Machine Learning Techniques for Neurotechnology with Applications for Healthy Users and Patients

vorgelegt von
M.Sc.
Johannes Höhne
geb. in Leverkusen

von der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
– Dr. rer. nat. –

genehmigte Dissertation

Promotionsausschuss

Vorsitzender	Prof. Dr. Klaus Obermayer
Gutachter	Prof. Klaus-Robert Müller
Gutachter	Prof. Benjamin Blankertz
Gutachterin	Prof. Andrea Kübler

Tag der wissenschaftlichen Aussprache: 09.12.2014

Berlin 2015

To my father, Manfred Höhne

ABSTRACT

ADVANCES in Neurotechnology are based on the recording and analysis of brain activity. Brain-Computer Interfaces (BCIs) constitute a very active research area within Neurotechnology. BCIs make it possible for brain activity to be directly translated into control commands and thus enable a communication channel that is independent of muscle control. One major goal of this research is to help people who cannot communicate independently, due to neural diseases such as stroke or Amyotrophic Lateral Sclerosis (ALS). BCIs may help these patients with advanced paralysis regain their communication abilities by using their minds to interact with their surroundings. This thesis contributes to the developments of BCIs in three ways.

Firstly, novel auditory BCI paradigms – named PASS2D and CharStreamer – are described and evaluated in online studies with healthy users. Both paradigms are based on Event Related Potentials (ERPs) and provide an intuitive and fast communication with BCI for users with impaired vision. While prior auditory BCI paradigms are rather complicated to use, the CharStreamer can be operated with instructions as simple as "please attend to the letter that you want to spell". Additionally, two offline studies investigate the impact of stimulus properties on the ERPs, and the performance and usability of a BCI system. However, the above mentioned studies also indicate that the state-of-the-art analysis pipeline for ERP-based BCI paradigms might be suboptimal, as ERPs exhibit additional label information which is not exploited in a Linear Discriminant Analysis (LDA).

Therefore, the second main contribution of this thesis deals with methodological improvements which yield more accurate data analysis than state-of-the-art methods. It is shown that neuroimaging data – in particular EEG data arising from BCI paradigms – exhibit intrinsic subclass structure, which can be exploited in a meaningful way. A novel Machine Learning method – called Relevance Subclass LDA (RSLDA) – is developed and tested on multiple EEG and fMRI data sets. It is shown that RSLDA yields increased classification accuracy, as well as a better interpretation of the underlying structure in the data. Both aspects are highly favorable, suggesting that RSLDA is suitable for various classification problems within neuroimaging and beyond.

Thirdly, a BCI study is conducted with severely motor-impaired individuals. It is shown that the application of modern Machine Learning methods allows to set up a highly flexible BCI system for patients with severe paralysis. This enables to achieve significant BCI control within a very small number of sessions. Moreover, this study shows that communication via BCI can be faster and more robust than communication with other assisted technology which is based on muscle activity. This shows for the first time that the neuronal signals of an attempted motor execution can be detected prior to the muscular movement of a patient.

ZUSAMMENFASSUNG

FORTSCHRITTE in der Neurotechnologie basieren auf der Aufnahme und Analyse von Hirnaktivität. Gehirn-Computer Schnittstellen (engl. Brain-Computer Interfaces, BCIs) stellen ein sehr aktives Forschungsfeld innerhalb der Neurotechnologie dar. BCIs ermöglichen es, Hirnaktivität direkt in Steuersignale zu übersetzen und schaffen damit einen Kommunikationsweg, der unabhängig von Muskelaktivität ist. Ein Hauptziel dieser Forschung ist es, Menschen zu helfen die aufgrund von neuronalen Erkrankungen wie dem Schlaganfall oder der Amyotrophe Lateralsklerose (ALS) nicht mehr eigenständig kommunizieren können. BCIs können diesen gelähmten Patienten ermöglichen, einen Teil ihrer Kommunikationsfähigkeit zurück zu erlangen, indem sie über ihre Hirnströme mit der Umwelt interagieren. Diese Dissertation trägt dem Fortschritt von BCIs auf drei verschiedenen Weisen bei.

Zunächst werden zwei neuartige auditorische BCI Paradigmen – genannt PASS2D und CharStreamer – beschrieben und über Online-Studien mit gesunden Versuchspersonen evaluiert. Beide Paradigmen basieren auf Ereigniskorrelierten Potentialen (engl. Event Related Potentials, ERPs) und ermöglichen eine intuitive und schnelle Kommunikation mit dem BCI für Nutzer mit Sehstörungen. Während bisherige auditorische BCI Paradigmen in ihrer Anwendung sehr kompliziert sind, kann der CharStreamer mit einer Anweisung so einfach wie "bitte konzentrieren Sie sich auf den Buchstaben den Sie auswählen möchten" genutzt werden. Zusätzlich untersuchen zwei Offline-Studien, wie sich die Stimuluseigenschaften sowohl auf die ERPs, als auch auf die Genauigkeit und die Benutzerfreundlichkeit eines BCIs auswirken. Die genannten Studien weisen weiterhin darauf hin, dass die bisher üblichen Datenanalyseverfahren zu ERP-basierten BCI Paradigmen suboptimal sind. ERP Daten weisen bestimmte Informationen auf, die bei allgemein verwendeten Linearen Diskriminanzanalyse (LDA) nicht berücksichtigt werden.

Der zweite maßgebliche Beitrag dieser Thesis befasst sich daher mit neuartigen Methoden, die zu einer verbesserten Datenanalyse führen. Es wird gezeigt, dass Neuroimaging Daten – insbesondere EEG Daten aus BCI Experimenten – eine intrinsische Subklassenstruktur aufweisen. Dazu wird eine neue Methode des Maschinellen Lernens entwickelt, die eine solche Subklassenstruktur verwerten kann – Relevance Subclass LDA (RSLDA). RSLDA ermöglicht sowohl eine verbesserte Klassifikationsgenauigkeit als auch eine verbesserte Interprätation der zugrundeliegenden Struktur in den Daten. Beide Aspekte sind sehr vorteilhaft und zeigen, dass RSLDA für eine Vielzahl von Klassifikationsproblemen im Bereich Neuroimaging und darüber hinaus geeignet ist.

Als dritter Beitrag wird eine BCI Studie mit schwer gelähmten Patienten durchgeführt. Es wird gezeigt dass es mit der Anwendung von modernen Methoden des Maschinellen Lernens möglich ist, ein hoch flexibles BCI System bei diesen Patienten anzuwenden. Dadurch kann eine zuverlässig BCI Steuerung innerhalb von nur wenigen Sitzungen ermöglicht werden. Diese Studie zeigt außerdem erstmals, dass die Kommunikation über das BCI schneller und zuverlässiger sein kann als über andere Unterstützungstechnologien, welche auf Muskelaktivität basieren. Es wird dabei für einen Patienten erstmalig gezeigt, dass die neuronalen Signale einer versuchten Bewegung erkannt werden können, bevor die muskuläre Aktivität messbar ist.

ACKNOWLEDGEMENTS

WHEN browsing through this thesis, I see the major outcomes of my work which would not have been possible without the support of many people. Above all, I want to thank my supervisors, Prof. Dr. Klaus-Robert Müller and Prof. Dr. Benjamin Blankertz. Prof. Müller introduced me into the world of Machine Learning and he gave me the opportunity to achieve a PhD. With his positive and enthusiastic approach to each kind of topic as well as his empathy, he constantly motivated me to follow my own scientific interest. Prof. Blankertz introduced me to the challenges of Brain-Computer Interfacing which have been fascinating me ever since. With his calm and precise working style and his endless efforts to care about everyone, he shaped a positive working environment which is surely unique. Thank you for all the support.

Both, Prof. Müller and Prof. Blankertz set up the BBCI group which has been a wonderful place to work at, with great colleagues and friends. It has been a privilege to be part of this international and interdisciplinary research group.

Special thanks go to Dr. Michael Tangermann who was stimulating my interest in auditory BCIs. Our common passion led to dozen of brainstorming sessions and also yielded several publications and a significant number of sections in this thesis. Together with Sven and Martijn, the four of us pursued the daily work in the TOBI project for three years, which I truly enjoyed.

Thank you Sven, not only for being the very first person to discuss research ideas with, but more importantly for being such a great friend, travel mate and surf buddy. Thanks Martijn and Basti for sharing perspective beyond our research careers. I want to thank Daniel for our intense methodological discussions with such fruitful outcomes and for the opportunity to sometimes also ask rather senseless questions.

Thank you Claudia, Maci, Stefan, Anne, Matthias T, Matthias SK, Javier, Han-Jeong, Felix, Janne, Markus, Paul, Daniel M, Irene, Xing-Wei and many more for their positive impact on the last four years – especially for the BBCI evenings which I will always remember. I want to thank Andrea, Imke and Dominik for their organizational and technical support which I could always rely on.

This work would not have been possible without those people who participated in the more than 150 EEG experiments, which I conducted within the time of my PhD research. I am especially grateful that I had the opportunity to work with severely motor-impaired people and I am heavily indepted to you for your motivation, patience and endurance. Thank you Pit, Kathrin and also Elisa for sharing your expertise with me and for making the time in Bad Kreuznach a unique experience. I am very thankful to Prof. Andrea Kübler for her kind support with the patient study and for her time and efforts in reading and evaluating this dissertation.

I want to thank Mirjam, Bernd, Chris, Elisa, Christian and many more people for all the good times we had together in the past years. I further owe my deepest gratitude to my parents and my family for their unconditional support for whatever I did. Our family ties enabled me to explore the world and to live an independent life while always having a place called home.

Contents

PREFACE

THE term Neurotechnology unifies a broad spectrum of technologies and methods that can be employed for various applications, such as neural rehabilitation, neural prosthesis, neuromodulation, as well as gaming and entertainment. Therefore, Neurotechnology is a highly interdisciplinary research field that combines expertise from Neuroscience, Computer Science, Psychology, Clinical Neurology and other disciplines.

Starting with the discovery of the human electroencephalography (EEG) in the early 20th century and the first human EEG recording by Hans Berger in 1926, researchers have been enabled to record and analyze human brain activity. During the late 1960's, the advances in Engineering and Computer Science allowed to monitor and analyze brain signals in real-time.

In 1973, Jacques J. Vidal then proposed a tool which translates brain signals into a control command for an external electronic device. In his landmark paper, this tool was called "Brain-Computer Interface" (BCI) and it has been a topic of extensive research ever since. Figure 1.1 depicts the basic functioning of a BCI in an intuitive manner: brain signals are recorded and the BCI system is analyzing the neuronal data in real-time. The output of the data analysis is used to control an application such as a speller. Thus, the outcome of the analysis (e.g. a spelled symbol) is shown to the user, which closes the "BCI feedback loop".

Today, there are more than 100 active research groups studying this topic throughout the world (Wolpaw, 2007). Thus, BCIs became an essential research field within Neurotechnology – with multiple BCI textbooks (Dornhege et al., 2007; Wolpaw and Wolpaw, 2012). Amongst other application scenarios, the BCI research has mainly been driven by the aspiration to provide an alternative communication solution for patients who lost voluntary muscle control. Thus, the major goal of BCI research is to restore communication and control to people with severe paralysis arising from diseases such as amyotrophic lateral sclerosis (ALS), brainstem stroke, spinal cord injury, muscular dystrophy, and cerebral palsy. Therefore, numerous experimental paradigms have been proposed that exploit distinctive brain activity. Such BCI systems are commonly relying on elaborate methods for signal processing and classification. These machine learning methods are necessary in order to enhance the neural signals of interest while suppressing of the rest of the "cerebral cocktail party" signals in real-time (Müller and Blankertz, 2006).

Despite significant advances in various aspects, such as the signal acquisition (Nicolas-Alonso and Gomez-Gil, 2012), data processing and classification (Blankertz et al., 2008b) as well as the understanding of the neuronal underpinnings of BCI control (Grosse-Wentrup et al., 2011; Halder et al., 2011), patient studies are still very rare. Kübler (2013) recently pointed out that "fewer than 10% of the papers published on brain-computer interfacing deal with individuals presenting motor restrictions, although many authors mention these as the purpose of their research". Thus, the BCI technology was tested and optimized mostly with healthy users, while its applicability with patients is rarely shown.

This thesis contributes to the research field of BCI and Neurotechnology in several ways. It is shown that novel machine learning techniques can increase the usability and performance of the BCI technology. Therefore, two novel BCI paradigms are proposed, which are both suitable for patients with impaired communication abilities. It is shown that such paradigms allow the communication with
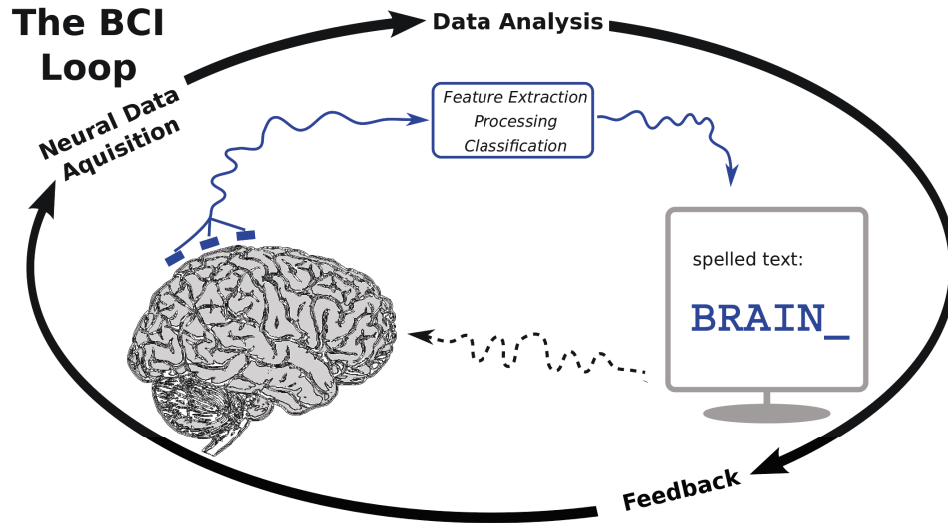
Figure 1.1: *The BCI feedback loop.*

BCI to become more effective and more user-friendly than previous approaches. However the studies with healthy subjects reveal shortcomings of state-to-the-art analysis techniques. Thus, new methods are derived which improve the classification accuracy for neuroimaging data – in particular for BCI data. Moreover, a motor imagery based BCI system is tested with patients, showing that individuals with severe motor impairments are able to gain significant control.

1.1 Outline of this Thesis

This section gives a brief outline of the remaining chapters of this thesis. Substantial parts of this thesis were published previously in peer-reviewed journals and conferences. The corresponding publications are referred by numbers and listed in the following section.

Chapter 2 provides the background information, which is essential to pursue the content of this dissertation. Firstly, the basic principles of the generation and acquisition of EEG data are introduced. The most relevant EEG signals for BCI control (i.e. Event Related Potentials (ERPs) and the sensorimotor rhythm (SMR)) are discussed and the technical core of a BCI system is explained in detail. Moreover, details of the fundamental algorithms and procedures for feature extraction, classification and artifact rejection are discussed. This chapter also introduces existing BCI paradigms which are most relevant for this thesis. Finally, the requirements and needs of a successful BCI application with patients are examined.

Chapter 3 presents experimental work with healthy subjects. Two auditory BCI spelling paradigms – named PASS2D and CharStreamer – are introduced [1, 3]. Both paradigms are based on ERPs and aim to provide a fast and intuitive BCI spelling application to the user. This is achieved by shifting complexity from the user to the BCI system, such that the user is confronted with a simple interface while the internal data processing pipeline deals with an increased amount of complexity in the data. Additionally, two offline studies investigate how stimulus properties impact the performance and usability of a BCI. Therefore, the use of naturalistic auditory stimuli is compared against artificial tones within the PASS2D paradigm [2]. The impact of the stimulation speed is analyzed for a simple auditory oddball paradigm [6].

Chapter 4 contains the core methodological contributions of this thesis. A novel classification

method, called "Relevance Subclass LDA" (RSLDA) is introduced, which is optimized for binary classification problems in the presence of additional label information [4,7]. This method is motivated by the classification problem of BCI data, as the studies described in Chapter 3 found ERPs to exhibit an internal subclass structure. Relevance Subclass LDA exploits such subclass structure and while being computationally highly efficient, it is also shown that RSLDA outperforms state-of-the-art methods for both, fMRI data and BCI data based on ERPs.

Chapter 5 describes a motor imagery study with severely motor impaired individuals [5]. It is shown that the application of machine learning methods allows to set up a highly flexible BCI system for patients with severe paralysis within a very small number of sessions. The individual needs and preferences of each patient are addressed by a user-centered design approach, which comprises of automatically adapting classifiers, as well as hybrid data processing and classification techniques. This study also describes one patient, for whom the BCI control could outperform his existing assistive technology solution in terms of accuracy, reaction times and information transfer. Therefore, this work reveals that the neuronal pattern detection of an attempted motor execution can be faster than the muscular output. This finding can be considered a significant success in the field of brain-computer interfacing research.

Chapter 6 summarizes the findings and discusses the impact of the work presented in this thesis.

1.2 List of Author Contributions

As it was mentioned above, significant parts of this thesis have previously been published in peer-reviewed journals and conferences. The following subsections list those articles, divided into main contributions and and additional contributions. The number of citations was specified with Google Scholar on the 12th of October 2014 for all articles that have been published for at least 6 months and which have been cited at least 5 times.

1.2.1 Main Contributions

Journal Articles

[1] Höhne J, Schreuder M, Blankertz B, Tangermann M (2011a). "A novel 9-class auditory ERP paradigm driving a predictive text entry system". In: *Front Neuroscience* 5, p. 99, **cited 70 times**

[2] Höhne J, Krenzlin K, Dähne S, Tangermann M (2012). "Natural Stimuli improve Auditory BCIs with respect to Ergonomics and Performance". In: *J Neural Eng* 9.4, p. 045003, **cited 23 times**

[3] Höhne J, Tangermann M (2014). "Towards User-Friendly Spelling with an Auditory Brain-Computer Interface: The CharStreamer Paradigm". In: *PLoS ONE* 9.6, e98322

[4] Höhne J, Holz EM, Staiger-Sälzer P, Müller KR, Kübler A, Tangermann M (2014c). "Motor Imagery for Severely Motor-Impaired Patients: Evidence for Brain-Computer Interfacing as Superior Control Solution". In: *PLoS ONE* 9.8, e104854

Journal Articles in Preparation

[5] Höhne J, Bartz D, Müller KR, Blankertz B (2014a). "Analyzing Neuroimaging Data with Subclasses: a Shrinkage Approach". In: in preparation

Peer-reviewed Conference Articles

[6] <u>Höhne J</u>, Tangermann M (2012). "How stimulation speed affects Event-Related Potentials and BCI performance". In: *Conf Proc IEEE Eng Med Biol Soc*. Vol. 2012. IEEE, pp. 1802–1805

[7] <u>Höhne J</u>, Blankertz B, Müller KR, Bartz D (2014b). "Mean shrinkage improves the classification of ERP signals by exploiting additional label information". In: *Proceedings of the 2014 International Workshop on Pattern Recognition in Neuroimaging*. IEEE Computer Society, pp. 1–4

1.2.2 Additional Contributions in Chronological Order

Journal Articles and Book Chapters

[8] Quek M, <u>Höhne J</u>, Murray-Smith R, Tangermann M (2012). "Designing future BCIs: Beyond the bit rate". In: *Towards Practical Brain-Computer Interfaces*. Ed. by BZ Allison, S Dunne, R Leeb, J del R. Millán, and A Nijholt. Berlin Heidelberg: Springer, pp. 173–196, **cited 6 times**

[9] Schreuder M, <u>Höhne J</u>, Blankertz B, Haufe S, Dickhaus T, Tangermann M (2013a). "Optimizing ERP Based BCI - a Systematic Evaluation of Dynamic Stopping Methods". In: *J Neural Eng* 10.3, p. 036025, **cited 19 times**

[10] Dähne S, Meinecke FC, Haufe S, <u>Höhne J</u>, Tangermann M, Müller KR, Nikulin VV (2014b). "SPoC: a novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters". In: *Neuroimage* 86.0, pp. 111–122, **cited 9 times**

[11] Holz EM, <u>Höhne J</u>, Staiger-Sälzer P, Tangermann M, Kübler A (2013b). "Brain-computer interface controlled gaming: Evaluation of usability by severely motor restricted end-users". In: *Artificial Intelligence in Medicine* 59.2. Special Issue: Brain-computer interfacing, pp. 111 –120, **cited 14 times**

[12] An X, <u>Höhne J</u>, Ming D, Blankertz B (2014). "Exploring Combinations of Auditory and Visual Stimuli for Gaze-Independent Brain-Computer Interfaces". In: *PLoS ONE* 9.10, e111070

[13] Bartz D, <u>Höhne J</u>, Müller KR (2014). "Multi-Target Shrinkage". submitted - available on arXiv

[14] Venthur B, Dähne S, <u>Höhne J</u>, Heller H, Blankertz B (2014). "Wyrm: A Brain-Computer Interface Toolbox in Python". In: *Neuroinformatics* in review

[15] Castano-Candamil JS, <u>Höhne J</u>, Castellanos-Dominguez G, Haufe S (2015). "Solving the EEG Inverse Problem based on Space-Time-Frequency Structured Sparsity Constraints". in review

Peer-reviewed Conference Articles

[16] <u>Höhne J</u>, Schreuder M, Blankertz B, Tangermann M (2010). "Two-dimensional auditory P300 Speller with predictive text system". In: *Conf Proc IEEE Eng Med Biol Soc*. Vol. 2010, pp. 4185–4188, **cited 43 times**

[17] <u>Höhne J</u>, Tangermann M (2011a). "Natural stimuli for auditory BCI". in: *Neurosc Let*. Vol. 500. Supplement 1, e11

[18] <u>Höhne J</u>, Tangermann M (2011b). "Stimulation Speed Boosts Auditory BCI Performance". In: *Proc. 5th Int. BCI Conf. Graz*. Ed. by GR Müller-Putz, R Scherer, M Billinger, A Kreilinger, V Kaiser, and C Neuper. Graz: Verlag der Technischen Universität Graz, pp. 16–20

[19] Höhne J, Schreuder M, Blankertz B, Müller KR, Tangermann M (2011b). "Novel Paradigms for Auditory ERP Spellers with Spatial Hearing: Two Online Studies". In: *Int J Bioelectromagnetism* 13.2, pp. 96–97

[20] Dähne S, Höhne J, Schreuder M, Tangermann M (2011b). "Slow Feature Analysis - A Tool for Extraction of Discriminating Event-Related Potentials in Brain-Computer Interfaces". In: *Artificial Neural Networks and Machine Learning - ICANN 2011*. Ed. by T Honkela, W Duch, M Girolami, and S Kaski. Vol. 6791. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 36–43

[21] Dähne S, Höhne J, Tangermann M (2011a). "Adaptive Classification Improves Control Performance In ERP-Based BCIs". In: *Proceedings of the 5th International BCI Conference*. Graz, pp. 92–95, **cited 9 times**

[22] Schreuder M, Höhne J, Treder MS, Blankertz B, Tangermann M (2011b). "Performance Optimization of ERP-Based BCIs Using Dynamic Stopping". In: *Conf Proc IEEE Eng Med Biol Soc*, pp. 4580–4583, **cited 19 times**

[23] Tangermann M, Höhne J, Schreuder M, Sagebaum M, Blankertz B, Ramsay A, Murray-Smith R (2011a). "Data Driven Neuroergonomic Optimization of BCI Stimuli". In: *Proc. 5th Int. BCI Conf. Graz*. Graz, pp. 160–163

[24] Tangermann M, Schreuder M, Dähne S, Höhne J, Regler S, Ramsay A, Quek M, Williamson J, Murray-Smith R (2011b). "Optimized Stimulation Events for a Visual ERP BCI". in: *Int J Bioelectromagnetism* 13.3, pp. 119–120, **cited 17 times**

[25] Tangermann M, Höhne J, Stecher H, Schreuder M (2012a). "No Surprise — Fixed Sequence Event-Related Potentials for Brain-Computer Interfaces". In: *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, pp. 2501–2504, **cited 5 times**

[26] Holz EM, Zickler C, Riccio A, Höhne J, Cincotti F, Tangermann M, Halder S, Mattia D, Kübler A (2013a). "Evaluation of Four Different BCI Prototypes by Severely Motor-Restricted End-Users". In: *Proceedings of the Fifth International Brain-Computer Interface Meeting 2013*. Ed. by J d. R. Millán, S Gao, GR Müller-Putz, JR Wolpaw, and JE Huggins. Verlag der Technischen Universität Graz, pp. 362–363

1.3 List of Abbreviations

- ALS: Amyotrophic Lateral Sclerosis

- AT: Assistive Technology

- AUC: Area Under the ROC curve

- BCI: Brain-Computer Interface

- CSP: Common Spatial Patterns

- EEG: Electroencephalogram

- ERP: Event Related Potential

- fMRI: functional Magnetic Resonance Imaging

- ICA: Independent Component Analysis

- ITR: Information Transfer Rate

- LDA: Linear Discriminant Analysis

- MI: Motor Imagery

- PASS2D: Predictive Auditory Spatial Speller with two-dimensional stimuli

- PyFF: Pythonic Feedback Framework

- RLDA: Regularized Linear Discriminant Analysis

- ROC: Receiver Operating Characteristic

- SMR: Sensorimotor Rhythms

- SOA: Stimulus Onset Asynchrony

- ssAUC: signed and scaled Area Under the ROC curve

FUNDAMENTALS IN BRAIN-COMPUTER INTERFACING

2.1 Neurophysiology of EEG signals

The electroencephalographic (EEG) signal is an electric potential that is measured on the scalp. Thus, in order to acquire EEG signals, electrodes are placed on the head and voltage fluctuations are recorded in the range of micro Volts. Fig. 2.1 depicts a standard electrode setup on the scalp.

Professor Hans Berger was the first scientist, who described the EEG as a tool to measure human brain activity (Berger, 1929). The EEG is generated by electric activity of (mainly) cortical neurons. The neuronal signal is however superimposed by several types of physiological and non-physiological artifacts. In the following, the mechanisms underlying the transformation of cerebral electrical activity to EEG potentials are briefly described. A thorough introduction in the field of EEG signal generation is given in Neuroscience textbooks (Kandel et al., 2000).
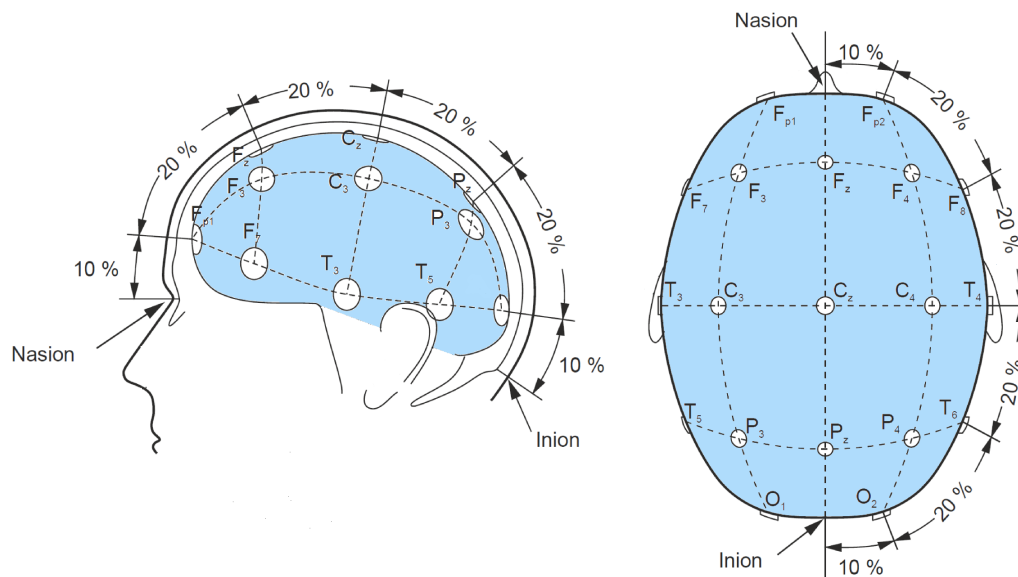
Figure 2.1: *Visualization of the EEG electrode placement on the scalp, corresponding to the 10-20 system. Figure modified from (Nicolas-Alonso and Gomez-Gil, 2012), with permission.*

2.1.1 EEG Signal Generation

Neurons are the core functional units in the brain. The human brain comprises of approximately 100 billion neurons, which are heavily interconnected with each other. Neuron cells have ionic pumps in their membranes. Ionic pumps are transmembrane proteins which actively transport molecules through the membrane. Due to their activity, a high concentration of potassium (K^+) is maintained inside the cell. Moreover, an increased concentration of sodium (Na^+) and calcium (Ca^{2+}) is maintained outside of the cell. In the resting state of the cell, such concentrations lead to a voltage difference of approximately -70 mV across the membrane – known as the *resting potential*. Information processing within and between neurons is performed by manipulating the resting potential, leading to action potentials. Action potentials can propagate throughout the axon of a neuron and stimulate/inhibit neighboring neurons by generating an excitatory/inhibitory postsynaptic potential (EPSP/IPSP) at the synapse. While the action potential lasts less than two milliseconds, an EPSP/IPSP can last tens of milliseconds, allowing a temporal and spatial summation. If a large population of neurons is simultaneously active, the summation of action potentials or the EPSP/IPSP potentials can form an electrical dipole in the brain. Depending on its strength, location and orientation, such a dipole might also be measurable as a scalp potential in the EEG.

Generally, it should be noted that compared to the electric potential on the cell membrane level (range of millivolts, mV), EEG potentials on the scalp are measured in the range of microvolts, μV, thus weakened by factor 1000. Due to varying electrical conductivity properties throughout the brain, the scull and the skin, the projection of neuronal dipoles to a scalp potential represents a highly complicated mathematical problem, called EEG forward problem (Baillet et al., 2001). However, it is known that EEG signals are mainly generated by EPSP/IPSP potentials of pyramidal cells. Pyramidal cells are cortical neurons that are aligned orthogonal to the scalp.

2.2 Neural Signals that Enable BCI Control

The EEG contains several types of neural signals that can be exploited by a BCI system. Generally, these discriminative EEG signals can be divided in two categories: (I) signals that are elicited in response to an external stimulus and (II) signals that are self-elicited. Event related potentials (ERPs) and steady state evoked potentials (SSEPs) are the most commonly used signals for BCIs that are based on external stimulation. BCI paradigms which rely on such EEG features are also called "synchronous BCI" systems. The other type of brain signals does not require external stimulation. Instead, spectral brain activity that is associated to a mental state can be identified directly. One EEG signal, which is frequently used for BCI, is the sensorimotor rhythm (SMR). The SMR can be controlled voluntarily by performing and also imagining motor movements. Thus, a BCI system is driven by perturbations of the SMR, which are also considered as event related desynchronizations (ERDs). Thus, one can apply "motor imagery" paradigms, in which ERDs are induced by a user who imagines movements of the left and right hand. Such motor imagery (MI) BCI paradigms are also called "asynchronous BCI" systems, as they operate independently of any external trigger.

In the following paragraphs, ERPs as well as the SMR will be introduced with more detail, as both concepts resemble important aspects throughout this thesis.

2.2.1 Event Related Potentials

The term event related potential (ERP) refers to voltage fluctuations in the EEG which are triggered by an event. Such events can be exogenous (i.e. externally generated) or endogenous (i.e. initiated internally). Examples for exogenous events are stimuli from the visual, auditory or tactile domain,

such that a subject is listening to auditory stimuli, or perceiving visual stimuli. An example for an endogenous event is when the subject decides to performs a motor action (e.g. moving the left hand). All these events trigger cascades of brain activity which result in voltage fluctuations in the EEG, being measured as ERPs.

As the brain response to an event often initiates cascades of ERP components with short duration ($i.e.$ $< 10ms$), it is important that the precise onset of the event is known. Moreover, ERPs are superimposed by neural activity which is unrelated to the event (also called "biological noise" or "background activity"). Additionally, physiological (e.g. heart beat) and non-physiological (e.g. 50 Hz line noise) artifacts might be present in the data. In order to average out such background activity and thus increase the signal to noise ratio (SNR), the EEG response is averaged over numerous (e.g. ≥ 100) events.

Various aspects of ERPs have been extensively studied within the research field of Psychophysics. This research area mainly investigates the relationship between a physical stimulus and its perception. Based on highly standardized and controlled studies, ERP components were analyzed corresponding to their temporal and spatial properties (Hillyard et al., 1973; Polich, 1989; Pritchard et al., 1991). Moreover, the impact of various properties such as stimulus intensity and timing were researched (Näätänen et al., 1981; Polich, 1989; Polich et al., 1996; Gonsalvez and Polich, 2002). Based on this research, a common terminology was established, describing an ERP with its polarity (P/N for positive/negative) and its latency after the event (time in ms).

BCI research is typically focusing on ERP components which are modulated by the subjects' attention. Thus, the ERP response to attended stimuli (referred to as "targets") differs from the ERP response to stimuli which the user is not attending to ("non-targets"). The ERP-based BCI can exploit this difference in the brain response to each stimulus. When repetitively presenting several stimuli and analyzing the corresponding ERP responses, the BCI is able to uncover to which stimulus the user is attending to.

N200 and P300

There are two ERP components which are mainly used to drive a BCI: N200 and P300. Both components were first described in Psychophysics literature by Sutton et al. (1965). There is a vast amount of literature, analyzing such components with respect to various aspects (Pritchard et al., 1991; Polich et al., 1996; Sellers et al., 2006a; Hill et al., 2004). In the following, both components are briefly described.

N200   The N200 component is a negative deflection in the ERP that occurs about 200 ms after stimulus onset. It lasts about 40-100 ms and arises within the neural processing in cortical brain areas (Hillyard et al., 1973; Pritchard et al., 1991). It is induced by those brain areas which are involved in the modality-specific stimulus processing. Thus, the N200 component for visual stimuli arises in the visual cortex, while auditory stimuli evoke a N200 which originates from the auditory cortex. Therefore, the spatial pattern of a visual N200 and an auditory N200 are substantially different. The N200 component can be subdivided into several subcomponents: a Mismatch Negativity (also called MMN or N2a), an attentional component (also called N2b) and a classification component (N2c) – for details, see Pritchard et al. (1991) and Näätänen et al. (2007, 1978). The attentional N2b is most interesting for BCI applications. Since each subcomponent contributes to the N200 response, targets as well as non-targets typically evoke a negative deflection in the EEG. However, target stimuli evoke a N200 which is more negative and eventually also longer-lasting than the N200 of non-targets – see Fig. 2.2A.

P300 The P300 component is a positive deflection in the EEG which occurs with a latency of approximately 300–500 ms after stimulus onset. It has stronger amplitude than the N200 and it may last for up to 400 ms. The P300 arises from a central cortical activity which is independent of the stimulus modality. Therefore it is also described to be the "A-HA response" in the brain (Kübler et al., 2009). Being thoroughly investigated in Psychophysics literature (for a review, see Picton (1992) and Polich (2007)), the P300 component was found to be a highly complex cascade of neural activity. The P300 can also be subdivided into several subcomponents: the novelty component (P3a) and an attentional component (P3b) have been widely studied in various scenarios (Polich, 2007). It was found that the P300 component depends on many experimental factors, such as the intensity and duration of a stimulus as well as the target-to-target interval (Polich, 1989; Picton, 1992; Gonsalvez and Polich, 2002; Allison and Pineda, 2003; Gonsalvez et al., 2007). Moreover, the first ERP-based BCI (see Farwell and Donchin (1988)) exploited the P300 component. Hence, ERP-based BCIs are also referred to as "P300 based BCI".

Habituation As mentioned above, stimulus properties as well as other experimental factors can highly impact the pattern, latency or amplitude of ERP components. This holds also for the N200 and the P300 component. This can be illustrated with two simple examples.

1. Gonsalvez and Polich (2002) found that the amplitude of the P300 (more precisely the P3b component) is correlated with the target-to-target interval. Thus, an elongation of the time between two target events yields an intensified P300 component.

2. Humans are highly trained to detect and analyze a facial expression. In comparison to other visual stimuli, the ERP response to a face stimulus is stronger. Moreover, additional face-specific ERP components such as N400f can be found in the EEG (Bentin and Deouell, 2000).

Based on the findings in the Psychophysics literature, various studies investigated suitable stimuli that could optimally drive a BCI. This was done for the visual domain (Hill et al., 2009; Kaufmann et al., 2011) and recently also for the auditory domain (Matsumoto et al., 2013; Lopez-Gordo et al., 2012; Käthner et al., 2012). Within this thesis, Section 3.2 discusses the usage of optimized auditory stimuli in detail, describing one of the first successful studies that utilized non-artificial stimuli for auditory BCI.

Visualization of ERPs

Depending on the individual background of the researcher, there are different conventions of how to visualize ERP. Therefore the following paragraph describes how ERPs are visualized within this thesis.

Generally, the traces of selected EEG electrodes are plotted as time series. Such traces depict the average amplitude modulation which is recorded in response to an event (such as an auditory stimulus). The x-axis shows the temporal course in milliseconds [ms] with $t = 0$ defining the onset of the stimulus. The y-axis displays the amplitude modulation in μV, with positive deflections plotted upwards (in Psychophysics, there is a convention to flip the y-axis). In order to visualize the difference between two conditions, an EEG trace is plotted for each condition. Fig. 2.2A shows an exemplary ERP response with two traces per channel, with the color of the traces coding for the class (i.e. target/non-target). It is common practice to also depict the spatial patterns of ERP components for each class with a scalp plot – Fig. 2.2C. To generate such a scalp plot, the ERP response is averaged in a manually defined time interval (e.g. 300-500 ms) for each channel.

The class discriminative information between targets and non-targets is commonly computed and visualized. Therefore, ssAUC values or signed r^2 values (both introduced in Section 2.4.5) are used within this thesis. Such univariate measures for class separability are often more informative than the

difference of the means, which is plotted with the ERP traces and scalp plots. Thus, they are plotted additionally in order to facilitate the interpretation of differences in the ERP in two conditions. A color bar below the EEG traces (see Fig. 2.2B) visualizes the temporal distribution of class separability for one channel. The spatial distribution of class separability can be depicted as scalp plots – see Fig. 2.2D.
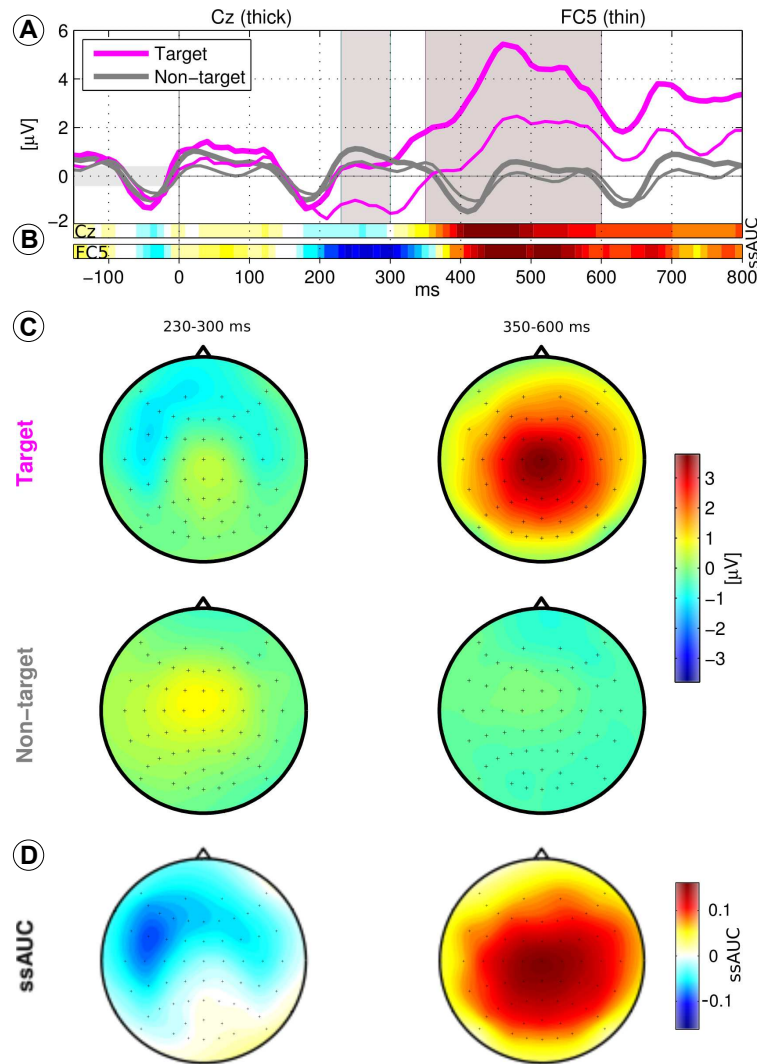
Figure 2.2: *Explanation of how to visualize ERPs.* **(A)** *The traces show the ERPs at electrodes Cz (thick lines) and FC5 (thin lines). The ssAUC bars* **(B)** *quantify the discriminative information for the two channels. The averaged ERP scalp maps of target and non-target stimuli for the two marked time intervals are shown in* **(C)**. *The spatial distribution of class discrimination (ssAUC values) is depicted in* **(D)**. *The scales for* **(B)** *and* **(D)** *are equal. The plot shows a grand average response of auditory BCI experiment. Data was taken from Experiment 4. Note that the repetitive pattern in the ERP arises from a fast sequence of stimuli, with one auditory stimulus every 225 ms.*

2.2.2 Sensorimotor Rhythm and Event Related Desynchronization

Rhythmic activity of neural sources can also be measured in the EEG. Such oscillations are usually divided in specific frequency bands that are associated with functional behavior. The strongest (i.e. showing the highest amplitudes) and most famous neural rhythm in the EEG is the α rhythm, which is the idle rhythm arising from the visual cortex. The α rhythm was also the first brain signal which was found in the EEG by Berger (1929). For adults, the α rhythm is typically observed in the 8-13 Hz frequency band. It is strongest, when the subject is relaxed while closing the eyes.

Another brain rhythm which is highly relevant for BCI application arises from the motor cortex: the sensorimotor rhythm (SMR). In analogy to the α rhythm, an increased SMR amplitude is observed when the corresponding sensorimotor area is in an idle state. The SMR can be generated across the entire motor cortex. This leads to variations in its spatial distribution, depending on which motor areas are active/inactive. It can moreover be subdivided into several spectral components, with the μ rhythm being the strongest. The μ rhythm is considerably weaker than the α rhythm, such that spatial filters (see Section 2.3.2 for details) might be required to visualize and analyze the μ rhythm. However, the μ rhythm is also observed in the 8-13 Hz frequency band for most adults. It is strongest (i.e. showing the highest amplitudes), when the subject is relaxed and not involved with any motor activity. It is important to note that both, motor execution (muscular activity) and motor imagery (imagining a motor action) are processed in the motor cortex. Therefore, the SMR (e.g. the μ rhythm) desynchronizes when the subject is performing motor execution or motor imagery. Moreover, focal areas of the motor cortex can be the desynchronized while other areas are synchronized. Such differences lead to distinct spatial signatures which can be associated to different parts of the body (left hand / right hand / foot).

