Advancing Pattern Recognition Techniques for Brain-Computer Interfaces: Optimizing Discriminability, Compactness, and Robustness

Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften

von der Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Dominic Heger
aus Karlsruhe

Tag der mündlichen Prüfung:	16. 07. 2015
Erste Gutachterin:	Prof. Dr.-Ing. T. Schultz
Zweiter Gutachter:	Prof. Dr.-Ing. R. Stiefelhagen

Acknowledgements

Without the inspiration and support of many people it is, generally, not possible to undertake and successfully complete a challenging endeavor, such as this dissertation. Therefore, I would like to gratefully thank Prof. Tanja Schultz for the great opportunity to work in this exciting field of research. Under her supervision I enjoyed a maximum amount of freedom to realize my own ideas, which allowed me to explore and build expertise in multiple directions of BCI research, data analysis, and machine learning that greatly enabled me grow as a research scientist. Next, I would like to thank Prof. Rainer Stiefelhagen for co-advising my dissertation. Thank you for all your support and advice. A very special thanks goes to my colleagues and friends at the Cognitive Systems Lab for the inspiring discussions and for all the laughs and frustrations we shared. It was a pleasure to work with you: Alfred Schmidt, Christian Herff, Christoph Amma, Dirk Gehrig, Dominic Telaar, Felix Putze, Helga Scherer, Jochen Weiner, Marcus Georgi, Matthias Janke, Thang Vu, and Tim Schlippe. From this list I would like to point out Christian and Felix, with whom I collaborated particularly closely on two important research topics that became parts of this dissertation, namely Brain-to-Text and EEG-based workload recognition. Furthermore, I would like to thank all students who contributed to my research by participating in the experiments and who worked extremely hard to create inspiring final theses under my supervision. I would like to gratefully thank my family for their unconditional love and support in all situations of life. Finally, an extraordinary thanks to Ellen for her never-ending patience and understanding during the challenging moments while working on this dissertation. Thanks for reminding me of the other important things beyond research.

Zusammenfassung

Die Vision von Schnittstellen zwischen Mensch und Maschine, die auf der Gehirnaktivität des Benutzers basieren, inspiriert Forscher und Science-Fiction Autoren gleichermaßen. Seit mehr als 15 Jahren wird international intensiv an Gehirn-Computer Schnittstellen (Brain-Computer Interfaces, BCIs) geforscht. BCIs werden zunehmend in klinischen Anwendungen eingesetzt, z.B. zur Kommunikation für Locked-in Patienten oder bei der Rehabilitation nach einem Schlaganfall. Als Benutzerschnittstellen für gesunde Personen finden BCIs vorwiegend in der Unterhaltungselektronik erste Anwendungen. Trotz zahlreicher Fortschritte in der BCI Forschung und vielfältiger potentieller Anwendungen für BCIs, unterliegen gehirnaktivitätsbasierte Schnittstellen immer noch großen Einschränkungen bezüglich ihrer praktischen Anwendbarkeit. Aktuelle BCIs haben in der Regel lange Kalibierungsphasen und ermöglichen nur einen sehr geringen Informationsdurchsatz von wenigen Bits pro Sekunde. Ihre Zuverlässigkeit ist, aufgrund des kleinen Signal-Rausch-Abstands von Gehirnaktivitätsmustern und aufgrund von Nichtstationaritäten der Signale, gering. Darüber hinaus sind die aktuell verwendeten Interaktionsparadigmen, die beispielsweise auf externen visuellen Stimuli basieren, oft unnatürlich und unflexibel. Um diesen Problemen entgegen zu treten, werden in dieser Dissertation wichtige Beiträge zur Mustererkennungs-Komponente, der zentralen Komponente eines modernen BCIs, entwickelt. Die Hauptergebnisse der Arbeit lassen sich wie folgt zusammenfassen:

Zentrale Zielkriterien für die Mustererkennung von BCIs
Um die Mustererkennung für BCIs systematisch weiterzuentwickeln, formulieren wir die Hypothese, dass drei zentrale Zielkriterien, Discriminative, Compact und Robust für die Mustererkennung von BCIs notwendige Bedingungen darstellen (Kapitel 3):

- Discriminative: Identifikation von Gehirnaktivitäts-Mustern, die die Unterscheidung von verschiedenen mentalen Aktivitäten und Benutzerzuständen erlauben.

- COMPACT: Kompakte Modellierung relevanter Aspekte von Gehirnaktivitäts-Mustern in interpretierbaren und generalisierenden Strukturen.
- ROBUST: Robustheit gegenüber Signal-Variabilitäten in Gehirnaktivitätssignalen, die nicht durch das BCI Paradigma moduliert werden.

Wir untersuchen diese drei Zielkriterien hinsichtlich der besonderen Relevanz für BCIs und setzen sie in Bezug zu fundamentalen Prinzipien aus der Maschinellen Lerntheorie. Während die Prinzipien DISCRIMINATIVE und COMPACT in der Mustererkennung weitreichend bekannt sind, wird unserem Wissen nach hier zum ersten Mal die Notwendigkeit der drei Zielkriterien DISCRIMINATIVE, COMPACT und ROBUST, sowie ihre wechselseitige Abhängigkeit, die eine gemeinsame Optimierung impliziert, analysiert.

Entwicklung eines generischen Rahmenwerks für die Mustererkennung von BCIs

Ein zentraler Beitrag dieser Dissertation besteht in der Entwicklung eines BCI Optimierungs-Rahmenwerks (*DCR Framework*), das die drei oben genannten Zielkriterien für die Musterkerkennung zum ersten Mal in einem gemeinsamen Optimierungsalgorithmus vereint (Kapitel 4). Dazu werden die drei Zielkriterien in einem konvexen Optimierungsproblem durch eine Least-Squares Regression mit ℓ_1-Norm und Sum-of-Norms Regularisierungen formalisiert. Zur Lösung dieses Problems wurde ein effizienter Optimierungsalgorithmus entwickelt, der auf der Alternating Direction Method of Multipliers basiert. Dabei nutzt das *DCR Framework* generische, hochdimensionale Merkmale im Zeit- und Frequenzbereich und bietet die innovative Funktionalität, Richtungen im Merkmalsraum zu definieren, gegenüber denen die gelernten Modelle invariant und somit robuster gegenüber Einflüssen von Signalvariabilitäten werden.

Wir evaluieren das *DCR Framework* durch acht unterschiedliche BCI Datensätze mit EEG, fNIRS und ECoG Signalen und zwei synthetisch generierten Datensätzen (Kapitel 5 und 6). Dabei zeigen wir, dass die Erkennungsleistung der vorgeschlagenen Methoden dem neusten Stand der Forschung entsprechen und zahlreiche, aktuelle, alternative Methoden übertreffen. Die Auswertungen beinhalten mehrere, öffentlich verfügbare Benchmark-Datensätze, wie beispielsweise den Datensatz der BCI Challenge @ NER2015. In diesem Wettbewerb wurden 260 Systeme miteinander verglichen. Unsere Einreichung, die auf dem *DCR Framework* basiert, konnte das sechstbeste Ergebnis erzielen und gewann den zweiten Preis bei der internationalen IEEE Neural Engineering Konferenz 2015. Eine stringente mathematische Umsetzung der drei Zielkriterien im *DCR Framework* ermöglicht es, das *DCR Framework* auf eine Vielzahl verschiedener BCI Erkennungsprobleme mit sehr unterschied-

lichen Signalcharakteristiken ohne grundlegende Anpassungen anzuwenden. Diese Flexibilität deutet darauf hin, dass der generische Ansatz der drei Zielkriterien nicht nur notwendige, sondern auch hinreichende Bedingungen für eine Vielzahl von BCI Mustererkennungsproblemen darstellt.

Erkennungssysteme für innovative BCI Paradigmen
Ein weiterer Hauptbeitrag dieser Dissertation sind zwei Studien, anhand derer wir zeigen, wie die drei Zielkriterien bei der Entwicklung von Erkennungssystemen für neue BCI Paradigmen erfolgreich umgesetzt werden können für die es bisher keine etablierten Abfolgen von Mustererkennungsschritten gibt. Dabei erfordern die beiden BCI Paradigmen keinen Lernaufwand auf der Seite des Benutzers, stattdessen wird die spontan entstehende Gehirnaktivität des Benutzers analysiert und interpretiert. Wir gehen auf die EEG-basierte Erkennung mentaler Belastungszustände (Workload-Erkennung) und die Erkennung von Lauteinheiten bei kontinuierlich gesprochener Sprache, anhand invasiv gemessener ECoG-Signale ein.
In der Workload-Studie konnte durch die Anpassung des Systems an die erkannte Workload-Intensität eine bessere Leistungen der Benutzer in den zu bearbeitenden Primär- und Sekundär-Aufgaben erzielt werden, und die Ergebnisse der Selbstauskunft der Benutzer zeigen signifikante Vorteile durch die Workload-Adaption. Darüber hinaus konnten wir mit dem *DCR Framework* Verbesserungen der Erkennungsleistung gegenüber unserem bisherigen System zeigen. Mit dem zweiten BCI Paradigma tragen wir zu dem sich aktuell schnell entwickelnden Forschungsbereich der Erkennung von Sprache, anhand invasiv gemessener Hirnaktivität bei. Dabei zeigen wir ein erstes System zur Erkennung von Vokalen während kontinuierlich artikulierter Sprache auf Basis von Elektrokortikographie (ECoG) Signalen. Die gelernten Modelle erlauben neben der Erkennung eine detaillierte Analyse der am Sprachprozess beteiligten Gehirnregionen und ihrer Interaktionen.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides general foundations for this dissertation. It introduces the basic concepts of modern Brain-Computer Interfaces (BCIs), including the relationship between BCI applications, BCI paradigms, and brain activity patterns. The chapter gives an overview on the state-of-the-art and current challenges in BCI research. It ends with a summary of the objectives of this dissertation.

1.1 Brain-Computer Interfacing

The vision of mind-reading machines and the control of machines by pure thought has stimulated the fantasy of both, researchers and science fiction authors for many decades. Brain-Computer Interfaces (BCIs) measure and interpret the users' brain activity with the goal to derive information about his or her intentions and mental states. They provide a communication channel that enables Human-Machine Interaction by only using signals emitted by the users' brain.

The primary goal of this dissertation is to advance pattern recognition of BCIs on the basis of three core objectives that we introduce and discuss in chapter 3 in detail. In this chapter, we lay the basic foundations and

introduce briefly what the current state of BCI research is and which role pattern recognition plays for BCIs.

1.1.1 Brain-Computer Interface Research

Jacques J. Vidal [Vidal, 1973, Wolpaw and Wolpaw, 2011] coined the term "Brain-Computer Interface" in the early 1970s at University of California Los Angeles. Thereafter, few pioneering papers followed in the 1970s and 1980s. BCI research started to gain more and more attention in the late 1990s as recording technology, signal processing and computational power had advanced. Since the millennium, one can see a linear increase in the number of published research articles.

Figure 1.1 shows the number of scientific articles listed by Google Scholar for the search terms[1] "Brain-Computer Interface" and "Brain-Machine Interface" over time since the 1970s.

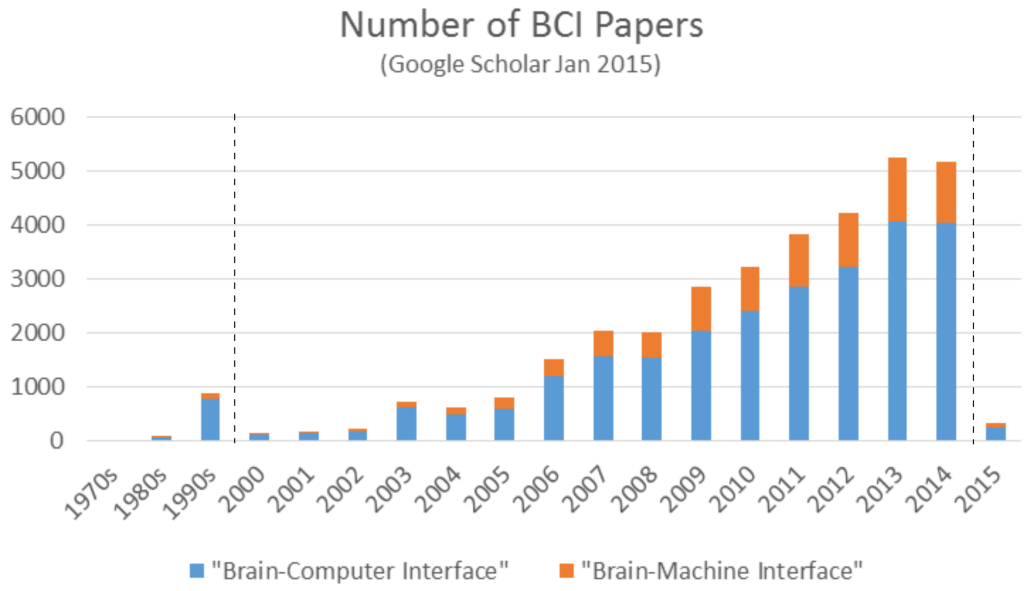


Figure 1.1 – Number of BCI papers since 1970 according to Google Scholar using search terms "Brain-Computer Interface" and "Brain-Machine Interfaces" (as of January 2015).

[1]Both search terms are synonymously used in literature. Throughout this work, we use "Brain-Computer Interface".

Today, BCI research is a highly interdisciplinary field that spans a wide range of different areas of science and engineering, including Neuroscience, Cognitive Psychology, Medicine, Human Factors, Statistics, Signal Processing, and Machine Learning.

While Brain-Computer Interfaces is a young and highly active research discipline that is rapidly evolving and pursued by numerous labs all over the world[2], BCI researchers have not yet agreed on common standards and a common terminology. For example, there are different opinions on what a "Brain-Computer Interface" actually is and what it is not (see next section). Currently, there are efforts towards a roadmap for future BCI research, such as the initiative to form a BCI community that has started at the BCI Meeting 2013 [Huggins et al., 2014] and the Brain/Neural Computer Interaction Horizon 2020 project [Brunner et al., 2015].

1.1.2 Definition of Brain-Computer Interfaces

The BCI research community has not agreed on a single common definition of the term Brain-Computer Interface [Brunner et al., 2015]. Several authors have proposed definitions that all share similar ideas but have a sightly different focus (see e.g. [Graimann et al., 2010] for a collection of definitions). One of the most frequently employed definition is the one by Wolpaw et al. [Wolpaw and Wolpaw, 2011], who define a BCI as

Definition: *BCI [Wolpaw and Wolpaw, 2011]*

> *A system that measures Central Nervous System (CNS) activity and converts it into artificial output that replaces, restores, enhances, supplements, or improves natural CNS output and thereby changes the ongoing interactions between the CNS and its external or internal environment.*

Throughout this thesis, we employ our own, more general BCI definition:

Definition: *BCI*

> *Brain-Computer Interfaces are Human-Machine interaction systems whose operation depends on the single-trial analysis and interpretation of their users' brain activity signals.*

[2]BCI research can still be seen as a comparably small field. For example, Google Scholar lists about 50000 papers per year for the search term "Computer Vision" in the last 5 years.

While the definition by Wolpaw et al. is centered around applications in relationship to CNS activity, our definition decouples the applications of a BCI from those that are typically performed by the CNS[3]. Furthermore, it emphasizes the focus of this dissertation on pattern recognition from brain activity signals of few seconds length (single-trials) as the central component of a BCI.

1.1.3 Components and Structure of a Brain-Computer Interface

The components of a modern BCI form a feedback loop that is also known as the "Brain-Computer Interface Cycle" [van Gerven et al., 2009]. Figure 1.2 illustrates an adapted version of this cycle (based on [Wolpaw et al., 2002, van Gerven et al., 2009], and others). It consists of the four main parts BCI user, brain signal acquisition, pattern recognition, and BCI program:

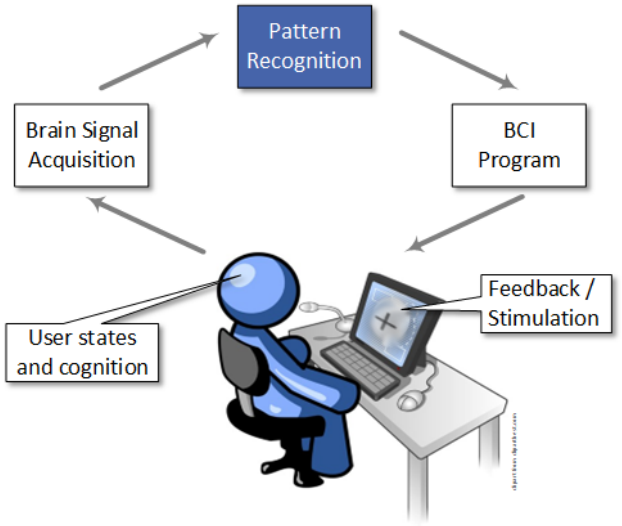


Figure 1.2 – BCI cycle (feedback loop) including the user, brain signal acquisition, pattern recognition and BCI program. Figure inspired by [Wolpaw et al., 2002, van Gerven et al., 2009].

[3]The central nervous system consists of the brain and the spinal cord. All BCIs discussed in this dissertation are brain activity based, therefore we refer to "brain activity" instead of "CNS activity" throughout this thesis.

BCI Users

Signals emitted by the BCI users' neural processes are the fundamental source of information for the BCI. These neural processes are intentionally controlled by the BCI users, for example by performing certain mental tasks, or correspond to their unintentionally occurring brain activity patterns.

In this thesis we target human BCI users, however the concepts and methods discussed in this thesis may, in principle, be applied in non-human BCIs, such as animal studies with primates or rodents.
Clinical patients have traditionally been a target user group of BCI research [Wolpaw et al., 2002]. For example, BCIs have been developed to provide means of communication and control to patients suffering from spinal cord injury and severe motor degenerative diseases. Specifically, BCIs were designed to support amyotrophic lateral sclerosis (ALS) patients, who in the final stages of the disease suffer from a nearly complete loss of muscle control, which is often referred to as *locked-in syndrome*[4].
In the last few years, healthy users have gained increasing attention in BCI research [Allison et al., 2007]. The direct measurement of neural signals that are related to the user's mental processes can be a valuable source of information for a wide range of intelligent systems that adapt their current state to the user [Frey et al., 2013]. This is, in particular relevant, as BCIs can infer information about the user, such as covert user states [Zander and Jatzev, 2009], that cannot easily be derived by audio, video and other biophysiological sensors.

More background on physiological foundations of human brain activity signals and their use in BCIs is discussed in section 2.1.1.

Brain Signal Acquisition

Brain signal acquisition systems for BCIs consist of sensors and amplifiers to measure electrophysiological or metabolic activity emitted by the neural processes of the BCI users. One can distinguish between *invasive* and *non-invasive* brain signal acquisition. In non-invasive measurements, sensors are extracraneal, i.e. placed outside the head, such as electrodes attached to the scalp for Electroencephalography (EEG). Invasive measurements require

[4]The term "locked-in syndrome" is not only used for subject conditions with an absence of all voluntary movements. For example, [Laureys et al., 2005] defined a condition with intact abilities to perform vertical eye movements and blinks as "classical locked-in syndrome".

neurosurgery to place sensors intracraneally, i.e. placed inside the head on top of the cortex (epidural or subdural recordings by Electrocorticography, ECoG), or microarrays placed within the cortical tissue (local field potentials (LFPs) or single cell recordings). Because of the high health risks involved in neurosurgery, most BCI research uses non-invasive measurements, while invasive BCI experiments with human users are only conducted within clinical interventions.

Sensors are typically placed at multiple locations distributed over specific cortical areas or over the whole scalp (e.g. according to the international 10/20 system [Homan et al., 1987]). The brain signal acquisition systems amplify and digitize the signals into a digital multivariate time series for further processing. A selection of different types of sensor modalities is available to acquire brain activity signals for BCIs of which EEG is the most frequently used modality. EEG performs a direct measurement of the electrical potentials that origin in the synaptic activity of the BCI users' neural processes. Figure 1.3 shows a BCI user wearing an EEG-cap (Brain Products actiCap, 16 channels) during a BCI experiment. In addition to direct measurement of signals produced by the neural processes, correlates of neural activity can be used as brain activity signals. For example, functional Near-Infrared Spectroscopy (fNIRS) uses optical measurements to acquire relative changes of cerebral blood oxygenation that is generally associated with brain activity (Blood-Oxygen Level Dependent, BOLD effect [Ogawa et al., 1990]). Magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) are less frequently used for BCIs, as they require a high technical effort, have high operation cost, require special facility, and strongly restrict users in their movements.

More background on brain signal acquisition and, in particular, on Electroencephalography (EEG), functional Near-Infrared Spectroscopy (fNIRS), and Electrocorticography (ECoG), can be found in section 2.1.2.

Pattern Recognition

The pattern recognition component is the central component of a BCI. In this dissertation, we refer to pattern recognition as the term for all signal processing and machine learning operations involved to transform digital brain activity signals into a recognition output. This output corresponds to a machine interpretable estimate of the class or intensity of a user state or mental task condition represented as a discrete or continuous variable.

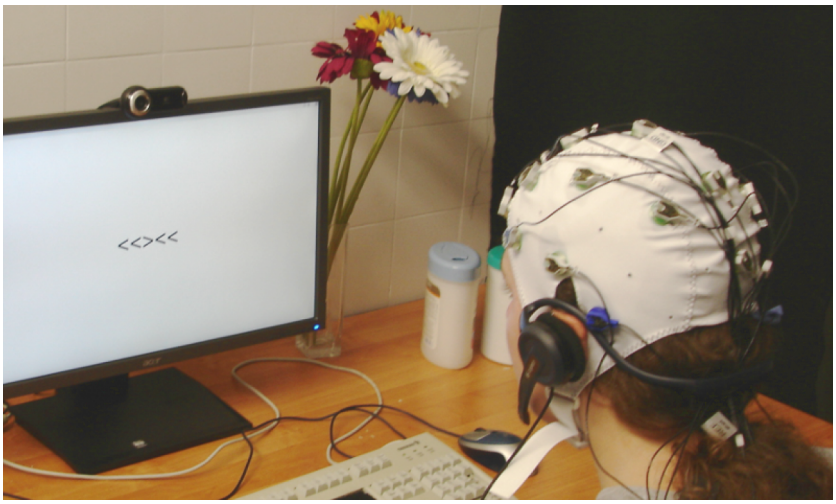


Figure 1.3 – A BCI user wearing an EEG-cap (Brain Products actiCap, 16 channels) during a BCI experiment, in which visual feedback is provided on the screen by the BCI program according to the Eriksen flanker task [Eriksen, 1995].

Pattern recognition in early BCIs was rather straight-forward and relied strongly on the fact that people can learn to modulate their brain activity patterns, in particular, as the feedback in the BCI cycle supports learning (i.e. operant conditioning [Rockstroh et al., 1984, Birbaumer, 2006, Skinner, 1938]). After intensive user training, the recorded brain signals can operate a BCI Program by using a simple static rule-based translation of the measured signals as control commands. Advanced signal processing and machine learning for BCIs substantially ease the effort of novice BCI users to operate a BCI and shift learning effort towards automatic pattern recognition methods [Müller et al., 2004, Vidaurre et al., 2011, Kindermans et al., 2014a]. Modern BCIs usually have a calibration phase that takes less than an hour instead of up to several months of user training. Moreover, pattern recognition methods enable to analyze the naturally occurring brain activity patterns that are not specifically intended to operate the BCI.

Common pattern recognition approaches for BCIs are described in section 2.2. In chapter 3, we introduce and discuss three core objectives to advance pattern recognition of BCIs. These objectives lead to a new pattern recognition framework for BCIs (*DCR Framework*) that is described in chapter 4 in detail and evaluated for multiple BCI problems in chapters 5 and 6.

BCI Program

The BCI program is the component of a BCI that processes the information about the recognized brain activity patterns to affect the operation of a computer application or an embedded device. For example, a BCI program can transform the pattern recognition output into a control command for an artificial device, such as a neuroprosthesis, or it can move a cursor on the computer screen in a certain direction according to the recognized class.

In addition to that, the BCI program often shows or triggers a feedback response to the user, such as a visual output on the screen or the movement of a mechanical device. The perception and associated cognition processes to such a feedback event influence the user's brain activity patterns and, in turn, provide the basis for the measurements of the brain signal acquisition system, which closes the BCI cycle.
During the calibration phase of a BCI, often predefined stimuli are presented to the user, such as instructions to perform a specific mental task. Additionally, a certain class of BCIs, called *dependent BCIs*, relies on the users' overt attention towards sensory stimulation by external sources that are controlled by the BCI program (see e.g. Oddball paradigm in the next section).

Systems that measure and derive information about the users' brain activity but do not close the loop by providing feedback to the user are sometimes called *cognitive monitoring* systems. Furthermore, interfaces with the primary purpose to present feedback on measured brain activity to the user are sometimes called *neurofeedback* systems. The techniques and methods used in cognitive monitoring systems, neurofeedback systems and BCIs do, in general, not differ. Consequently, we use only the term BCI for all such systems throughout this dissertation, which is in agreement with our definition in section 1.1.2.

The following section 1.1.4 gives more examples and details on current BCI applications and corresponding BCI paradigms.

1.1.4 BCI Applications, BCI Paradigms and Brain Activity Patterns

BCI is an umbrella term that encompasses a variety of different *applications*. Each of these BCI applications employs certain BCI *paradigms* that are known to modulate characteristic *brain activity patterns* (figure 1.4). The terms "BCI application", "BCI paradigm", and "brain activity pattern" are

not precisely defined in BCI literature [Brunner et al., 2015], therefore the following sections provide definitions for the usage of these terms within this dissertation and list multiple examples.

Figure 1.4 – Summary of the structural relationship between BCI applications, BCI paradigms and brain activity patterns.

BCI Applications

Researchers have proposed and implemented a wide range of different BCI applications. Traditionally, research effort has been put into communication systems for locked-in patients [Farwell and Donchin, 1988], control of prostheses [Müller-Putz et al., 2005, Hochberg et al., 2012], wheelchairs [Galán et al., 2008], or computer applications [Bensch et al., 2007]. More recently, stroke rehabilitation [Daly and Wolpaw, 2008, Ang et al., 2010] gained a lot of attention in the BCI research community. Additional applications that may become increasingly relevant in the near future are games and virtual reality applications [Nijholt et al., 2009, Bos et al., 2010], ergonomics or usability testing [Hirshfield et al., 2009, Frey et al., 2013], user state monitoring of drivers and pilots [Lin et al., 2005], education and tutoring systems, rehabilitation for attention deficit hyperactivity disorder patients [Lim et al., 2010], functional electrical stimulation for bowel and bladder control [Peckham and Knutson, 2005], approaches to make Human-Machine Interaction more natural and empathic [Heger et al., 2011a], rapid media tagging [Parra et al., 2008, Wang et al., 2009], security and surveillance applications [Müller et al., 2008, Hild et al., 2014].
Few BCI applications, like the P300 speller [Farwell and Donchin, 1988], have even been translated into clinical practice [Wolpaw and Wolpaw, 2011]. Furthermore, BCIs have reached the consumer market with brain games that are commercially available [Zhang et al., 2010].
In sum, there is a large and growing body of very different BCI applications. A primary limiting factor for the development of new BCI applications is the reliable recognition of the BCI paradigms they employ.

BCI Paradigms and Corresponding Brain Activity Patterns

The term "BCI paradigm" is regularly used in BCI literature (e.g. [Obermaier et al., 2003, Schalk et al., 2004, Neuper et al., 2003, Babiloni et al., 2007, Guger et al., 2009, Fazli et al., 2012]) but is not clearly defined. We refer to "BCI paradigms" throughout this thesis as

Definition: *BCI paradigm*

> *The experimental task or mental state of the user that modulates certain brain activity patterns to operate the BCI.*

In this definition, BCI paradigms are closely related to "brain activity patterns" to which we refer to as

Definition: *Brain activity patterns*

> *The characteristic properties of the neural processes that are modulated by the BCI paradigm and elicit measurable changes in parts of the brain activity signals.*

Only a few BCI paradigms are regularly employed in BCI research. The most frequently employed categories of BCI paradigms are:

- Cognitive tasks: Cognitive tasks are mental tasks which users perform to generate certain brain activity patterns. *Motor Imagery* [Pfurtscheller and Neuper, 1997], i.e. the imagined movements of body parts (hands, feet, etc.), is the most commonly used cognitive task in BCI research. Other tasks that have repeatedly been employed for BCIs are mental arithmetics [Anderson et al., 1995], mental rotation, or working memory tasks, such as the n-back task [Kirchner, 1958].
- User states: In comparison to cognitive tasks, cognitive or affective user states occur passively as part of the current mental condition of the user, i.e. they are often unrelated to a specific event and are not intentionally performed by the user. A typical example for user states are affective states that are typically induced in lab experiments using positive and negative emotional stimuli (see [Mühl et al., 2014a] for review). Other user states that have been used for BCIs are workload, vigilance and fatigue, selective attention to auditive and visual stimuli, and the perception of erroneous events, such as an unexpected feedback because of recognition errors (see [Frey et al., 2013] for review).
- Oddball paradigm: In the oddball paradigm, a sequence of repetitive (usually auditive or visual) stimuli is presented to the user, which

is interrupted by different rare stimuli, so-called oddballs (e.g. 20% of stimuli). The perception of an oddball stimulus generates a P300 event-related potential (ERP) that can be measured, e.g. by EEG and MEG. A BCI user can perform binary selections for communication and control by shifting his or her attention to or away from different oddball stimuli. For example, this paradigm is used in the matrix speller [Donchin and Coles, 1988], in which the user can shift his or her attention to different letters that are aligned in a grid and light up repeatedly.

- Perception of fast repetitive stimuli: The perception of fast periodic (usually visual) stimuli triggers a corresponding response in associated cortical areas. Typical stimuli are flashing light sources (e.g. LEDs or phase-reversing checkerboxes [Allison et al., 2008]). The user can select different options of the BCI by switching his or her focus of attention to one of multiple stimuli sources that have distinct stimulation sequences or stimulation frequencies. A particular stimuli source and, therefore a particular option for the BCI, can be detected from the brain response sequence that corresponds to the stimulation sequence.

Details about the brain activity patterns and the signal properties that are modulated by these BCI paradigms are described in section 2.1.3.

Figure 1.5 illustrates examples for the relation of applications, paradigms and patterns. It shows BCIs for communication and control with different brain activity patterns employing motor imagery, P300 speller or fast repetitive stimuli. Furthermore the figure illustrates adaptive human-machine interaction using workload recognition that can be assessed using the BOLD effect, brain rhythms, or event-related potentials.

1.1.5 State-of-the-Art and Challenges

In the previous section we have outlined diverse BCI applications for clinical and non-clinical use. However, brain-based user interfaces are still not widely used outside research labs. Currently, only less then 10 locked-in patients use BCIs regularly [Wolpaw and Wolpaw, 2011]. For healthy users, there are only a few commercially available BCIs on the consumer market, mainly for entertainment purpose [Zhang et al., 2010].

The following limitations of state-of-the-art BCIs are among the most important obstacles for a wider applicability and user acceptance of BCIs: (1) current BCIs require inconvenient setups, (2) their information throughput is

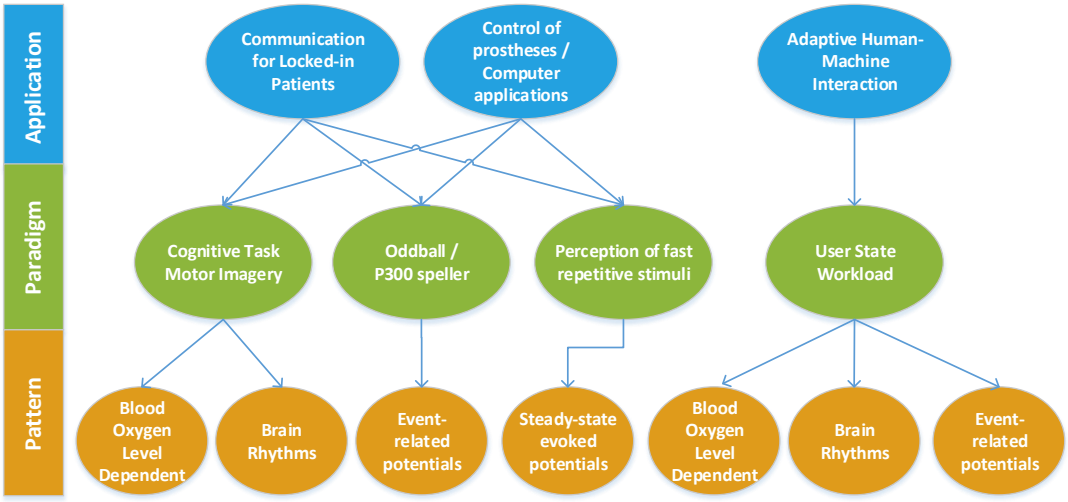


Figure 1.5 – Examples for the relation of BCI applications, paradigms and brain activity patterns. Lines connect BCI applications, paradigms and patterns to illustrate their relation for communication and control with different brain activity patterns employing motor imagery, P300 speller or fast repetitive stimuli. Furthermore the figure illustrates adaptive human-machine interaction using workload recognition that can be assessed using the BOLD effect, brain rhythms, or event-related potentials.

low in comparison to conventional user interfaces, (3) the recognition is unreliable, especially under not completely controlled conditions, and (4) many established BCI paradigms are not natural and not intuitive for BCI users.

Inconvenient Setups

Currently, BCIs are strongly restricted in their practical applicability. BCIs usually cannot be used right away, but require *calibration* that can take up to an hour. A calibration data collection has to be performed prior to each BCI session and the data cannot be reused for other sessions or other users, because of strong inter-session and inter-person variabilities of brain activity signals (section 2.1.3).
Traditional brain signal acquisition devices, such as EEG caps, are obtrusive as they require conductive gel that the users have to wash out of their hair after each BCI session. Furthermore, the conductive gel cannot easily be applied by the user himself or herself. Head-sets with dry electrodes and fNIRS optodes can be uncomfortable to wear, since they often impose a high pressure to achieve a good contact between sensor and scalp.

Additionally, many BCIs force the users to limit their natural movements to maintain a high signal quality. For example many BCIs do not operate properly when the user moves his or her eyes, when muscles are activated, or while he or she is speaking (see section 2.1.3).

Low Information Throughput

One of the most limiting factors of BCIs is their low information throughput. BCIs usually discriminate a small number of discrete classes of brain states and require signal segments (trials) of several seconds length to derive a recognition estimate. Therefore, the fastest non-invasive BCIs have information throughput usually below 2 bits/s [Spüler et al., 2012]. Such a low throughput prevents real-time control of complex systems and reduces the speed of communication devices to a level which is outperformed by nearly all conventional input modalities, by far.

Unreliable Recognition

In addition to low recognition accuracies, BCIs often cannot maintain persistent recognition rates over time. Brain activity signals are inherently characterized by artifacts and non-stationarities that can lead to a strong decline of recognition performance. Typically, only a few minutes of calibration data are available to train the BCI, which makes it challenging to learn robust pattern recognition models. Additional data from other recording sessions or different users cannot be used, in general, because of the strong inter-person and inter-session variabilities (section 2.1.3). The unreliable recognition becomes even more evident when BCIs are supposed to be used outside of controlled laboratory conditions.

Unnatural and Non-Intuitive BCI Paradigms

Most BCIs employ one of few well-studied BCI paradigms that have been outlined in section 1.1.4. These BCI paradigms are based on neuroscientific effects known to generate certain brain activity patterns that can be measured reliably in most users with a fair signal-to-noise ratio. However, the interaction protocols that the corresponding BCI applications impose are unnatural and non-intuitive for many users. Especially, BCI paradigms that require significant user learning or depend on external stimulation (dependent BCIs), such as blinking or flickering lights, appear inconvenient in

comparison to alternative modalities for Human-Machine Interaction, such as mouse, keyboard, speech, or gestures.

Approaching the Challenges by Pattern Recognition

The pattern recognition component, which plays a central role in a BCI, can significantly contribute to counteract the limitations discussed above. An important branch of research on pattern recognition methods for BCIs is concerned with methods that reduce the calibration time of BCIs. For example, transfer learning techniques have been proposed that enable to use calibration data from previously recorded sessions of the user or different users (see e.g. [Krauledat et al., 2008, Heger et al., 2013, Kindermans et al., 2014b] and section 4.2). Since the 1990s, numerous incremental advances in pattern recognition methods have contributed to the continuous increase in recognition rates for BCIs. Furthermore, to increase information throughput, pattern recognition methods enabled the single-trial recognition of brain activity signal segments of few seconds length and increased the number of different classes that can be discriminated [Wolpaw and Wolpaw, 2011]. Techniques that reduce the impact of artifacts (e.g. [Romero et al., 2008]) and more recently non-stationarities (e.g. [Samek et al., 2012]), such as linear subspace transformations and adaptive classification, are currently developed to increase the reliability of BCI pattern recognition. Developments in data analysis methods enabled to investigate brain activity patterns and have discovered novel BCI paradigms, such as monitoring of certain covert task-specific user states (e.g. [Reissland and Zander, 2009]). These efforts have significantly contributed to the field of BCI research, however they usually focus on isolated aspects of the four challenges described in the previous paragraphs. In this dissertation we approach each of the four challenges by systematically advancing generic methods for BCI pattern recognition as outlined in the next section.

1.2 Thesis Objectives

The goal of this thesis is to systematically advance pattern recognition for BCIs in order to yield BCIs with shorter setup times, higher information throughput, more reliable recognition and enable more natural and intuitive

BCI paradigms. To achieve this goal we pursue the following objectives in this dissertation:

1.2.1 Identifying Core Objectives of Pattern Recognition for BCIs

A major limiting factor for pattern recognition of BCIs is that there is no systematic theory on how the pattern recognition component of BCIs should be realized. Therefore, the first objective of this thesis is to identify core aspects that are necessary conditions for the pattern recognition component of a BCI. Thus, we formulate the following hypothesis for this dissertation:

For pattern recognition in BCIs it is necessary to implement and balance three core objectives:

- DISCRIMINATIVE*: Identification of brain activity patterns, that enable to discriminate between different classes or intensities of a BCI paradigm*
- COMPACT*: Compact modeling of relevant aspects of brain activity patterns in generalizing structures*
- ROBUST*: Robustness against signal variabilities in brain activity signals that are not modulated by the BCI paradigm*

To provide evidence for this hypothesis, we highlight the relevance of each of these aspects for BCIs and relate each of them to objectives of methods that are already used in the pattern recognition component of BCIs, furthermore, we show that the three objectives are indeed necessity conditions by relating each of them to essential principles of pattern recognition.

1.2.2 Generic Framework for Single-Trial Recognition based on Joint Optimization

The second objective of this thesis is to create a *BCI recognition framework that jointly optimizes the three identified pattern recognition objectives* DISCRIMINATIVE, COMPACT, and ROBUST in a principled and generic way, which we call the *DCR Framework*. For the first time, this challenge has been approached by formulating the three components as a convex optimization problem, which we solve using a new algorithm based on the Alternating Direction Method of Multipliers [Boyd et al., 2011].

We evaluate our framework using different brain activity signals (EEG, fNIRS, and ECoG). We show the great flexibility and state-of-the art performance for different established and novel BCI paradigms with various signal patterns with different characteristics, such as the recognition of oscillatory signals, event-related potentials, and hemodynamic activity. In addition to that, the successful evaluations using the principled approach of the *DCR Framework* show empirically that the three objectives Discriminative, Compact, and Robust are a sufficient set of conditions for a variety of tasks in BCI pattern recognition.

1.2.3 Novel BCI Paradigms

Improvements in the three pattern recognition objectives are especially relevant for recognizing *new BCI paradigms* that do not follow well-known neurophysiological effects but modulate brain activity patterns in a complex way. Therefore, the third objective of this thesis is to introduce two innovative systems to recognize BCI paradigms that have not been proposed in this form before. In these systems, the naturally occurring brain activity patterns are analyzed automatically and, particularly, no learning by the user is required prior to operate the BCI. For each of the BCIs, we highlight the particular aspects of how the above mentioned objectives Discriminative, Compact, and Robust are implemented.

Specifically, we discuss a novel self-paced BCI for *EEG-based workload adaptive human-machine interaction*, in which we could show significant benefits for the users by adapting an information presentation system to the recognized workload. Furthermore, we analyze ECoG-based *Brain-to-Text* vowel classification with non-stationarity reduction, which is a significant improvement for our recently introduced system to decode continuously spoken speech from ECoG signals.

1.3 Structure of this Thesis

Figure 1.6 summarizes the structure of this thesis. The first two chapters introduce the central aspects of pattern recognition based BCIs and provide the necessary foundations on brain activity signals. In chapter 3, we introduce and discuss the three core objectives of pattern recognition for BCIs and provide evidence for each of them by relating them to BCIs and to principles of pattern recognition (thesis objective 1.2.1). Furthermore, we discuss

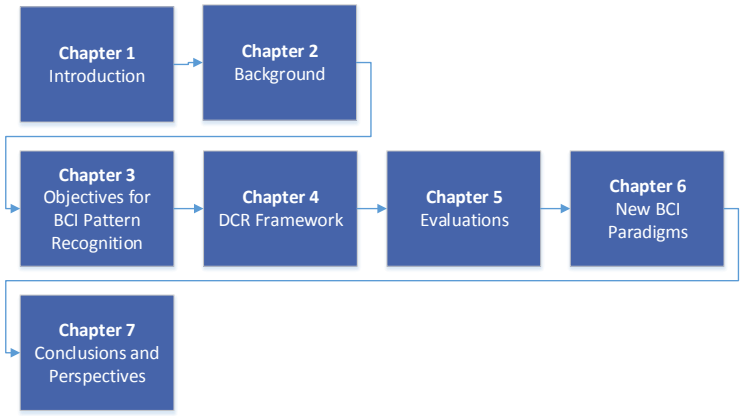


Figure 1.6 – Structure of this thesis

their interdependence. Chapter 4 describes in depth the *DCR Framework*, a general framework for single-trial recognition based on the joint optimization of the three objectives of pattern recognition for BCIs (thesis objective 1.2.2). Chapter 5 provides detailed evaluations of the *DCR Framework*, using multiple different BCI problems and synthetic data. Two studies on novel BCI paradigms are introduced in chapter 6 (thesis objective 1.2.3). Chapter 7 summarizes the main results and contributions of this dissertation and proposes directions for future research.

Parts of this dissertation have been published in international journals and conference proceedings. A list of own publications can be found in Appendix **??**.

Chapter 2

Background

The first part of this chapter gives an introduction into the neurophysiological background of brain activity signals and different techniques for brain signal acquisition. The second part covers general aspects of single-trial recognition of brain activity patterns in a modern BCI system.

2.1 Brain Activity Signals

The human brain is a highly complex information processing system that drives most functions of the human body and higher cognitive abilities, including thinking and acting. Although the functionality of specific areas and the principles on how information is encoded, processed and distributed are still not completely understood, it is generally accepted that brain function is based on electrical and chemical processes, which operate as an interconnected information processing system based on nervous cells (neurons). The neural processes and correlates thereof can (partly) be measured and exploited as brain signals for brain-computer interfacing.

2.1.1 Neural Information Transfer

A comprehensive introduction into the neurophysiological foundations of brain activity are beyond the scope of this dissertation. Here, we only give a short summary of the most relevant aspects. The interested reader may consult [Kandel et al., 2000, Zschocke and Hansen, 2011, Wolpaw and Wolpaw, 2011].

Neurons in the brain are comparably simple but highly interconnected units of information processing. The information transport in the neuronal network is based on neurons firing in response to electrical or chemical excitation. If the excitations of a neuron exceed a certain threshold, an action potential (spike) in form of a brief local current is released. It propagates along the axon to the synapse where neurotransmitters are released. The resulting movement of positive and negative ions (Na+, K+, Cl-) cause changes in the membrane conductance that lead to inhibitory or excitatory effects in the postsynaptic neuron, called excitatory and inhibitory postsynaptic potentials (EPSP and IPSP). These potentials can contribute to generate a succeeding action potential in the postsynaptic target cell, i.e. connected neuron or muscle cell.

The neural information transfer, described above, is accompanied by subsequent changes in the cerebral blood flow, which is commonly called *neurovascular coupling*. The neural processes require energy in the form of adenosine triphosphate, which is synthesized primarily from glucose and oxygen. The cerebral blood flow supplies the neural cells with both of these substrates. Consequently, neural activity increases cerebral blood flow and causes local changes in the level of blood-oxygenation. The brain signal acquisition modalities fNIRS and fMRI rely on these neurovascular effects (see next section).

2.1.2 Brain Signal Acquisition

In the following we briefly describe the background and characteristics of the brain acquisition modalities that have been used in the experiments of this thesis (chapters 5 and 6), namely Electroencephalography (EEG), functional Near-Infrared Spectroscopy (fNIRS) and Electrocorticography (ECoG).

Electroencephalography (EEG)

The summed activity of postsynaptic potentials (EPSP and IPSP, see section 2.1.1) of large populations of neurons[1] generate local (extracellular) field potentials (LFPs). Field potentials of large populations of neurons that are time-synchronously active and spatially aligned in parallel within a cortical area can generate an open dipole field that is strong enough to be measured non-invasively at the scalp. *Electroencephalography* (EEG) is the direct measurement of the variations of these summed cortical field potentials using electrodes attached to the scalp.

EEG is the most frequently used modality for brain signal acquisition for BCIs because it is comparably affordable, mobile, and delivers signals with a high temporal resolution. Traditionally, EEG is recorded using electrode caps with 16-256 electrodes attached according to standardized locations of the extended 10-20 system [Oostenveld and Praamstra, 2001]. In BCI research, EEG is usually measured using difference amplification with respect to a reference position[2]. Usually, electrode gel (conductive gel) is used to increase the conductivity between the electrodes and the scalp. In the last few years, dry electrode EEG systems are becoming increasingly popular, but EEG recordings with conductive gel are still standard as they provide the highest signal quality. Non-pathological EEG potentials vary between $\pm 100\mu V$ on the scalp relative to the reference voltage. Since this activity is rather weak the EEG is sensitive to artifacts (see section 2.1.3).

