.J. Till, *A simple generalisation of the area under the ROC curve for multiple class classification problems.* Machine Learning, 2001. **45**(2): p. 171-186.

[42] S.G. Mason and G.E. Birch, *A brain-controlled switch for asynchronous control applications.* Biomedical Engineering, IEEE Transactions on, 2000. **47**(10): p. 1297-1307.

[43] G. Townsend, B. Graimann, and G. Pfurtscheller, *Continuous EEG classification during motor imagery-simulation of an asynchronous BCI.* Neural Systems and Rehabilitation Engineering, IEEE Transactions on, 2004. **12**(2): p. 258-265.

[44] G.E. Birch, Z. Bozorgzadeh, and S.G. Mason, *Initial on-line evaluations of the LF-ASD brain-computer interface with able-bodied and spinal-cord subjects using imagined voluntary motor potentials.* Neural Systems and Rehabilitation Engineering, IEEE Transactions on, 2002. **10**(4): p. 219-224.

[45] Z. Yu, S.G. Mason, and G.E. Birch. *Enhancing the performance of the LF-ASD brain-computer interface*. in *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual*

Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint. 2002. IEEE.

[46] J.F. Borisoff, et al., *Brain-computer interface design for asynchronous control applications: improvements to the LF-ASD asynchronous brain switch.* Biomedical Engineering, IEEE Transactions on, 2004. **51**(6): p. 985-992.

[47] M. Fatourechi, et al. *A hybrid genetic algorithm approach for improving the performance of the LF-ASD brain computer interface*. in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*. 2005. IEEE.

[48] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, *Optimal spatial filtering of single trial EEG during imagined hand movement.* Rehabilitation Engineering, IEEE Transactions on, 2000. **8**(4): p. 441-446.

[49] R. Scherer, et al., *The self-paced Graz brain-computer interface: methods and applications.* Computational Intelligence and Neuroscience, 2007. **2007**.

[50] M. Pregenzer and G. Pfurtscheller, *Frequency component selection for an EEG-based brain to computer interface.* Rehabilitation Engineering, IEEE Transactions on, 1999. **7**(4): p. 413-419.

[51] Y. Chae, J. Jeong, and S. Jo. *Noninvasive brain-computer interface-based control of humanoid navigation*. in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*. 2011. IEEE.

[52] B.A.S. Hasan and J.Q. Gan, *Hangman BCI: An unsupervised adaptive self-paced Brain–Computer Interface for playing games.* Computers in Biology and Medicine, 2012.

[53] G. Schalk, et al., *Brain–computer interfaces (BCIs): detection instead of classification.* Journal of neuroscience methods, 2008. **167**(1): p. 51-62.

[54] H. Peng, F. Long, and C. Ding, *Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2005. **27**(8): p. 1226-1238.

[55] P.L. Nunez, et al., *EEG coherency: I: statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales.* Electroencephalography and Clinical Neurophysiology, 1997. **103**(5): p. 499-515.

[56] Z.J. Koles, M.S. Lazar, and S.Z. Zhou, *Spatial patterns underlying population differences in the background EEG*. Brain Topography, 1990. **2**(4): p. 275-284.

[57] B. Blankertz, et al., *Optimizing spatial filters for robust EEG single-trial analysis.* Signal Processing Magazine, IEEE, 2008. **25**(1): p. 41-56.

[58] J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg, *Designing optimal spatial filters for single-trial EEG classification in a movement task.* Clinical neurophysiology, 1999. **110**(5): p. 787-798.

[59] G.R. Arce, *Nonlinear signal processing: a statistical approach*2004: Wiley-Interscience.

[60] H. Zhang and C. Guan, *A maximum mutual information approach for constructing a 1D continuous control signal at a self-paced brain–computer interface.* Journal of Neural Engineering, 2010. **7**(5): p. 056009.

[61] H. Zhang, et al., *BCI competition IV–data set I: learning discriminative patterns for self-paced EEG-based motor imagery detection.* Frontiers in neuroscience, 2012. **6**.