It is a common BCI scenario to exploit those spatial and spectral signatures of the SMR for a BCI paradigm which is based on motor imagery (MI). Therefore, the BCI might evaluate the μ rhythm of two motor areas, such as the right hand and left hand. The user then controls the BCI by imaging movements with the left or the right hand while the BCI is evaluating fluctuations in the μ rhythm arising from the corresponding motor cortices.

2.3 The Online BCI Loop

In order to set up a reliable BCI system, it is highly important to extract the relevant information from the continuous stream of EEG data. Therefore, an elaborate signal processing is required in order to enhance the signals of interest while suppressing of the rest of the "cerebral cocktail party" signals in real-time (Müller and Blankertz, 2006).

The technical processing pipeline of any BCI can be divided into five steps: (1) Data Acquisition, (2) Preprocessing (3) Feature Extraction, (4) Classification and (5) Feedback – see also Fig. 1.1. Within this thesis, two types of BCI systems are discussed: a BCI based on event related potentials and a BCI based on motor imagery. The following paragraphs discuss the most relevant processing and analysis steps (steps 2–4) for both types of BCI systems.

2.3.1 Processing Steps for ERPs

The following paragraph describes the standard data processing steps which are performed for a BCI based on ERPs. A detailed review of the state-of-the-art data processing of ERPs is also given in Blankertz et al. (2011). One should note that the processing steps are equal for offline analysis and
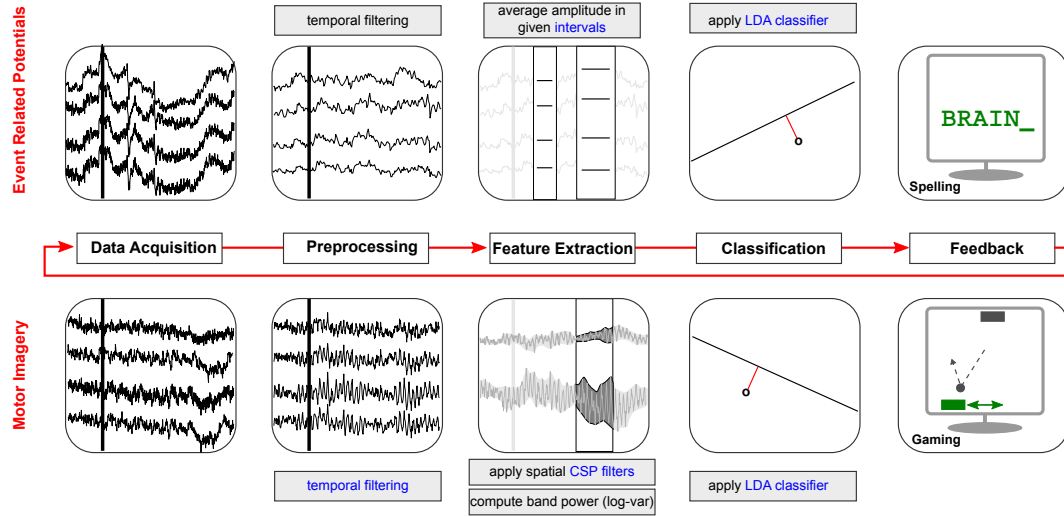
Figure 2.3: *The five major steps in the online BCI loop based on ERPs (top) and motor imagery (bottom). Note that various applications can be driven with both approaches.*

online BCI application. However, the time intervals and the classifier weights are determined within the offline analysis.

Preprocessing: Temporal Filtering and Subsampling   As the attention-related ERPs (e.g. N200 and P300) are found in a rather low-frequency domain (approx. 0.5-12 Hz) the EEG data is first filtered with a 20 Hz low-pass filter (Chebyshev filter, order 5) and a 0.3 Hz high-pass filter (Butterworth filter, order 5). This temporal filtering is advisable as one increases the signal to noise ratio by filtering out irrelevant information. However, temporal filters are not essential in a scenario with a high signal-to-noise ratio. Moreover, such filters should be chosen with care, as they may introduce filter artifacts and phase shifts. Those phase shifts may lead to controversial conclusions when interpreting ERP components. However, phase shifts can be canceled out by applying the filter from both directions, forwards and backwards. As the backwards filtering is only applicable for offline scenarios, such phase corrections are only performed when visualizing the ERP components – but not for online BCI application.
Note that state of the art EEG amplifiers have a default sampling rate of up to 1000 Hz. Therefore, signals are commonly subsampled to 100 Hz after filtering.

Feature Extraction: Computing ERP Amplitudes   For feature extraction, the ERP amplitudes are averaged in given time intervals for each channel. This is also illustrated in Fig. 2.3. This processing step results in a data point $x \in \mathbb{R}^{1 \times d}$ with $d = n_{channels} \times n_{ival}$ for each epoch. Such intervals have to be specified prior to the online BCI loop.
There are several ways to determine such intervals. The experimenter could manually inspect the ERP responses and choose intervals which contain discriminative ERP components. A second strategy is to use a heuristic to choose discriminative time intervals. A third strategy (also called "subsampling") is to use a dense array of small (e.g. 40 ms) intervals, disregarding any discriminative information. While the subsampling approach is simple to implement, it will produce a very high dimensional feature space which is unfavorable if there is only limited amount of calibration data available. All the three above mentioned approaches intervals selection are commonly used for ERP-based BCIs.

2.3.2 Processing Steps for Motor Imagery Features

The following paragraph describes the standard data processing steps which are performed for a BCI based on mental imagery. Note that the processing steps are equal for offline analysis and online BCI application.

Preprocessing: Temporal Filtering When extracting bandpower features from the EEG, temporal filtering is the first crucial step. It is common practice to first determine a band-pass filter based on analyzing calibration data. The frequency band which contains the most discriminative spectral features is therefore chosen. For motor imagery tasks, those features are mostly found within the μ-band (8-13 Hz) or the β-band (15-30 Hz). The resulting filters (e.g. order 5 Butterworth filter, 10-12 Hz) extract a rather narrow band signal which is further processed in the feature extraction step. There are also alternative approaches that apply not only one band-pass filter, but a predefined set of band-pass filters (also called "filter bank") in parallel – see Ang et al. (2009) and Ang et al. (2012) for details.

Feature extraction: Applying CSP filters For feature extraction, one can apply spatial CSP filters in order to extract class discriminative sources. The bandpower of the extracted sources is then considered as features. There are several approaches to estimate the bandpower of a signal. Within BCI research, the bandpower is commonly estimated by computing the $log(var(\boldsymbol{x}))$ – thus taking the logarithm of the variance of the band-pass filtered signals.

Common Spatial Patterns (CSP)

Common Spatial Patterns (CSP) is a popular signal processing algorithm that is commonly applied for BCIs based on mental imagery. CSP was first described by Ramoser et al. (2000) while its methodology was discussed in numerous review papers for data processing in BCI (Lotte et al., 2007; Blankertz et al., 2008b; Lemm et al., 2011). The goal of CSP is to determine spatial filters that optimally capture modulations of class-discriminative brain rhythms. CSP filters extract oscillations from the signal that feature a distinctive band power for two classes. From a more technical perspective, CSP finds the spatial projection from band-pass filtered EEG data such that the difference between the variances of the projected data for two classes is maximized. As it was mentioned above, the log-variance of band-pass filtered signals is an estimator of the band power.

As it is discussed in Blankertz et al. (2008b), there are several ways to formulate the CSP problem. In the following, the "discriminative view" is briefly described. Assume $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ to be the covariance matrices of the band-passed filtered EEG signals for left-hand and right-hand motor imagery. Then the class discriminative activity $\boldsymbol{S}_d$ and the common activity $\boldsymbol{S}_c$ can be determined with

$$\boldsymbol{S}_d \;\; = \;\; \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2 \tag{2.1}$$

$$\boldsymbol{S}_c \;\; = \;\; \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2. \tag{2.2}$$

CSP aims to find the filters which maximize the ratio of discriminative activity and common activity,

$$\max_{\boldsymbol{w} \in \mathbb{R}^c} \frac{\boldsymbol{w}' \boldsymbol{S}_d \boldsymbol{w}}{\boldsymbol{w}' \boldsymbol{S}_c \boldsymbol{w}} \tag{2.3}$$

The solution of this problem can be determined by solving the generalized eigenvalue problem

$$S_d \boldsymbol{w} = \boldsymbol{\Lambda} S_c \boldsymbol{w}. \tag{2.4}$$

The generalized eigenvalue decomposition yields a set of eigenvectors $\boldsymbol{w}_i$ and eigenvalues λ_i with $i \in \{1 \dots n_{channels}\}$. Those eigenvectors $\boldsymbol{w}_i$ with a large positive (or negative) eigenvalue ($|\lambda_i| \gg 0$) project the data to the class discriminative directions. Therefore, it is the common practice to use several eigenvectors from both ends of the eigenvalue spectrum as spatial filters. An example of projected EEG data is shown in Fig. 2.4.

Conceptually, it is important to note that CSP finds a linear projection, which is applied before estimating the band-power of the projected data (band power estimation is a nonlinear step). The order of first applying the linear and then the non-linear processing step is crucial – for a detailed discussion, see Dähne et al. (2014b) and Haufe et al. (2014).
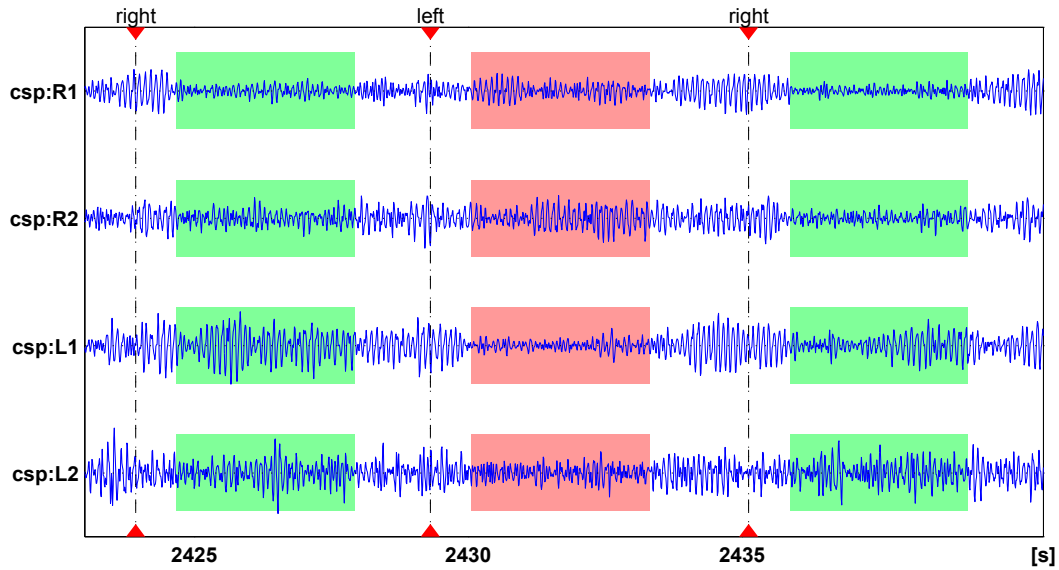
Figure 2.4: *Illustration for EEG data which is projected on CSP filters. A user performs left hand and right hand motor imagery for 4 seconds. The output of two CSP filters for each class is shown. The spectral µ rhythm desynchronizes when the user imagines moving the corresponding hand. Figure was taken from Blankertz et al. (2008b), with permission.*

Extensions of CSP CSP filters are commonly applied as a preprocessing step for BCIs which are based on the modulation of brain rhythms. CSP is popular for such applications as it leads to a high signal-to-noise ratio (respectively a high classification performance) while being computationally efficient and simple to implement. However, the major disadvantage of the standard CSP algorithm is that it is sensitive to noise and non-stationarities in the data. Moreover, CSP requires a considerate amount of training data, as it requires accurate estimators of the covariance matrices. This might lead to a long training/calibration phase within a BCI experiment. In order to approach the above mentioned shortcomings, several modifications and extensions have been developed. Most extensions add prior knowledge into the algorithm, which is formalized as a regularization term in the nominator or denominator of Equation 2.3. Such optimized CSP algorithms can be more invariant and robust to noise (Blankertz et al., 2008a; Kawanabe et al., 2014) and non-stationarities (Samek et al., 2012, 2014). Lemm et al. (2005) proposed spatio-spectral CSP filters, while Sannelli et al. (2011) introduced CSP Patches, which is a CSP variant that can be applied with less training data. A generalization of

CSP, called Source Power Comodulation (SPoC) was described in Dähne et al. (2014b) and extended in Dähne et al. (2014a). SPoC finds optimal spatial filters to extract continuous power modulations.

2.4 Classification

After signal acquisition and feature processing, a classification method has to be applied in order to decode the user's intention in the BCI framework. The specification of the data (i.e. number and dimensionality of the data points) depends heavily on the type of BCI paradigm. In an ERP-based BCI paradigm, the typical classification task is to separate between brain responses to target and non-target stimuli. A rather high-dimensional data point (e.g. $x \in \mathbb{R}^{1 \times 600}$) is generated for each stimulus. For motor imagery paradigms, the classification task is to differentiate between different motor imagery conditions such as left hand vs. right hand. In contrast to ERP data, the feature processing for MI data already comprises a supervised step to enhance class separability and to lower dimensionality – i.e. applying CSP filters. Thus, the classification task for MI data is facilitated to low dimensional problem.

However, numerous studies have investigated various classification techniques for the BCI data – for detailed reviews, see Garrett et al. (2003), Lotte et al. (2007), and Lemm et al. (2011). Although several highly elaborate, non-linear methods were proposed (Tomioka and Müller, 2010; Müller et al., 2003), most comparative studies found LDA with shrinkage regularization be amongst the best performing methods.

2.4.1 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a simple and powerful classification method which is frequently applied for BCI data. LDA is based on the following three assumptions:

- Data of each class are Gaussian distributed.

- Gaussians of all classes have the same covariance matrix.

- True class distributions (μ_i, Σ) are known.

If all three assumptions are met, LDA yields the Bayes' optimal classifier. Fig. 2.6 shows several 2D toy examples for LDA classification. Only for plot A, the above mentioned assumptions are met.

LDA seeks a linear projection w such that within-class variance is minimized while the between-class variance is maximized. Therefore, the LDA classifier maximizes the objective function

$$J(w) = \frac{w' S_B w}{w' S_W w},\tag{2.5}$$

with S_B and S_W specifying the between-class and within-class variance respectively. This general formulation is also called "Fisher Discriminant Analysis (FDA)". For a c-class problem, this Rayleigh coefficient can be maximized by solving the generalized Eigenvalue problem, as it was shown for CSP in Equation 2.4. For the two-class scenario, it can however be shown that the optimal projection w can be determined by

$$w = C^{-1}(\mu_1 - \mu_2).\tag{2.6}$$

Thus, in order to compute an LDA classifier, the class means μ_1 and μ_2 as well as the class-wise covariance C have to be estimated.
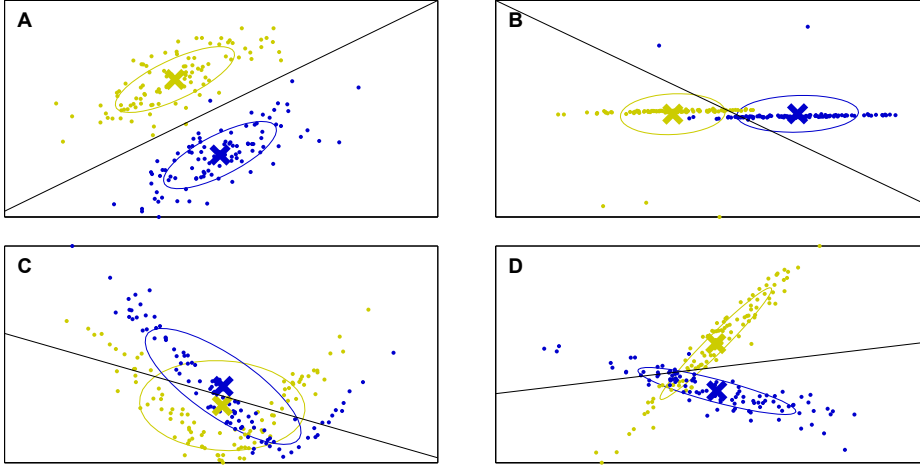
Figure 2.5: *Illustration of classification with LDA. Each scatter plot shows two-dimensional data from two classes (blue and yellow). Class means are marked with bold crosses and the class-wise covariances are depicted as ellipses. The LDA separation hyperplane is plotted as a black line. Plot A shows Gaussian distributed data, which can be well separated by LDA. Plot B shows two Gaussians which are contaminated by outliers. Plot C shows data which do not follow a Gaussian distribution and which are not suitable for LDA. Plot D depicts an scenario in which the covariance structures are substantially differing between the two Gaussian classes.*

2.4.2 Shrinkage Estimation of the Covariance

The sample estimate of the covariance matrix is an unbiased estimator with good properties in favorable conditions. However, the sample estimator might be distorted and unsuitable, if data are high dimensional and only a limited amount of data points are available. It is known, that this curse of dimensionality leads to sample estimates C^s of the unknown covariance C with a systematical distortion: directions with high variance are over-estimated, while low-variance directions are under estimated. For BCI data, this issue mainly affects LDA classifiers which are trained for ERP data – see Blankertz et al. (2011) for further discussions. In order to compensate for such distortions, one can introduce a regularization term when estimating the covariance

$$C^{reg}(\lambda) = (1 - \lambda)C^s + \lambda \nu I, \tag{2.7}$$

with λ and ν being regularization and scaling parameters and $\lambda \leq 1$. While choosing those parameters by cross validation resembles a high computational effort, the shrinkage method serves an analytical solution to find an optimal regularization parameter λ (Ledoit and Wolf, 2004). Shrinkage seeks for an estimate of the covariance matrix, such that the expected mean squared error (EMSE) is minimized,

$$\lambda^* = \underset{\lambda}{\mathrm{argmin}} \, \mathbb{E}\left[\sum_{i,j} \left(C_{ij}^{reg}(\lambda) - C_{ij} \right)^2 \right]$$

$$= \frac{\sum_{i,j} \left\{ \mathrm{Var}\left(C_{ij}^s \right) - \mathrm{Cov}\left(C_{ij}^s, \nu I_{ij} \right) \right\}}{\sum_{i,j} \mathbb{E}\left[\left(C_{ij}^s - \nu I_{ij} \right)^2 \right]}. \tag{2.8}$$

Replacing the expectations with sample estimates yields an analytical formula for an estimator λ^*, being highly favorable as model selection through cross-validation is not required. The scaling ν is commonly defined as the average eigenvalue of C^s.

2.4.3 Adaptation of LDA Classifiers

For an online system such as a BCI, it is important to consider that signals may change over time. There are several reasons for the non-stationary nature of the signals, namely biological effects (e.g. level of concentration or muscle artifacts) and technical issues (e.g. electrode movements, dried out electrodes). Therefore, novel adaptive processing methods have recently been researched (Vidaurre et al., 2011a; Vidaurre et al., 2011c,b; Kindermans et al., 2014). The most common approach is to track the changes of the covariance matrix and account for such rotations and drifts. As the LDA classifier solely depend on the means and the covariance of the data (see Eq. (2.6)), adaptive LDA classifier weights can be obtained by

$$w_{n+1} = C_{n+1}^{-1}(\mu_1 - \mu_2). \tag{2.9}$$
$$\text{with } C_{n+1} = (1-\alpha)C_n + \alpha C_{segment}, \tag{2.10}$$

where $C_{segment}$ stands for the covariance of the new data segment.

Compared to the fixed classifier, there are two additional parameters required. The strength of the adaptation is parameterized by α and the segment size (the number data points/trials in one segment) has to be determined. Adaptive classifiers have been shown to improve BCI performance for both motor imagery paradigms (Vidaurre et al., 2011a) and ERP paradigms (Dähne et al., 2011a).

2.4.4 Advantages and Shortcomings of LDA

LDA with covariance shrinkage estimation exhibits several favorable properties. It leads to high accuracies for most BCI data sets and it is simple to implement while being fairly robust to estimation errors.

Being a linear method, LDA allows an interpretation of the sources which are driving the classifier (Müller et al., 2003; Haufe et al., 2014). In the BCI framework this can be highly valuable as such an interpretation can lead to conclusions regarding the spatial or temporal origin of the neural signals of interest. It is however important to remark that the weights of the filter w are not suitable for any interpretation, as w is also driven by noise sources. Nevertheless, numerous researchers have been interpreting the weights of classifiers as this might appear to be a straight-forward step. However, there is an intuitive explanation, why LDA filters should not be interpreted[1]: the filter not only exploits the signal of interest, it also suppresses the noise from the data. Therefore, a considerate part of the filter weights corresponds to the noise suppression. Thus, interpreting filter vectors might lead to an erroneous interpretation of noise sources. Instead, Haufe et al. (2014) showed that the difference of the means $(\mu_1 - \mu_2)$ resembles the activation pattern of an LDA classifier which is suitable for interpretation.

Despite the above mentioned advantages of shrinkage LDA for BCI applications and Neurotechnology in general, LDA is a linear classifier which is not able to uncover non-linear characteristics in the data – see Fig. 2.5. In order to account for non-linearities in the data, appropriate feature processing steps are essential in order to transform the data to follow a Gaussian distribution (Müller et al., 2003). Therefore, both technical and biological artifacts are excluded, as outliers can heavily impact the LDA classifier – see Fig. 2.5. Techniques for artifact removal in EEG are described in Section 2.5. For motor

---

[1]The same concept holds for any other linear filters such as CSP filters: the filters themselves are not suitable for interpretation.

imagery data, log-variance features are computed as estimates of the spectral power in order to obtain Gaussian distributed data.

2.4.5 Measuring Class Discriminative Information

Within this thesis, numerous types of features will be described, that allow to driving a BCI system. Generally, successful BCI control is based on the exploitation of class-discriminative EEG features. This section describes how to quantify the class discriminative information of each feature. Two univariate statistical measures are introduced which are commonly used for BCI research: signed r^2 values and ssAUC.

Signed r^2 values

Signed r^2 values resemble a specific modification of a correlation coefficient of two variables x and y. However, x is assumed to continuous, while y is expected to be dichotomous. The signed r^2 value specifies how much variance of the joint distribution x can be explained by class membership y. The computation of signed r^2 values is based on the biserial correlation coefficient $r(x, y)$,

$$sgn - r^2(\boldsymbol{x}, \boldsymbol{y}) \quad := \quad sign(r(\boldsymbol{x}, \boldsymbol{y})) \cdot r(\boldsymbol{x}, \boldsymbol{y})^2 \tag{2.11}$$

$$r(\boldsymbol{x}, \boldsymbol{y}) \quad := \quad \frac{\sqrt{N_1 \cdot N_2}}{N_1 + N_2} \frac{MEAN\{x_i | y_i = 1\} - MEAN\{x_i | y_i = 2\}}{STD\{\boldsymbol{x}\}} \tag{2.12}$$

The sign of the r-value determines whether a correlation (positive sign) or an anticorrelation (negative sign) is found in the data.

ssAUC

The signed and scaled Area Under the Curve (ssAUC) is another measure for class separability. It is based on the Receiver-Operator Characteristics, also called ROC curve (Green and Swets, 1966). The ROC curve is a graphical representation of the quality of a binary classifier, as it relates "sensitivity" as function of "1−specificity". In other words, the ROC curve is created by plotting the "true positive rate" versus the "false positive rate" for various threshold settings.

The area under the ROC curve (AUC) can be regarded as a way to reduce ROC curve to a single value, being the expected classification performance. However, the AUC does not provide information about the direction of an effect. The ssAUC therefore resembles a simple modification of the AUC which is signed and linearly scaled to the range of $[-1, 1]$.

In comparison to other methods such as the signed r^2 values, ssAUC is that it does not rely on the assumption that the distributions are Gaussian.

2.5 Dealing with Artifacts

EEG signals are very prone to artifacts and correcting for such artifacts is an important processing step when analyzing EEG data. One has to differentiate between technical and biological artifacts. The majority of biological artifacts in EEG arise from eye movements and other muscular activity of the head and neck. Both types of biological artifacts generate electric fields (amplitudes $\gg 100\mu V$) which are several orders of magnitudes stronger than neuronal activity ($\leq 40\mu V$). There are various reasons for technical artifacts, such as electrodes loosing contact to the skin, amplifier clipping, external electric fields or malfunctions in the electric insulation. Fig. 2.6 shows several examples of artifacts in EEG data.
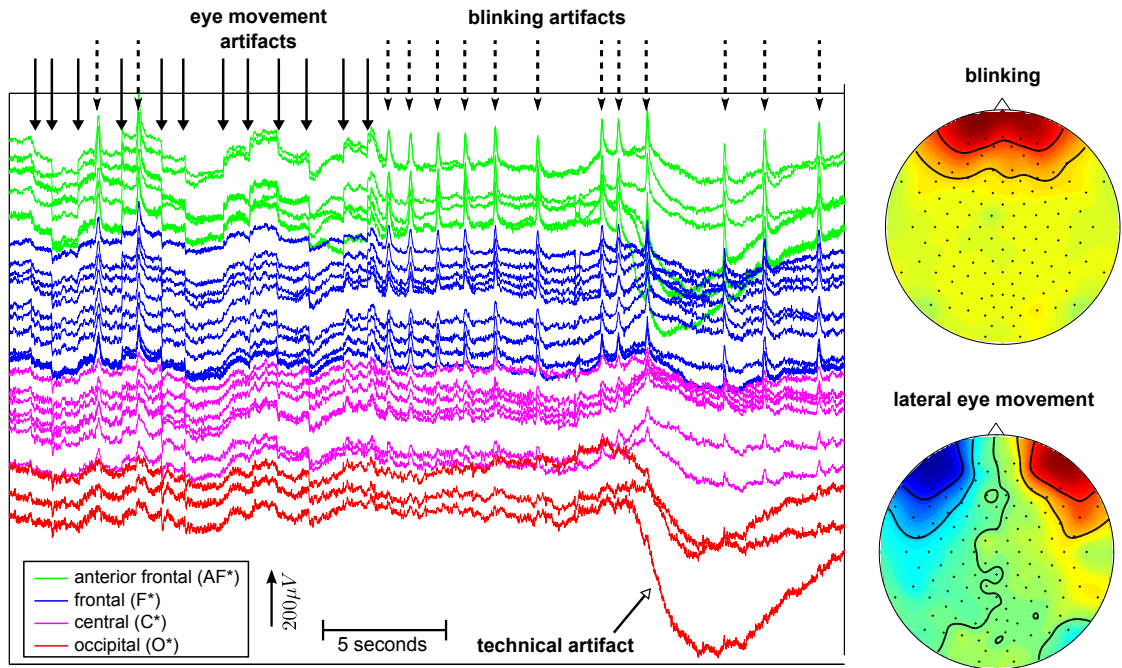
Figure 2.6: *Visualization of artifacts in EEG signals. The time series shows an excerpt (40 second) of EEG recording which is contaminated with numerous artifacts. EEG channels are colored corresponding to their location. Eye movement artifacts and blinking artifacts are marked by solid/dashed lines. A technical artifact is also marked. The scalp plots depict spatial patterns of eye artifacts.*

There are two major strategies of how to deal with artifacts when analyzing EEG data: rejection methods and projection methods. Both strategies are described in the following.

2.5.1 Rejection Methods

Artifact rejection methods seek for EEG epochs which contain artifacts and then exclude those epochs from the following data analysis. Rejecting artifact epochs will however "cost" data and more conservative parameter settings (or thresholds) will result in a reduced amount of data which remains for further analysis. The standard approach which is followed within this thesis is to reject an epoch if the features are out of the range of $[thres_{min}\ thres_{max}]$. Both, the features and the thresholds can be determined in multiple ways, mainly depending on which types of artifacts are to be removed.

In order to identify eye movement artifacts, a subset of channels which is most affected by horizontal and vertical eye movements (typically `F9`, `F10` for horizontal and `Fp1` and `Fp2` for vertical eye movements) is selected as preprocessing step. The amplitude difference within these electrodes is then evaluated and an epoch is rejected if the amplitude difference is out of the range of $[-80\ 80]\,\mu V$. In order to remove EEG epochs that are contaminated with an EMG artifact of muscular activity, the band-power in each electrode is estimated as the standard preprocessing step. As muscular activity – such as jaw muscle activity – elicits a broad-band EMG artifact, the spectral power in the band $[5\ 40]\,\text{Hz}$ resembles a suitable feature which allow to identify muscle artifacts. However, the spectral power of EEG signals may vary across subjects and channels even in absence of any artifact. Therefore, it is common practice to determine the threshold based on the spread of the data.

Instead of applying the thresholding as described above, there is a vast number of outlier detection algorithms which can also be applied with EEG data – see Hodge and Austin (2004) for a review. For instance, Harmeling et al. (2006) proposed a simple and fast rejection method which computes an ordering of the data from outliers to prototypes. Such ordering is based on the indices of a high-dimensional nearest neighbors clustering.

2.5.2 Projection Methods

Projection methods follow another approach to clean EEG data from artifacts. Instead of rejecting epochs that contain artifacts, the EEG data are decomposed into artifactual sources and non-artifact (i.e. neural) sources. Spatial filters are applied, which outproject artifactual sources. The resulting EEG data is then supposed to only contain neuronal sources. Generally, two steps are required to apply artifact projection methods.

Firstly, the EEG data has to be unmixed and thereby decomposed into sources. Independent component analysis (ICA) is mostly applied for unmixing. ICA is a blind source separation algorithm which seeks to decompose the data into a set of maximally independent sources. There are several mathematical formulations to quantify "independence", resulting in multiple objective functions and ICA methods such as Infomax (Bell and Sejnowski, 1995), TDSEP (Ziehe and Müller, 1998) and many more. For a review, see Hyvärinen et al. (2004).

As a second step, the artifactual source components have to be identified. The major problem of artifact projection methods is that the decomposition of the data might not lead to components that are "purely" artificial or "purely" neuronal. When discarding or considering those components, the resulting data will either lack some (maybe important) neural sources, or the data will still contain artifacts. A common strategy is to manually inspect each component – investigating the spatial pattern as well as the spectrum of the source. As this manual procedure requires expertise and time, an automatized procedure was described in Winkler et al. (2011). For the automatized framework, a subject independent classifier evaluates multiple features and thereby identifies neuronal and artifactual source components.

Based on these two steps, artifactual components can be projected out and a cleaned EEG is obtained. One should however note that due to dimensionality reduction, the cleaned EEG might not have full rank if it is projected back into the electrode space. This poses a problem for processing algorithms such as CSP – see Section 2.3.2.

Additionally, recent work by Bünau et al. (2009) proposed to apply spatial filters, which project the data into a stationary and a non-stationary subspace. Assuming artifacts to be non-stationary and neural activity to be rather stationary, they proposed to apply the Stationary Subspace Analysis (SSA) as a preprocessing step to exclude non-stationary sources.

2.6 Existing BCI Paradigms

This section aims to provide a coarse overview over existing BCI paradigms, while focusing on the most relevant paradigms for this thesis. Thus, the most influential ERP-based BCI paradigms using visual and auditory stimuli are introduced. Then, the concept of BCIs based on mental imagery and the general needs and requirements of patient applications are briefly discussed.
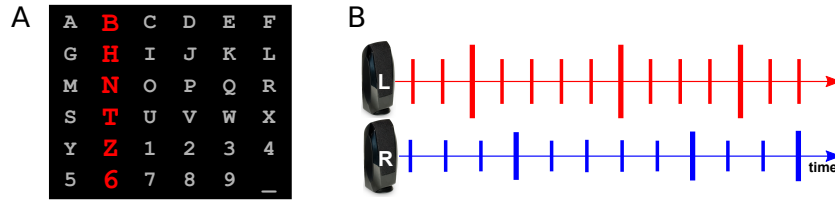
Figure 2.7: *Visualization of the visual MatrixSpeller and an auditory streaming paradigm. In the MatrixSpeller (**A**), the letters are organized in a 6 × 6 matrix and rows and columns are flashing in random order. The auditory streaming paradigm (**B**) is analog to (Hill et al., 2004) and (Hill and Schölkopf, 2012): two streams of auditory stimuli – plotted in red and blue – are presented in parallel, with each stream containing standard (non-target) and deviant (target) stimuli.*

2.6.1 Visual ERP Paradigms

It is important to note that very first visual BCI paradigm based on ERPs, the MatrixSpeller (Farwell and Donchin, 1988), is still frequently applied and researched. The MatrixSpeller was designed in a simple, intuitive but also highly effective design: the alphabet consists of 36 letters and symbols. It is displayed as a 6 × 6 matrix and the rows and columns of the matrix are flashing in a random order. There are numerous extensions of this paradigm which aim to increase the spelling speed by using optimized stimulus properties (Kaufmann et al., 2011; Townsend et al., 2012; Tangermann et al., 2011b), flashing patterns (Allison and Pineda, 2006; Sellers et al., 2006a; Hill et al., 2009) or by incorporating additional prior information in form of language models (Speier et al., 2012; Kindermans et al., 2012; Mainsah et al., 2014). The MatrixSpeller is a gaze-dependent visual paradigm, as the users need to move their gaze onto their desired letter. On the one hand, such paradigms (also called "overt" paradigms) generally evoke highly discriminative ERP signals, resulting in a high communication speed. On the other hand the performance and usability of such paradigms drops significantly when users are not able to control their gaze – see Treder and Blankertz (2010) and Brunner et al. (2010) for details. Therefore, the MatrixSpeller might be inapplicable for those impaired users who are in need for an alternative communication solution (also referred to as "end-users" in the following). Moreover, the latest achievements in the eye-tracking technology raised the debate of whether gaze-dependent BCI paradigms could generally be substituted with eye-trackers, which are considered to be both cheaper, and more robust (Treder and Blankertz, 2010).

Gaze-independent visual BCI paradigms have been proposed as an alternative solution for users with impaired gaze control. For such paradigms, users are not required to move their gaze, as all stimuli are presented at the same location. One example is the CenterSpeller (Treder and Blankertz, 2010), where a sequence of symbols (varying in shape and color) is presented. Acqualagna et al. (2010) presented the rapid serial visual presentation (RSVP) paradigm as an alternative, which displays the letters of the entire alphabet in a fast and random order. While such paradigms do not require the users to move their gaze, they however require the ability to maintain the gaze on a constant position while not closing the eyes. As this might be a problem for a substantial group of end-users, BCI paradigms were investigated which are completely independent of the visual domain and use auditory or tactile stimuli instead.

2.6.2 Auditory ERP Paradigms

There is an increasing awareness within the BCI research community that "traditional" visual ERP paradigms have limited use for that population of severely impaired users. This has been stimulating the

research of novel auditory BCI paradigms (Nijboer et al., 2008b; Kanoh et al., 2008; Furdea et al., 2009; Klobassa et al., 2009; Schaefer et al., 2010; Guo et al., 2010; Höhne et al., 2010; Halder et al., 2010; Höhne et al., 2011a; Schreuder et al., 2011a; Kim et al., 2011; Käthner et al., 2012; Hill and Schölkopf, 2012; Nambu et al., 2013).

While the earliest approaches (Hill et al., 2004; Hill et al., 2005) could show the basic feasibility of auditory BCIs, they suffered from very low theoretical or practical information transfer rates (ITR) of less the 1 bit/min. Some of the more recent paradigms (partly also presented within this thesis – see Sections 3.1.2 and 3.2.2) showed a break-through in performance (Schreuder et al., 2011a; Höhne et al., 2011a, 2012), with an ITR of more than 4 bits/min and an online spelling speed of up to one symbol/min. In online studies with healthy subjects, their communication rate came close to that of gaze-independent visual BCI paradigms.

Basic Streaming Paradigms   In auditory streaming paradigms, the user perceives multiple streams of stimuli and the BCI aims to determine which stimulus stream the user is attending to. Mostly, two parallel streams are presented, enabling the BCI to make a binary decision. Fig. 2.7 shows an example for a binary streaming paradigm. Hill et al. (2004) showed that it is possible to decode the users' attention by analyzing the ERP response of both target and the non-target stimuli from the attended direction. However, the major drawback of such basic streaming paradigms is that they only enable a binary class decision which results in a limited bandwidth of less than 1 bit/min.

AMUSE Paradigm   The AMUSE paradigm (Auditory Multi-class Spatial ERP) is a well-established auditory BCI paradigm which was first described in an offline study by Schreuder et al. (2010). It was then applied for online study with a spelling application in Schreuder et al. (2011a). The AMUSE paradigm introduced the use of spatial auditory cues, using brisk artificial stimuli varying in location and pitch. As it is shown in Fig. 2.8, the user is surrounded by six speakers that are positioned at distinct locations. While a pseudorandom stimulus sequence is presented, the user attends to only those stimuli coming from the target speaker. Evaluating the users' ERPs, a one-out-of-six class decision drives a speller application (see Fig. 2.8B). The spelling is implemented in a two-step procedure of first selecting the group and then the intended symbol.

Despite its comparably high performance, the major drawback of AMUSE – and gaze-independent BCI paradigms in general – becomes obvious when analyzing workload and usability. Compared to the MatrixSpeller, AMUSE is considerably more complex to use, as the two-step spelling procedure might be slow and difficult to follow. In addition, the bulky setup (the user must be surrounded by six speakers) might be obstructive for patient application. Therefore, there is a need for auditory BCI paradigms that allow a fast communication while being intuitive to use and portable. Two paradigms that aim for this objective are described in this thesis – see Chapter 3.

2.6.3 BCI Performance Evaluation: the Information Transfer Rate

This section describes how to measure the communication speed of a BCI system. Therefore, the most commonly used metric is the information tranfer rate (ITR). Arising from information theory, the ITR is a general metric that quantifies the amount of information which is transfered over a noisy channel (Shannon, 1949). While (Schlögl et al., 2007) reviewed multiple formalizations of the ITR, Wolpaw et al. (2002a) proposed simple formula which is widely used to evaluate BCI performance.
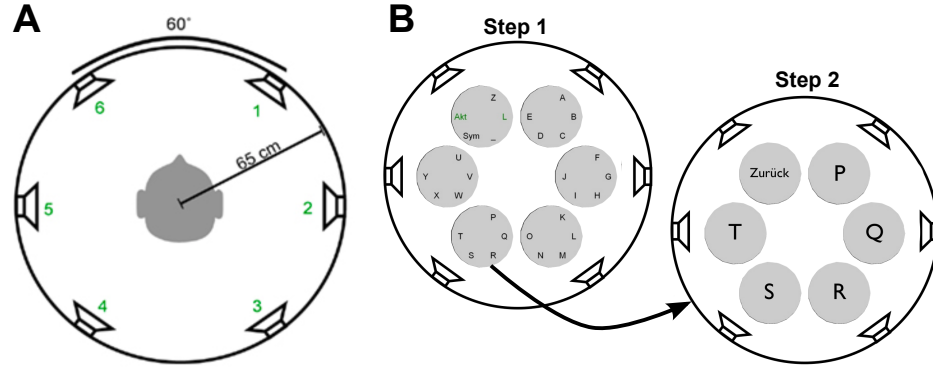
Figure 2.8: *Visualization of the AMUSE paradigm. Plot **A** depicts how the users are surrounded by six speakers at ear height. Speakers are equally spaced between neighbors with a circle diameter of 65 cm. Plot **B** depicts the AMUSE spelling application which allows to choose a letter in a 2-step approach. Plots modified from (Schreuder, 2014), with permission.*

They defined R as the bits per selection with

$$R = log_2 C + P \cdot log_2 P + (1-P)\frac{1-P}{C-1}, \tag{2.13}$$

and C being the number of classes and P the selection accuracy. The ITR is mostly estimated as bits/minutes with

$$ITR[bits/min] = V \cdot R \tag{2.14}$$

and V specifying the number of selections per minute. Thus, the ITR depends on the selection accuracy and the selection speed and yields an suitable BCI performance metric. The ITR formalization of Eq. 2.13 is also used in the remainder of this thesis.

However, it is a simplified measure for information transfer which is based on multiple assumptions that might not hold for numerous BCI applications (Schlögl et al., 2007). For instance, the ITR formulation in Eq. 2.13 assumes each class to have the same prior probability and the class-specific accuracy to be identical for each class. Moreover, it is an "all-or-nothing" metric, which only evaluates whether one decision is correct or incorrect. The ITR formula by Wolpaw et al. (2002a) neither considers the certainty of a decision, nor its underlying rank-distribution. This is troublesome in particular for problems with a high number of classes, as it does not reward the situation in which the true target class is identified as second-best (or third-best) class amongst $C = 30$ or more classes.

Approaching the above-mentioned problems, Section 3.3.2 describes a novel, rank-based measure for BCI performance and multiclass accuracy.

2.7 Requirements for Successful Patient Applications

For the last two decades, patients with highly limited or no means of communication represent one major target population for brain-computer interface research (Birbaumer et al., 1999; Birbaumer and Cohen, 2007). The ultimate goal is to provide a BCI system as a communication tool for patients in the completely locked-in state. Although a number of advances with respect to performance and usability have been achieved, the BCI technology has not yet met the requirements for a successful

application with patients. Mak et al. (2011) reviewed the current status, the limitations and further directions of BCI, concluding that "P300-BCI should be simple to operate, affordable, accurate, and efficient for communication on a daily basis". The same requirements hold for any other BCI which aims to be applied for communication.

The following paragraph discusses a number of aspects that need to be improved in order to enable a successful patient application.

Usability and Workload   The great majority of non-visual BCI paradigms lacks in usability and simplicity. Especially gaze-independent paradigms require the users to deal with a rather complex spelling interface. This can be illustrated by comparing the gaze-dependent MatrixSpeller (Farwell and Donchin, 1988) with gaze-independent visual paradigm, or non-visual spelling paradigms. To operate the MatrixSpeller, users do not require instructions beyond the hint to mentally focus on the desired symbol. All available symbols are present on the screen at all times and the paradigm follows the concept of "what you see is what you get". Other gaze-independent paradigms have a lower bandwidth, enabling the user to select between a reduced number (e.g. 6) classes. The selection of a symbol thus requires the execution of a series of control steps, which is not intuitive to naive users. While healthy study participants in good condition may be able to use such "indirect" interfaces despite of the enhanced workload, it remains a problem and a high entrance barrier for many patients (Küber et al., 2013; Schreuder et al., 2013b).

Hardware issues: Costs and Accessibility   Besides one exception[2], BCI systems are not yet commercially available. Hence, BCI systems are customized by each research laboratory, using expensive research products, resulting in costs for hardware only of more than 10,000€ and up to 100,000€. Being designed for scientific use only, current BCI systems are very expensive and barely accessible for patients, doctors or centers for assistive technology.