Figure 2.1 shows a 5 second segment of a 16 channel EEG of a healthy and awake person in time domain.

Functional Near-Infrared Spectroscopy (fNIRS)

Functional Near-Infrared Spectroscopy (fNIRS) measures changes of oxygenation in regional cerebral blood flow. According to the blood oxygen level dependent (BOLD) effect, oxygenated (HbO_2) and deoxygenated (HbR) hemoglobin are functional indicators for brain activity. An increase of neural activity in a cortical area is accompanied with an increase in the consumption of oxygen (cf. section 2.1.1) which causes a rising ratio of HbO_2 and HbR in

[1] In [Makeig et al., 2012] the number of cortical neurons is estimated to be twenty billion, of which the activity sources underlying the EEG can span several square centimeters of the cortical surface.

[2] Signals recorded using such a reference derivation can be rereferenced to bipolar signal derivation or common average reference in the signal processing stage of the BCI.

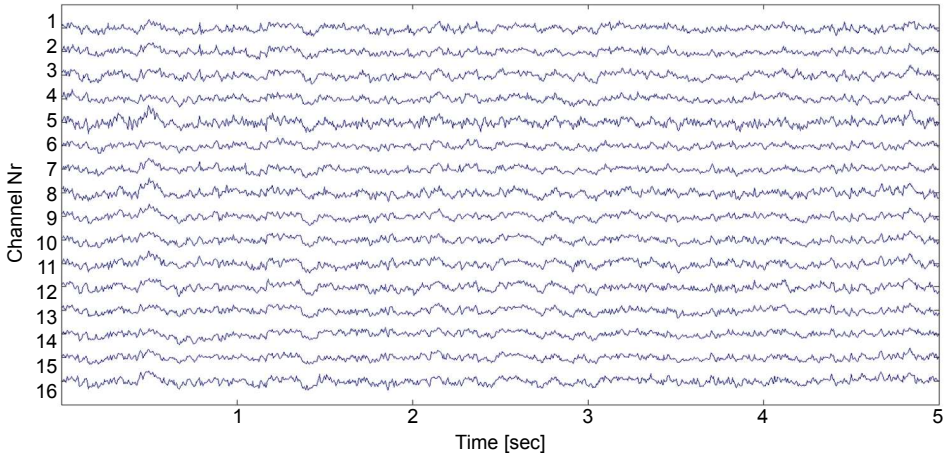


Figure 2.1 – Example of EEG signals (5 seconds, 16 channels)

that specific area (neurovascular coupling). Continuous wave fNIRS exploits the fact that HbO_2 and HbR have different absorption characteristics for light in the near-infrared spectrum. While near-infrared light (700-900 nm) penetrates easily through biological tissue, it is absorbed by hemoglobin in the cortex, light sources can send near-infrared light through the scull and light detector optodes measure the intensity of the scattered light at nearby locations. Using the modified Beer-Lambert Law [Sassaroli and Fantini, 2004] changes in the cerebral blood oxygenation, and thereby brain activity, can be estimated from the changes in light intensities.

fNIRS is an emerging optical brain imaging modality gaining rising attention in the BCI community. In contrast to functional magnet resonance imaging (fMRI), which also measures BOLD responses, fNIRS is non-invasive, is comparably cheap, portable and does not confine the subjects. In contrast to EEG, fNIRS is not susceptible to electrical artifacts from environmental and physiological sources. Furthermore, no conductive gel needs to be used. Also, frontal fNIRS recordings, where measurements are not obstructed by hair, have very short setup times (about one minute). The major disadvantage of fNIRS is its low temporal resolution (section 2.1.2).
fNIRS is usually assessed by multiple sensors that are located at position defined by the BCI researcher depending on the BCI paradigm. Light emitting optodes and sensors optodes are attached to the scalp in usually 2-4 cm interoptode distance. Measurement positions are located roughly in the middle between emitter and sensor in a depth of half the interoptode distance.

Figure 2.2 shows an example of a 120 seconds segment of a oxygenated (blue) and deoxygenated (green) hemoglobin changes in time-domain measured by 8 channels of fNIRS. The data have been measured from a healthy person during resting and the performance of a working memory task. Oscillations in measured HbO signals that correspond to cardiac activity (frequency around 1 Hz). Slow increases of HbO and decreases of HbR activity can be associated with hemodynamic responses (e.g. starting at second 75 in figure 2.2). More details on fNIRS signal characteristics can be found in section 2.1.3.

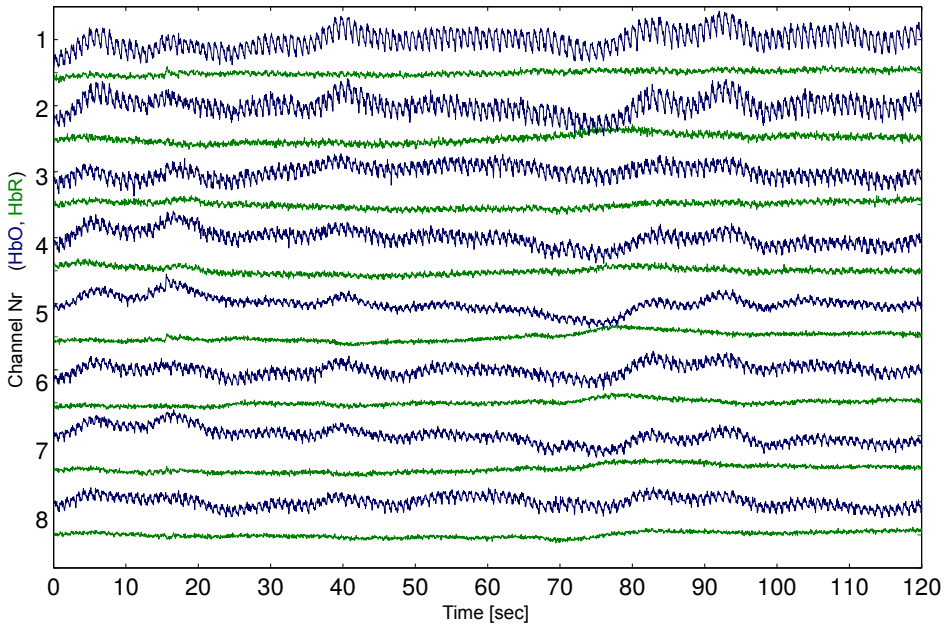


Figure 2.2 – Example of an fNIRS signal segment (120 seconds). The figure shows unfiltered signals in time-domain of 8 channels. For each channel oxygenated (blue) and deoxygenated (green) hemoglobin changes are shown.

Electrocorticography (ECoG)

Electrocorticography (ECoG) measures electrical brain activity similar to EEG but uses electrode grids that are placed directly on the surface of the cortex. Therefore, the characteristics of ECoG signals are similar to those of EEG. However, non-invasive signals usually contain strong artifacts and are filtered by brain tissues, skull and scalp, which leads to volume conduction and low signal-to-noise ratio (see section 2.1.2). ECoG signals have less artifacts, a higher spatial resolution and high frequency activity (e.g. high gamma broadband 70-170 Hz) can be measured that is associated with local-

ized activity of functional processing (cf. section 2.1.3). ECoG grids usually consist of 8×1 or 8×8 electrodes, more recently, high-density grids with more than 128 electrodes, wireless transmitters and non-contact power supply have been developed. The location of the grids is generally determined by the clinical needs of the patients. ECoG grids are usually implanted only for few days to 1-2 weeks [Wolpaw and Wolpaw, 2011, chapter 15] and are removed after the clinical intervention without damaging cortical structures.
Figure 2.3 shows an example of an implanted subdural ECoG grid placed over the left fronto-parietal and temporal lobes.

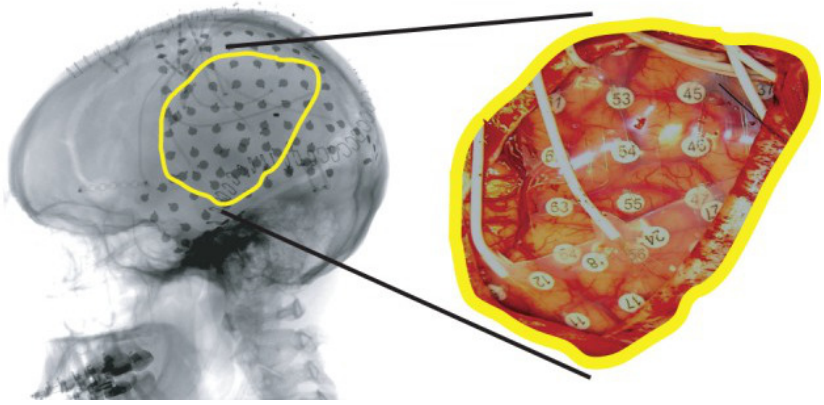


Figure 2.3 – Example of an implanted subdural ECoG grid placed over the left fronto-parietal and temporal lobes. [Wang et al., 2012]

BCI research based on *invasive* recordings, such as ECoG, is becoming increasingly popular as impressive applications have been demonstrated (e.g. [Hochberg et al., 2012, Herff et al., 2015]). Because of the high risks involved in neurosurgery, invasive BCI experiments with humans are not conducted with healthy users. ECoG recordings for BCI research most frequently come from epilepsy patients who had to undergo neurosurgery because of their disease and have agreed to participate in research studies within the time of their clinical interventions.

Microelectrode Arrays

Microarrays are another invasive brain signal acquisition modality. They are most widely used in animal research and rarely used for BCIs with human users, because of the high health risks involved in implanting wires that penetrate the cortical tissue. Microarrays can record the electrical activity of single neurons (spikes) or small groups of neurons and therefore allow for

a very high resolutions in time and space. They can usually only record activity of a small part of the brain where the array is located.

Temporal and Spatial Resolution of Brain Signal Acquisition Modalities

The temporal and spatial resolution that can be reached by a certain brain signal acquisition modality is an important design factor in the development of a BCI. In general, electrical brain activity measurements (EEG, ECoG, MEG, Microarrays) have a high temporal resolution, whereas measurements based on cerebral blood oxygenation (fNIRS, fMRI) have a low temporal resolution. The spatial resolution of non-invasive electrical measurements (EEG, MEG) is usually low because of volume conduction effects, i.e. due to conductivity, each sensor measures the superposition of the activity of multiple neural sources. Therefore, signals measured at different locations can be highly correlated and represent a mixture of different activity sources that may only be partly located directly below the sensors. It is important to keep in mind that the brain activity signal usually used in BCI research can only reflect the cortical activity of large neural populations. The activity in most parts of the brain, in particular subcortical regions, can not be measured directly by EEG, ECoG, and fNIRS.

In sum, the brain signal acquisition modalities discussed in the previous section have in common that their temporal and spatial resolution can be magnitudes larger than the actual neural processes in the brain. Therefore, their signals can only roughly capture and represent the information about the processes that occur in the brain in response to BCI paradigms.

Figure 2.4 summarizes temporal and spatial resolution of different brain acquisition modalities.

Modern Brain Signal Acquisition Devices

In the last few years, miniaturized and mobile EEG and fNIRS sensor devices have become available (e.g. [Filipe et al., 2011, Debener et al., 2012]). Additionally, there are first devices designed for the consumer market (e.g. [Badcock et al., 2013]). In contrast to traditional hardware they have shorter setup times, can easily be self-attached and are designed to be unobtrusive and comfortable to wear. However, the signal quality is often not as good as the one of clinical or research hardware. Besides professional and

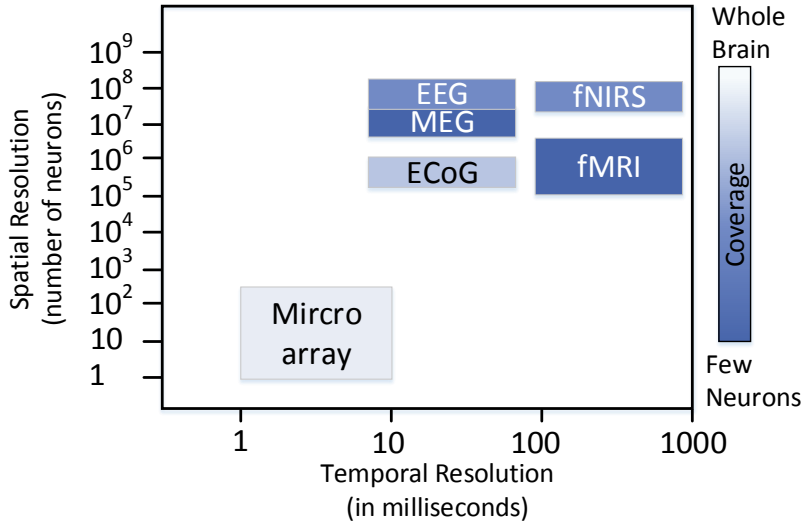


Figure 2.4 – Temporal and spatial resolution of different brain acquisition modalities. Smaller values correspond to higher resolutions (figure based on [Wolpaw and Wolpaw, 2011]).

consumer brain acquisition devises, open designs have been published that come with the complete instructions on how to build a brain acquisition device on your own, including circuit designs, and part lists. For example, the openEEG project [Griffiths et al., 2003] has developed several open hardware EEG devices. Recently, we developed a modular open fNIRS design at the Cognitive Systems Lab that is available by the openNIRS project[3] [von Lühmann, 2014]. Figure 2.5 shows the design of our openNIRS prototype.

2.1.3 Characteristics of Brain Activity Patterns

Information in Brain Activity Patterns

As defined in section 1.1.4, brain activity patterns are characteristic properties of the neural processes that are modulated by a BCI paradigm. Information contained in such brain activity patterns are primarily encoded in the temporal, frequency, and spatial characteristics of the multivariate brain activity signals.

[3] www.opennirs.org

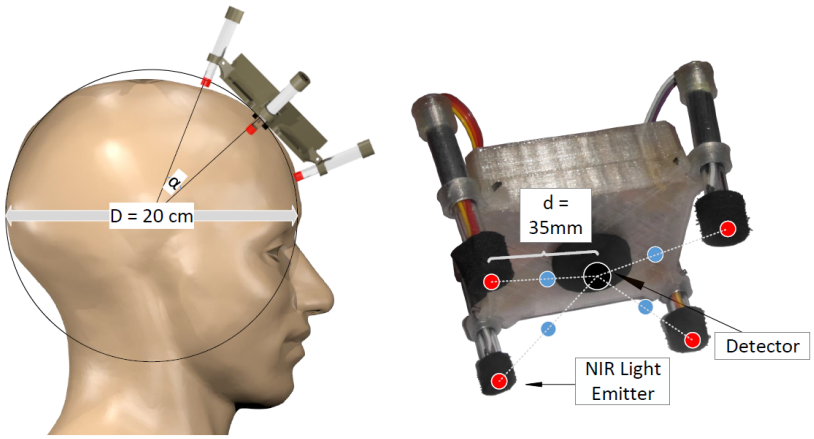


Figure 2.5 – The openNIRS prototype developed at the Cognitive Systems Lab [von Lühmann, 2014].

- Time: Information encoded in the amplitudes of the time series represent activity of specific neural populations over time. Timings of the brain activity signals are, in particular, relevant if they occur in response to an event (event-related), for example in stimulus dependent BCIs.
- Frequency: Bursts of the activity of neural populations and the repetitive interaction between neural populations in different brain areas can generate oscillatory signal parts and brain rhythms. Changes in the frequency characteristics of brain activity signals primarily occur if the activations of large neural populations synchronize or desynchronize (called event-related (de)synchronization ERD/ERS, whereby the event can e.g. be the execution of a cognitive task).
- Space: A general neuroscientific principle is that the brain is organized in functional regions. Therefore, the locations of brain activity sources and the pathways between specific brain areas, i.e. their connectome, are a major source of information on brain activity patterns.

Brain Activity Patterns used in BCIs

The following list summarizes the brain activity patterns that are modulated by different BCI paradigms, including those introduced in section 1.1.4. Event-related potentials (in particular P300 potentials) and neural

oscillations (in particular sensorimotor rhythms) are traditionally the most frequently used brain activity patterns in BCI research.

Event-related potentials (ERPs): ERPs correspond to the brain's electrical response shortly after an event. Events are typically external auditory, visual, or tactile sensory stimuli, but can also be internal stimuli (e.g. associated with the execution of a motor action, cognitive activity, or psychophysiological events). ERP correspond to the activity of large populations of neurons that are involved in the brain's perceptive and cognitive processes.
ERPs are commonly acquired by electrical brain activity measurements, such as EEG and MEG. The Oddball paradigm is usually applied to elicit P300 ERPs. Cognitive tasks and user states are known to modulate components of ERPs that may be used for BCIs.
Signal Properties: ERPs are usually analyzed in time-domain in a signal interval shortly before or after the corresponding event. The ERP complex that can be observed in response to such events consists of multiple positive and negative signal variations (waves), called ERP components. Figure 2.6 shows a prototypical ERP complex. In general, this waveform can not easily be observed in single-trial signals by visual inspection because of the noisiness of the signals. The most important parameters of ERP components are their peak amplitude (in μV, often normalized to the amplitude of the signal in an interval shortly before the event), their peak latency (in ms), and the location of strongest activity (spatial focus).
ERPs are modulated by different BCI paradigms. The most frequently ones used in EEG-based BCIs are:
P300 (or P3): A positive deflection approximately 300ms after oddball stimuli (see 1.1.4). According to the Context-Updating Model hypothesis [Donchin and Coles, 1988, Debener et al., 2005, Polich, 2007], a mental update causes the P300 component and its amplitude reflects the degree to which the event was consistent with a current mental model of the context.
Error potentials: Error potentials consist of a frontal negative component (Ne/ERN, error-related negativity) and a later centro-parietal positive component (Pe, error-related positivity). They are caused by evaluation of an error event, i.e. when the user detects that an outcome is different from what is expected. Error events can, for example, be mistakes of the user or the machine. Error potentials can occur in response to the observation of an error, in response to errors in an interaction, or in response to a feedback event.

Neural Oscillations: Brain rhythms and oscillatory activity can be observed on different levels of neural processing, for example in large neural populations (e.g. measurable by EEG) or in single cell activity. Multiple

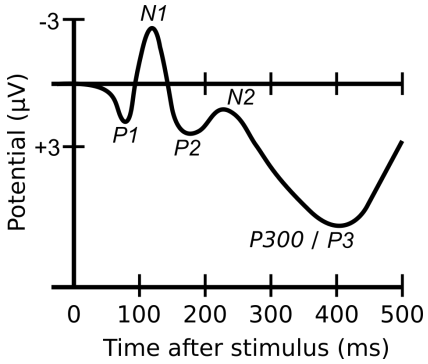


Figure 2.6 – Prototypical ERP complex (waveform of ERP components). Positive (P) and negative (N) components are numbered consecutively. The third positive component P3 is the P300 component. Note that negative voltages are plotted upwards (based on [Wikimedia Commons, 2008]).

different brain rhythms and oscillatory activity can be observed in spontaneous brain activity, i.e. fluctuations in brain activity that are not related to a certain event or cognitive task. One of the most well known rhythms is the α-rhythm that occurs if a person relaxes with closed eyes [Niedermeyer and da Silva, 2005].

Signal Properties: The EEG is traditionally divided into frequency bands[4] δ (<4 Hz), θ (4-7 Hz), α (8-13 Hz), β (15-30 Hz), and γ (>30 Hz). User states and cognitive tasks can modulate the energy in these frequency bands and sub-bands thereof. For example, executed and imagined movements of body parts (motor execution and motor imagery) lead to an inhibition of sensorimotor rhythms. Specifically, oscillatory activity in the μ (7-12 Hz) and β (18-25 Hz) frequency bands is reduced predominantly near the primary motor cortex [Pfurtscheller and Neuper, 1997]. These characteristic signal changes are generally associated with a desynchronization of neural populations associated with motor processes (event related (de)synchronization, ERD/ERS). In ECoG signals, activity in the high γ frequency broadband (70-170 Hz) is generally associated with specific information about cortical functional processes [Crone et al., 2006, Roland et al., 2010]. Other cognitive tasks and user states that modulate neural oscillatory patterns include workload, vigilance, and affective states.

Hemodynamics and Blood-Oxygen Level Dependent (BOLD): Neural activity is generally associated with brain metabolic activity, which includes oxygen consumption of active neural cells (section 2.1.1, neurovascular coupling).

[4]The exact bounds of the frequency bands vary between users and are specified differently by different researchers (e.g. [Anokhin and Vogel, 1996]).

Therefore, changes in cerebral blood oxygenation are functional indicators for brain activity. Hemodynamics and the BOLD effect can be observed during many cognitive tasks, user states, and the perception of fast repetitive stimuli using fNIRS and fMRI.

Signal Properties: Figure 2.7 shows a prototypical hemodynamic response as measured by fNIRS (filtered and averaged activity over multiple trials). HbO_2 levels in a cortical area rise with brain activity and peek approximately 5 to 10 seconds after the beginning of activation, HbR levels fall in the same intervals.

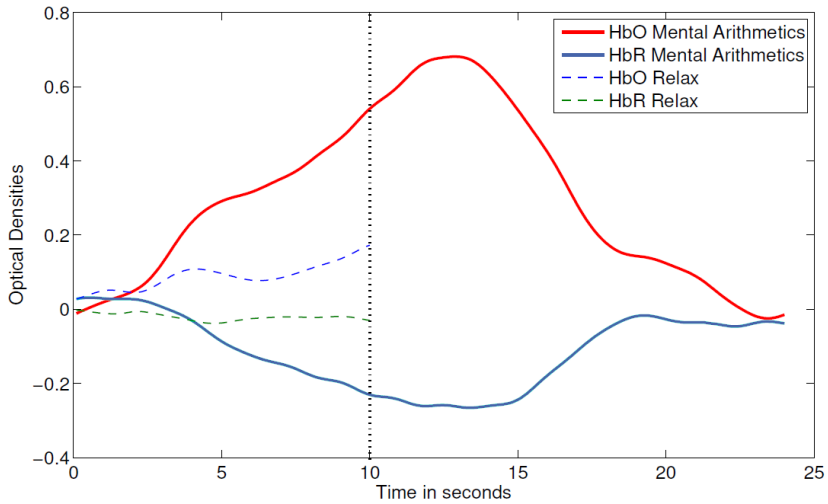


Figure 2.7 – Prototypical hemodynamic response measured by fNIRS. The red solid line corresponds to oxygenated hemoglobin concentrations (HbO_2), the solid blue line corresponds to deoxygenated hemoglobin (HbR) in response to cortical activity induced by mental arithmetics. The dashed lines correspond to HbO and HbR during relaxation phases. Signals of 30 times mental arithmetics or relaxation for 10 seconds (vertical line) have been frequency filtered (0.01-0.6 Hz) and averaged. The figure also shows the relaxation time after the 10 seconds of mental arithmetics where the activity returns to baseline [Herff et al., 2013b]

Steady-state evoked potentials: Steady-state (usually visual) evoked potentials correspond to the brains' response to the BCI paradigm 'perception of fast repetitive stimuli' (section 1.1.4). Evoked potentials can be acquired using electrical brain activity measurements, such as EEG and MEG.

Signal Properties: Activity patterns corresponding to the stimulation sequence can be observed at cortical areas of perceptive processing. For example the visual perception of repetitive stimuli, such as flickering lights causes corresponding activity patterns at occipital areas related to visual processing.

If the user attends one of multiple flickering light sources with different flickering frequencies, the attended source can be identified by analyzing peaks (and their harmonics) in the frequency spectrum at the occipital regions [Wolpaw and Wolpaw, 2011, chapter 14].

Slow cortical potentials: Slow cortical potentials are associated with a general increase and decrease cortical activation that can be used to control a BCI (e.g. thought translation device (TTD) [Kübler et al., 1999]). Users can learn to modulate slow cortical potentials by feedback training using different cognitive tasks [Hinterberger et al., 2004].
Signal Properties: Slow cortical potentials are characterized by direct current shifts (DC shifts) of the EEG signal and activity changes in the δ frequency band at large areas of the scalp with highest amplitudes around the vertex (position Cz in the international 10/20 system).

Variabilities in Brain Activity Signals

Noise in brain activity signals is primarily caused by the fact that brain signal acquisition modalities cannot capture the neural processes in detail because of their temporal and spatial resolution (see section 2.1.2). In addition to this general *measurement noise*, brain activity signals are inherently characterized by strong *inter-person*, *inter-session*, and *intra-session variabilities*.

A primary reason for the person specificity (inter-person variabilities) of brain activity signals is that different persons have different anatomies of their brains and skulls. Furthermore, there are variations in the functional organization of the brains due to neuroplasticity (learning). Inter-session variabilities can have technical reasons, such as inexact sensor repositioning or differences in environmental noise. Non-stationarities (next section), that include changes of psychological factors, can change between session. Additionally, different strategies in performing mental activities contribute to differences in the signal characteristics of brain activity patterns between persons and sessions.

In addition to these variabilities, artifacts and non-stationarities have a major influence on the signal-to-noise ratio of the acquired signals. They can be regarded as intra-session or trial-to-trial variabilities.

Artifacts and Non-Stationarities

We define artifacts in brain activity signals throughout this dissertation as

Definition: *Artifacts*

> *Signal parts that do not origin in the neural activity of the brain.*

In general, artifacts are specific to the brain signal acquisition modality. Here, we briefly discuss artifacts in EEG and fNIRS. ECoG signals are less prone to artifacts, but have similar signal characteristics as EEG signals, including the same types of artifacts. For all modalities used for BCIs, the amplitudes of artifacts in brain signals can be magnitudes higher than the amplitudes of neural activity.

One can distinguish between *biological artifacts* that are produced by the user's organism and *technical artifacts* that origin in the technical devices directly or indirectly involved in the measurement of brain activity signals.

Biological artifacts in EEG signals, include ocular (electrooculography, EOG) artifacts, i.e. artifacts from eye movements and blinks. One can recognize EOG artifacts in the measured signals by visual inspection because of their large amplitudes in low frequencies that are predominantly present at the frontal electrodes. EOG artifacts are primarily caused by the movement of the retinal or cornea-retinal dipole and the eyelids [Croft and Barry, 2000]. The electrical potentials induced by active muscle cells (electromyography, EMG) have a strong influence on a wide frequency range of the EEG signal, with peak energy between 20 and 30 Hz [Goncharova et al., 2003, Heger et al., 2011b]. Movements of the tongue, for example while speaking, cause glossokinetic artifacts. Similar to EOG, they are induced as the tongue has the physiological properties of a dipole and they are characterized by large amplitudes in low frequencies [Vanhatalo et al., 2003]. Other biological EEG artifacts come from cardiac activity (around 1-2 Hz) and sweating (below 0.1 Hz).
Technical artifacts in the EEG are induced by AC power lines (50 or 60 Hz), electromagnetic fields from external technical devices (e.g. fMRI), electrostatic chargings (e.g. office chair), movements of electrodes and cables, changes of electrode conductivity, and contact loss of electrodes.

Biological artifacts in fNIRS signals are primarily caused by cardiovascular activity, such as heart beats (primarily around 1-2 Hz) and slow waves (e.g. Mayer Waves, below 0.1 Hz), and changes of cardiac activation (around 1-2 Hz), breathing (below 0.5 Hz), and differences in blood pressure due to

changes in head orientation [Matthews et al., 2008, Cooper et al., 2012]. Technical artifacts in the fNIRS include spikes caused by optode movements and changes in environmental lighting conditions.

Throughout this dissertation we define non-stationarities as

Definition: *Non-stationarities*

> *Changes of the statistical distribution of brain activity signals over time that do not correspond to the characteristic signal properties modulated by the BCI paradigm.*

Usually researchers discriminate between artifacts and non-stationarities as they occur on different time scales. Artifacts are usually permanent or short-time effects, while non-stationarities are gradual changes that usually last for longer periods of time. Thus, after some time, brain activity patterns are rarely identically distributed to the ones observed during the initial calibration. Non-stationarities can have psychological or technical origin.

Technical reasons include physical properties of sensor acquisition changing over time, such as conductivity changes because of drying electrode gel. Psychological variables are concurrent neural processes that are not caused by the BCI paradigm, including changes in user states, such as vigilance, fatigue, emotions, mood, getting bored, neural plasticity, i.e. learning of the user, or a switch of task strategy.

2.2 Pattern Recognition for BCIs

2.2.1 Single-Trial Recognition

BCI experiments usually consist of multiple repetitions of the user generating brain activity signals that correspond to different classes or intensities of a BCI paradigm. The individual signal segments are commonly called *trials*. For example, a typical calibration session of a motor imagery BCI consists of EEG recordings of around 100 trials (2-3 seconds per trial) that correspond to the users' imagined movement of the left and right hand in randomized order.

There are two different protocols for single-trial recognition of BCIs, usually called synchronous and self-paced (or asynchronous) [Wolpaw and Wolpaw, 2011, chapter 10]. The majority of current BCIs

are so-called *synchronous* BCIs. With a synchronous BCI the user can only interact with the system at periods of time that are imposed by the system (trials). The system indicates by stimuli when a trial starts and user's brain activity patterns are evaluated. In contrast, in *self-paced BCIs* (or synonymously asynchronous BCIs), the user determines the points of time of the interaction. The self-paced interaction scheme is more natural and flexible, however it is more challenging for pattern recognition, as the point of time and the duration of relevant brain activity are unknown to the system. Therefore, stimulus locked evaluations, e.g. to evaluate ERPs, can usually not be performed. Furthermore, brain activity during periods of time in which the user does not intend to interact with the BCI can, in general, be very diverse and highly non-stationary, which makes their modeling challenging.

Many of the signal properties that characterize brain activity patterns (section 2.1.3) cannot easily be identified by visual inspection of the time series signals, even when performed by experts. In classical neurophysiological research, insights are obtained by averaging multiple trials of the same condition to infer an average model that reveals general signal properties of this condition. Averaging over a large number of trials significantly improves the signal-to-noise ratio of the brain activity signals (c.f. signal variabilities, section 2.1.3). However, this processing approach is not applicable for BCIs when they need to operate in real-time.

BCI research usually requires *single-trial recognition*, i.e. to derive an estimate for the current mental state or intention of the user from brain activity signals corresponding to a short segment of time. This is challenging and thus advanced machine learning techniques need to be applied to cope with the signal variabilities.

2.2.2 Pattern Recognition for BCIs - General Aspects

The pattern recognition component is the central component of a BCI (section 1.1.3). In general, pattern recognition is "concerned with the automatic discovery of regularities in data through the use of computer algorithms" [Bishop et al., 2006]. In this dissertation, we refer to the term 'pattern recognition' for all operations involved to *transform measured digital brain activity signals into a discrete or continuous recognition output* that corresponds to a class or intensity of a BCI paradigm.

Linear methods are widely used for pattern recognition in BCI research. Comparative evaluations have shown that linear methods often outperform non-linear methods in different BCI tasks (e.g. [Müller et al., 2003, Garrett et al., 2003]). Furthermore, as brain activity signals can be regarded as the superposition, i.e. a linear mixture, of multiple cortical activity sources, linear processing can be regarded a suitable approach. There is a large repertoire of linear pattern recognition methods that leverage the strong theoretical background of linear algebra. Linear methods are computationally efficient, understandable and interpretable. Furthermore, they can approximate many complex real-world processes adequately, are usually robust against outliers, and are often more stable than non-linear methods [Lotte et al., 2007].

Pattern recognition is often decomposed into three stages: *signal preprocessing*, *feature extraction*, and *machine learning*. In the signal preprocessing stage, the recorded signals are conditioned, for example, by filtering outliers and artifacts. The feature extraction stage extracts and selects relevant aspects of brain activity signals in time, frequency, and space that correspond to characteristics of the brain activity patterns modulated by the BCI paradigm (section 2.1.3). The resulting representations that are used to derive the class or intensity of a BCI paradigm are commonly called *features*. In the machine learning stage, statistical models are learned from calibration data that allow the recognition of unseen features from brain activity signals. Figure 2.8 summarizes the structural relationship between pattern recognition, brain activity signals, brain activity patterns, and the BCI paradigm (cf. section 1.1.4).

The next two sections discuss pattern recognition methods for BCIs in more detail. It should be noted that the transitions between the classical three stages of pattern recognition (signal pre-processing, feature extraction, and machine learning) are blurred and become even less relevant with modern pattern recognition techniques, which will be discussed in more detail in the context of the pattern recognition objective DISCRIMINATIVE (sections 3.2.1 and 3.3.1).

2.2.3 Signal Processing and Feature Extraction Methods

The main goal of signal processing and feature extraction in a BCI is to make the information contained in the data that should be recognized accessible

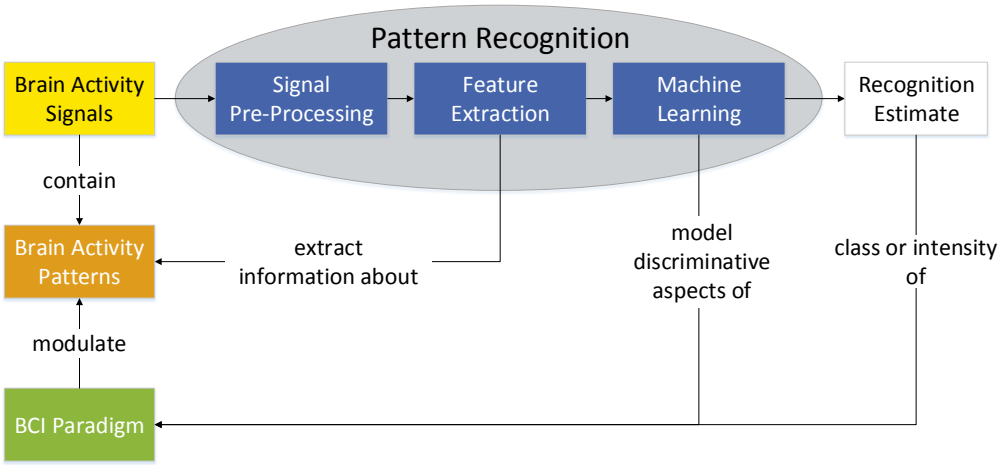


Figure 2.8 – Structural relationship between brain activity signals, pattern recognition, brain activity patterns, and the BCI paradigm.

from the raw brain activity signals. It is generally assumed, that brain activity patterns have spatial, temporal and spectral characteristics that are modulated by a BCI paradigm. Signal pre-processing and feature extraction can be regarded as the application of a series of filters to represent relevant signal parts (for review see [Wolpaw and Wolpaw, 2011], chapter 7).

Artifact Filtering

Eye movement artifacts can be subtracted from EEG signals using additional Electrooculography (EOG) recordings and weights calculated by regression analysis (EOG regression [Schlögl et al., 2007]). Furthermore, blind source separation methods are regularly used to remove artifacts from brain activity signals. Independent component analysis (ICA) can decompose the multivariate signals into statistical independent components. Components that correspond to artifact activity can be identified from their temporal, spectral, and spatial signal characteristics. When the inverse transform is applied, without considering the artifact components, cleaned signals can be reconstructed [Jung et al., 2000]. Additionally, wavelet-based methods have successfully been used to remove movement artifacts in fNIRS data [Molavi and Dumont, 2012].

Methods to remove non-stationarities have become a hot topic in BCI research. For example, the stationary subspace analysis (SSA) [von Bünau et al., 2009] decomposes multivariate brain activity signal into a stationary and a non-stationary subspace, such that the brain activity

signals can be projected into the stationary subspace.
In general, artifacts and non-stationarities can be filtered out from brain activity signals by frequency filters (next paragraph) if their frequency characteristics do not overlap with those relevant for the brain activity patterns.
More information on artifact reduction methods is provided in the review papers [Jung et al., 2000, Croft and Barry, 2000, Goncharova et al., 2003, Fatourechi et al., 2007, Matthews et al., 2008, Cooper et al., 2012, Molavi and Dumont, 2012] and in the context of the pattern recognition objective ROBUST (sections 3.2.3 and 3.3.3).

Filtering in Time, Frequency, and Space

- Temporal filtering
  The selection of relevant time windows within a trial can be used to isolate relevant time periods, e.g. ERPs components. Additionally, temporal filtering can remove irrelevant signal segments, for example at the beginning of the trial where no activity is present due to the reaction time of the user. Another temporal filtering technique is downsampling, which is regularly performed to reduce the amount of data that has to be processed and to remove high-frequency signal variations.

- Frequency filtering
  The removal and isolation of signal parts corresponding to specific frequency bands is performed by finite impulse response filters (FIR) and infinite impulse response filters (IIR), i.e. high-pass, low-pass, band-pass, and notch filters. They are commonly applied to brain activity signals to filter undesired activity (e.g. artifacts) or isolate oscillatory activity in specific frequency bands of interest.

- Spatial filtering
  Spatial filters are transformations of the multivariate brain activity signals in the dimension that corresponds to different sensors (different recording locations). Spatial filtering includes to select a subgroup of sensors that, for example corresponds to a particular cortical area. Other well-known spatial filters are the *common average reference*, which subtracts the average of all other channels from each channel and *surface Laplacians*, which subtract the average of its immediate neighbors from each channel.
  Furthermore, *linear transformations* can be regarded as spatial filters, i.e. transformations of the form $Y = WX$, where $Y \in \mathbb{R}^{a \times l}$ are

the spatially filtered signals, $W \in \mathbb{R}^{a \times c}$ is the spatial filter matrix consisting of a spatial filters and $X \in \mathbb{R}^{c \times l}$ are the unfiltered brain activity signals with c channels and l samples. Therefore, for example, unsupervised signal compression methods, such as principle component analysis (PCA), can also be regarded as spatial filters.
Common Spatial Patterns (CSPs) [Koles, 1991, Blankertz et al., 2008b] and their variants are among the most frequently applied algorithms in BCI research. They are primarily used with event-related (de)synchronization in oscillatory activity, such as for motor imagery classification (section 2.1.3). The fundamental idea of CSPs is to learn spatial filters that are designed to optimally discriminate between two classes according to their variance, i.e. the variance of the CSP filtered signal is large for class 1 while it is small for class 2 or vice versa. A CSP filter w^* is defined by maximizing the ratio between the covariances of the two classes Σ_1 and Σ_2 of the transformed data:

$$w^* = arg\max_w \frac{w^\top \Sigma_1 w}{w^\top \Sigma_2 w},$$
$$\text{subject to } \; w^\top(\Sigma_1 + \Sigma_2)w = 1$$

A matrix of CSP filters $W \in \mathbb{R}^{c \times c}$ can be calculated by common diagonalization of Σ_1 and Σ_2, for example by solving the generalized eigenvalue problem

$$W^\top \Sigma_1 W = W^\top(\Sigma_1 + \Sigma_2)W.$$

CSP filters correspond to columns in W. Usually the most discriminative CSP filters (first and last columns in W) are used for feature extraction.

The weights of linear spatial filters can be visualized in topographical plots. Figure 2.9 shows four CSP filters that have been calculated to discriminate motor imagery of the left hand and both feet in [Heger et al., 2013]. Characteristic dipoles can be recognized in sensorimotor regions.

Feature Extraction Methods

After artifact removal and filtering the signals in time, frequency, and space, a variety of different features can be calculated and represented in a feature vector for each trial:

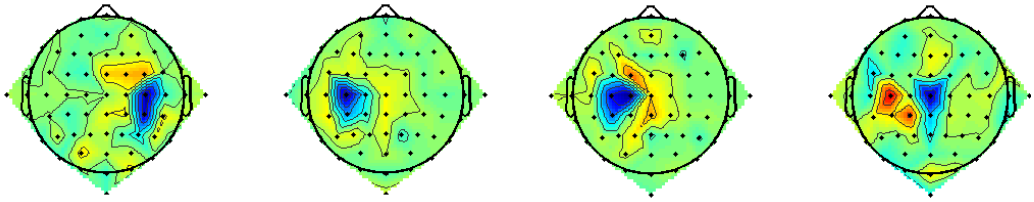


Figure 2.9 – Topographical scalp maps of four CSP filters trained to discriminate two classes of motor imagery (left hand versus both feet) from the experiment [Heger et al., 2013].

Time-domain feature extraction algorithms include the calculation of first and second order statistics, such as mean amplitudes and covariance matrices.
For oscillatory patterns, frequency features can be extracted using fast Fourier transform, Hilbert transforms, multitaper approaches, Welch's periodigram method, wavelet analysis, or autoregressive models.
Features that exploit more specialized knowledge of the signals include the slope of a hemodynamic response (e.g. [Herff et al., 2013a]), Hjorth parameters [Hjorth, 1970], or the spectral power in different frequency bands and ratios thereof for vigilance detection [Berka et al., 2007].
Inverse solution-based, connectivity-based features [Heger et al., 2014c] and feature from chaos theoretical measures have been proposed but are less frequently used for BCIs, they include phase synchronization [Gysels and Celka, 2004], spectral coherence, fractal dimension [Kulish et al., 2006] or Hurst exponents [Phothisonothai and Nakagawa, 2007].

To reduce the number of extracted features, automatic feature selection and compression methods can be used. Usually, separability and dependence measures, such as Fisher scores, correlation coefficients, and mutual information are most commonly applied. Additionally, wrapper approaches, such as sequential forward and backward selection or genetic algorithms have been applied (see [Bashashati et al., 2007, Tangermann, 2007] for review).

2.2.4 Machine Learning Methods

The goal of machine learning algorithms is to identify structures in the set of feature vectors extracted from calibration data that enable to recognize unseen feature vectors and infer the corresponding classes or intensities of

the BCI paradigm. In general, one can distinguish *classification* algorithms, that assign a discrete, categorical class label to each feature vector, and *regression* algorithms that assign a continuous variable to each feature vector [Bishop et al., 2006]. Class labels usually represent a type of mental activity the BCI user has performed, whereas continuous variables usually correspond to an intensity or a probability score of a mental activity.

In BCI research, there is traditionally more effort on the feature extraction stage, to filter relevant activity and represent informative features, than on the machine learning stage. Many standard machine learning techniques have been applied to BCI problems, among which Linear Discriminant Analysis (LDA) and Support Vector Machines (SVMs) are most widely used. Details about common machine learning algorithms can be found in numerous very good text books [Mitchell, 1997, Duda et al., 2001, Bishop et al., 2006, Hastie et al., 2009]. In their review article, Lotte et al. [Lotte et al., 2007] present a taxonomy of classifiers and discuss the most frequently ones used in the context of BCIs.

Chapter 3

Core Objectives of Pattern Recognition for BCIs

This chapter introduces three core objectives for pattern recognition in BCIs. We define each of the objectives, highlight their particular relevance to BCIs and relate each of them to principles of pattern recognition. Finally, we investigate the interdependence of the three objectives and conclude that all three objectives need to be implemented and balanced for successful pattern recognition for BCIs.

In the previous chapter, we have seen that in the pattern recognition component of a BCI, raw brain activity signals are typically filtered in time, frequency, and space, relevant features are extracted and learning algorithms are applied to recognize different classes or intensities of a BCI paradigm. This way, processing chains have been developed as best-practices for few established BCI paradigms, such as CSP-based motor imagery classification [Blankertz et al., 2008b] or classification of the event-related potentials during the Oddball paradigm [Blankertz et al., 2011] (see also section 1.1.4).

A major part of research on pattern recognition for BCIs is concerned with advancing the individual methods in these pattern recognition chains and with adopting new methods to improve the recognition results. Another important research direction of pattern recognition for BCIs is to explore new BCI paradigms. As a young discipline, BCI research offers the opportunity of a large number of unexplored paradigms that can potentially be employed to

create innovative BCIs. For such paradigms, it is, in general, not clear which patterns in the signal are discriminative, how features can be extracted, and whether they can be recognized in single-trials. Furthermore, if the results of a recognition system are not well above chance level, it is unclear whether the pattern recognition chain is unsuitable or the BCI paradigm cannot be recognized[1].

BCIs are a very challenging field for pattern recognition (e.g. [Nicolas-Alonso and Gomez-Gil, 2012]) and currently, there is no systematic theory on how pattern recognition components for BCIs should be developed. Therefore, it would particularly be advantageous to have a pattern recognition framework for BCIs that imposes minimal prior assumptions on the BCI paradigm and is known to be successful for recognizing many different BCI paradigms and many different brain activity signals.

In the following, we develop a very principled approach and analyze what the critical building blocks for successful pattern recognition for BCIs are. For this purpose, we identify three *core objectives* of pattern recognition methods (called DISCRIMINATIVE, COMPACT, and ROBUST) that are *necessary conditions of pattern recognition for BCIs*. In the next chapter we will see that a principled implementation of the three objectives using convex optimization leads to a new generic framework for BCI pattern recognition, which we call the *DCR Framework*. With the successful evaluations of the *DCR Framework* for multiple different BCI paradigms and brain activity signals, we provide empirical evidence that these three objectives are a *sufficient set of conditions* for BCI pattern recognition (chapters 5 and 6).

3.1 Definitions

In the following, we hypothesize that the three objectives DISCRIMINATIVE, COMPACT, and ROBUST are necessary conditions for BCI pattern recognition. We motivate why each of the objective is relevant for BCIs and which methods are used in current BCIs to implement the three objectives (section 3.2). To provide evidence that the three objectives are indeed necessity conditions, we relate each of them to principles of pattern recognition and show that if one of the objectives is not implemented, pattern recognition cannot be successful (section 3.3). Furthermore, we will see that the objectives depend on each other (section 3.4), which motivates that they

[1] Significant differences in statistical analyzes of the features do in general not imply that different classes or intensities of the BCI paradigm can be recognized in single-trials.

should be optimized jointly.

Hypothesis: *Three Core Objectives are Necessary Conditions*

For pattern recognition in BCIs it is necessary to implement and balance three core objectives:

- DISCRIMINATIVE: Discriminative brain activity patterns,
- COMPACT: Compact modeling, and
- ROBUST: Robustness against signal variabilities.