[62] A Satti, D Coyle, and G Parasad, *Optimal frequency band selection with particle swarm optimization for a brain computer interface*. IEEE Computational Intelligence Society Workshop and Summer School in Evolutionary Computation, 2008.

[63] K.K. Ang, et al. *Filter bank common spatial pattern (FBCSP) in brain-computer interface*. in *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*. 2008. IEEE.

[64] H. Zhang, et al. *Spatio-spectral feature selection based on robust mutual information estimate for brain computer interfaces*. in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. 2009. IEEE.

[65] Z. Szabó, B. Póczos, and A. Lőrincz, *Separation theorem for independent subspace analysis and its consequences.* Pattern Recognition, 2012. **45**(4): p. 1782-1791.

[66] Z. Szabó, B. Póczos, and A. Lőrincz, *Undercomplete blind subspace deconvolution.* The Journal of Machine Learning Research, 2007. **8**: p. 1063-1095.

[67] E. Alpaydin, *Introduction to machine learning*2004: MIT press.

[68] F. Lotte, et al., *A review of classification algorithms for EEG-based brain–computer interfaces.* Journal of Neural Engineering, 2007. **4**.

[69] C. Vidaurre, et al. *Unsupervised adaptation of the LDA classifier for Brain-Computer Interfaces*. in *Proceedings of the 4th International Brain-Computer Interface Workshop and Training Course*. 2008.

[70] C.J.C. Burges, *A tutorial on support vector machines for pattern recognition.* Data Mining and Knowledge Discovery, 1998. **2**(2): p. 121-167.

[71] C.C. Chang and C.J. Lin, *LIBSVM: a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology (TIST), 2011. **2**(3): p. 27.

[72] K.R. Muller, C.W. Anderson, and G.E. Birch, *Linear and nonlinear methods for brain-computer interfaces.* Neural Systems and Rehabilitation Engineering, IEEE Transactions on, 2003. **11**(2): p. 165-169.

[73] D. Garrett, et al., *Comparison of linear, nonlinear, and feature selection methods for EEG signal classification.* Neural Systems and Rehabilitation Engineering, IEEE Transactions on, 2003. **11**(2): p. 141-144.

[74] H.W. Sorenson, *Least-squares estimation: from Gauss to Kalman.* Spectrum, IEEE, 1970. **7**(7): p. 63-68.

[75] S. Haykin, *Adaptive filter theory. 3rd ed*1996, Upper Saddle River, NJ: Prentice-Hall International.

[76] A. Ben-Hur and J. Weston, *A user's guide to support vector machines.* Methods in Molecular Biology, 2010. **609**: p. 223-239.

[77] S. Fazli, et al., *Using rest class and control paradigms for brain computer interfacing.* Bio-Inspired Systems: Computational and Ambient Intelligence, 2009: p. 651-665.

[78] A. Bashashati, R.K. Ward, and G.E. Birch, *Towards development of a 3-state self-paced brain-computer interface.* Computational Intelligence and Neuroscience, 2007. **2007**: p. 9.

[79] M. Fatourechi, RK Ward, and GE Birch, *A self-paced brain–computer interface system with a low false positive rate.* Journal of Neural Engineering, 2007. **5**(1): p. 9.

[80] B. Libet, E.W. Wright, and C.A. Gleason, *Readiness-potentials preceding unrestricted 'spontaneous' vs. pre-planned voluntary acts.* Electroencephalography and Clinical Neurophysiology, 1982. **54**(3): p. 322-335.

[81] G. Dornhege, et al., *Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms.* Biomedical Engineering, IEEE Transactions on, 2004. **51**(6): p. 993-1002.

[82] J. Gu and R. Ward. *Novel feature generation and classification for a 2-state Self-paced Brain Computer Interface system*. in *Electrical & Computer Engineering (CCECE), 2012 25th IEEE Canadian Conference on*. 2012. IEEE.