   In analogy to the previous paragraph, usability constraints also apply for hardware issues. The majority of the currently applied hardware is bulky, sensitive to external noise and not intuitive to set up. Moreover, electrode systems based on conductive gel are inconvenient for each user, especially for patients. Therefore, dry electrodes (Popescu et al., 2007; Volosyak et al., 2010; Grozea et al., 2011) as well as water-based systems (Volosyak et al., 2010) have lately been researched with promising outcomes.

Speed and Robustness   The ultimate goal is to provide locked-in patients with a BCI, which is both fast and robust to use. However, those two aspects are opposing, since increased robustness of a system can often be achieved by accumulating more evidence, leading to a slower communication rate. Therefore it is advisable to discuss the speed-accuracy trade-off with each patient individually. While some end-users might prefer a more reliable and rather slow solution, other patients might prefer the challenge of a faster and less reliable system.

Calibration Time   Comparing most state-of-the-art BCI systems with other assistive technology such as eye-trackers, it becomes obvious that the calibration time of a BCI is considerably longer. Firstly, multiple hardware issues lead to an increased time demand to acquire neuronal signals in general. Secondly, the internal data analysis methods of most BCI paradigms are mostly based on calibration data.

   The general goal is to reduce the calibration time to a minimum. Therefore, novel (dry or water-based) EEG systems have been researched, as described in the previous paragraph. Moreover, advances

---

[2]The exception is the intendix-system by g.Tec, which offers an implementation of the MatrixSpeller http://www.intendix.com/

in the data analysis framework allowed reducing the amount of calibration data which is necessary to operate a BCI (Blankertz et al., 2007, 2011; Sannelli et al., 2011). One recent approach for ERP-based paradigms also enables BCI control without any calibration data (Kindermans et al., 2012, 2014).

Experience with Patients Patients in need for a BCI as communication pathway display a variety of individual needs and characteristics. However, Kübler (2013) recently pointed out, "fewer than 10% of the papers published on brain-computer interfacing deal with individuals presenting motor restrictions, although many authors mention these as the purpose of their research". Even within patient studies, the patients who were chosen to participate were rarely in-need of a BCI, since their residual communication abilities with assisted technology (AT) were higher than the best state-of-the-art BCI could ever provide. There are many possible reasons for this mismatch, with some listed below:

1. Increased organizational, technical and temporal effort is necessary to conduct patient experiments

2. Ethical issues have to be intensely discussed for each patient study when dealing with (completely) locked-in individuals.

3. There is a lack of access to these patients.

Consequently, there is a lack of knowledge of what exact problems – be it on a global or individual scale – one has to address in order to provide an effective and useful tool for patients. Additionally, in contrast to data from healthy users (Sajda et al., 2003; Blankertz et al., 2004, 2006a; Tangermann et al., 2012b), EEG data of patients is not publicly available. Therefore, it is troublesome to evaluate and optimize novel computational tools for the needs of patients.

TOWARDS USER-FRIENDLY AUDITORY BCIS: SHIFTING COMPLEXITY FROM THE USER TO THE BCI SYSTEM

W HILE the general applicability of ERP-based BCIs has already been proven more than twenty years ago (Farwell and Donchin, 1988), researchers have recently been studying BCI paradigms that can be operated by users with an impaired oculomotor function – see Section 2.6 for details.

This chapter describes four EEG studies with healthy subjects. All four studies investigate auditory event related potentials (see Section 2.2.1 for an introduction) for brain-computer interfacing. The field of auditory BCIs is relatively young, with the proof of concept being established in Hill et al. (2004) and Schreuder et al. (2010). However, compared to (gaze-dependent) visual paradigms, state-of-the art auditory BCI spelling paradigms are suffering from two major shortcomings:

- **Speed.** Auditory BCI paradigms feature a slower information transfer rate and spelling speed.

- **Complexity.** Most auditory BCI spellers are highly complex to use. The user is required to be very focused in order to navigate through the spelling application, as two consecutive multiclass selections (i.e. (1) group selection and (2) letter selection) have to be performed in order to spell a letter.

This chapter therefore addresses those two shortcomings of auditory BCIs, aiming to reduce complexity for the user while improving spelling speed. Section 3.1 describes an online study with the PASS2D paradigm, which combines a nine-class ERP paradigm with a predictive text system. In Section 3.2, the use of naturalistic auditory stimuli is investigated within the PASS2D paradigm. It is found that natural stimuli can improve the usability and performance of the PASS2D paradigm, although the stimuli themselves are less standardized. In Section 3.3, the CharStreamer paradigm is introduced, which strives for the maximally user-friendly auditory spelling paradigm. The CharStreamer can be operated with instructions as simple as "please attend to the letter that you want to spell". This is achieved by implementing a 30-class auditory BCI paradigm, which combines natural stimuli and sequential stimulation. Finally, Section 3.4 addresses the importance of choosing an appropriate stimulation speed for ERP based BCIs. Based on the results of a simple auditory ERP experiment, it is shown that the choice of stimulation speed highly impacts the ergonomics, neurophysiology, as well as the classification accuracy and the resulting BCI performance.

3.1 Combining a 9-class Auditory ERP Paradigm with Predictive Text System: PASS2D

I N this part, an auditory BCI paradigm with a spelling application is introduced. Compared to existing auditory paradigms, several novel approaches were taken in order to increase usability of the BCI, while shifting complexity from the user into the system. As the first auditory BCI speller, the system enabled the user to select a letter with a 1-step approach. Such 1-step selection is enabled by an internal language prediction framework, which is commonly used in mobile phones. Additionally, the number of classes was increased to nine, leading to a more complex internal multiclass decision. Those nine stimuli were presented with headphones in order to allow a small and portable setup. The paradigm - called PASS2D - was investigated in an online study with twelve healthy participants. Users spelled with more than 0.8 characters per minute on average (3.4 bits per minute) which makes PASS2D a competitive method, being more accurate and faster than most of the auditory ERP spellers previously reported. Thus, PASS2D enriches the toolbox of existing ERP paradigms for BCI end users such as people with ALS disease in a late stage. The data and results were previously published in (Höhne et al., 2010).

3.1.1 Motivation

As discussed in Section 2.7, there is a need for communication solutions which are independent of any muscular activity. While most visual paradigms rely on the user's ability to control the eye-gaze with its corresponding muscles, auditory BCI paradigms could enable a communication pathway which is entirely independent of muscular abilities. The objective of this study was to design a user-friendly and portable setup for an auditory BCI that enables fast and intuitive spelling.

A variety of auditory paradigms were described in Section 2.6.2. This work presents an approach to extent some characteristics of the AMUSE paradigm by Schreuder et al. (2010) and Schreuder (2014). Analog to the AMUSE paradigm, we also use auditory stimuli that vary in both, pitch (high, medium, low pitch) and direction (stimuli from the left, middle, right). However, the information transmitted by those two dimensions was redundant within AMUSE, such that the tones which were presented from a specific direction had a unique pitch. We aim to unlock this relation such that information transmitted by the two dimensions was independent. This means that several tones with varying pitch were presented from the same direction. The resulting 3×3 design offered an arrangement of nine stimuli that were easy to discriminate from each other.

The novel spatial auditory stimuli were implemented into a 9-class BCI paradigm. The chosen setup had the advantages of

- offering a graphical representation which is easy to understand and memorize,

- being applicable with light headphones only

- enabling an intuitive spelling procedure,

- yielding a competitive spelling speed.

The paradigm was named 'Predictive Auditory Spatial Speller with two-dimensional stimuli', or PASS2D.

3.1.2 Experiment 1: PASS2D Online Study

Experimental Protocol

Twelve healthy volunteers (9 male, mean age: 25.1 years, range: 21 – 34, all non-smokers) participated in a single session of a BCI experiment. Table A.1 provides details about the age and sex of the participants. A session consisted of a calibration phase and a subsequent online spelling part – as shown in Fig. A.1. It lasted three to four hours.

EEG signals were recorded monopolarly using a Fast'n Easy Cap (EasyCap GmbH) with 63 wet Ag/AgCl electrodes. Signals were amplified using two 32-channel amplifiers (Brain Products). Feature extraction and classification was performed as described in Section 2.3.1 and 2.4. Fig. A.1 shows the course of the experiment. For the calibration part as well as for the online part, participants were asked to focus on target stimuli while ignoring all non-target stimuli. Auditory stimuli were presented on light neckband headphones (Sennheiser PMX 200).

Collection of Calibration Data   Three calibration runs were recorded per subject. To differentiate target and non-target subtrial in later experimental stages, the collected data were used to train a binary classifier (RLDA) as described in Section 2.4. Each calibration run consisted of nine trials (i.e. nine multiclass selections) with each of the nine sounds being target during one of the trials, see Fig. A.1. In addition, one practice-run (run 0) was performed initially without EEG recording. Prior to the start of each calibration trial, the target cue was presented to the subject three times while in addition the corresponding number on the 3×3 grid was highlighted on the screen.

During the calibration phase, each trial consisted of 13 or 14 pseudo-random iterations of all nine auditory stimuli. Visual stimuli were not given during these trials. While the last 12 iterations were used to train the classifier, the first one or two iterations were dismissed to ensure a balanced distribution of stimuli in the calibration data. One trial provided 9×12 subtrial epochs (12 target subtrials and 8×12 non-target subtrials) for the classifier training. The combined training data from all runs comprised $108 \times 27 = 2916$ subtrial epochs for each subject (minus a small fraction of artifactual epochs that were discarded). Participants were asked to count the targets and to report the number of occurrences at the end of each trial (counting task).

Online Spelling Task   Two online spelling runs were performed. Subjects were asked to spell a short German sentence ('Klaus geht zur Uni') composed of 18 characters (including space characters) and a long sentence composed of 36 characters ('Franz jagt im Taxi quer durch Berlin') in separate runs. The task was to finish both sentences without mistakes and each false selection had to be corrected.

Auditory Stimuli   The selection of stimuli is a crucial element for any kind of BCI system which is based on evoked potentials. An in-depth investigation of the impact of stimulus properties on ERPs and BCI performance is given in Section 3.2. For the PASS2D paradigm, three artificially generated tones that varied in pitch (high/medium/low) and tonal character were carefully chosen such that they were - on a subjective scale - as different as possible from each other. The tones were generated artificially with 708 Hz (high), 524 Hz (medium) and 380 Hz (low) as base frequencies. Each tone was presented on the headphones with three different directions: only on the left channel, only on the right channel, and on both channels. With its two independent dimensions, the stimuli can be visualized in a 3×3 array with pitch specifying the row and direction coding for the column – see Fig. 3.1). This 3×3 design obeys a close analogy to the number pad of a standard mobile phone, where e.g. key 4 is represented by the medium tone pitch (used for keys 4, 5 and 6) and was presented on the left channel only (used for keys 1, 4 and 7).
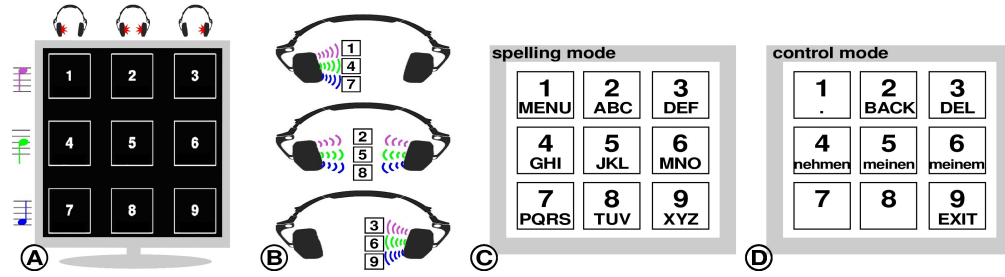
Figure 3.1: *Visualization of the nine auditory stimuli, varying in pitch and direction. The 3×3 design depicted in plot **A** was shown on the screen. Plot **C** and **D** show the spelling mode and the control mode, respectively.*

Each stimulus lasted 100 ms, SOA was 225 ms and a low-latency USB sound card (Terratec DMX 6Fire USB) was used to reduce latency and jitter. The pseudo-random iterations of stimuli were generated such that two subsequent stimuli did not have the same pitch. Moreover, the same stimulus was repeated only after at least three other stimuli had appeared.

In this study, the visual domain was only used to report which selections were made, which text had already been spelled, or which words were available to choose from in the so called *control mode* (see Fig. 3.1D and Section 3.1.2 for an explanation of the two system modes during the online spelling phase). The paradigm was implemented such that all visual information could be read out to the user. Thereby PASS2D can be operated completely independent of the visual domain and it could even be used by blind users.

Predictive Text System For the presented ERP speller, the commonly used T9 predictive text system from mobile phones (discussed in Dunlop and Crossan (2000)) was applied in a modified version. A similar approach to incorporate application intelligence into a BCI system was presented by Jin et al. (2010) in order to effectively communicate Chinese characters in a visual BCI paradigm.

The standard T9 system uses more than nine keys: key '1' codes for dot/comma, keys '2' to '9' code for the alphabet, '0' for space, '+' and '#' for symbols or further functions. The system was modified such that instead of the twelve keys mentioned above, only nine keys were needed for spelling while remaining an intuitive control scheme. The system was constrained to words in a corpus of about 10,000 frequently used words of the German language, which can be arbitrarily extended.

To overcome the problem of having a spelling scheme that is easy and intuitive to use and on the other hand flexible and fast with only nine keys, two modes were implemented: A *spelling mode* and a *control mode*, see Fig. 3.1.

In analogy to the predictive text system in mobile phones, a word was spelled by entering a sequence of keys. To spell a character, the user had to select the corresponding key ('2' to '9') in the *spelling mode*. Each key codes for three or four characters, see Fig. 3.1C.

After selecting the correct sequence of keys for a specific word, the user chose key '1' to switch the system into the *control mode*. In this mode, he sees the desired word in a list — together with all other words which can be represented by the entered sequence. By choosing one of the keys '4' to '8' he determines the desired word with one additional selection step. The list of matching words is ordered such that more frequent words (i.e. they have a higher rank according to the underlying corpus) are represented by smaller key numbers. As an example: after entering the keys '6346361' the user can choose from 'nehmen', 'meinem', 'meinen' (see Fig. 3.1D) as all these words can be represented by the entered sequence of keys. In the *control mode*, the system is limited to present a maximum list size of 5 words, which was sufficient to spell each word of the underlying corpus.

In case the user performed an erroneous multiclass selection, he could correct it by one to three addi-

tional selections. If the selection of the last sequence key did not conform to the corpus (there was no word in the corpus that fitted the entered code), it was not accepted and could be corrected with only one selection. If the mode was changed by mistake, it took one selection to return to the correct mode and another to choose the right key. In all other cases of erroneous selections, it took the user two selections to delete an erroneous key (change the mode by entering '1' and then delete last the key by entering '3') and a third selection to enter the intended key.

3.1.3 Findings

Offline Data

The most relevant findings of the offline analysis described below. Section A.1.3 presents additional behavioral data.

Binary Classification Accuracy The accuracy of a binary decision (based on the epoch of one subtrial) was estimated on the calibration data for each participant. Based on the estimated errors, participants *VPnx* and *VPmg* were excluded from the following online experiments due to the poor binary classification performance of less than 70 % (classwise balanced). A cross-validation analysis (see Table A.1) revealed that on average over all ten remaining subjects, 77.7 % of the stimuli were correctly classified. To account for the imbalance between non-targets and targets, the classwise balanced accuracy was used, which is the average decision accuracy across classes (target vs non-target) with a chance level of 50 %.

Spatial and Temporal Distribution of N200 and P300 Fig. 2.2 depicts the grand average ERPs at electrodes 'Cz' and 'FC5' together with the corresponding scalp maps for two time intervals. As expected, the ERPs for the non-target stimuli (grey lines) show a regular pattern that reflects the neural processing of the auditory stimuli. It occurs every 225 ms and is dominated by a N200 component. Moreover, those plots illustrate the different EEG signatures of the non-targets and the targets. At frontal electrodes a lateral and symmetric class-discriminative negativity is observed 230-300ms after stimulus onset. It directly follows up on the N200 component. For simplicity reasons the class-discriminant component will be referred to as the *N200* component in the following.

Starting from approximately 350 ms after the stimulus onset, a second class-discriminant interval is observed for target stimuli. It is a symmetric positive component located at central electrodes and will be referred to as the *P300* component in the following. The amount of class discrimination that is contained in the two electrodes during different time intervals is represented by two colored bars (Figure 2.2b). They depict ssAUC values (see Section 2.4.5 for details). Positive ssAUC values are colored in red and represent time intervals where target ERP amplitudes are larger than non-target ERP amplitudes. Negative ssAUC values are colored in blue and represent time intervals with target ERP amplitudes smaller than non-target ERP amplitudes.

Due to the contra-lateral processing of auditory stimuli (Langers et al., 2005), the N200 was expected to vary for each stimulus. Fig. 3.2a depicts the grand average ssAUC scalp maps of N200 for each of the nine stimuli, illustrating that the early negative deflection is spatially varying for different auditory stimuli, but not the P300 (Fig. 3.2b). In most multiclass ERP paradigms the classification is based on a 2-class problem (target vs non-target). Thus, the fact that there might be variability in the spatial (or temporal) distribution of discriminant information for different stimuli is mostly disregarded. Although the classification procedure in the presented approach is also based on 2-class decision, Fig. 3.2a shows that there is some spatial class-discriminant information, which is not yet

exploited by the LDA classifier[1]. Generally, the discriminative information of the N200 component seems to be stronger on the left hemisphere (cp. the grand average in Fig. 2.2d). In addition, this spatial distribution seems to follow a slight contra-lateral tendency (Fig. 3.2a): stimuli presented on the right audio channel (right column in the grid of scalp maps) induce a more discriminative N200 component on the left hemisphere. On the contrary, the class discriminative N200 components induced by stimuli on the left audio channel (left column in the grid of scalp maps) are located rather on the right hemisphere.
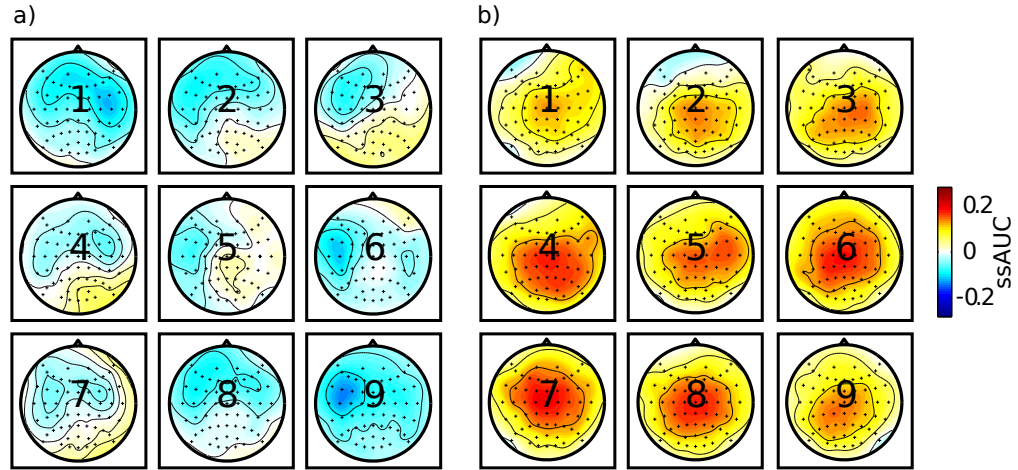

Figure 3.2: *Grand average of the spatial distribution of the stimulus specific N200 component (**a**) and P300 component (**b**). Area under the ROC curve (ssAUC) values are signed such that positive values stand for a positive components and negative values represent negative components. Plots are arranged corresponding to the 3x3 design of the PASS2D paradigm.*

Discriminative Information in the Spatial and Temporal Domain The impact of the spatial and the temporal domain on the classifier was investigated separately (Fig. 3.3) by analysis of isolated data segments of individual channels or small time intervals. The most discriminative information was found 400–500 ms after the stimulus onset, which reflects the importance of the P300 component. The most discriminative channels were found at central-lateral locations such as *C4*/*C5*, when averaging intervals were selected heuristically. Comparing this to the grand average ERP scalpmaps in Fig. 2.2, one can find an overlap of N200 and P300 in the mentioned areas. This stresses the importance of the N200, although a stimulus-specific variation (see Fig. 3.2 and results above) was found. It can be concluded that both components N200 and P300 can be used for classification, but the P300 component contains more discriminative information.

Online data

BCI Performance: Bitrate and Spelling Speed It took 15 min to 26 min (μ=20.9) to spell the short sentence and 31 min to 76 min (μ=43.5) for the long sentence. Variation in the number of multiclass selections originates from the different number of false selections (see Table A.1) which then had to be corrected. Since the sentences were not spelled word by word but in one go, all kinds of pauses

---

[1]The Relevance Subclass LDA (cp. Section 4.2.5) is designed to exploit such stimulus-specific features. It was however not applied for this online study.
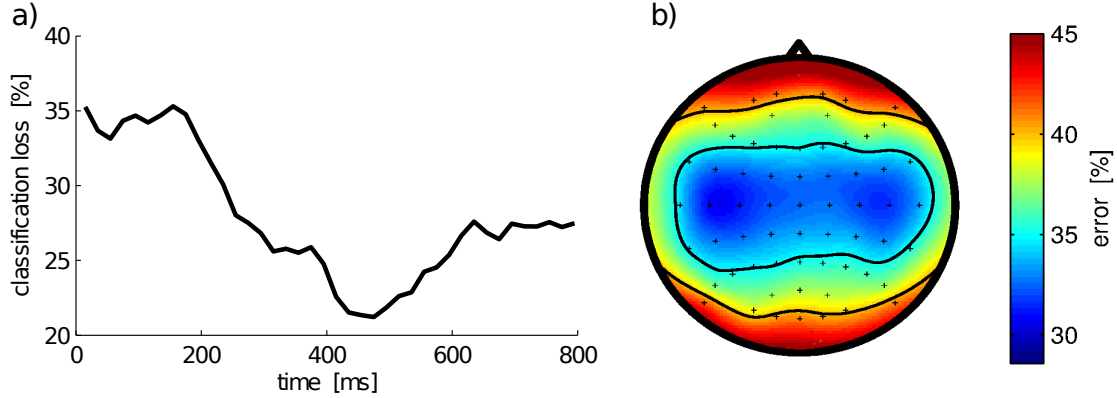
Figure 3.3: *Grand average temporal (**a**) and spatial (**b**) distribution of discriminative information. Reported loss values in the temporal domain are obtained for a sliding window of 50 ms. The loss values obtained for each electrode separately are depicted as a scalp topography (**b**).*

Table 3.1: *Spelling speed in the online condition, averaged over both sentences. The bitrate neglects the beneficial effect of the predictive text system; all individual pauses are taken into account.*

	avg	min	max
characters per minute	0.89	0.65 (*VPnz*)	1.17 (*VPoc*)
ITR[bits/min]	3.4	2.7 (*VPnz*)	4.4 (*VPoc*)

are taken into account. However, the time for individual relaxation and fixed intertrial periods are among the main influence factors for the spelling speed. Fig. 3.4 shows that neglecting the time for individual relaxation, and thereby only considering the stimulation time ($\sim$ 31 seconds) and a fixed inter trial time (4 seconds), results in an average benefit of more than 1 bit/min or 0.25 char/min.

In general, a higher multiclass accuracy can be obtained by increasing the number of subtrials. The rate of communication (Wolpaw et al., 2002b) counterbalances this effect, enabling to compare different studies more accurately.

On average, subjects achieved an Information Transfer Rate (ITR – see Section 2.6.3) of 3.4 bits/min in the online condition (based on the nine class decision, incl. all pauses), see Table 3.1. An average online spelling speed of 0.89 characters/minute was observed – see also Table A.1 and Fig. 3.4.

In general, the information of one character is coded by at least 4.75 bits (1 out of 27, 26 letters plus space). Considering that the BCI controlled speller application presented here enables an average spelling speed of 0.89 characters/minute, the ITR could also be quantified with 4.23 bits/minute (0.89 × 4.75) in a hypothetical BCI paradigm with 27 classes. The discrepancy between 3.4 bits/min and 4.23 bits/min can be explained with the predictive text system, which thus increases the ITR by at least 0.83 bits/min or 24%.

Multiclass Accuracy Averaged over all trials and participants of the online experiments, 89.37% of the multiclass decisions were correct (chance level is 1/9, 11.11%). Table A.2 reveals that none of the nine stimuli has a significantly increased or decreased accuracy.

In the presented paradigm, the nine auditory stimuli are not completely independent: for each target there are 4 non-targets being equal in one dimension, i.e. two stimuli with the same pitch (same row) and two stimuli with the same direction (same column). Since these similarities could influence the results, it was tested if this is reflected in the binary classifier outputs or multiclass decisions.
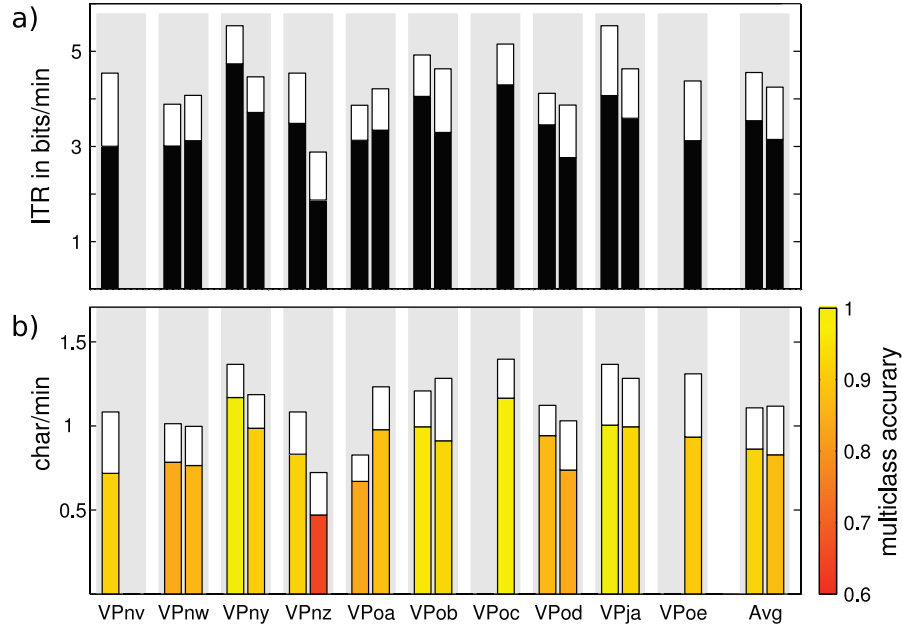
Figure 3.4: *For each subject and both sentences the barplots show the multiclass accuracy and the resulting Information Transfer Rate (**a**) and the spelling speed in characters per minute (**b**). The white extensions of the bars mark the potential increase that could result if individual pauses are disregarded for the computation of the spelling speed and ITR. For each subject, the left (right) bar represents the performance of the short (long) sentence. For three subjects there is only one bar, because the spelling of the second sentence was canceled or not even started.*

1. False positives for non-targets with the *same pitch* as the target: The probability of false positives in single-epoch binary classification for these non-targets was in fact higher than for other non-targets, as the classifier outputs were significantly more negative and therefore more similar to target outputs ($p < 10^{-20}$). An increased probability for erroneous multiclass selections with the correct pitch but wrong direction was observed as well. Table A.2 reveals that 47 out of 79 multiclass errors had an equal pitch. Assuming no dependency, one would expect 19.75 (2 out of 8). This is a significant deviation (χ^2 Test with $p < 10^{-11}$). This dependency is referred to as "systematic confusion" in the following and it is discussed in more detail in Section 3.2.2.

2. False positives for non-targets with the *same direction* as the target: No significant effect was found for non-targets with a correct direction but with different pitch in comparison to other non-targets ($p = 0.13$), although the average classifier outputs were again more negative. Multiclass selection errors toward a decision with a correct direction but wrong pitch were not accumulated.

According to these results, the classifier could resolve the dimension 'pitch' better than the dimension 'direction' which also stands in line with the findings by Halder et al. (2010).

Four subjects (*VPnv, VPnz, VPoc, VPoe*) had a sudden drop of multiclass accuracy within the online phase. The exact reason for that effect remains unclear. Technical problems as well as physiological instabilities or lack of concentration may have caused this effect, but could be neither found, nor excluded. Experiments for *VPnv, VPoc* and *VPoe* were stopped after the drop of accuracy.

3.1.4 Conclusions

It is clear, that the stimulus characteristics have a strong impact on the BCI performance. The decision for a 3x3 design was partially driven by the possibility to use a T9-like text encoding system, even though other designs could potentially be better in terms of signal-to-noise ratio. Prior work (Schreuder et al., 2010) showed, that both stimulus types (direction and pitch) contain valuable information for a discrimination task, and that a redundant combination can enhance the separability compared to the single stimulus types.

The results show that the PASS2D paradigm offers fast spelling speed (avg. of 0.89 chars/min and 3.4 bits/min) and an intuitive interaction scheme while being driven by simple stimuli from headphones.

Using state-of-the-art machine learning approaches for ERP classification (Müller et al., 2008; Blankertz et al., 2011), the individual discriminative ERP signatures of subjects could be exploited reasonably well and in real-time and most participants could spell two complete sentences during a single session.

Although among the fastest currently available audio paradigms for BCI, the present work is not reaching the ITR level of these visual paradigms yet, but it is not far from this performance. As the line of research of auditory BCI is relatively young, the potential future development is promising. Moreover — as pointed out in Section 3.1.1 — it represents a qualitatively new solution for end users with visual impairments.

Moreover, the presented paradigm follows principles of user-centered design (Zickler et al., 2011). Firstly, this is expressed by the decision to use a T9-like text entry method. The spelling process in PASS2D is easy to understand and widely known to naïve users because of its similarity to T9-spelling in mobile phones. Moreover, it implements a predictive text entry system, which improves the spelling speed and usability.

Secondly — although the spatial dimension as a class discriminative cue could be exploited more fine-grained (cp. to the approach of (Schreuder et al., 2010) using up to 8 spatial directions) — the PASS2D approach was restricted to three directions only. Taking this decision, the hardware complexity and space requirements for the setup of the system at a patient's home can be reduced, as three directions can be implemented by off-the-shelf headphones and simple stereo sound cards.

Thirdly, the PASS2D paradigm has the potential to adapt to its user in terms of the underlying language model: the predictive text system can consider individual spelling profiles via updates of the text corpus. This implements an important aspect of flexibility, as patients tend to use a lot of individual abbreviations of frequently used terms in order to speed up their communication.

Fourthly, the presented speller design is flexible with respect to the sensory modality. Although operated as a spelling interface with auditory feedback, the interaction scheme is well suited also for visual ERP stimuli or control via eye-tracking assistive technology with full visual feedback. In combination with suitable visual highlighting effects (Hill et al., 2009; Tangermann et al., 2011b; Kaufmann et al., 2012), the graphical representation of the speller (see Fig. 3.1) can directly be used to elicit ERP effects by a visual oddball. Thus, patients in LIS with remaining gaze control could use both, a visual or a hybrid (R. Millán et al., 2010) visual-auditory version of the speller. With a progressing neurodegenerative disease, a further decrease in gaze control or daily changing conditions, the patient has the opportunity to switch from the visual to the hybrid or to the purely auditory setting. As the elicited ERPs are expected to change during this transition, the underlying feature extraction and classification should of course be adapted. If it is possible to perform this transition in a transparent manner, patients can simply continue to use the same interaction scheme independent of the stimulus modality in action.

It is concluded that this auditory ERP Speller enables BCI users to kick-start communication within a single session and thereby offers a promising alternative for patients in LIS or CLIS. The next step will be to further simplify the spelling procedure such that it allows a purely auditory navigation.

Future work will also be conducted to further improve the paradigm with respect to spelling speed, pleasantness, intuitiveness and applicability for patients in locked-in state and complete locked-in state. Experiments with patients are planned.

3.1.5 Lessons Learned

⋆ *Stimuli that vary in the two dimensions (here: pitch and directions) can be used for a BCI. Thus, the user is able to combine those two cues and a class discriminative ERP response is elicited only, if both dimensions match the target (e.g. high-pitch tone, presented from the left).*

⋆ *Distinctive stimulus properties such as the direction of an auditory stimulus can be reflected in the discriminative patterns of the ERP. Therefore, it might be beneficial to incorporate such information into the classification – see also Chapter 4*

⋆ *By introducing a language model into the BCI system, one can increase the information transfer rate of the BCI system by at least 24%.*

3.2 Natural Stimuli can Improve Performance and Neuroergonomics

T<small>HIS</small> part describes how to improve auditory BCI paradigms with respect to ergonomics and performance by using natural stimuli. Moving from well-controlled, brisk artificial stimuli to natural and less controlled stimuli seems counter-intuitive for event-related potential (ERP) studies. As natural stimuli typically contain a richer internal structure, they might introduce higher levels of variance and jitter in the ERP responses. Both characteristics are unfavorable for a good single-trial classification of ERPs in the context of a multi-class Brain-Computer Interface (BCI) system, where the class discriminant information between target stimuli and non-target stimuli must be maximized.

For the application in an auditory BCI system, however, the transition from simple artificial tones to natural syllables can be useful despite of the variance introduced. In the presented study healthy users (N=9) participated in an offline auditory 9-class BCI experiment with artificial and natural stimuli. It is shown that the use of syllables as natural stimuli does not only improve the users' ergonomic ratings, also the classification performance is increased. Moreover, natural stimuli obtain a better balance in multi-class decisions, such that the number of systematic confusions between the nine classes is reduced. Those findings may contribute to make auditory BCI paradigms more user-friendly and applicable for patients. The data and results were previously published in Höhne et al. (2012).

3.2.1 Motivation

The AMUSE paradigm (Schreuder et al., 2011a) and the PASS2D paradigm (see Section 3.1 and Höhne et al. (2011a)) are amongst the best performing (i.e. featuring a fastest information transfer rate) auditory paradigms. Both approaches utilize rather brisk and artificially generated tones to elicit auditory ERP responses, with 6 tones of 40 ms duration (AMUSE) and 9 tones of 100 ms duration (PASS2D). The spatial direction of stimulus presentation as well as the pitch of stimuli were used to code for the multi-class paradigm. Though suited for relatively fast text entry, two practical drawbacks were observed that were related to this choice of stimuli.

Firstly, these highly controlled and very uniform tone sets were perceived as little intuitive and were – by single users – even described as unpleasant (Höhne et al., 2011a). Taking into consideration that such ratings might indicate a limited overall acceptance of a final BCI spelling system, but also that the motivation of users is correlated with BCI performance (Kleih et al., 2010; Tangermann et al., 2011b), an improvement of such subjective user ratings must be sought.

Secondly, a posterior analysis of the online spelling performance in both paradigms revealed a number of systematic multi-class confusions in the classification of target vs. non-target stimuli. A systematic confusion is present if there are two classes i and j which are confused by the BCI more often than other pairs of classes. Systematic confusions might arise from flaws in the stimulus design: for the AMUSE paradigm, these mis-classifications were related to front-back confusions. In the PASS2D paradigm, stimuli were confused that share some characteristics (e.g. pitch or direction).

Even though visual ERP paradigms have undergone improvements by stimulus optimization (Allison and Pineda, 2003; Sellers et al., 2006a; Hill et al., 2009; Tangermann et al., 2011b; Kaufmann et al., 2011), the stimulation principles for auditory BCI paradigms – as a relatively young line of research – were only rarely investigated. Initial attempts to compare different auditory stimulation principles can be found in Schreuder et al. (2009), Halder et al. (2010), and Höhne and Tangermann (2011a).

The goal of this section is to tackle both of the above mentioned problems (low user acceptance and confusion) simultaneously by improvements on the level of stimulation. For this purpose a comparison between three auditory stimulus sets is performed within the PASS2D paradigm.

3.2.2 Experiment 2: Improving auditory BCIs with Natural Stimuli

Participants

Nine healthy subjects (age: 24–26) participated in an offline BCI experiment comprising a single session of EEG recording. Two of the participants (*VPmg* and *VPlg*) had already participated in earlier BCI experiments. Each participant provided written informed consent, did not suffer from a neurological disease and had normal hearing. Subjects were not paid for participation.

Experimental Design

Within a single session, three conditions (i.e. three different sets of stimuli, see Sec. 3.2.2) were compared. The session was divided into several blocks that lasted approx. 10 min including a short break. Subjects were asked to perform six blocks at least, but they could decide to extend the recording in steps of three blocks. Two subjects performed nine blocks, while seven subjects chose to perform 6 blocks only.

Every block consisted of nine trials, with three consecutive trials showing the same type of stimulus (same condition). The order of conditions within trials was block-randomized. A trial was defined as a sequence of 135–144 auditory stimuli, subdivided into 15–16 iterations. With a single iteration resembling a complete set of nine stimuli in random order, each trial contained 15–16 target stimuli and 120–128 non-target stimuli.

During data preprocessing (see Sec. 2.3.1) only the last 14 iterations of each trial were considered. Removing the initial one or two iterations compensated for starting effects, such as orientation time necessary to direct spatial auditory attention to the target tone.

Participants were asked to concentrate on the occurrences of the target stimuli and to neglect all other (non-target) stimuli. In addition, they were asked to count the targets and to report the number of occurrences at the end of each trial. Prior to the start of a new trial, the target stimulus was cued by three repetitive presentations. Targets were pseudo-randomized between trials, such that the number of targets was balanced between the nine classes.

Behavioral Data

After the EEG recording, the participant filled out a questionnaire and rated each condition on a visual-analog scale, answering six questions per condition (translated from German):

Q1: "How motivating does condition x appear to you?" (**Motivation**)
Q2: "How do you judge your concentration while attending to stimuli in condition x?" (**Concentration**)
Q3: "How tiring is condition x?" (**Tiring**)
Q4: "How difficult was it to discriminate the stimuli in condition x?" (**Discrimination**)
Q5: "How exhausting is condition x?" (**Exhaustion**)
Q6: "What is your overall impression of condition x?" (**Overall**)

The scales were designed such that negative features (such as "hard to discriminate", "very exhausting" or "very tiring") were assigned low scores. To deliver a rating, subjects had to set a mark on a

line of 10 cm length, which represented a continuous scale between the most negative and positive outcomes of each question.

Stimuli

Three different sets of auditory stimuli were used, forming the three conditions: (1) artificially generated tones, (2) spoken syllables and (3) sung syllables. Each set consisted of nine stimuli with spatial characteristics.

- The stimuli of condition 1 had already been successfully applied in an online study (Höhne et al., 2011a). The nine artificially generated stimuli consisted of three tones with different pitch (high/medium/low) and also a varying tonal character. Each of the three tones was presented from three different directions (left/middle/right), leading to the 3 × 3 design shown in Fig. 3.5a.

- For condition 2, short spoken syllables were recorded by three speakers, visualized in Fig. 3.5b. Each speaker recorded three stimuli: syllables that either contained the vowel "i", an "æ" or an "o", like {ti, tæ, to, it æt, ot}. To obtain an intuitive separation of the stimuli, every speaker was presented only from one fixed direction (base: from the left, tenor: from the middle, soprano: from the right). Thereby the 3×3 design of the PASS2D paradigm (Höhne et al., 2011a) was maintained since a column represented a speaker/direction and each row represented the vowel {"i", "æ" or "o"}, see Fig. 3.5b. The three different vowels lead to an intrinsic difference in the higher order harmonics, but the stimuli in condition 2 were all spoken and had no explicit pitch differences.

- For condition 3, the stimuli were recorded similar to condition 2. The only difference was that the syllables were not spoken, but sung by the same voices as in condition 2. Syllables with an "i" were sung with high pitch ($A\#$), syllables with an "æ" were sung with medium pitch (F), syllables with an "æ" were sung with low pitch ($C\#$)[2].

All stimuli were generated/recorded such that they lasted 100 ms (condition 1) or 125 ms (condition 2-3). Condition 2 was considered as an intermediate condition and the transition from condition 1 to condition 3 is focused in the following, as it denotes the step from the maximally standardized artificial stimuli to the most complex natural stimuli.
Stimuli for multi-class auditory BCI paradigms are generally designed such that they are easy to discriminate on the one hand, but also similar enough to evoke at least similar target and non-target responses for each stimulus. In contrast to the artificial stimuli with a well-defined onset (condition 1), the natural stimuli used in condition 2 and 3 had an intrinsic temporal diffuse characteristic, as shown in Fig. 3.6. Thus, the uniform and artificial stimuli in condition 1 didn't vary over time, while the stimuli in condition 3 (syllables) had a rather complex and heterogeneous temporal structure. However, the syllables in condition 3 were recorded and aligned such that vowels started at the same time (i.e. 30 ms) in each stimulus. This alignment had two advantages: a sequence of stimuli was then perceived to be rhythmic and the class discriminative information in the stimuli was aligned. The time-frequency spectrograms in Fig. 3.6b show this alignment.