The objectives DISCRIMINATIVE, COMPACT, and ROBUST in this hypothesis are defined as follows:

Definition: *Discriminative brain activity patterns*

DISCRIMINATIVE*: The identification and modeling of representations of brain activity patterns that enable to disciminate between different classes or intensities of a BCI paradigm.*

Definition: *Compact modeling*

COMPACT*: The implementation of relevant aspects of discriminative brain activity patterns in generalizing structures with restricted complexity.*

Definition: *Robustness against signal variabilities*

ROBUST*: The implementation of discriminative brain activity patterns that is invariant and stable against those variabilites in brain activity signals that are not modulated by the BCI paradigm.*

Figure 3.1 illustrates the three objectives DISCRIMINATIVE, COMPACT, and ROBUST that form the pattern recognition component of a BCI (cf. figure 2.8). The arrangement of the three objectives is meant to indicates that the objectives depend on each other and the influence of the objectives has to be balanced, such that there is no bias towards one or two of the objectives, as will be discussed in section 3.4.

3.2 Goals and Relevance for BCIs

In this section, we describe each of the three objectives DISCRIMINATIVE, COMPACT, and ROBUST, that have been introduced in the previous section.

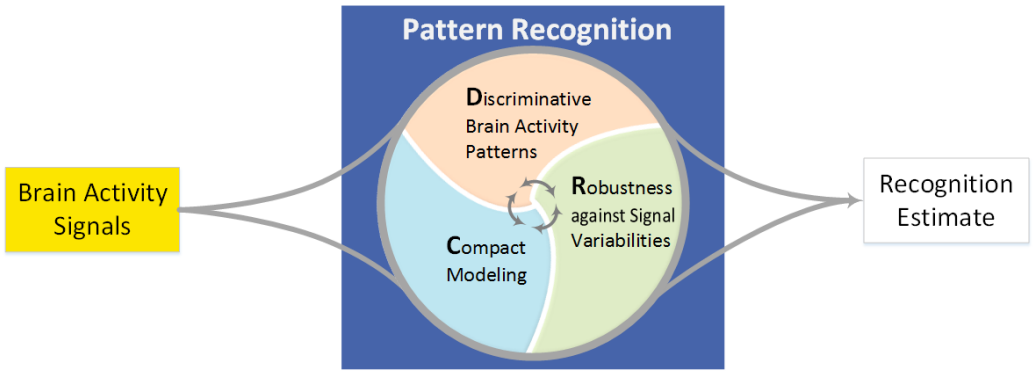


Figure 3.1 – The three core objectives DISCRIMINATIVE, COMPACT, and ROBUST.

We motivate their relevance for BCI pattern recognition and outline methods that are typically used to implement them.

3.2.1 Discriminative Brain Activity Patterns

The primary goal of a BCI is to discriminate different brain activity patterns that are modulated by a BCI paradigm and distinguish them from brain activity that is not related to the BCI paradigm. Therefore, DISCRIMINATIVE corresponds to the identification of different classes or different intensities of a BCI paradigm from brain activity patterns, i.e. to *identify and represent relevant information* of brain activity patterns, and to *learn discriminative aspects* of these representations. This enables the BCI program to use information about differences in states or intentions of the user for the purpose of the BCI application (section 1.1.3).

It is a common approach of pattern recognition to extract informative features and model their discriminative aspects by learning algorithms. However, the identification and representation of relevant information is especially important to BCIs, as the brain activity signals reflect the activity that is generated during the communication of the neural structures in the brain at a very coarse level (section 2.1.2). Therefore, the information is, in general, not directly accessible, i.e. the class or intensity of a BCI paradigm can usually not be inferred from the raw brain activity signals directly due to their low signal-to-noise-ratio (see section 2.1.2 temporal and spatial resolution). Furthermore, it is difficult to model the disciminative aspects of brain activity patterns by expert knowledge (e.g. neuroscientific insights), because

there are strong differences between different persons and sessions (section 2.1.3). Usually, brain activity signals cannot easily be interpreted by visual inspection, therefore, it is often difficult to make sure that the data quality is high. The recorded data is labeled based on task instructions and it is generally assumed that the user actually performs the task properly. This way, a significant amount of mislabeled data can usually be expected (label noise). In [Müller et al., 2008] the authors concluded that for BCI pattern recognition "the signal-to-noise ratio is highly unfavorable, in fact, it typically is even ill-defined what signal and, respectively, what noise are".
In summary, DISCRIMINATIVE corresponds to those aspects of pattern recognition for BCIs that represent the characteristics of the brain activity patterns in a way that is accessible to learning algorithms and the modeling of the particular discriminative aspects from these representations.

Methods for Discriminative Brain Activity Patterns

As described in section 2.2.3, a variety of different feature extraction algorithms have been used in BCI research that calculate representations of the information in brain activity signals in the form of feature vectors. Furthermore, section 2.2.4 discussed the typical machine learning approaches for BCIs that learn models of these representations and allow to discriminate between different classes or intensities of a BCI paradigm.

3.2.2 Compact Modeling

Calibration data acquisition is time consuming and fatiguing for the users and due to inter-session and inter-person variabilities, calibration data is usually recorded directly before the BCI can be used. Therefore, pattern recognition for BCIs has to deal with very small amounts of calibration data (typically few minutes). More precisely, there is a *low number of calibration instances compared to a large amount of potential features* [Müller et al., 2004], which is a challenge for the generalization abilities of learning algorithms (see also section 3.3.1).

Extracting a small number of features based on expert knowledge is often difficult, as signals are usually recorded by many sensors and brain activity patterns are subject to variabilities in time, frequency, and location. Therefore, an approach to overcome this problem is to extract a variety of potential features and apply compact modeling techniques to *restrict the complexity*

of their representation, for example, by identifying a subset of informative features. This often increases recognition performance as it avoids the inductive bias that comes from the assumptions made by the different operations. Furthermore, it is difficult to optimize the parameters of multiple operations in the pattern recognition component of a BCI as they usually depend on each other (e.g. [Farquhar and Hill, 2013]).

In neuroscience and BCI research, scientists are particularly interested in finding the neural substrate associated with particular brain functions. Compact modeling techniques are important to identify the most relevant structures in the high-dimensional representations of brain activity patterns to make the learned models more *interpretable*. This is, in particular, relevant to BCIs as models should be validated to make sure that they rely on neurophysiologically plausible properties of brain activity patterns and not on artifacts.

Methods for Compact Modeling

In a traditional BCI processing pipeline, compact modeling is performed by only extracting a small number of discriminative features (hand-tuned), i.e. *knowledge-based selection* of informative time periods, frequency bands and sensor locations. Additionally, automatic *feature selection* and *feature compression* algorithms can be applied as filter steps. In recent years, implicit feature selection by sparse modeling techniques, such as sparsity inducing regularizations, have become popular in machine learning and have also been applied to BCI for feature extraction and learning algorithms (e.g. [Lotte and Guan, 2011]). For example, ℓ_1-norm regularization, such as in Lasso regression [Tibshirani, 1996] is a central technique in multiple fields of machine learning, including sparse coding, dictionary learning, and compressive sensing. A regularized machine learning algorithm that is regularly used in BCI research and has shown to be suitable for multiple BCI problems is shrinkage Linear Discriminant Analysis (shrinkage LDA) [Friedman, 1989, Schäfer and Strimmer, 2005].

3.2.3 Robustness against Signal Variabilities

Brain activity signals are inherently subject to strong inter-subject, inter-session, and inter-trial variabilites caused by subjective differences, non-stationarities, and artifacts (see section 2.1.3). For example, artifacts often

overlap in their time, frequency and spatial characteristics with the actual brain activity patterns and can have significantly more energy than the brain activity signal parts. Therefore, they are a typical source of recognition errors and pattern recognition for BCIs should be invariant and stable against such variabilities. More precisely, strong changes in the signals should not effect the recognition output strongly (invariance against artifacts) and small changes in the signals should lead to small changes in recognition output (stability [Bousquet and Elisseeff, 2002]). The problem of signal variabilities becomes even more evident for the development of BCIs for not completely controlled laboratory conditions. Therefore, real-world BCI applications based on brain activity information and not on (task correlated) artifacts, may not be realized without approaches that carefully implement ROBUST.

Methods for Robustness against Signal Variabilities

In strongly restricted laboratory experiments, users can be instructed such that artifacts remain at a minimum level and recording conditions can be chosen to minimize non-stationarities. In such cases, the influences of signal variabilities may be ignored. In not completely controlled environments, signal pre-processing methods (section 2.2.3) need to be applied to reduce artifacts and to increase robustness against non-stationarities. Methods that increase the robustness against signal variabilities can also be integrated in spatial filtering and machine learning methods, such as extensions of the common spatial patterns algorithm presented in section 2.2.3 (e.g. [Wang and Zheng, 2008, Kang et al., 2009]). Additionally, adaptation techniques for learning algorithms [Vidaurre et al., 2006, Sugiyama and Kawanabe, 2012] and second-order baselining [Reuderink et al., 2011] have been proposed that adjust their models continuously over time to counteract non-stationarities.

Furthermore, data space adaptation methods [Arvaneh et al., 2013, Arvaneh et al., 2014] and multi-task or transfer learning techniques [Pan and Yang, 2010] (see also section 4.2) are required to reduce the impact of inter-session and inter-person variabilities (e.g. [Lotte et al., 2009, Heger et al., 2013]), as naïve approaches of combining brain activity signals from different persons and sessions (pooling of data sets) are usually not successful.

3.3 Relationship to Pattern Recognition Principles

In this section, we relate each of the three objectives to fundamental principles of pattern recognition to show that they are essential for a successful recognition (sections 3.3.1-3.3.3). As a consequence, the three objectives are necessary conditions for BCI pattern recognition, i.e. DISCRIMINATIVE, COMPACT, and ROBUST are a set of objectives that has to be implemented for successful pattern recognition for BCIs (section 3.3.4).

3.3.1 Discriminative Brain Activity Patterns

The objective DISCRIMINATIVE is concerned with the identification and modeling of representations of brain activity patterns. From a learning theoretical perspective DISCRIMINATIVE finds an informative mapping from representations of the information contained in the brain activity signals to a discriminative target space following the principles of statistical learning theory (e.g. Vapnik-Chervonenkis (VC) theory [Vapnik, 2000] or probably approximately correct (PAC) learning [Valiant, 1984]). Specifically, this modeling is performed by supervised machine learning methods that apply empirical risk minimization [Von Luxburg and Schölkopf, 2008], i.e. they learn a mapping $\hat{f}$ that minimizes the error indicated by a loss function L on n training instances $(X_i, Y_i), \ldots, (X_n, Y_n)$ with features X_i and labels Y_i corresponding to different classes or intensities of the BCI paradigm: $\hat{f} = arg\min_f \frac{1}{n} \sum_{i=1}^{n} L(X_i, Y_i, f(X_i))$. The features X_i are extracted to reduce irrelevant or redundant signal parts and provide access to latent information contained in the raw brain activity signals, e.g. with the help of integrating domain knowledge. The feature extraction can support the learning algorithm to find a suitable mapping $\hat{f}$ that models the characteristic properties of the actual recognition problem given a (small) sample of training data.

The two aspects 'identification' and 'modeling' of discriminative representations are subsumed by one objective (DISCRIMINATIVE), as they are inherently connected to each other [Bishop et al., 2006]. A machine learning algorithm can, in principle, learn feature representations automatically, if they are extracted only from information given in the data, i.e. without including additional expert or domain knowledge.

Note, in this context, the term DISCRIMINATIVE does not primarily correspond to discriminative classification. We refer in DISCRIMINATIVE to *dis-*

criminative and generative learning algorithms that perform *classification or regression*.

3.3.2 Compact Modeling

The *Bias-Variance Tradeoff* [Hastie et al., 2009] is a well-known machine learning principle on the complexity of a learned model in relation to its generalization abilities and the amount of training data available. In this context, *bias* corresponds to a systematic error of the learning algorithm due to ineffectively modeling the dependency between the features and the target variable. *Variance* denotes the sensitivity towards modeling small fluctuations of the training data. There is a tradeoff between bias and variance, which can be illustrated by decomposing the least-squares problem into the sum of three terms: bias, variance and an irreducible error (e.g. [Bishop et al., 2006]). On the one hand, high bias and low variance models may fail to fit the underlying distribution of the data accurately and tend to underfit, i.e. no matter how much data is available there will be a general error. On the other hand, low bias and high variance models can learn spurious patterns or idiosyncratic properties of the specific training data set and are generally susceptible to overfitting, i.e. they do not generalize well to unseen data.
The generalization error can be inferred from the difference between the recognition error on the training data and the recognition error on an unseen test data set. Minimizing the training error can increase the test error, if the complexity of the learning algorithm is high. The *VC dimension* (for Vapnik-Chervonenkis dimension [Vapnik, 2000]) gives a probabilistic upper bound for the test error given the training error and training sample size. This is related to the *Curse of Dimensionality* problem [Bellman and Dreyfus, 1962] that states that the volume of a vector space increases exponentially with the number of dimensions. According to this principle, the number of training instances should be large compared to the number of dimensions, i.e. features, in the feature space, otherwise the learned models have low generalization abilities. However, this relationship strongly depends on the learning algorithm used. In particular, the complexity of the model (function class), rather than the number of dimensions, has to be low for successful learning [Vapnik, 2000]. Recent examples in the context of sparse modeling have shown that learning can be highly successful, even if the number of features is much larger than the number of training instances. For example, under appropriate assumptions sparse modeling approaches can handle a number of irrelevant features that is exponential in the number of training instances [Zhao and Yu, 2006, Wainwright, 2009, Bach and Obozinski, 2010].

In sum, compact modeling techniques have to be applied in pattern recognition for BCIs to model the relationship between the training features and the target variables exactly and to have good generalization abilities on unseen data.

3.3.3 Robustness against Signal Variabilities

Usually, pattern recognition methods rely on the assumption that there exists an unknown joint probability distribution $P(X, Y)$ of the feature vectors X and corresponding class labels or intensities Y, from which the calibration features and corresponding classes or intensities (X_i, Y_i) are independently sampled (independent and identically distributed, iid.). In general, supervised machine learning algorithms assume that the *distribution of the features does not change over time* (e.g. PAC-learning [Valiant, 1984, Russell and Norvig, 2009]). This assumption is violated due to the non-stationary nature of brain activity signals. For example, the distribution of features during calibration of the BCI and the distribution during the operation of the BCI may not be identically distributed, which is a major problem for the supervised machine learning methods commonly used [Müller et al., 2008].

In addition to the principles of computational learning theory outlined in the previous paragraph, ROBUST is related to the *stability of a learning algorithm* [Bousquet and Elisseeff, 2002]. When using a stable learning algorithm, small changes in the composition of the calibration data have little impact on the learned models and therefore the recognition output.

3.3.4 The Objectives are Necessary Conditions

In the previous sections, we have described the relationship of the three objectives to principles of pattern recognition. If these principles, are not respected, pattern recognition is, in general, not successful: If DISCRIMINATIVE is not implemented, informative structures are not identified and modeled and remain hidden in the brain activity signals with very unfavorable signal-to-noise ratio. Thus, the classes or intensities of a BCI paradigm cannot be discriminated. If COMPACT is not implemented with the typically small data sets and the large amount of potential features in BCI research, high variance models are calculated by the learning algorithm and overfitting occurs that can strongly degrade recognition performance. If ROBUST is not

implemented, recognition is not successful, because the distribution of brain activity patterns changes over time, between recording sessions and different users. Furthermore, there is a high risk that the recognition is based on task related artifacts rather than on brain activity patterns.

3.4 Interdependence of the Objectives

The three core objectives DISCRIMINATIVE, COMPACT, and ROBUST are not independent of each other. Changes in the BCI's pattern recognition component with regard to one objective usually influence the effects of the other objectives. Therefore, it is important to balance the impact of the three objectives such that there is no bias towards one or two of the objectives.

In this section we outline the interdependence of all three core objectives, for both directions of each pairwise interdependence. We describe that emphasizing or neglecting one objective influences the other objective and may lead to essential problems in pattern recognition, i.e. low recognition performance.

Table 3.1 summarizes the interdependencies, i.e. how a bias towards implementation of one objective influences the others.

Discriminative and Compact

A bias towards implementing DISCRIMINATIVE influences COMPACT, as high variance models are susceptible to overfitting [Vapnik, 2000, Von Luxburg and Schölkopf, 2008]. For example, extracting a very large number of features that may generate a variety of discriminative information (high model complexity), however the information cannot be represented well by a machine learning model that has to be trained from a small number of training instances.

Neglecting DISCRIMINATIVE influences COMPACT as it leads to underfitting, i.e. large bias and low variance [Vapnik, 2000, Von Luxburg and Schölkopf, 2008]. For example, a single feature may not represent different classes of a BCI paradigm well and therefore cannot discriminate it precisely.

Influences \ Bias towards	Discriminative	Compact	Robust
Discriminative		Model complexity ↓ ⇒ risk of underfitting ↑	Invariance against relevant brain activity patterns ↑
Compact	Model complexity ↑ ⇒ risk of overfitting ↑		Explicit modeling / avoid modeling of signal variabilities ⇒ compact modeling ↓/↑
Robust	Modelling of non-informative signal patterns ↑ ⇒ risk of learning artifact patterns ↑	Invariance against signal variabilities ↑, Strong sparsity ⇒ stability ↓	

Table 3.1 – Summary of interdependencies of the three pattern recognition core objectives. A bias towards implementing the objectives (columns) influences the objectives (rows) as described in the corresponding cell. The symbol "⇒" indicates a consequence, "↑" and "↓" indicate an increase and decrease, respectively.

Discriminative and Robust

Implementing Discriminative but neglecting Robust leads to recognition errors caused by signal variabilities, such as artifacts and non-stationarities. Additionally, discriminative activity may not be modeled by the machine learning algorithms as it might be hidden by such non-informative variabilities. Neglecting Robust can also lead to learning of artifacts patterns instead of brain activity patterns, which implies that the user controls the BCI by generating systematic artifact instead of brain activity patterns.

Neglecting Discriminative influences Robust as it can lead to an increasing invariance against relevant non-stationarities of the brain activity patterns that are modulated by the BCI paradigm. For example, this problem is discussed in [Samek et al., 2014] in the context of early selection ap-

proaches to filter out non-stationarities, such as the stationary subspace analysis [von Bünau et al., 2009].

Compact and Robust

Neglecting or implementing COMPACT influences ROBUST as compact models have been found to improve generalization abilities and robustness characteristics (e.g. [Lotte and Guan, 2011] using regularized models for person transfer). On the other hand strongly COMPACT models can be less stable [Bousquet and Elisseeff, 2002]. For example, a sparse model that only uses the information of a single EEG electrode over motor cortex can be more robust against variabilities from eye movement artifacts (higher invariance), but can also be less robust in the case of electrode failure (lower stability). Furthermore, neglecting or implementing ROBUST influences COMPACT as robust models either ignore feature space parts that contain signal variabilities and can, therefore, be modeled more compact (lower variance), or have to model signal variabilities explicitly and are, therefore, less compact (higher variance).

3.5 Discussion

In this chapter, we introduced three core objectives for pattern recognition in BCIs, which we call DISCRIMINATIVE, COMPACT, and ROBUST (sections 3.2.1-3.2.3). We motivated their relevance for BCI pattern recognition and outlined typical methods that are applied for the purpose of one of the three objective in BCI research. Each of these objective can be related to essential pattern recognition principles (sections 3.3.1-3.3.3), which provides theoretical evidence for the hypothesis (section 3.1) that the three objectives are necessary conditions for BCI pattern recognition.

The first two objectives DISCRIMINATIVE, and COMPACT are related to well known and generally accepted principles of machine learning theory and practice. They are most widely known by effects that might occur if they are not implemented adequately, such as *overfitting* and *underfitting*, and have also been discussed in BCI literature (e.g. [Müller et al., 2004, Lotte et al., 2007]). The third objective also corresponds to a known principle in learning theory (namely *iid. sampling*) but is usually implemented independently from the other objectives in BCIs.

Pattern recognition methods that implement one or two of the three objectives are well-known. However, it is not established that the ROBUST objective should be implemented and balanced together with the DISCRIMINATIVE and COMPACT objectives that realize the bias-variance tradeoff. More recently, non-stationarity reduction has been implemented in combination with the objective DISCRIMINATIVE in the context of the Common Spatial Patterns algorithm for BCIs [Blankertz et al., 2007, Samek et al., 2014], but all three objectives have not been implemented jointly. Therefore, we also discussed the interdependencies of the three objectives in this chapter. To the best of our knowledge, it is the first time that the interdependence of these objectives have been formulated. Consequently, the three core objectives should be optimized in a joint optimization which leads to the *DCR Framework* that is discussed in the next chapter.

Note that the intention of this chapter was to identify core objectives of existing pattern recognition methods that are necessary conditions for successful BCI pattern recognition and not to introduce new methods. The innovation we provide is to outline the particular relevance of the interplay of all the three objectives for BCI pattern recognition in the form of the triad DISCRIMINATIVE, COMPACT, and ROBUST, which is not commonly known in BCI research.

Chapter 4

The DCR Framework

This chapter introduces the DCR Framework*, a generic pattern recognition framework for BCIs that jointly optimizes the three objectives* Discriminative*,* Compact*, and* Robust *discussed in the previous chapter. We formulate the problem of jointly optimizing the three objectives in a principled way and provide its theoretical background. Furthermore, we present an efficient optimization algorithm for the framework based on the Alternating Direction Method of Multipliers.*

According to the hypothesis that we formulated and discussed in chapter 3, pattern recognition for BCIs has to implement the three objectives Discriminative, Compact, and Robust. Furthermore, we have discussed that the three objectives are inherently interdependent (section 3.4). Therefore, each objective should be implemented in the pattern recognition component of a BCI with respect to the other objectives.

With the *DCR Framework*, we present a new generic framework for BCI pattern recognition, in which the three objectives are formulated as three terms of a convex objective function. To the best of our knowledge, no other BCI pattern recognition framework has been proposed before that performs a joint optimization of the three objectives. The framework is based on the idea to minimize efforts for pre-processing and feature extraction and let the learning algorithm be responsible to find relevant and robust information about the BCI paradigm. Therefore, we extract a large number of simple

features from the brain activity signals in time-domain or frequency-domain without performing spatial filtering and feature selection. The optimization problem to learn models for the recognition framework can be solved efficiently using the Alternating Direction Method of Multipliers (ADMM) [Boyd et al., 2011], an approach that has regained popularity for large-scale, distributed optimization in recent years. We present a new ADMM-based algorithm to perform the joint optimization that has not been proposed before. The models learned by the algorithm are linear transformations that can be visualized and are interpretable for experts. Our evaluations (chapters 5 and 6) show that the *DCR Framework* achieves competitive performance for multiple BCI paradigms and different brain activity patterns. To the best of our knowledge no other BCI pattern recognition framework has been evaluated with as many different types of brain signals and BCI paradigms before.

In summary, the goals of the *DCR Framework* are:

- Joint optimization of Discriminative, Compact, and Robust
- Unified approach combining isolated operations in BCI pattern recognition with little prior assumptions (i.e. no specialized feature extraction, but generic high-dimensional features)
- Efficient optimization algorithm based on ADMM
- Interpretable Models
- Generic framework applicable to multiple BCI paradigms and brain activity signals with competitive performance

In this chapter, we first formulate and discuss the objective function for joint optimization (sections 4.1.1 and 4.1.2), including the use of high-dimensional feature spaces and the concept of so-called robustness directions (sections 4.1.3 and 4.1.4). Section 4.2 discusses related work on optimization-based pattern recognition for BCIs and section 4.3 provides basic foundations on the Alternating Direction Method of Multipliers (ADMM). We present the ADMM-based algorithm for joint optimization in section 4.4. We discuss extensions of the algorithm for hyperparameter estimation, multi-class classification, and visualization in section 4.5. Finally, a discussion of the goals of the *DCR Framework* listed above, its limitations and a summary of contributions can be found in section 4.6. Figure 4.1 summarizes the structural relationship of the different components of the *DCR Framework*, including the section numbers where they are discussed.

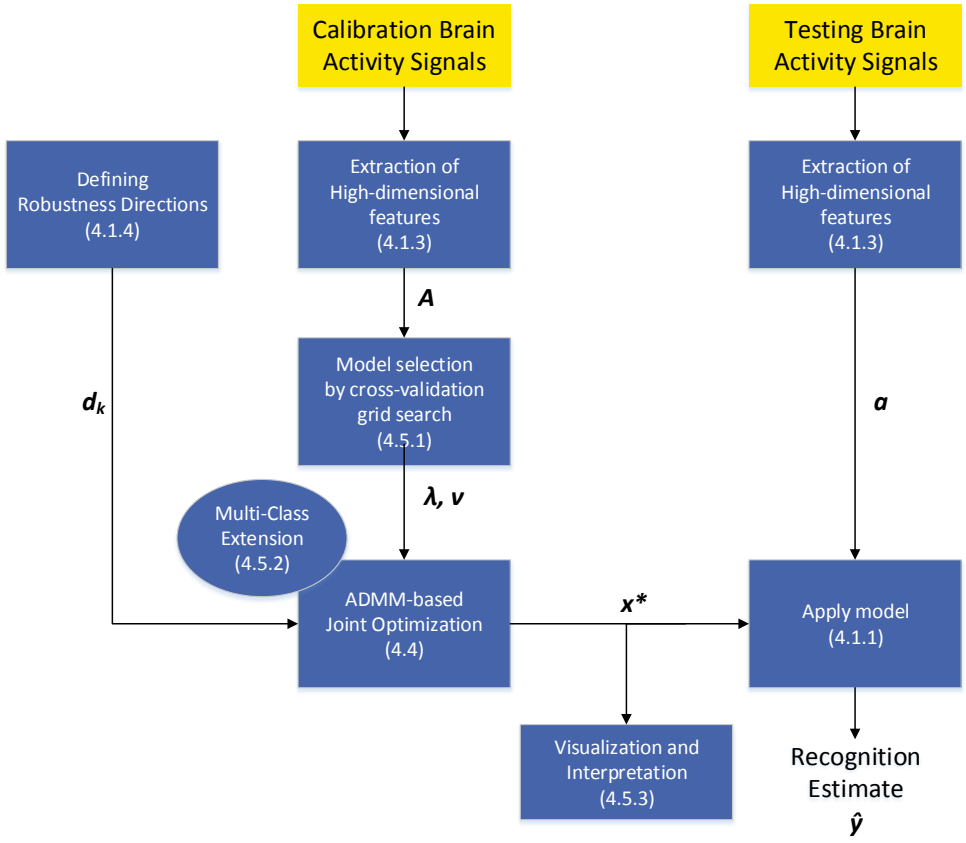


Figure 4.1 – Structural relationship of the components of the *DCR Framework*. The corresponding section numbers are shown in brackets.

4.1 Problem Formulation

4.1.1 Objective Function

The optimization problem of the *DCR Framework* is defined by the following unconstrained objective function [Heger et al., 2015], which consists of one additive term for each of the three objectives: DISCRIMINATIVE $f(x)$ using a least-squares regression based term to represent the relationship of input features and target values, COMPACT $g(x)$ a sparsity inducing ℓ_1-norm regularization term that performs an implicit feature selection, and ROBUST $h(x)$ a sum-of-norms based regularization term that corresponds to a flexible method to include directions in the feature spaces to which the

learned model is invariant and therefore robust against variabilities in these directions (robustness directions, see next section and section 4.1.4).

$$\begin{aligned} \underset{x}{\text{minimize}}\ & f(x) + g(x) + h(x), \\ \text{where}\ f(x) &= \|A \cdot x - y\|_2^2 , \\ g(x) &= \lambda \|x'\|_1 , \\ h(x) &= \nu \sum_{k=1}^{K} \left\| d_k^\top \cdot x \right\|_2 . \end{aligned} \tag{4.1}$$

The vector $x \in \mathbb{R}^n$ is the model to be learned, the matrix $A \in \mathbb{R}^{m \times n}$ consists of m feature vectors extracted from the brain activity signals of the calibration trials to which an additional 1 was added to cover additive biases (augmented feature vectors), m is the number of trials, n is the augmented feature vector length. The target vector $y \in \mathbb{R}^m$ represents the target values for each trial, e.g. indicating the ground truth class labels of each trial. The scalar factors $\lambda \in \mathbb{R}_+$ and $\nu \in \mathbb{R}_+$ determine the influence of the COMPACT and ROBUST terms and are estimated by cross-validation on the training data (see section 4.5.1). The vector $x' \in \mathbb{R}^{n-1}$ corresponds to the first n dimensions of vector x to avoid that the additive bias offset is removed by the implicit feature selection of the ℓ_1-norm regularization. The vectors $d_k \in \mathbb{R}^n$ correspond to the *robustness directions* in the *DCR Framework* (next section).

After optimization, a solution x^* of the optimization problem (4.1) can be used to predict the target value $\hat{y}$ of an unseen feature vector a by

$$\hat{y} = a^\top x^* . \tag{4.2}$$

Alternatively, for a classification problem, a binary classification result $\hat{y}$ instead of a regression result can be obtained by

$$\hat{y} = sgn(a^\top x^* - \tau), \tag{4.3}$$

where τ is a decision offset (as discussed in the next section).

4.1.2 Rationale of the Formulated Problem

The DISCRIMINATIVE-term in the optimization problem (4.1) estimates a least-squares regression model x. This model is a linear transform that

can be regarded as a spatio-temporal filter directly including the prediction of the target y. It has the advantage that it can be used for regression, but also for binary classification, when the target values in y are chosen appropriately and the sign function is applied. In the case of classification, prior knowledge of the distribution of the different classes can be incorporated in a flexible way by setting the target values in y for the first class to $\tilde{y}_+ = \frac{n_+ + n_-}{n_+}$, and for the second class to $\tilde{y}_- = -\frac{n_+ + n_-}{n_-}$, where n_+ and n_- are the numbers of training instances for the first class and second class, respectively. The corresponding decision offset is set to $\tau = \tilde{y}_+ + \frac{1}{2}(\tilde{y}_- - \tilde{y}_+)$. This choice of y makes the least-squares formulation equivalent to Linear Discriminant Analysis (LDA, more details on this relationship can be found in [Duda et al., 2001], chapter 5.8.2).
Besides least-squares regression, other loss functions can be implemented in the *DCR Framework*, such as logistic regression [Hosmer and Lemeshow, 2000]. However, we decided for linear regression because of the flexibility to use it for both, regression and classification problems and its equivalence in the classification case to LDA, which is among the most successful classification algorithms for BCIs (see e.g. [Lotte et al., 2007], section 2.2.4).

The Compact-term in the optimization problem (4.1) is an ℓ_1-norm regularization, which is well-known for its sparsity inducing properties [Tibshirani, 1996]. Increasing the weight $\lambda \in \mathbb{R}_+$ for the Compact-term (see equation 4.1) causes the number of zero coefficients in x to increase, which corresponds to an implicit feature selection due to the sparsity of the model. Therefore, selecting an appropriate value for λ allows for an optimal bias-variance trade-off.
Besides the ℓ_1-norm regularization, other regularization terms to restrict the model complexity are possible, such as Tikhonov regularization [Hoerl and Kennard, 1970], dual spectral norm regularization [Fazel et al., 2001], and others (see e.g. [Hurley and Rickard, 2009] for review). We decided for the ℓ_1-norm, as it outperformed alternative regularization approaches for BCI problems (e.g. [Heger et al., 2014b]). Furthermore, it has been shown for ℓ_1-norm regularization that effective models can be learned, if the number of features is much larger than the number of instances [Zhao and Yu, 2006, Wainwright, 2009, Bach and Obozinski, 2010], which enables the *DCR Framework* to be used with high-dimensional features (cf. section 4.1.3). Using sparse models is computationally efficient. Additionally, in comparison with many alternative feature selection approaches to implement the Compact-objective (cf. section 2.2.3), sparsity inducing reg-

ularization does not analyze features independently (as in greedy selection) but takes their redundancy and relevance for the learned model into account.

The robustness direction vectors d_k are an innovative and generic approach to integrate the objective ROBUST in the framework. They can be regarded as K directions in the feature space that can be chosen and whose influence is minimized in the joint optimization. Therefore, they enable to learn models that are invariant towards variabilities in these directions. To the best of our knowledge, a method to include feature space directions to which the learned models are invariant has not been proposed for other pattern recognition approaches before.

The two terms for DISCRIMINATIVE and COMPACT are commonly known from the Lasso [Tibshirani, 1996]. The three terms for DISCRIMINATIVE, COMPACT, and ROBUST may remind the reader of the elastic net [Zou and Hastie, 2005], which adds a ridge regularization (also called ℓ_2-norm penalty, Tikhonov regularization, or equivalent to Frobenius norm penalty) to the Lasso. Yet, instead of a ridge regularization, the optimization problem (4.1) includes a sum-of-norms regularization term for ROBUST (not squared), which is similar to the recently published sparse group Lasso [Simon et al., 2013]. However, in ROBUST the sum-of-norms term is not a penalty on x, but on $d_k^\top x$, i.e. the robustness directions projected by x. Therefore, the sum-of-norms regularization in ROBUST induces sparsity on the robustness directions. This means, it reduces the influence of the robustness directions to the model and performs an automatic selection of a subset of robustness directions to which the learned model is invariant. This is similar to the effect of group sparsity on the estimated model in group Lasso.
The intensity of the influence of the ROBUST term can be determined by choosing $\nu \in \mathbb{R}_+$ (see equation 4.1). As x can be interpreted in a classification task as the orthogonal vector of the separating hyperplane, more weight on the robustness term, causes x to become more and more orthogonal to the robustness directions d_k, i.e. the separating hyperplane becomes increasingly parallel to the robustness directions. Therefore, the classification is increasingly invariant to variabilities in the direction of robustness directions. Note that the separating hyperplane corresponding to x can be orthogonal to multiple or all robustness directions (i.e. $d_k^\top x = 0, \forall k$), if the nullspace[1] of the matrix consisting of all robustness directions has rank less than the number of dimensions n in the feature space.

[1]The nullspace of a matrix M is defined as $Null(M) = \{v \in \mathbb{R}^n | M \cdot v = 0\}$.

The optimization problem (4.1) is an unconstrained convex problem as all operations in the objective function preserve convexity [Boyd and Vandenberghe, 2004]. The Alternating Direction Method of Multipliers (ADMM) [Boyd et al., 2011] that can exploit the structure of the problem to solve the optimization efficiently. We present an algorithm and a detailed discussion for the *DCR Framework* in sections 4.3 and 4.4. Note that several standard solving approaches, such as gradient descent and (Quasi-)Newton approaches cannot be applied, since the ℓ_1-norm, which is important to implement the implicit feature selection for the COMPACT objective, is non-differentiable or they are not computationally efficient, especially for high-dimensional features (e.g. subgradient approaches).

4.1.3 High-Dimensional Feature Spaces

The *DCR Framework* is a generic recognizer for classification and regression problems that can be used with various kinds of features. However, it is in particular suitable to be used with generic high-dimensional feature spaces (see section 3.3.2, number of irrelevant features), which is in line with the approach to shift the burden of modeling relevant information of brain activity patterns from feature extraction to the learning algorithm stage (section 3.2.2).

As discussed in section 2.1.3, brain activity patterns encode information in time, frequency, and space. In BCI research typically learned spatial filters are applied to integrate spatial information in the extracted features [Blankertz et al., 2008b, Nicolas-Alonso and Gomez-Gil, 2012]. However, relevant spatial structures can be learned automatically in the joint optimization of the *DCR Framework* and applying spatial filters is not necessary.

Therefore, depending on the characteristics of the BCI paradigm, one or both of the following elementary features should be used with the *DCR Framework*:

- Time-domain features: Downsampled raw signals from each channel, and/or
- Frequency-domain features: Logarithmic Power Spectral Density estimates, for example calculated by Welch's method [Welch, 1967].

These features are extracted for each trail from the signals of each channel and vectorized to form a high-dimensional feature vector.

Downsampling of the raw time-domain signals can often be applied without loosing information, as brain activity signals are typically recorded using sampling rates that exceed the dynamics of the brain activity signals by multiple times. For example, modern EEG[2] acquisition devices can record signals at a sampling rate of 1 kHz, but the relevant dynamics of ERP components are usually below 20 Hz. Downsampling can decrease the number of dimensions tremendously, for example for a trial of ERP data with 32 channels and 1 second length, sampled at a rate of 1 kHz consists of 32000 samples, whereas it has 1280 dimensions if the data is downsampled to 40 Hz sampling rate. Downsampling may not be necessary to identify and learn appropriate models by the *DCR Framework*, but removes high frequency noise and can speed up computational time.
Frequency information usually cannot be learned directly form the raw time-domain data and should be extracted using one of the available approaches (section 2.2.3). Welch's methods is especially relevant for noisy data and has been found useful for many BCI problems (e.g. [Heger et al., 2011a, Heger et al., 2015]). Applying the logarithm to power spectral density estimates has been found to improve recognition performance as it makes the spectral features follow an approximately Gaussian distribution [Gasser et al., 1982].

The features discussed above are not commonly used for typical BCI pattern recognition tasks. The primary reason may be that they create very high-dimensional feature spaces that cannot be used with traditional learning algorithms and they require sparse methods for efficient processing. However, we found that generic high-dimensional features can outperform alternative approaches and can be applied to very different BCI problems and brain activity signals with different characteristics (see evaluations in chapters 5 and 6). Furthermore, they can be regarded as generic features as their extraction requires only a minimum of expert and domain knowledge and they are not specialized to a specific BCI paradigm.

4.1.4 Robustness Directions

As introduced in section 4.1.2, the robustness directions d_k in equation (4.1) are an innovative and generic approach to learn models with the *DCR Framework* that are invariant towards variabilities in the defined directions.

[2] A similar oversampling can be present in fNIRS recordings, depending on the acquisition hardware and recording settings.

Robustness direction can be chosen to reduce the impact of different kinds of variabilites. For example, they can be applied to perform supervised (session or person) transfer with the *DCR Framework*: Let $\mu_k^{\mathfrak{s}}$ be the mean vector of the feature vectors of class k extracted from the calibration data (source data set) and $\mu_k^{\mathfrak{t}}$ be the set of feature vectors extracted from the transfer data (target data set). Then, the robustness directions d_k can be set to the differences between the calibration data means and the transfer means:

$$d_k = \mu_k^{\mathfrak{s}} - \mu_k^{\mathfrak{t}}, \quad \forall k \in \mathcal{C}, \tag{4.4}$$

where $\mathcal{C}$ is the set of class indices. Analogue to this supervised transfer, an unsupervised transfer can be performed, if class labels are not available. In this case, a robustness direction can simply be set to the difference between the means of both data sets.

To remove non-stationarities (changes of the features over time, see section 2.1.3), the calibration data feature vectors of each class c are split chronologically into blocks $\mathcal{B}_c$ of equal size. The robustness directions d_k can be set to the difference between global mean feature vector μ_c of class c in the calibration data and the mean vector μ_c^i of the i-th block in $\mathcal{B}_c$:

$$d_k = \mu_c - \mu_c^i, \quad \forall k = (i, c) \in (\{1, \ldots, |\mathcal{B}|\} \times \mathcal{C}). \tag{4.5}$$

Analog to the person or session transfer, this can also be calculated in an unsupervised fashion, if class labels are not available.

In addition to person transfer, session transfer, and robustness against non-stationarities, other ROBUST-schemes may be realized by designing appropriate robustness directions, such as robustness against influences that are determined by expert knowledge or by influences from the experimental design.

4.2 Related Work on Optimization-based Pattern Recognition Frameworks for BCIs

Optimization Approaches

Many problems in pattern recognition, signal processing, machine learning, statistics and related fields can be formulated as optimization problems.

Since the 1950s intensive research led to the development of standardized techniques to solve different classes of convex optimization problems, such as linear programs, quadratic programs, second-order cone programs, semidefinite programs, and others (see [Boyd and Vandenberghe, 2004] for review). Solutions to these problems can be calculated by generic solvers, for example Glmnet (using a coordinate descent algorithm, [Friedman et al., 2010]), L-BFGS (using a quasi Newton approach, [Liu and Nocedal, 1989]), SDPT3 (using an interior-point method, [Toh et al., 1999]), and DAL (using an augmented Lagrangian approach, [Tomioka and Sugiyama, 2009]).
In comparison to these solvers, the Alternating Direction Method of Multipliers is a framework to implement optimization algorithms for many convex problems in a principled way that can be highly customized. ADMM is, in particular, suitable for our problem (4.1) as the objective function can be decomposed into two terms that can be solved extremely efficiently by exploiting the structure of the sub-problems (see section 4.4).

Unified Discriminative Frameworks for BCIs

A central concept of the *DCR Framework* is that the operations for preprocessing, feature extraction and classification in the pattern recognition component of a BCI are not regarded as isolated processes (see also section 3.2.2). In line with this basic idea [Tomioka and Müller, 2010, Mak et al., 2011, Makeig et al., 2012, Heger et al., 2014b], a few frameworks have been developed that unify multiple of the pattern recognition operations of a BCI:
Li and Guan [Li and Guan, 2006] proposed an extended expectation-maximization algorithm that iteratively updates the parameters of a Bayes classifier and common-spatial patterns filters.
Tomioka and Müller [Tomioka and Müller, 2010] proposed a regularized discriminative BCI framework. They used first-order or second-order features that are classified using a logistic regression predictor function and evaluated different regularization methods. Kothe et al. [Kothe and Makeig, 2013] followed the basic ideas of their approach and integrated BCI pipelines based on the Dual Augmented Lagrangian optimization algorithm [Tomioka and Sugiyama, 2009] into BCILAB, a plugin of the widespread EEG processing toolbox EEGLAB [Delorme et al., 2011].
Christoforou et al. [Christoforou et al., 2010] developed the second-order bilinear discriminant analysis. It uses a bilinear model that includes spatial and temporal filtering of first and second order features. The authors formulated a non-convex (bi-convex) problem based on a logistic regression that

is solved separately for spatial and spectral filters by a coordinate descent algorithm that requires careful initialization.
Barachant et al. [Barachant et al., 2012] proposed a generic approach for BCI recognition based on Riemannian geometry. They perform the classification directly on the manifold of covariance matrices using Riemannian distances in the space of symmetric positive-definite matrices.
Recently, Santana et al. [Santana et al., 2014] proposed to jointly optimize temporal filtering, spatiotemporal projection and classification using a Deep Neural Network (DNN). They were able to train the DNN with small amounts of data by initialization of the individual layers using a CSP-based processing pipeline. After initialization the network was optimized using error backpropagation.
In comparison to our *DCR Framework*, these approaches integrate the Discriminative and the Compact objectives but do not include mechanisms to implement the Robust objective, such as the robustness directions.

Recently, Samek et al. [Samek et al., 2014] proposed the divergence Common Spatial Patterns framework, in which CSPs are formulated as optimization problems based on the Kullback-Leibler divergence or beta divergence. Divergence CSPs have shown competitive performance to several previously published CSP variants, which are more robust against non-stationarities, perform session transfer or person transfer.
In comparison to our *DCR Framework*, divergence CSPs combine the Discriminative and Robust objectives in one joint optimization. However, they do not optimize Compact and do not include classification within their approach.

Transfer and multi-task learning

In the last few years, learning schemes, commonly called transfer learning and multi-task learning, gained increasing attention in machine learning research. The basic idea of multi-task learning is to learn a general model that can be applied to several related problems, whereas the goal of transfer learning is to exploit knowledge from a target domain or problem to improve learning for this domain or problem. Pan and Yang [Pan and Yang, 2010] provided a thorough survey on the substantial body of machine learning literature that has been published on transfer learning. Many of the approaches are based on advanced machine learning techniques, that have not been applied to BCI problems or other real-world problems. Multi-task and transfer learning

methods in BCI research are primarily suitable to overcome the session and person variabilites which can reduce calibration times for BCIs.

Alamgir el al. [Alamgir et al., 2009] proposed a framework using Bayesian multi-task learning. They formulated a problem based on a logistic regression loss function that can be solved by coordinate descent. The problem is non-convex and can therefore, in general, only be optimized to a local optimal solution and not efficiently to a global optimum. They showed improvements in classification performance for person transfer learning of a motor imagery task using two electrodes at sensorimotor locations. Their method also improved classification in a setting with slightly different experimental setups. Another Bayesian multi-task learning system was developed by Kang et al. [Kang and Choi, 2011], who used it with a common spatial patterns based feature extraction.
Krauledat et al. [Krauledat et al., 2008] proposed a so-called zero-training approach, which identified prototypical spatial filters in multiple previously recorded sessions of a user using a clustering approach. The prototype filters have good generalization abilities and can allow BCI use directly after a very short recalibration time for bias adaptation. However, a large number of sessions of the user have to be available before he or she can benefit from the zero training approach.
Falzi et al. [Fazli et al., 2009] used an ensemble learning based approach to create a subject-independent BCI. They constructed subject-dependent classifiers for different frequency bands and sparsely combined their outputs using quadratic regression with ℓ_1 norm penalty regularization. They could achieve results comparable to subject-dependent reference methods using a bias-correction that was applied as an offline post-processing step.
Tu and Sun [Tu and Sun, 2012] presented a framework for subject-to-subject transfer on feature extraction and classification level. They extracted generalizing and subject-specific filters banks from a set of candidate filters generated using extreme energy ratio features by solving optimization problems. They employed a two level ensemble learning strategy. In the first level, they generated learners for both filter banks and combined both learners in the second level. Evaluations showed a successful subject-to-subject transfer.

These session and person multi-task or transfer learning approaches relate to aspects of the objective ROBUST. However, they are not unified discriminative frameworks that perform a joint optimization. Usually, they consist of multiple isolated operations that are not jointly optimized, such as specialized BCI features that are independently extracted. Furthermore, these general learning frameworks usually cannot handle intra-session variablilites, such as modeling invariances against non-stationarities. In contrast, the *DCR*

Framework has a great flexibility by designing robustness directions to implement ROBUST. This feature enables to perform transfer learning, multi-task learning, robustness against non-stationarities, and other schemes.