All stimuli were presented with a stimulus onset asynchrony (SOA) of 130 ms. A Terratec DMX 6Fire USB sound card was used for stimulation, and light neckband headphones (Sennheiser PMX 200) enabled a comfortable audio perception. The mean latency of 51.4 ms (median: 50.5 ms, std: 4.46 ms,

---

[2]The chosen pitches would result in a consonant chord, when they were played together.
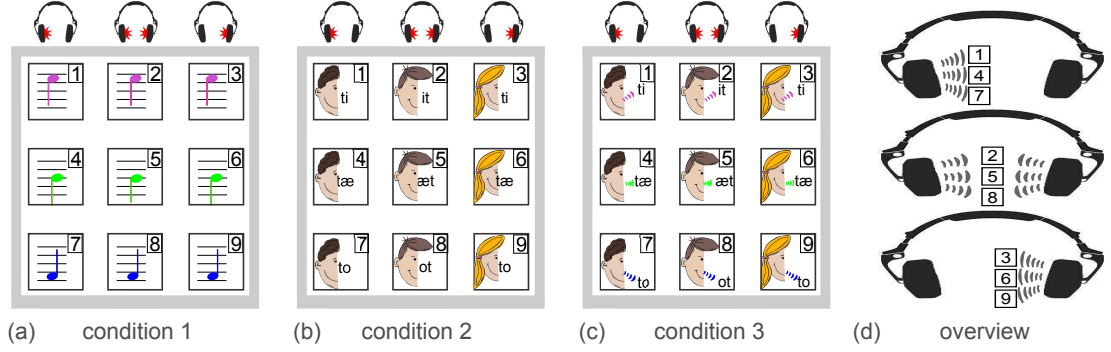
Figure 3.5: *Graphical representation of the three sets of auditory stimuli used for Experiment 2.*
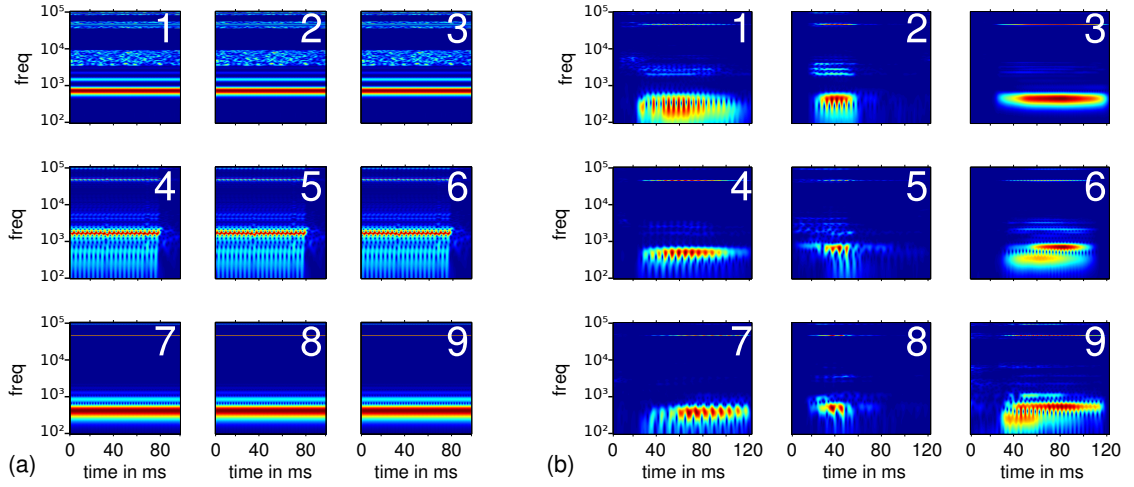
Figure 3.6: *Spectrograms of auditory stimuli used for Experiment 2. Subplot (**a**) shows the spectrograms of three artificial tone stimuli used for condition 1. Stimuli were the same for the left, right and binaural presentation. Subplot (**b**) shows nine different stimuli used in condition 3, which consisted of sung syllables. In this condition, the directional presentation was supported by the use of different singers for the left, right and binaural stimuli.*

min: 41.2 ms, max: 61.8 ms) was corrected before the start of the data analysis. Pseudo-random sequences of stimuli were generated such that two subsequent stimuli were neither in the same row, nor in the same column (cp. to the 3 × 3 design shown in Fig. 3.5). As an example, none of the stimuli {5,6,1,7} was presented immediately after stimulus 4 had been presented. This constraint was implemented to prevent a consecutive presentation of two stimuli that share the speaker identity, pitch or direction. The stimulus presentation was programmed in Python and embedded in the PyFF framework (Venthur et al., 2010).

Data Acquisition and Preprocessing

EEG signals were recorded with a Fast'n Easy Cap (EasyCap GmbH) using 63 monopolar, wet Ag/AgCl electrodes placed at symmetrical positions based on the extended international 10-20 system. Channels were referenced to the nose. EOG signals were recorded via bipolarly referenced electrodes (vertical

EOG: electrode Fp2 vs. an electrode directly below the right eye; horizontal EOG: F9 vs. F10). Two 32-channel amplifiers (Brain Products BrainAmp) processed the signals by an analog bandpass filter between 0.1 Hz and 250 Hz before digitalization (sampling rate 1 kHz). After applying the analog filter, the EEG raw data was first high-pass filtered at 0.2 Hz, then low-pass filtered at 25 Hz, both by a causal Chebyshev filter. For details, see also Section 2.3.1.

The EEG response to one stimulus is called subtrial in the following and comprises the most informative time period of 800 ms starting with the stimulus onset. DC offsets were subtracted based on the mean offset in a baseline interval of -150 ms to 0 ms relative to the stimulus onset. As 14 iterations of nine stimuli were contained in one trial, and three trials of each condition belonged to one block, the number of subtrials (before artifact rejection) was $14 * 9 * 3 = 378$ per block and condition, summing up to $6 * 378 = 2268$ subtrials for seven of the subjects, and $9 * 378 = 3402$ for two subjects.

Eye-artifacts were excluded by applying a moderate min/max-threshold criterion: subtrials were rejected if their peak-to-peak activity in at least one of the EOG channels exceeded 80 μV. On average over the subjects, this criterion lead to a rejection of 5.5 % of artifactual subtrials, while approximately maintaining the 1:8 ratio of targets and non-targets.

Features and Classification

For each subtrial of the preprocessed EEG signals, a feature vector was obtained by computing the average amplitudes of 19 predefined intervals for all electrodes, resulting in 19 intervals × 63 channels = 1197 features per subtrial. The intervals are marked in the top plot of Fig. 3.9. Short time intervals of 30 ms length were chosen to cover earlier ERP components, while broader late components are sampled more coarsely by intervals of 60 ms length.

Binary classification of target and non-target epochs was performed using a linear Discriminant Analysis (LDA) regularized with covariance shrinkage – see Section 2.4 for details. All subtrial epochs that survived the previous artifact rejection step (see Sec. 3.2.2) were used to estimate the classification performance. In order to account for the imbalance between targets and non-targets (ratio 1:8), the class-wise balanced accuracy is reported. It describes the average decision accuracy across classes (target vs. non-target) and has a chance level of 50 %. The binary classification accuracy was estimated by a 5-fold cross-validation procedure, which itself was repeated five times with random shuffling of the epochs (5 × 5 cross-validation).

Any performance comparison between the three conditions is based on EEG channels only. In addition, classification performance is estimated exclusively for EOG channels and for EOG combined with EEG. The latter two combinations are used only to upper-bound the unwanted influence of potential eye-related artifacts to an EEG-based system.

Simulation of Information Transfer Rates

It is noteworthy that the stimuli were presented in a rapid sequence (SOA of 130 ms), therefore already lower binary classification accuracies may result in sophisticated communication rates. To compare the communication speed across several BCI paradigms, the Information Transfer Rate (ITR) is widely used metric – see Section 2.6.3 for details.

The simulation targeted the ITR of a single block of online use of a BCI system. An online multi-class BCI experiment of 100 hours duration was simulated for each subject and each condition. Therefore, classifier outputs for target and non-target events were generated according to the binary accuracy, which was derived from the offline data analysis (see Sec. 3.2.3). Based on generated classifier outputs, trials were simulated and a multiclass decision was made as soon as an early-stopping criterion was fulfilled, at the latest after 20 iterations. For details on the dynamic stopping method, see Schreuder et al. (2013a) [Höhne method]. A fixed inter-trial pause of 7 seconds was added in the simulation

each trial, assuming that subjects need time to shortly relax and re-orient their attention to the next tone. The ITR was then computed based on the number of correct and incorrect decisions after the simulated online BCI experiment.

Quantification of Systematic Confusions
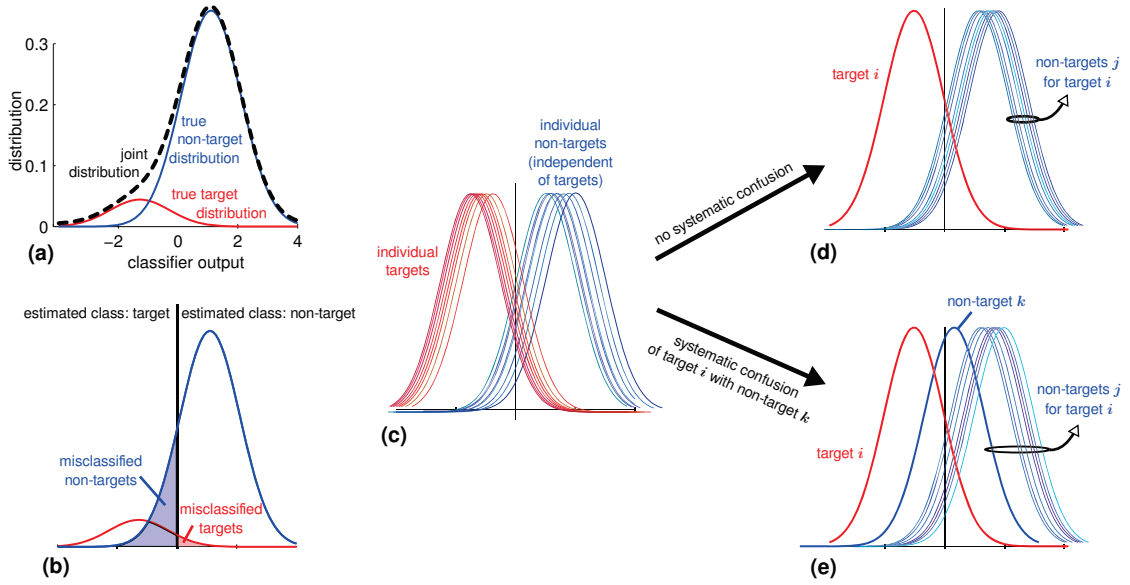
Figure 3.7: *Schematic visualization of distributions of classifier outputs. Plot (**a**) depicts the distributions of classifier outputs for targets and non-targets, when all nine stimuli are pooled together. The distributions of misclassified stimuli are visualized in plot (**b**). Plot (**c**) shows the (rescaled) distributions for the 9 possible target and non-target stimuli. These distributions of classifier outputs disregard the trial structure, i.e. the distribution of non-targets j are relatively independent of a specific choice of a target i. In contrast, the plots (d) and (e) consider the trial structure. Here, distributions of non-targets j do depend on the choice of a specific target i. Plot (**d**) depicts a situation where there is no systematic confusion (no increased probability of a misclassification) between target i and any of the non-targets j. Plot (**e**) shows another example, where there is a systematic confusion between target i and the non-target k.*

This section deals with the question, whether or not there are pairs of stimuli that are more difficult to discriminate than others. One can pose the same question from the classifier's point of view by asking, whether or not there are pairs of stimuli that are more likely to be confused by the classifier than others. This phenomenon will be called "systematic confusion" in the following.

The problem of systematic confusions cannot be investigated with a measure for binary (target vs. non-target) classification accuracy alone, as shown in Fig. 3.7: in the depicted simulations, a binary classification accuracy of 90% is simulated in a 9-class paradigm. This is visualized in plot (a-c), where the red and blue curves show the distributions of classifier outputs (cl_{out}) for target and non-target stimuli. The shaded areas in (b) depict the fraction of binary misclassifications which is 10% for both classes, with value 0 being the classification threshold. Investigating only plots (a-c) – which is a visualization of the binary classification accuracy – it is not possible to evaluate systematic confusions, since both situations plotted in (d) and (e) can evolve from distributions described in Fig. 3.7a-c.

In order to evaluate systematic confusions, multi-class confusion matrices might be of limited help. Reflecting the worst cases (misclassifications) only, those matrices are unable to provide information

about systematic similarities between some target- and non-target subclasses. Instead, one can perform an introspection of the distributions of cl_{out}: the cl_{out} for all non-targets j have to be considered, when the user was focusing on target i. If there are no systematic confusions, then the distributions of cl_{out} for any non-target j is equally distributed and independent of the target stimulus i, as shown in Fig. 3.7d. If there are systematic confusions, then the distributions of cl_{out} for a non-targets j depend of the target stimulus i. Thus, when the user is attending to target i, there will be a non-target k which the BCI will classify more likely as target than other non-targets j (see Fig. 3.7e).

In the following paragraph, it will be described how to statistically quantify the systematic confusions that were described above. In a typical BCI scenario, stimuli are presented in iterations, where in one iteration, each stimulus is presented exactly once in a pseudo-random order. In the given 9-class scenario, there is one classifier output for a target stimulus i and a classifier output for each of the 8 non-target stimuli in every iteration. The non-target j with the smallest (i.e. most negative) classifier output is denoted as the "worst non-target"($wNT_{j|i}$) in the following, as it is seen by the classifier most likely as the target. In the ideal case without systematic confusions, $wNT_{j|i}$ is independent of target stimulus i, as shown in plot 3.7c and the probability of being the "worst" non-target is ($p(wNT_{j|i}) = 1/8$) for each pair {i,j} and in each iteration. This can be described with a Bernoulli distribution [3]

By accumulating $wNT_{j|i}$ across iterations, one can obtain the number of times that non-target j is the "worst" for target i, being referred to as $nNT_{j|i}$ in the following. In the situation without systematic confusions, $nNT_{j|i}$ is a binomial distributed random variable with $p = 1/8, k = nNT_{j|i}, n = nT_i$, where nT_i denotes the number of sequences with i being target.

Hence, if there is a systematic confusion between target m and non-target l, then $nNT_{m|l}$ does not follow a Binomial distribution[4],

$$f(k;n,p) \quad = \quad \binom{n}{k}p^k(1-p)^{n-k} \tag{3.1}$$

$$\binom{n}{k} \quad = \quad \frac{n!}{k!(n-k)!} \tag{3.2}$$

It is tested across all iterations and all subjects, whether or not $k = nNT_{i|j}$ follows a binomial distribution for any pair {i,j}. This can be tested by a significance test with a p-value of 0.05. While this test assumes the iterations to be independent, it does not require of the overall classification accuracy to be equal for each subject or iteration.

3.2.3 Findings

Behavioral Data

The subjective ergonomic ratings for the three stimulus conditions were assessed by questions Q1–Q6. These ratings as well as the objective counting accuracy show a clear trend: it was easier for the subjects to concentrate on natural stimuli (conditions 2, 3) than on artificial stimuli (condition 1). Participants rated the stimuli of condition 3 significantly more positive than condition 1 for each

---

[3]The Bernoulli distribution describes a probability of a binary random variable. This distribution is mostly used to model the outcome of a binary coin toss, where p denotes the probability of a success (e.g. "head") and $q = (1-p)$ p denotes the probability of a failure ("tail"). The single event of flipping a coin is also referred to as a "Bernoulli trial".

[4]The Binomial distribution is the discrete probability distribution which describes the outcome of sequence of independent Bernoulli trials. In the example of coin toss, f(k;n,p) describes the probability of observing k successes (e.g. "head") when flipping the coin n times with a success probability p
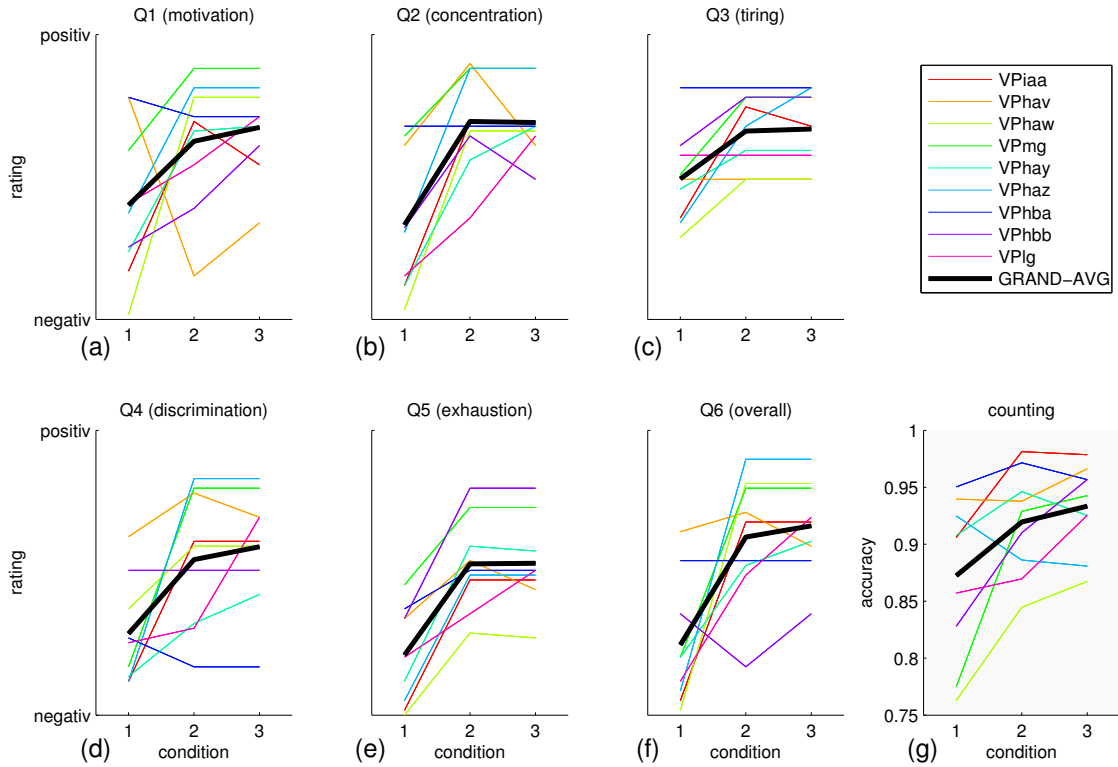
Figure 3.8: *Overview over the behavioral data collected from each subject and the grand average. The subjective ratings of ergonomic aspects of three conditions are shown in subplots (a)–(f). Relative differences between reported counts and the true number of target stimuli are depicted in subplot (g).*

question (paired t-test, $p < 0.05$). Exemplarily (Fig. 3.8d), all participants except *VPhba* rated the stimuli of condition 3 to be easier to discriminate than the stimuli of condition 1. The same trend can be seen for all other ergonomic ratings and also in the counting performances for the three conditions: the participants gave better ergonomic ratings and reported the number of targets more accurately for natural stimuli than for artificial stimuli.
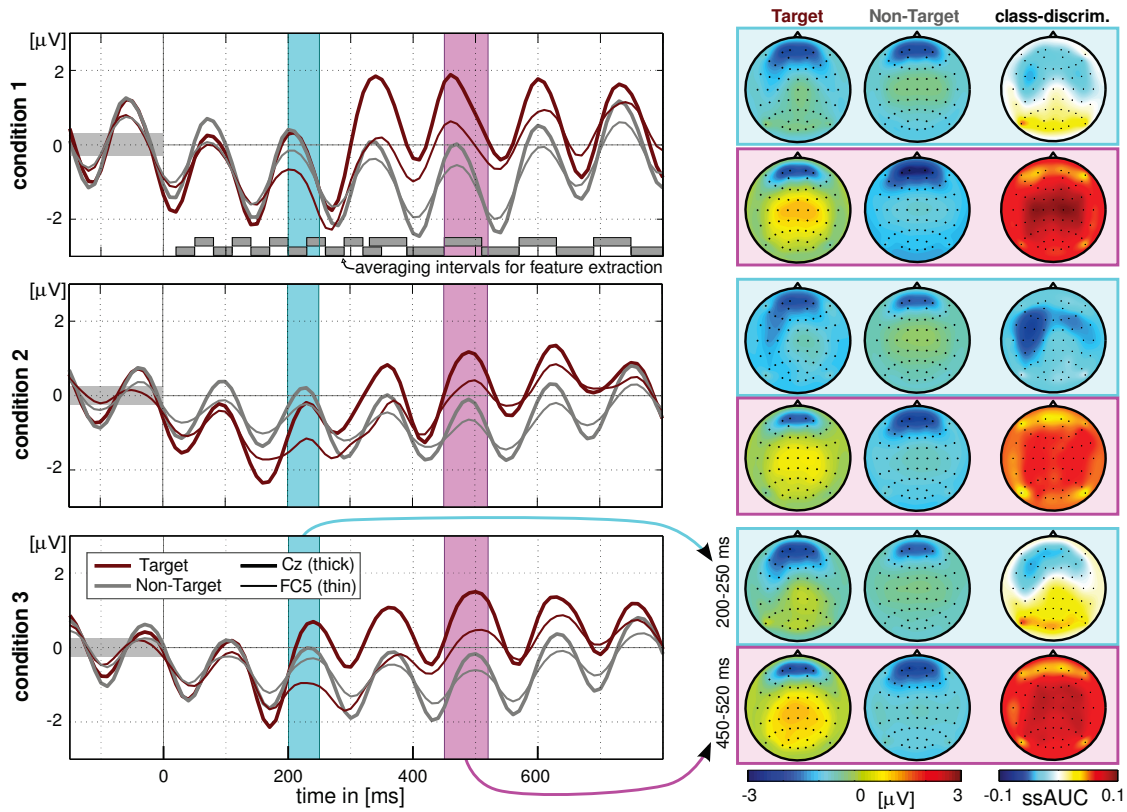
ERPs

Fig. 3.9 shows time series of the event related potentials (ERPs) averaged over all participants for the three conditions. As expected, ERPs for non-target stimuli (gray lines) in all three conditions show a regular pattern that mainly reflects early processing of the auditory stimuli. However, this regular pattern is not the same between the conditions, as a phase shift can be observed: regular responses for condition 1 have a shorter latency (approx. 30 ms) than responses for conditions 2 and 3. As a result, the peaks of steady state responses are slightly shifted to the left in condition 1.

The grand average spatial distribution of target- and non-target responses are shown in the scalp maps of Fig. 3.9. In addition, the class-discriminant information between target- and non-target responses was quantified with a signed and scaled measure of *area under the ROC curve,* called ssAUC. It is visualized as a third scalp map per condition and interval.

A common observation for all conditions is the appearance of a class-discriminant early negative component. It is centered in fronto-temporal areas around 200 ms post stimulus onset. Recent psychophysiological literature (Gamble and Luck, 2011) describes the same – or similar – early

Figure 3.9: *Grand average ERPs for target and non-target responses of the three conditions, observed for Experiment 2.* **Time series plots (left):** *From top to bottom, conditions 1–3 are visualized. For each condition, the average target and non-target responses are depicted for two EEG channels (FC5 and Cz). Two time intervals were marked in the time series plots: light blue intervals with a range of 200–250 ms after stimulus onset and light magenta intervals ranging 450–520 ms. The gray blocks in the top plot mark the 19 time intervals used for feature extraction in the classification task.* **Scalp plots (right):** *For each condition and both colored intervals, three scalp plots are provided. They depict the average ERP activity for targets, non-targets, and of the distribution of class discriminative information (ssAUC).*

negative discriminative components in a spatial auditory multiclass paradigm as N2ac components. Common is also the existence of a class discriminative late positive component. It shows a centro-parietal distribution starting around 300 ms and extends up to 700 ms. Its distribution resembles that of a P3b component, but appears much later than in standard oddball paradigms with slower stimulus presentation and less different classes.

Although largely similar, the scalp plots vary in details between conditions. One can observe a trend of increased lateralization of class discriminative early negative components to the left hemisphere for the natural stimulus conditions 2 and 3. The rightmost column of Fig. 3.9 suggests that this effect is strongest for spoken syllables (condition 2).
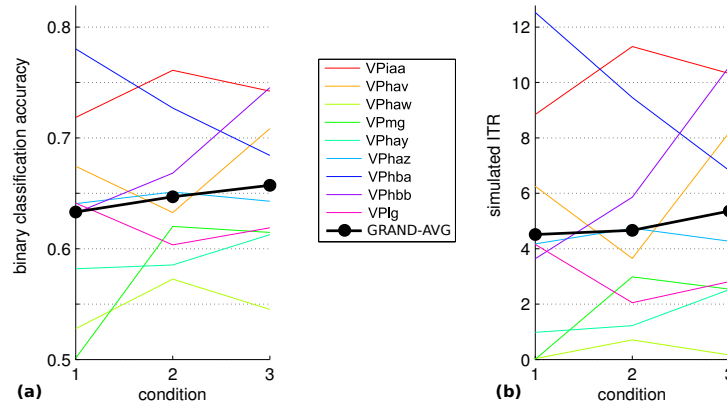
Figure 3.10: *In subplot (**a**) the estimated binary classification accuracy is depicted for each subject (thin lines) and for the grand average (thick black line) for three conditions. Subplot (**b**) compares the resulting simulated Information Transfer Rate (ITR) in bits per minute for each subject and the grand average.*

Classification Accuracy

The binary classification accuracy was computed for each subject and each condition with the finding that stimuli of condition 3 obtained a higher average accuracy than condition 1 and 2, see Fig. 3.10a. Over all participants, the class-wise balanced accuracy was between 50% (chance level) and 78%. Among the nine tested subjects, *VPmg* would profit in a special way from the two new conditions: evoked potentials can be classified clearly above chance level using natural tones, while this was not possible for the artificial tones. It can be observed that the performance curve of subject *VPhba* behaves against the general trend. This participant was also the only one who reported that it was easier for him/her to concentrate on stimuli of condition 1 than on the natural stimuli in conditions 2 and 3 (see Fig. 3.8d).

Individual scalp maps of class-discriminant activity did not give rise to conjecture a substantial influence of EOG activity to the classification results. However, since unconscious saccades and head movements in response to spatial auditory targets were already discussed in Röttger et al. (2007), the impact of EOG activity was double-checked by estimating (1) the classification performance on the two EOG channels only, and (2) on the combined EEG+EOG channels.

In scenario (1), average classification performances were 54.1%, 53.8% and 54.6% for the three conditions. Although located close to chance level (50%), the two EOG channels seemed to contain a small amount of task-relevant information. For comparison, the average absolute performance for EEG channels was about 10% higher (63.4%, 64.7% and 65.8%). The difference between scenario (1) and only EEG channels is significant.

In scenario (2), the difference in classification performance between using EEG channels only, and EEG channels plus EOG channels was very small and not significant for any condition.

Taken together, these results point out that EOG channels did not provide any additional information compared to EEG channels. The small amount of class discriminative information contained in the EOG channels probably represents EEG activity that is picked up by the electrodes Fp2, F9 and F10.

## Simulated Information Transfer Rate

The simulated ITR values (see Section 2.6.3) for each condition were based on two assumptions: (1) the binary classification is constant over time and (2) there are no systematic confusions. Fig. 3.10b

depicts the outcome of the simulation: the average simulated ITR increases from 4.51 bits/min to 5.31 bits/min by the transition from artificial stimuli (condition 1) to natural stimuli (condition 3), being highly competitive for gaze-independent BCIs.

Temporal and Spatial Distribution of Discriminative Components
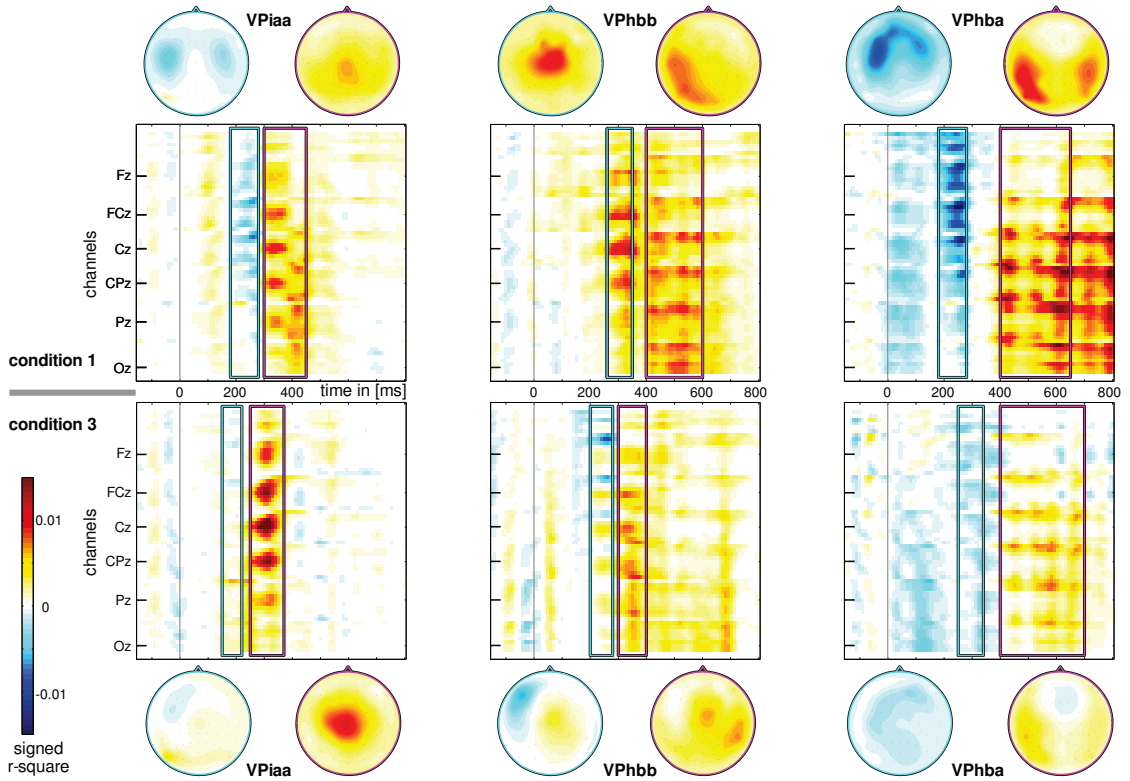
Fig. 3.11 shows, how class discriminative information is distributed over time. A comparison between the shapes of lines reveals that the time structures are more similar between the two types of natural stimuli (red vs. blue) than between the artificial and the natural stimuli (black vs. red/blue). The blue line is above the red line for most subjects, which indicates a generally increased class-discrimination for condition 3 compared to condition 2. The black curve shows different peaks than the blue and red curve for several participants. This either indicates a temporal shift in components or the existence of different components when comparing the artificial stimulus condition to natural stimulus conditions. Noteworthy are the differences visible in the curves of subjects *VPiaa*, *VPhbb*, and *VPhba*.
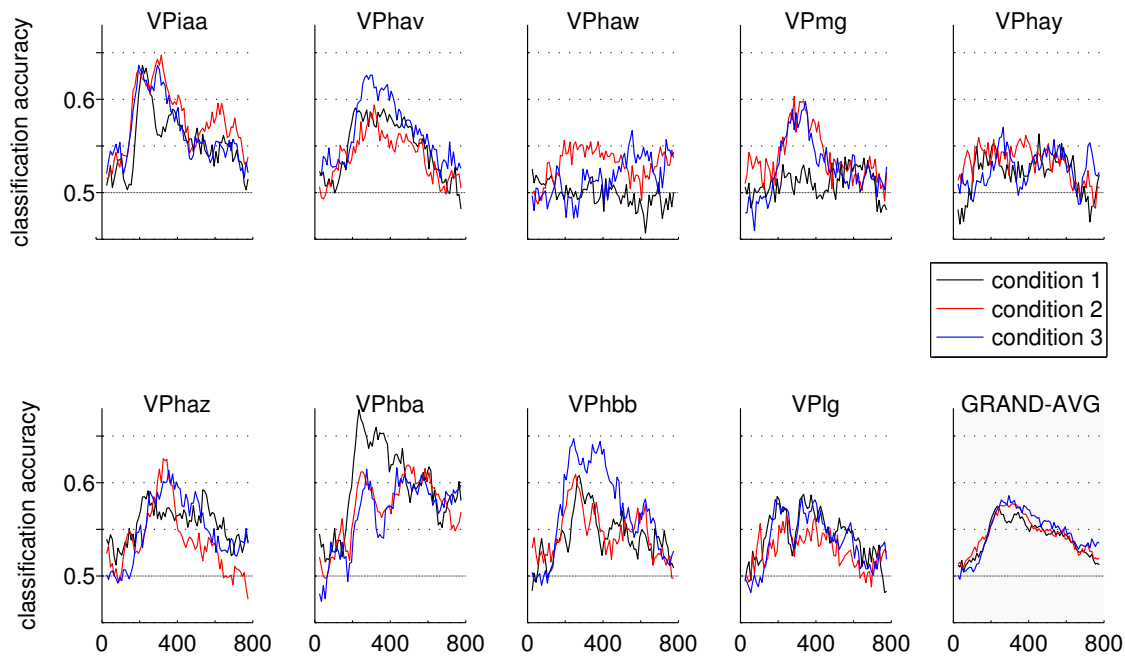
Figure 3.11: *Distribution of discriminative information over time. For each subject and the grand average, class discriminative information contained in all channels is estimated over time and compared for the three conditions. Classification accuracy is estimated within a sliding window of 50 ms width, which is used to scan the epoch. Most information is contained in the time windows around 300 ms after stimulus onset.*

In Fig. 3.11, it can be seen that *VPiaa* shows two distinct discriminative peaks for the natural sound condition. The first peak is centered at approximately 200 ms and the second about 150 ms later, at 350 ms. The second peak is strongly attenuated in the artificial sound condition and also about 50 ms delayed compared to the natural sound conditions. Subject *VPhbb* also shows two distinct peaks in the time-resolved classification plots in Fig. 3.11. For this subject it is the earlier peak that is attenuated in the artificial sound condition. Finally, in subject *VPhba*, an effect that is contrary to other subjects was observed: the transition from condition 1 to condition 3 leads to a weakening of both components.

For the three mentioned subjects, the individual spatio-temporal dynamics of class discriminative

Figure 3.12: *Spatio-temporal distribution of class discriminative information for three selected subjects (arranged in columns) and for two conditions (arranged in rows). For each combination, one matrix plot and two scalp plots are provided. All plots share the same color scale. A matrix plot shows signed r-square values for each EEG channel (y-axis) and time bin (x-axis). Channels are sorted from front to back and left to right, with occipital channels located in the bottom rows. The two scalp plots depict averaged r-square values for two individually chosen time intervals, capturing early and late class discriminative components. Their positions in time are marked by light blue and light magenta rectangles in the corresponding matrix.*

information are plotted in Fig. 3.12. The plots show r-square values for each channel and each time point. For two selected time intervals, the temporal averages of r-square values are visualized as scalp plots.

For subject *VPhbb*, an early negative discriminative component is found in condition 3, which is entirely absent in condition 1. While a positive component (P300) was found in both conditions, the spatial distributions vary. Changing from condition 1 to 3, neither the position on the scalp, nor timing or intensity of class discriminant components are maintained for subject *VPhbb*.

For subject *VPiaa*, the transition from condition 1 to 3 leads to an earlier appearance of discriminative components. While the approximate spatial distribution is maintained for both components, the intensity of class discrimination varies for the two conditions: the early component is slightly weakened in condition 3, while the positive component is considerably increased in condition 3 compared to condition 1. For *VPhba*, the intensity of both components decreases by far, while the temporal and spatial characteristics are maintained.

Another interesting aspect that becomes evident in the scalp maps of Fig. 3.12 is a change of shape in early discriminative components. The spatial distribution of r-square values in the early interval is

rather symmetric in all subjects in condition 1. However, in condition 2 and condition 3 the maps are more asymmetric, as the center of mass of the early negative components is shifted towards left fronto-temporal regions. This shift is also visible in the grand-average ssAUC maps of early time intervals in Fig. 3.9, right-most column.

Joint Effects on Classification Performance and Behavioral Data
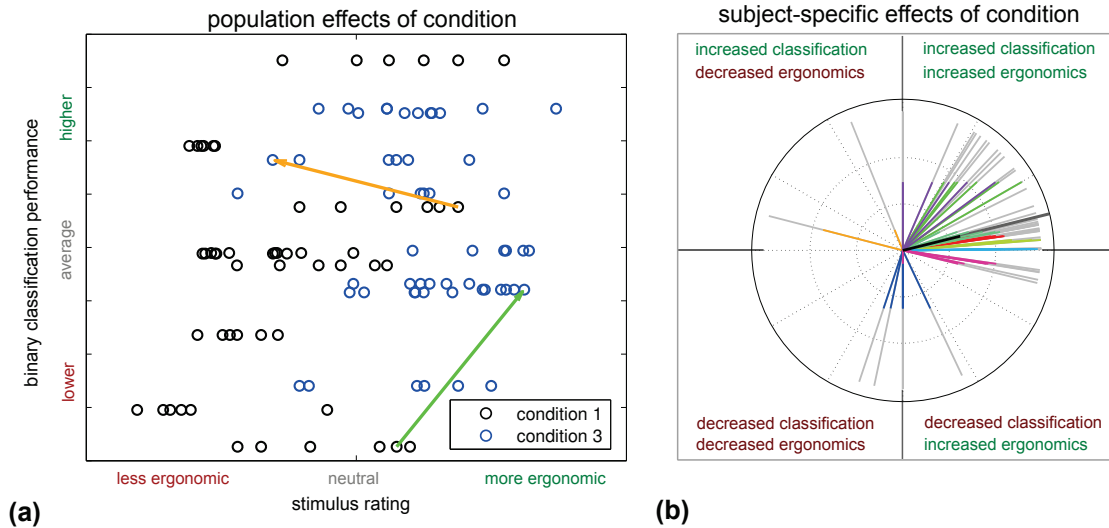
Figure 3.13: *Joint effects of the stimulus condition (1 and 3) on both, classification accuracy and ergonomic ratings. One classification accuracy and six ergonomic ratings expressed by the VAS scores for questions Q1–Q6 were available per subject and condition. All values were standardized by z-scoring and entered into the scatter plot (**a**). The subject-specific changes from condition 1 to condition 3 are depicted in plot (**b**). Gray bars indicate directions of change, while the colored portion of the bar indicates its magnitude (relative to the maximum over all subjects). Subject identity is color-coded as in Fig. 3.8 and 3.10, with black representing the grand average.*

So far, the presented results show that classification performance and ergonomic rating of the stimuli increased when making the transition from the artificial stimuli in condition 1 to more natural stimuli in condition 3. It remains to be shown however, that this effect occurs *simultaneously* in the majority of subjects.

Fig. 3.13 shows the classification performance as well as the stimulus ratings for condition 1 and 3, pooled over subjects and rating questions. In order to preserve visibility, only two conditions (1 and 3) are shown and it is not differentiated between individual subjects or questions. It is difficult to make assertions about joint effects in the ratings, because their respective ranges differ. Thus, the classification performances and the stimulus ratings were standardized by z-scoring (i.e. removing the mean and dividing by the standard deviation). Fig. 3.13a shows that for condition 1 the majority of subjects rated the stimuli less ergonomic and the classification performance was lower, compared to condition 3. Subjects rated each condition with respect to six categories, as shown in Fig. 3.8. This resulted in six data points for each subject and condition. Having the same classification performance, those sample points appear on a horizontal line in Fig. 3.13a. Two sample points (classification/stimulus rating) of condition 1 are connected with their corresponding sample point in condition 3. The arrows always point from condition 1 to condition 3 and thereby mark the effect of the transition from artificial to more natural stimuli. In Fig. 3.13b, such transition arrows are plotted for all subjects

and rating questions, with the color indicating the identity of the subjects (same color code was used as in Fig. 3.10). The vast majority of transition arrows points into the upper right quadrant, which represents a simultaneous increase in stimulus ergonomics and classification performance.

Systematic Confusions

The systematic confusions by the classifier were analyzed according to the method described in Section 3.2.2. Significant systematic confusions are present in all three conditions, but the number of confusions is reduced by natural stimuli in conditions 2 and 3 compared to artificial stimuli in condition 1 (see Fig. 3.14). Condition 3 (as the condition with the best neuroergonomic design) exhibits the smallest number of confusions. It should be noted that the number of systematic confusions is independent of the binary classification accuracy, as shown in Fig. 3.7.
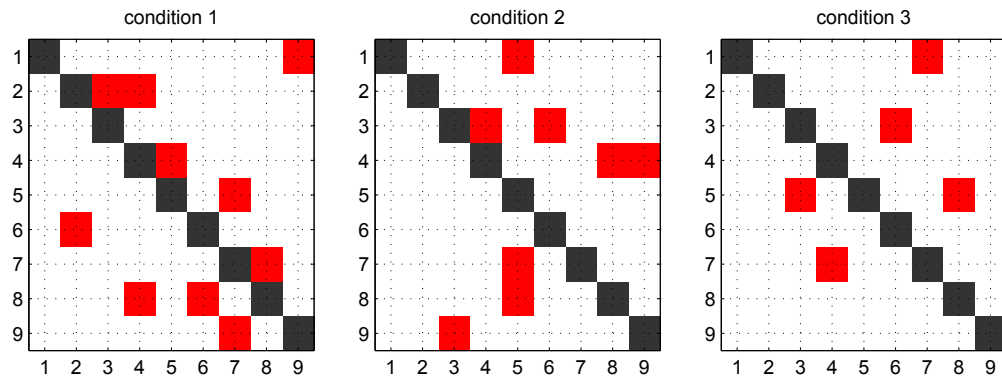
Figure 3.14: *Systematic confusions of stimuli for each condition. A row of a confusion matrix corresponds to one stimulus in the role of a target. A red square with index i, j marks systematic confusions of the non-target j with this target i (p-value of binomial distribution ≤ 0.05). Example: in condition 1, target stimulus 4 is systematically confused with non-target stimulus 5.*

3.2.4 Conclusions

Auditory BCI paradigms are a potential solution for severely motor-disabled patients, as they can be utilized independent of gaze control or eye blinks. Aiming to improve existing auditory BCI paradigms with respect to usability and performance, this study investigates the use of natural auditory stimuli.

The transition from artificial to natural stimuli was motivated by the idea to utilize the humans' over-trained ability of speech processing. First, this comprises the decomposition of a complex auditory stream into relevant components, such as syllables. Second, humans are trained to focus on one out of several voices that are perceived from different directions.

Tones produced by singers offer a large number of class-discriminant cues for the BCI user (e.g. harmonics, pitch, direction, voice-characteristics). Even though the syllables used in this study are more complex and less standardized than the artificial tones, they allow for better classification rates and lead to increased subjective ergonomic ratings. In short, the auditory BCI became "faster" and was considered "more pleasant" when using these more natural stimuli.

Of course it is an interesting question, whether or not the syllables evoke ERP components that are different from those ERP components evoked by artificial tones. In fact one could observe considerable differences between the two stimulus conditions in the individual ERP responses (Sec. 3.2.3) as well as in the grand average responses (Sec. 3.2.3). The delay of 30 ms in the grand average for the time

series of natural stimuli might best be explained best by a delayed perception of natural stimuli. An alternative explanation would be based on an increased mental processing demand for the natural stimuli due to their higher complexity. In the grand average, the trend of increased lateralization of early negative components to the left hemisphere was observed especially for the spoken syllables (Fig. 3.9). This lateralization of language-related processing in the human brain was observed before (Friederici and Alter, 2004) and it is plausible that language-related brain areas become increasingly involved in the processing syllables compared to tones. As the lateralization is best reflected in the ssAUC scalp maps (see rightmost column of Fig. 3.9), this suggests an active role of language-related areas during the discrimination of target and non-target stimuli. Latencies and amplitudes of late positive class discriminative components were rather unstable between conditions, when compared on an individual basis. Assuming that these components might represent P3b components, which are known for their stability, this variation comes unexpected on the one hand. On the other hand, the multiclass setup with short SOA is far from the standard oddball paradigm.

Even though a rather fast stimulation speed (SOA: 130 ms) was applied, the results show that users can handle such a rapid sequence of adequately designed stimuli. It can be assumed, that auditory stimuli can in principle be presented with at least the speed as visual BCI paradigms.

Moreover, this study demonstrates the problem of systematic confusions, which was mostly disregarded by the BCI community so far. A data driven approach to identify and quantify those confusions is presented. Based on this method, it is shown that next to increasing the classifier performance, also the number of systematic confusions can be reduced by a design of stimuli that follows neuroergonomic principles.

Several auditory BCI paradigms for text spelling were recently developed and successfully tested with healthy subjects (Klobassa et al., 2009; Höhne et al., 2011a; Schreuder et al., 2011a). The question, whether or not auditory BCIs are applicable with end-users such as patients suffering from ALS for daily use, remains an open question. However, the above mentioned improvements in the experimental paradigm resemble an important step for the transfer of multiclass auditory BCIs from the lab into the real world and into patients' homes.