4.3 Alternating Direction Method of Multipliers

The Alternating Direction Method of Multipliers (ADMM) is a framework for solving large-scale distributed optimization problems. It recently gained popularity as it has shown very competitive performance for many large-scale problems in signal processing, machine learning and related areas. Boyd et al. detail about designing ADMM algorithms in their excellent review article [Boyd et al., 2011] and present some generic design patterns that we applied in our optimization algorithm.

4.3.1 General Form of ADMM Problems

The basic problem formulation for ADMM is given by

$$\begin{aligned} &\text{minimize } f_1(x) + f_2(z), \\ &\text{subject to } Mx + Nz = c, \end{aligned} \tag{4.6}$$

with variables $x \in \mathbb{R}^{n_1}$ and $z \in \mathbb{R}^{n_2}$ and convex functions $f_1(x)$, $f_2(z)$, matrices $M \in \mathbb{R}^{p \times n_1}$, $N \in \mathbb{R}^{p \times n_2}$, and $c \in \mathbb{R}^p$.

This formulation is a general formulation for linear equality-constrained convex optimization problems with the optimization variable split in two parts $x \in \mathbb{R}^{n_1}$ and $z \in \mathbb{R}^{n_2}$, and the objective function split in two corresponding parts $f_1(x)$ and $f_2(z)$.

To solve the ADMM optimization problem (4.6) it is expressed in augmented Lagrangian form:

$$L_\rho(x, z, \gamma) = f_1(x) + f_2(z) + \gamma^\top (Mx + Nz - c) + (\rho/2) \left\| Mx + Nz - c \right\|_2^2 . \tag{4.7}$$

The augmented Lagrangian form is based on the Lagrangian form of problem (4.6). It only adds the additive term $(\rho/2) \left\| Mx + Nz - c \right\|_2^2$. This augmentation improves the robustness of a (sub)gradient ascent strongly and ensures convergence under less strict conditions, such as f_1 and f_2 do not have to

be finite or strictly convex [Boyd et al., 2011]. As the augmentation term represents the equality constraint of the optimization problem and vanishes if this constraint is satisfied, the choice of the augmented Lagrangian parameter $\rho \in \mathbb{R}_+$ does not change the optimization. Nonetheless, it can effect the convergence speed of the optimization algorithm (see dynamic ρ-updating, section 4.4.5).

The variable γ in equation (4.7) is the dual variable or Lagrange multiplier. For simplicity of notation, the dual variable γ is often scaled by the reciprocal value of ρ, i.e. the scaled dual variable u is defined by $u = \frac{1}{\rho} \cdot \gamma$. We will use this notation in the following sections.

The dual problem of the augmented Lagrangian (4.7) is given by

$$\begin{aligned} \underset{\gamma}{\text{maximize}} \quad & \inf_{x,z} L_\rho(x, z, \gamma), \\ \text{subject to} \quad & \gamma > 0. \end{aligned} \tag{4.8}$$

Assuming strong duality (see e.g. Slater's condition [Borwein and Lewis, 2010]), any solution of the dual function $\inf_{x,z} L_\rho(x, z, \gamma)$ is a lower bound of the primal optimization problem. Therefore, by maximizing the dual function (equation 4.8), an optimal solution of the optimization problem (4.6) can be calculated, which is a concave optimization problem.

4.3.2 Iterative Variable Updating

ADMM solves the problem (4.6) by iteratively calculating updates of the variables x, z, and u of the augmented Lagrangian (4.7). In each iteration these updates optimize the variables x, z, and u, respectively, while the other variables are held fixed.

The optimal value x^{k+1} in the $(k+1)$-th iteration is calculated by the x-update

$$x^{k+1} = \arg\min_x f_1(x) + (\rho/2) \left\| Mx + Nz^k - c + u^k \right\|_2^2 .$$

The optimal value z^{k+1} in the $(k+1)$-th iteration is calculated by the z-update

$$z^{k+1} = \arg\min_z f_2(z) + (\rho/2) \left\| Mx^{k+1} + Nz - c + u^k \right\|_2^2 .$$

The dual variable γ is updated as in dual (sub)gradient ascent [Boyd et al., 2011] using a step size equal to the augmented Lagrangian parameter ρ

$$\gamma^{k+1} = \gamma^k + \rho(Mx^{k+1} + Nz^{k+1} - c).$$

Therefore, the value u^{k+1} of the scaled dual variable in the $(k+1)$-th iteration is calculated by the u-update

$$u^{k+1} = u^k + Mx^{k+1} + Nz^{k+1} - c.$$

4.3.3 Convergence and Stopping Criteria

The updates described in the previous section are iteratively calculated until a stopping criterion is reached. The convergence of ADMM can be proven under mild assumptions, in general, it is necessary and sufficient that an optimal solution is primal and dual feasible (see section 3.3 of [Boyd et al., 2011]), i.e. the primal and dual residuals converge to zero norm. In the $(k+1)$-th iteration, one can calculate the primal residual by $r^{k+1} = Mx^{k+1} + Nz^{k+1} - c$ corresponding to the equality constraint of problem (4.6) and the dual residual corresponding to the dual feasible condition by $s^{k+1} = \rho M^\top N(z^{k+1} - z^k)$.

The stopping criteria of the iterative algorithm are determined by bounds on the suboptimality of the current point. Following [Boyd et al., 2011], we stop iterating if the primal and dual residuals are below bounds ϵ^{pri} and ϵ^{dual} that are determined by small absolute and relative tolerances, i.e. $\left\|s^k\right\|_2 \leq \epsilon^{pri}$ and $\left\|r^k\right\|_2 \leq \epsilon^{dual}$. In addition to these criteria, the algorithm usually stops after a given maximum number of iterations.

In theory it can be shown that it can take a large number of iteration until ADMM converges to high accuracy. The convergence rate of ADMM is still subject to current research. In [He and Yuan, 2012] a sublinear convergence rate of $O(1/k)$, where k is the number of iterations, has been shown for non-distributed computation, [Shi et al., 2014] showed linear convergence for more general and distributed cases. However, in practice ADMM converges comparably fast for many problems and a reasonably accuracy can be reached computationally very efficiently. For the BCI problems discussed in this thesis, it converges usually in less than 100 iterations. Furthermore, the calculations in each iteration are often very fast, e.g. if the problem structure can be exploited. Therefore, for our problems an optimal solution can be calculated in few seconds or less using a standard desktop computer.

4.4 Joint Optimization using the Alternating Direction Method of Multipliers

We can formulate our optimization problem (4.1) according to the standard ADMM form (4.6) by setting

$$\begin{aligned} f_1(x) &= f(x) + h(x), \\ f_2(z) &= g(z), \\ M &= I, \\ N &= -I, \\ c &= 0, \end{aligned} \tag{4.9}$$

where I is the identity matrix.

The settings for $f_1(x)$ and $f_2(z)$ define, that DISCRIMINATIVE and ROBUST are optimized during the x-updates and COMPACT is optimized during the z-update. The settings for M, N, and c define an equality constrained that enforces x and z to be identical.
Setting the variables in this way allows to exploit structures in the objective function, which enables a very efficient calculation: The DISCRIMINATIVE and the ROBUST term can be combined as both terms are convex quadratic and differentiable. Therefore, an optimal solution of this part of the optimization can be found by calculating the derivative and setting it to zero (see next section). Note, that this is only possible as ADMM allows to isolate this part of the problem. Furthermore, isolating the COMPACT term into the function $f_2(z)$, enables to calculate the optimal solution for this part by soft-thresholding [Daubechies et al., 2004], which is a very elegant and fast proximal gradient method [Combettes and Wajs, 2005] (see section 4.4.2).

As mentioned above, the optimization problem (4.1) can be solved in ADMM form by iteratively calculating the following x-, z-, and u-updates until the stopping criterion is reached:

4.4.1 x-Update

The x-update can be calculated by

$$\begin{aligned} x^{k+1} &= \arg\min_x f(x) + h(x) + (\rho/2)\left\|x - z^k + u^k\right\|_2^2, \\ &= \arg\min_x \|A \cdot x - y\|_2^2 + \nu \sum_{k=1}^{K} \left\|d_k^\top \cdot x\right\|_2 + (\rho/2)\left\|x - z^k + u^k\right\|_2^2, \end{aligned}$$

which has a closed form solution that can be obtained by calculating the derivative[3] with respect to x and setting the result to zero:

$$\begin{aligned} \nabla(\|A \cdot x - y\|_2^2 + \nu \sum_{k=1}^{K} \left\|d_k^\top \cdot x\right\|_2 + \quad (4.10) \\ (\rho/2)\left\|x - z^k + u^k\right\|_2^2) = 0 \iff \\ \frac{1}{2}A^\top(Ax^{k+1} - y) + \nu \sum_{k=1}^{K} \frac{d_k^\top}{||d_k^\top x^{k+1}||_2} d_k x^{k+1} + \\ (\rho/2)(x^{k+1} - z^k + u^k) = 0 \iff \\ \frac{1}{2}A^\top A x^{k+1} + \nu \sum_{k=1}^{K} \frac{d_k^\top}{||d_k^\top x^{k+1}||_2} d_k x^{k+1} + \\ (\rho/2)Ix^{k+1} - A^\top y - (\rho/2)(z^k + u^k) = 0 \iff \\ (\frac{1}{2}(A^\top A + \nu \sum_{k=1}^{K} \frac{d_k^\top}{||d_k^\top x^{k+1}||_2} d_k + (\rho/2)I)x^{k+1} = A^\top y - (\rho/2)(z^k + u^k) \quad (4.11) \end{aligned}$$

The equation above can be interpreted as a linear system of equations of the form

$$Dx^{k+1} = q, \tag{4.12}$$

with $D \in \mathbb{R}^{n \times n}$ and $q \in \mathbb{R}^n$. Solving this equation, for example by Gauss elimination or algorithms that exploit the sparsity of the problem, yields the value for x in iteration $k + 1$.

[3]The calculated gradient has been validated using empirical gradient checking, i.e. we estimated the slope a large series of random points (difference of the function value to the function values plus a small constant for each dimension) and compared it to the calculated partial derivatives, which results in a small accumulated error.

4.4.2 z-Update

The z-update can be calculated using a closed form solution:

$$\begin{aligned} z^{k+1} &= \arg\min_z g(z) + (\rho/2) \left\| x^{k+1} - z + u^k \right\|_2^2 \\ &= \arg\min_z \lambda \left\| z \right\|_1 + (\rho/2) \left\| x^{k+1} - z + u^k \right\|_2^2 \\ &= S_{\lambda/\rho}(x^{k+1} + u^k), \end{aligned} \tag{4.13}$$

where S_κ is the componentwise soft-thresholding function [Boyd et al., 2011] defined by

$$(S_\kappa(a))_i = \begin{cases} a_i - \kappa & : a_i > \kappa \\ 0 & : |a_i| \leq \kappa \\ a_i + \kappa & : a_i < -\kappa. \end{cases} \tag{4.14}$$

In this equation, the subscript i denotes the i-th element of the vector. The soft-thresholding function can be derived from the subgradient formulation and is the proximal operator of the ℓ_1-norm [Daubechies et al., 2004, Parikh and Boyd, 2013]. Its calculation is very simple and computationally extremely efficient.

4.4.3 u-Update

The u-update, i.e. the update of the scaled dual variable, can directly be calculated by

$$u^{k+1} = u^k + x^{k+1} - z^{k+1}. \tag{4.15}$$

This calculation is a simple sum of vectors and, therefore, very efficient.

4.4.4 ADMM Algorithm for Jointly Optimizing Discriminative, Compact, and Robust

The following algorithm shows the pseudocode of the ADMM-based algorithm for joint optimization of DISCRIMINATIVE, COMPACT, and ROBUST in the *DCR Framework*:

The input to the algorithm is a matrix composed of the calibration features A, the vector of target values for the regression or classification problem y (cf. section 4.1.2), and the robustness directions d_k (section 4.1.4). The hyperparameters λ and ν are estimated by cross-validation on the calibration

Input: A: matrix of calibration feature, y: vector of target values,
d_k: robustness directions,
λ: COMPACT-term weight, ν: ROBUST-term weight
Result: z: solution of the optimization
$x \leftarrow 0$, $z \leftarrow 0$, $u \leftarrow 0$;
$\rho \leftarrow 1$;
while *not convergence criteria met (section 4.3.3)* **do**
  x-update (equation 4.11 or equation 4.17 - section 4.4.5);
  z-update (equation 4.13);
  u-update (equation 4.15);
  dynamic ρ-update (section 4.4.5);
end

Algorithm 1: ADMM-based algorithm for joint optimization in the *DCR Framework*.

data as described in section 4.5.1. The initial values of x, z and u can be set to zero, ρ is initialized by 1.

In each iteration, the x-update, the z-update and the u-update are performed. Furthermore, the augmented Lagrangian parameter ρ is updated to increase convergence speed (see next section). The iterations are performed until the convergence criteria are met. The (global) optimal solution of the optimization problem is the output of the algorithm, which corresponds to the variable z after the last iteration.

4.4.5 Improving Memory Consumption and Computational Time

Optimizing the x-update for High-Dimensional of Features

To use the algorithm presented above for the BCI problems we intend to solve, it is essential to optimize calculations for the case that the number of dimensions in the feature space is much larger than the number of training instances ($n \gg m$). Calculating the matrix D in equation (4.12) (left hand side of equation (4.11)) can have extensive memory requirements and can be computationally expensive as D is an n-dimensional square matrix (in many practical examples, the number of features n can well exceed 10000, as discussed in section 4.1.3). One can exploit the structure of the problem,

i.e. the fact that the rank of the m-by-n-matrix A cannot be larger than m. The Woodbury matrix identity [Woodbury, 1950, Higham, 2002] can be applied to equation (4.11), which makes the calculation efficient.

Equation (4.11) can be rewritten as

$$\frac{1}{2}(\begin{bmatrix} A^\top & \nu\frac{d_1}{||d_1^\top x^{k+1}||_2} \cdots \nu\frac{d_K}{||d_K^\top x^{k+1}||_2} \end{bmatrix} \begin{bmatrix} A \\ d_1 \\ \vdots \\ d_K \end{bmatrix} + (\rho/2)I)x^{k+1} = A^\top y - (\rho/2)(z^k + u^k).$$

Substituting

$$\begin{aligned} Q &= \begin{bmatrix} A^\top & \nu\frac{d_1}{||d_1^\top x^{k+1}||_2} \cdots \nu\frac{d_K}{||d_K^\top x^{k+1}||_2} \end{bmatrix}, \\ R &= \begin{bmatrix} A^\top & d_1 \cdots d_K \end{bmatrix}^\top, \end{aligned}$$

results in

$$\frac{1}{2}(QR + (\rho/2)I)x^{k+1} = A^\top y - (\rho/2)(z^k + u^k). \tag{4.16}$$

Applying the Woodbury matrix identity to equation (4.16) can speed up calculations and memory consumption tremendously. Therefore, we calculate the x-update by solving the following linear system:

$$(\frac{1}{\rho}I^n - Q(I^m + \frac{1}{\rho}RQ))x^{k+1} = \frac{1}{\rho^2}RA^\top y + \rho(z^k - u^k), \tag{4.17}$$

where I^l is the l-by-l identity matrix. This transformation achieves a speedup in comparison to equation (4.11) as no matrix calculations have to be performed using $n \times n$ matrices.

Additional speedups can be achieved by caching constant terms. Furthermore, naïve approaches to solve the linear system of equations (4.17) should be avoided, such as those using matrix inversion. Instead, a solution should be calculated exploiting the sparsity of the problem.

Dynamic ρ-update

To accelerate convergence, we use a dynamic ρ-updating, which is a well-known extension for ADMM. Specifically, we applied the following updating heuristic (as suggested in [Boyd et al., 2011]):

$$\rho = \begin{cases} 2\rho & : \left\|r^k\right\|_2 > 10 \left\|s^k\right\|_2 \\ \frac{1}{2}\rho & : \left\|s^k\right\|_2 > 10 \left\|r^k\right\|_2 , \end{cases} \tag{4.18}$$

i.e. we doubled ρ, if the norm of the primal residual exceeds 10 times the norm of the dual residual, and set ρ to its half, if the norm of the dual residual exceeds 10 times the norm of the primal residual.

4.5 Extensions of the Joint Optimization algorithm

4.5.1 Model selection

The optimization problem (4.1) has two free parameters, λ and ν, that have to be chosen appropriately to balance the influence of the DISCRIMINATIVE, COMPACT, and ROBUST-terms. Grid search using a range of different parameters for λ and ν can be applied to find an optimal parameter configuration for the given data set. This procedure is very common for hyperparameter estimation, for example, it is commonly used in support vector machines (e.g. [Lin et al., 2003]). For each parameter configuration, the recognition performance is estimated by cross-validation on the training data and the best performing values are selected for λ and ν.

Figure 4.2 shows an example of such a cross-validation grid search taken from the evaluation in section 5.3. The classification accuracies for each cross-validation on the training data are shown as a heat map.

One can see that the recognition accuracy depends on the appropriate choice of both parameters, i.e. it requires all three term for DISCRIMINATIVE, COMPACT, and ROBUST to achieve an optimal performance. Particularly, the optimal performance cannot be achieved if either λ or ν are set to 0. In this example, an optimal choice for the two parameters increases the performance from 55% (DISCRIMINATIVE-term only, $\lambda = 0$ and $\nu = 0$) to 85% (e.g. $\lambda = 0.0399$ and $\nu = 0.067$).

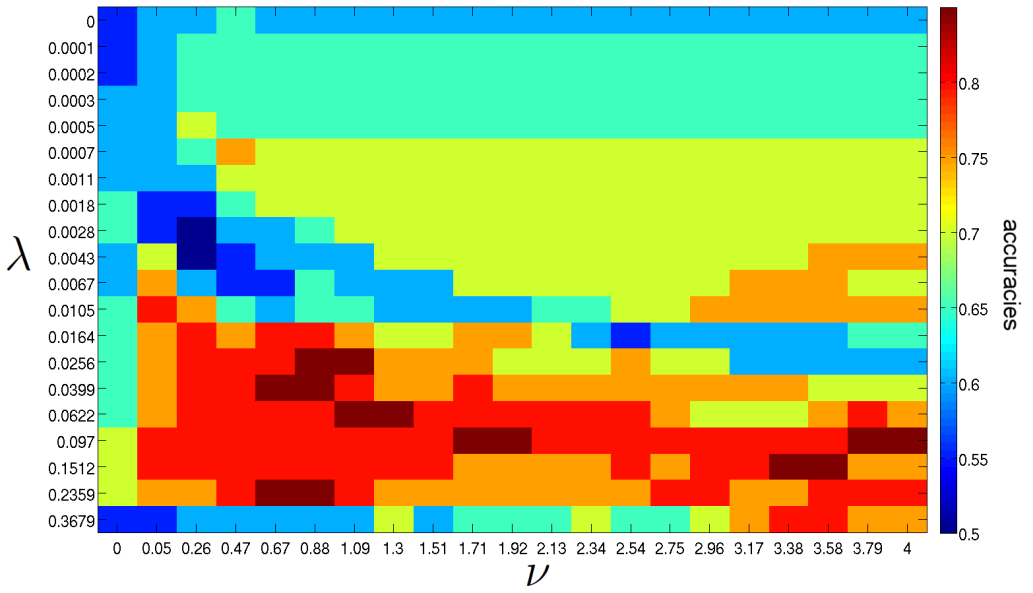


Figure 4.2 – Example for the estimation of optimal parameters for λ and ν by grid search. The figure shows the heat map of the cross-validation accuracies for different choices of the parameters.

4.5.2 Multi-Class Classification

The common one-vs-rest and one-vs-one classification strategies can be used to apply the *DCR Framework* to multi-class (multinominal) classification problems.

The one-vs-rest strategy trains one classifier for each class that discriminates the class from all other classes. It predicts the class with the highest confidence score, whereby the predicted target value without applying the sign function can be used as confidence score, i.e. $\tilde{y} = a^\top x^* - \tau$ (cf. last paragraph of section 4.1.1).

The one-vs-one strategy trains $|\mathcal{C}| \cdot (|\mathcal{C}| - 1)/2$ classifiers to discriminate between each pair of classes, where $|\mathcal{C}|$ is the number of classes. It predicts the class with the highest score using a majority voting of the classification results.

4.5.3 Interpretability and Visualization

The objective function of the *DCR Framework* is a principled mathematical formulation that can be directly related to the three core objectives Discriminative, Compact, and Robust. The solution by convex optimization leads to a global optimum, i.e. the learned model x reflects an optimal representation of the data that can be used to draw conclusions about the characteristic signals patterns of the BCI paradigm.

A learned linear model x can be visualized, which allows to interpret what has been learned. For example, this can be used to validate that the model has learned neural patterns and is not based on artifacts. Furthermore, it can give insights into relevant structures in the brain activity data, which is in particular relevant for the development of recognition systems for novel BCI paradigms and to use BCI as a research tool for cognitive neuroscience.

The model x can be considered as a backward model, in which each value is a weight for the corresponding value in the feature vector that indicates how this feature contributes to the prediction of a target value. However, these backward model weights cannot be directly interpreted to determine the origin of neural processes in time, frequency, or space.

Haufe et al. [Haufe et al., 2014] recently proposed the following method to convert a linear backward model x into a forward model w (generative model), which allows neurophysiologcial interpretation of the weight vector:

$$w = \Sigma_A \cdot x \cdot \Sigma_{Ax}^{-1}, \tag{4.19}$$

where Σ_A is the covariance matrix of the zero mean feature matrix A and Σ_{Ax} is the covariance matrix of the predicted features.

The linear forward and backward models can be visualized as each coefficient in the model x can be related to a particular sensor location and time offset or frequency bin. Thus, for time-domain features, the spatial distribution can be plotted over time using multiple topographical maps. For frequency features, the spatial distribution for multiple specific frequency bands can be visualized.

Figure 4.3 shows an example of forward models from an fNIRS experiment using time-domain signals. Models for 8 channels oxygenated hemoglobin (HbR) and deoxygenated hemoglobin (HbO) of two users (S1 and S2) are shown. One can clearly see from the HbO models that recognition is based on the difference between the first 20 and the second 20 seconds of the trial,

which is a reasonable model to classify the increase of a typical hemodynamic response (cf. figure 2.7 section 2.1.3).

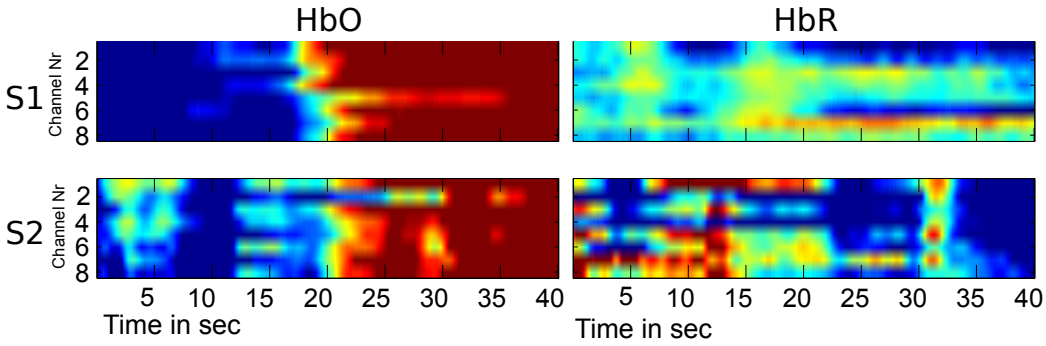


Figure 4.3 – fNIRS forward models of 8 channels oxygenated hemoglobin (HbR) and deoxygenated hemoglobin (HbO) of two users S1 and S2.

Figure 4.4 shows an example of a series of topographical scalp maps[4] of a backward model (lower row) and the corresponding forward model (upper row). The models correspond to the classification of motor imagery EEG for the user AV in the data set BCI3IVa that was recorded using 118 channels (see section 5.1.1). The model is shown for 6 different frequencies bands (model coefficients corresponding to a frequency band have been averaged). The forward model plots clearly show activity at motor regions particularly in the frequency bands 8-12 Hz and 20-24 Hz, which corresponds to the typical neuroscientific findings (cf. event-related (de)synchronization in motor imagery, section 2.1.3). In contrast to the forward models, the backward models contain activity that should not be interpreted as motor imagery related brain activity. For example, in the frequency band 12-16 Hz the backward model shows activity at frontal medial and left temporal regions that are not present in the corresponding topographical plot for the forward model.

4.6 Contributions and Discussion

The *DCR Framework* is a new BCI pattern recognition framework for *joint optimization* of the three core objectives Discriminative, Compact, and Robust. It follows a strictly principled approach to implement Discriminative, Compact, and Robust by joint optimization of the three ob-

[4]Plots indicate the spatial distribution of the activity over the scalp as viewed from top (depicted nose indicates frontal direction).

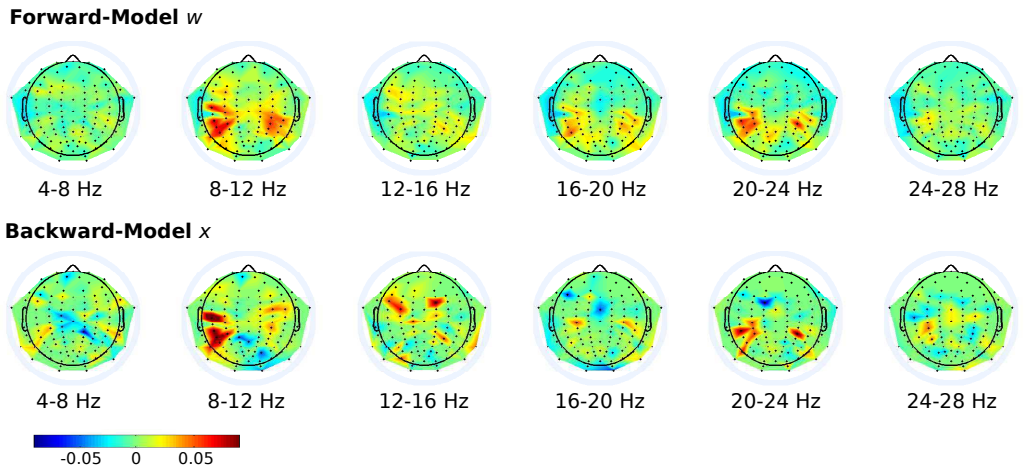


Figure 4.4 – Example of topographical scalp maps showing forward and backward models learned by the *DCR Framework* for EEG motor imagery classification.

jectives that are formulated as three terms in the objective function (equation (4.1)). In particular, this joint optimization enables a *unified approach* that combines multiple, typically isolated, processing steps for pattern recognition. For example, no specialized operations for spatiotemporal filtering, feature extraction, feature selection and classification are performed. Instead, generic high-dimensional features are extracted, which impose minimal prior assumptions on the BCI paradigm and require a minimum of expert and domain knowledge. Additionally, the automatic estimation of all parameters avoids inductive biases and can increase recognition performance compared to specialized BCI processing chains.
The learned linear models of the *DCR Framework* are *interpretable* and visualizing the models allows to verify what has been learned (previous section), which is a major advantage in comparison to other advanced pattern recognition approaches. They can give insights into relevant structures in the data, which is in particular relevant for the development of recognition systems for novel BCI paradigms.
To the best of our knowledge, no other BCI recognition framework optimizes the three objectives jointly. Furthermore, an optimization that includes a direct learning of both, regression or classification problems, and a flexible design for implementing the ROBUST objective, in supervised (i.e inductive) and unsupervised (i.e transductive) settings, has not been proposed before.

We presented an algorithm based on ADMM that allows for a computationally *efficient calculation* of the joint optimization. For all evaluations in this

thesis, the optimization can be performed on a desktop computer within few seconds. The algorithm presented here is a centralized algorithm, although a major advantages of ADMM is that large-scale problems can be optimized in distributed and decentralized settings (e.g. cluster computing). However, the amount of data in BCI research is currently very limited and, in practice, the major computational effort of using the *DCR Framework* comes from the grid search for model selection, which is trivial to calculate in parallel in a distributed setting as these calculations are independent. Nonetheless, the proposed optimization algorithm can easily be modified for distributed calculation, which may become relevant if the typical amount of data in BCI research will grow in the future.
To the best of our knowledge, it is the first time that ADMM has been applied to create a unified framework for BCI pattern recognition.

In BCI research, different BCI paradigms (section 1.1.4) are, typically, processed with dedicated pattern recognition approaches. The *DCR Framework* is designed to be a flexible and generic framework applicable to *various BCI paradigms and brain activity signals* with *competitive performance*.
To the best of our knowledge, no other generic framework for BCI pattern recognition has been evaluated with different BCI paradigms and different brain activity signals, which we will show in section 5 and chapter 6. The evaluations of the *DCR Framework* illustrate its advanced performance for many different BCI paradigms. Thereby, the *DCR Framework* follows a principled approach to translate the three objectives directly into a mathematical formulation that can be solved by joint convex optimization for very generic high-dimensional features. This suggests that, in addition to being necessary conditions (section 3.3.4), the objectives Discriminative, Compact, and Robust can be regarded a sufficient set of conditions for pattern recognition of a variety of BCI problems.

Limitations

In comparison to typical advanced pattern recognition systems, such as deep neural networks, multi-kernel learning, and others, the *DCR Framework* can only learn linear transformations, i.e. non-linear structures in the features may not be optimally modeled. However, one should keep in mind that linear models have been the most successful methods for BCI pattern recognition (see section 2.2.2 for advantages of linear operations). The primary reasons for the design decision to use linear models are the efficient calculation of optimal transformations, the interpretability of linear models and

high recognition performance for various BCI problems, especially when using high-dimensional features.

The proposed algorithm is computationally very efficient for practical problems. However, if the x-update (equation 4.17) is implemented using a naïve, dense Gauss elimination, the optimization algorithm has a computational worst-case complexity of $O(kn^3)$, where k is the number of iterations and n is the number of dimensions in the feature space. Nonetheless, due to the sparsity of the problem, the x-update can usually be calculated very fast, if efficient solvers and the speedups discussed in section 4.4.5 are applied. Usually, the algorithm converges in less than 100 iterations.

The robustness directions in the *DCR Framework* are a flexible way to guide the optimization towards learning models that are invariant against variabilities in these directions in the feature space. However, the robustness directions may not allow to learn invariances towards all kinds of artifacts and variabilities that may occur in brain activity signals. Robustness directions, such as those typically used for transfer learning or to learn invariances against non-stationarities as discussed in section 4.1.4, ignore the variances and higher order statistics of the variabilities. Therefore, multiple robustness directions are needed to be defined to be able to learn invariance against variabilites that correspond to changes of the feature distribution, including changes in mean and variances. Nonetheless, in the evaluations (chapters 5 and 6) we could show that, in practice, the transfer directions significantly increase performance for subject-transfer learning, session-transfer learning, and reduction of non-stationarities for different BCI problems and different brain activity signals. To the best of our knowledge, no algorithm has been proposed before in transfer learning, multi-task learning, or BCI literature, where robustness directions (or alike) can be defined to improve the robustness against signal variabilities of the learned models.

Chapter 5

Evaluation of the DCR Framework

In this chapter we evaluate the DCR Framework*using data. First, we isolate the* Discriminative *and* Compact *terms and analyze these two aspects of the* DCR Framework *using two different BCI problems. Then, we illustrate the characteristics of the* Robust*-term using synthetic data. Finally, we evaluate the joint optimization of* Discriminative*,* Compact *and* Robust *with the* DCR Framework *in two different BCI tasks.*

In this chapter, we first investigate the two terms for Discriminative and Compact of the *DCR Framework*, i.e. $f(x)$ and $g(x)$ in equation (4.1), to show that they can produce state-of-the-art performance for typical EEG and fNIRS data sets. Specifically, we set the weight for the Robust term to zero ($\nu = 0$), evaluate two different publicly available benchmark data sets and compare the performance of the *DCR Framework* with results of multiple alternative state-of-the-art BCI pattern recognition approaches that do not integrate specific pattern recognition mechanisms for Robust. Note, that evaluating $f(x)$ and $g(x)$ only, does not contradict the hypothesis that the three objectives are necessary conditions for pattern recognition in BCIs, as the implementation of $f(x)$ and $g(x)$ also includes some robustness against signal variabilities, because of the interdependence of Discriminative, Compact, and Robust. Furthermore, the analyzed data sets are recorded under controlled conditions and contain rather clean signals.

Second, we investigate the influences of the ROBUST-term, i.e. $h(x)$ in equation (4.1), in more detail using synthetic data. We show the effects of different values of the parameter ν on the separating hyperplane in two different classification tasks. In the first task, we set the robustness directions according to a simulated shift in the data distribution. In the second task, we use the robustness directions to learn models that are invariant against the activity of non-stationary sources in the data.

Third, we evaluate the complete *DCR Framework* using two different BCI data sets to show its advanced performance for person transfer. The first data set is a comparably large data set that consists of motor imagery data of 106 users. The second data set corresponds to an error potentials (event-related potentials) recognition task that was used in the BCI Challenge at the IEEE Neural Engineering conference 2015. In this competition the *DCR Framework* achieved a price winning performance (see section 5.4.4).

5.1 Evaluating the *DCR Framework* using DISCRIMINATIVE and COMPACT

We first evaluate the *DCR Framework* using generic high dimensional features (section 4.1.3), for two different BCI problems and compare it with common state-of-the-art BCI pattern recognition methods. In these evaluations, we want to highlight that the *DCR Framework* yields competitive recognition performance out-of-the-box for very different BCI paradigms without using specific optimizations.

5.1.1 Motor Imagery Classification

In the first evaluation, we perform motor imagery classification, which is one of the most frequently investigated BCI paradigms and a classical benchmark in BCI research.

Related Approaches

There is a substantial body of BCI literature on motor imagery classification. Most frequently, EEG-based BCIs apply the Common Spatial Patterns (CSP) algorithm or one of its numerous variants (see section 2.2.3). It is usually

applied to raw signals filtered using a frequency broadband, such as 8-30 Hz, to calculate a small number of discriminative spatial filters. Commonly, the variances are calculated from the CSP filtered signals and the logarithm is applied to Gaussianize the features. This way, CSPs extract a small set of specialized features for motor imagery classification.

One of the most successful CSP variants that was the winning approach for multiple tasks in BCI Competition IV [Tangermann et al., 2012] is the *Filter-Bank Common Spatial Patterns* (FBCSP) algorithm [Ang et al., 2008]. It applies a filter bank of, usually 4 Hz wide bandpass filters between 4 and 32 Hz (4-8 Hz, 8-12 Hz, ..., 28-32 Hz), to the raw data. The basic CSP algorithm is applied to each of the frequency filtered signals and the most discriminative spatial filters (e.g. the 2 first and 2 last columns of the CSP transformation matrix per class) are used to transform the data. The spatial filters that have the highest mutual information between the corresponding logarithmic variance features and the class labels are selected using kernel density estimation based mutual information feature selection [Ang et al., 2008].

The combination of CSP-based features and classification by Linear Discriminant Analysis (LDA) can be seen as the standard approach for EEG-based motor imagery recognition. LDA finds a linear transformation that minimizes the ratio of within-class scatter and between-class scatter of the calibration features. The LDA transformation can be calculated by the generalized eigenvalue analysis of the two scatter matrices [Duda et al., 2001].
If only small amounts of calibration data are available, shrinkage LDA [Schäfer and Strimmer, 2005] can be applied instead of the vanilla LDA, which implements the COMPACT objective. It employs Ledoit and Wolf's method for regularized empirical covariance matrix estimation [Ledoit and Wolf, 2004] to improve the estimates of the scatter matrices for high dimensional features, when only small amounts of data are available. The Ledoit-Wolf covariance estimator interpolates the sample covariance matrix with a unity matrix, whereby an asymptotically optimal interpolation weight is analytically determined. This generates an invertible, well-conditioned and more accurate estimate than the sample covariance matrix.

As CSPs are pattern recognition methods that are learned from training data, they should implement the COMPACT objective, in addition to the regularization of the classifier. There are different variants of regularized Common Spatial Patterns (rCSPs), a comparison can be found in [Lotte and Guan, 2011]. In the following evaluation we used *rCSPs with diagonal loading*. In this approach, the empirical covariances involved in the CSP calculation are linearly interpolated with a unity matrix, where the regularization parameter,

i.e. the weighting of the unity matrix, is estimated using cross-validation (see [Lotte and Guan, 2011]).

Besides CSP-based methods, high-dimensional spectral features (HDspec) can be used for motor imagery classification. They are not commonly used in literature but are well suited for the *DCR Framework*. In the evaluations below, they are extracted in 1 Hz wide frequency bands using Welch's method, as described in section 4.1.3.

Data Corpus: BCI Competition III data set IVa (BCI3IVa)

In this evaluation, we use one of the most frequently employed benchmark data sets in BCI research for motor imagery classification. It was published[1] as data set IVa for the BCI Competition III (BCI3IVa) [Blankertz et al., 2008a].
BCI3IVa contains EEG data of five users, which performed two classes of motor imagery: imagined right hand and right foot movements. For each trial in the experiment, a visual cue was shown for 3.5 seconds on the screen to indicate which class of motor imagery the user should perform. Between the trials there were pauses of random length between 1.75 and 2.25 seconds. Figure 5.1 illustrates the timings for a trial.

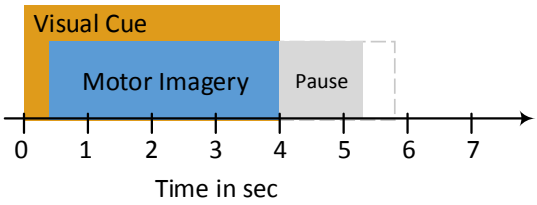


Figure 5.1 – Timings of a trial in BCI3IVa.

A different number of calibration and test trials is available for each user, which is shown in table 5.1. This way, a particular challenge of BCI3IVa is that there is only a little amount of calibration data available. The data set contains signals recorded using 118 channels downsampled to 100 Hz.

BCI3IVa consists of quite clean motor imagery data of trained BCI users. Therefore, additional specific ROBUST methods are not mandatory to obtain very good recognition results. For the DISCRIMINATIVE and COMPACT objectives we evaluate the different approaches outlined in the previous section.

[1] http://www.bbci.de/competition/iii/desc_IVa.html

User	# calibration trials	# test trials
AA	168	112
AL	224	56
AV	84	196
AW	56	224
AY	28	252

Table 5.1 – Number of trials for calibration and testing for each of the five users in BCI3IVa.

Because of the low number of calibration trials and high number of channels the COMPACT objective is particularly relevant for this data set.

Evaluation of BCI3IVa

In this evaluation, we discriminate between the two different classes of motor imagery in BCI3IVa. We compare the performance of generic high-dimensional spectral features (HDspec, as proposed for the *DCR Framework* in section 4.1.3) and Common Spatial Patterns based features, i.e. vanilla CSPs and FBCSPs (section 2.2.3).
The feature extraction approaches have been applied with parameter settings that typically achieve high recognition performance as follows. HDspec features correspond to 3304 features[2] in this experiment, whereas the 8 most discriminative CSP-based features were used and for FBCSP-based features the 22 most discriminative features were selected (see below). We compare classification by LDA based approaches with classification by the *DCR Framework* using only the DISCRIMINATIVE and COMPACT terms. For CSP and LDA we also evaluate regularized variants, called rCSP and sLDA, respectively.

We evaluate the following seven approaches that include well-established baseline approaches and approaches that are expected to achieve state-of-the-art performance in BCI3IVa (cf. section 5.1.1): We evaluate Linear Discriminant Analysis classification of features extracted by the Common Spatial Patterns algorithm (*LDA-CSP*). Furthermore, we evaluate the regularized variant of this approach, i.e. shrinkage Linear Discriminant Analysis classification of features extracted by regularized Common Spatial Patterns

[2]Corresponding to 28 frequency bands and 118 channels

based features (*sLDA-rCSP*). Moreover, we evaluate shrinkage Linear Discriminant Analysis classification of features extracted by the filter-bank Common Spatial Patterns algorithm using regularized CSPs (*sLDA-rFBCSP*). In addition to CSP-based features, we evaluate Linear Discriminant Analysis classification and high-dimensional spectral features (*LDA-HDspec*). Additionally, we evaluate shrinkage Linear Discriminant Analysis Classification and generic high-dimensional spectral features (*sLDA-HDspec*). Finally, we apply the *DCR Framework* to high-dimensional spectral features (*DCFram-HDspec*) and compare it to regularized Common Spatial Patterns based features (*DCFram-rCSP*).

Results and Discussion

Table 5.2 lists the feature extraction and classification approaches for comparison and summarizes how the objectives DISCRIMINATIVE and COMPACT are implemented. It also shows the number of features extracted (DISCRIMINATIVE column) and number of features modeled for the COMPACT approaches.

Table 5.3 shows the recognition accuracies of the different methods for each of the five users in BCI3IVa and figure 5.2 summarizes the results averaged over the five users.

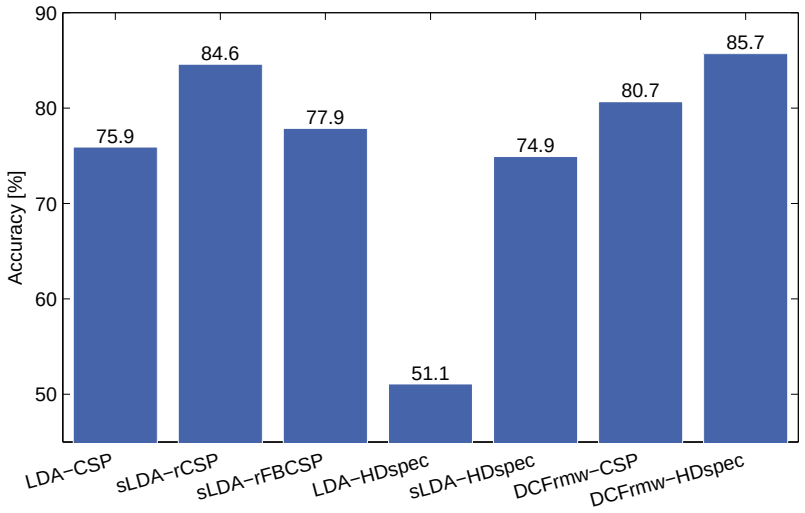


Figure 5.2 – Recognition results for BCI3IVa averaged over the five users.

	DISCRIMINATIVE		COMPACT	
	classification	features	classification	spatial filtering
LDA-CSP	LDA classification	8 CSP based features	-	-
sLDA-rCSP	LDA classification	8 CSP based features	Shrinkage	Diagonal loading
sLDA-rFBCSP	LDA classification	Automatic frequency band selection, $7 \cdot 8 = 56$ CSP based features	Shrinkage, Mutual information based selection of 22 features	Individual filter calculations with diagonal loading
LDA-HDspec	LDA classification	3304 dimensional, generic spectral features	-	n/a
sLDA-HDspec	LDA	3304 dimensional, generic spectral features	Shrinkage classification	n/a
DCFrmw-rCSP	Least-squares classification	ℓ_1-norm regularized 3304 dimensional generic spectral features	optimization with on avg. 10 active model weights	Diagonal loading
DCFrmw-HDspec	Least-squares classification	ℓ_1-norm regularized 3304 dimensional generic spectral features	optimization with on avg. 540 active model weights	n/a

Table 5.2 – Overview over the different features extraction and classification approaches for BCI3IVa. "-" stands for no particular method applied, and "n/a" denotes compact modeling techniques are not applicable for the extraction of HDspec features.

Method \ User	AA	AL	AV	AW	AY	Mean (std)
LDA-CSP	74.1	**100**	52.6	61.6	**91.3**	75.9 (19.8)
sLDA-rCSP	**78.6**	100	69.9	**95.5**	79	84.6 (12.6)
sLDA-rFBCSP	80.4	**100**	62.8	80.4	65.9	77.9 (14.77)
LDA-HDspec	50	57.1	46.4	48.2	53.6	51.1 (4.3)
sLDA-HDspec	74.1	**100**	69.9	79	51.6	74.9 (17.4)
DCFrmw-rCSP	75.0	**100**	70.4	67.4	90.5	80.7 (14.0)
DCFrmw-HDspec	**78.6**	**100**	**74.5**	93.3	82.1	**85.7 (10.6)**

Table 5.3 – Comparison of different pattern recognition methods for the five users of BCI3IVa. The results are shown in percent accuracy.

As expected, the results for all of the CSP-based approaches show that they can effectively discriminate between the two motor imagery classes. A more detailed investigation shows that LDA-CSP suffers from the small amount of calibration data as it does not implement appropriate mechanisms for the Compact objective. In contrast, rLDA-sCSP, the corresponding approach that particularly implements the Compact objective, outperforms the not regularized approach by 12.7% relative (8.7% absolute). Regularized FBCSPs were calculated with 2 to 56 features[3] selected by the mutual information feature selection. The best performance was achieved with 22 features. Surprisingly, the rFBCSP based approach did only perform slightly better than the vanilla CSP in this evaluation, although regularized CSPs were used and the approach automatically identifies relevant frequency bands, however the feature selection may overfit in the presence of little data.

The devastating effect of neglecting the Compact objective can be seen in the approach LDA-HDspec, where recognition accuracies drop to 51.1% which corresponds to chance level performance for each user. This can be attributed to the fact that the (not regularized) LDA overfits when it is applied to the 3304-dimensional features.

One can clearly see that the regularized approaches using the high dimensional spectral features (sLDA-HDspec, DCFrmw-HDspec) have competitive performance to the CSP-based approaches. This is particularly interesting as a major body of EEG-based motor imagery BCI literature uses CSP-based features (section 2.2.3).

The *DCR Framework* (DCFrmw-HDspec) shows superior perfor-

[3] 7 frequency bands with 8 filters each

mance in comparison to all other methods in this evaluation. Additionally, it outperforms[4] numerous other published approaches that have been evaluated using BCI3IVa, such all results reported in [Lotte and Guan, 2011, Arvaneh et al., 2011, Samek et al., 2012, Zhang et al., 2013, Santana et al., 2014, Brandl et al., 2015] and many others. The results achieved in this evaluation would correspond to the third best results in BCI competition III [Blankertz et al., 2008a]. Thereby, one should keep in mind that the submissions for the BCI competition can be regarded as highly optimized for BCI3IVa. For example, the competition winner used different feature sets for the different users in the data set, whereas we just applied the basic *DCR Framework* here and did not perform any specific optimizations for the data set or individual users and did not use a system combination to increase the performance.

5.1.2 fNIRS n-back Classification

After the evaluation of the *DCR Framework* using motor imagery EEG data, we investigate the classification of fNIRS data, which have fundamentally different signal characteristics than EEG signals. For this evaluation we classify different levels of workload induced by the n-back task [Kirchner, 1958], as described below.

Related Approaches

Workload has been studied by multiple BCI researchers using the n-back task. In these studies, predominantly EEG was used to measure brain activity (e.g. [Berka et al., 2007, Heger et al., 2010c, Brouwer et al., 2012]). Workload induced by the n-back task has been investigated with fNIRS measures [Ayaz et al., 2007, Ayaz et al., 2012] but their analyses are limited to properties of averaged hemodynamic responses and they did not perform recognition.
fNIRS BCIs are an emerging field of research but single-trial studies are still rare compared to EEG-based BCIs. In [Herff et al., 2013a], we showed for the first time that fNIRS can be used to discriminate between different

[4]It should be noted that a rigorous comparison of the results with those published in other papers is often not possible and not common in BCI literature, as the signal processing is usually not exactly identical, which can have a strong impact on the performance. Nonetheless, the *DCR Framework* achieves a very good performance for BCI3IVa that we expect to outperform currently published state-of-the-art results.

workload levels of the n-back task by single-trial recognition ($n \in \{1, 2, 3\}$). We investigated signals measured from the prefrontal cortex to quantify the users' workload using binary classification tasks and the three class problem of discriminating between 1-back, 2-back, and 3-back. The hemodynamic responses are consistent enough to classify all of these tasks on a single-trial basis with classification accuracies significantly above chance level.
For single-trial recognition of fNIRS signals, there is currently no standard feature extraction method. Commonly, features are calculated from the oxygenated (HbO) and de-oxygenated (HbR) hemoglobin concentration changes to represent informative properties of cerebral hemodynamics. Typically, simple statistical properties of the time-domain signal amplitudes, such as mean, variance, kurtosis, skewness or laterality have been calculated as features [Tai and Chau, 2009, Moghimi et al., 2012, Herff et al., 2013a, Heger et al., 2014a]. An effective feature extraction method for fNIRS single trial analysis is the slope calculated by fitting a line to the measured HbO and HbR signals of a hemodynamic response. If the slopes of all channels (HbO and HbR) are concatenated into feature vectors, the number of features in a feature vector corresponds to twice the number of measurement locations, which is a strong knowledge-based reduction of the recorded signal data of each trial. In [Heger et al., 2014b], we showed that high-dimensional raw time-domain amplitude signals (HDfeat) can improve fNIRS recognition results when combined with regularized least-squares classification (COMPACT).
Recently, Bauernfeind et al. [Bauernfeind et al., 2014] compared different classifiers for fNIRS BCIs. Their evaluation included classifiers that are well established in BCI research, such as Support Vector Machines and different Linear and Quadratic Discriminant Analysis based classifiers. They recommended using shrinkage LDA (sLDA) because of its simplicity, small computational costs and good recognition performance.

Data Corpus: fNIRS n-back (NBACK)

In this evaluation, we use the fNIRS n-back data corpus (NBACK) that has been made available by the Cognitive Systems Lab [Herff et al., 2013a].

The data corpus consists of recordings of 10 users (4 female, 6 male) with a mean age of 22 years. None of the participants had taken part in an n-back study before to ensure that no training effects are present. The data were recorded by an Oxymon Mk III continuous wave fNRIS system (Artinis medical systems, Netherlands) using a sampling rate of 25 Hz. Optodes were

located at a distance of 3.5 cm measuring activity of the prefrontal cortex as shown in figure 5.3. Using this montage, the data set consists of 8 channels with oxygenated (HbO) and deoxygenated (HbR) hemoglobin concentration signals (16 time series).

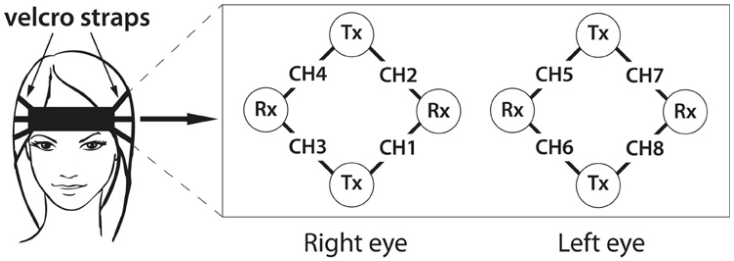


Figure 5.3 – Optode placement in NBACK. Transmitter optodes are marked as Tx, while Rx indicates receiver optode positions. Figure from [Herff et al., 2013a]

The n-back task requires the user to continuously memorize the last n of a series of rapidly presented letters (stimuli). If a stimulus appears that has been presented exactly n stimuli before, the user should react by a button press. Increasing n increases the memory effort and the task difficulty for the user. During the experiment n was varied between 1, 2 and 3, which corresponds to an easy, a demanding and a very challenging task, respectively.

A trial consisted of 5 seconds for instruction, informing the user about the next task (1-, 2- or 3-back in pseudorandomized order), followed by 22 stimuli presented on the screen every 2 seconds for 0.5 seconds (followed by a blank screen) within each trial. Subsequently, a cross was displayed for 15 seconds to ensure that hemoglobin levels returned to baseline. Signals of the first 20 seconds after the instructions of each trial were used in this evaluation.

Analog to section 5.1.1, NBACK has been recorded using a strongly controlled experimental setup and only small amounts of artifacts and non-stationarities are contained in the data. Therefore, we do not model robustness directions here and concentrate on the objectives DISCRIMINATIVE and COMPACT. The joint optimization of all three objectives will be evaluated in sections 5.3 and 5.4.

Evaluation of NBACK

In this evaluation, we analyze the three binary problems of discriminating between the three n-back conditions, i.e. $n \in \{1, 2, 3\}$ using 10-fold cross-validation.

Similar to the evaluation of BCI3IVa in section 5.1.1, we compare generic high dimensional features with specialized features. More precisely, we classify high-dimensional time-domain raw signal amplitudes as described in section 4.1.3 (*HDfeat*) and specialized fNIRS features based on the slope of the hemodynamic response (*Slope*). We employ LDA based classifiers or the *DCR Framework*, with only using the DISCRIMINATIVE and COMPACT terms in equation (4.1) (*DCFrmw*).

We evaluate the following five approaches: As a basic approach we use Linear Discriminant Analysis classification of features based on the signals' slopes (*LDA-Slope*), which can still be regarded as a current state-of-the-art approach for fNIRS based BCIs (section 5.1.2). In addition to slope features, we evaluate generic high-dimensional time-domain features using Linear Discriminant Analysis classification (*LDA-HDfeat*) and shrinkage Linear Discriminant Analysis classification (*sLDA-HDfeat*), which uses the HDspec features that we have proposed in [Heger et al., 2014b]. Furthermore, we apply the *DCR Framework* with DISCRIMINATIVE and COMPACT only, to evaluate slope features (*DCFram-Slope*), and using generic high-dimensional time-domain features (*DCFram-HDfeat*).

Results and Discussion

Table 5.4 lists the evaluated feature extraction and classification approaches for comparison and summarizes how the objectives DISCRIMINATIVE and COMPACT are implemented. It also shows the number of features extracted (DISCRIMINATIVE column) and number of features modeled for the COMPACT approaches.

Table 5.5 shows the average recognition accuracies of the five different methods for each of the three binary classification tasks (1-back vs. 2-back (1-2), 1-back vs. 3-back (1-3), 2-back vs. 3-back (2-3)). A star indicates classification results that are significantly above chance level (one sided, paired Wilcoxon signed rank tests on the results of the 10 users, $p < 0.05$). One can see that only the results of the *DCR Framework* are significantly above

	Discriminative		Compact	
	classification	features	classification	features
LDA-Slope	LDA classification	16 slope-based features	-	Knowledge-based compact features
LDA-HDfeat	LDA classification	8000 dim. generic time-domain features	-	n/a
sLDA-HDfeat	LDA classification	8000 dim. generic time-domain features	Shrinkage	n/a
DCFrmw-Slope	Least-squares classification	16 slope-based features	ℓ_1-norm regularized optimization with on avg. 15 active model weights	Knowledge-based compact features
DCFrmw-HDfeat	Least-squares classification	8000 dim. generic time-domain features	ℓ_1-norm regularized optimization with on avg. 1963 active model weights	n/a

Table 5.4 – Overview over the different features extraction and classification approaches to evaluate NBACK. "-" stands for no particular method applied, and "n/a" denotes compact modeling techniques are not applicable for the extraction of high-dimensional time domain features.

Approach \ Task	**1-2**	**1-3**	**2-3**	**Mean**
LDA-Slope	56.5 (14.5)	64.5* (11.2)	56.0 (12.0)	59.0
LDA-HDfeat	55.8* (7.8)	54.5* (6.1)	47.5 (10.1)	52.6
sLDA-HDfeat	53.5 (14.4)	66.5* (8.9)	**61.8*** (14.2)	60.6
DCFrmw-Slope	**60.5*** (13.3)	63.3* (10.6)	60.3* (5.6)	61.3
DCFrmw-HDfeat	59.0* (12.0)	**67.5*** (8.3)	58.8* (9.4)	**61.8**

Table 5.5 – Comparison of different pattern recognition approaches for the three different binary classification tasks of NBACK and the mean over the results of all tasks for each method. The results are shown in percent accuracy, standard deviations across users are shown in parentheses. Stars indicate results that are significantly above chance level (one-sided Wilcoxon signed rank tests, $p < 0.05$).

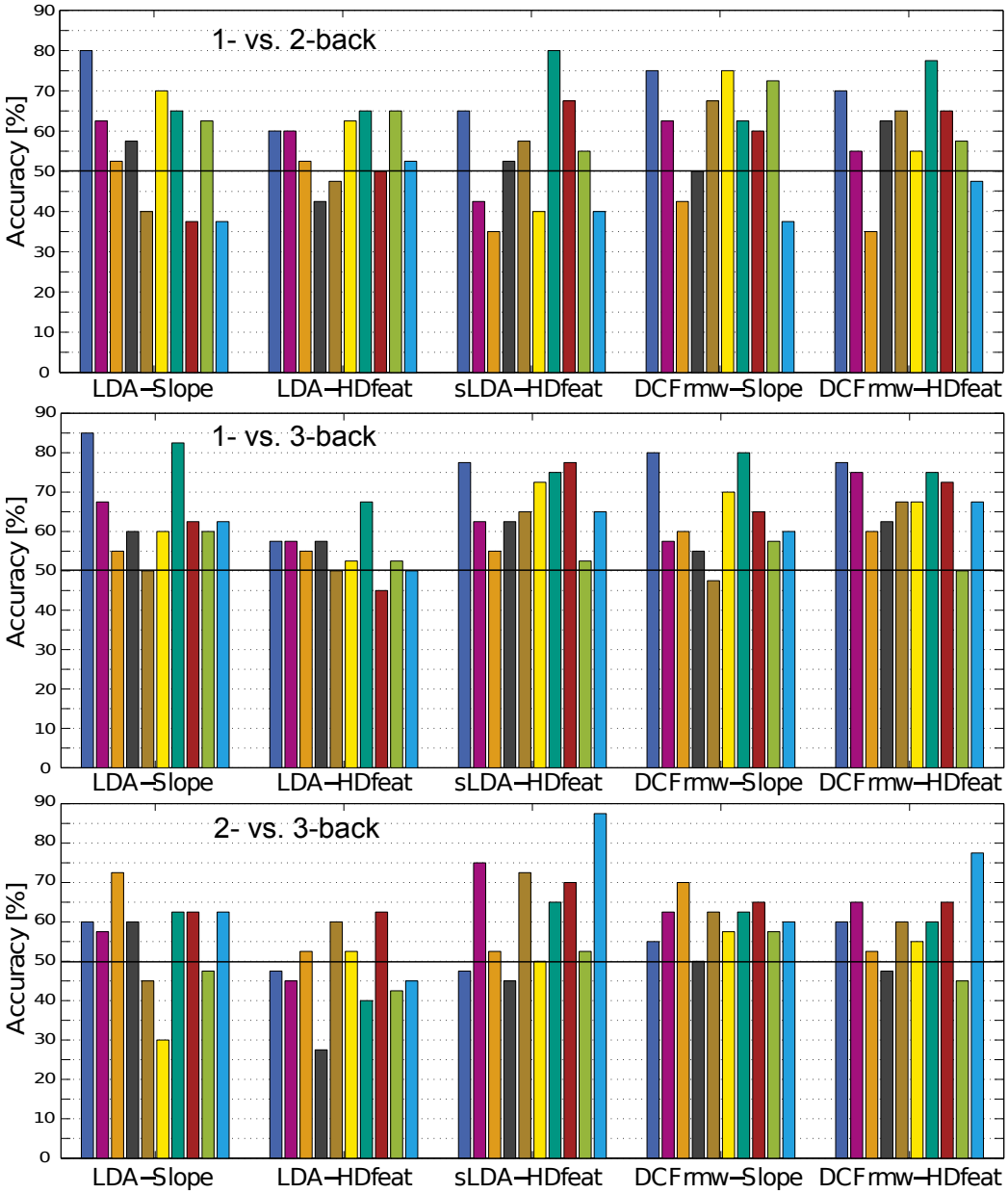


Figure 5.4 – Accuracies of 10-fold cross-validations of the three binary classification tasks for the different pattern recognition approaches. Each bar corresponds to the result of one of the 10 users.

chance level for each of the three classification tasks. Figure 5.4 shows the individual results for the 10 users for each of the three tasks.

For all approaches besides *LDA-HDfeat*, the classification of 1-3 performs best, which is expected as the largest difference in workload among the binary classification tasks can be assumed. The baseline approach *LDA-Slope* does only perform well for classifying the task 1-3 and cannot achieve results significantly above chance level for the other two tasks. Similar to the evaluation in section 5.1.1, not regularized LDA achieved the overall lowest results for classifying high-dimensional features (*LDA-HDfeat*) which can be attributed to overfitting. Using shrinkage LDA (*sLDA-HDfeat*) shows strong performance gains for the tasks 1-3 and 2-3.
The evaluations using the *DCR Framework* with only the DISCRIMINATIVE and COMPACT terms (*DCFrmw-Slope*, *DCFrmw-HDfeat*) show the best recognition accuracy averaged over all three tasks, i.e. 61.3% and 61.8%, respectively. One can clearly see that the regularized approaches using high-dimensional spectral features (*HDspec*) have competitive performance in comparison to the *Slope*-based approaches.
The *DCR Framework* (*DCFrmw-HDfeat*) shows a superior performance (averaged results) in comparison to the alternative methods in this evaluation, however, pairwise Wilcoxon tests show no significant differences between the different approaches (apart from *LDA-HDfeat*, which performs significantly worse than all other approaches).

5.2 Evaluation of ROBUST using Synthetic Data

After the evaluation of the *DCR Framework* using only the DISCRIMINATIVE and COMPACT terms, we investigate the ROBUST term in this section. The robustness directions of the *DCR Framework* are a new pattern recognition concept to implement the ROBUST objectiv, therefore, we perform isolated analyzes for ROBUST in this section. Variabilities in real brain activity signals, such as non-stationarities, are usually difficult to analyze. Using synthetic data enables to illustrate the effects of the robustness directions in a controlled setting, i.e. on data that are created with variabilities according to a certain well-defined model. We illustrate the concept of ROBUST for transfer learning (section 5.2.1) and for the reduction of non-stationarities 5.2.2). Evaluations of the *DCR Framework* using real brain activity signals are discussed in sections 5.3 and 5.4.

5.2.1 Robustness against Data Shift

Signal variabilities across users and recording sessions can cause a mean shift of the features, as discussed in section 2.1.3.

In this example we illustrate the concept of the robustness directions and how they make the *DCR Framework* robust against such variabilities. We use synthetic toy data to simulate the problem that calibration data to train the BCI is given from one person or session (training data) and the BCI should be applied to another target person or session where the data distribution has changed. From the target domain, we have few data available to calculate the robustness directions (transfer data). In contrast to real data that is high-dimensional, we generate synthetic 2-dimensional data that can easily be visualized to illustrate the behavior of the *DCR Framework*.

Synthetic Data Generation

We sampled two classes of synthetic training features from normal distributions. For the first class, features were drawn from $\mathcal{N}(\begin{bmatrix}10 & 10\end{bmatrix}^\top, I)$ and for the second class, features were drawn from $\mathcal{N}(\begin{bmatrix}20 & 10\end{bmatrix}^\top, I)$, where $\mathcal{N}(m, \Sigma)$ is the multivariate normal distribution with mean m and covariance matrix Σ and $I \in \mathbb{R}^{2\times 2}$ is the unity matrix. The transfer data were sampled from a normal distribution with shifted mean, i.e. from $\mathcal{N}(\begin{bmatrix}20 & 20\end{bmatrix}^\top, I)$ for class one and from $\mathcal{N}(\begin{bmatrix}30 & 20\end{bmatrix}^\top, I)$ for class two.

Evaluation

The robustness directions were chosen as the differences between the training means $\mu_k^{\mathfrak{s}}$ and the transfer means $\mu_k^{\mathfrak{t}}$ for each class k, as described in section 4.1.4:

$$\mathcal{D} = \left\{ \mu_k^{\mathfrak{s}} - \mu_k^{\mathfrak{t}} \mid k \in \mathcal{C} \right\},$$

where $\mathcal{C} = \{1, 2\}$ is the set of class indices.

Figure 5.5 shows the synthetic features and the different separating hyperplanes[5] learned by the *DCR Framework* when increasing the weight for the ROBUST-term ν linearly between 0 and 1.
One can see that the separating hyperplane increasingly turns towards the

[5]The model x in equation (4.1) corresponds to the normal vector of the separating hyperplane.

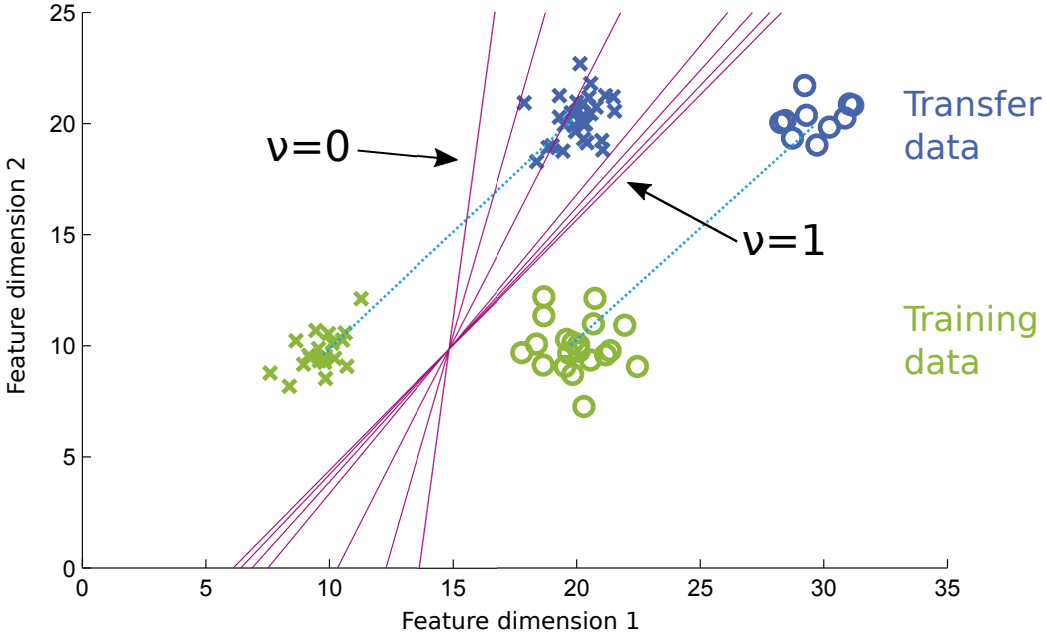


Figure 5.5 – Synthetic features and the different separating hyperplanes (red) when increasing the weight for the ROBUST-term ν linearly. Crosses indicate features of class 1 and circles indicate features of class 2. Training data are shown in green, transfer data in blue. The lightblue dotted lines indicate the transfer directions.

direction parallel to the defined robustness directions (vectors between the training and transfer data). Setting $\nu = 1$ in this example results in a separating hyperplane that is parallel to the defined robustness directions, i.e. the sum-of-norms in the ROBUST-term in equation (4.1) becomes zero. Therefore, setting higher values for ν does not change the separating hyperplane. The recognition accuracies of the transfer data using the classifiers trained with different values for ν gradually increases from 0% to 100% in this example. This means that without the ROBUST-term shifted test data that are distributed as the transfer data cannot be classified correctly at all, while setting the appropriate weight for the robustness direction makes the learned model invariant towards the data shift, which results in a perfect classification.

In this evaluation we generated 2-dimensional random data to be able to visualize the data and illustrate the influence of the robustness directions. Note that the concept of robustness directions in the *DCR Framework* can be more powerful in high-dimensional feature spaces, where a separating

hyperplane can be parallel to multiple directions in the feature space, as outlined in the last paragraph of section 4.1.2.

5.2.2 Robustness against Non-Stationarities

Signal variabilities in brain activity signals occur often within a recording session due to the non-stationary nature of activity sources that contribute to the measured brain activity signals, as discussed in section 2.1.3.

In this evaluation we illustrate the robustness against signal variabilities caused by non-stationarities using synthetic toy data. We simulate the problem that the calibration and test data are subject to changing distributions over time (non-stationarities) and show that the *DCR Framework* can improve the classification of unseen test data due to its ROBUST-term.

The synthetic data is generated according to a simplified model of the EEG. This way, we can analyze the behavior of the *DCR Framework* under controlled conditions and show that the *DCR Framework* can improve the robustness against the non-stationarities that our model generates.

Synthetic Data Generation

The idea behind the data generation is that the measured signals are a linear mixture of multiple activity sources. These sources can either be stationary and encode the actual information that should be recognized, or can be non-stationary and non-informative, i.e. they contain random variabilities that change over time. This corresponds to a simplified model of the physiology underlying the EEG, which is a mixture of informative and non-informative cortical activity sources that are mixed due to volume conduction effects (section 2.1.2).

We generated a data set that consists of 400 trials, of which the first 100 trials were used for training and the remaining 300 for testing. Each trial consists of 500 samples.

We generated three stationary sources $X_t^{\mathfrak{s}} \in \mathbb{R}^{3\times 500}$ and 29 non-stationary sources $X_t^{\mathfrak{n}} \in \mathbb{R}^{29\times 500}$ that were linearly mixed in each trial:

$$X_t = M \cdot \begin{bmatrix} X_t^{\mathfrak{s}} \\ X_t^{\mathfrak{n}} \end{bmatrix}, \ \forall t \in \{1, \ldots, 400\},$$

where $M \in \mathbb{R}^{32\times32}$ is the mixing matrix with entries randomly chosen from a uniform distribution between -0.5 and 0.5 and columns normalized to one. A total number of 32 sources has been used here, as this corresponds to a typical number of channels in EEG data. Thereby, using three stationary sources generates data with strong non-stationarities that are roughly comparable with EEG data and recognition rates are similar to those typical achieved in BCIs. Similar effects can be shown with different numbers of stationary and non-stationary sources.

To generate trials of two different classes, the variances of the stationary sources for each trial were randomly sampled from a normal distribution with zero mean and two different variances that were randomly switched between trials. In this evaluation, we chose $\mathcal{N}(0, 0.3)$ and $\mathcal{N}(0, 4)$ to generate different variances (the choice of values is not critical here as long as the variances are sufficiently different). For each block of 10 consecutive trials, the variances for the non-stationary sources were randomly sampled from $\mathcal{N}(0, 1)$ and the sign of the log-variance of 50% percent of randomly chosen non-stationary sources was flipped to generate strong changes of the variance that correspond to non-stationarities. For each trial, a diagonal matrix was composed from the sampled variances and mixed by M. A covariance matrix for each trial was generated by the outer product of the mixed variance matrices. The signals of each trial $X_t \in \mathbb{R}^{32\times500}$ were sampled form a multivariate normal distribution with zero mean and the generated covariance matrix.

Figure 5.6 shows the first 35 trials of the generated synthetic signals. In the first row above the 32 signal time series, the corresponding labels are shown encoded as a time series (green).

Evaluation

For each of the generated trials X_t we extracted the variance of each channel (row in X_t) and concatenated them into feature vectors. We trained the *DCR Framework* using the 100 training trials and predicted the test trials. To calculate the set of robustness directions $\mathcal{D}$, blocks of 5 consecutive trials of the same class c were extracted from the training data and the set $\mathcal{B}_c$ was calculated containing the mean feature vectors of each block. The robustness directions were set to the differences between the mean vectors μ^i of each block and the average of all block mean vectors in $\mathcal{B}_c$:

$$\mathcal{D} = \left\{ \mu^i - \frac{1}{|\mathcal{B}_c|} \sum_{\mu^j \in \mathcal{B}_c} \mu^j \;\middle|\; \mu^i \in \mathcal{B}_c,\ c \in \mathcal{C} \right\}.$$

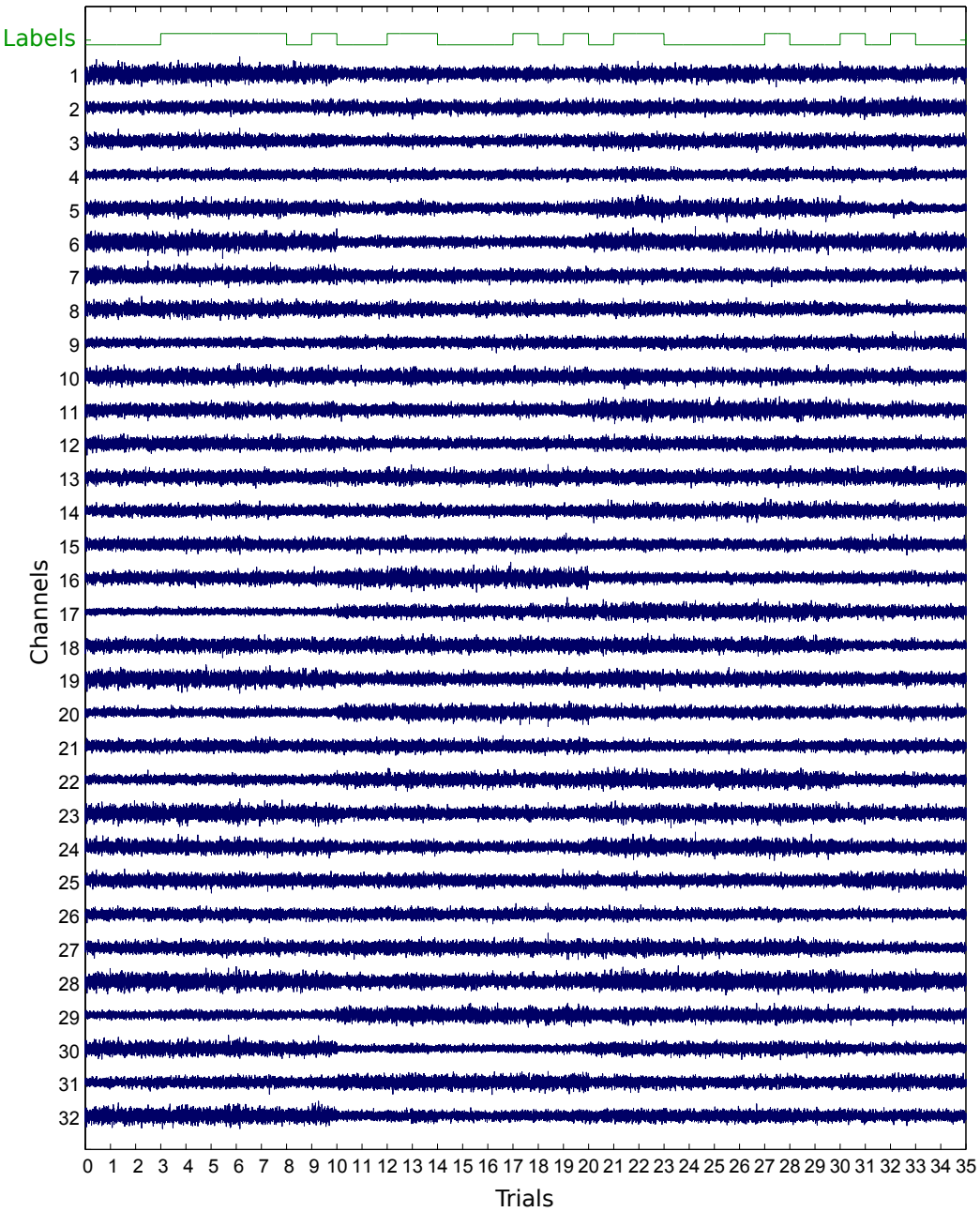


Figure 5.6 – The first 35 trials of the generated synthetic time-domain data. Figure shows 32 channels of generated data and corresponding labels encoded as time series (green).

With this setup, we evaluated the performance of classifying the test data using different values for the weight of the ROBUST-term ν. The weight for the COMPACT-term λ was chosen by hand to achieve an adequate bias-variance tradeoff, i.e. high recognition rates for both, training and test data.

Increasing the weight for the ROBUST term ν linearly from 0 to 30 steadily increases the recognition performance of the test data from 70% to 84.3%, which can be attributed to the increasing robustness against non-stationarities. Analogue to the evaluation in the previous section, setting $\nu \geq 30$ results in a separating hyperplane that is parallel to all defined robustness directions and gives identical results.

Visualizing high-dimensional data is not trivial, as information is generally lost. For example, visualizing the data and their separating hyperplane in an arbitrary two-dimensional subspace may not indicate that the classes are well separated. Therefore, we chose the basis K of a two-dimensional subspace to include the mean of the three directions of most variance determined by principal component analysis (PCA) as the first basis vector and the normal vector of the separating hyperplane calculated by the *DCR Framework* as the second basis vector. This ensures that the discriminative aspects of the separating hyperplane are visible in the two-dimensional subspace.

The projection of a point $x \in \mathbb{R}^2$ in the subspace spanned by K can be projected into the high-dimensional space by

$$Proj(x) = Kx,$$

where the two basis vectors are the column vectors of $K \in \mathbb{R}^{32\times 2}$.
The corresponding orthogonal projection of the data to the subspace can be performed by

$$Proj_K(x) = (K^\top K)^{-1} K^\top x.$$

This can, for example be seen as

$$\begin{aligned} Proj_K(Proj(x)) &= (K^\top K)^{-1} K^\top Proj(x) \\ &= (K^\top K)^{-1}(K^\top K)x \\ &= x. \end{aligned}$$

Figure 5.7 shows six plots of the test data and the separating hyperplane trained when increasing ν linearly from 0 to 30. The plots show the 32 dimensional data projected onto the 2-dimensional subspace as described above. One can see that with increasing ν, the data are better separated by the hyperplane and recognition accuracies increase from 70% to 84.3%.

This shows that non-stationarities generated by our physiologically motivated model can successfully be reduced by the *DCR Framework*, i.e. the non-stationarities fall in a subspace that is spanned by the robustness directions that describe the variabilites of the features of each class over time.

5.3 Evaluation of Motor Imagery with User Transfer

After the evaluations of the individual parts of the *DCR Framework*, i.e. DISCRIMINATIVE and COMPACT in section 5.1 and ROBUST in section 5.2, we evaluate the complete *DCR Framework* using the joint optimization of DISCRIMINATIVE, COMPACT and ROBUST in this section. We analyze the classical motor imagery BCI paradigm as in section 5.1.1, but this time in a transfer learning setting to show the full potential of the *DCR Framework*. The idea of this evaluation is to analyze the situation where there is only little calibration data available from a BCI user (cf. section 1.1.5, inconvenient setups), but a database of different users can be employed to learn user-independent effects of the recognition problem.
More introductory details and related work regarding motor imagery classification has been discussed in sections 1.1.4 and 5.1.1.

Parts of this section have been published in [Heger et al., 2015].

5.3.1 Description of the Data Corpus

For this evaluation we employed the EEG Motor Movement/Imagery Dataset that is freely available from PhysioNet[6] [Goldberger et al., 2000, Schalk et al., 2004]. The data set consists of EEG recordings of 109 different users. We selected the runs 6, 10, and 14 of the recording sessions, where users performed two classes of motor imagery: moving both fists versus moving both feet. This way, the data set consists of 45 trials per user. The data of four users has not been used in the following evaluations as their recordings contain fewer trials.

We split the data set for our evaluations into three parts (Figure 5.8): The first 50 users were used only for transfer learning, i.e. a disjoint set of users that are not used for training or testing. Each of the remaining recordings

[6] www.physionet.org/physiobank/database/eegmmidb/

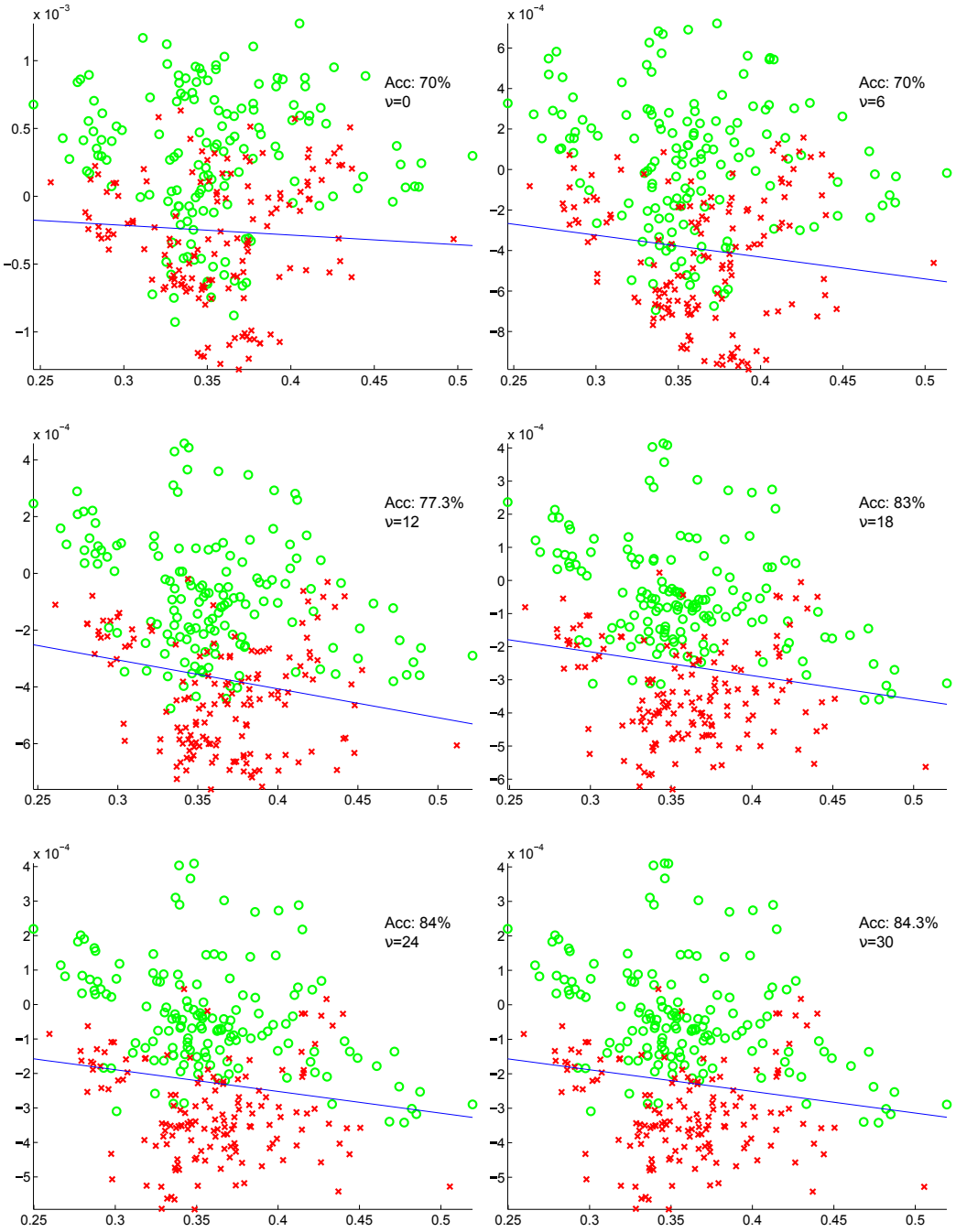


Figure 5.7 – Synthetic data and separating hyperplane of the reduction of variabilities from non-stationarities projected to a 2-dimensional subspace. The recognition accuracies for varying ν linearly between 0 and 30 increase steadily from 70% to 84.3%.

were split into the first 20 trials for training (10 trials per class) and the remaining trials were used for testing.

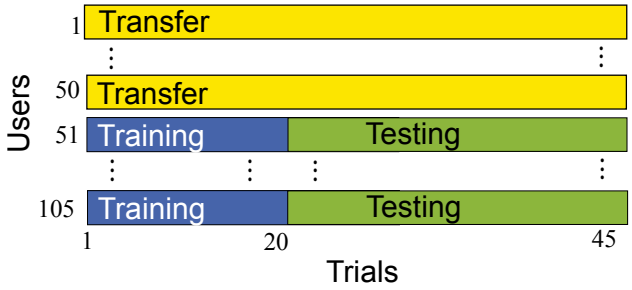


Figure 5.8 – Partitioning of the data corpus into three sets: transfer set (users 1-50), training set (first 20 trials of users 51-105) and test set (trials 21-45 of users 51-105).

5.3.2 Motor Imagery Recognition System

We used the *DCR Framework* to discriminate between the two different classes of motor imagery. Our recognition system was designed as follows:

Pre-processing: We extracted trials from the EEG signals between 0.5 and 4 seconds after each stimulus and re-referenced the data to common average reference. We removed signal offsets and linear trends from each trial and decorrelated the signals by applying a whitening transform. The whitening transform was calculated using a Ledoit-Wolf robust covariance estimator [Ledoit and Wolf, 2004] on training or transfer data and estimated transformations were applied to test data. This pre-processing was also used for all evaluated alternative pattern recognition methods (next section).

Discriminative: For each trial and each of the 64 channels, we calculated high-dimensional power spectral density features using Welch's method (HDspec). We selected the frequency bins in the range between 8-30 Hz and stacked them into a 1408-dimensional feature vector[7]. We selected this frequency range for comparability with CSP-based approaches, although wider frequency bands showed similar and even better performance with the *DCR Framework*.

Robust: To perform user transfer with our framework, the transfer directions d_k were set to the difference between the transfer mean and the training

[7]Corresponding to 64 channels and 22 frequency bins each.

mean:

$$d_k = \mu_k - \mu, \quad \forall k \in \{1, \ldots, 50\}$$

where μ_k is the mean of the features from user k in the transfer set and μ is the mean of the training features from the current user.

We trained the *DCR Framework* for each of the test users with the features of the 20 training trails (matrix A, section 4.1.1). The hyperparameters λ and ν were estimated by cross-validation on the training data (section 4.5.1).

5.3.3 Alternative Pattern Recognition Approaches

We compared the results of the *DCR Framework* (in the following called *'DCRFrmw (user transfer)'*) to the following four alternative BCI pattern recognition pipelines:

rCSP+sLDA (regularized Common Spatial Patterns with diagonal loading [Lotte and Guan, 2011] in combination with a shrinkage Linear Discriminant Analysis (LDA) classifier [Lotte et al., 2007]): This pattern recognition approach can be regarded as the standard approach for motor imagery classification when little training data is available (as discussed in section 5.1.1). The CSP filters and the classifier were trained using only the 20 training trials of each user and the transfer data was not used in this setup. Regularization weights for the CSPs were estimated by 5-fold cross-validation on the training data. The six most discriminative CSP filters were applied to the pre-processed and frequency filtered EEG signals (8-30 Hz) and logarithmic variance features were extracted. For shrinkage LDA, the analytical estimator [Ledoit and Wolf, 2004] for the shrinkage parameter was used.

divCSP-AS+sLDA (divergence CSPs[8] for across subject learning): This recently proposed state-of-the-art method has shown competitive performance for many BCI problems including user transfer [Samek et al., 2014]. Training data were used for the divCSP term and the transfer sessions were used for the divCSP-AS regularization term. Parameter settings were applied as suggested in [Samek et al., 2014], i.e. it was configured for 6 spatial filters, deflation mode, and estimation of the regularization parameter by cross-validation. Pre-processing, feature extraction and classification were performed as for *rCSP+sLDA*.

[8]The implementation that is provided by its authors at www.divergence-methods.org has been used.

HDspec+sLDA (high-dimensional frequency features classified by shrinkage LDA): HDspec features were extracted as described in the previous section and classified by shrinkage LDA. This pattern recognition approach does also not use the transfer data but only learns from the 20 training trials of each user.

DCFrmw (no transfer) (*DCR Framework* with only the DISCRIMINATIVE and ROBUST-terms): To evaluate the effects of ROBUST and to show that the *DCR Framework* effectively performs person transfer, this pipeline calculates the baseline results without user transfer, i.e. setting $\nu = 0$.

5.3.4 Evaluations and Results

	Mean (std.)	**Median**
rCSP+sLDA	66.4 (16.5)	64.0
divCSP-AS+sLDA	67.1 (13.9)	64.0
HDspec+sLDA	74.1 (15.0)	72.0
DCFrmw (no transfer)	73.9 (15.2)	72.0
DCRFrmw (user transfer)	**75.6** (15.3)	**76.0**

Table 5.6 – Recognition accuracies in percent of the *DCR Framework* with (DCR-Frmw) and without (DC-Frmw) user transfer in comparison to alternative approaches.

Table 5.6 shows the recognition accuracies for the different approaches averaged across the 55 test users and the corresponding median recognition rates. All results are significantly better than chance level performance (one-sided Wilcoxon signed rank tests $p < 10^{-6}$). The data set includes high and low performers. Therefore, the results of the individual users range from chance level to perfect classification for each approach and standard deviations are typically high (e.g. [Allison and Neuper, 2010]).

HDspec+sLDA performed significantly better than the *rCSP+sLDA* pipeline (one-sided, paired Wilcoxon signed rank test $p < 10^{-6}$). This shows that HD frequency features in combination with compact modeling can outperform state-of-the-art CSP-based methods. The mean recognition rate of *HDspec+sLDA* is even slightly better in mean than the result of our optimization framework without user transfer (*DCFrmw no transfer*), but these results are not significant ($p > 0.12$).

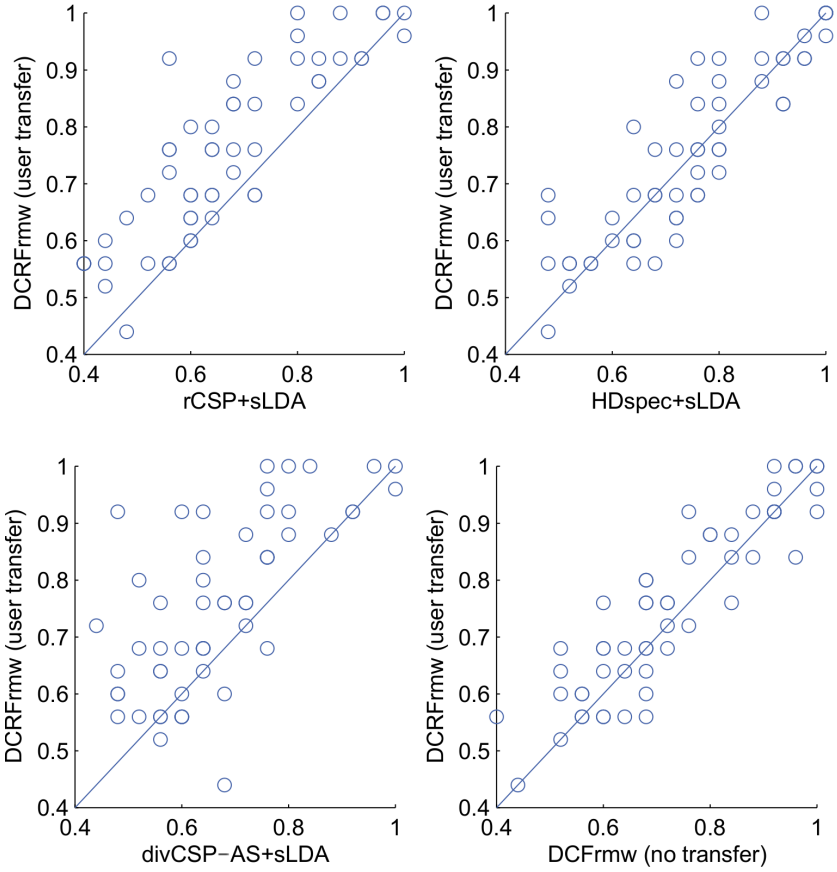


Figure 5.9 – Scatter plots of the recognition accuracies of the 55 test users for the proposed approach (DCRFrmw user transfer) in comparison with regularized CSPs classified by shrinkage LDA (rCSP+sLDA), HDspec features classified by shrinkage LDA (HDspec+sLDA), divergence Common Spatial Patterns with across user learning (divCSP-AS+sLDA), and our optimization framework without user transfer (DCRFrmw no transfer).

divCSP-AS, which has achieved a very good performance in multiple other tasks [Samek et al., 2014], shows rather weak performance in our evaluation. This can be explained by the small amount of calibration data and neglecting the Compact objective.