3.2.5 Lessons Learned

⋆ *The stimulus design directly impacts the ERPs and the ergonomic ratings.*

⋆ *Sung syllables serve as suitable stimuli for auditory BCI paradigms. Even when they are presented in a rapid sequence (SOA 130 ms), such brisk stimuli can be differentiated due to the human's overtrained ability to listen to speech.*

⋆ *Comparing the results of natural and artificial stimuli, the auditory BCI became "faster" and was considered "more pleasant" when using natural stimuli.*

⋆ *Flaws in the stimulus design might lead to systematic confusions, such that there are pairs of stimuli which are more difficult to discriminate than others. This can be quantified with a statistical test, described in Section 3.2.2.*

3.3 Towards the Simplest Auditory ERP Speller: the CharStreamer

R EALIZING the decoding of brain signals into control commands, brain-computer interfaces (BCI) aim to establish an alternative communication pathway for locked-in patients. In contrast to most visual BCI approaches which use event-related potentials (ERP) of the electroencephalogram, auditory BCI systems are challenged with ERP responses, which are less class-discriminant between attended and unattended stimuli. Furthermore, these auditory approaches have more complex interfaces that impose a substantial workload on their users.

Aiming for a maximally user-friendly spelling interface, this study introduces a novel auditory paradigm: "CharStreamer". The speller can be used with an instruction as simple as "please attend to the character what you want to spell". The stimuli of CharStreamer comprise 30 spoken sounds of letters and actions. As each of them is represented by the sound of itself and not by an artificial substitute, it can be selected in a one-step procedure. The mental mapping effort (sound stimuli to actions) is thus minimized. Usability is further accounted for by an alphabetical stimulus presentation: contrary to random presentation orders, the user can foresee the presentation time of the target letter sound.

Healthy, normal hearing users (n=10) of the CharStreamer paradigm displayed ERP responses that systematically differed between target and non-target sounds. Class-discriminant features, however, varied individually from the typical N1-P2 complex and P3 ERP components found in control conditions with random sequences. To fully exploit the sequential presentation structure of CharStreamer, novel data analysis approaches and classification methods were introduced. The results of online spelling tests showed that a competitive spelling speed can be achieved with CharStreamer. With respect to user rating, it clearly outperforms a control setup with random presentation sequences.

Substantial parts of the data and results were published in Höhne and Tangermann (2014).

3.3.1 Motivation

In addition to prevailing BCI concepts, which make use of visual event-related potentials (ERPs) and self-driven imagery tasks, recent studies proposed tactile and auditory paradigms to broaden the applicability of BCI (for further discussion, see Klobassa et al. (2009), Käthner et al. (2012), Riccio et al. (2012), Schreuder et al. (2012), Kaufmann et al. (2013), De Massari et al. (2013), De Vos et al. (2013), and Gao et al. (2013)). However, such paradigms which are independent of the visual pathway tend to be more complex and less intuitive to use compared to their visual counterparts. This becomes obvious, when comparing existing auditory (or generally non-visual) BCI paradigms to the most frequently used and probably the most successful visual BCI paradigm, the MatrixSpeller (Farwell and Donchin, 1988): to operate it, users do not require instructions beyond the hint to mentally focus on the desired symbol. All available symbols are present on the screen at all times. As the paradigm is following the concept of "what you see is what you get", only low workload is imposed onto the user to select a symbol. If users are capable of directing their gaze to the desired symbol and keep it there, all symbols of a full alphabet are reachable within one logical selection step. Thus, there is a one-to-one mapping from stimuli to the intended action, which is very intuitive.

Existing non-visual spelling paradigms are far from such simple concepts. Their control only has a low degree of freedom and an intrinsically lower communication bandwidth. Thus, the complex options offered in most real-world situations (or a spelling task) cannot be controlled directly. For

this reason, a user interface of BCI communication software typically needs to restrict the number of possible control actions at each step to a small, but feasible set. As a result, the selection of a symbol requires the execution of a series of control steps. Determining a suitable mapping from (few) BCI control options to the (high) complexity of an application is a critical design decision and has been approached in many different ways (Treder et al., 2011; Treder and Blankertz, 2010; Waal et al., 2012).

The mapping introduces an extra level of vicariousness, which bears a number of difficulties in terms of usability. Firstly, sub-steps – e.g. along trees, into the depth of menus etc. – that are necessary to reach a goal within a BCI application conflict with the imperfect control signals, as errors accumulate. Secondly, a spelling tree either needs to be memorized or presented constantly to the user. Thirdly, the user needs to cope with a large cognitive distance between low-level control actions (e.g. selecting the third class) and high-level goals (e.g. spelling "M"). Obviously those three aspects can introduce a non-negligible extra workload for the BCI user. Although these kind of mappings have been optimized in various ways for spelling applications (Schreuder et al., 2010; Höhne et al., 2010; Wills and MacKay, 2006), the resulting interfaces are far more complex than the logical one-step procedure of the visual MatrixSpeller and in the RSVP paradigm (Acqualagna and Blankertz, 2013). In the latter paradigm, however, the user needs to at least memorize the desired symbol during the full duration of the selection step, which may comprise tens of seconds of stimulus presentation. While healthy study participants in good condition may be able to use such "indirect" interfaces despite of the enhanced workload requirements, it remains a problem and a high entrance barrier for many patients (Küber et al., 2013; Schreuder et al., 2013b). But also for healthy persons, it severely limits the usability of the application (Quek et al., 2012).

This observation motivates a novel auditory BCI approach which is introduced in the presented study. The "CharStreamer" paradigm was designed in order to eliminate the above-mentioned mapping problems. Aiming for a simple-to-use auditory paradigm, the CharStreamer strives to realize two main goals:

1. Every symbol can be selected within a single step.

2. Every symbol is represented by the sound of itself, not by an artificial substitute.

Moreover, a third aspect of complexity was challenged. Typically, BCI paradigms which evaluate evoked potentials in the EEG follow the principles of the oddball paradigm with random sequences of target and non-target stimuli. Motivated by the goal to further increase usability and to reduce mental workload, the CharStreamer presents stimuli in a sequential order. Due to this design decision, a user is not required to be alert constantly, as it is exactly known when the desired symbol will be presented. While removing the randomness may lead to atypical and slightly less discriminative EEG features, it also introduces additional temporal structure to the ERP responses (Tangermann et al., 2012a), which can be exploited by an adapted data analysis procedure. Therefore, novel principles for data processing and classification are introduced.

3.3.2 Experiment 3: CharStreamer Online Study

Paradigm Design

The *CharStreamer* paradigm was designed such that it is very easy to understand and usable in an intuitive manner. The whole alphabet, i.e. 26 characters plus 4 command items was split into three groups, $group_{L,M,R}$ with L, M, R representing left, middle, and right respectively. The letters which were contained in one group were read out by the same voice and from the same direction (left, middle and right side). The exact division is shown in Fig. 3.15 A. Stimuli from all three groups were alternately presented, such that every third stimulus belonged to the same group, originating from

the same voice and direction. It should however be noted that the number of characters in the groups differed (9, 10, 11 characters in $group_{L,M,R}$) [5]. Stimuli were presented in group-wise iterations, while each stimulus was presented exactly once in one iteration (see Fig. 3.15 C). The length of one iteration varied between 9 and 11 stimuli for the three groups. Each stimulus lasted 200-250 ms. Although stimuli from all three groups were presented in parallel, two characters never had the exact same onset. Due to the regular temporal distribution of stimuli into the three groups, the perceived stimulus onset asynchrony (SOA_{all}) was three times as fast as the group-wise SOA (SOA_{group}), see Fig. 3.15 C. The paradigm was tested in three experimental conditions (condition A-C, see Fig. 3.15 C) with varying parameterization:

- **Condition A** is a slow oddball condition ($SOA_{group} = 750$ ms, $SOA_{all} = 250$ ms, pseudo-random stimulus order).

- **Condition B** is a fast oddball condition ($SOA_{group} = 250$ ms, $SOA_{all} = 83.3$ ms, pseudo-random stimulus order).

- **Condition C** is a fast sequential condition ($SOA_{group} = 250$ ms, $SOA_{all} = 83.3$ ms, fix stimulus order): the stimulation order was not random but instead following the fix order of the alphabet. Thus, the user always knew exactly when the target letter would appear.

Due to the split of the alphabet into three parts with unequal number of letters, there was no fixed neighborhood of letters across groups (see Fig. 3.15 C). For example, when F is the target letter and O and Y are the following stimuli after the first occurrence of F, then N and W will follow after the second occurrence of F.

Exemplary audio files for each condition are also published, see Höhne and Tangermann (2014) [Supplementary Information A, B, C]. While condition A and B can be regarded as control conditions, condition C is finally named "CharStreamer paradigm", as it is the most advanced and most user-friendly setup.

Auditory Stimuli

The selection and optimization of stimuli for BCI paradigms based on evoked potentials is a crucial aspect. For visual paradigms, the effects of stimulus properties have been described by various authors (Sellers et al., 2006a; Kaufmann et al., 2011; Townsend et al., 2012; Geuze et al., 2012). In the field of auditory BCI, the impact of stimulus properties has been studied by Höhne et al. (2012), Matsumoto et al. (2013), and Lopez-Gordo et al. (2012). Moreover, polyphonic music has recently been explored as a novel stimulation approach for BCIs (Treder et al., 2014). The authors underline the importance to carefully select and optimize stimuli. The optimization criteria are partially contradictory, as stimuli should have natural characteristics while being highly distinguishable, highly standardized and should not be too arousing.

For our study, the spoken alphabet was recorded by three speakers with naturally differing voices (2 male, 1 female), and two of them with an obvious accent. The recording was processed such that an individual auditory stimulus (with a maximum duration of 250 ms) was obtained for each letter. While compressing some sounds in time became necessary, the natural characteristics of the voice, the pitch and the individual intonation was preserved as far as possible. The alphabet was recorded with German intonation and pronunciation. In order to prevent confusions, the vowel color of single letters was slightly altered, if there was another letter with a similar sound in the same group. This applies to the letters (C, D, E) of the first group and (M and N) of the second. Spectrograms of six selected auditory stimuli are shown in Fig. 3.15 B.

---

[5] Please note, that the commands for a whitespace (-, "leer"), a pause ("paus", the final vowel "e" was omitted for brevity), to read aloud ("lies") and for delete ("del") appear in an English translation.
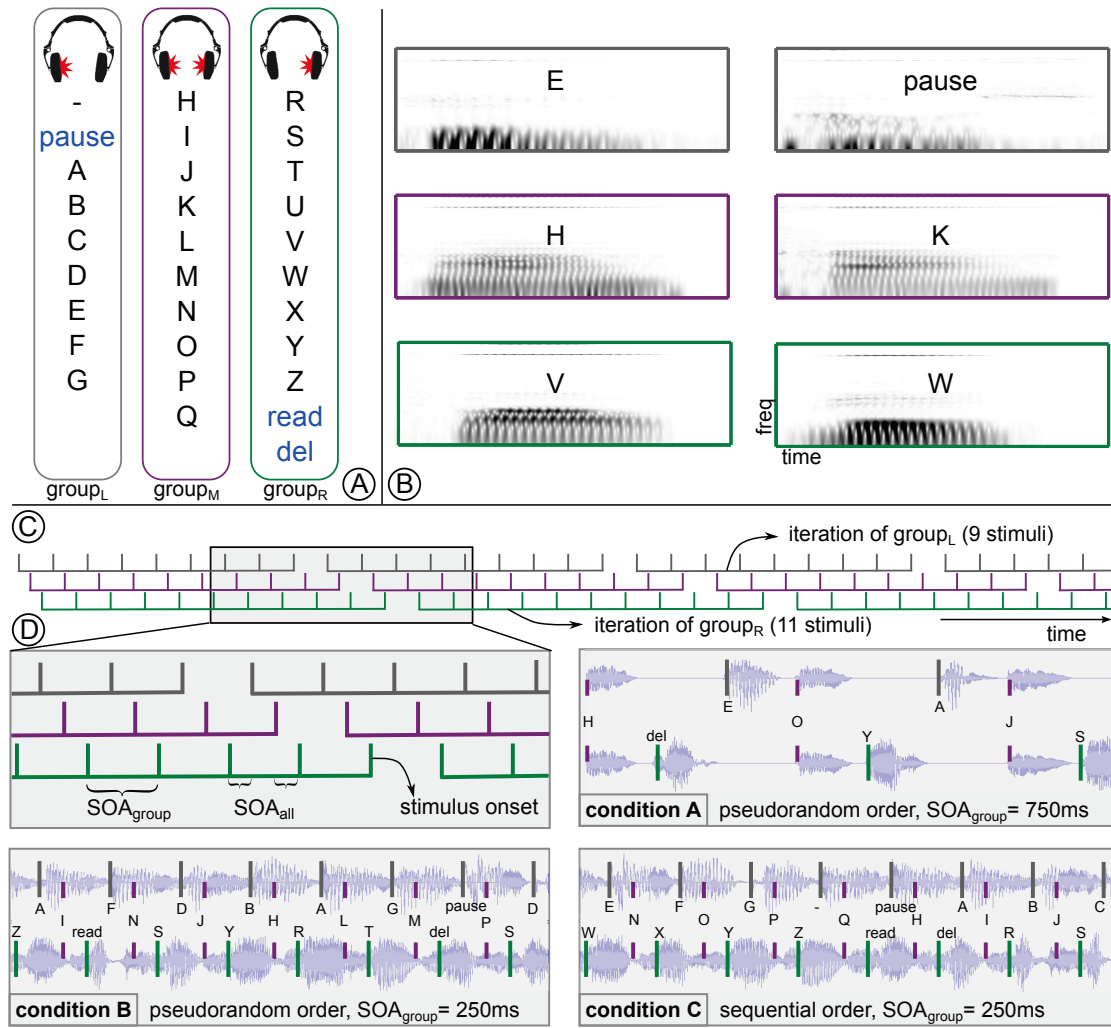
Figure 3.15: *Visualization of the CharStreamer paradigm. The alphabet consisting of 30 characters and symbols was split into 3 consecutive groups (**A**). Each group of letters is presented from a different direction. Spectrograms of six selected auditory stimuli are shown in **B**. The course of a trial is shown in **C**, depicting a sequence of several consecutive iterations. Part **D** visualizes excerpts with a duration of approx. ∼2 seconds. To illustrate the mapping of the three groups to the stereo headphone tracks, the corresponding waveforms for each condition are displayed in the background. Moreover, a magnification of plot **C** is provided in the top-left corner of **D**.*

Study Design

Ten participants were enrolled for the study with a single session of approx. 3-4 hours duration. Each participant had normal hearing and no history of neurological disease. The study was performed in accordance with the declaration of Helsinki. The study was approved by the Ethics Committee of the Charité University Hospital (number EA4/110/09) and all participants gave written consent prior to the start of their session. The study protocol consisted of a calibration phase and an online copy-spelling phase. During recordings, participants were asked to sit still and to avoid eye-movements while focusing a fixation cross. In the calibration phase, the three conditions (A-C) were applied in a

block-randomized order. In each condition, 15 characters were used for calibration and the subjects had the task to mentally focus on the target letter. They were allowed to count the target occurrences, but not explicitly asked to do so as the counting was identified to be a distracting task in a pilot experiment. At the end of each trial, participants reported with a visual analog scale, how easy/hard it was to focus on the target letter. With 14 iterations per trial, ~ 210 target stimuli and ~ 6100 non-target stimuli were collected for each condition and subject.

After the calibration phase, participants were asked for subjective usability ratings on a visual-analog scale for the three conditions. Furthermore they were asked, which of the conditions they would prefer to use on a daily basis, if they had to rely on the BCI system for communication.

In the second part of their session, participants performed an online copy spelling task. It was performed exclusively in stimulus condition C. To decode target vs. non-target epochs, a classifier was trained on the calibration data of condition C, following a "standard" procedure for feature extraction and linear classification (for details, see Sections 3.3.2-3.3.2). Participants were asked to spell the sentence MIT GEDANKEN SCHREIBEN IN BERLIN, (consisting of 32 characters incl. whitespace) without error correction. In the online spelling, a dynamic stopping method was applied (for details see Schreuder et al. (2013a), *Höhne method*) such that within one trial each letter was presented at least five times and maximally 12-15 times[6].

EEG Acquisition and Preprocessing

EEG signals were recorded with a Fast'n'Easy Cap (EasyCap GmbH) using 63 monopolar, wet Ag/AgCl electrodes placed at symmetrical positions based on the extended international 10-20 system. Channels were referenced to the nose. Electrooculogram (EOG) signals were recorded via bipolarly referenced electrodes (vertical EOG: electrode Fp2 vs. an electrode directly below the right eye; horizontal EOG: F9 vs. F10). Two 32-channel amplifiers (Brain Products BrainAmp) processed the signals by an analog bandpass filter between 0.1 Hz and 250 Hz before digitalization (sampling rate 1 kHz). After applying the analog filter, the EEG raw data were first high-pass filtered at 0.2 Hz, then low-pass filtered at 25 Hz, both by a causal Chebyshev filter.

Artifact Correction

EEG signals are generally very prone to muscle and eye artifacts. Correcting for these artifacts was of special interest for this study, as a novel experimental paradigm is researched which might induce unknown or unexpected neural components with atypical temporal and spatial distribution. In this study, two different types of methods for artifact correction were used: a rejection method and a projection method. Bothe methods are introduced in Section 2.5.

To train the classifier which was applied during the original online experiment, an artifact rejection method was applied: EEG epochs violating a min–max threshold difference were rejected. This simple rejection criterion has been described in more detail in a previous study (Höhne et al., 2012).

However, an offline analysis of the EEG data revealed that the above-mentioned rejection method was insufficient for the current study. Although being instructed differently, some users exhibited (unconscious) eye-movements which were partly correlated to the presentation of target stimuli. Thus, either too many target epochs were rejected (using a conservative threshold) or amplitude modulations originating from eye-movements were considered as discriminative features by the classifier when using a more liberal threshold. To circumvent both unfavorable options, an artifact projection method (Winkler et al., 2011) was applied during offline analysis. This elaborate projection method automatically detects neuronal and artifactual source components derived from independent component analysis (ICA). Based on its result, artifactual components were projected out and a cleaned EEG was obtained, which was assumed to be free of eye-movement artifacts.

---

[6]The varying number of maximal repetitions was caused by different group sizes in $group_{l,m,r}$.

Feature Extraction and Classification

This paragraph describes the BCI data processing pipeline that was applied for the online experiment. It should be noted that only *condition C* was applied online. All target and non-target events were analyzed with a "standard" ERP processing pipeline, which is typically applied in the BBCI group for evoked potentials. This pipeline is described in detail in Blankertz et al. (2011): EEG data were band-passed filtered (0.2-25 Hz) and epoched between [-1000 +1000 ms]. Artifacts were removed based on the artifact rejection method described above. Compared to other ERP-paradigms in BCI, the information contained in pre-stimulus EEG intervals could be considered for classification, since the user knew the stimulus order and class-discriminative EEG signals might be elicited before the stimulus onset (Tangermann et al., 2012a). Three to five class-discriminative time intervals were selected by a heuristic. The channel-wise mean amplitudes in those intervals were used as features. A binary linear discriminant analysis (LDA) classifier with shrinkage regularization of the covariance matrix was trained using these features – see Section 2.4.

Optimized Feature Extraction and Classification

The CharStreamer paradigm (condition C) exhibits an intrinsic sequential structure. Fig. 3.16 depicts this temporal structure and the resulting classification problem for sequential data. Therefore, the standard ERP classification procedure described above is likely to be suboptimal – as illustrated in Fig. 3.16D.

Thus, the BCI pipeline was optimized using a meta classifier as depicted in Fig. 3.16E. The meta classifier evaluates a sequence of outputs from several sub-classifiers. This procedure is visualized in Fig. 3.17. Those sub-classifiers were designed in order to uncover two characteristics that were specific for the CharStreamer paradigm:

- Stimuli were presented in a sequential order with every 9th, 10th or 11th stimulus being a target. The user knew when the next target stimulus would be presented.

- Stimuli were presented from thee directions (left, middle or right).

Each sub-classifier was calibrated with the exact same automatized procedure. The main difference between these classifiers arises from the selection of data points which were used to calibrate the respective classifier. This selection resulted in varying weights for feature extraction and classification. Given a set of training data points (EEG recording, epoched from 1000 ms before stimulus onset to 1000 ms after stimulus onset) and labels (class 1 and class 2), a "standard" binary classification approach was taken for each sub-classifier: (I) Class-discriminative time intervals were selected by a heuristic. (II) The averaged EEG data in those intervals were taken as features. (III) Classifier weights for the LDA classifier were trained with covariance shrinkage regularization (Blankertz et al., 2011). The sub-classifiers are described below:

- **global cls:** the standard classification procedure was applied globally. All available target stimuli and all non-target stimuli were used for calibration. This global classifier is typically used for ERP-based BCI paradigms, since it exploits high-level class-relevant information. The ratio between target and non-target stimuli in our paradigm was 1/29.

- **groupwise cls:** the standard classification procedure was applied individually for each of the three groups. This resulted in three classifiers, which were trained and applied for disjoint sets of stimuli. All target- and non-target stimuli from the same group (e.g. $group_L$, as shown in Fig. 3.17A) were used to calibrate a group-wise classifier. The classifier extracted class-relevant information (target vs. non-target) which is specific to the group. The ratio between the number of data points in class 1 and 2 was approximately 1/9.
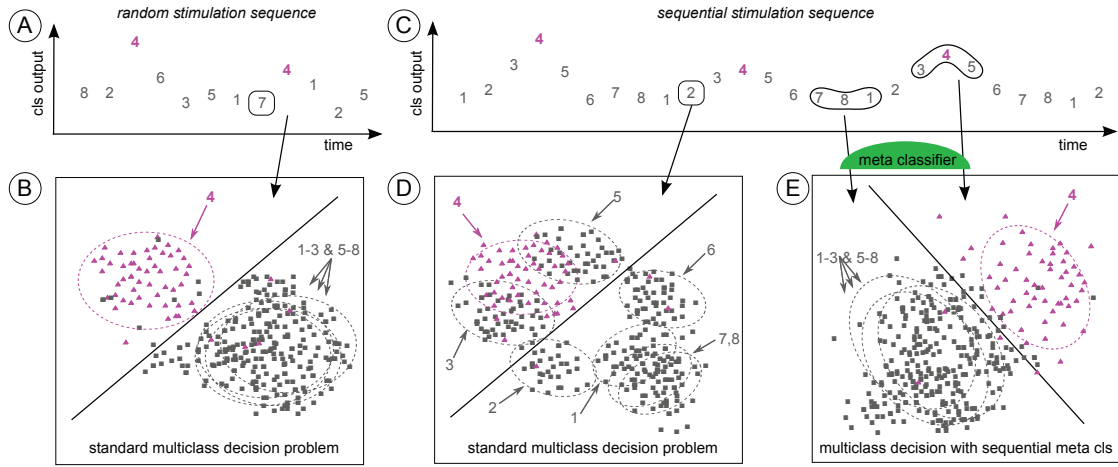
Figure 3.16: *Graphical illustration of the classification problem with sequential stimuli compared to randomly ordered stimuli. The typical oddball scenario with the classification of random stimulation sequences is depicted in plot **A** and **B**. For sequential stimuli, it can be observed that classifier outputs of non-targets before or after a target behave similar to target responses (plot **C**). This leads to systematic structural distortions in the standard multi-class decision (**D**). Plot **E** depicts how a meta classifier can make explicit use of the sequential information and thereby improve the multi-class decision.*

- **pretarget cls:** the standard binary classification procedure was applied to contrast the difference between a target stimulus and its predecessor. While all available target stimuli (class 1) were taken for calibration, only those non-targets that were presented 250 ms before a target (non-targets from the same direction which preceded targets) were considered as class 2. The ratio between class 1 and 2 stimuli was 1/1.

- **posttarget cls:** the standard binary classification procedure was applied to contrast the difference between targets and their directly following non-targets. While all available targets (class 1) were taken for calibration, only those non-targets that were presented 250 ms after the target (i.e. non-targets from the same direction which followed a target) were considered as class 2. The ratio between class 1 and 2 was 1/1.

- **spatial cls:** the standard binary classification procedure was applied to exploit whether the user is attending to the left, middle or right. Thus, a binary classifier was trained for each direction/group. To calibrate each of these classifiers (e.g. for the attended left direction), all stimuli from $group_L$ (targets and non-targets) were distributed into class 1 and 2. Those stimuli that were presented while the user was attending to the intended direction (e.g. left) were considered as class 1. All other stimuli which were presented while the user was attending to a different direction were considered as class 2. The ratio between class 1 and 2 was approximately 1/2 for each direction.

The meta classifier evaluated the outputs of the above mentioned sub-classifiers[7]. However, the meta classifier was trained to also uncover sequential effects (see Fig. 3.16). The meta classifier response of the ith stimulus depended on the sub-classifier outputs of the stimulus sequence $i-m$ to $i+m$. Thus, m preceding and m following stimuli were also considered. An example with $m{=}9$ is shown in

---

[7]To reduce the number of noisy features in the meta classifier, each sub-classifier had to fulfill a minimum binary classification accuracy: only those sub-classifiers featuring a binary classification accuracy of more than 65% (assessed by cross-validation on the training data) were evaluated by the meta classifier.
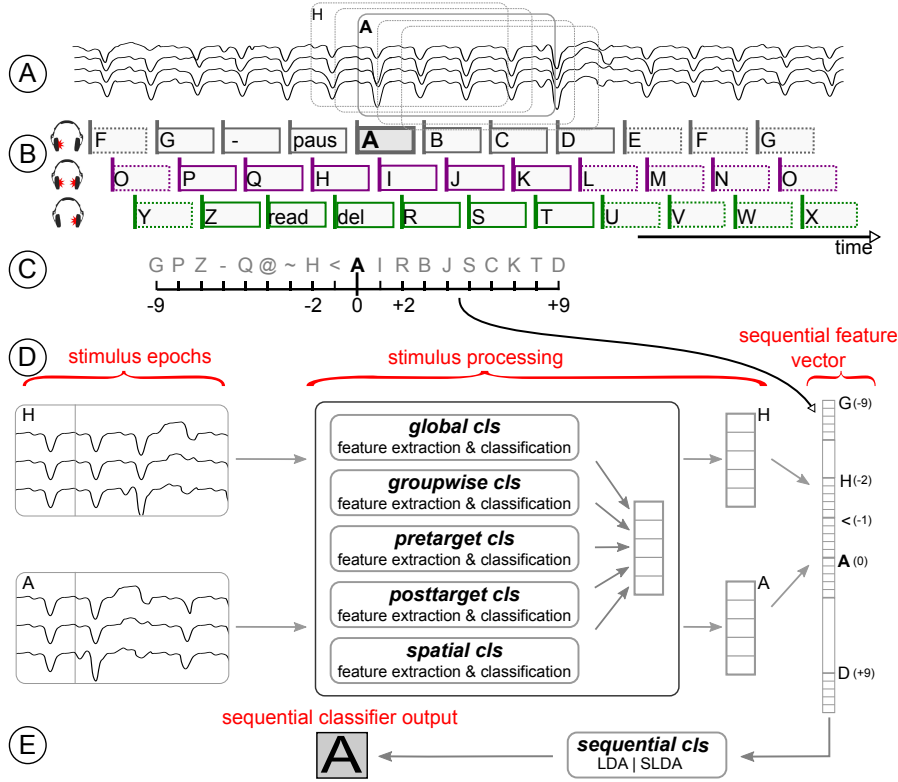
Figure 3.17: *Design of the meta classifier which is optimized for sequential stimuli. Plots **A** and **B** illustrate the EEG epochs and the stimulation sequence in condition C. Plot **C** shows the range of EEG epochs which were considered in order to compute the sequential classifier output with $len = 9$. Plot **D** depicts the processing pipeline of stimulus epochs: each epoch was evaluated by up to five classifiers and the resulting classifier outputs were considered as features of the sequential classifier. The sequential feature vector is evaluated by a meta classier which computes a sequential classifier output for the epoch of interest (**E**).*

Fig. 3.17. This design resulted in a meta classifier (called "sequential classifier" in the following) which considered up to $5 \times ((2 \times m) + 1)$ dimensions. As model selection, the hyperparameter $m \in \{0..9\}$, and the classification algorithm (LDA, sparse LDA (Clemmensen et al., 2011)) were chosen by 5-fold cross-validation.

The calibration data were used to train the sequential classifier. Moreover, the resulting binary classification accuracy was assessed by nested cross-validation. To assess the performance for the online experiment, the EEG data from the copy spelling task was re-analyzed. Therefore, the artifact projection filter as well as the sequential classifier was trained on the calibration data only. Note that during the actual online experiment, a standard ERP classifier (see Section 3.3.2) was applied without the artifact projection method.

Quantification of Multiclass Accuracy based on the Rank

This section deals with the problem of how to quantify the multiclass classification accuracy. When dealing with a small number of classes – e.g. $k = 6$ – then the fraction of correct decisions resembles a quantity, which is highly intuitive and easy to compute. This measure is often applied to describe BCI
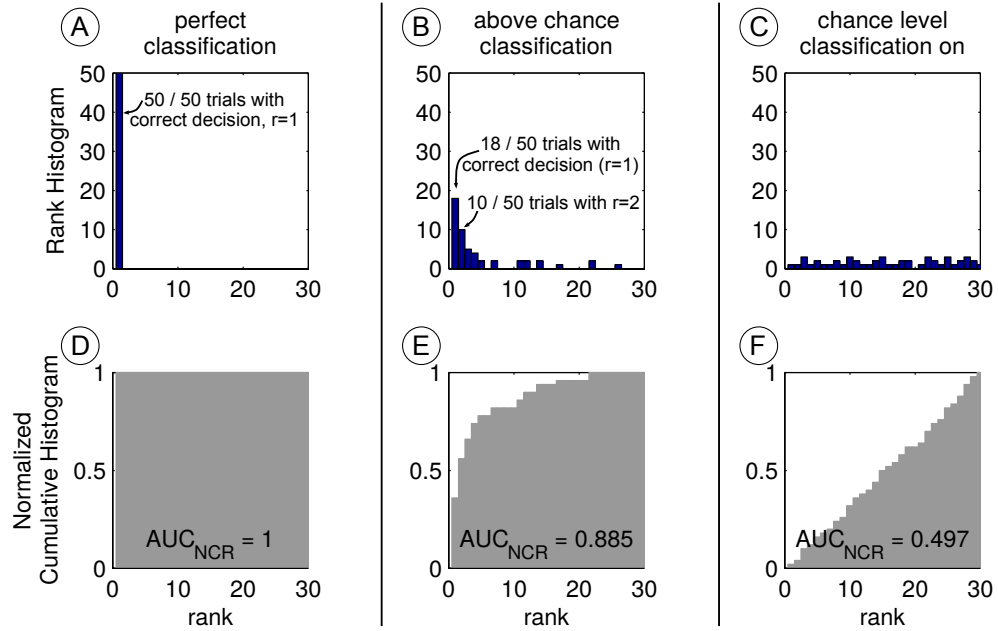
Figure 3.18: *Assessing multiclass accuracy with the AUC_{NCR} for a classification problem with 30 classes. Plots **A-C** depict the multiclass rank histograms for three scenarios: perfect classification, above chance classification and random classification. Fifty trials were simulated for each scenario. Plots **D-F** depict the normalized cumulative rank for each scenario and the resulting AUC_{NCR}.*

performance (Riccio et al., 2012).

However, when dealing with a high number of classes – e.g. $k > 20$ – the fraction of correct decisions might be a troublesome all-or-nothing metric. It does not reward the situation in which the true target class is identified as second-best (or third-best) class. The same holds for the ITR calculation by Wolpaw's formular (see Section 2.6.3 for details) which is also based on the fraction of correct decisions.

Instead, the classification accuracy for multiclass problems with a high number of classes can be assessed with a rank-based method, which is described in the following. In general, the rank (or ranking r) refers to the relative position within ordered scores. Rank-based performance measures are commonly applied in other field with high number of classes, such as website search engines (Langville and Meyer, 2011).

For the multiclass BCI problem based on ERP data, the rank determines the number classes which have received "better" (i.e. more discriminant) classifier outputs. Thus, at the end of each trial, the binary classifier outputs are grouped to the corresponding classes and a score is computed for each class (e.g. character). The class with the highest score has the most evidence to be the target class. This class gets assigned the rank $r = 1$. The class with the second highest score obtains the rank $r = 2$ and so on. Therefore, the fraction of correct decisions can be computed with the fraction of trials in which the target class had rank 1.

Fig. 3.18A-C shows the distribution of r for several scenarios: perfect classification, above chance classification and classification on chance level. In order to quantify this distribution, one can generate normalized cumulative rank histograms (see Fig. 3.18D-F). For all possible rank positions e.g. $r = 1..30$ on the x-axis, these graphs accumulate, how often the target class was contained within the first r ranks. The area under such cumulative histograms (AUC_{NCR}) then gives a suitable overall assessment

of multiclass accuracy. In close analogy to the AUC over the ROC, the AUC_{NCR} can be intuitively interpreted as the probability that the target class is ranked higher than a uniformly drawn non-target class. For a perfect multiclass accuracy with all trials being correctly classified, $AUC_{NCR} = 1$. A random classifier yields a uniform distribution of ranks of the target class which results in $AUC_{NCR} = 0.5$ (see Fig. 3.18). Thus, the higher the AUC_{NCR}, the better the multiclass classification accuracy.

3.3.3 Findings

Usability Ratings

Fig. 3.19A depicts the behavioral ratings for the three experimental conditions, which was assessed after the calibration phase of the experiment[8]. Despite the fast stimulation speed, participants clearly rated condition C to be the preferred condition, being the least tiring condition with a clear target stimulus. This finding was supported by the average trial-wise behavioral rating (Fig. 3.19B) which indicate, how easy it was for the user to focus on the target letter. The usability ratings thus show that condition C was the preferable condition for most subjects.
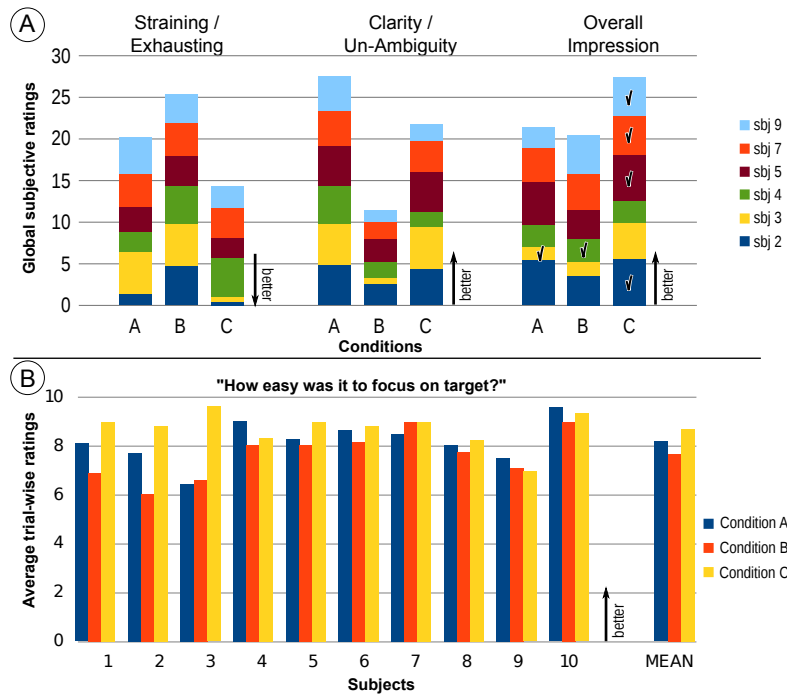
Figure 3.19: *Usability obtained for the three conditions. Plot **A** shows the global subjective ratings for each condition. The overall preference for daily use is indicated for each participant by a tick mark. Arrows indicate, if larger or smaller ratings are better. Plot **B** depicts the average rating, of how well the user could focus on the target letter during each trial in the calibration.*
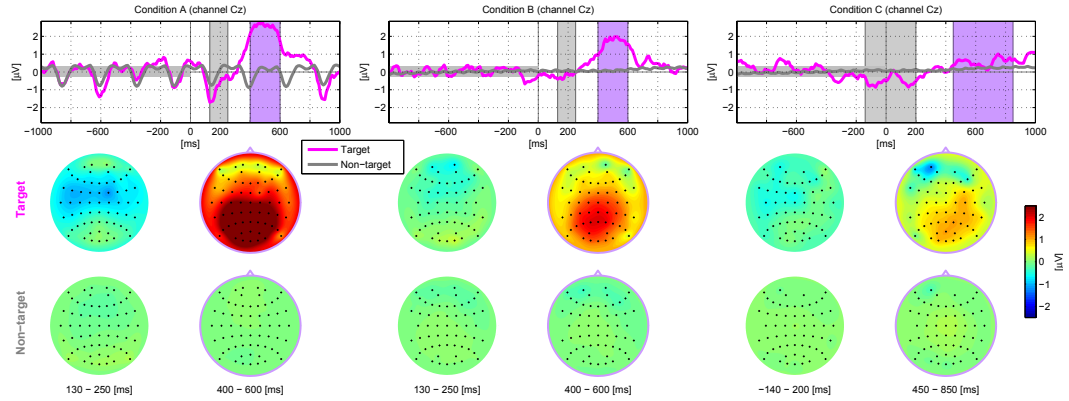
Figure 3.20: *Grand averaged ERPs for conditions A, B and C. It should note be noted that the stimulation speed of condition A is slower than in condition B and C.*

Physiology

Fig. 3.20 shows the ERPs for each condition A, B and C averaged across all subjects. As it was expected for auditory oddball paradigms, typical N200 and P300 responses were found for conditions A and B. Due to the slower stimulation speed (SOA), both components were more discriminative in condition A than in condition B (Höhne and Tangermann, 2012). For the sequential condition C, neither the classical N200 nor the P300 component was present in the grand-average. Instead, a slow, class-discriminative negativity between -200 and +200 ms was observed in the grand average. However, EEG responses of condition C showed a high variation between subjects - with multiple components having their individual temporal and spatial distribution. The ERPs of three exemplary subjects are shown in Fig. A.3.

Offline Analysis of Calibration Data

All following analyses were performed after removing artifacts caused by muscle activity and eye movements. Therefore, the artifact projection method as well as the artifact rejection method was applied as described in Section 3.3.2.

Binary Accuracy   Fig. 3.21A reveals that condition A yields the highest average binary accuracy. The slower timing leads to ERPs with larger amplitudes which can be classified more accurately (Höhne and Tangermann, 2012). On average, the sequential condition C elicits an equal classification accuracy compared to the oddball condition B. However, there is a high variance across subjects: For subject 3, condition B clearly outperforms condition C. Subjects 2 and 6 display the contrary behavior with condition C outperforming condition B. Moreover, the meta classifier leads to an improved classification performance compared to the standard classification approach with subjects 1 and 2 featuring an extraordinary improvement.

Multiclass Accuracy   As discussed in Section 3.3.2, the rank of target class was quantified for this study. Fig. 3.21B visualizes multi-class accuracy as cumulative rank histogram providing additional information compared to the pure accuracy. In analogy to Fig. 3.18, the first entry on the x-axis (multi-class rank = 1) gives the "standard" multi-class performance, as it resembles the fraction of trials with a correct class decision.

---

[8]Only six out of the ten subjects are shown as the remaining four data sets were not saved due to data loss.

Accordingly, the average multi-class performance was 47% for the meta classifier and 41% for the standard classifier (chance level is $1/30 = 3.3\%$). However, the graphs in 3.21B depict the normalized rank distribution which is a powerful tool to assess the multiclass accuracy. It can be seen that on average, 77% (72% for the standard classifier) of the trials obtain a rank better than or equal to 5. The AUC_{NCR} can be computed and it is found that the meta classifier has an improved AUC_{NCR} ranking score compared to the standard approach.
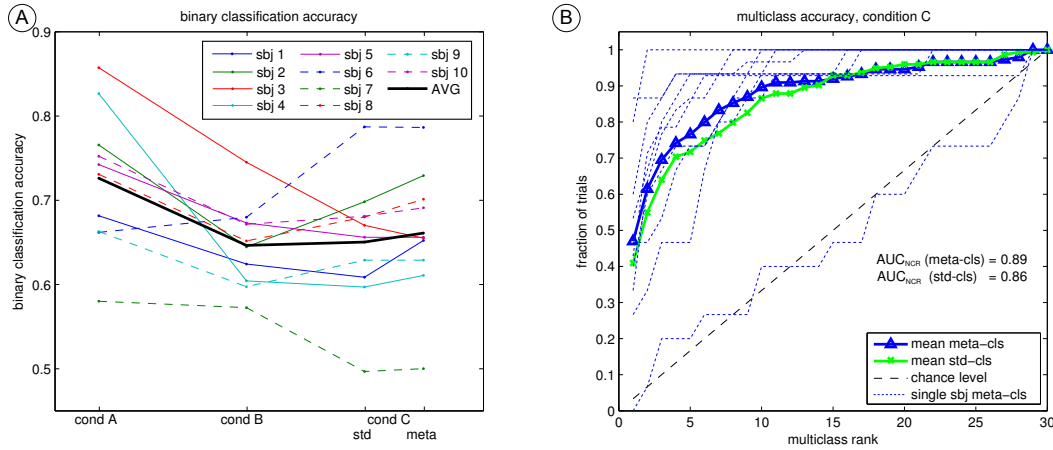
Figure 3.21: *Classification accuracy for the calibration data of three conditions. The binary classification accuracy, estimated with cross-validation is plotted for each condition and subject (**A**). The thick black line marks the mean. Plot **B** depicts cumulative rank histograms which describe the multi-class accuracy for the two classification approaches ("std" and "meta"). This was estimated by cross-validation on calibration data, using entire trials as test sets. Precisely, the point for rank $= i$ quantifies the fraction of trials with a rank of the target class equal or lower than i. Thus, the mean multi-class performance (correct decision – rank=1) was 47% (41%) for the meta (std) classifier. One can observe that 77% (72%) of the trials have a multi-class rank better or equal than 5. While perfect BCI control (each 30-class decision is correct) would result in a straight line with y=1, the dashed line marks the multi-class accuracy based on chance level.*

Time Intervals showing class-discriminative ERP Responses Fig. A.2 depicts discriminative time intervals for each subject and condition. It can be observed that epochs of condition A contain more discriminative features, as the estimated classification accuracy is generally higher than for the other conditions. This stands in line with the results described in Fig. 3.21a. Condition A moreover exhibits discriminative time intervals primarily between 200 and 800 ms after stimulus onset, which corresponds to the N200 and P300 component. Compared to condition A, data from condition B has generally fewer discriminative features that are also shorter - between 250 and 600 ms after stimulus onset. As the stimulation speed is the only difference between the two conditions (condition B exhibits a three times faster stimulation speed than condition A), it can be argued that the SOA has a high impact on the discrimination of evoked potentials (Höhne and Tangermann, 2012). For condition C, discriminative EEG components are observed considerably earlier - even before the stimulus was presented. Moreover, the components are not as temporally concise as one would expect for an oddball experiment (condition B).