The *DCR Framework* achieved a successful user transfer with an increase in median performance by 4% absolutely in comparison to both *DCFrmw no transfer* and *HDspec+sLDA*. Particularly, the results of our framework with user transfer (*DCRFrmw user transfer*) are significantly better than without user transfer (*DCFrmw no transfer*) ($p < 0.02$).

Figure 5.9 summarizes the performance of *DCRFrmw user transfer* in comparison with the four alternative approaches using scatter plots. Each point corresponds to the recognition performance of one user. If a point is located above the diagonal, the proposed approach *DCRFrmw user transfer* outperforms the approach shown at the x-axis.

Figure 5.10 shows the topographical plots of the weights of the model learned by *DCRFrmw user transfer* averaged across users for the frequency bands 8-13 Hz, 14-19 Hz and 20-25 Hz (averaged across channels) and the corresponding forward model (see section 4.5.3). The most influential discriminative regions are at sensorimotor areas, fairly localized at areas corresponding to hand and feet motor imagery. One can see that the resulting models are not strongly affected by eye or muscle artifacts.
Figure 5.11 show the corresponding weights of the forward model according to their frequency distribution (averaged over users and sensorimotor channels C3 and C4). One can clearly see that the most influential frequencies are in the μ-band (around 10 Hz) and the β-band (around 22 Hz), which is in agreement with the well-known neurophysiological effects of motor imagery [Pfurtscheller and Neuper, 1997].

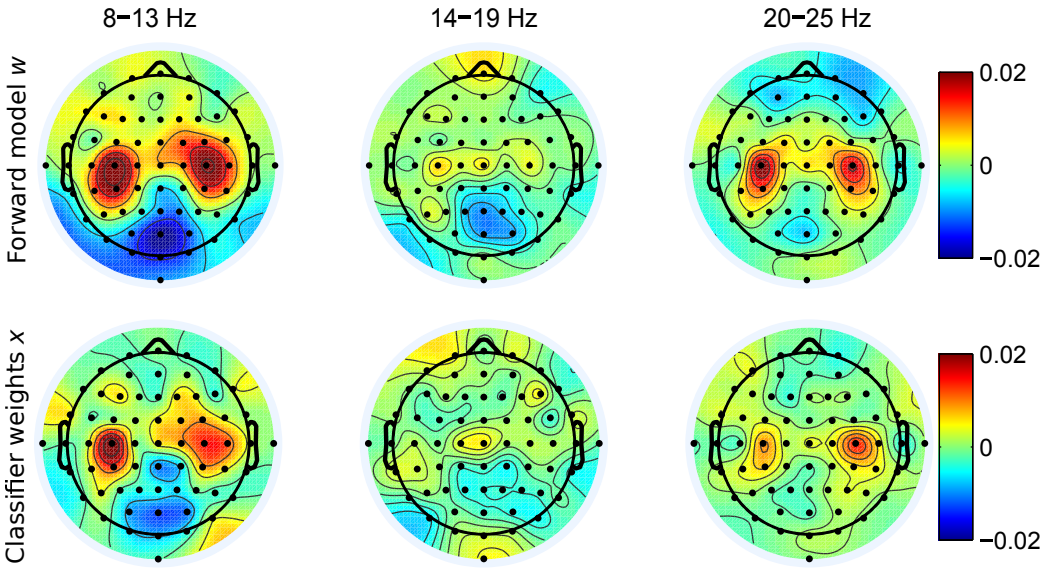


Figure 5.10 – Topographical plots of the forward models (top row) and weight vectors (bottom row) averaged across users and for the frequency bands 8-13 Hz, 14-19 Hz and 20-25 Hz.

This evaluation shows that the *DCR Framework* can outperform current BCI pattern recognition methods in a typical BCI setting. In particular, HDspec

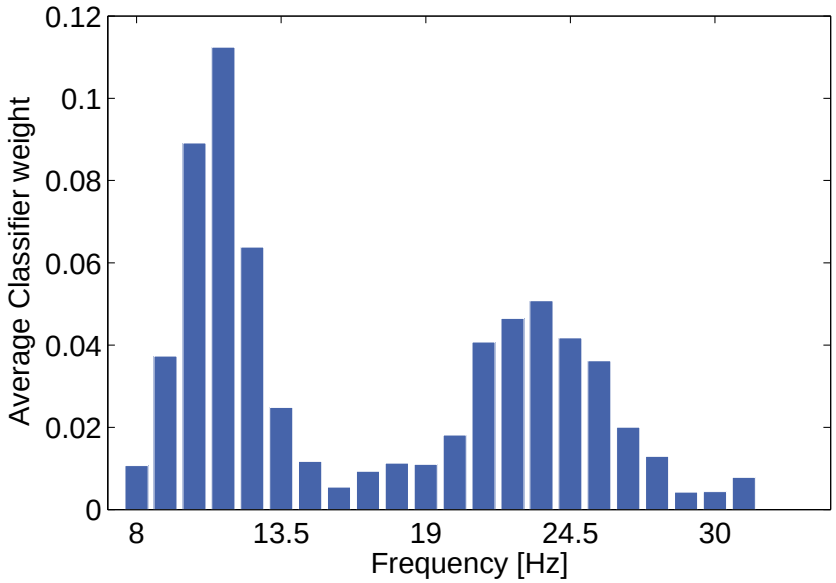


Figure 5.11 – Frequency distribution of forward model weights averaged across all users and sensorimoter areas (channels C3 and C4).

features showed superior performance to CSP-based features and the *DCR Framework* successfully performed user transfer, which lead to an increase in median performance by 4% above the best of the state-of-the-art alternative approaches.

5.4 Recognition of Error Potentials with User Transfer

In this section, we present our system for detecting error potentials in the BCI Challenge that was organized as part of the 7th International IEEE EMBS Neural Engineering Conference 2015. Our prize-winning recognition approach is based on the *DCR Framework* and uses a combination of different types of features, including time-domain, frequency-domain and meta-data features, and a post-processing step to account for different types of trials in the experiment.

5.4.1 Related Work

In the last few years, multiple BCI research groups pursued the recognition of error potentials (ErrPs). ErrPs are event-related potentials that are evoked by errors due to mistakes of the user [Falkenstein et al., 2000] or erroneous behavior of an operated system. Detecting ErrPs can be interesting for various intelligent applications, such as systems that proactively correct errors, for example by reprompting the user. A typical interaction ErrP occurs in a window of about 150 ms to 600 ms after a stimulus, with its most pronounced components being a negative peak around 250 ms and a positive peak around 350 ms [Ferrez and del R Millan, 2008]. Note that these latencies can differ from those of error-related negativity ERPs triggered without external feedback.

Schalk et al. [Schalk et al., 2000] were among the first that investigated error potentials in the context of BCIs. In their study, four users performed a motor imagery based cursor control task. The authors characterized differences in grand averages of data following successful and unsuccessful trials.
Ferrez et al. [Ferrez and del R Millan, 2008] classified ErrPs from EEG data recorded during the operation of a simulated BCI for spatial control with a predefined error rate of 20%. Using temporal features they achieve classification accuracies of up to 82% and were able to maintain this accuracy for multiple sessions of the same user recorded at different days.
Spüler et al. [Spüler et al., 2012] used the detection of error-related potentials for online adaptation of the classifier in a code-modulated visual evoked potentials BCI. With this system they achieved an average information transfer rate of 144 bit/min, which was the highest bitrate reported so far for a non-invasive BCI.
In [Putze et al., 2013] we showed that classification accuracies for ErrP recognition can be improved using user-adapted classifiers which are trained using selected data from other users in addition to the user-specific calibration data.
In [Putze et al., 2015] we integrated a self-correction using ErrP recognition in an online gesture interface, which significantly improved its recognition accuracy. The ErrP recognition provided lower costs and higher user acceptance than a manual correction.
Margaux et al. [Margaux et al., 2012] used ErrP recognition during the online operation of an P300 speller BCI to compensate for recognition errors if an ErrP was detected using the second best guess. They found that this automatic correction yielded a higher bit rate than a respelling strategy. The dataset they used is the basis for the evaluation below.

5.4.2 Description of the BCI Challenge @NER15 and its Data Corpus

The objective of the BCI Challenge at IEEE EMBS Neural Engineering Conference 2015 was to recognize error potentials in response to feedback events that occurred during usage of a P300-Speller [Donchin and Coles, 1988]. The data provided for the competition consist of 56 channel EEG recordings and one EOG channel sampled at 100 Hz. A training data sets (16 users) was provided together with the ground-truth labels and a testing data set without ground-truth labels (10 users). The ground-truth labels contain the binary information whether a trial contains a spelling error or not, i.e. whether an ErrP can be expected in the data or not. For the competition only the signal data, timings of the feedback events and ground-truth labels for the training data were provided, i.e. no information of the P300 spelling or the spelled words were available. As a particular challenge, the group of users in training was disjoint from the group of users in the test set and thus demanding user-transfer for the classification of test data. For each user, 5 sessions with a total number of 240 trials were provided. In the data set there are two different kinds of trials: a slow mode of the P300 speller, which is a less error-prone condition (each letter was flashed 8 times) and a fast mode, which is a more error-prone condition (each letter was flashed 4 times). After the last flash of a trial, the recognized letter was displayed in the middle of the screen in large font (feedback event). Figure[9] 5.12 shows the interface of the P300-speller used in the experiment during a feedback event. More details on the data set can be found in [Margaux et al., 2012].

The competition lasted for 97 days and had more than 260 competing teams. During the competition, three submissions per day could be uploaded to the competition platform at kaggle.com. The performance of the submission was immediately evaluated on 20% of the test data (2 test users) and a preliminary ranking (called public leaderboard) of all competing teams was updated and provided online. The performance criterion was the area under the receiver operating characteristic curve (AUC). AUC is a reasonable criterion for ErrP recognition for applications that tradeoff true-positive rate and false-positive rate (i.e. sensitivity and specificity).
After the competition ended, the performance was evaluated on the complete test data and the final ranking was calculated (called private leaderboard).

[9]The image was extracted from the video at the competition website https://www.kaggle.com/c/inria-bci-challenge

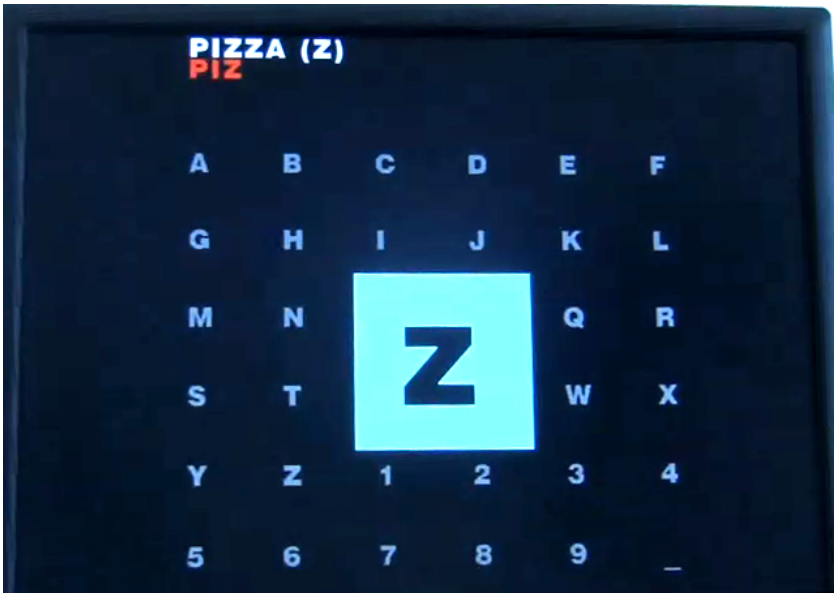


Figure 5.12 – Interface of the P300-speller used in the experiment during a feedback event.

5.4.3 Error Potentials Recognition System

We trained the *DCR Framework* to discriminate between trials that contain ErrPs and trials that do not contain ErrPs. Our recognition system was designed as follows:

Pre-processing: We pre-processed the provided EEG signals (without the EOG channel) using bandpass filters (1-30 Hz) to remove trends and high frequency noise. Then we down-sampled the data by a factor of 10 and applied a whitening transform to decorrelate the signals.

Discriminative: We combined time-domain amplitudes (0-800ms) of each channel, power spectral density features (i.e. HDspec features with 1 Hz wide bins, Welch's method) of each channel, and meta-data features (session numbers and delay to previous trial) into a 1570 dimensional feature vector and normalized the features to unit power.
We intentionally did not include the data leakage information in our system[10]. For recognition we used the least-squares regression that is part of the *DCR Framework* to predict a real-valued output score that indicates whether a

[10] A data leakage has been detected during the competition that allowed to infer information about the labels of the 5th session of each user. Using this information can strongly increase recognition performance, however this is not available in realistic conditions.

trial contains an ErrP or not. Target values for the regression were chosen as described in section 4.1.2.

COMPACT: The ℓ_1-norm regularization that is part of the *DCR Framework* was used to control for model complexity.

ROBUST: Unsupervised transfer learning was used by the sum-of-norms regularization of the *DCR Framework*. The transfer directions d_k have been chosen as the differences between the mean feature vectors μ_k (all sessions of the user combined) of each of the 16 training users Tr and each of the 10 testing users Te:

$$d_k = d_{ij} = \mu_i - \mu_j, \quad \forall k \in Tr \times Te,$$

where $\times$ denotes the Cartesian product and $Tr = \{s_{tr1}, \ldots, s_{tr16}\}$, and $Te = \{s_{te1}, \ldots, s_{te10}\}$ are the sets of training and test users. This way, 160 transfer direction vectors were calculated.

Priorshift: In a post-processing step, we adjusted the predictions of sessions 1-4 for each user to integrate prior knowledge about the different types of copy spelling conditions (fast and slow mode trials). We identified long trials by the time to the previous feedback event and added a small constant to the prediction output. The fifth session was not corrected as it contains only short mode trials.

5.4.4 Evaluation and Results

We used 4-fold cross-validations on the training data with splitting on user bounds to estimate the parameters, such as regularization weights, priorshift and filter characteristics. This evaluation scheme takes the user transfer characteristics of the data set into account and avoids biases to the preliminary ranking (public leaderboard).

We evaluated our system using leave-one-person-out cross-validations on the training data. Figure 5.13 shows the recognition results of the person-wise cross-validation of (1) the proposed system without the ROBUST-term (DC-Frmw) and without priorshift correction, (2) the proposed system without the ROBUST-term but with the priorshift correction, i.e. post-processing to account for the different prior probabilities for ErrPs in the different spelling conditions (DC-Frmw priorshift), (3) the proposed system with the complete *DCR Framework* (with DISCRIMINATIVE, COMPACT, and ROBUST terms)

without priorshift (DCR-Frmw) and (4) the *DCR Framework* with priorshift (DCR-Frmw priorshift).

Our final system (DCR-Frmw priorshift) could achieve an area under the ROC curve (AUC) of 0.761 in this evaluation. Figure 5.14 shows the results (DCR-Frmw priorshift) individually for the 16 users.

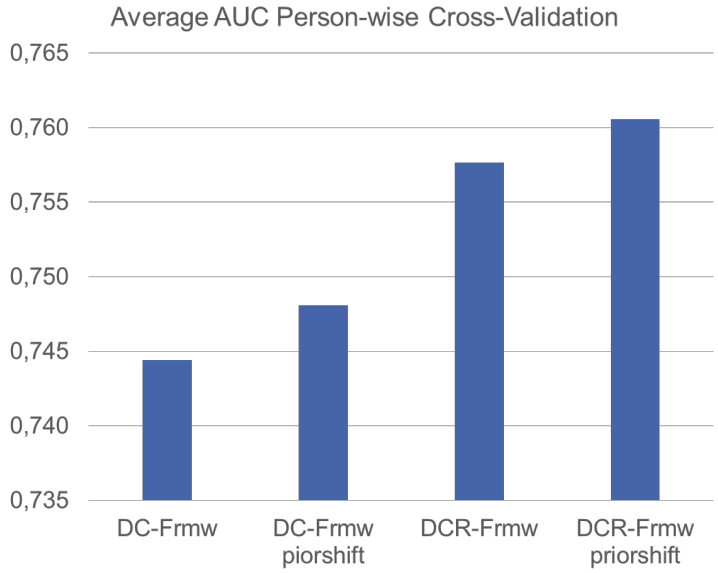


Figure 5.13 – Person-wise cross-validation results of the proposed system without and with ROBUST-term (DC-Frmw, DCR-Frmw) and without and with priorshift correction.

In the preliminary ranking on a subset of the test data (public leaderboard), our system achieved an AUC of 0.81124, whereby this can be regarded a biased estimate of the final evaluation results on all data, as the score is only based on two test users (20% of test data) and strong inter-individual performance differences can be expected (also indicated by figure 5.14).

The described system achieved an area under the ROC curve (AUC) of 0.7457 in the final competition ranking on the full data set at kaggle.com (private leaderboard score). This corresponds to the 6th best performance of 260 teams (top 3%) and was awarded the *second prize winner* at the 7th IEEE EMBS Conference on Neural Engineering. To be eligible for a prize it was required to present the system used for the submission at the IEEE Neural Engineering Conference.

Figure 5.15 shows the top 10 of the final competition ranking of 260 teams. The details about most of the competing systems have not officially been

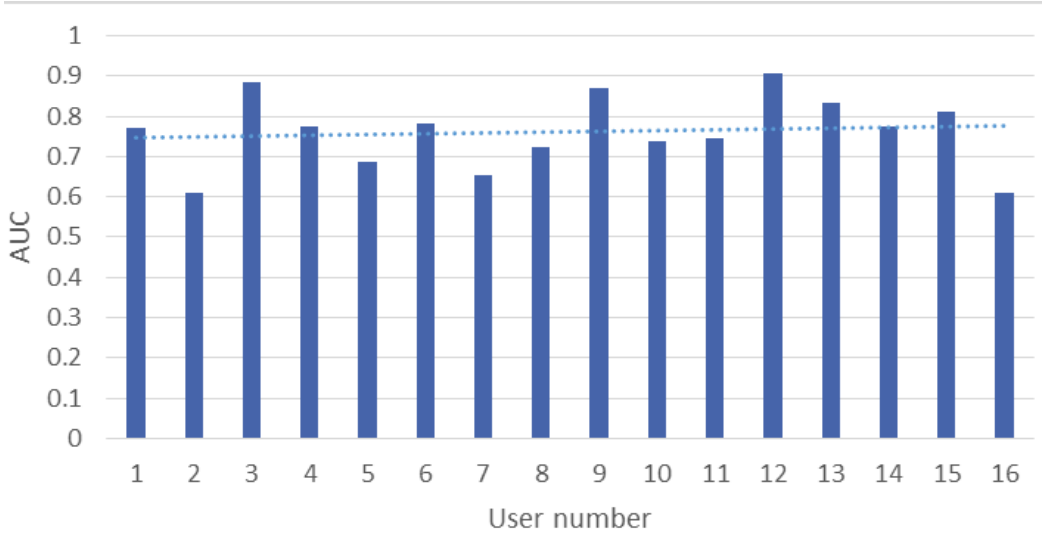


Figure 5.14 – Results of DCR-Frmw priorshift for each user using person-wise cross-validation on the training data. Dotted horizontal line shows average AUC accorss all users.

published. It should be noted that the there had been a data leakage that made it possible to infer information about the true labels of the 5th session of each subject. The top four competitors have stated in the competition forum that they exploited the leaked data to improve their system. We intentionally omitted this information as it is not available in realistic BCI settings. Discussions in the competition forum further indicate that other teams could successfully improve their recognition results by extracting features from multiple time lags, furthermore, multi-stage classification and ensemble learning methods may help to improve the recognition results.

The BCI Challenge @NER 2015 was an international competition at kaggle.com that attracted both, members of the BCI community and pattern recognition experts. Our successful submission validates the state-of-the-art performance of the *DCR Framework*.

#	Δrank	Team Name * in the money	Score	Entries	Last Submission UTC (Best - Last Submission)
1	–	the overfitting avengers *	0.87224	32	Tue, 24 Feb 2015 19:20:22
2	–	phalaris *	0.85669	44	Tue, 24 Feb 2015 10:12:06
3	↑17	H2O.ai *	0.81850	30	Tue, 24 Feb 2015 23:50:05 (-0.2h)
4	↑2	barrack_d(NER)	0.76921	110	Tue, 24 Feb 2015 21:32:58 (-4.6h)
5	↑53	Jose M.	0.74790	3	Sun, 07 Dec 2014 16:26:59 (-9.3d)
6	↑8	CSL (NER)	0.74570	27	Tue, 24 Feb 2015 16:42:43
7	↑46	Daniel Yoo	0.73175	6	Tue, 24 Feb 2015 21:58:30 (-0.2h)
8	↓1	Vivien	0.72325	38	Mon, 16 Feb 2015 06:55:54 (-21.7h)
9	↑3	khyh	0.72304	56	Mon, 09 Feb 2015 13:42:03 (-5d)
10	↓1	clustifier	0.71651	70	Tue, 24 Feb 2015 17:03:56 (-3h)

Figure 5.15 – Final ranking of the BCI Challenge @NER 2015. Figure shows the top-ten results of 260 teams.

Chapter 6

Novel BCI Paradigms

This chapter introduces two recognition systems for novel BCI paradigms that apply the principles and methods discussed in this dissertation. First, we evaluate the EEG-based workload recognition during the interaction with an adaptive interaction system. The second evaluation is the recognition of vowels during continuous speech from brain activity signals invasively measured by ECoG.

The objectives Discriminative, Compact, and Robust are, in particular, relevant for the development of new BCIs that employ novel BCI paradigms. Such BCI paradigms may not be particularly designed to modulate brain activity patterns following well-known neurophysiological effects that can easily be measured, but modulate brain activity patterns in a complex way. Therefore, little may be known about the discriminative patterns and no specialized feature extraction methods exist that can be used to recognize different classes or intensities of the BCI paradigm in single-trials.

In this chapter we analyze two systems to recognize BCI paradigms that have not been proposed in this form before. They are especially relevant as they automatically analyze spontaneously generated brain activity patterns, i.e. the systems analyze the complex patterns in naturally occurring brain activity and do not require any learning by the user to operate the BCI.

The first evaluation is a closed-loop online system for EEG-based workload recognition that we analyze in an experiment where users interact with a

simulated interaction system during single and dual task situations. The evaluations in previous chapters have been performed in typical BCI laboratory environments, i.e. the have been performed with standardized tasks and signals were recorded under strongly controlled conditions. In contrast, the experiments in this chapter are not completely controlled to simulate a real-world interaction. This way, the measured signals are generally affected by various strong artifacts. In addition, the occurring brain activity patterns consist of task specific workload related activity. In the evaluation we apply the *DCR Framework* for the recognition of low and high workload states in a self-paced fashion (asynchronous BCI) and show that it has superior performance than our previous system based on support vector machines.

The second evaluation in this chapter is based on the decoding of the cortical activity during speech, which is a novel BCI paradigm that we call Brain-to-Text [Herff et al., 2015]. In comparison to the evaluations that have been presented in the previous chapters, this evaluation uses invasive brain activity signals, i.e. subdural electrocorticography (ECoG) recordings. The brain activity patterns that underlie the neural processes of speech are still not completely understood. In this evaluation we apply the *DCR Framework* for the classification of vowels during continuous speech and show that it has superior performance than our baseline system based on Kullback-Leibler based feature selection and Gaussian classification.

This chapter provides additional evaluations to show that the *DCR Framework* can successfully be applied to the recognition of new BCI paradigms for which no specialized feature extraction methods or benchmark data sets exist. In both evaluations, we highlight the particular aspects of how Discriminative, Compact, and Robust are implemented and show that the *DCR Framework* outperforms previous systems.

6.1 EEG-based Workload during BCI adaptive Human-Machine Interaction

In contrast to human-human interaction, where the theory of mind plays a major role (e.g. [Carruthers and Smith, 1996]), machines are widely unaware of the mental states of their users. The development of user-centered technology is addressed by researchers in different disciplines (e.g. [Picard, 2000, Fong et al., 2003, Schultz et al., 2013]). In BCI research, passive BCIs [Zander and Kothe, 2011] approach this issue by analyzing the

brain activity signals of the users and recognizing their mental states (in the following referred to as "user states", cf. 1.1.4). This additional information source can be used to adapt intelligent systems to the current situation of the user for an improved human-machine interaction.

Workload is a valuable paradigm for passive BCIs as multiple applications may strongly benefit from the ability to adapt automatically to a detected workload level of their users. For example, the level of workload of air traffic control operators could be monitored and kept at an appropriate level by balancing the task load to maintain an optimal efficiency and safety. Intelligent assistants, such as robots that interact with humans could adapt their behavior according to the user's workload, for example to identify the right points of time to take the initiative for actions in collaborative work to increase productivity.

In this section we discuss a workload recognition system for the real-time adaptation of intelligent human-machine interface systems. In our evaluation we adapt the dialog behavior of a simulated humanoid robot such that it better suits high or low workload states of the user. The system is an asynchronous BCI (section 2.2) that recognizes short segments of EEG data and continuously estimates the current workload level of its user in real-time.

6.1.1 Related Work

Starting in the 1960s, there is a line of research on biophysiological measuring and modeling of workload (see e.g. [Kramer, 1990, Brookings et al., 1996, De Waard and Studiecentrum, 1996] for review). Mental workload of a person is regularly defined as the *usage of multiple mental resources that have limited capacity during the performance of a task* (e.g. [Moray, 1979, Wickens, 2008]). Workload is therefore dependent on the availability of different mental resources that are consumed by multiple different cognitive processes, such as attention, memory retrieval, planning, and many others. Therefore, workload can be seen as a rather general concept that involves very different brain processes. From a neural perspective, workload includes different brain activity patterns that are related to the activity of different resources and their interaction over time while performing a particular task. Thus, workload can be seen as a task and user specific concept that may include complex dynamics in brain activity patterns, especially when complex tasks are involved.

Multiple researchers have analyzed EEG features that correlate with workload. Oscillatory activity within multiple frequency bands have been identified as workload correlates. For example, an increase of frontal θ (4-8 Hz) activity and a decrease of parietal α (8-13 Hz) activity have been proposed as measures of workload (e.g. [Gundel and Wilson, 1992, Klimesch et al., 1993, Gevins and Smith, 2003]). However, also other frequency ranges, such as β activity and spectral power ratios, such as $\beta/(\alpha+\theta)$, have been proposed as workload measures ([Brookings et al., 1996, Berka et al., 2004, Holm et al., 2009, Walter et al., 2013]). The inconsistencies among researchers, may be caused by subjective differences and the different aspects and resources involved in the tasks that have been investigated. Therefore, many current workload recognizers use features based on spectral power within a wide frequency range, such as 4-30 Hz [Berka et al., 2004, Heger et al., 2010c, Kothe and Makeig, 2011, Mühl et al., 2014b].

Spectral power based workload recognition systems have been applied in different experiments, including standardized cognitive tasks [Gevins and Smith, 2003, Heger et al., 2010c, Brouwer et al., 2012, Walter et al., 2013], simulated and real driving [Kohlmorgen et al., 2007, Jarvis et al., 2011], and operator monitoring [Berka et al., 2005].

In [Heger et al., 2010c, Heger et al., 2010b, Heger et al., 2011a] we developed an EEG-based, self-paced, real-time workload recognition system. We introduced high-dimensional spectral features (HDspec) for workload recognition, i.e. power spectral density based features from a wide frequency band extracted from 1-2 seconds of EEG data and proposed the temporal smoothing of the classification output to generate more stable recognition results. We integrated the workload recognizer in the closed-loop interaction with an adaptive information system of a humanoid robot head. To the best of our knowledge, our study [Heger et al., 2011a] was the first evaluation of an EEG-based workload recognition based adaptive interaction system in the domain of human-robot interaction. The following evaluations are based on this study.

6.1.2 Description of the Data Corpus

During the experiment, participants had to perform two different tasks, partly by multitasking, i.e. handling two tasks at the same time. In the first task (*information task*) the participants were asked to manually fill in a paper form according to information given by speech synthesis embodied by a humanoid robot head. In a secondary task, participants performed a vari-

ant of the Eriksen flanker task [Eriksen and Schultz, 1979] (*flanker task*), in which five arrows were displayed (e.g. <<><<). Participants were expected to indicate the orientation of the middle arrow by pressing the corresponding left or right button on the keyboard (right button in the given example).

Figure 6.1 shows the experimental setup. The information system was represented by a humanoid robot head which talked to the participants using text-to-speech synthesis. The participants faced paper forms to be filled in for the *information task* as well as a desktop computer to execute the *flanker task*.

Figure 6.1 – Recording setup with the robot head and speech synthesis (left side), the laptop computer for the secondary task (center) and participant wearing an EEG cap (right side) while performing the tasks, i.e. writing down the information on paper using one hand and pushing buttons on the keyboard using the other hand.

In total 20 subjects participated in the experiment. Each participant completed five sessions that were recorded consecutively in one sitting. Each session consisted of four parts: First, 1 minute of the *information task* (single task), followed by 1 minute of performing the *information task* and the *flanker task* in parallel (dual task). After that, the single task and the dual task were performed one more time. This way, each session alternates twice between a low workload period (single task) and a high workload period (induced by dual tasking). Transitions between segments were marked by an acoustic signal. Table 6.1 summarizes the experimental design.

Single (1 min)	**Dual** (1 min)	**Single** (1 min)	**Dual** (1 min)

Table 6.1 – Experimental setup of alternating single and dual tasks that has been used for each of the five sessions of each participant.

The first session was used to train the workload recognizer. The workload recognition system was used in one of the four subsequent session that automatically adapted the speaking style of the humanoid robot according to the recognized workload (EegAdaptive). In this session, the speaking style for the *information task* was selected from two different strategies, appropriate to high and low workload of the user, as described in the next section. The other three sessions were performed using different speaking styles for the *information task* as baselines (AlwaysHigh, AlwaysLow, Oracle, see next session). To eliminate the impact of bias effects such as fatigue, the order of sessions 2-5 was randomly chosen.

Adaptive Information System

The information for the *information task* were reported to the user via text-to-speech synthesis. The information were listings of a database containing attributes of students, such as name, id, and telephone number.

There were two different speaking *styles* to present this information: The Low style designed for low mental workload, and the High style designed for high workload conditions. Although the style of presentation differed between Low and High, the content of the information stayed the same.

The Low style focused on high information throughput, i.e. only short pauses between utterances and between different database entries were made. Whenever possible, multiple information chunks were merged into one utterance and phone numbers were presented in a block-wise fashion. However, maximizing efficiency was not the only criterion but Low takes the time to convey information in complete sentences to mimic a polite communication.

The High style on the other hand was tuned towards situations in which the user has to divide his or her cognitive resources between two tasks that he or she executes in parallel (dual task). As this multitasking may cause memory capacity reduction, split attention, and limited processing capabilities, the High style accommodated the situation by presenting information in a separated fashion, giving only one attribute at a time and reporting phone numbers as single digits. Furthermore, pauses were extended between utter-

ances and database entries such that the user has more time to deal with the secondary task. Reporting time was conserved by limiting the information to the attribute name and value, thus minimizing utterance duration.

Speaking style	Low	High
Pause duration	short (500 ms)	long (2000 ms)
Number presentation	block-wise	isolated
Items per utterance	multiple	single
Formulations	polite	concise
Example utterances	The name of the next person is Heidi Kundel. Her telephone number is 52-11-66-3.	Heidi Kundel Telephone: 5-2-1-1-6-6-3

Table 6.2 – Low and High styles for information presentation.

Table 6.2 summarizes the two speaking styles. The output of the workload recognizer (next section) selected the appropriate style (i.e. High when the workload recognition corresponds to high mental workload, and Low otherwise). The speaking style could be switched seamlessly between two spoken utterances. Besides the information on the user's workload level from workload recognition system, the adaptive interaction system that controls the information presentation and speech synthesis had no information on the secondary task.

Session	Speaking style
AlwaysLow	Low
AlwaysHigh	High
EegAdaptive	Low / High according to recognized workload
Oracle	Low / High according to single / dual tasking

Table 6.3 – Different speaking styles during the four sessions.

In addition to the session EegAdaptive that adapts the speaking style to the recognized workload, sessions were recorded that use consistently one of the speaking styles for the whole session, called AlwaysLow and AlwaysHigh. Additionally, a session called Oracle switched between speaking styles according to the reference information on the secondary task, i.e. instead of relying on potentially noisy information from EEG workload recognition, it selects the suitable speaking style for each utterance according to

the contextual information of whether the secondary task is currently running or not. This session is called ORACLE and can be regarded as a gold standard that performs an optimal adaptation to the task for comparison with EEGADAPTIVE. Table 6.3 summarizes the different speaking styles for the four sessions.

6.1.3 Workload Recognition System

During the experiment, EEG data were recorded by an active EEG-cap (BrainProducts actiCap) with 16 electrodes sampled at 256 Hz using BiosignalsStudio [Heger et al., 2010a]. The impedances of each electrode were kept below 20 kΩ during all recordings.

The workload recognition system implemented the three objectives DISCRIMINATIVE, COMPACT, and ROBUST as follows:

DISCRIMINATIVE: To enable the workload recognition continuously over time (asynchronous BCI), short windows of 2 seconds length were continuously extracted from the signals with an overlap of 1.5 seconds to get a workload estimate every 0.5 seconds. Logarithmic power spectral density features (HDspec) between 4 and 30 Hz were calculated from the EEG data of all channels. Support Vector Machines (SVMs) [Chang and Lin, 2011] with linear kernels or the *DCR Framework* were employed to discriminate the two different brain activity patterns corresponding to two different levels of mental workloads, i.e. with and without secondary task.

COMPACT: The dimensionality of the HDspec features was reduced by averaging over three adjacent frequency bins. Furthermore, both the SVM and the *DCR Framework* learn models using regularized optimization. The SVM implements the regularization using slack variables to penalize missclassification of the training data and relies only on a small number of support vectors. The *DCR Framework* implements the ℓ_1-norm regularization to learn sparse models. The regularization parameters were estimated using 5-fold cross-validation on the training data for both approaches.

ROBUST: Robustness against artifacts is among the most challenging problems for training and operation of the workload recognition systems in this study. Predominantly, eye movement and muscular artifacts are present when EEG signals are recorded under not strictly controlled conditions as in this experiments. Therefore, we applied fully automatic artifact reduction methods based on the combination of two blind source separation techniques: Independent Component Analysis (ICA) and Canonical Correlation

Analysis (CCA). ICA is well-known to be very effective for artifact removal of eye blinks and saccades. The infomax algorithm [Makeig et al., 1996] is applied to the training data to calculate a transformation matrix that decomposes the 16-channel EEG signal into 16 independent components. The components related to eye movement activity were identified by their frequency and power characteristics. During the operation of the workload recognition system the EEG signals are transformed using the precalculated transformation matrix. For muscular artifacts, we applied a blind source separation based on canonical correlation analysis that has shown to be more effective than low pass filtering or ICA-based methods [De Clercq et al., 2006]. It leverages the fact that muscle activity has, in general, a lower autocorrelation than brain activity. After decomposition, the components with an autocorrelation below a certain threshold are set to zero. Thereafter, the signals are recomposed into cleaned EEG signals by back transformation into the original signal space.

In addition to artifact reduction techniques, the recognition results were integrated by averaging over the past s recognition outputs (linear temporal smoothing). This temporal smoothing procedure increases the robustness of the recognition results (reduces variabilities over time) with the cost of a reduction of temporal resolution. It was applied to the binary classification outputs of the SVM or the real-valued regression outputs of the *DCR Framework* to generate a task specific workload estimate over time.

The calculated workload estimates were thresholded to control switching between the two different speaking styles. To determine a subject specific threshold from recognition results of the training session we calculate the average workload estimation for the training parts without the secondary task (w_1) and the average workload with both tasks (w_2). The subject specific threshold t is calculated by $t = \frac{w_1+w_2}{2}$.

The described workload recognition system features a speedometer to visualize the recognized workload (see Figure 6.2). However, we did not reveal this information to the subjects during the experiments to avoid distractions and influences by self-regulation.

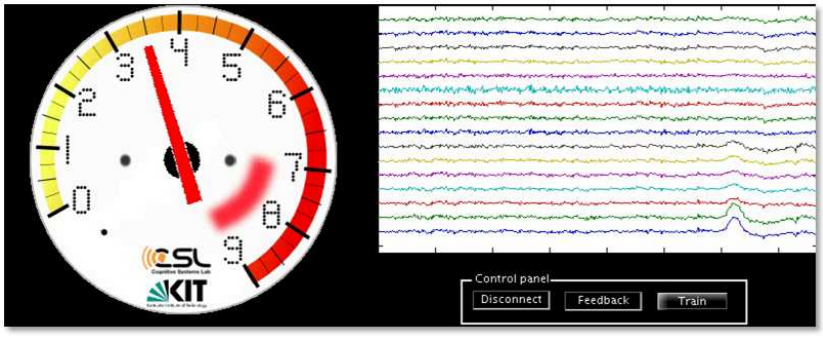


Figure 6.2 – Speedometer visualizes the recognized workload estimates in real-time.

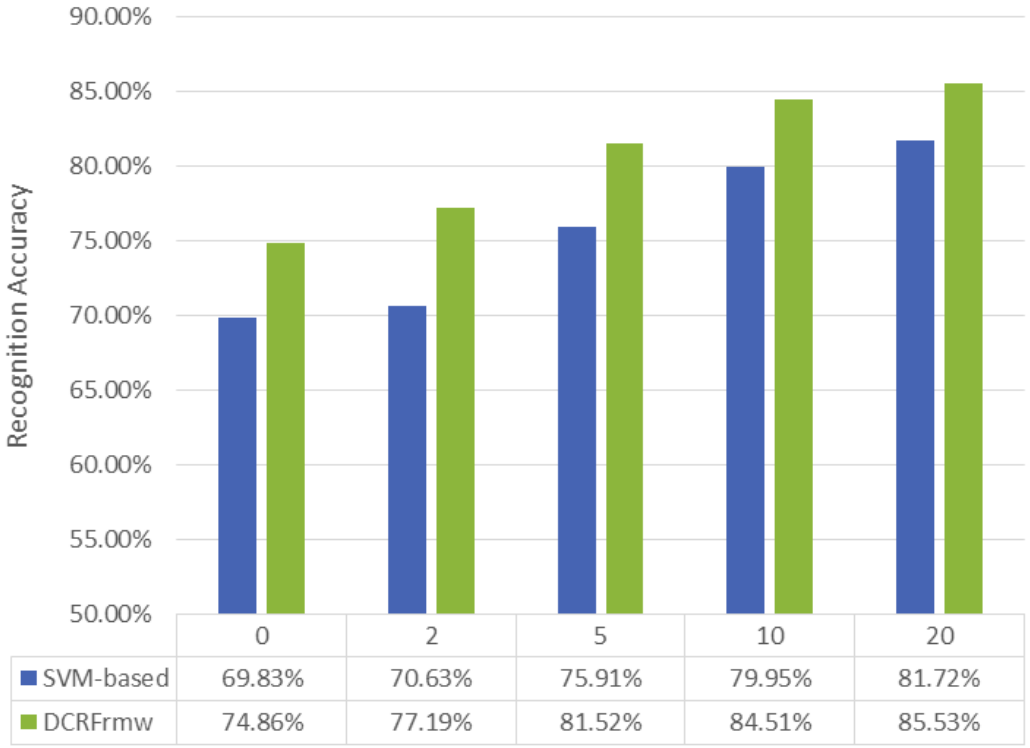
6.1.4 Evaluations and Results

Workload Recognition Performance

During the online experiment the workload recognition system based on SVM classification was used as described in the previous section. In the following evaluations we compare the results of the SVM-based system with the *DCR Framework* in an offline analysis of recognizing workload during the EEGADAPTIVE session.

We evaluated the recognition of high and low workload periods during the workload adaptive session of the experiment (EEGADAPTIVE). This means we assume low workload in single task and high workload in dual task. We varied the number of recognition outputs used for the linear temporal smoothing from 0 to 20. Figure 6.3 shows the corresponding recognition results of the SVM classification (SVM) based system and the *DCR Framework* using regression (DCRFrmw) using the linear temporal smoothing with $s \in \{0, 2, 5, 10, 20\}$ averaged over all subjects.

One can clearly see the superior performance of the *DCR Framework* in this evaluation, which can be attributed to benefits of regression in comparison to the classification based SVM approach. Both recognition approaches benefit from the linear temporal smoothing. It achieved an increase in recognition accuracy of more than 10% (absolute) for both classification approaches, whereby the estimate includes information from the EEG signals of the last 12 seconds instead of the last two seconds when no temporal smoothing is applied. The maximum recognition accuracy of 85.5% was achieved with the *DCR Framework* using the recognition results from the last 20 recognition outputs for temporal smoothing. In comparison, increasing the window



	0	2	5	10	20
SVM-based	69.83%	70.63%	75.91%	79.95%	81.72%
DCRFrmw	74.86%	77.19%	81.52%	84.51%	85.53%

Figure 6.3 – Recognition accuracies of the SVM classification (SVM) based system and the *DCR Framework* regression based system (DCRFrmw) for lengths $s \in \{0, 2, 5, 10, 20\}$ of the linear temporal smoothing.

length to 12 seconds (11.5 seconds overlap) only achieved a performance of 79.63% and 78.63% for the SVM-based system and for the *DCR Framework*-based system, respectively. Figure 6.4 shows the classification results of all 20 users with $s = 20$ recognition outputs for temporal smoothing.

The system could achieve recognition accuracies of individual users up to 95%, whereby the chance level of the (binary) classification task is about 50%. The performance of nearly all users is above 80%, with the exception of user 13 that achieved only 65%, which can be seen as an outlier. Overall, the workload recognition performance in this study was suitable to enable an automatic adaptation of the intelligent human-machine interaction system.

Task performance and Subjective Ratings

In addition to the recognition results, we assessed the performance in the human-machine interaction scenario, i.e. the performance of the users in the

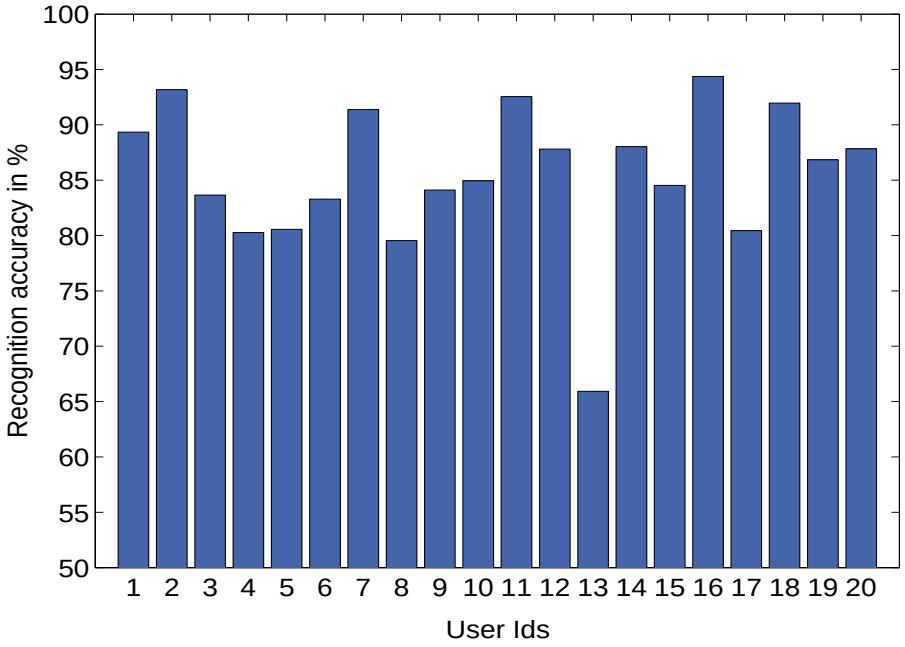


Figure 6.4 – Person-specific workload recognition accuracies of the 20 users of the *DCR Framework* regression based system with linear temporal smoothing using 20 output results.

information task and the flanker task during the 5 sessions. Table 6.4 summarizes the task performance results. Comparing the EegAdaptive sessions with the baseline session AlwaysLow shows that EegAdaptive could increase the correctness rates in both tasks in comparison to the (fast) AlwaysLow. Comparing the EegAdaptive sessions with the baseline AlwaysHigh showes a higher throughput (higher completion rate with the same correctness rate) in the information task for EegAdaptive. Furthermore, the adaptive strategies EegAdaptive and Oracle maintained the correctness rate of the (slow) AlwaysHigh.

After each session we assessed subjective ratings using a questionnaire with 11 items using 6 point scale (table 6.5). Table 6.6 summarizes the subjective evaluation results.

The evaluations showed that users clearly recognized the adaptive behavior (Q1) and preferred the adaptive speaking styles (Q2-Q7). Overall, the results of EegAdaptive are a promising approximation to the optimal adaptation strategy Oracle.

Strategy	Correctness rate information task	Completion rate information task	Correctness rate flanker task
AlwaysLow	86%	98%	69%
AlwaysHigh	96%	58%	87%
EegAdaptive	96%	85%	82%
Oracle	94%	85%	86%

Table 6.4 – Average completion and correctness rates for the robot instruction and the flanker task.

Q1	How strongly did the robot adapt to the switch between the conditions with and without secondary task?
Q2	How appropriate was the behavior of the robot in conditions without secondary task?
Q3	How appropriate was the behavior of the robot in conditions with secondary task?
Q4	Would you like to work together with a robot with this behavior?
Q5	How do you judge the behavior of the robot concerning "friendliness"?
Q6	How do you judge the behavior of the robot concerning "empathy"?
Q7	How do you judge the behavior of the robot in general?
Q8	Experienced time pressure*
Q9	Experienced accomplishment*
Q10	Experienced effort*
Q11	Experienced frustration*

Table 6.5 – Questionnaire for subjective evaluation of presentation strategies. Items marked with * are extracted from the NASA TLX workload scale.

Item	AlwaysLow	AlwaysHigh	EegAdaptive	Oracle
Q1	2.0 (0.97)	2.5 (1.66)	4.5 (1.10)	5.4 (1.09)
Q2	4.9 (1.00)	4.1 (1.75)	4.9 (1.14)	5.1 (1.07)
Q3	2.3 (1.10)	4.3 (1.03)	3.9 (1.25)	5.1 (0.89)
Q4	2.2 (1.15)	3.3 (1.18)	3.6 (1.14)	4.8 (0.69)
Q7	2.8 (0.95)	4.0 (0.71)	3.9 (0.87)	4.8 (0.61)
Q5	3.1 (1.15)	3.8 (0.80)	3.7 (1.22)	4.3 (0.86)
Q6	2.2 (0.93)	2.6 (1.19)	3.4 (0.99)	4.4 (0.87)
Q8	5.3 (0.66)	3.2 (1.14)	4.0 (0.99)	3.5 (1.23)
Q9	3.0 (1.19)	3.8 (1.16)	3.7 (1.04)	4.0 (1.27)
Q10	5.1 (1.05)	3.5 (1.12)	4.4 (0.75)	4.0 (1.09)
Q11	4.0 (1.25)	2.5 (1.05)	3.0 (1.00)	2.5 (0.61)

Table 6.6 – Subjective evaluation of the different strategies and experienced mental workload; average score (standard deviation).

The evaluations of the perceived workload (questions Q8-Q11 from NASA TLX workload scale [Hart and Staveland, 1988]) showed that users experienced low workload in AlwaysHigh. Oracle nearly reaches this workload level, while AlwaysLow generated higher workload. EegAdaptive could reduce workload and could achieve workload ratings nearly as low as Oracle, but is dependent on noisy recognition results.

We also investigated the relationship between recognition performance in the EegAdaptive session and the subjective ratings. We found a strong correlation between the difference of the user ratings of EegAdaptive and Oracle and the recognition rates of the workload recognition system. Therefore, improvements in recognition accuracy can be expected to further improve subjective experiences of the workload adaptive system towards the results of Oracle. Table 6.7 summarizes the correlation results.

Item	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Pearson correlation	-0.40	-0.18	-0.51*	-0.35	-0.51*	-0.74*	-0.54*
Item	Q8	Q9	Q10	Q11			
Pearson correlation	0.29	-0.24	0.46*	0.24			

Table 6.7 – Pearson correlation coefficients beween recognition accuracy and the difference between user ratings of Oracle and EegAdaptive (EegAdaptive-Oracle). Statistically significant correlations are marked by a star ($\alpha = 0.05$).

6.1.5 Conclusions

We described the design, implementation, and evaluation of a workload adaptive information system for closed-loop human-robot interaction in a study with single and dual tasking. The workload recognition system was able to adapt the speaking styles according to brain activity patterns of its user in real-time. We highlighted how DISCRIMINATIVE (HDspec), COMPACT (regularization) and ROBUST (artifact reduction, temporal smoothing) were implemented and could show performance improvements by using the *DCR Framework* regression compared to the SVM based classification. The linear temporal smoothing further increased the robustness against short-time variabilities. The *DCR Framework* could achieved a mean recognition rate of 85.53% for the discrimination between low and high mental workload, which outperformed our previous SVM-based system that achieved 81.72%. Furthermore, we could show that the adaptive strategy using the workload recognition improved task performance and user satisfaction in comparison to static interaction strategies.

6.2 ECoG-based Brain-to-Text Classification of Vowels

Continuous speech production is a highly complex process involving multiple parts of the human brain. The fundamental building blocks of continuous speech have been studied by scientists from different disciplines, including linguists, speech processing technologists, and computational neuroscientists. However, a fundamental representation that allows for decoding speech from neural signals has not been presented, yet. In this chapter we contribute to our research on Brain-to-Text [Herff et al., 2015], which is the first system to decode brain activity of continuously spoken speech into text that we have recently proposed. While our paper describes the decoding of continuously spoken speech, in this thesis we show that vowels produced during continuous speech can be classified from invasively measured brain activity, i.e. intracranial electrocorticographic (ECoG) recordings. ECoG measures electrical potentials directly on the brain surface with high temporal and spatial resolution. Due to the location directly on the brain surface, signals are unfiltered by skull and scalp.

We compare two different classification approaches (i) generative Gaussian models as used in our continuous decoding system [Herff et al., 2015] and (ii)

the *DCR Framework* with regularized discriminative models that are trained to be robust against non-stationarities.

In addition to their discriminative abilities, we show that the learned models can give insights into timings and locations of neural processes associated with the continuous production of speech.

6.2.1 Related Work

The high complexity and agile dynamics of the activity in cortical networks make it challenging to investigate speech production with traditional neuroimaging techniques, such as functional magnetic resonance imaging (fMRI), or non-invasive brain activity measurements, such as EEG. In the last few years, researchers started to investigate speech using ECoG recordings. To date, previous work has mostly focused on isolated aspects of speech in the brain, but so far not on the analysis and fully automatic decoding of brain activity during continuously produced natural speech.

Studies provided evidence for a neural representation of phones and phonetic features during speech perception [Chang et al., 2010, Mesgarani et al., 2014], but did not investigate continuous speech production. Furthermore, studies investigated the dynamics of the general speech production process [Crone et al., 2001a, Crone et al., 2001b]. Neural activity during the production of isolated phones [Leuthardt et al., 2011, Guenther et al., 2009, Formisano et al., 2008, Blakely et al., 2008, Pei et al., 2011] or words [Kellis et al., 2010] has been classified using different brain imaging techniques but not during continuous speech. First attempts to classify brain activity during the imagined production of isolated phones are reported in [Brumberg et al., 2011]. [Mugler et al., 2014] recently demonstrated the classification of a full set of phones within manually segmented boundaries during isolated word production.

In [Herff et al., 2015] we showed for the first time that techniques from automatic speech recognition can be applied to decode a textual representation of spoken words from neural signals. For this system we employed techniques from automatic speech recognition, such as statistical bigram language models, a restricted English dictionary that maps words to phone sequences, a Gaussian modeling of the cortical signals corresponding to phones, and a Hidden Markov Model based decoding. When restricting the dictionary to

small subsets, Word Error Rates as low as 25% could be achieved. The following evaluations are based on the dataset for this study.

6.2.2 Description of the Experiment and Data Corpus

The data set for this evaluation was recorded and provided to us in a collaboration with Dr. Gerwin Schalk and his team from the Wadsworth Center, New York State Department of Health, Albany, USA. Our collaboration in this project has been established more than three years ago and was initially funded by the KIT International Excellence fund, which supported our research visit to Albany for one month.

Participants and electrode placement

Seven epileptic patients (4 female, 3 male) who underwent neurosurgical procedures for epilepsy treatment at the Albany Medical Center (Albany, New York, USA) participated in this study. The age of the participants varied between 18 and 56 (mean age of 31.0). All participants gave informed consent and the study was approved by the Institutional Review Board of Albany Medical College and the Human Research Protections Office of the US Army Medical Research and Materiel Command.

Electrodes were implanted depending only on clinical needs of the patients. All participants had electrode grids placed on the left hemisphere that covered parts of the frontal and temporal lobes. Electrode grids (Ad-Tech Medical Corp., Racine, WI; PMT Corporation, Chanhassen, MN) consisted of platinum-iridium electrodes (4 mm in diameter, 2.3 mm exposed) with distances of 0.6-1 cm embedded in silicone. In a post-operative computer tomography scan, electrode positions were registered and co-registered with a pre-operative magnetic resonance imaging scan.

To be able to compare activations across subjects, electrode positions of all subjects were co-registered in a common Talairach space [Talairach and Tournoux, 1988]. Activation maps were rendered using the NeuralAct software package [Kubanek and Schalk, 2014]. See Figure 6.5 for electrode placement of all subjects that passed the data pre-selection process (see next section).

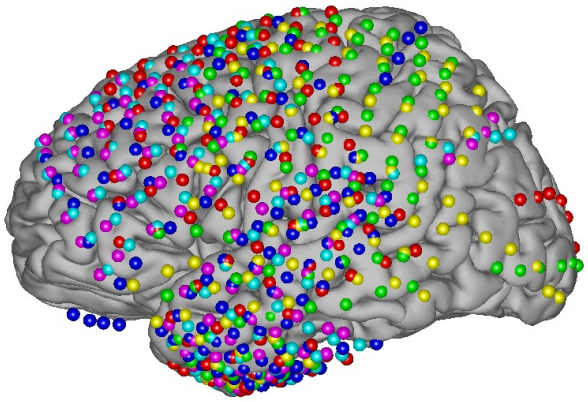


Figure 6.5 – Combined electrode montage of all participants after preselection. Participant 1 (yellow), participant 2 (magenta), participant 3 (cyan), participant 5 (red), participant 6 (green) and participant 7 (blue). Participant 4 did not yield sufficient activations related to speech activity and thus was excluded from the analyzes presented here.

Experiment Description

In this study, brain activity during overt speech production of the participants was recorded using electrocorticography (ECoG) grids that had been implanted as part of presurgical procedures preparatory to epilepsy surgery. Additionally, we recorded the acoustic waveform of the participants' speech in synchronization with the ECoG signals. Both ECoG and acoustic signals were digitized at 9600 Hz. BCI2000 [Schalk et al., 2004] and eight 16-channel g.USBamp biosignal amplifiers (g.tec, Graz, Austria) have been used to record the signals in this study.

During the experiment, participants had to read out text excerpts that consisted of historical political speeches, i.e. the Gettysburg Address [Roy and Basler, 1955] and the JFK's Inaugural Address [Kennedy, 1989], a childrens' story, i.e. Humpty Dumpty [Crane et al., 1867], and fan-fiction of the television serial *Charmed* [fanfiction.net, 2009].

The texts were displayed on a screen located in about one meter distance in front of the participant and scrolled through the screen from right to left at a constant rate, which was adjusted to the participants comfort (rate of scrolling text: 42-76 words/min). The participants had to read the displayed text aloud as it appeared. Each participant took part in two or three

recording sessions. Table 6.8 summarizes the data recording details for every session.

Participant	Session	Text	|Phrases|	Recording length (s)
1	1	GA	36	279.87
	2	JFK	38	326.90
2	1	HD	21	129.87
	2	HD	21	129.07
	3	HD	21	126.37
3	1	Charmed	42	310.27
	2	Charmed	40	310.93
	3	Charmed	41	307.50
4	1	GA	38	299.67
	2	GA	38	311.97
5	1	JFK	49	341.77
	2	GA	39	222.57
6	1	GA	38	302.83
7	1	JFK	48	590.10
	2	GA	38	391.43

Table 6.8 – Details for every recording session. Texts are abbreviated as follows: GA is the Gettysburg address, JFK is the John F. Kennedy's inaugural speech, HD is Humpty Dumpty and Charmed are Charmed fan-fiction texts.

Cross-modality Phone Labeling

We used our in-house speech recognition toolkit BioKIT [Telaar et al., 2014] to phone-label the audio recordings. As ECoG and audio data were recorded in synchronization, this enables to mark the ECoG signals corresponding to the production of any given phone as identified by the speech recognition system from the audio data.

We segmented the recorded texts along pauses into 21 to 49 phrases, depending on the session length. An English automatic speech recognition system, which was trained on broadcast news was applied to the segmented acoustic recordings. The sequence of phones was calculated by Viterbi forced alignment [Huang et al., 2001] given the transcribed texts and acoustic models of the automatic speech recognition system. We then adapted the Gaussian mixture model-based acoustic models of the system using maximum likelihood linear regression (MLLR) [Gales, 1998]. Finally, we repeated the

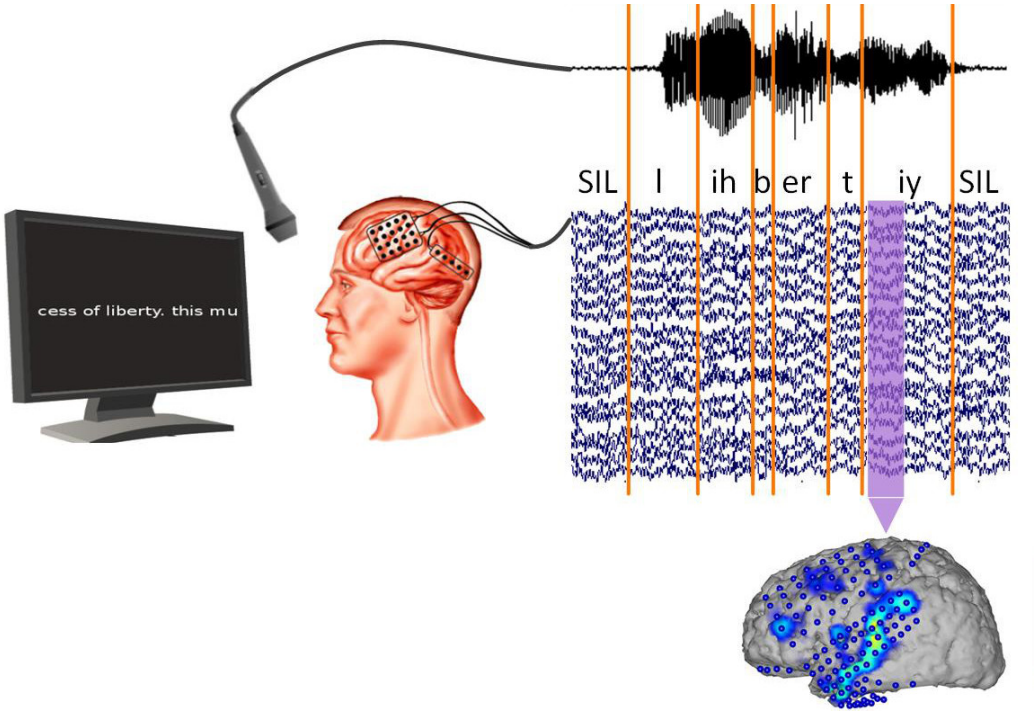


Figure 6.6 – Data recording and phone labeling. Acoustic data and ECoG data were recorded synchronously. Acoustic data is labeled on phone-level using BioKIT and labels from the acoustic data are then imposed on the neural data.

Viterbi forced alignment using the adapted models of each session. These final phone alignments were then imposed on the ECoG data.

Figure 6.6 illustrates the experimental setup and labeling of the neural data.

Data Pre-selection

To pre-select recordings, we analyzed whether speech activity segments could be distinguished from segments with no speech activity. Therefore, we fitted a Gaussian model to all feature vectors (see next section) containing speech activity and one to feature vectors when the participant was not speaking. Timings of speech and non-speech segments were extracted from the audio recordings.

In a leave-one-phrase-out validation, we then evaluated whether these models could be used to identify speech activity above chance level. Both sessions of participant 4 and session 2 of participant 5 did not show classification rates

significantly above chance level (paired t-test, $p > 0.05$) and were excluded based on this analysis.

6.2.3 Feature Extraction

The neural signal data was downsampled to 600 Hz and continuously segmented into 50 ms intervals with 25 ms overlap. This enabled to capture the fast cortical processes underlying continuous speech while providing signal segments long enough to extract spectral power in the gamma broadband between 70 and 170 Hz robustly. Each signal segment was labeled with the corresponding phone from the audio labeling.

To calculate features, we first removed linear trends in the raw signals from each channel. The signals were then down-sampled to 600 Hz. Noisy channels were identified and excluded from the evaluations. Specifically, we calculated the energy in the frequency band 58-62 Hz (line noise) and removed channels with more noise energy than two interquartile ranges above the third quartile of the energy of all channels in the data set. We used common average re-referencing on the remaining channels and applied elliptic IIR low-pass and high-pass filters to represent broadband gamma activity. To attenuate the first harmonic of 60 Hz line noise, which is within the high-gamma frequency range, we applied an elliptic IIR notch filter (118-122 Hz).

We calculated the signal energy $E_{i,c}$ for each channel c and interval i and applied the logarithm to Gaussianize the feature distribution [Gasser et al., 1982]. Then, the logarithmic broadband gamma power of all channels were stacked into one vector $E_i = [E_{i,1}, \ldots, E_{i,d}]$.

The vectors of neighboring segments up to 200 ms prior and after the current interval were concatenated to include the temporal dynamics of the context of each ECoG interval. Contexts of similar sizes have been found relevant in other speech perception studies [Sahin et al., 2009]. The resulting feature vectors thus include the four feature vectors in the past and four in the future, i.e. $F_i = [E_{i-4}, \ldots, E_i, \ldots, E_{i+4}]^\top$. The stacked feature vectors were extracted every 25 ms.

6.2.4 Vowel Classification from ECoG Data

In this evaluation, we investigate the frame-wise classification of the five vowels (/a/, /e/, /i/, /o/, /u/) from neural data. Brain activity based vowel

classification is a particularly challenging task as vowels share multiple articulatory properties and their production involves similar motor actions. To the best of our knowledge, vowel classification from neural data has not been investigated for continuous speech before. Furthermore, the frame-wise modeling of phones is an important aspect of Brain-to-text, our system for the decoding of text from neural activity ([Herff et al., 2015] and improvements in vowel classification can be expected to improve the decoding of word sequences.

In this evaluation, we compare two different classification approaches:
(i) generative Gaussian models as used in our Brain-to-Text decoding system [Herff et al., 2015] using features selected by a Kullback-Leibler divergence based feature selection and
(ii) discriminative models learned by the *DCR Framework* that are regularized for sparsity and trained to be robust against non-stationarities.

In both approaches, the vowel recognition is only based on the ECoG data and does not use the acoustic information, i.e. the acoustic data were only used to create the phone labeling that is required to train the neural models and to validate the recognition estimates.

Gaussian Model Training

To limit model complexity of the Gaussian models (COMPACT), we selected features using a Kullback-Leibler divergence (KL-div) based feature selection as follows: We estimated the relevance of the features $E_{i,c}$ (log broadband gamma at a recording position for a specific time interval) by calculating the mean KL-div [Duda et al., 2001] between all phone-pairs for this feature. The number of selected features was automatically determined based on the distributions of KL-div values. i.e. features with the largest normalized mean KL-div values were selected until the difference between subsequent values in the sorted sequence of KL-div values was smaller than -0.05. The feature selection was purely based on KL-divs in the training data and did not include any prior knowledge about suitable brain regions or time offsets.

We modeled the selected features for each vowel by a Gaussian distribution. Thus, each vowel is characterized by the mean broadband gamma activity and the variance of the neural activity measurements at each of the selected electrodes and time offsets. The Gaussian models were used in a multi-class Gaussian classifier [Bishop et al., 2006] to discriminate between the five different vowels.

DCR Framework Model Training

In addition to Gaussian models, we applied the *DCR Framework* for vowel classification. It differs from the Gaussian classification approach, as it learns discriminative models instead of the generative Gaussian models. Furthermore, it the feature selection using KL-div is not applied but an implicit feature selection by ℓ_1-norm regularization is performed. Additionally, it learns models that are robust against changes of the feature distributions over time and between multiple recording sessions (non-stationarities). For multi-class classification we employed the one-vs-rest classification scheme as described in section 4.5.2.

To incorporate invariance against non-stationarities, we chose the robustness directions as follows (cf. section 4.1.4):
For each of the five vowel classifiers in the one-vs-rest classification, we split the training features of vowel c and the other vowels into 10 blocks per session, each, and calculated the mean feature vectors B_k^c for each block. The number of blocks has been evaluated for different numbers between 1 and 16, of which 10 was a reasonable tradeoff for all participants. We set robustness directions d_k^c to the difference between the average feature vector in the training data for the vowel to be detected μ^c and the average feature vector of each of the blocks in the training data set B_k^c, i.e. $d_k^c = \mu^c - B_k^c$.

Summary of the Vowel Classification Approaches

The frame-wise vowel classification approaches implemented the three objectives Discriminative, Compact, and Robust as follows:

Discriminative: For both classification approaches we used logarithmic high-gamma power features (70-170Hz) extracted from each channel. High-gamma activity is known to correlate with language function, however the channel number and channel locations differ between the participants. Features from 50 ms windows of all channels were stacked using features from up to 200 ms before and up to 250 ms after the interval of each feature.

Compact: For the Gaussian modeling approach a KL-div based feature selection was applied. This filter based feature selection identified features whose distribution is most dissimilar between the different phones.
The *DCR Framework* does not require a specialized feature selection but uses its ℓ_1-norm penalty regularization for implicit feature selection.

ROBUST: The robustness directions of the *DCR Framework* were chosen to reduce the impact of non-stationarities. For the participant with multiple recording sessions, the choice of robustness directions also reduces between-session variabilities.

6.2.5 Evaluations and Results

ECoG recordings have high temporal and spatial resolution which allows us to trace the temporal dynamics of speech production in the brain. The topography and temporal information of the models learned by the *DCR Framework* can be visualized for neurophysiological interpretation. Figure 6.7 shows topographical maps of a model to discriminate all vowels from the other phones (without silence) for participant 7. The backward model learned by the *DCR Framework* was converted into a forward model to represent interpretable regions of cortical activity with high relevance [Haufe et al., 2014] (see section 4.5.3). Heat maps show the temporal course of regions of high discriminability (red) according to the learned model.
Starting 200 ms before the actual vowel production, early differences are present in diverse areas. Concurrent with the vowel production and shortly after the production onset, high discriminability in sensorimotor areas can be observed. 150 ms after production, the regions of highest discriminabilities correspond to auditory regions of the superior temporal gyrus.

For vowel classification, features were calculated as described in section 6.2.3. As the data available for the evaluations is very limited we combined multiple sessions of the participants. The feature vectors were restricted to vowel frames and classified frame-wise using a 10-fold cross-validation with splits between phrases.

Figure 6.8 shows the recognition results for the six evaluated participants in terms of f-scores weighted by the prior distribution of the phones. Whiskers indicate standard deviations across the different vowels.
Randomization tests showed that all recognition results were significantly above chance level, except for participant 6 (one-sided, paired Wilcoxon signed rank tests, $p < 0.05$). The classification using the *DCR Framework* shows improvements in recognition rates over Gaussian models for all participants except participant 1. It achieved significant improvements in weighted f-score over Gaussian models by up to 6.8% absolute (participants 2, 3, and 5, paired Wilcoxon signed rank tests, $p < 0.05$). The average performance improvements were 2.8% absolute.

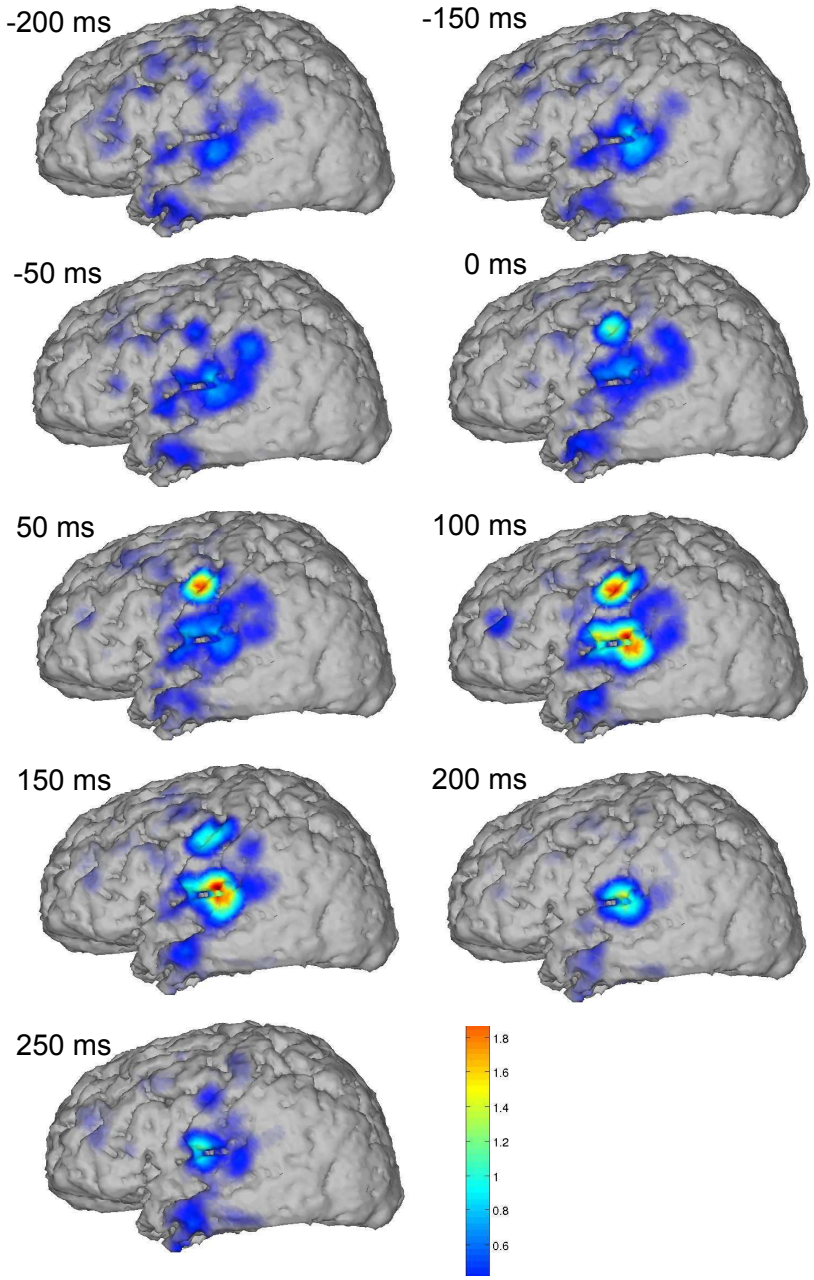


Figure 6.7 – Topographical maps showing discriminative regions of vowel speech production on the brain over time.

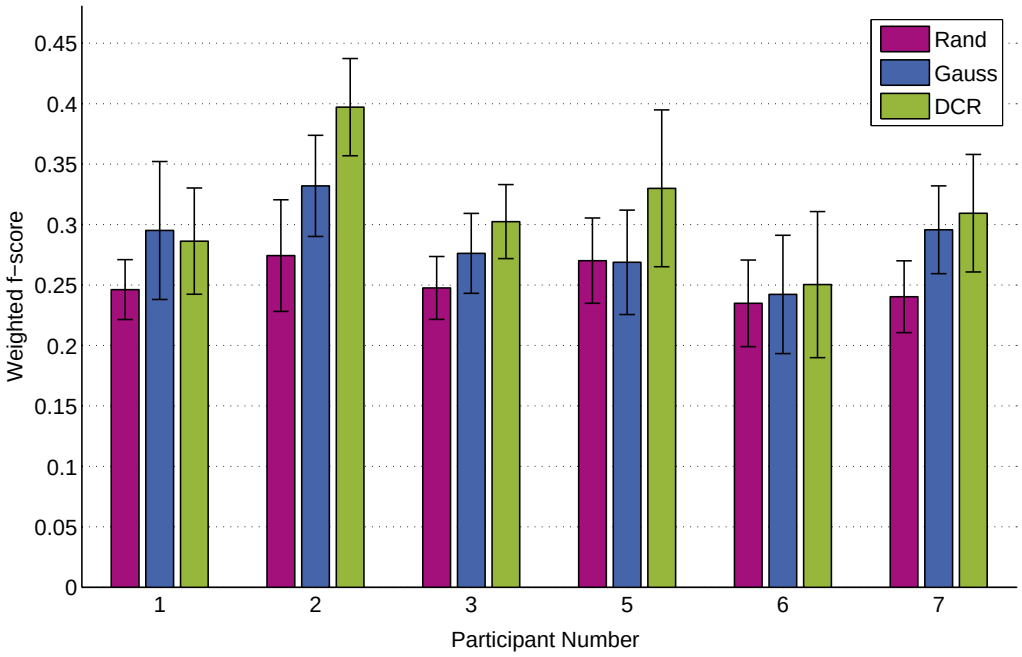


Figure 6.8 – Weighted f-scores of frame-wise vowel classification. Figure shows results of the *DCR Framework* models (green) in comparison to Gaussian models (blue) and randomized models (purple). Whiskers indicate standard deviations across the different vowels.

6.3 Conclusions

We described the evaluation of two approaches for the classification of vowels during continuous speech from ECoG data. Both vowel recognition systems were able to classify the five vowels significantly above chance level. We highlighted how Discriminative (based on high-gamma power), Compact (KL-div feature selection and ℓ_1-norm regularization) and Robust (reduction of non-stationarities) were implemented. The models learned by the *DCR Framework* show neurophysiological meaningful interactions of different brain areas involved in the speech production process. Furthermore, we could show performance improvements by using the *DCR Framework* instead of Gaussian models by up to 6.8% in weighted f-score. These results suggest that frame-wise modeling of phones in Brain-to-Text by the *DCR Framework* may increase the performance of decoding of word sequences from neural activity, which is an important step towards the recognition of imagined speech.

Chapter 7

Conclusion and Perspectives

The final chapter of this dissertation briefly summarizes the presented work. We summarize the main contributions and explain how our findings may foster future developments in BCI research.

7.1 Discussion, Contributions and Main Results

The primary goal of this dissertation was to systematically advance pattern recognition for BCIs. In the introductory chapter (section 1.2) we analyzed the current state-of-the-art of BCIs and identified four important challenges that pattern recognition for BCIs should approach: enable shorter setup times, higher information throughput, more reliable recognition and more natural and intuitive BCI paradigms. Furthermore, we found that there is a lack of principled approaches for pattern recognition in BCIs which allow to recognize many different BCI paradigms and different brain activity signals (chapter 3).

As a consequence, we formulated the hypothesis that for pattern recognition in BCIs it is necessary to implement the three objectives Discriminative, Compact, and Robust (section 3.1). These objectives have special relevance for BCIs (section 3.2) and can be related to principles of pattern recognition (section 3.3). To the best of our knowledge the three objectives

have not been formulated as a triad before, and in particular, it has not been discussed that the three objectives are dependent on each other and have to be optimized with respect to each other for optimal pattern recognition results.

The insights on the interdepencences of the three objectives lead to the *DCR Framework*, which is, to the best of our knowledge, the first BCI pattern recognition framework that jointly optimizes the three objectives (chapter 4). It makes use of generic high-dimensional features in time and frequency domain and can be used for classification and regression problems. A particular novelty of the *DCR Framework* are the robustness directions that provide an elegant method to incorporate directions in the feature space to which the learned models become invariant. To the best of our knowledge, an approach that includes robustness directions or alike has not been proposed in pattern recognition or BCI research before.
We proposed an algorithm based on the Alternating Direction Method of Multipliers (ADMM) that allows to calculate the updates in each step by elegant closed-form solutions (section 4.4). This way, calculations are very efficient, especially if the number of dimensions is much larger than the number of calibration instances.

Using the *DCR Framework*, we evaluated 8 different BCI data sets with EEG, fNIRS, and ECoG data and 2 synthetically generated data sets (chapters 5 and 6). We showed that the proposed methods can achieve state-of-the-art recognition performances outperforming numerous current alternative methods. The evaluations include different publicly available benchmark data sets, such as the BCI Challenge @ NER2015 in which our submission based on the *DCR Framework* won the 2nd prize at the IEEE Neural Engineering Conference 2015. The application of the *DCR Framework* to very different kinds of BCI problems using classification or regression problems of different brain signal types shows that it is indeed a quite generic approach for BCI pattern recognition.

With this work, we contributed to each of the four challenges listed above as follows: We showed that the *DCR Framework* can be used when only small amounts of calibration data have been recorded (section 5.1). Furthermore, we showed that it can be applied to reduce the setup times by transfer learning using only a small amount of user specific calibration data and additional data from different subjects (in section 5.3). Furthermore, we evaluated its multi-subject learning capabilities (section 5.4). Therefore, we conclude that the *DCR Framework* improves inconveniences of BCI setups

by *reducing setup times* (challenge 1) before BCI use.
In the evaluations in chapters 5 and 6, we showed improvements in recognition accuracies for multiple BCI problems in comparison to many alternative approaches. By definition [Kronegg et al., 2005], improvements in recognition rates directly translate to *improvements in information throughput* (challenge 2). Furthermore, because of its generic data-driven approach, we assume that the *DCR Framework* will benefit from future advancements in new measurement technologies that may lead to further increases in information throughput of BCIs (see next section).
We have highlighted that the proposed robustness directions technique of the *DCR Framework* contributes to *more reliable recognition* (challenge 3) results. Specifically, we have shown that the robustness directions can improve recognition when signal variabilities are present that are caused by using data of different users (sections 5.2.1, 5.3 and 5.4) or caused by non-stationarities (sections 5.2.2 and 6.2).
Following the principles of DISCRIMINATIVE, COMPACT, and ROBUST, we advanced pattern recognition for *novel BCI paradigms* (challenge 4) for which there are no well established pattern recognition methods. Specifically, we introduced two examples for novel BCI paradigms in which spontaneous naturally occurring brain activity of the user is analyzed and interpreted (chapter 6). For workload recognition we could show advanced recognition performance using the *DCR Framework*. Furthermore, the analysis of the task performance and self-reports of the users showed measurable benefits of EEG-based workload adaptation (section 6.1). Additionally, we contributed to the emerging research of recognizing speech from invasive neural signals by contributing a first recognition system for vowels during continuously articulated speech from ECoG signals (section 6.2).

In summary, this dissertation contributed the following key achievements that are important foundations for both, practical BCI application and future BCI research:

- Formulation of three objectives DISCRIMINATIVE, COMPACT, and ROBUST as theoretical concepts that are necessary for pattern recognition of BCIs (chapter 3)
- Design and development of the *DCR Framework*, a generic pattern recognition framework for BCIs that optimizes the three objectives in a single joint convex optimization (chapter 4)
- Extensive empirical validation of our theory and the *DCR Framework* using multiple different brain activity signals and multiple different BCI problems (chapter 5)

- Advancing pattern recognition for novel BCI paradigms, i.e. EEG-based workload recognition during the interaction in a semi-controlled task and the classification of vowels during continuous speech from ECoG signals (chapter 6)

7.2 Perspectives and Future Directions

BCI research is still in its infancy and major obstacles have to be overcome to bring BCIs to real-life applications with a widespread user acceptance. In general, traditional brain signal recording devices, such as clinical EEG-caps with conductive gel, are not acceptable for most users in daily life. Therefore, research has to provide new brain activity sensors to overcome such inconvenient setups of current systems. Furthermore, it is crucial to design innovative BCI applications that bring the users measurable benefits, which cannot easily be achieved by less obtrusive interfaces. Additionally, new invasive technologies that may become available in the near future are of particular interest as they enable to assess much richer information from the brain, as outlined below.

In the remainder of this section, we outline a number of directions for future research related to pattern recognition for BCIs that emerge from the theoretical and practical advancements that have been developed in this thesis.

A major limiting factor for pattern recognition of BCIs is the small amount of calibration data that is currently available for BCI research. Therefore, a highly important aspect of BCI pattern recognition is a flexible implementation of the Compact objective in order to model as much information as possible from the available data. Increasing the amount of calibration data, has shown to strongly impact recognition rates and robustness in other pattern recognition based research disciplines, such as automatic speech recognition and computer vision. Recently, a first large-scale EEG data corpus has been published by the Temple University Hospital (e.g. [Obeid and Picone, 2013, Obeid et al., 2014]). It comprises about 22,000 EEG recordings from approximately 15,000 different persons, including their medical histories and clinical diagnoses. Although this is not a typical BCI data set, BCI pattern recognition methods can be applied here as research tools (cf. [Brunner et al., 2015]) to analyze and infer new insights on brain function. The basic algorithm of the *DCR Framework* (section 4.4.4) can easily be extended for large-scale distributed optimization (cluster com-

puting) with moderate modifications. Therefore, such distributed extensions should be developed for intelligent analysis of big neural data.

BCIs for communication and control, which are currently a major part of BCI research, may not be the primary applications in the future. Instead, new BCI paradigms need to be investigated, in which the user interacts naturally with the environment and the BCI contributes only additional information about the user (cf. passive BCIs). Because of the current limitations of BCIs, the highest potential for such applications may arise from combining BCIs with other sensory modalities. These so-called hybrid BCIs [Pfurtscheller et al., 2010] have not been discussed in this work, however the *DCR Framework* supports large amounts of features, which enables its use for feature fusion of information from BCIs and additional sensory modalities. For example, applications that combine BCIs with augmented reality interfaces, such as Google Glass and Microsoft Hololens, should be developed. They have the opportunities to provide innovative real-world applications in which the BCI enables completely new ways of interacting with the device that can have significant benefits for the users.

In current BCI research, brain activity patterns are typically associated with simple behavioral patterns of the user to train supervised learning models. However, complex interactions in real-world applications require some form of context awareness to be able to associate the brain activity patterns with the corresponding events that occur in a certain situation. Such information may be provided by an underlying cognitive model that integrates basic principles about user behavior, the user's task, and environmental knowledge. Putze [Putze, 2014] recently proposed ways to combine empirical cognitive models, such as BCIs, with computational cognitive models. The study discussed in section 6.1 can be seen as a first promising step into this interesting research direction.

Invasive BCIs using ECoG and microarrays are an important research direction, in particular to address the limitations of current BCIs due to their low information throughput. Invasive systems have shown to be applicable to complex tasks, such as the control of robotic arms for reaching and grasping [Hochberg et al., 2012]. Currently, new sensors and new technology to measure brain activity are developed that may allow long term implantation of sensors with low health risks for the user. For example, micro-scale, free-floating sensor networks [Seo et al., 2013] or sensors based on neuro-nanotechnology [Alivisatos et al., 2013], may enable to infer more relevant and precise information about neural processes for the next generation of BCIs. The generic data-driven approach using high-dimensional features

and its scalability in terms of number of features (COMPACT-objective) and number of instances (possibility by distributed optimization) are important factors that may allow the *DCR Framework* and similar methodologies to be suitable pattern recognition approaches for these new technologies. Overall, we believe that advancements in measuring cortical information are a key to enable BCIs with high information throughput for many real-world applications and we hope that the fundamental ideas proposed in this dissertation will be highly valuable for the development of pattern recognition methodologies for such future BCIs.

Bibliography

[Alamgir et al., 2009] Alamgir, M., Grosse-Wentrup, M., and Altun, Y. (2009). Multitask learning for brain-computer interfaces. In Teh, Y.W., M. T., editor, *JMLR Workshop and Conference Proceedings Volume 9: AISTATS 2010*, pages 17–24, Cambridge, MA, USA. Max-Planck-Gesellschaft, JMLR.

[Alivisatos et al., 2013] Alivisatos, A. P., Andrews, A. M., Boyden, E. S., Chun, M., Church, G. M., Deisseroth, K., Donoghue, J. P., Fraser, S. E., Lippincott-Schwartz, J., Looger, L. L., et al. (2013). Nanotools for neuroscience and brain activity mapping. *ACS nano*, 7(3):1850–1866.

[Allison et al., 2007] Allison, B., Graimann, B., and Gräser, A. (2007). Why use a BCI if you are healthy. In *ACE Workshop-Brain-Computer Interfaces and Games*, pages 7–11.

[Allison et al., 2008] Allison, B. Z., McFarland, D. J., Schalk, G., Zheng, S. D., Jackson, M. M., and Wolpaw, J. R. (2008). Towards an independent brain–computer interface using steady state visual evoked potentials. *Clinical neurophysiology*, 119(2):399–408.

[Allison and Neuper, 2010] Allison, B. Z. and Neuper, C. (2010). Could anyone use a BCI? In *Brain-computer interfaces*, pages 35–54. Springer.

[Anderson et al., 1995] Anderson, C. W., Devulapalli, S. V., and Stolz, E. A. (1995). EEG signal classification with different signal representations. In *Neural Networks for Signal Processing [1995] V. Proceedings of the 1995 IEEE Workshop*, pages 475–483. IEEE.

[Ang et al., 2008] Ang, K. K., Chin, Z. Y., Zhang, H., and Guan, C. (2008). 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*, pages 2390–2397. IEEE.

[Ang et al., 2010] Ang, K. K., Guan, C., Sui Geok Chua, K., Ang, B. T., Kuah, C., Wang, C., Phua, K. S., Chin, Z. Y., and Zhang, H. (2010). Clinical study of neurorehabilitation in stroke using EEG-based motor imagery brain-computer interface with robotic feedback. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 5549–5552. IEEE.

[Anokhin and Vogel, 1996] Anokhin, A. and Vogel, F. (1996). Eeg alpha rhythm frequency and intelligence in normal adults. *Intelligence*, 23(1):1–14.

[Arvaneh et al., 2011] Arvaneh, M., Guan, C., Ang, K. K., and Quek, C. (2011). Optimizing the channel selection and classification accuracy in EEG-based BCI. *Biomedical Engineering, IEEE Transactions on*, 58(6):1865–1873.

[Arvaneh et al., 2013] Arvaneh, M., Guan, C., Ang, K. K., and Quek, C. (2013). EEG data space adaptation to reduce intersession nonstationarity in brain-computer interface. *Neural computation*, 25(8):2146–2171.

[Arvaneh et al., 2014] Arvaneh, M., Robertson, I., and Ward, T. E. (2014). Subject-to-subject adaptation to reduce calibration time in motor imagery-based brain-computer interface. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 6501–6504. IEEE.

[Ayaz et al., 2007] Ayaz, H., Izzetoglu, M., Bunce, S., Heiman-Patterson, T., and Onaral, B. (2007). Detecting cognitive activity related hemodynamic signal for brain computer interface using functional near infrared spectroscopy. In *Neural Engineering, 2007. CNE'07. 3rd International IEEE/EMBS Conference on*, pages 342–345. IEEE.

[Ayaz et al., 2012] Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., and Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *Neuroimage*, 59(1):36–47.

[Babiloni et al., 2007] Babiloni, F., Cichocki, A., and Gao, S. (2007). Brain-computer interfaces: towards practical implementations and potential applications. *Computational Intelligence and Neuroscience*, 2007.

[Bach and Obozinski, 2010] Bach, F. and Obozinski, G. (2010). Sparse methods for machine learning theory and algorithms. *ECML/PKDD Tutorial*.

[Badcock et al., 2013] Badcock, N. A., Mousikou, P., Mahajan, Y., de Lissa, P., Thie, J., and McArthur, G. (2013). Validation of the emotiv epoc® eeg gaming system for measuring research quality auditory erps. *PeerJ*, 1:e38.

[Barachant et al., 2012] Barachant, A., Bonnet, S., Congedo, M., and Jutten, C. (2012). Multiclass brain–computer interface classification by Riemannian geometry. *Biomedical Engineering, IEEE Transactions on*, 59(4):920–928.

[Bashashati et al., 2007] Bashashati, A., Fatourechi, M., Ward, R. K., and Birch, G. E. (2007). A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. *Journal of Neural engineering*, 4(2):R32.

[Bauernfeind et al., 2014] Bauernfeind, G., Steyrl, D., Brunner, C., and Muller-Putz, G. R. (2014). Single trial classification of fNIRS-based brain-computer interface mental arithmetic data: A comparison between different classifiers. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 2004–2007. IEEE.

[Bellman and Dreyfus, 1962] Bellman, R. E. and Dreyfus, S. E. (1962). *Applied dynamic programming*. Rand Corporation.

[Bensch et al., 2007] Bensch, M., Karim, A. A., Mellinger, J., Hinterberger, T., Tangermann, M., Bogdan, M., Rosenstiel, W., and Birbaumer, N. (2007). Nessi: an EEG-controlled web browser for severely paralyzed patients. *Computational intelligence and neuroscience*, 2007.

[Berka et al., 2004] Berka, C., Levendowski, D. J., Cvetinovic, M. M., Petrovic, M. M., Davis, G., Lumicao, M. N., Zivkovic, V. T., Popovic, M. V., and Olmstead, R. (2004). Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. *International Journal of Human-Computer Interaction*, 17(2):151–170.

[Berka et al., 2007] Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Olmstead, R. E., Tremoulet, P. D., and Craven, P. L. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, 78(Supplement 1):B231–B244.

[Berka et al., 2005] Berka, C., Levendowski, D. J., Ramsey, C. K., Davis, G., Lumicao, M. N., Stanney, K., Reeves, L., Regli, S. H., Tremoulet, P. D., and Stibler, K. (2005). Evaluation of an eeg workload model in

an aegis simulation environment. In *Defense and security*, pages 90–99. International Society for Optics and Photonics.

[Birbaumer, 2006] Birbaumer, N. (2006). Breaking the silence: brain–computer interfaces (bci) for communication and motor control. *Psychophysiology*, 43(6):517–532.

[Bishop et al., 2006] Bishop, C. M. et al. (2006). *Pattern recognition and machine learning*, volume 1. springer New York.

[Blakely et al., 2008] Blakely, T., Miller, K. J., Rao, R. P., Holmes, M. D., and Ojemann, J. G. (2008). Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 4964–4967. IEEE.

[Blankertz et al., 2007] Blankertz, B., Kawanabe, M., Tomioka, R., Hohlefeld, F., Müller, K.-r., and Nikulin, V. V. (2007). Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing. In *Advances in neural information processing systems*, pages 113–120.

[Blankertz et al., 2011] Blankertz, B., Lemm, S., Treder, M., Haufe, S., and Mller, K.-R. (2011). Single-trial analysis and classification of {ERP} components a tutorial. *NeuroImage*, 56(2):814 – 825. Multivariate Decoding and Brain Reading.

[Blankertz et al., 2008a] Blankertz, B., Müller, K.-R., Krusienski, D., Schalk, G., Wolpaw, J. R., Schlögl, A., Pfurtscheller, G., Millán, J. d. R., Schröder, M., and Birbaumer, N. (2008a). The BCI Competition III.

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

[Borwein and Lewis, 2010] Borwein, J. M. and Lewis, A. S. (2010). *Convex analysis and nonlinear optimization: theory and examples*, volume 3. Springer Science & Business Media.

[Bos et al., 2010] Bos, D. P.-O., Reuderink, B., van de Laar, B., Gürkök, H., Mühl, C., Poel, M., Nijholt, A., and Heylen, D. (2010). Brain-computer interfacing and games. In *Brain-Computer Interfaces*, pages 149–178. Springer.