Online Spelling Accuracy

An online copy spelling with the sequential condition C was performed with nine out of ten subjects[9]. A thorough reanalysis of the offline and online data revealed that the classifiers which were applied during the online experiment of multiple subjects were severely distorted and partially driven by involuntary eye movements. Therefore, the results obtained during the online experiment are not shown here.

However, both calibration data and online spelling data were reanalyzed (in an offline investigation after the experiment) using an ICA projection method (see Section 3.3.2) to filter out artifacts related to eye movements. Therefore, all projection filters and classification weights were trained solely on calibration data. Online data was evaluated only once, in order to realistically simulate an online experiment in technically plausible conditions. The resulting spelling accuracy of each subject is shown in Fig. 3.22. It was found that seven users were able to use the CharStreamer paradigm with above-chance accuracy. Displaying the strongest class discrimination in the offline data (see Fig. A.2), subject 6 is also the best performing subject in the online spelling with 24/32 (75%) correctly spelled characters. Having an average of 1.5 multi-class selections per minute, subject 6 showed an information transfer rate based on Wolpaw's formula (Wolpaw et al., 2002a) of 4.3 $bits/min$, which is highly competitive for an auditory ERP paradigm, see Fig. 3.22E. One should however note that for the ITR calculation, only the number of correct and incorrect multi-class decisions are considered, disregarding any other information in the rank of incorrect decisions (see also Figure 3.18).

Subject 1 and 7 failed to obtain online control. Exhibiting a very low binary classification accuracy upon calibration data (see Fig. 3.21), a failure of online control was expected for subject 7. For subject 1, a satisfying accuracy was observed on the calibration data, which could however not be transferred to online control. Spelling results shown in Fig. 3.22A-C are based on the sequential classifier. Investigating the top-3 ranked letters by two well performing subjects, Fig. 3.22A reveals that the sequential classifier has still the tendency to assign a high rank to those non-targets which follow or precede the target stimulus. However, from the 32 letters to spell, 11.1 (34.7%) were correctly chosen on average across all subjects, while 4.7 (14.6%) were second-ranked, see Fig. 3.22B. Disregarding subjects 1 and 7 from the average, 13.5 letters (42.4%) were correctly spelled and 5.7 (17.8%) were second-ranked, which points out a considerable spelling accuracy for such a user-friendly BCI paradigm. Fig. 3.22D-E depict how the sequential classifier generally obtains either equal or improved accuracy and compared to the standard classifier on the online data. Fig. 3.22D shows the AUC_{NCR} for each subject for the two approaches, 3.22E shows the fraction of trials with correct class decision. An equal behavior of both approaches could rise from the fact that the sequential classifier might use a parameterization (i.e. m=0, weights only on the global classifier) such that it behaves equally to the standard classifier.

3.3.4 Conclusions

In this study, a novel auditory ERP paradigm (called "CharStreamer") is introduced, which represents a significant step towards more user-friendly brain-computer interfaces. The CharStreamer enables enormous simplifications in terms of the user interface and the workload for the user. It is shown that complexity can be shifted from the user to the system, such that the user is exposed to the simplest and most convenient BCI setup, while the internal data processing pipeline is dealing with atypical and maybe less discriminative EEG signals. The design of the CharStreamer questions two foundations of successful ERP paradigms:

- Is a randomized stimulation order necessary to elicit class-discriminative EEG components?

---

[9]For subject 10, there were technical problems which prevented the copy spelling run, such that online data was not recorded
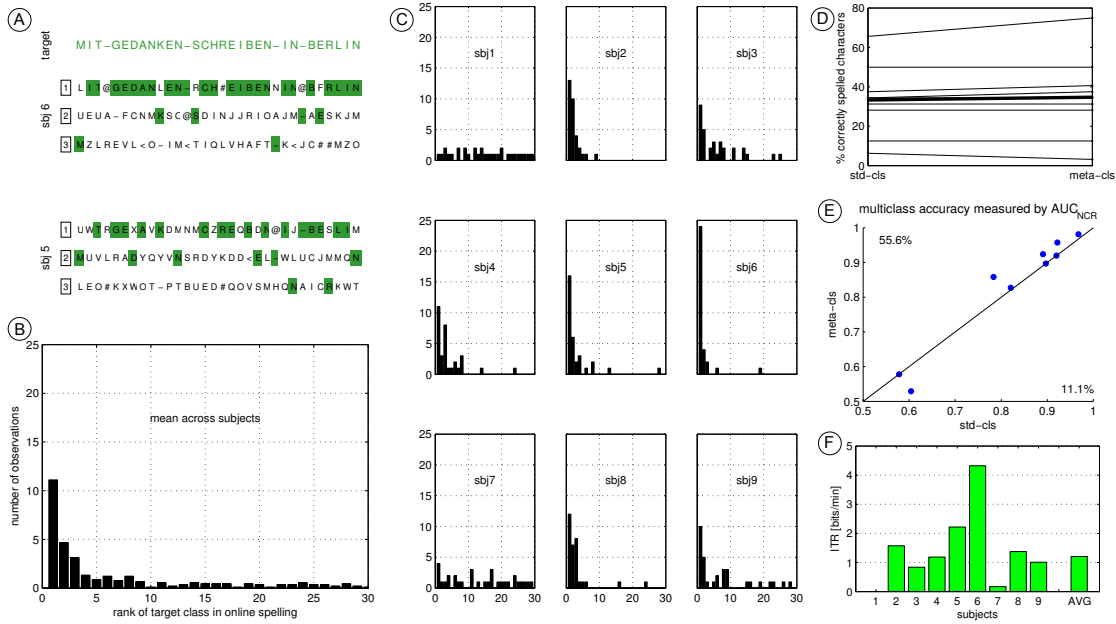
Figure 3.22: *Online spelling accuracy. Plots **A**-**C** describe the spelling accuracy obtained by the sequential classifier. The target sentence and the top-3 ranked characters of two users are shown in **A**. Histogram **B** depicts the rank of the target letter averaged across subjects. The individual rank-histograms are shown in C. Plot **D** depicts the spelling accuracy (rank=1) of the standard classifier and the sequential classifier for each subject and the grand average (thick line). Plot **D** shows a scatter plot of the AUC_{NCR} for the online data of all subjects for meta classifier and the standard classifier. Plot **F** depicts the information transfer rate for each subject.*

- Are the "classical" N200 and P300 components indispensable to drive an ERP-based BCI system?

The CharStreamer paradigm is based on an alphabetical, sequential auditory stimulation such that the user knows when the target letter will be presented. The fast and sequential design of the CharStreamer evoked neuronal components which are significantly distinct from N200 and P300 components of oddball-based auditory ERP paradigms. A central negativity before the onset of the target stimulus was observed for most subjects. It can be speculated that this EEG component may be related to an increased alertness of the subject. Moreover, it may obey a similar neurophysiological origin to the Bereitschafspotential (Kornhuber and Deecke, 1965), which is known to precede a (motor) execution.

Comparing existing auditory BCI paradigms to visual paradigms, another three limits of auditory paradigms are scrutinized:

- The number of classes for auditory BCI paradigms is considerably lower than for visual paradigms. While the visual MatrixSpeller (Farwell and Donchin, 1988) as well as the rapid serial visual presentation (RSVP) speller (Acqualagna and Blankertz, 2013) can deal with 30 classes or more, existing auditory BCI paradigms were so far limited to nine classes (Höhne et al., 2011a). This limitation is mostly due to complexity, since differentiating between short auditory stimuli is more complicated and demanding than differentiating between visual stimuli. The CharStreamer paradigm tries to overcome that limitation by using 30 carefully recorded stimuli. Those stimuli are simple to recognize and easy to distinguish, as they consist of the spoken alphabet, recorded from several voices. As already mentioned, the stimulus differentiation is moreover simplified by presenting stimuli in an alphabetical order.

- Due to the reduced number of available classes, auditory ERP spellers were so far incapable of presenting the entire alphabet to the user. While several visual spellers allow a 1-step approach with the letters themselves being stimuli, auditory BCI spellers either implement a 2-step spelling system (Furdea et al., 2009; Klobassa et al., 2009; Schreuder et al., 2011a) or they combine a 1-step approach with application intelligence (Höhne et al., 2011a). The letter is thus represented in a highly indirect and complicated manner. For example, in the AMUSE paradigm, the letter "L" is spelled by selecting "the second letter of the third group", which is considerably more complicated than focusing on the "L" being highlighted on the screen. As this complex structure might be a major obstacle when applying BCI paradigms with patients in need for a communication solution, the CharStreamer is the first auditory paradigm that enables direct relation between stimulus and letter. Thus, following the principle "what you see/hear is what you get", the user only needs to focus on the presentation of letter "L" in order to spell the letter "L".

- The stimulation speed of ERP paradigms is a crucial aspect which directly effects neurophysiology and communication rate (such as ITR) as discussed in Höhne and Tangermann (2012): although visual paradigms are usually confronted with technical limits such as the frame rate of the screen, Acqualagna and Blankertz (2013) showed that a stimulus onset asynchrony (SOA) of 83,3 ms – corresponding to ∼12 stimuli per second – is possible. However, the fastest auditory paradigm had a SOA of 130 ms (Höhne et al., 2012) – corresponding to ∼7.7 stimuli per second. The CharStreamer design shows that auditory paradigms are not necessarily slower than visual paradigms. By arranging the stimuli in 3 streams presented from different directions, an overall SOA of 83.3 ms –∼12 stimuli per second – was enabled, while the user was still able to identify each stimulus. With such rapid sequences of stimuli, the CharStreamer paradigm is extending the limits of stimulation speed. For future studies, it might however be beneficial to use a slower stimulation as this may further increase usability as well as ERP amplitudes and classification accuracy.

All aspects mentioned above were considered to design the most user-friendly and simple-to-use auditory ERP speller. While most aspects have been individually implemented and discussed in other studies, the CharStreamer paradigm unifies those aspects into one BCI paradigm. Serial presentation of the whole alphabet was first described in the visual RSVP speller (Acqualagna et al., 2010; Acqualagna and Blankertz, 2013). Spatially distinct stimuli for auditory ERP paradigms were proposed with the auditory AMUSE paradigm (Schreuder et al., 2010) and later on implemented in various other approaches (Höhne et al., 2011a; Schreuder et al., 2011a; Käthner et al., 2012). Auditory streaming paradigms, where multiple concurrent streams are presented to the user were suggested by Hill et al. (2004). Moreover, Hill and Schölkopf (2012) showed that one can detect the users' attended stream based on the analysis of evoked potentials of single trials. In order to reduce workload and to increase comfort level and BCI performance of auditory BCI paradigms, it was suggested to utilize natural stimuli instead of highly standardized artificial tones (Höhne et al., 2012; Lopez-Gordo et al., 2012; Xu et al., 2013). The first ERP paradigm with non-random order of stimulation was presented in Tangermann et al. (2012a).

Behavioral data showed that the chosen simplifications tremendously improve the usability of the BCI paradigm. However, such simplifications also raise the need for novel computational methods in order to establish a functioning system. Firstly, it was found that the raw EEG data was contaminated with involuntary eye-movement artifacts, which had to be projected out. The sequential nature of the CharStreamer paradigm triggered involuntary vertical eye-movements: although instructed not move the eyes, multiple subjects raised their gaze just before the target stimulus would appear. When working with completely locked in patients, this problem would not arise due to their inability to perform directed eye movements. However, such artifacts have to be removed for a valid analysis of the neuronal sources which drive the CharStreamer paradigm. Therefore, an ICA-based artifact

projection method was applied in an offline analysis of both calibration and online spelling data. It should be noted that this linear projection was applied as a preprocessing step, prior to feature selection and classification. The parameters of the projection were assessed based on the calibration data only, which is essential in order to obtain a technically plausible online system.

Secondly, it was observed that due to the sequential structure in the data, the classifier had problems to differentiate neighboring stimuli, thus confusing targets with their preceding or following non-targets. Therefore, a meta classifier was developed in order to improve classification accuracy for sequential ERP data. The concept of applying an meta classifier in the BCI framework is far from novel, as meta classifiers were already suggested for motor imagery (Dornhege et al., 2003a; Holz et al., 2013b) or hybrid BCIs (Fazli et al., 2012; Leeb et al., 2010). However, the presented data illustrates that one can apply a meta classifier on ERP data, in order to account for intrinsic sequential effects in the data.

Restoring communication solutions for locked-in patients is the ultimate goal of most BCI research. Due to several reasons, paradigms which are simple to use and easy to understand are favorable when applying BCI with patients. Firstly, complicated interaction systems might be deterring and communication barriers could impede mandatory explanation steps. Secondly, patients might also be frustrated by the complexity of the BCI before even starting to use it.

The Charstreamer paradigm finally demonstrates that it is possible design such a user-friendly auditory BCI spelling system. Elaborate artifact projection methods as well as innovative classification approaches for sequential stimuli enable such a novel paradigm, which features a comfortable and intuitive usage as well as a competitive spelling speed.

3.3.5 Lessons Learned

* *It is possible to set up an auditory BCI with 30 classes, if the stimuli are chosen in an appropriate way.*

* *The AUC_{NCR} is a rank-based measure to assess the multiclass accuracy for classification problems with a high number of classes.*

* *The spoken letters of the entire alphabet can be used as auditory stimuli for a BCI. This enables an intuitive 1-step spelling process with an auditory BCI.*

* *Sequential stimuli elicit class-discriminative ERP components. However, such stimuli introduce a temporal dependency in the data, which gives rise to novel classification approaches.*

* *Artifact projection methods (based on ICA) can be a valuable signal processing tool, if there are muscular artifacts in the data, which partly correlate with task.*

* *Subjects rate an ERP paradigm to be more user-friendly, if stimuli are presented sequentially rather than in a random order.*

3.4 Finding Individually Optimized Stimulation Speed

THIS part addresses the importance of choosing the stimulation speed in ERP based BCIs. In most such paradigms, stimuli are presented with a pre-defined and constant speed. Based on the results of a simple auditory ERP experiment, it is shown that the choice of stimulation speed highly impacts the ergononomics, neurophysiology, as well as the classification accuracy and the resulting BCI performance quantified by the information transfer rate. These findings quantify the improvement in BCI performance when optimizing a very basic experimental parameter. The data and results were previously published in Höhne and Tangermann (2011b) and Höhne and Tangermann (2012).

3.4.1 Motivation

Various paradigms were proposed using the visual (Acqualagna et al., 2010) or auditory (Höhne et al., 2010, 2011a; Schreuder et al., 2011a) modality of stimulation. Most of those ERP paradigms follow the oddball principle of rare target and frequent non-target events. But they differ in the choice and presentation mode of stimuli. Thus, it is reasonable to boost the classification accuracy and BCI performance by optimizing the stimulus characteristics. For the visual and auditory modality, this can be achieved by finding stimulation procedures that elicit the strongest possible class-discriminative components (Tangermann et al., 2011b; Kaufmann et al., 2011; Hill et al., 2009; Höhne et al., 2012).

Another parameter that can be modified is the stimulation speed, which is often described by the stimulus onset asynchrony (SOA) or inter stimulus intervals (ISI). The SOA specifies the time between the onsets of two consecutive stimuli. Most BCI paradigms are applied with a SOA value between 83 ms (Acqualagna et al., 2010) and 500 ms (Farwell and Donchin, 1988). Comparing the visual BCI performance of two SOA levels (175 ms and 350 ms), Sellers et al. (2006a) already stated in 2006 that the choice of SOA highly affects the BCI performance, concluding that "it appears to be worthwhile to test multiple ISI values and thereby determine the optimal value for each user". Nevertheless, the exact choice of stimulation speed has not yet been considered to be crucial, thus it was not optimized by any means.

In the present study, the parameter SOA was investigated with respect to the impact on classification accuracy and BCI performance in a simple auditory oddball paradigm. Classical ERP literature (Gonsalvez and Polich, 2002) describes decreasing amplitudes of class-discriminative ERP components such as P300 for decreasing SOA values and target-to-target intervals (TTI). Consequently, it is expected that the binary classification accuracy (target vs. non-target) correlates with the SOA, such that fast SOA conditions result in a lower accuracies than slow SOA conditions. But although speeding up the stimulation might lead to a reduced separability per stimulus, evidence is acquired with an increased rate. Thus, there may be more stimuli, with each stimulus carrying less discriminative information, which could result in an increased BCI performance. Accordingly, finding the best SOA for a BCI user corresponds to finding the optimal trade-off between the rate of stimulation and the evidence which is provided by each stimulus.

3.4.2 Experiment 4: How Stimulation Speed affects ERPs

Experimental Protocol

Within a single session of about 3 hours, a simple auditory oddball paradigm was tested in 14 SOA conditions. The same type of experiment was performed with varying stimulation speed: a SOA between 50 ms and 1000 ms. The exact SOA conditions are shown at the bottom of Fig. 3.25c. The experiment was divided into four parts, each part consisting of eight blocks with randomized order of conditions. Within each block, there were four consecutive trials of the same condition. In each trial, participants had to concentrate on a rare target tone while neglecting the frequent (83.4 %) non-target tone. Both types of stimuli were sinusoidal with a duration of 40 ms. The target tone had a high pitch (1000 Hz) and the non-target tone had a low pitch (500 Hz). Each trial consisted of 72–90 stimulus presentations (16.6 % targets), and the participant had the task to mentally count the occurrences of the target stimulus. In total, this leads to 1296 events (216 targets and 1080 non-targets) in each condition. Within one trial, the sequence of targets and non-targets was randomized, while it was assured that there were at least three non-targets between two consecutive target stimuli. While attending to the auditory stimuli, the participants were asked to fixate a fixation cross and to not use any muscles. After the first block, the subjects were asked which stimulation speed they preferred.

EEG Acquisition   EEG signals were recorded using a Fast'n Easy Cap (EasyCap GmbH) with 61 wet monopolar Ag/AgCl electrodes placed at symmetrical positions. Channels were referenced to the nose. Additionally, Electrooculogram (EOG) was acquired under the right eye. Signals were amplified using two 32-channel amplifiers (Brain Products), sampled at 1 kHz and band-pass filtered between 0.4 and 40 Hz. The data was epoched between -150 ms and 1000 ms relative to each stimulus onset.

Analysis

All ERP analyses were performed in Matlab and the EEG data was downsampled to 200 Hz. In total, 216 target epochs and 1080 non-target epochs were obtained for each participant and each condition. To remove artifacts, epochs were excluded if their peak-to-peak voltage difference in any EEG or EOG channel exceeded 100μV. For classification, the mean potentials in 12 globally selected intervals at each channel were taken as features, leading to a 732-dimensional (12×61) feature vector for each epoch. The intervals were chosen between 100 ms and 700 ms after stimulus onset with shorter intervals for early responses. A binary RLDA classifier (see Section 2.4 for details) was trained to discriminate between target and non-target epochs for each participant and condition. The classification accuracy was estimated by a cross-validation with 5 folds and 5 shuffles. To account for the imbalance between non-targets and targets, the classwise balanced classification accuracy was calculated, which is the average decision accuracy across classes (target vs. non-target, chance level 50 %).

Simulating the ITR   Based on the empirically obtained binary classification accuracy for each SOA condition, the corresponding BCI performance (in bits/minute) was assessed by simulation. A BCI experiment with a 6-class ERP paradigm was simulated for each subject and SOA condition. Therefore, classifier outputs for target and non-target events were generated according to the binary accuracy, which was determined for the two-class oddball data. Thus, it is assumed that the binary classification accuracy (targets vs. non-targets) of the 6-class paradigm corresponds to the classification accuracy of the 2-class paradigm with equal stimulation speed. Based on the generated classifier outputs, trials were simulated and a multiclass decision was made as soon as an early-stopping criterion was fulfilled, at the latest after 15 presentations of each stimulus (Schreuder et al., 2011b). The duration of a trial and the selection accuracy of the corresponding one-out-of-six decision thus depended on the SOA and
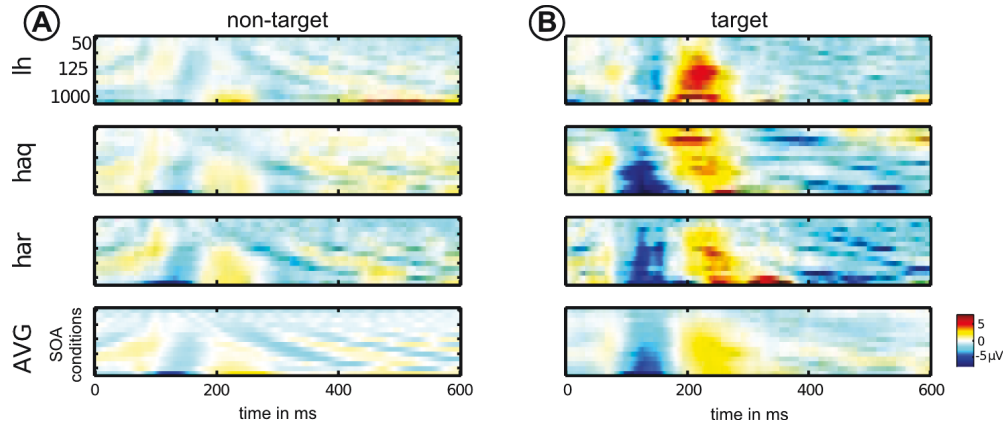
Figure 3.23: *Target and non-target ERPs maps for three subjects and the grand average over all subjects at electrode Fz. Each image depicts the course of an ERP over time and each row corresponds to one SOA condition. All color legends are equal, with red colors coding for positive amplitudes and blue colors coding for ERP negative amplitudes.*

the binary classification accuracy. To account for pauses in between trials, a fixed time of 7 seconds was added after each selection. The ITR (as defined in Section 2.6.3) was then computed based on the number of correct and incorrect decisions after the simulated BCI session, which lasted 60 minutes.

3.4.3 Findings

An analysis of the EEG data revealed that the stimulation time strongly impacts the shape of ERP components for non-target and target epochs. Fig. 3.23 depicts ERP responses to target and non-target stimuli for three subjects and the grand average. The ERP response is color-coded with blue (red) colors coding negative (positive) amplitudes. Each of the 14 rows in the image corresponds to one SOA condition where the top row shows the fastest stimulation (SOA = 50 ms) and the bottom row reflects the slowest stimulation (SOA = 1000 ms). As a general trend, the amplitudes of the ERPs increase with slower stimulation speed, which is in line with classical ERP literature (Gonsalvez and Polich, 2002). This holds particularly for non-target ERPs.

For target and non-target responses, one can observe a negative deflection 150 ms after stimulus onset. This leads to a vertical blue pattern in the images. For the target events, this N150 component is considerably stronger which is often referred to as Mismatch Negativity (MMN) in neurophysiology literature (Näätänen et al., 2007). Target responses show a positive deflection that starts 200 ms after stimulus onset. Amplitude and duration of this P200 component increase with increasing SOA (and decreasing stimulus speed, respectively).

For non-targets, one can additionally find a diagonal pattern between 200 ms and 400 ms after stimulus onset. This pattern reflects the shift in the steady state response, caused by consecutive stimuli. Thus, those responses are directly affected by stimulation speed.

Fig. 3.24 depicts the class discrimination between targets and non-targets over time. Fig. 3.24A shows the course of class discrimination for electrode *Fz*, while Fig. 3.24B displays a measure of class discrimination that incorporates all 61 EEG channels. To quantify class discrimination for one channel over time, the area under the ROC-curve (AUC) was computed and slightly modified (signed and linearly scaled to the range range of [0, 1]). The resulting measure (called ssAUC, see also (Höhne et al., 2011a)) provides information about the strengths and the direction of an effect. In Fig. 3.24A, an early negative class-discriminative component (MMN) and a later positive discriminative component
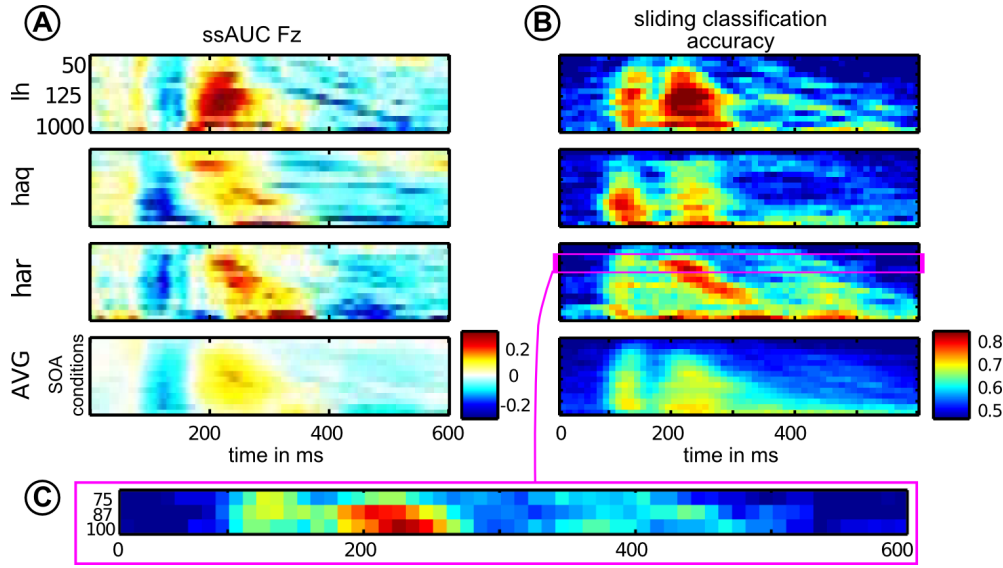
Figure 3.24: *Class discrimination maps over time for each SOA condition: ssAUC values at electrode Fz over time (**A**) and binary classification accuracy based on the mean amplitude of a sliding 50 ms EEG epoch with all electrodes (**B**). A close-up of the binary classification accuracy har for the SOA conditions 75, 87, 100 is shown in plot **C**.*

(P2) can be observed at *Fz*.

To obtain a measure for class discrimination that considers all 61 EEG channels, classification accuracy was estimated with a sliding window as features: mean amplitudes of a 50 ms interval were computed for all electrodes, resulting in a 61-dimensional feature vector for each stimulus. Based on those features, the classification accuracy (targets vs. non-targets) was computed for the given interval. The averaging interval was sliding between 0 ms and 600 ms after stimulus onset. Fig. 3.24B depicts the classification accuracy, with red (blue) coding for high (low) classification accuracy.

Fig. 3.24A-B reveals that the latency of the class discriminative N150 component is the same for all conditions. Thus, stimulation speed does not affect the latency of the N150. In contrast, the latency of the class discriminative P200 component is affected by the stimulation speed, in particular for subject *har* and *haq*. Moreover, one can observe the general trend of increasing amplitudes and class discrimination with increasing SOA for both the N150 and the P200 component, which is known from classic ERP literature (Gonsalvez and Polich, 2002).

This correlation of class discrimination and SOA is also reflected in Fig. 3.25A, where classification accuracy is plotted for each subject and each condition. On average, the binary classification accuracy is highest for a SOA of 1000 ms (SOA_{1000}). Although this observation is in line with classic ERP literature, classification accuracy is not decreasing monotonously with faster stimulation. For example, Fig. 3.25A shows clear peaks for subject *har* at SOA_{87} and SOA_{175}, which means that those stimulation conditions induce evoked potentials that can be classified more accurately than other (even slower) stimulation speeds. For *har*, the classification accuracy at SOA_{87} (0.84) is considerably higher than the accuracy for SOA_{75} (0.73) and also higher than SOA_{100} (0.78). The reason for that increase is explained in Fig. 3.25C, showing that for SOA_{75}, there is only early discriminative information centered at 120 ms after stimulus onset. For SOA_{87}, a strong P200 component is observed additionally, which explains the increase in classification accuracy from 73 % to 84 %. Reducing the stimulation speed from 87 ms to 100 ms (SOA_{100}), the P200 latency increases, but more importantly, the early component at 120 ms diminishes, which results in a reduction of overall class discrimination and
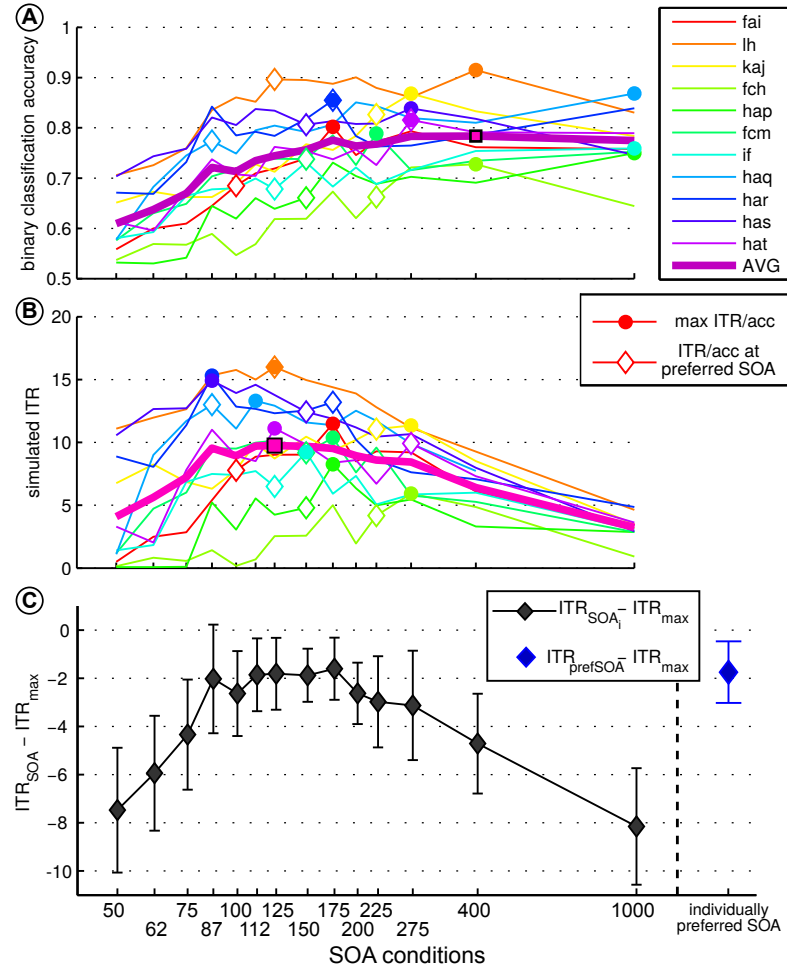
Figure 3.25: *Classwise balanced binary classification accuracy which was observed for each subject and SOA condition (**A**). Plot **B** shows the simulated ITR for each subject and SOA condition. Individual maximum values are marked with colored circle, individually preferred conditions are marked with a diamond. The average absolute difference between the ITR of the individual optimal SOA and all other SOA conditions are depicted in plot **C**. Plot **C** also depicts the average absolute difference between the ITR of the individual optimal SOA and the individually preferred SOA condition. The whiskers show the standard deviation across subjects.*

classification accuracy (84 % to 78 %). This is only one example for individual variability in ERP components and classification accuracies for slightly different stimulation speeds.

Fig. 3.25B shows the ITR that was simulated for each subject and condition as described above. One can observe that the optimal stimulation speed (with respect to ITR) is between 87 ms and 200 ms for most subjects. The maximum ITR value for each subject is marked with a circle. Due to considerable variability in the binary classification accuracy, the ITR is also varying for single subjects, leading to peaks in the curve, such as SOA_{87} for *har*. Fig. 3.25C quantifies how much BCI performance is lost by a globally defined stimulation speed that is used for all subjects: the individual maximum ITR (ITR_{max}) is subtracted from the individual ITR (ITR_{SOA_i}) for each SOA condition *i*. Thus, the curve in Fig. 3.25C can only reach the value 0 if all subjects have their maximum ITR at the same stimulation

speed. The graph shows that if the stimulation speed is globally chosen between 87 ms and 200 ms, the average BCI performance is ∼2 bits/min lower than the individually optimized ITR. Across all conditions, SOA_{175} performs best with a bitrate of 1.6 bits/min. Thus, if the individually optimal SOA was used as stimulation speed, the average increase in ITR would be at least 1.6 bits/min, even if the globally optimum was known.

Moreover, it was found that using the individually preferred stimulation speed leads to a very good performance as well (loss of $SOA_{pref SOA} = 1.74$).

3.4.4 Conclusions

In typical BCI paradigms based on ERPs, the stimulation speed (here SOA) is pre-defined and thus equal for each subject. Changing the stimulation speed, one observes varying ERPs as shown in Fig. 3.23. In the study presented here, it is demonstrated, that even in one of the simplest types of ERP paradigms (2-class auditory oddball), a slight change in stimulation speeds may result in non-linear variations of class-discriminative ERP components and the resulting classification accuracy. Discriminative ERP component are suppressed or enhanced for specific stimulation speeds, as it is shown for one subject in Fig. 3.24.

Consequently, this study points out that an individual choice of the stimulus onset asynchrony is highly beneficial with respect to BCI performance. The analyses of a simulated online BCI experiment with 14 SOA conditions reveal that BCI performance (assessed by ITR) is increased by ∼2 bits/min, if the SOA is defined for each subject individually.

The work by Sellers et al. (2006a) already showed that the choice of SOA highly impacts the BCI performance. The presented study underlines these findings and quantifies the systematic loss of performance due to the global selection of the SOA. Moreover, it is shown that the individually preferred stimulation speed also leads to a very good BCI performance, being almost as good as the (mostly unknown) global optimum.

3.4.5 Lessons Learned

> ⋆ *The timing of the stimulation directly impacts the ERP.*
>
> ⋆ *Mainly the duration of late ERP components such as P300 is affected when changing the stimulation speed.*
>
> ⋆ *The average BCI performance can be increased by ∼2 bits/min or ∼ 10%, if the optimal stimulation speed is applied for each user individually.*
>
> ⋆ *For the simple binary oddball experiment, the optimal information transfer rate can be achieved with a stimulation speed of 175 ms*

3.5 Critical Assessment or the Contributions for Auditory BCI

This chapter describes several approaches that guide towards more user-friendly auditory BCI paradigms. Several novel paradigms were proposed and extensively tested with healthy subjects. However, the final evaluation of a novel auditory BCI paradigm can only be done once it is applied with end-users – i.e. individuals who can improve their mean of communication with such BCI paradigm. These patient/end-user studies are subject to further research and they require a significant amount of additional work and collaborations with multiple clinical institutions. Therefore, the presented paradigms *PASS2D* and *CharStreamer* have been implemented and published within an open-source software framework PyFF (Venthur et al., 2010). Other BCI researchers can modify or extend these paradigms and apply them with healthy subjects as well as with end-users.

Moreover, the concept of sequential stimulation which was fist described in the *CharStreamer* paradigm should be investigated in follow-up studies. Sequential stimuli drastically simplify the complexity of (auditory) BCI paradigms, but the impact on the ERP signals needs to be evaluated in a more systematic way, with a large number of subjects.

Auditory BCI is a rather young line of research which has been arousing significant interest by the BCI community. However, auditory BCI paradigms cannot be seen as a general solution for all enduser scenarios. Therefore, it is important to extend the concepts and paradigms which were presented in this chapter and to integrate them into other domains, leading to hybrid BCI approaches (Pfurtscheller et al., 2010). A first approach has been described in An et al. (2014), where the concept of *PASS2D* has been combined with a visual BCI speller, yielding an audio-visual speller.

ANALYZING NEUROIMAGING DATA WITH SUBCLASSES: A SHRINKAGE APPROACH

NEUROIMAGING data is subject to numerous data analysis methods. Amongst them, Linear Discriminant Analysis (LDA) is commonly applied for binary classification problems. The popularity of LDA arises from its simplicity and competitive classification performance which was described for various types of neuroimaging data.

However, Chapter 3 describes several studies indicating the standard LDA approach to be suboptimal for binary classification problems in the presence of additional label information (i.e. subclass labels). This chapter discusses that problem and illustrates how neuroimaging data feature subclass labels that are disregarded by an LDA classifier.

We introduce a novel method that allows to incorporate such subclass labels in an efficient manner. The novel method, called Relevance Subclass (RSLDA) LDA is based on regularized estimators of the subclass mean, while using other subclasses as regularization target. The applicability and performance of our method is demonstrated on data arising from two different neuroimaging modalities: (I) EEG data from brain-computer interfacing with event related potentials and (II) fMRI data in response to different levels of visual motion. We show that RSLDA outperforms the standard LDA approach for both types of datasets. These findings illustrate that it is beneficial to exploit such subclass structure in neuroimaging data. Finally, we show that our classifier also outputs regularization profiles, which can be interpreted in a meaningful way.

Thus, RSLDA yields increased classification accuracy as well as a better interpretation for neuroimaging data. Both aspects are highly favorable, suggesting to apply RSLDA for classification problems within neuroimaging and beyond.

Parts of the data and results were published in Höhne et al. (2014b). A second journal article is in preparation (Höhne et al., 2014a)

4.1 Motivation

Multivariate analysis techniques are commonly applied in order to investigate neuroimaging data. The main objective behind such analysis is to study the temporal and spatial properties of neural processes that are initiated within the experimental paradigm. In a typical analysis scenario, a binary classifier is trained on the neural responses to two types of stimuli, which can be measured with neuroimaging techniques such as EEG or fMRI. Various machine learning methods have been proposed for this classification task (Garrett et al., 2003; Pereira et al., 2009; Lemm et al., 2011). They differ in complexity (linear / non-linear) as well as in additional assumptions on the distribution of the data.

However, neuroimaging studies can have a rather complex experimental paradigm, which might not qualify for simple binary classification methods. Such complexity can arise from several subcondi-
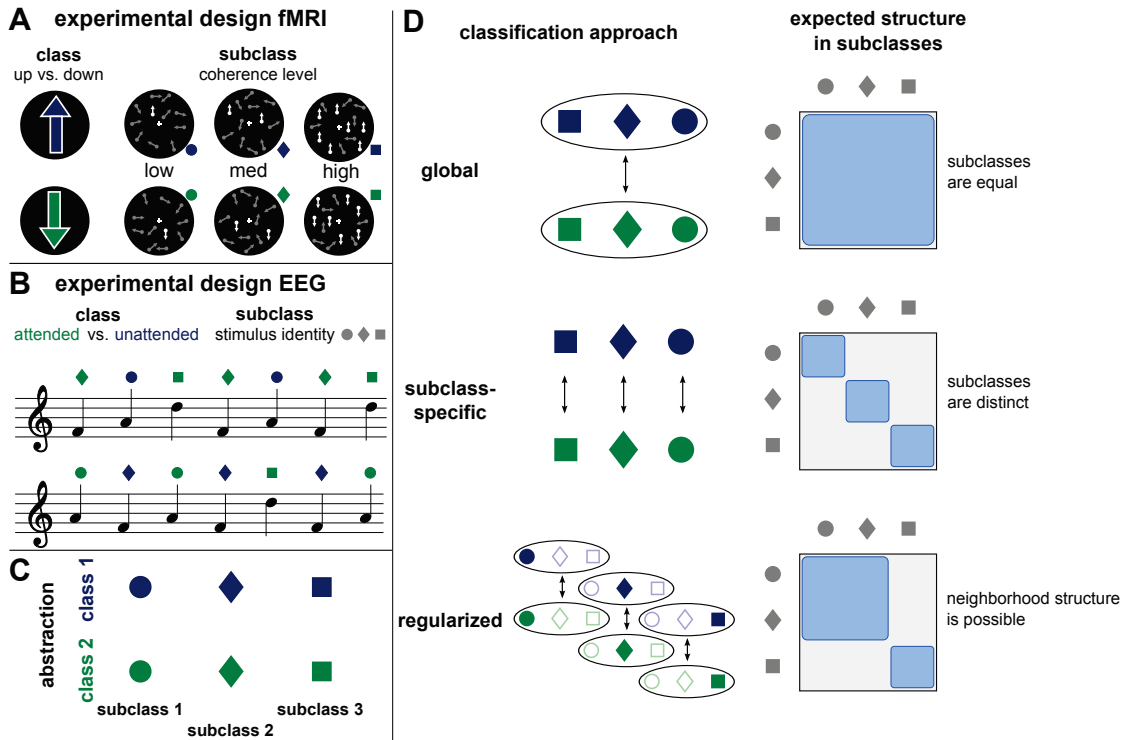
Figure 4.1: *Illustration of subclass structure in neuroimaging studies. Plot A shows the experimental paradigm of an fMRI study investigating upwards and downwards motion with several coherence levels. The coherence level can be considered as subclass/subcondition. Plot B depicts the design of an EEG study where subjects had the task to attend to specific stimuli. The data of both studies can be analyzed as a binary problem with subclass structure, as shown in plot C. Plot D visualizes three classification approaches for such data. The right column shows the underlying assumptions on the subclasses for each approach.*

tions/subclasses, as stimuli from the same type (i.e. same condition/class) might be presented with multiple peculiarities. Examples for an fMRI and an EEG study are therefore depicted in Fig. 4.1A-B and briefly described below.

Fig. 4.1A shows the experimental paradigm of a visual motion fMRI study with two conditions/ classes[1]. Neural correlates of upwards and downwards motion are investigated, while the visual stimuli either had low, medium or high motion coherence. The coherence level can therefore be regarded as subclass.

Fig. 4.1B shows the experimental paradigm of an auditory EEG study with two classes: attended and unattended stimuli. While random sequences of three types of stimuli are presented, subjects have the task to attend to only one of them and ignore the other two stimuli – as described in Chapter 3. When training a classifier on the single-trial event-related potentials (ERPs) for attended vs. unattended stimuli, the stimulus identity can be regarded as subclass information.

Both above mentioned studies seek for neural correlates of a binary classification problem (i.e. upwards vs. downwards visual motion and attended vs. unattended auditory stimuli). However, subclass information is available (i.e. coherence level and stimulus identity respectively) and considering such

---

[1]The terms "class" and "condition" are considered to be equivalent. The same holds for "subclass" and "subcondition". For the remainder of this chapter, the terms "class" and "subclass" will be used.

information might be favorable for the classification task. Therefore, Fig. 4.1C depicts three classification approaches that can be applied for this data.

The global approach disregards any subclass information and thereby assumes the subclasses of each class to be equal. Data are pooled across all subclasses and only one classifier is computed for the entire data.

The subclass-specific classification approach is based on one classifier for each subclass and thereby assumes each subclass to be distinct. This approach is confronted with a reduced amount of data which is available to train each classifier.

The regularized approach presents a trade-off between the global and the subclass-specific approach. A classifier is computed for each subclass separately, while the remaining subclasses are used for regularization. Thus, the regularized approach is able to exploit some dependency or neighborhood structure which might be present in the data. However, this approach is based on additional regularization parameters, which have to be estimated.

The aim of this work is to discuss the binary classification problem with subclass information in the context of neuroimaging data. We compare the three above mentioned approaches based on a reanalysis of existing EEG and fMRI data. Moreover, a novel regularization approach – called Relevance Subclass LDA – is derived, which is able to exploits subclass information in a highly efficient way. We show that the proposed method outperforms the global and subclass-specific approach. Moreover, we show that Relevance Subclass LDA also delivers a distribution of regularization parameters. Such parameters can serve as a valuable tool to interpret the underlying subclass structure in the data.