[Bousquet and Elisseeff, 2002] Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526.

[Boyd et al., 2011] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

[Brandl et al., 2015] Brandl, S., Müller, K.-R., and Samek, W. (2015). Robust common spatial patterns based on Bhattacharyya distance and gamma divergence. In *Brain-Computer Interface (BCI), 2015 3rd International Winter Conference on*, pages 1–4. IEEE.

[Brookings et al., 1996] Brookings, J. B., Wilson, G. F., and Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological psychology*, 42(3):361–377.

[Brouwer et al., 2012] Brouwer, A.-M., Hogervorst, M. A., Van Erp, J. B., Heffelaar, T., Zimmerman, P. H., and Oostenveld, R. (2012). Estimating workload using EEG spectral power and ERPs in the n-back task. *Journal of neural engineering*, 9(4):045008.

[Brumberg et al., 2011] Brumberg, J. S., Wright, E. J., Andreasen, D. S., Guenther, F. H., and Kennedy, P. R. (2011). Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex. *Frontiers in neuroscience*, 5.

[Brunner et al., 2015] Brunner, C., Birbaumer, N., Blankertz, B., Guger, C., Kübler, A., Mattia, D., Millán, J. d. R., Miralles, F., Nijholt, A., Opisso, E., et al. (2015). BNCI Horizon 2020: towards a roadmap for the BCI community. *Brain-Computer Interfaces*, (ahead-of-print):1–10.

[Carruthers and Smith, 1996] Carruthers, P. and Smith, P. K. (1996). *Theories of theories of mind*. Cambridge Univ Press.

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

[Chang et al., 2010] Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature neuroscience*, 13(11):1428–1432.

[Christoforou et al., 2010] Christoforou, C., Haralick, R., Sajda, P., and Parra, L. C. (2010). Second-order bilinear discriminant analysis. *The Journal of Machine Learning Research*, 11:665–685.

[Combettes and Wajs, 2005] Combettes, P. L. and Wajs, V. R. (2005). Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200.

[Cooper et al., 2012] Cooper, R. J., Selb, J., Gagnon, L., Phillip, D., Schytz, H. W., Iversen, H. K., Ashina, M., and Boas, D. A. (2012). A systematic comparison of motion artifact correction techniques for functional near-infrared spectroscopy. *Frontiers in neuroscience*, 6.

[Crane et al., 1867] Crane, W., Gilbert, John, S., McConnell, W., Tenniel, John, S., Weir, H., and Zwecker, J. B. (1867). *Mother Gooses Nursery Rhymes. A Collection of Alphabets, Rhymes, Tales and Jingles.* London: George Routledge and Sons.

[Croft and Barry, 2000] Croft, R. and Barry, R. (2000). Removal of ocular artifact from the EEG: a review. *Neurophysiologie Clinique/Clinical Neurophysiology*, 30(1):5–19.

[Crone et al., 2001a] Crone, N., Hao, L., Hart, J., Boatman, D., Lesser, R., Irizarry, R., and Gordon, B. (2001a). Electrocorticographic gamma activity during word production in spoken and sign language. *Neurology*, 57(11):2045–2053.

[Crone et al., 2001b] Crone, N. E., Boatman, D., Gordon, B., and Hao, L. (2001b). Induced electrocorticographic gamma activity during auditory perception. *Clinical Neurophysiology*, 112(4):565–582.

[Crone et al., 2006] Crone, N. E., Sinai, A., and Korzeniewska, A. (2006). High-frequency gamma oscillations and human brain mapping with electrocorticography. *Progress in brain research*, 159:275–295.

[Daly and Wolpaw, 2008] Daly, J. J. and Wolpaw, J. R. (2008). Brain–computer interfaces in neurological rehabilitation. *The Lancet Neurology*, 7(11):1032–1043.

[Daubechies et al., 2004] Daubechies, I., Defrise, M., and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on pure and applied mathematics*, 57(11):1413–1457.

[De Clercq et al., 2006] De Clercq, W., Vergult, A., Vanrumste, B., Van Paesschen, W., and Van Huffel, S. (2006). Canonical correlation anal-

ysis applied to remove muscle artifacts from the electroencephalogram. *Biomedical Engineering, IEEE Transactions on*, 53(12):2583–2587.

[De Waard and Studiecentrum, 1996] De Waard, D. and Studiecentrum, V. (1996). *The measurement of drivers' mental workload*. Groningen University, Traffic Research Center.

[Debener et al., 2005] Debener, S., Makeig, S., Delorme, A., and Engel, A. K. (2005). What is novel in the novelty oddball paradigm? functional significance of the novelty p3 event-related potential as revealed by independent component analysis. *Cognitive Brain Research*, 22(3):309–321.

[Debener et al., 2012] Debener, S., Minow, F., Emkes, R., Gandras, K., and Vos, M. (2012). How about taking a low-cost, small, and wireless EEG for a walk? *Psychophysiology*, 49(11):1617–1621.

[Delorme et al., 2011] Delorme, A., Mullen, T., Kothe, C., Acar, Z. A., Bigdely-Shamlo, N., Vankov, A., and Makeig, S. (2011). EEGLAB, SIFT, NFT, BCILAB, and ERICA: new tools for advanced EEG processing. *Computational intelligence and neuroscience*, 2011:10.

[Donchin and Coles, 1988] Donchin, E. and Coles, M. G. (1988). Is the P300 component a manifestation of context updating? *Behavioral and brain sciences*, 11(03):357–374.

[Duda et al., 2001] Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern classification*. John Wiley & Sons,.

[Eriksen, 1995] Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition*, 2(2-3):101–118.

[Eriksen and Schultz, 1979] Eriksen, C. W. and Schultz, D. W. (1979). Information processing in visual search: A continuous flow conception and experimental results. *Perception & Psychophysics*, 25(4):249–263.

[Falkenstein et al., 2000] Falkenstein, M., Hoormann, J., Christ, S., and Hohnsbein, J. (2000). ERP components on reaction errors and their functional significance: a tutorial. *Biological psychology*, 51(2):87–107.

[fanfiction.net, 2009] fanfiction.net (2009). *"Traitor among us" and "Split Feelings"*. available on `https://www.fanfiction.net/`.

[Farquhar and Hill, 2013] Farquhar, J. and Hill, N. J. (2013). Interactions between pre-processing and classification methods for event-related-potential classification. *Neuroinformatics*, 11(2):175–192.

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

[Fatourechi et al., 2007] Fatourechi, M., Bashashati, A., Ward, R. K., and Birch, G. E. (2007). EMG and EOG artifacts in brain computer interface systems: A survey. *Clinical neurophysiology*, 118(3):480–494.

[Fazel et al., 2001] Fazel, M., Hindi, H., and Boyd, S. P. (2001). A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pages 4734–4739. IEEE.

[Fazli et al., 2012] Fazli, S., Mehnert, J., Steinbrink, J., Curio, G., Villringer, A., Müller, K.-R., and Blankertz, B. (2012). Enhanced performance by a hybrid NIRS–EEG brain computer interface. *Neuroimage*, 59(1):519–529.

[Fazli et al., 2009] Fazli, S., Popescu, F., Danóczy, M., Blankertz, B., Müller, K.-R., and Grozea, C. (2009). Subject-independent mental state classification in single trials. *Neural networks*, 22(9):1305–1312.

[Ferrez and del R Millan, 2008] Ferrez, P. W. and del R Millan, J. (2008). Error-related EEG potentials generated during simulated brain–computer interaction. *Biomedical Engineering, IEEE Transactions on*, 55(3):923–929.

[Filipe et al., 2011] Filipe, S., Charvet, G., Foerster, M., Porcherot, J., Beche, J., Bonnet, S., Audebert, P., Régis, G., Zongo, B., Robinet, S., et al. (2011). A wireless multichannel eeg recording platform. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 6319–6322. IEEE.

[Fong et al., 2003] Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3):143–166.

[Formisano et al., 2008] Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "Who" Is Saying "What"? Brain-based decoding of human voice and speech. *Science*, 322(5903):970–973.

[Frey et al., 2013] Frey, J., Mühl, C., Lotte, F., and Hachet, M. (2013). Review of the use of electroencephalography as an evaluation method for human-computer interaction. *arXiv preprint arXiv:1311.2222*.

[Friedman et al., 2010] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

[Friedman, 1989] Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175.

[Galán et al., 2008] Galán, F., Nuttin, M., Lew, E., Ferrez, P. W., Vanacker, G., Philips, J., and Millán, J. d. R. (2008). A brain-actuated wheelchair: asynchronous and non-invasive brain–computer interfaces for continuous control of robots. *Clinical Neurophysiology*, 119(9):2159–2169.

[Gales, 1998] Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech & language*, 12(2):75–98.

[Garrett et al., 2003] Garrett, D., Peterson, D. A., Anderson, C. W., and Thaut, M. H. (2003). Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 11(2):141–144.

[Gasser et al., 1982] Gasser, T., Bächer, P., and Möcks, J. (1982). Transformations towards the normal distribution of broad band spectral parameters of the EEG. *Electroencephalography and clinical neurophysiology*, 53(1):119–124.

[Gevins and Smith, 2003] Gevins, A. and Smith, M. E. (2003). Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science*, 4(1-2):113–131.

[Goldberger et al., 2000] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C., and Stanley, H. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

[Goncharova et al., 2003] Goncharova, I., McFarland, D. J., Vaughan, T. M., and Wolpaw, J. R. (2003). EMG contamination of EEG: spectral and topographical characteristics. *Clinical Neurophysiology*, 114(9):1580–1593.

[Graimann et al., 2010] Graimann, B., Allison, B., and Pfurtscheller, G. (2010). Brain–computer interfaces: A gentle introduction. In *Brain-Computer Interfaces*, pages 1–27. Springer.

[Griffiths et al., 2003] Griffiths, D., Nelo, J. P., Robinson, A., Spaar, J., Vilnai, Y., and Veigl, C. (2003). The ModularEEG design. openeeg.sourceforge.net.

[Guenther et al., 2009] Guenther, F. H., Brumberg, J. S., Wright, E. J., Nieto-Castanon, A., Tourville, J. A., Panko, M., Law, R., Siebert, S. A., Bartels, J. L., Andreasen, D. S., et al. (2009). A wireless brain-machine interface for real-time speech synthesis. *PloS one*, 4(12):e8218.

[Guger et al., 2009] Guger, C., Daban, S., Sellers, E., Holzner, C., Krausz, G., Carabalona, R., Gramatica, F., and Edlinger, G. (2009). How many people are able to control a P300-based brain–computer interface (BCI)? *Neuroscience letters*, 462(1):94–98.

[Gundel and Wilson, 1992] Gundel, A. and Wilson, G. F. (1992). Topographical changes in the ongoing EEG related to the difficulty of mental tasks. *Brain topography*, 5(1):17–25.

[Gysels and Celka, 2004] Gysels, E. and Celka, P. (2004). Phase synchronization for the recognition of mental tasks in a brain-computer interface. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 12(4):406–415.

[Hart and Staveland, 1988] Hart, S. G. and Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning*, volume 2. Springer.

[Haufe et al., 2014] Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., and Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110.

[He and Yuan, 2012] He, B. and Yuan, X. (2012). On the O(1/n) convergence rate of the douglas-rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709.

[Heger et al., 2014a] Heger, D., Herff, C., Putze, F., Mutter, R., and Schultz, T. (2014a). Continuous affective states recognition using functional near infrared spectroscopy. *Brain-Computer Interfaces*, 1(2):113–125.

[Heger et al., 2015] Heger, D., Herff, C., Putze, F., and Schultz, T. (2015). Joint optimization for discriminative, compact and robust brain-computer

interfacing. In *7th International IEEE EMBS Conference on Neural Engineering*. IEEE.

[Heger et al., 2014b] Heger, D., Herff, C., and Schultz, T. (2014b). Combining feature extraction and classification for fNIRS BCIs by regularized least squares optimization. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 2012–2015. IEEE.

[Heger et al., 2010a] Heger, D., Putze, F., Amma, C., Wand, M., Plotkin, I., Wielatt, T., and Schultz, T. (2010a). Biosignalsstudio: a flexible framework for biosignal capturing and processing. In *KI 2010: Advances in Artificial Intelligence*, pages 33–39. Springer.

[Heger et al., 2013] Heger, D., Putze, F., Herff, C., and Schultz, T. (2013). Subject-to-subject transfer for CSP based BCIs: feature space transformation and decision-level fusion. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 5614–5617. IEEE.

[Heger et al., 2010b] Heger, D., Putze, F., and Schultz, T. (2010b). An adaptive information system for an empathic robot using EEG data. In *Social Robotics*, pages 151–160. Springer.

[Heger et al., 2010c] Heger, D., Putze, F., and Schultz, T. (2010c). Online workload recognition from EEG data during cognitive tests and human-machine interaction. In *KI 2010: Advances in Artificial Intelligence*, pages 410–417. Springer.

[Heger et al., 2011a] Heger, D., Putze, F., and Schultz, T. (2011a). An EEG adaptive information system for an empathic robot. *International Journal of Social Robotics*, 3(4):415–425.

[Heger et al., 2011b] Heger, D., Putze, F., and Schultz, T. (2011b). Online recognition of facial actions for natural EEG-based BCI applications. In *Affective Computing and Intelligent Interaction*, pages 436–446. Springer.

[Heger et al., 2014c] Heger, D., Terziyska, E., and Schultz, T. (2014c). Connectivity based feature-level filtering for single-trial EEG BCIs. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2064–2068. IEEE.

[Herff et al., 2015] Herff, C., Heger, D., de Pesters, A., Telaar, D., Brunner, P., Schalk, G., and Schultz, T. (2015). Brain-to-text: Decoding spoken

phrases from phone representations in the brain. *Frontiers in Neuroscience*, 9(217).

[Herff et al., 2013a] Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., and Schultz, T. (2013a). Mental workload during n-back taskquantified in the prefrontal cortex using fNIRS. *Frontiers in human neuroscience*, 7.

[Herff et al., 2013b] Herff, C., Heger, D., Putze, F., Hennrich, J., Fortmann, O., and Schultz, T. (2013b). Classification of mental tasks in the prefrontal cortex using fNIRS. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 2160–2163. IEEE.

[Higham, 2002] Higham, N. J. (2002). *Accuracy and stability of numerical algorithms*. Siam.

[Hild et al., 2014] Hild, J., Putze, F., Kaufman, D., Kühnle, C., Schultz, T., and Beyerer, J. (2014). Spatio-temporal event selection in basic surveillance tasks using eye tracking and EEG. In *Proceedings of the 7th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Eye-Gaze & Multimodality*, pages 3–8. ACM.

[Hinterberger et al., 2004] Hinterberger, T., Schmidt, S., Neumann, N., Mellinger, J., Blankertz, B., Curio, G., and Birbaumer, N. (2004). Brain-computer communication and slow cortical potentials. *Biomedical Engineering, IEEE Transactions on*, 51(6):1011–1018.

[Hirshfield et al., 2009] Hirshfield, L. M., Solovey, E. T., Girouard, A., Kebinger, J., Jacob, R. J., Sassaroli, A., and Fantini, S. (2009). Brain measurement for usability testing and adaptive interfaces: an example of uncovering syntactic workload with functional near infrared spectroscopy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2185–2194. ACM.

[Hjorth, 1970] Hjorth, B. (1970). EEG analysis based on time domain properties. *Electroencephalography and clinical neurophysiology*, 29(3):306–310.

[Hochberg et al., 2012] Hochberg, L. R., Bacher, D., Jarosiewicz, B., Masse, N. Y., Simeral, J. D., Vogel, J., Haddadin, S., Liu, J., Cash, S. S., van der Smagt, P., et al. (2012). Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, 485(7398):372–375.

[Hoerl and Kennard, 1970] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

[Holm et al., 2009] Holm, A., Lukander, K., Korpela, J., Sallinen, M., and Müller, K. M. (2009). Estimating brain load from the EEG. *The Scientific World Journal*, 9:639–651.

[Homan et al., 1987] Homan, R. W., Herman, J., and Purdy, P. (1987). Cerebral location of international 10–20 system electrode placement. *Electroencephalography and clinical neurophysiology*, 66(4):376–382.

[Hosmer and Lemeshow, 2000] Hosmer, D. W. and Lemeshow, S. (2000). Introduction to the logistic regression model. *Applied Logistic Regression, Second Edition*, pages 1–30.

[Huang et al., 2001] Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall PTR.

[Huggins et al., 2014] Huggins, J. E., Guger, C., Allison, B., Anderson, C. W., Batista, A., Brouwer, A.-M., Brunner, C., Chavarriaga, R., Fried-Oken, M., Gunduz, A., et al. (2014). Workshops of the fifth international brain-computer interface meeting: defining the future. *Brain-Computer Interfaces*, 1(1):27–49.

[Hurley and Rickard, 2009] Hurley, N. and Rickard, S. (2009). Comparing measures of sparsity. *Information Theory, IEEE Transactions on*, 55(10):4723–4741.

[Jarvis et al., 2011] Jarvis, J., Putze, F., Heger, D., and Schultz, T. (2011). Multimodal person independent recognition of workload related biosignal patterns. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 205–208. ACM.

[Jung et al., 2000] Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., Mckeown, M. J., Iragui, V., and Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(02):163–178.

[Kandel et al., 2000] Kandel, E. R., Schwartz, J. H., Jessell, T. M., et al. (2000). *Principles of neural science*, volume 4. McGraw-Hill New York.

[Kang and Choi, 2011] Kang, H. and Choi, S. (2011). Bayesian multi-task learning for common spatial patterns. In *Pattern Recognition in NeuroImaging (PRNI), 2011 International Workshop on*, pages 61–64. IEEE.

[Kang et al., 2009] Kang, H., Nam, Y., and Choi, S. (2009). Composite common spatial pattern for subject-to-subject transfer. *Signal Processing Letters, IEEE*, 16(8):683–686.

[Kellis et al., 2010] Kellis, S., Miller, K., Thomson, K., Brown, R., House, P., and Greger, B. (2010). Decoding spoken words using local field potentials recorded from the cortical surface. *Journal of neural engineering*, 7(5):056007.

[Kennedy, 1989] Kennedy, J. F. (1989). *Inaugural Addresses of the Presidents of the United States. (Washington, DC)*. Available online at: `www.bartleby.com/124/`.

[Kindermans et al., 2014a] Kindermans, P.-J., Schreuder, M., Schrauwen, B., Müller, K.-R., and Tangermann, M. (2014a). True zero-training brain-computer interfacing–an online study. *PloS one*, 9(7):e102504.

[Kindermans et al., 2014b] Kindermans, P.-J., Tangermann, M., Müller, K.-R., and Schrauwen, B. (2014b). Integrating dynamic stopping, transfer learning and language models in an adaptive zero-training ERP speller. *Journal of neural engineering*, 11(3):035005.

[Kirchner, 1958] Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, 55(4):352.

[Klimesch et al., 1993] Klimesch, W., Schimke, H., and Pfurtscheller, G. (1993). Alpha frequency, cognitive load and memory performance. *Brain topography*, 5(3):241–251.

[Kohlmorgen et al., 2007] Kohlmorgen, J., Dornhege, G., Braun, M., Blankertz, B., Müller, K.-R., Curio, G., Hagemann, K., Bruns, A., Schrauf, M., and Kincses, W. (2007). Improving human performance in a real operating environment through real-time mental workload detection. *Toward Brain-Computer Interfacing*, pages 409–422.

[Koles, 1991] Koles, Z. J. (1991). The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG. *Electroencephalography and clinical Neurophysiology*, 79(6):440–447.

[Kothe and Makeig, 2011] Kothe, C. A. and Makeig, S. (2011). Estimation of task workload from EEG data: new and current tools and perspectives. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 6547–6551. IEEE.

[Kothe and Makeig, 2013] Kothe, C. A. and Makeig, S. (2013). BCILAB: a platform for brain–computer interface development. *Journal of neural engineering*, 10(5):056014.

[Kramer, 1990] Kramer, A. F. (1990). Physiological metrics of mental workload: A review of recent progress. Technical report, DTIC Document.

[Krauledat et al., 2008] Krauledat, M., Tangermann, M., Blankertz, B., and Müller, K.-R. (2008). Towards zero training for brain-computer interfacing. *PloS one*, 3(8):e2967.

[Kronegg et al., 2005] Kronegg, J., Voloshynovskiy, S., and Pun, T. (2005). Analysis of bit-rate definitions for brain-computer interfaces. In *Csrea hci*, pages 40–46.

[Kubanek and Schalk, 2014] Kubanek, J. and Schalk, G. (2014). NeuralAct: A Tool to Visualize Electrocortical (ECoG) Activity on a Three-Dimensional Model of the Cortex. *Neuroinformatics*, pages 1–8.

[Kübler et al., 1999] Kübler, A., Kotchoubey, B., Hinterberger, T., Ghanayim, N., Perelmouter, J., Schauer, M., Fritsch, C., Taub, E., and Birbaumer, N. (1999). The thought translation device: a neurophysiological approach to communication in total motor paralysis. *Experimental Brain Research*, 124(2):223–232.

[Kulish et al., 2006] Kulish, V., Sourin, A., and Sourina, O. (2006). Human electroencephalograms seen as fractal time series: Mathematical analysis and visualization. *Computers in biology and medicine*, 36(3):291–302.

[Laureys et al., 2005] Laureys, S., Pellas, F., Van Eeckhout, P., Ghorbel, S., Schnakers, C., Perrin, F., Berre, J., Faymonville, M.-E., Pantke, K.-H., Damas, F., et al. (2005). The locked-in syndrome: what is it like to be conscious but paralyzed and voiceless? *Progress in brain research*, 150:495–611.

[Ledoit and Wolf, 2004] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411.

[Leuthardt et al., 2011] Leuthardt, E. C., Gaona, C., Sharma, M., Szrama, N., Roland, J., Freudenberg, Z., Solis, J., Breshears, J., and Schalk, G. (2011). Using the electrocorticographic speech network to control a brain–computer interface in humans. *Journal of neural engineering*, 8(3):036004.

[Li and Guan, 2006] Li, Y. and Guan, C. (2006). An extended EM algorithm for joint feature extraction and classification in brain-computer interfaces. *Neural computation*, 18(11):2730–2761.

[Lim et al., 2010] Lim, C. G., Lee, T.-S., Guan, C., Fung, D. S., Cheung, Y. B., Teng, S., Zhang, H., and Krishnan, K. (2010). Effectiveness of a

brain-computer interface based programme for the treatment of ADHD: a pilot study. *Psychopharmacol Bull*, 43(1):73–82.

[Lin et al., 2003] Lin, C. J., Hsu, C.-W., and Chang, C.-C. (2003). A practical guide to support vector classification. *National Taiwan University*.

[Lin et al., 2005] Lin, C.-T., Wu, R.-C., Liang, S.-F., Chao, W.-H., Chen, Y.-J., and Jung, T.-P. (2005). EEG-based drowsiness estimation for safety driving using independent component analysis. *Circuits and Systems I: Regular Papers, IEEE Transactions on*, 52(12):2726–2738.

[Liu and Nocedal, 1989] Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

[Lotte et al., 2007] Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B., et al. (2007). A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of neural engineering*, 4.

[Lotte and Guan, 2011] Lotte, F. and Guan, C. (2011). Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms. *Biomedical Engineering, IEEE Transactions on*, 58(2):355–362.

[Lotte et al., 2009] Lotte, F., Guan, C., and Ang, K. K. (2009). Comparison of designs towards a subject-independent brain-computer interface based on motor imagery. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 4543–4546. IEEE.

[Mak et al., 2011] Mak, J., Arbel, Y., Minett, J., McCane, L., Yuksel, B., Ryan, D., Thompson, D., Bianchi, L., and Erdogmus, D. (2011). Optimizing the P300-based brain–computer interface: current status, limitations and future directions. *Journal of neural engineering*, 8(2):025003.

[Makeig et al., 1996] Makeig, S., Bell, A. J., Jung, T.-P., Sejnowski, T. J., et al. (1996). Independent component analysis of electroencephalographic data. *Advances in neural information processing systems*, pages 145–151.

[Makeig et al., 2012] Makeig, S., Kothe, C., Mullen, T., Bigdely-Shamlo, N., Zhang, Z., and Kreutz-Delgado, K. (2012). Evolving signal processing for brain–computer interfaces. *Proceedings of the IEEE*, 100(Special Centennial Issue):1567–1584.

[Margaux et al., 2012] Margaux, P., Emmanuel, M., Sébastien, D., Olivier, B., and Jérémie, M. (2012). Objective and subjective evaluation of on-

line error correction during P300-based spelling. *Advances in Human-Computer Interaction*, 2012:4.

[Matthews et al., 2008] Matthews, F., Pearlmutter, B. A., Ward, T. E., Soraghan, C., and Markham, C. (2008). Hemodynamics for brain-computer interfaces. *Signal Processing Magazine, IEEE*, 25(1):87–94.

[Mesgarani et al., 2014] Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, page 1245994.

[Mitchell, 1997] Mitchell, T. M. (1997). Machine learning.

[Moghimi et al., 2012] Moghimi, S., Kushki, A., Power, S., Guerguerian, A. M., and Chau, T. (2012). Automatic detection of a prefrontal cortical response to emotionally rated music using multi-channel near-infrared spectroscopy. *Journal of neural engineering*, 9(2):026022.

[Molavi and Dumont, 2012] Molavi, B. and Dumont, G. A. (2012). Wavelet-based motion artifact removal for functional near-infrared spectroscopy. *Physiological measurement*, 33(2):259.

[Moray, 1979] Moray, N. (1979). *Mental workload: Its theory and measurement*, volume 8. Plenum Publishing Corporation.

[Mugler et al., 2014] Mugler, E. M., Patton, J. L., Flint, R. D., Wright, Z. A., Schuele, S. U., Rosenow, J., Shih, J. J., Krusienski, D. J., and Slutzky, M. W. (2014). Direct classification of all American English phonemes using signals from functional speech motor cortex. *Journal of Neural Engineering*, 11(3):035015.

[Mühl et al., 2014a] Mühl, C., Allison, B., Nijholt, A., and Chanel, G. (2014a). A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain-Computer Interfaces*, 1(2):66–84.

[Mühl et al., 2014b] Mühl, C., Jeunet, C., and Lotte, F. (2014b). EEG-based workload estimation across affective contexts. *Frontiers in neuroscience*, 8.

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

[Müller et al., 2004] Müller, K.-R., Krauledat, M., Dornhege, G., Curio, G., and Blankertz, B. (2004). Machine learning techniques for brain-computer interfaces. *Biomedical Engineering*, 4:11–22.

[Müller et al., 2008] Müller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., and Blankertz, B. (2008). Machine learning for real-time single-trial EEG-analysis: from brain–computer interfacing to mental state monitoring. *Journal of neuroscience methods*, 167(1):82–90.

[Müller-Putz et al., 2005] Müller-Putz, G. R., Scherer, R., Pfurtscheller, G., and Rupp, R. (2005). Eeg-based neuroprosthesis control: a step towards clinical practice. *Neuroscience letters*, 382(1):169–174.

[Neuper et al., 2003] Neuper, C., Müller, G., Kübler, A., Birbaumer, N., and Pfurtscheller, G. (2003). Clinical application of an EEG-based brain–computer interface: a case study in a patient with severe motor impairment. *Clinical neurophysiology*, 114(3):399–409.

[Nicolas-Alonso and Gomez-Gil, 2012] Nicolas-Alonso, L. F. and Gomez-Gil, J. (2012). Brain computer interfaces, a review. *Sensors*, 12(2):1211–1279.

[Niedermeyer and da Silva, 2005] Niedermeyer, E. and da Silva, F. L. (2005). *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins.

[Nijholt et al., 2009] Nijholt, A., Bos, D. P.-O., and Reuderink, B. (2009). Turning shortcomings into challenges: Brain–computer interfaces for games. *Entertainment Computing*, 1(2):85–94.

[Obeid et al., 2014] Obeid, I., Harati, A., Jacobson, M., and Picone, J. (2014). A big-data approach to automated eeg labeling. *Frontiers in Neuroinformatics*, (94).

[Obeid and Picone, 2013] Obeid, I. and Picone, J. (2013). Bringing big data to neural interfaces. In *Proc. Fifth Int. BCI Meeting, June*, pages 3–7.

[Obermaier et al., 2003] Obermaier, B., Müller, G., and Pfurtscheller, G. (2003). "virtual keyboard" controlled by spontaneous EEG activity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(4):422–426.

[Ogawa et al., 1990] Ogawa, S., Lee, T.-M., Kay, A. R., and Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872.

[Oostenveld and Praamstra, 2001] Oostenveld, R. and Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical neurophysiology*, 112(4):713–719.

[Pan and Yang, 2010] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359.

[Parikh and Boyd, 2013] Parikh, N. and Boyd, S. (2013). Proximal algorithms. *Foundations and Trends in optimization*, 1(3):123–231.

[Parra et al., 2008] Parra, L. C., Christoforou, C., Gerson, A. D., Dyrholm, M., Luo, A., Wagner, M., Philiastides, M. G., and Sajda, P. (2008). Spatiotemporal linear decoding of brain state. *Signal Processing Magazine, IEEE*, 25(1):107–115.

[Peckham and Knutson, 2005] Peckham, P. H. and Knutson, J. S. (2005). Functional electrical stimulation for neuromuscular applications*. *Annu. Rev. Biomed. Eng.*, 7:327–360.

[Pei et al., 2011] Pei, X., Barbour, D. L., Leuthardt, E. C., and Schalk, G. (2011). Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal of neural engineering*, 8(4):046028.

[Pfurtscheller et al., 2010] Pfurtscheller, G., Allison, B. Z., Brunner, C., Bauernfeind, G., Solis-Escalante, T., Scherer, R., Zander, T. O., Mueller-Putz, G., Neuper, C., and Birbaumer, N. (2010). The hybrid BCI. *Frontiers in neuroscience*, 4.

[Pfurtscheller and Neuper, 1997] Pfurtscheller, G. and Neuper, C. (1997). Motor imagery activates primary sensorimotor area in humans. *Neuroscience letters*, 239(2):65–68.

[Phothisonothai and Nakagawa, 2007] Phothisonothai, M. and Nakagawa, M. (2007). Fractal-based EEG data analysis of body parts movement imagery tasks. *The Journal of Physiological Sciences*, 57(4):217–226.

[Picard, 2000] Picard, R. W. (2000). Toward computers that recognize and respond to user emotion. *IBM systems journal*, 39(3.4):705–719.

[Polich, 2007] Polich, J. (2007). Updating p300: an integrative theory of p3a and p3b. *Clinical neurophysiology*, 118(10):2128–2148.

[Putze, 2014] Putze, F. (2014). *Adaptive Cognitive Interaction Systems*. PhD thesis, Karlsruhe Institute of Technology.

[Putze et al., 2015] Putze, F., Amma, C., and Schultz, T. (2015). Design and evaluation of a self-correcting gesture interface based on error potentials from EEG. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3375–3384. ACM.

[Putze et al., 2013] Putze, F., Heger, D., and Schultz, T. (2013). Reliable subject-adapted recognition of EEG error potentials using limited calibration data. In *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*, pages 419–422. IEEE.

[Reissland and Zander, 2009] Reissland, J. and Zander, T. O. (2009). Automated detection of bluffing in a gamerevealing a complex covert user state with a passive BCI. In *Proceedings of the Human Factors and Ergonomics Society Europe Chapter Annual Meeting, Linkoeping, Sweden.*

[Reuderink et al., 2011] Reuderink, B., Farquhar, J., Poel, M., and Nijholt, A. (2011). A subject-independent brain-computer interface based on smoothed, second-order baselining. In *Engineering in medicine and biology society, EMBC, 2011 annual international conference of the IEEE*, pages 4600–4604. IEEE.

[Rockstroh et al., 1984] Rockstroh, B., Birbaumer, N., Elbert, T., and Lutzenberger, W. (1984). Operant control of EEG and event-related and slow brain potentials. *Biofeedback and Self-regulation*, 9(2):139–160.

[Roland et al., 2010] Roland, J., Brunner, P., Johnston, J., Schalk, G., and Leuthardt, E. C. (2010). Passive real-time identification of speech and motor cortex during an awake craniotomy. *Epilepsy & Behavior*, 18(1):123–128.

[Romero et al., 2008] Romero, S., Mañanas, M. A., and Barbanoj, M. J. (2008). A comparative study of automatic techniques for ocular artifact reduction in spontaneous EEG signals based on clinical target variables: a simulation case. *Computers in biology and medicine*, 38(3):348–360.

[Roy and Basler, 1955] Roy, E. and Basler, P. (1955). *The gettysburg address, in The Collected Works of Abraham Lincoln.* New Brunswick, NJ: Rutgers University Press.

[Russell and Norvig, 2009] Russell, S. J. and Norvig, P. (2009). *Artificial intelligence: a modern approach (3rd edition).* Prentice Hall.

[Sahin et al., 2009] Sahin, N. T., Pinker, S., Cash, S. S., Schomer, D., and Halgren, E. (2009). Sequential processing of lexical, grammatical, and phonological information within brocas area. *Science*, 326(5951):445–449.

[Samek et al., 2014] Samek, W., Kawanabe, M., and Müller, K.-R. (2014). Divergence-based framework for common spatial patterns algorithms. *Biomedical Engineering, IEEE Reviews in*, 7:50–72.

[Samek et al., 2012] Samek, W., Vidaurre, C., Müller, K.-R., and Kawanabe, M. (2012). Stationary common spatial patterns for brain–computer interfacing. *Journal of neural engineering*, 9(2):026013.

[Santana et al., 2014] Santana, E., Brockmeier, A. J., and Principe, J. C. (2014). Joint optimization of algorithmic suites for EEG analysis. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 2997–3000. IEEE.

[Sassaroli and Fantini, 2004] Sassaroli, A. and Fantini, S. (2004). Comment on the modified Beer Lambert law for scattering media. *Physics in Medicine and Biology*, 49(14):N255.

[Schäfer and Strimmer, 2005] Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1).

[Schalk et al., 2004] Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N., and Wolpaw, J. R. (2004). BCI2000: a general-purpose brain-computer interface (BCI) system. *Biomedical Engineering, IEEE Transactions on*, 51(6):1034–1043.

[Schalk et al., 2000] Schalk, G., Wolpaw, J. R., McFarland, D. J., and Pfurtscheller, G. (2000). EEG-based communication: presence of an error potential. *Clinical Neurophysiology*, 111(12):2138–2144.

[Schlögl and Brunner, 2008] Schlögl, A. and Brunner, C. (2008). BioSig: a free and open source software library for BCI research. *Computer*, 41(10):44–50.

[Schlögl et al., 2007] Schlögl, A., Keinrath, C., Zimmermann, D., Scherer, R., Leeb, R., and Pfurtscheller, G. (2007). A fully automated correction method of EOG artifacts in EEG recordings. *Clinical neurophysiology*, 118(1):98–104.

[Schultz et al., 2013] Schultz, T., Amma, C., Wand, M., Heger, D., and Putze, F. (2013). Biosignale-basierte Mensch-Maschine Schnittstellen. *at–Automatisierungstechnik at–Automatisierungstechnik*, 61(11):760–769.

[Seo et al., 2013] Seo, D., Carmena, J. M., Rabaey, J. M., Alon, E., and Maharbiz, M. M. (2013). Neural dust: An ultrasonic, low power solution for chronic brain-machine interfaces. *arXiv preprint arXiv:1307.2196*.

[Shi et al., 2014] Shi, W., Ling, Q., Yuan, K., Wu, G., and Yin, W. (2014). On the linear convergence of the ADMM in decentralized consensus optimization. *Signal Processing, IEEE Transactions on*, 62(7):1750–1761.

[Simon et al., 2013] Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

[Skinner, 1938] Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. Appleton-Century.

[Spüler et al., 2012] Spüler, M., Rosenstiel, W., and Bogdan, M. (2012). Online adaptation of a c-VEP brain-computer interface (BCI) based on error-related potentials and unsupervised learning. *PloS one*, 7(12):e51077.

[Sugiyama and Kawanabe, 2012] Sugiyama, M. and Kawanabe, M. (2012). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT Press.

[Tai and Chau, 2009] Tai, K. and Chau, T. (2009). Single-trial classification of NIRS signals during emotional induction tasks: towards a corporeal machine interface. *Journal of neuroengineering and rehabilitation*, 6:39.

[Talairach and Tournoux, 1988] Talairach, J. and Tournoux, P. (1988). *Coplanar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging*. Thieme.

[Tangermann et al., 2012] Tangermann, M., Müller, K.-R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Miller, K. J., Müller-Putz, G. R., et al. (2012). Review of the BCI competition IV. *Frontiers in neuroscience*, 6.

[Tangermann, 2007] Tangermann, M. W. (2007). *Feature selection for brain-computer interfaces*. PhD thesis, Universität Tübingen.

[Telaar et al., 2014] Telaar, D., Wand, M., Gehrig, D., Putze, F., Amma, C., Heger, D., Vu, N. T., Erhardt, M., Schlippe, T., Janke, M., Herff, C., and Schultz, T. (2014). BioKit - Real-time decoder for biosignal processing. In *Interspeech*.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[Toh et al., 1999] Toh, K.-C., Todd, M. J., and Tütüncü, R. H. (1999). SDPT3a MATLAB software package for semidefinite programming, version 1.3. *Optimization methods and software*, 11(1-4):545–581.

[Tomioka and Müller, 2010] Tomioka, R. and Müller, K.-R. (2010). A regularized discriminative framework for EEG analysis with application to brain–computer interface. *NeuroImage*, 49(1):415–432.

[Tomioka and Sugiyama, 2009] Tomioka, R. and Sugiyama, M. (2009). Dual-augmented lagrangian method for efficient sparse reconstruction. *Signal Processing Letters, IEEE*, 16(12):1067–1070.

[Tu and Sun, 2012] Tu, W. and Sun, S. (2012). A subject transfer framework for EEG classification. *Neurocomputing*, 82:109–116.

[Valiant, 1984] Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

[van Gerven et al., 2009] van Gerven, M., Farquhar, J., Schaefer, R., Vlek, R., Geuze, J., Nijholt, A., Ramsey, N., Haselager, P., Vuurpijl, L., Gielen, S., et al. (2009). The brain–computer interface cycle. *Journal of Neural Engineering*, 6(4):041001.

[Vanhatalo et al., 2003] Vanhatalo, S., Voipio, J., Dewaraja, A., Holmes, M., and Miller, J. (2003). Topography and elimination of slow EEG responses related to tongue movements. *Neuroimage*, 20(2):1419–1423.

[Vapnik, 2000] Vapnik, V. (2000). *The nature of statistical learning theory*. Springer Science & Business Media.

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

[Vidaurre et al., 2011] Vidaurre, C., Sannelli, C., Müller, K.-R., and Blankertz, B. (2011). Co-adaptive calibration to improve BCI efficiency. *Journal of neural engineering*, 8(2):025009.

[Vidaurre et al., 2006] Vidaurre, C., Schlogl, A., Cabeza, R., Scherer, R., and Pfurtscheller, G. (2006). A fully on-line adaptive BCI. *Biomedical Engineering, IEEE Transactions on*, 53(6):1214–1219.

[von Bünau et al., 2009] von Bünau, P., Meinecke, F. C., Király, F. C., and Müller, K.-R. (2009). Finding stationary subspaces in multivariate time series. *Physical Review Letters*, 103(21):214101.

[von Lühmann, 2014] von Lühmann, A. (2014). Design and evaluation of a system for mobile brain activity measurement using functional near-infrared spectroscopy. Master's thesis, Karlsruhe Institute of Technology.

[Von Luxburg and Schölkopf, 2008] Von Luxburg, U. and Schölkopf, B. (2008). Statistical learning theory: models, concepts, and results. *arXiv preprint arXiv:0810.4752*.

[Wainwright, 2009] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5):2183–2202.

[Walter et al., 2013] Walter, C., Schmidt, S., Rosenstiel, W., Gerjets, P., and Bogdan, M. (2013). Using cross-task classification for classifying workload levels in complex learning tasks. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 876–881. IEEE.

[Wang and Zheng, 2008] Wang, H. and Zheng, W. (2008). Local temporal common spatial patterns for robust single-trial EEG classification. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 16(2):131–139.

[Wang et al., 2009] Wang, J., Pohlmeyer, E., Hanna, B., Jiang, Y.-G., Sajda, P., and Chang, S.-F. (2009). Brain state decoding for rapid image retrieval. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 945–954. ACM.

[Wang et al., 2012] Wang, Z., Gunduz, A., Brunner, P., Ritaccio, A. L., Ji, Q., and Schalk, G. (2012). Decoding onset and direction of movements using electrocorticographic (ECoG) signals in humans. *Frontiers in neuroengineering*, 5.

[Welch, 1967] Welch, P. (1967). The use of Fast Fourier Transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, pages 70–73.

[Wickens, 2008] Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3):449–455.

[Wikimedia Commons, 2008] Wikimedia Commons (2008). Components of ERP. CC BY-SA 3.0.

[Wolpaw and Wolpaw, 2011] Wolpaw, J. and Wolpaw, E. W. (2011). *Brain-computer interfaces: principles and practice*. Oxford University Press.

[Wolpaw et al., 2002] Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791.

[Woodbury, 1950] Woodbury, M. A. (1950). Inverting modified matrices. *Memorandum report*, 42:106.

[Zander and Jatzev, 2009] Zander, T. O. and Jatzev, S. (2009). Detecting affective covert user states with passive brain-computer interfaces. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–9. IEEE.

[Zander and Kothe, 2011] Zander, T. O. and Kothe, C. (2011). Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *Journal of neural engineering*, 8(2):025005.

[Zhang et al., 2010] Zhang, B., Wang, J., and Fuhlbrigge, T. (2010). A review of the commercial brain-computer interface technology from perspective of industrial robotics. In *Automation and Logistics (ICAL), 2010 IEEE International Conference on*, pages 379–384. IEEE.

[Zhang et al., 2013] Zhang, R., Xu, P., Liu, T., Zhang, Y., Guo, L., Li, P., and Yao, D. (2013). Local temporal correlation common spatial patterns for single trial EEG classification during motor imagery. *Computational and mathematical methods in medicine*, 2013.

[Zhao and Yu, 2006] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

[Zschocke and Hansen, 2011] Zschocke, S. and Hansen, H.-C. (2011). *Klinische Elektroenzephalographie*. Springer-Verlag.

Appendix A

A.1 BCI and Biosignals MATLAB Toolbox

During the work for this dissertation, a toolbox for processing of BCI and biosignals time series data has been developed. All of the presented pattern recognition methods and evaluations have been implemented in this toolbox. It consists of a modular and light-weight architecture of object-oriented operations implemented in MATLAB.

Figure A.1 illustrates the features of the toolbox. The key features are:

- Nearly 100 operators for signal processing and machine learning that can be flexibly combined
- Object-oriented MATLAB implementation
- Capable of online real-time processing

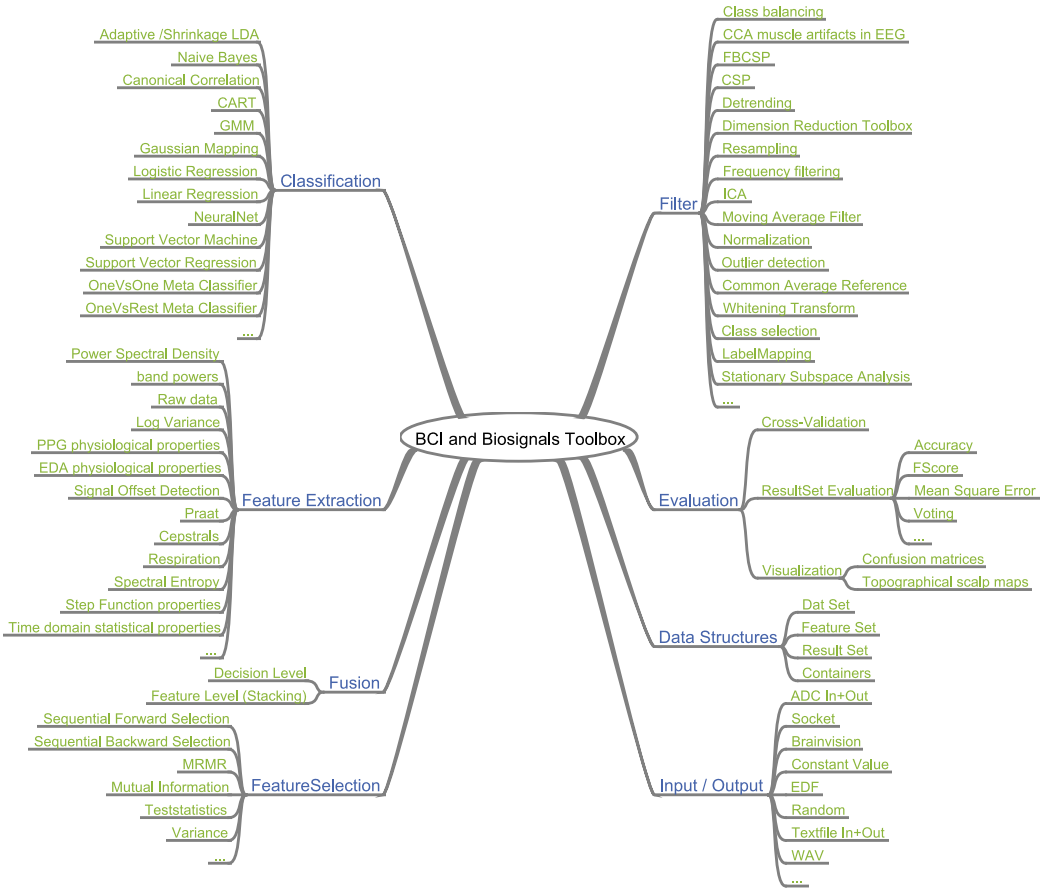


Figure A.1 – Features of the BCI and Biosignals MATLAB Toolbox