The remainder of this chapter is organized as follows. Section 4.2 introduces the methodological details of state-of-the-art classification methods and their suboptimality in the presence of subclass structure. Then, the concept of shrinkage is reviewed, which is an algorithm that can be applied to find regularized estimators of the covariance and the mean. The novel classification method "Relevance Subclass LDA" (RSLDA) is introduced which is based on shrinkage. Two evaluation data sets are described. Results are presented in Section 4.3 and we conclude with a discussion in Section 4.4.

4.2 Methods

4.2.1 Linear Classification for Neuroimaging data

Linear methods such as linear support vector machines (SVMs) (Vapnik, 1995; Müller et al., 2001) or linear discriminant analysis (LDA) are commonly applied to analyze neuroimaging data. There are three main reasons, why linear methods are often preferred to more elaborate nonlinear methods (Müller et al., 2003; Misaki et al., 2010).

- (**Performance**) After applying suitable steps for feature extraction and processing, the classification performance of linear methods is on the same level as non-linear methods – or even better (Misaki et al., 2010; LaConte et al., 2005; Krusienski et al., 2006).

- (**Overfitting**) Linear methods are commonly based on less parameters, which is favorable when analyzing a highly limited amount of data points featuring a high dimensionality (Duda et al., 2001).

- (**Computation**) The computational effort of linear methods is significantly lower, which resembles an important factor when investigating large-scale data set in which thousands of classification problems have to be solved (e.g. fMRI searchlight analysis (Kriegeskorte et al., 2006)).

While linear SVMs are often applied in neuroimaging, recent comparison studies found LDA with covariance shrinkage to perform equally on fMRI data (Misaki et al., 2010). For EEG data, LDA

with covariance shrinkage was described to even outperform linear SVMs (Krusienski et al., 2006). Moreover, training LDA classifiers takes less computation than training SVM classifiers, as LDA does not require any additional parameter selection and a second-level cross-validation. Therefore, we will focus on LDA for the remainder of the manuscript, using LDA with covariance shrinkage as a baseline method, being reported as the best-performing methods by several studies (Krusienski et al., 2006; Misaki et al., 2010; Blankertz et al., 2011).

4.2.2 Analyzing Binary Classification with Subclass Structure

As it was depicted in Fig. 4.1, the experimental design of a neuroimaging study might give rise to subclass structure in the data. Subclasses are the intersections of the classes of the actual classification problem (green vs. blue in Fig. 4.1) with another grouping of the data (square, diamond, and circle in Fig. 4.1) that is independent of the class labels. Here, we consider the case, that the grouping is known for all data points, i.e. including test data. Thus we investigate classification approaches that incorporate the implied subclass information.

Fig. 4.2 depicts two-dimensional toy data which illustrate a binary classification problem with three subclasses. The global average (i.e. mean across all subclasses) for each class is marked with a star. In the following, several classification approaches are introduced in a conceptual manner. The detailed, mathematically sound description is presented in Section 4.2.3 - 4.2.5.

Fig. 4.2B and C depict the global and subclass-specific classification scenario with LDA. Fig. 4.2D and E show approaches to regularized classification scenarios with LDA.

Fig. 4.2D illustrates a regularized classification framework for one subclass (diamonds), in which the global mean is regarded as regularization target. The red areas mark the range of separation hyperplanes, as the final orientation of the hyperplane directly depends on the two regularization parameters – one parameter for each class.

Fig. 4.2E illustrates the regularized framework, in which the means of all remaining subclasses are regarded as individual regularization targets. The blue/green shaded areas depict the range of regularized estimates of the means. Due to the increased number of parameters, the estimation of the means and the resulting separation hyperplanes (i.e. red area) feature an increased degree of freedom.

4.2.3 The Global Approach: LDA with Covariance Shrinkage

LDA is a multivariate linear classification method that is frequently applied to analyze neuroimaging data (Blankertz et al., 2011; Pereira et al., 2009; Misaki et al., 2010). LDA assumes the data to follow a normal distribution with all classes having the same covariance structure (i.e. homoscedasticity). For a methodological introduction of LDA with covariance shrinkage, see Section 2.4.1–2.4.2.

4.2.4 Subclass-specific Approach: LDA Classifier for each Subclass

The subclass-specific approach computes a binary classifier for each pair of corresponding subclasses (e.g. blue diamonds vs. green diamonds in Fig. 4.2). The class-wise means are computed for each subclass individually. This leads to a highly reduced amount of data points which is available, compared to the global class mean. When assuming the covariance of the data to reflect the background noise
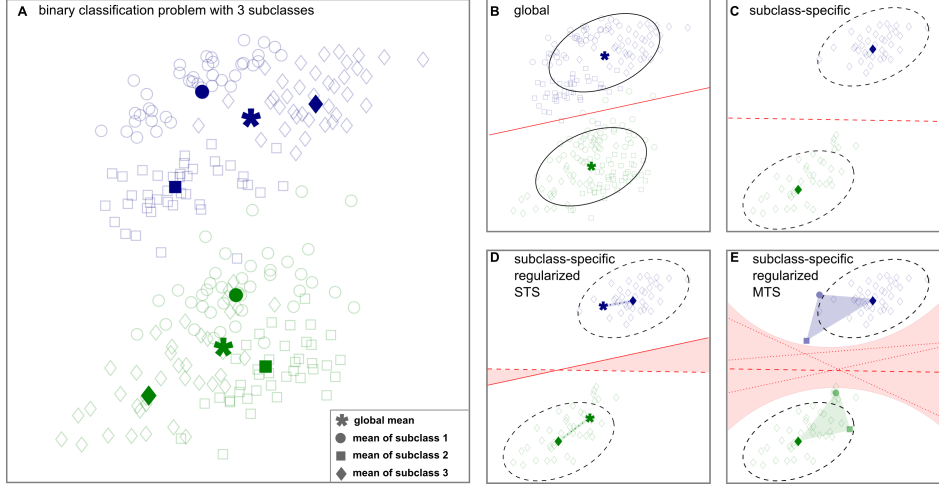
Figure 4.2: *Example for a binary classification task with subclasses. Plot **A** shows the distribution of data points with the color/symbol specifying the class/subclass respectively. The means are shown in bold. Plot **B** depicts the LDA separation hyperplane of the global LDA approach (solid line). Plot **C** depicts one subclass-specific LDA classifier (diamonds) with a dashed line. Plots D and E describe the regularized classification scenarios. Plot **D** shows the approach which uses the global mean as regularization target, referred to as Single-Target Shrinkage (STS). Plot **E** depicts the Multi-Target Shrinkage (MTS) approach, using all remaining subclasses as separate regularization target. The shaded green and blue areas denote the range of mean estimators when regularizing between the sample subclass mean and the regularization targets. The red areas denote the range of classification hyperplanes that can be obtained by regularization.*

which is independent of the subclass, it is valid to estimate the covariance C pooled over all subclasses. Besides, due to the increased amount of data, the computation of C on pooled data yields a decreased amount of systematical distortion (Blankertz et al., 2011).

4.2.5 Regularized Approach: Subclass-specific Classifiers that may incorporate Data from other Subclasses

In order to obtain a robust estimator for the subclass mean, one can regularize the sample estimator towards the mean of other subclasses (see Fig. 4.2D-E). Thus, one can define the regularized estimator for the mean of class i and subclass g by

$$\boldsymbol{\mu}_{i,g}^{MTS}(\boldsymbol{\lambda}) := (1 - \sum_{l \neq g} \lambda_l)\boldsymbol{\mu}_{i,g} + \sum_{l \neq g} \lambda_l \boldsymbol{\mu}_{i,l}. \tag{4.1}$$

The range of possible outcomes of this regularization with multiple targets ($\boldsymbol{\mu}_{i,g}^{MTS}$) is depicted in Fig. 4.2E.

When assuming each subclass l to have the exact same regularization parameter, one can rewrite Eq. (4.1) to

$$\boldsymbol{\mu}_{i,g}^{STS}(\lambda) := (1 - \lambda)\boldsymbol{\mu}_{i,g} + \lambda \boldsymbol{\mu}_{i,\bar{g}} \tag{4.2}$$

with $\boldsymbol{\mu}_{i,\bar{g}}$ denoting the sample mean of class i, excluding the data points from subclass g. The classification scenario using $\boldsymbol{\mu}_{i,g}^{STS}$ is depicted in Fig. 4.2D.

Mean shrinkage

The regularized approaches, introduced before, depend on a set of regularization parameters that is required to compute the estimator of the mean, see Equation (4.1)-(4.2). In this paragraph we describe how to determine such parameters in an accurate and computationally efficient way.

Historic Remark and Intuitive Introduction   It should be noted that Mean Shrinkage (also referred to as James-Stein shrinkage), has been controversially discussed in the past. Especially in the 1960's and 1970's, there was a public debate that resulted in the term "Stein's Paradox" (Efron and Morris, 1977). In the following a short intuitive introduction is given to this paradox.

"Taking an average is an easy and fairly familiar process that seems to need no justification", stated Efron and Morris (1977). However, they also stated that "Stein's paradox defines circumstances in which there are estimators better than the arithmetic average". The unintuitive nature of the Stein's paradox can be demonstrated with two simple real-world examples.

Suppose our task is, to estimate the average number of tacklings per game of football players in the Bundesliga[2]. We randomly select 100 players and we want to estimate the average number of tacklings of each of these players for the entire season, based on the data of the first four games. Stein's paradox then says that we can obtain a better estimate for the vector of those one hundred values, if we simultaneously use the data of all 100 players, instead of estimating the average number of tacklings for each player separately. The first step of the James-Stein shrinkage is to define the average of the average (the "grand average"), thus the average number of tacklings for all 100 players in the first four games. The second and most important step is called "shrinking" towards the grand-average: the James-Stein estimator reduces the number of tacklings for players that tackled more than the grand average. Those players that tackled less than the grand average obtain an increased estimator compared to their individual sample means. The shrinkage factor can be determined analytically, as it is described in the following Section.

On average across all players, the James-Stein estimator is better (i.e. has a reduced mean squared error) than the sample mean (i.e. arithmetic average) of each individual player.

The second example of James-Stein shrinkage seems even more counterintuitive: suppose our task is to estimate three unrelated quantities, such as the average weight of a newborn (weight), the average number of bricks in university buildings (bricks) and the average number passengers on a flight from Berlin Tegel to Cologne/Bonn Airport (passengers). Assume we have independent measurements of each of these quantities. Stein's paradox then says that we can obtain a better estimate for the vector of the three quantities, if we simultaneously use the entire data, instead of estimating the three averages separately.

This appears very unrealistic on the first glance, as the three quantities are completely unrelated. However, one should note that for this example we do not obtain a better estimator of each individual quantity itself (e.g. the number of passengers). Instead we can obtain a better estimator for the vector of the means of all three quantities. Thus, if we are to only estimate the average weight of a newborn, it does not help to include data from the bricks or passengers into the estimation. But again, if we want to estimate the vector of the mean of all three quantities, the James-Stein estimator is better than the individual arithmetic averages. In this case, "better" means that the James-Stein estimator has a reduced mean squared error.

Single-Target Shrinkage (James-Stein Shrinkage)   The shrinkage algorithm allows for improved estimation of the mean with respect to expected mean squared error (EMSE) [3]. James-Stein shrinkage

---

[2]This example was constructed based on the "Baseball example" in Efron and Morris (1977)

[3]The analog shrinkage estimation for covariance matrices (also called Ledoit-Wolf shrinkage) is discussed in Section 2.4.2

(James and Stein, 1961) yields an estimator for the optimal shrinkage intensity in Eq. (4.2),

$$\lambda^{JS} = \operatorname*{argmin}_{\lambda} \mathbb{E} \left\| \boldsymbol{\mu}_{i,g} - \hat{\boldsymbol{\mu}}_{i,g}^{reg}(\lambda) \right\|^2 = \frac{\sum_d \operatorname{Var}(\hat{\mu}_{d,i,g}^s)}{\sum_d \mathbb{E}\|\hat{\mu}_{d,i,g}^s - \hat{\mu}_{d,i,\bar{g}}\|^2}. \tag{4.3}$$

Replacing with sample estimates, we obtain

$$\hat{\lambda}^{JS} = \frac{\sum_d \widehat{\operatorname{Var}}(\hat{\mu}_{d,i,g}^s)}{\sum_d \|\hat{\mu}_{d,i,g}^s - \hat{\mu}_{d,i,\bar{g}}\|^2}. \tag{4.4}$$

Compared to computational expensive cross-validation, the optimal shrinkage strength can be calculated according to Eq. 4.4 with very low computational cost which makes substitution attractive. Thus, James-Stein shrinkage can be applied to obtain the regularization parameters for the Single-Target Shrinkage scenario (cf. Fig. 4.2C), respectively the $\boldsymbol{\mu}_{i,g}^{STS}$ in Equation (4.2).

Multi-Target Shrinkage   A recently proposed generalization of the shrinkage framework to multiple targets (Multi-Target Shrinkage, MTS) allows for the simultaneous estimation of the $k-1$ parameters in Eq. (4.1) (Bartz et al., 2014). Minimizing the EMSE of the regularized estimator leads to a quadratic program:

$$\boldsymbol{\lambda}^{\star} = \operatorname*{argmin}_{\lambda} \mathbb{E} \left\| \boldsymbol{\mu}_{i,g} - \hat{\boldsymbol{\mu}}_{i,g}^{reg}(\lambda) \right\|^2 = \operatorname*{argmin}_{\lambda} \frac{1}{2} \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{A} \boldsymbol{\lambda} + \mathbf{b}^{\mathrm{T}} \boldsymbol{\lambda}, \tag{4.5}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{g-1}, 0, \lambda_{g+1}, \ldots, \lambda_k)$. Sample estimates for the parameters $\mathbf{A}$ and $\mathbf{b}$ are given by

$$\hat{A}_{qr} = \sum_{d=1}^{D} \left(\hat{\mu}_{d,i,q} - \hat{\mu}_{d,i,g} \right) \left(\hat{\mu}_{d,i,r} - \hat{\mu}_{d,i,g} \right) \tag{4.6}$$

$$\hat{b}_q = \sum_{d=1}^{D} \widehat{\operatorname{Var}}(\hat{\mu}_{d,i,q}) \tag{4.7}$$

The formulation as a quadratic program allows for imposing additional constraints. Since MTS can be seen as providing a weighting of data points, it makes sense to constrain the weight of the data points in $\hat{\mu}_{l \neq g}$ to be lower than the weights of the data points in $\hat{\mu}_g$:

$$\forall l \neq g: \quad \lambda_l n_l^{-1} \leq (1 - \sum_{l \neq g} \lambda_k) n_g^{-1},$$

where $n_{g/l}$ is the number of data points in subclass g/l.

   While both, cross-validation and MTS permit the estimation of multiple regularization parameters, the time-demand of cross-validation increases exponentially in the number of parameters and quickly becomes infeasible. As an example, determining the ten optimal parameters in a simulated binary classification task (3000 data points, 700 dimensions, 6 subclasses, 4 folds, choosing parameters from 11 possible values) with cross-validation takes approx. 70.000 years, compared to 2.4 seconds necessary to train a classifier using MTS shrinkage. Determining the two optimal parameters in the STS scenario of the same classification task takes 48 minutes with cross-validation, while the shrinkage approach yields a classifier in 2.2 seconds.

   Thus, Multi-Target Shrinkage can be applied to efficiently estimate the regularization parameters for $\boldsymbol{\mu}_{i,g}^{MTS}$ in Eq. (4.1).

Using Mean Shrinkage to determine regularization parameters of subclass classifiers

When estimating the mean of subclass g, the data points of other subclasses resemble reasonable shrinkage targets to compute subclass-specific LDA classifiers. There are two possible regularization approaches which are visualized in Fig. 4.2D and E respectively. Fig. 4.2D depicts the Single-Target Shrinkage (STS) approach that aims to find an estimator which is regularized towards the pooled mean across all remaining data points. The STS approach assigns the same weight to all remaining data points. The STS approach yields a single regularization parameter, which can be determined by James-Stein shrinkage estimation (James and Stein, 1961).

Contrarily, Fig. 4.2E depicts the Multi-Target Shrinkage (MTS) approach. MTS considers several regularization targets, leading to an extended degree of freedom which is represented by $(n_{subclasses} - 1) \times n_{classes}$ parameters. In a binary classification scenario with six subclasses, this would result in ten parameters to be estimated.

The LDA classifier can then be computed with either the STS or the MTS mean estimator. The resulting classifier for STS is denoted "STS Subclass LDA". As the MTS approach is able to exploit structural neighborhood information of the data, the corresponding classifier is denoted "MTS subclass LDA", or "**Relevance Subclass LDA**" (**RSLDA**). Both methods also compute a regularized estimate for the covariance matrix. Data are first centered (i.e. subtracting the subclass-wise mean) and then the shrinkage covariance matrix is computed as described in Ledoit and Wolf (2004) and Blankertz et al. (2011).

However, as it was shown by Bartz and Müller (2013), high-variance directions tend to dominate the estimation of the shrinkage strength – for both covariance and mean estimation. As the eigenspectrum of neuronal data might be skewed (for ERP data, this was illustrated in Blankertz et al. (2011), Figure 7), it is advisable to down-weight the impact of high-variance directions. Therefore, data were whitened before applying the mean shrinkage algorithm.

4.2.6 Additional Baseline methods

As it was illustrated in Figure 4.2C, one straight-forward way to extract such subclass-specific information is to train the classifiers solely on the subclass-specific data. This approach might however be suffering from the highly reduced number of data points. Estimates for the mean μ and especially for the covariance C might become inaccurate. Therefore, C can be estimated on pooled data across all subclasses, while computing subclass-specific μ. This baseline method is called "Sample Subclass LDA" in the following.

In total, five classifiers were compared in this study, see Table 4.1.

4.2.7 Analyzing EEG Data with Subclasses

To evaluate and compare the novel classification approaches on real neuroimaging data, existing data sets of several ERP experiments were reanalyzed. All data sets were recorded for brain-computer interfacing studies. In each of these experiments, there is a set of k stimuli which are repetitively presented in a pseudorandom order, as it was also described in Experiment 1 – 4 of this thesis. The subjects had the task to attend to one specific stimulus (target), while neglecting all other stimuli (non-targets). Each stimulus was target stimulus at least once. Thus, for each data point (i.e. EEG epoch corresponding to one stimulus), the classifier needs to estimate whether or not the user was attending (target vs. non-target stimulus). However, the stimulus identity represents a meaningful subclass.

Table 4.1: *List of all LDA classifiers used in the EEG data analysis.*

Global LDA	LDA classifier with covariance shrinkage estimation (shrC); subclass information is disregarded (see Fig. 4.2B).
Sample Subclass LDA	The mean is computed on subclass specific data and shrC is done based on data from all subclasses (see Fig. 4.2C).
STS Subclass LDA	This novel classifier is based on a regularized mean for each subclass. Two regularization parameters $(\lambda_T, \lambda_{NT})$ are estimated by shrinkage with the mean over all remaining subclasses as shrinkage target (see Fig. 4.2D). The shrC is calculated based on data from all subclasses.
Relevance Subclass LDA (RSLDA)	This novel classifier is based on a regularized mean for each subclass. A set of $(n_{subclasses} - 1) \times n_{classes}$ regularization parameters λ_l^c are estimated by shrinkage with the means of each subclass as shrinkage targets (see Fig. 4.2E). The shrC is calculated based on data from all subclasses.
xval-STS Subclass LDA	The same as the STS Subclass LDA, while the regularization parameters were chosen with cross-validation instead of the shrinkage algorithm. Model selection was done with a 4-fold cross-validation with 11 candidate parameters {0, 0.1,...1} for each parameter (λ_T and λ_{NT}). The best performing parameters setting out of 121 configurations were selected.

In the example of an auditory BCI paradigm with k different stimuli, there are k subclasses and for each data point (i.e. EEG epoch) we know which stimulus was presented. As stimuli may differ in pitch, direction or intensity, it is highly plausible that those differences lead to subclass-specific features in the ERPs. For the PASS2D study, this was already illustrated in Figure 3.2 on page 32. Therefore, considering subclass-specific features might improve the classification performance.

Evaluation Data

We reanalyzed the calibration data of five ERP-based BCI experiments. Each data set exhibited specific characteristics, as they were differing in the stimulus modality as well as in the number of trials and subjects – see Table 4.2 for details. Note that the RSVP data set features a remarkable number of 30 subclasses, which can be considered as a rather extreme experimental design. Across all experiments, the data of 74 subjects were analyzed, providing a representative evaluation.

Table 4.2: *Details of the ERP data sets which were reanalyzed to evaluate the classifiers.*

Data Set	AMUSE	PASS2D	CenterSpeller	MVEP	RSVP
Modality	auditory	auditory	visual	visual	visual
# Subclasses	6	9	6	6	30
# Subjects	21	12	13	16	12
# Epochs	4320	2916	2040	2100	7200
# Targets	720	324	340	350	240
Reference	(Schreuder et al., 2011a)	(Höhne et al., 2011a)	(Treder and Blankertz, 2010)	(Schaeff et al., 2012)	(Acqualagna et al., 2010)

Feature Extraction

The widely used "subsampling approach" was taken (Höhne et al., 2012; Kindermans et al., 2014) for feature extraction: the EEG data were first epoched [-150 800] ms relative to the stimulus onset and baselined between [-150 0] ms. EEG epochs containing eye artifacts were excluded by an heuristic,

cf. (Höhne et al., 2012) and Chapter 2.5.1. Then, for each channel the mean amplitude value was computed in a fixed set of 12 intervals. Those intervals had a length of 40-60 ms and they were densely placed between 140 ms and 650 ms after stimulus onset. It should be noted that the global selection of such intervals circumvents any additional parameter selection, while the feature space becomes high-dimensional (e.g. 63 channels × 12 intervals = 756 dimensional feature space).

Based on those features, the classification accuracy was estimated with a 5-fold cross-validation (with 4 repetitions). Classifier weights and all additional parameters were solely estimated on the training data. To ensure the highest possible comparability, the artifact rejection and the division into training and test data were performed globally. Thus, each method was trained and applied on the exact same data and the binary classification accuracy was assessed with the area under the ROC curve, see Blankertz et al. (2011) for details.

4.2.8 Analyzing fMRI Data with Subclasses

To see whether the proposed Relevance Subclass LDA approach is also suitable for other types of neuroimaging data, we applied RSLDA to an fMRI data set published previously (for methodological details, see Hebart et al. (2012)). In this experiment, subjects (N=21) had to judge the dominant direction of motion of random dot kinematograms with different levels of motion coherence. The highest level of motion coherence had been set to 50% and was clearly discernible. The remaining two levels of motion coherence had been adjusted to each subject's 65% and 85% correct threshold and had a mean coherence of 7.9% and 13.4%, respectively. In terms of motion coherence, the two lower levels were closer to each other than to the highest level of motion coherence. For our analysis, we used the dominant direction of motion (up vs. down) as the main class and the three coherence levels as subclasses.

For each subject, we fitted a general linear model to 8-10 runs of preprocessed fMRI data (except for spatial normalization and smoothing). Each trial was fit by a canonical hemodynamic response function regressor with onset and duration of stimulus presentation. This yielded a total of 128-160 parameter estimates per subclass and 384-480 parameter estimates in total. A searchlight analysis (Kriegeskorte et al., 2006) was conducted to detect brain regions where the classifier can exploit information about the direction of motion. This approach runs a classification analysis within a sphere around a given voxel, while the classification output (e.g. cross-validation accuracy) gets assigned to the center voxel. This process is repeated for each voxel in the brain. The sphere had a radius of 10 mm, encompassing 139 voxels. This means that per subject we approached a total of approximately 40.000 binary classification problems with subclasses. For each of these problems, a classifier was trained upon 192-240 data points per class (384-480 data points in total, 128-160 data points per subclass) with 139 dimensions. Both, the global LDA and the RSLDA classifiers were applied in the cross-validation procedure. The estimated classification accuracies for both approaches were reported.

4.3 Results

4.3.1 Classification Performance on ERP data

Fig. 4.3 shows a comparison of the binary accuracy on ERP data for all five classification approaches. Each scatter plot relates the RSLDA (y-axis) to one other approach (x-axis). It can be seen that RSLDA outperforms all other approaches except the STS Subclass classifier, which showed an equal performance. Importantly, it can be seen that none of the 74 subjects exhibited a notably worsened performance with RSLDA. Especially the poorly performing subjects featuring below 70% binary
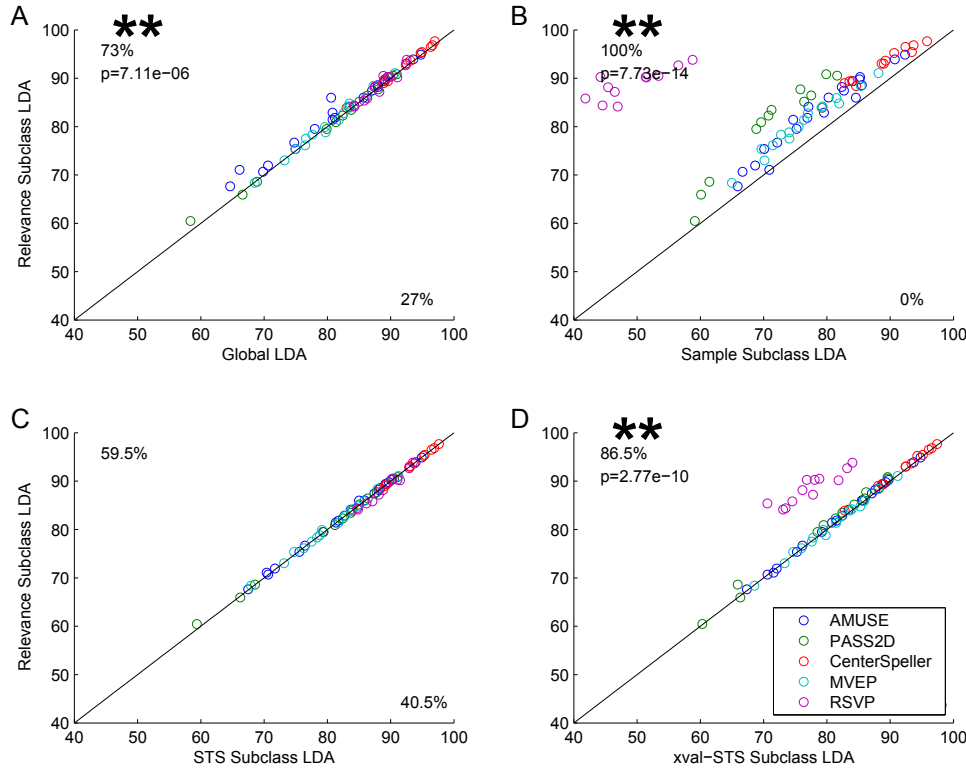
Figure 4.3: *Overview of the classification performances with RSLDA and other baseline methods. Each scatter plot shows the accuracies of the Relevance Subclass LDA approach (y-axis) against one of the three other approaches (x-axis). Five data sets were analyzed and marked with an individual color. A circle corresponds to one subject. Significant differences (2-sided Wilcoxon signed rank test with $p < 0.05/p < 0.01$) are marked with */**.*

classification accuracy could benefit from the subclass-specific classifiers. Exemplarily, one subject from the AMUSE data set featured a binary accuracy of 66.7% with the global LDA approach, while improving to 71.4% with RSLDA. Fig. 4.3 also reveals that the Sample Subclass LDA is not suitable for this data. This holds especially for the RSVP data, which features 30 subclasses.

To further explore the differences within the subclass-specific classifiers, the discriminative spatial LDA patterns for one exemplary subject of the AMUSE paradigm are shown in Figure 4.4. Such patterns reflect the discriminative neural source which the classifier is exploiting. They can be computed by $P = \mu_{\text{target}} - \mu_{\text{non-target}}$ for each classifier, see Haufe et al. (2014) for details.

For our approach with spatio-temporal features, we obtain one scalp topography per time interval. In order to limit the number of scalp maps to inspect, we trained classifiers here on three time intervals only, while we used 12 time intervals for the classification results reported in Figure 4.3. The resulting scalp maps are shown in Figure 4.4, where each row corresponds to one time interval. The global LDA approach yields only one classifier and therefore only one pattern for each interval. RSLDA computes one classifier for each of the six subclasses of the AMUSE paradigm. This results in 18 scalp patterns with each column corresponding to one subclass. Investigating such scalpmaps, we find that subclass 3 and 4 exhibited a distinct neural response compared to the other subclasses. RSLDA can exploit such differences, which yields in a superior classification performance.
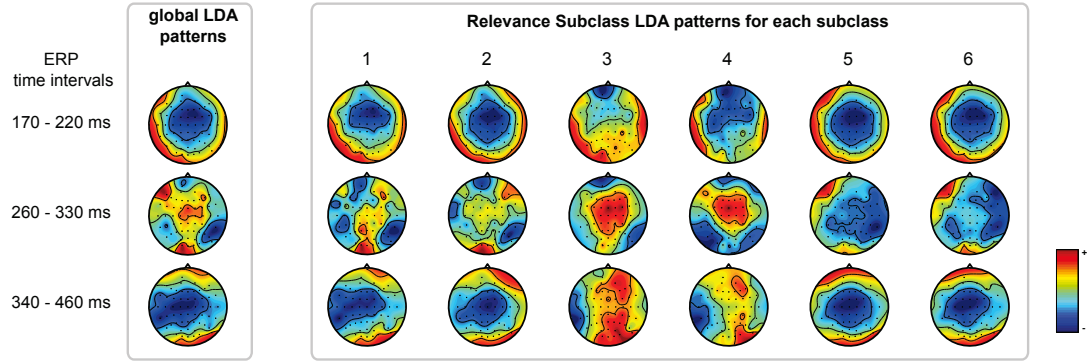
Figure 4.4: *Scalpmaps of the LDA and RSLDA patterns ($\mu_2 - \mu_1$) for three ERP time intervals. Data were taken from one subject of the AMUSE data set. The left plot shows the class discriminative patterns of the global classifier. The subclass-specific patterns extracted by Relevance Subclass LDA are shown in teh right plot. It can be observed that the subclasses 3 and 4 are highly distinct from the remaining four subclasses.*

Table 4.3: *Average classification accuracies and standard deviations across subjects. The individual data points are also plotted in Fig. 4.3 and listed in Table A.3.*

Data set	RSLDA	STS Subclass LDA	Sample Subclass LDA	Global LDA	xvalSTS
AMUSE	82.74 ± 7.8	82.55 ± 7.8	78.61 ± 7.5	81.46 ± 8.5	82.42 ± 7.6
PASS2D	80.15 ± 9.5	80.14 ± 9.7	71.03 ± 7.3	80.16 ± 9.8	79.17 ± 9.4
CenterSpeller	92.41 ± 3.9	92.41 ± 3.9	88.24 ± 4.8	91.97 ± 4	91.89 ± 4.1
MVEP	80.61 ± 5.7	80.50 ± 5.7	76.33 ± 5.9	80.33 ± 5.5	80.56 ± 5.7
RSVP	88.56 ± 3.1	89.10 ± 2.8	48.36 ± 5.7	88.38 ± 2.7	77.36 ± 4

4.3.2 Reanalyzing Online BCI Data

The shortcomings of state-of-the art classifier are described in Chapter 3 – in particular in Section 3.1.3 on page 31. Those shortcomings motivate the development of the RLSDA method, which is introduced in this chapter. The previous sections describe the performance of RSDLA on the offline calibration data of multiple ERP data sets. However, we also perform a pseudo-online experiment and reanalyze the online data of the PASS2D experiment (see Experiment 1 on page 28) and the AMUSE data (Schreuder et al., 2011a). Both, RSLDA classifiers and the global LDA classifiers were trained on calibration data and applied on online spelling data. Figure 4.5 depicts scatter plots of the multiclass accuracy, showing that RSLDA outperformed global LDA for both data sets. Thus, while being motivated by the outcome of Experiment 1, the applicability of RSLDA could also be demonstrated on online BCI data.

4.3.3 Classification Performance on fMRI data

The outcome of the searchlight analysis of RSLDA and global LDA is shown in Fig. 4.6. Results were thresholded at $p < 0.001$ uncorrected (minimum cluster size = 30). Both RSLDA and global LDA yielded the same two regions with significant discrimination in occipital in the calcarine sulcus (MNI: [-18, -102, -3]; [3, -87, 12]) and another region in right ventrolateral prefrontal cortex (MNI: [42, 21, 3]), probably related to the decision-making task on the stimuli (Hebart et al., 2014). Importantly, RSLDA found an additional region in left lateral mid-occipital gyrus (MNI: [-30, -84, 33]) superior to
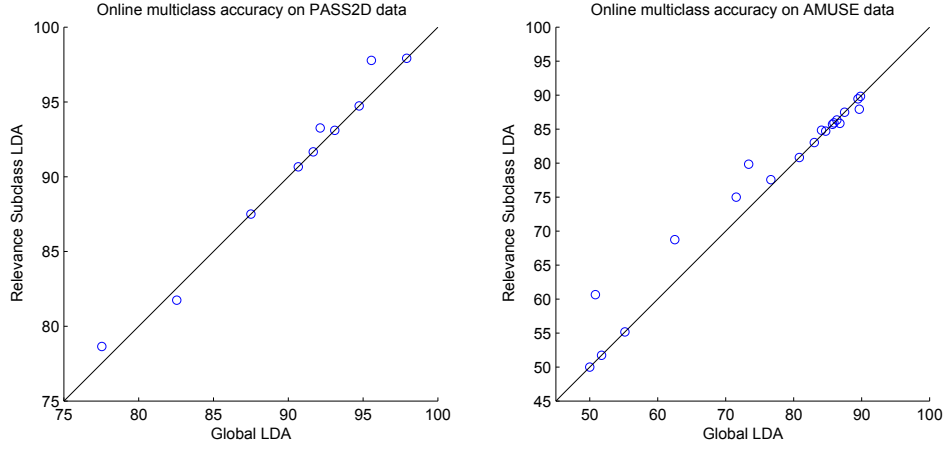
Figure 4.5: *Multiclass classification accuracy on the pseudo-online data from the PASS2D and the AMUSE data.*

motion-sensitive area MT+/V5.

This demonstrates that RSLDA was more sensitive than standard LDA in detecting information buried in brain activity patterns which otherwise would have remained below significance threshold.

4.3.4 Interpretation of Regularization Parameters

The preceding paragraphs describe how Relevance Subclass LDA outperforms the global LDA approach for both, ERP and fMRI data. In this section, we uncover the underlying characteristics of the novel RSLDA approach by analyzing the regularization parameters. The distribution of these parameters reveals the internal subclass-specific structure in the data, which RSLDA can exploit.

As described in Eq. (4.1), the regularized mean estimate for each subclass comprises of $k-1$ parameters. This can be visualized as a matrix $\boldsymbol{L} \in k \times k$ per class. This matrix is called "regularization profile" in the following, and L_{ij} specifies how much the mean of subclass i is regularized towards subclass j. Thus, each row i corresponds to the regularization parameters used for subclass i, averaged across all subjects and cross-validation folds. The diagonal elements of the matrix resemble the weight of the sample mean, thus $L_{ii} = 1 - \sum_{l \neq i} \lambda_l$.

Fig. 4.7 shows the regularization profile (matrix $\boldsymbol{L}$), obtained by RSLDA for the ERP data of three data sets. Notably, the structural information which is observed in L can directly be related to the experimental design. Thus, for any subclass/stimulus i, the MTS algorithm chose subclasses j as regularization target, such that the stimuli i and j were sharing some physical properties (e.g. direction, pitch for an auditory stimulus). It should be stressed that the MTS algorithm chose such regularization parameters in a purely data-driven way, without any manual labeling of the meaning of the subclasses.

The choice of the regularization parameter for subclass i towards subclass j in the MTS algorithm corresponds to the similarity of the neural data of those subclasses. Based on the literature on neural processing of visual and auditory stimuli (Langers et al., 2005), it can thus be expected that the regularization parameters reflect physical properties of the stimuli (e.g. direction or pitch for an auditory stimulus). It should be stressed that the MTS algorithm chooses such regularization parameters in a purely data-driven way, without any manual labeling of the meaning of the subclasses.

For the AMUSE dataset, auditory stimuli were presented from the left and right side of the subject. This structure is also reflected in the regularization profile. In order to estimate the subclass-specific
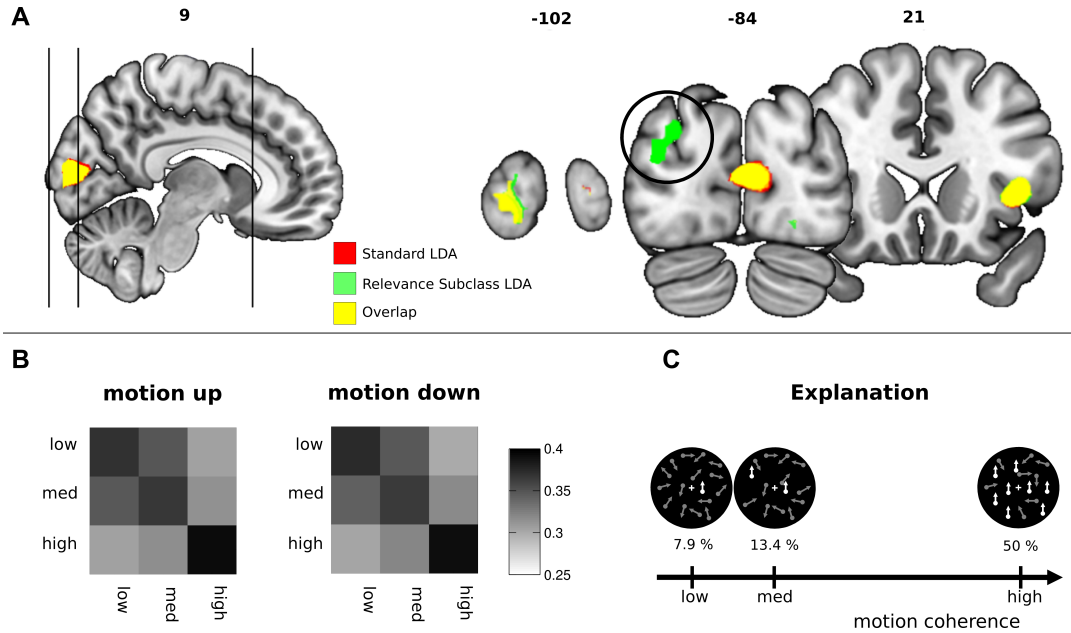
Figure 4.6: *Localization of discriminative brain areas (p < 0.001) found with global LDA and Relevance Subclass LDA (RSLDA) (**A**). While the overlap between both approaches is marked in yellow, the green areas mark discriminative brain activity which was only found with RSLDA. Plot (**B**) depicts the regularization profiles – obtained by RSLDA – for both classes, upwards and downwards motion. The regularization profiles in (**B**) can be explained by motion coherence used in the task design, as illustrated in (**C**).*

mean of a stimulus which was presented from the left side (e.g. subclass 5), the MTS algorithm assigns higher weights to those stimuli that were also presented from the left (i.e. subclass 4 and 6), compared to the remaining subclasses.

Also for the paradigms PASS2D and MVEP, the regularization profiles reflect stimulus characteristics that are common between several stimuli/subclasses. For the MVEP data, RSLDA classifiers regularize towards subclasses with visual motion stimuli that have a similar orientation. This leads to increased weights for neighboring subclasses in the regularization profiles. The regularization profiles of the PASS2D data highlight block-structures (of 3×3 blocks) as well as diagonal structures. They correspond to common properties of the auditory stimuli in this experiment: stimuli with the same pitch are represented by one block and the stimuli which were presented from the same direction are grouped in one diagonal – see Section 3.1.2 for details.

The regularization profiles of the fMRI data are shown in Figure 4.6B. Such profiles yield additional information about the similarity of the different motion stimuli (see Figure 4.6C). There were only three subclasses (i.e. levels of coherence) and the 3×3 matrices are depicted for both classes (i.e. up-wards/downwards motion). It can clearly be seen, that the two lower levels of motion coherence were regularized towards each other. Moreover, the high level of motion coherence was regularized stronger towards the intermediate level than to the lowest level of motion coherence. Thus, the the physical similarity of the stimuli is directly reflected in the regularization profile, which is computed by RSLDA and the MTS algorithm.
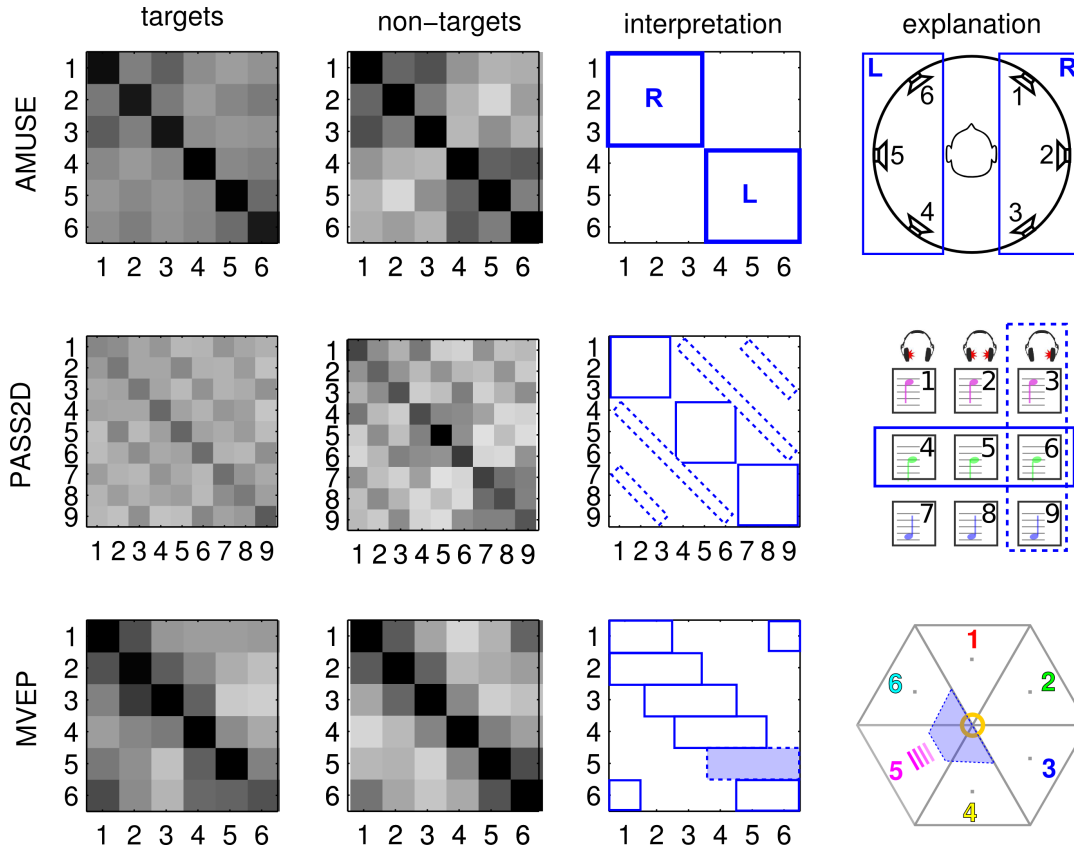
Figure 4.7: *Regularization profiles (i.e. distribution of the MTS regularization parameters) and their interpretation for three ERP data sets. The first two columns show the regularization profiles for targets and non-targets, averaged across subjects. The third column highlights the structure in the profiles. The forth column shows the details of the experimental paradigm, which explains the structure in the parameters.*

4.3.5 Limits

Whenever proposing a novel method, it is advisable to investigate its limits. Therefore, we investigated the performance of RSLDA in classification settings that are unfavorable for RSLDA, and compared it to the performance of global LDA. The ERP data (described in Table 4.2 and Fig. 4.3) was artificially modified in order to investigate the robustness of RSLDA to (1) noise in the subclass labels and (2) a low number of data points with a high dimensionality.

Noisy Subclass Labels   There might be scenarios in which the subclass labels are noisy or it is unknown whether or not such additional label information should be considered for the classification. This scenario was simulated with a permutation of the subclass labels of the ERP data. Thus, the subclass labels are completely random and do not reflect any plausible structure.

Fig. 4.8A shows a scatter plot for such data. As expected, it can be seen that RSLDA does not outperform standard global LDA in data with random subclass labels. However, RSLDA performs equally for all data sets except RSVP. It should be noted that RSVP represents generally the most extreme data set with 30 subclasses for the binary task, yielding a total number of $30 \times 29 = 870$
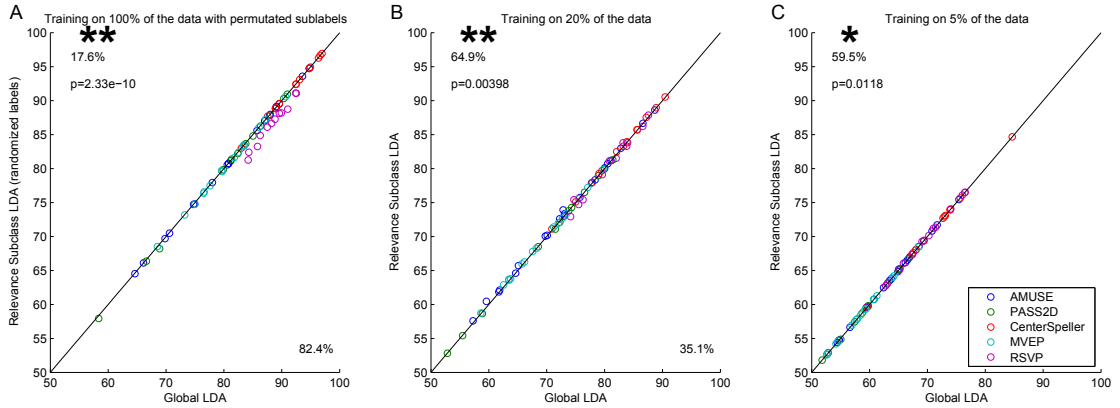
Figure 4.8: *Classification performance for different limit scenarios. Each scatter plot compares the classification accuracy of RSLDA and global LDA for the ERP data with artificial manipulations. Plot **A** shows the results for randomized subclass labels. Plot **B** and **C** show the results when computing the classifier with the correct subclass labels but using only subset (20% or 5%) of the training data.*

regularization parameters to be estimated. Hence, even if the subclass labels are describing irrelevant information, RSLDA does not worsen the classification accuracy for classification problems with a feasible number of subclasses.

Low Number of Data Points   As RSLDA is internally estimating numerous parameters, it is important to investigate the limits with respect to the number of data points which are available for training. Therefore, we evaluated the ERP data set with a reduced amount of training data. Fig. 4.8B-C depict the classification accuracy for all data sets, using only 20% or 5% of the data to train the classifier. As an example, 20% of the CenterSpeller data results in 408 data points per subject in total (68 for each subclass) with 72 targets (12 per subclass)[4]. Each data point comprises a feature vector with approximately 700 dimensions.

As expected, we found that the classification accuracy of both methods LDA and RSLDA decreases when using less training data. In such conditions with less training data RSLDA however (slightly) outperforms LDA when using only 20% or 5% of data to train the classifier. Thus, we could show that the novel method RSLDA is applicable for data sets with a low number of data points and high-dimensional features. The reason for this good performance of RSLDA in such scenarios might be counter intuitive on the first glance, since numerous parameters have to be estimated on a little amount of data points. However, investigating Eq. 4.5-4.7 it becomes obvious that the MTS algorithm finds the optimal regularization parameters by averaging over both, data points and feature dimensions. Hence, the MTS algorithm benefits from the high-dimensional feature space and suitable regularization parameters can also be estimated in such limit scenarios. However, the "suitable estimators" for this data will most likely be very close to the global average, such that RSLDA and LDA perform very similar for these limit scenarios.

---

[4]In the second example, 5% of the CenterSpeller data results in 102 data points per subject in total (17 for each subclass) with 18 targets (3 per subclass). Each data point comprises a feature vector with approximately 700 dimensions

4.4 Discussion

This chapter discusses binary classification problems for neuroimaging data. We investigate the shortcomings of existing methods in the presence of additional label information. Such additional label information can be formalized with the concept of "subclasses", if each data point is associated to exactly one class and one subclass. The exact meaning of a subclass depends on each individual problem.

Existing methods either disregard such subclass information (global approach), or focus on each subclass individually, which disregards the information that is shared between subclasses (subclass-specific approach). We propose a regularized approach and introduce the novel method "Relevance Subclass LDA" (RSLDA), which yields subclass-specific classifiers that exploit the relation between subclasses. The underlying regularization parameters can be estimated in a highly efficient manner, using the Multitarget Shrinkage algorithm (Bartz et al., 2014).

The proposed approach can be expected to improve classification performance, whenever the neural data of the subclasses is expected to be different on one hand, but also to exhibit information that is shared between subclasses. This is typically the case, if some parameters of the physical properties of the stimuli is varied within the experimental design. For instance, auditory evoked potentials depend on the direction of the sound source (difference between subclasses), and they are similar for sounds coming from neighboring directions (shared information). We also describe an fMRI study, in which the subjects perceived two conditions of stimuli (upwards/downwards moving dots) with varying characteristics (three levels of coherence). While a classifier is trained to identify the dominant direction of motion, the coherence levels can be regarded as subclasses. Moreover, the analysis of the limits of the proposed method has shown that the fulfillment of this condition is not critical. The proposed approach does not break down if the subclass information is not exploitable.

Reanalyzing an extensive amount of fMRI data (21 subjects) and EEG data (74 subjects), we show that RSLDA could outperform other state-of-the-art methods. Moreover, the RSLDA classifier also outputs regularization profiles, which can be interpreted in a meaningful way. Thus, RSLDA yields increased classification accuracy as well as a better interpretation for neuroimaging data. Both aspects are highly favorable, suggesting to apply RSLDA for classification problems within neuroimaging and beyond.

4.5 Lessons Learned

⋆ *Neuroimaging data might feature additional label information that can be formalized as subclasses.*

⋆ *The experimental structure can lead to subclass-specific features in neuroimaging data.*

⋆ *Exploiting subclass-specific features with RSLDA can result in an increased classification accuracy compared to baseline methods such as standard LDA.*

⋆ *RSLDA also outputs regularization profiles that allow interpreting the underlying subclass structure in the data.*

⋆ *RSLDA can be applied in classification scenarios with little amount of data with high-dimensional features, as the internal parameter estimation is gaining confidence when increasing the dimensionality of the data.*

⋆ *RSLDA is shown to be very robust. Even if the subclass labels do not contain valuable information, RSLDA does not worsen the classification accuracy compared to a global LDA. This holds, if the number of subclasses is not extremely high.*

LOCKED-IN PATIENTS CAN USE A BCI BASED ON MOTOR IMAGERY

ALTHOUGH one major goal of BCI research has always been to provide a communication pathway for locked-in individuals, studies are very rare. This chapter presents a BCI study with severely severely motor-impaired patients with little or no means of muscular control. A highly flexible BCI system is presented, that enables to establish BCI control for such patients within a very short time. Within only six experimental sessions, three out of four patients were able to gain significant control over the BCI, which was based on motor imagery or attempted execution. For the most affected patient, the BCI could outperform the best assistive technology (AT) of the patient in terms of control accuracy, reaction time and information transfer rate. We credit this success to the applied user-centered design approach and to a highly flexible technical setup. State-of-the art machine learning methods allowed the exploitation and combination of multiple relevant features contained in the EEG, which rapidly enabled the patients to gain substantial BCI control. Thus, we could show the feasibility of a flexible and tailorable BCI application in severely impaired users. This can be considered a significant success for two reasons: Firstly, the results were obtained within a short period of time, matching the tight clinical requirements. Secondly, the participating patients showed, compared to most other studies, very severe communication deficits. They were dependent on everyday use of AT and two patients were in a locked-in state. For the most affected patient a reliable communication was rarely possible with existing AT. The data and results were previously published in Höhne et al. (2014c).

5.1 Motivation

BCIs strive to decode brain signals into control commands, such that even severely handicapped people with no means of muscular control are enabled to communicate. A vast amount of studies have demonstrated the proof of concept, showing that healthy users are able to control noninvasive BCIs with a high accuracy and a communication rate of up to 100 bits/min (Bin et al., 2011). Translating brain signals into digital control commands, BCI systems can be applied for communication (Sellers and Donchin, 2006), interaction with external devices (e.g. steering a wheelchair) (Millan et al., 2009), rehabilitation (Daly and Wolpaw, 2008) or mental state monitoring (Blankertz et al., 2006b; Müller et al., 2008). While recent studies also investigated the neuronal underpinnings of BCI control (Halder et al., 2011; Grosse-Wentrup et al., 2011), the main objective of BCIs has always been to provide an alternative communication channel for patients that are in the locked-in state (Birbaumer et al., 1999; Kübler and Birbaumer, 2008; Kübler, 2013).

Although the proof-of-concept for noninvasive BCI technology has already been shown more than twenty years ago, patient studies are still very rare. Kübler (2013) recently pointed out that "fewer than 10% of the papers published on brain-computer interfacing deal with individuals presenting motor restrictions, although many authors mention these as the purpose of their research". Moreover,

within patient studies, those patients who were chosen to participate were rarely in need of a BCI, since their residual communication abilities with assisted technology (AT) were higher than the best state-of-the-art BCI could ever provide. Thus, there is a lack of studies with patients who are in a state that allows the BCI to become the best available communication channel. Some examples can be found in Birbaumer et al. (1999, 2000), Kübler et al. (2001), Murguialday et al. (2011), Kübler et al. (2005), Neuper et al. (2003), Birbaumer et al. (2008), Neumann and Kübler (2003), Birbaumer (2006), Hinterberger et al. (2005), Sellers and Donchin (2006), and Nijboer et al. (2008a), also being reviewed in Birbaumer and Cohen (2007), Kübler and Birbaumer (2008), and Mak and Wolpaw (2009). However, recent clinical studies have shown that it is even possible to set up BCI systems with patients in the complete locked-in condition. De Massari et al. (2013) introduced the idea of semantic conditioning as a potential alternative paradigm with completely paralyzed patients, and Cruse et al. (2011) applied a MI paradigm with patients diagnosed as being in the vegetative state. Moreover, patients with disorders of consciousness were trained to use BCI (Lulé et al., 2013), however, no functional communication could be achieved. These studies reveal that it may be possible to obtain significant classification accuracies for those patients, but it has not yet been shown that patients in complete paralysis can "reliably" use a BCI system (Sellers, 2013).

Our contribution describes the results of a MI-BCI study with four patients who showed severe brain damage. While all four patients had substantial difficulties with communication, two patients had a communication rate with their individually adapted AT of less than 5 bits/min. This means that for these participants, a BCI has the chance to become their individually best available communication channel, with all the beneficial implications for the Quality-of-Life of these patients (Holz et al., 2013b; Lulé et al., 2008).

The objective of this study is to show that the application of state-of-the-art machine learning methods allows to set up a MI-BCI system for patients in need of communication solutions within a very small number of sessions. We addressed this issue within a BCI gaming paradigm, which was specifically adapted to the needs of each patient according to user-centered design principles (Zickler et al., 2011). Both, the BCI system and the feedback application were optimized in an iterative procedure in order to account for the users' individual preferences. For the first time, automatically adapting classifiers, as well as hybrid data processing and classification approaches were applied online with (locked-in) patients. Moreover, a thorough psychological evaluation was done (Holz et al., 2013b).

More precisely, we demonstrate that by following the principle `"let the machine learn"` (Blankertz et al., 2002), patients gained significant BCI control within six sessions or less.

5.2 Experiment 5: Motor Imagery with Locked-in Patients

5.2.1 Patient Participants

The BCI system was tested with four severely disabled users in the information center of assistive technology, Bad Kreuznach, Germany. The patients were diagnosed with different diseases causing hemi- or tetraplegia. All patients were in a generally constant condition with no primary progress in their disease. No cognitive deficits were known. Table 5.1 summarizes disease- and demographic-related information. All patients had severe communication deficits and were using an AT solution on a daily basis. They had been continuously provided with individually optimized and cutting-edge AT (such as customized switches or eye-trackers) for more than five years. Only patient 3 had previously participated in BCI with MI training in a different study more than ten years ago - without gaining significant control (see patient *KI* in Kübler (2000) and Kübler and Birbaumer (2008)). It should be

noted that the patient numbering was ordered with decreasing residual communication abilities. Two of the four patients (patients 3 and 4) were in the locked-in state. Patients in the locked-in state are restricted in their voluntary motor control to such an extent that they are not able to communicate. This definition however makes an exception for one remaining communication channel. For most patients in the locked-in state, eye movements are the last remaining form of muscular control. If no remaining form of voluntary muscular activity is available (including the control of eye gaze, blink or button press), patients are considered to be in the "complete locked-in state".

Since different disagreeing definitions of the (complete) locked-in state exist, Table 5.1 also provides the communication rate with AT (measured as Information Transfer Rate (ITR) – cp. Section 2.6.3) as an additional measure. Communication rates with AT were empirically estimated by quantifying the time that the users needed to answer yes/no questions or ratings on a visual analog scale (VAS) in the evaluation process of this study. In the following paragraphs, each individual patient and his current physical condition is described in further detail.

Table 5.1: *Demographic and disease related data of all patients.*

	Patient 1	**Patient 2**	**Patient 3**	**Patient 4**
Age	47	48	45	45
Diagnosis	Tetraparesis after pons infarct	Hemiplegia after cerebral bleeding	Infantile cerebral palsy	Tetraparesis after cerebral bleeding
Artificial Ventilation	No	No	No	No
Artificial Nutrition (PEG)	No	No	No	Yes
Wheelchair	Yes	Yes	Yes	Yes
Residual muscular control	Eye-movement Speech Residual movement of right hand	Eye-movement Residual movement of left arm, hand and head Mimic	Eye movement (unreliable) Mimic Residual movement of right hand/arm	Eye-movement (highly unreliable) Residual movement of left thumb (depending on physical state)
Computer input device	Keyboard PC	Keyboard PC	Joystick/switch with hand letterboard with eye movements	Button press with thumb (yes/no): yes: 1 button press no: 2 button presses
Use of ICT on a daily basis	Yes	Yes	Yes	Yes
Experience with AT since	2006	1982	1986	2000
ITR with AT ICT	>30 bits/min	>30 bits/min	1-5 bits/min	0-2 bits/min
Experience with MI -BCI	No	No	Yes	No

Patient 1.   Amongst all patients enrolled in this study, patient 1 had the least impaired communication ability – being able to speak. Due to a stroke, his pronunciation is slurred, his language is considerably slowed down and needs to be amplified in volume. Although he has limited control over his left hand, he can reliably control his right hand to write, type or steer an electric wheelchair.

Patient 2.   Although lacking the ability to speak, patient 2 has high residual communication abilities since he can voluntarily control the left hand, left arm and his facial muscles. Thus, he can gesture

and also use a standard computer keyboard.

Patient 3.   Patient 3 is communicating with trained caregivers (partner-scanning) by controlling his eye gaze. He has been trying to use numerous ixeye-tracking systems, without gaining sufficient control. However, he can control a computer with a slow, weak but reliable control of his right forearm through the press of a button. Being highly motivated to use BCI technology, he already participated in a BCI study more than ten years ago (Kübler, 2000), which tested the control via slow cortical potentials (SCP) of the EEG. Unfortunately, he was not able to gain reliable control over the SCP-based BCI system in any session. Due to highly limited means of communication, a functioning BCI system would directly improve the quality of life of patient 3.

Patient 4.   Having the goal to provide communication solutions for people who can hardly communicate with AT or otherwise, patient 4 represents the ultimate end-user target group for BCI technology. The one exclusively known voluntary muscular control is a rather unreliable movement of his right thumb. He thus uses his thumb to press a button (pinch grip), which reflects the only available communication channel.

When starting the study, he had been in this condition for more than nine years. His communication is very slow and unreliable to the extent, that he is sometimes completely unable to communicate at all for several hours. In principle, he uses the button press in order to communicate an answer upon a question. A single button press would represent a *yes*-answer/agreement, while disagreements are expressed by two consecutive button presses. He shows a high variation within and across days of his attentiveness (he spontaneously falls asleep), of his mood, and of his responsiveness. The median time for a single button press is estimated to be 12 s, but delays of tens of seconds appear frequently (approx. 40%). The variation of responsiveness is the biggest communication hurdle: whenever patient 4 wishes to provide a negative response or disagreement, the second button press might be heavily delayed or not executed. Then the caregiver erroneously assumes an agreement. Given this communication quality and a communication rate at its best of 2 bits/min, patient 4 can be regarded to be close to the complete locked-in condition.

5.2.2 Study Protocol

The study protocol was approved by the Ethical Review Board of the Medical Faculty, University of Tübingen, Germany (case file 398/2011BO2). Written informed consent was obtained from each patient or their legally authorized representative. The study consisted of six EEG sessions per patient. There was not more than one EEG session per day and depending on the patient's condition, the session took 1-3 hours - including preparation time. Additionally, one introductory interview was conducted before the study and two interviews for evaluation were held after the last BCI session. Fig. 5.1A depicts details of the individual sessions. The psychological evaluation, with respect to the interview and questionnaires, is described in a separate article (Holz et al., 2013b).

In the first EEG session, every patient was screened to explore individual brain patterns and to select the two MI classes (left-hand, right-hand and foot imagery) which resulted in highest and most robust class-discriminability. Moreover, standard auditory oddball ERP recordings and a labeled recording for eye-movements, blinking artifacts and eyes open/closed measurements were performed during this screening session. MI training with feedback was not performed during this first EEG session, but only during the following five BCI sessions.

Each feedback session (2-6) was split in two parts: patients first executed a copy task (CopyTask), afterwards they received full control of the application in the free game mode (FreeMode). Patients 3 and 4 attempted to perform a motor action, while patients 1 and 2 used motor imagery. In each trial, the task was visually cued by an arrow, e.g. pointing rightwards or downwards (for right-hand

or foot imagery), see Fig. 5.1B. During both the CopyTask and the FreeMode, patients received online feedback (see Fig. 5.1C) of their targeted brain activation. However, in the CopyTask the outcome of a trial did not initiate an action in the game. In the FreeMode, the directional cue was replaced by a question mark and the gaming application was fully controlled by the BCI with two available actions: "select next column" and "place coin". Each action was represented by one MI class. The FreeMode was only started if the patient had reached sufficient control ($\geq$ 70%) in the CopyTask (leading to less frequent and shorter FreeMode phases for early sessions).

In order to reduce the number of unintended actions in the FreeMode, an action (placement of a coin or selection of the next column) was only performed if a predefined threshold had been exceeded by the BCI classifier. This resulted in "noDecision" trials if the threshold was not exceeded. Consequently no action was elicited for these trials. Introducing "noDecision" trials lead to a decreased fraction of incorrect decisions, yet at the same time to a reduction of communication rate (here: actions per minute and ITR). The ITR values reported throughout this study were calculated such that all pauses were taken into account (Höhne et al., 2011a).

Within the entire study, long durations of trials and inter-trial pauses led to an approximate speed of 4 trials/minute. Since one bit can be coded within one trial, the maximum achievable bit rate with this system was about 4 bits/min (with 100% correct trials). Although speeding up the communication rate by shortening the durations of trials and pauses would have been possible, we did not make use of this option in order to minimize the stress level and workload. Moreover, it should be noted that a reliable slow control might be preferable compared to a fast communication solution which is less reliable.

5.2.3 Application

Gaming applications represent a playful way to practice and improve the use of BCI systems, because they may provide long-term and short-term motivation. Moreover, we considered the frustration of erroneous actions in a game to be lower than erroneous selections of letters in a spelling task. Therefore, a computer version of the game "Connect-4" was used within all sessions. "Connect-4" is a strategic game, in which two players take turns in filling a matrix of free slots with coins. The objective of the game is to connect four of one's own coins of the same color vertically, horizontally, or diagonally. The two players are alternately placing their coins in one of the seven columns. The gaming application can be controlled by a 2-class motor imagery BCI, since only two actions are needed to play the game: (1) select the next column, or (2) place the coin in the current column. The software was implemented as a standalone java-application. Fig. 5.1C shows a screen shot of the application.

5.2.4 EEG Acquisition

Two different EEG systems were used within this study, both systems utilized passive gel electrodes. In the screening session, a 63-channel EEG system was used with most electrodes placed in motor-dense areas (cap: EasyCap, amplifier: BrainProducts, 2×32 channels, 1000 Hz sampling rate). One EOG channel was recorded additionally below the right eye. In sessions 2–6, a 16-channel EEG system was used (cap&amplifier: g.Tec, 1200 Hz sampling rate), while electrodes were placed symmetrically in areas close to the motor cortex. All EEG signals were referenced to the nose. Impedances were kept below 10 $k\Omega$, if possible. Data analysis and classification was performed with MATLAB (The MathWorks, Natick, MA, USA) using an inhouse BCI toolbox. For online processing and offline analysis, the EEG data was low-pass filtered to 45 Hz and down-sampled to 100 Hz.
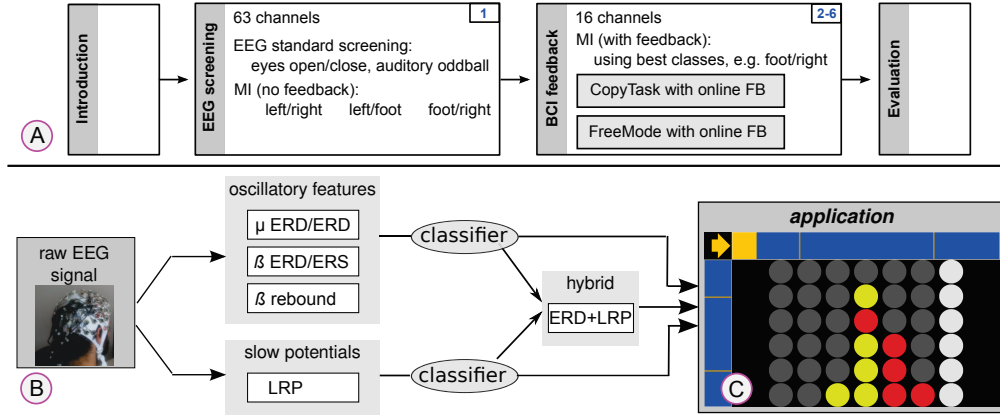
Figure 5.1: *The experimental design is shown in plot (A). Plot (B) depicts the architecture of the flexible BCI system which simultaneously considers oscillatory features and slow potentials. Two classifiers are applied and the feedback application is receiving simultaneous output of both classifiers and their weighted combination. A screen shot of the "Connect-4" application in mode FR (foot vs. right hand) is plotted in (C). In the top-left corner, the cue is presented (an arrow pointing to the right) and based on the BCI output, the yellow bar is either extending rightwards or downwards. The rightmost column is currently selected and visually highlighted.*

5.2.5 BCI Setup

This study focused on patients with severe brain injuries, thus the EEG signals and class-discriminative features were expected to be different to those known for healthy users. For this reason, the BCI was designed such that it could be driven by a wide range of features and their combinations. The incorporation of multiple features of the EEG or from other modalities into the BCI system is called a "hybridBCI" system (Dornhege et al., 2003b; Pfurtscheller et al., 2010; R. Millán et al., 2010; Fazli et al., 2012). Fig. 5.1B shows the architecture of the BCI system used for this patient study. The BCI simultaneously delivered three control signals to the application. Spectral features (event related desynchronization (ERD) in μ, β, δ band or β rebound) as well as slower movement-related potentials (i.e. lateralized readiness potential, LRP) were processed and classified. The two classifier outputs and their individually weighted sum were received by the application. The experimenter could then choose (based on a prior offline analysis of the data), which of the three output signals should be used to control the application.

5.2.6 Feature Extraction and Classification

To extract oscillatory features, signals were band-pass filtered by a Butterworth filter of order 5 in the individually defined spectral band. After visual inspection of the channel-wise ERD, a discriminative time interval was defined to compute optimized spatial filters with the Common Spatial Patterns (CSP) method (see Section 2.3.2 and Blankertz et al. (2008b)) and to train the classifier, a shrinkage-regularized linear discriminant analysis (LDA) (Blankertz et al., 2011). In analogy to Blankertz and colleagues (2008) (Blankertz et al., 2008b), offline classification accuracy was estimated using a (standard) cross-validation procedure, where the CSP filters and LDA weights were computed on the training set, and binary accuracy was assessed on the test set.

For the feature extraction of non-oscillatory slow potentials, raw EEG was band-pass filtered with a Butterworth filter (0.2–4 Hz) with a subsequent channel-wise baselining step (the interval of 300 ms

duration before trial onset). In analogy to ERP classification (Blankertz et al., 2011), the mean amplitude in a manually selected (class-discriminative) time interval was taken from each channel in order to form the feature vector of a trial. A binary classifier (again LDA) was trained based on those features.

Both LDA classifiers were automatically adapted during the CopyTask phase. As described in Vidaurre et al. (2011a), the pooled covariance matrix and the mean of the features was re-estimated after each trial, using the known labels (adaptation rate of 0.03). This also resulted in an implicit bias correction. In the FreeMode, no adaptation was performed. Besides the internal adaptation, the research team could recalibrate and fine-tune the classifiers between and within sessions. This was important in order to account for unstable features in the EEG data.

5.3 Findings

5.3.1 Standard Screening

The outcome of the standard screening (session 1) is depicted in Fig. 5.2. For patients 3 and 4 we found very atypical EEG signatures without any alpha or beta rhythms in the eyes-open and eyes-closed condition. It should be noted that these patients were unable to voluntarily open and close their eyes in response to an instruction/cue. Thus, eye-closure was supported by the caregiver who carefully moved the eyelids by hand.

5.3.2 ERD Features and BCI Performance

The BCI performance in this study was assessed for the two experimental conditions: during the Copy-Task, the labels are known and the BCI performance can easily be evaluated using the fraction of correct trials (called "binary accuracy" in the following). A trial is correct, whenever the accumulated BCI output is pointing to the correct direction at the end of the trial, thus chance level is 50%.
For the FreeMode, labels are unknown, unless the patient is able to report his intention with AT in each trial. Moreover, the number of games which were won against a computer heuristic can also be assessed as a complex and very high-level performance measure for the FreeMode. Playing the game with random control was simulated with the finding that a random player won 10% of the games and 20% of the games ended with a draw. Thus, the computer heuristic would win 70% of the games when playing against a player with random control.

Offline Analysis One interesting question was whether or not class discriminant features are found consistently across sessions. Therefore, Fig. 5.3 shows the results of an offline analysis of the CopyTask data. For all patients except patient 3, we found at least one discriminative feature (e.g. β ERD) which was consistently present in all sessions. Patients 3 did not present any reliable feature with discriminative information. Notably, none of the patients featured a consistent ERD component in the α band. However, the spatial distribution of such features was observed to be variable for some patients. Fig. A.5 visualizes the spatial distribution of class discriminative information for each patient across all sessions as scalp maps. This finding underlines the necessity of a flexible BCI system like it was used for this study. It should also be noted that the offline accuracy described in Fig. 5.3 cannot be directly translated into online BCI performance, as the cross-validation procedure was performed for each session separately. The resulting online BCI performance can be lower, if the features changed between sessions (Samek et al., 2014). In a scenario of rather stable features across sessions, the
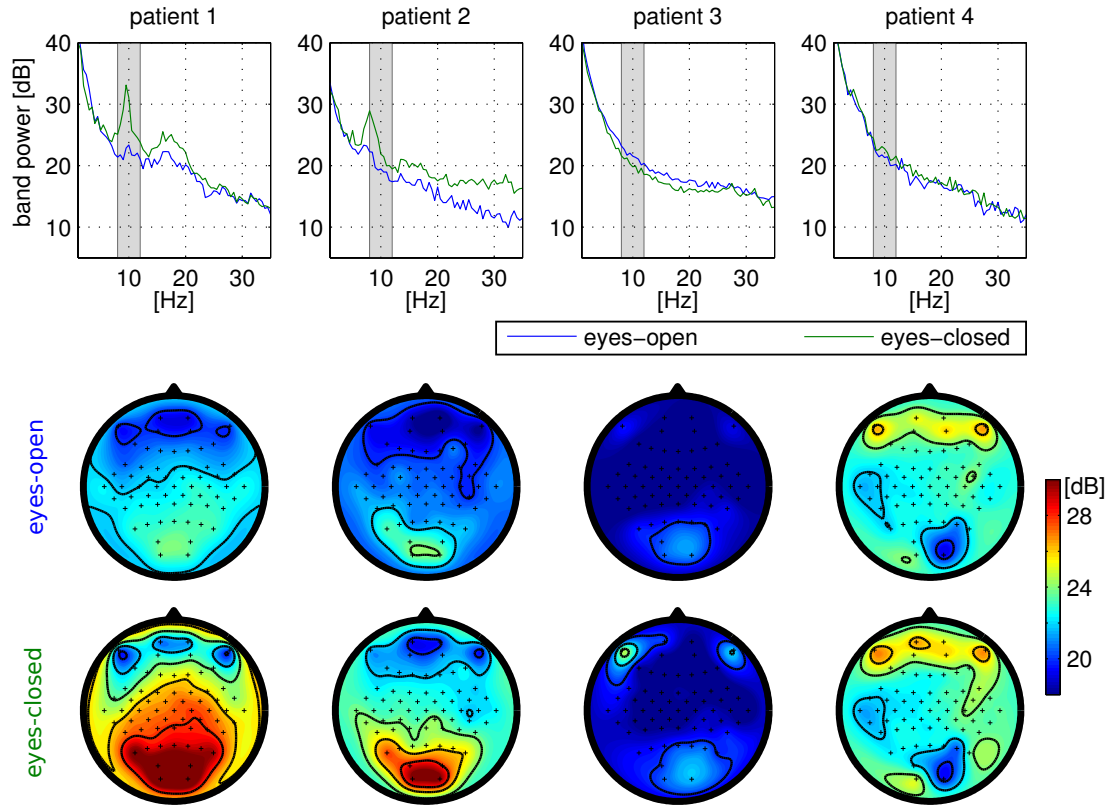
Figure 5.2: *Standard physiological screening of the four patients. The top row shows the spectra at electrode 'Cz' in the conditions eyes-open and eyes-closed. The spatial distribution of the channel-wise spectral power in the alpha-band [8-12 Hz] is depicted in the scalp maps of the lower row.*

online performance can also be higher, as the online classifier was trained with more data (from previous sessions).

Online BCI Control   Fig. 5.4 and Fig. 5.5 show the online performance of the CopyTask for all four patients. All patients except patient 3 could gain significant control over the BCI. Excluding patient 3, we obtained 10/14 sessions with an online binary accuracy being significantly better than chance. Again, one should stress that this was done with a patient population and there were no more than six EEG sessions with each patient, and five of these with BCI feedback. Fig. A.6 depicts the online accuracy in the FreeMode, which could only be assessed for patient 1 and 2.

In the following, EEG features and the resulting BCI performance for each of the four patients are discussed separately. After previously discussing offline results, we will only discuss online performances in the following.

Patient 1   Within the motor imagery study, a beta rebound as well as an LRP was found to be class-discriminant features for left-hand vs. right hand imagery, see Fig. 5.3. In the online framework, the beta-rebound was used to drive the system in session 4 and all following sessions. The LRP feature was not used, because it was more prone to (eye) artifacts and the patient featured involuntary eye-movements in the directions of the arrow. Although the beta-rebound was found quite consistently, the spatial distribution differed across sessions, see Fig. A.5. Therefore, it was required to retrain CSP
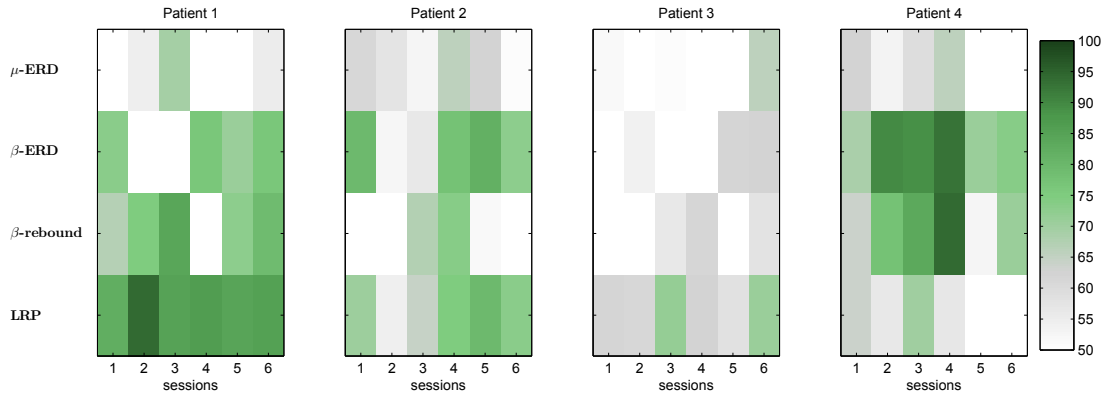
Figure 5.3: *Discriminative power of each feature across sessions, obtained with offline reanalysis of the CopyTask data. Global parameters such as the frequency band and time interval were chosen individually for each patient after manually inspecting the data from all sessions. For each session, the same global parameters were taken – which might be suboptimal. The classification accuracy was then estimated with cross-validation using the same parameters for each session. Note that the number of trails was varying across sessions with later sessions featuring fewer trials. Moreover, a β rebound was defined to as a discriminative feature in the β band, which was observed more than 500 ms after the end of a trial. As the β ERD of patient 4 was heavily delayed, it is also considered as β rebound in this analysis. Fig. A.5 shows the corresponding spatial distribution of discriminative information as scalp maps.*

filters and to use LDA with adaptation. The user was then able to gain significant[1] online control over the BCI, as shown in Fig. 5.4A. One can also observe that the BCI accuracy increased within sessions, resulting in the most reliable control towards the end of each session. The level of control was not perfect, but sufficient to drive the application in the FreeMode (cp. Fig. A.6). Patient 1 played the game Connect-4 five times in total, and he could win three of those games.

Patient 2   A beta ERD as well as a LRP were found to be class-discriminant features for left-hand vs. foot imagery, see Fig. 5.3. Since the beta ERD had a more consistent spatial pattern and was also less susceptible to artifacts, either the beta classifier or the meta classifier (beta + LRP) was used in the online BCI framework. However, although the ERD feature in the beta-band was found in almost every session, one could observe a high variation in class discrimination, spatial patterns as well as in BCI performance across and within sessions (see Fig. 5.4B and Fig. A.5). Due to the adaptive methods mentioned above, patient 2 was nevertheless able to control the game in the FreeMode at the end of session 4 and all following sessions (Fig. A.6). In total, he played four games in the FreeMode (winning two of them).

Patient 3   In analogy to a previous study (Kübler, 2000), reliable class discriminant features could not be found in the EEG data of patient 3 (cp. Fig. 5.3). He was thus not able to control the BCI system, as shown in the CopyTask performance in Fig. 5.4C. For the online framework, either the meta classifier or the LRP classifier were applied. None of them performed reliably above chance level.

---

[1]For a single block with #trials = 20, the significance criterion (one-sided χ^2 test with $p < 0.05$) for non-random control is 14 (=70%) correct trials. Thus, if only one block was observed, significant non-random BCI control would be shown if at least 70% of the trials we correct. However, when considering trials of all blocks within a session, the significance criterion of the average classification accuracy is considerably lowered. One example: for 100 trials (5 blocks × 20 trials), significance criterion for significant control is 59 correct trials (59%).
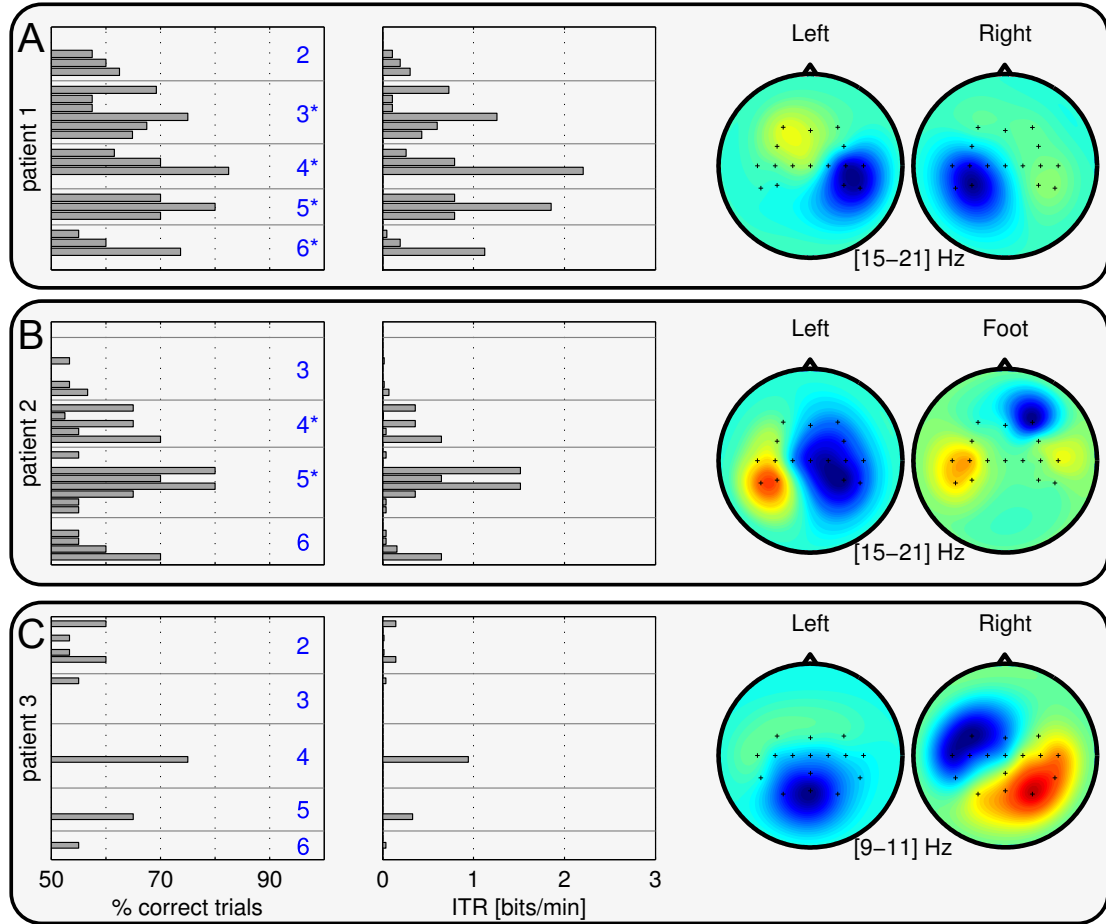
Figure 5.4: *Binary online accuracies (left column) and estimated bit rates (middle column) in the CopyTask for patients 1-3. Each bar represents one block of at least 20 trials. Session numbers are specified in blue color (left column). Session numbers with a * mark sessions with significant online BCI control across all trials (χ^2 test with $p < 0.05$). For patient 2, results for session 3 had to be disregarded due to technical problems. The right column depicts the scalp patterns of the most discriminant spectral features, based on data from all sessions. Results for Patient 4 are shown in Fig. 5.5.*

Recall, that this user displayed very atypical EEG spectra at rest (Fig. 5.2): during the eyes-open and eyes-closed conditions, no alpha or beta peaks were present. Due to the lack of BCI control, patient 3 did not officially enter the FreeMode (see study protocol). However, although featuring insufficient BCI control, patient 3 insisted in attempting to play the BCI game in the FreeMode ("for the fun of it"). He could neither gain control, nor was the resulting data analyzed in the present evaluation.

Patient 4 A highly discriminative β ERD component was present during each session of patient 4 (cp. Fig. 5.3). His motor-related EEG patterns exhibited typical spatial distributions (see Fig. 5.5A). This finding is even more surprising, since patient 4 revealed very atypically EEG signatures in the resting state – stereotypical brain rhythms such as α and β were absent (cf. Fig. 5.2).

Despite his physical condition, patient 4 achieved the best BCI control amongst the four patients.

Fig. 5.5A shows the online binary performance, revealing that he gained highly accurate online control (up to 90% binary accuracy) over the BCI system within the third EEG session (which was the second session with BCI feedback), and all following sessions. Even when pooling across all six sessions, his BCI control was highly significant (χ^2 test with $p < 0.001$). He exhibited very typical EEG activity during the right-hand and foot tasks of attempted motor execution, even though he had been unable to move his feet for more than nine years.

For this patient we could directly compare the communication rate of the BCI to his residual communication abilities with AT, by asking him to execute a button-press as soon as the corresponding cue appeared: we found, that the BCI-controlled feedback became discriminant after 1–3 seconds, while the button-press had a delay of 5–20 seconds — and sometimes the muscle contraction did not occur at all. As an example for this unbalanced communication behavior, a representative time window of 77 s was extracted for Fig. 5.5B. The interval contains six trials (three hand and three foot trials). The patient was requested to perform a button press in hand movement trial (marked in light magenta), but not during foot trials (marked in green). The BCI output and successful button presses are visualized. Patient 4 could only initiate a thumb muscle contraction successfully in two of the three trials. Moreover, any resulting button presses during this test were considerably delayed and occurred after the trial period of 7 s. The BCI, however, indicated the correct decisions at the end of each trial and even earlier in most cases. For the foot class, no motor action (i.e. muscle movement) was available; nevertheless the BCI could reliably detect the intention of a foot movement. Thus, to the best knowledge of the authors, this is the first quantitative report that shows that a BCI can uncover a patient's intention quicker and more reliable than the best available non-BCI AT.

Due to fatigue, temporal constraints and severe attention deficits, patient 4 entered the FreeMode only twice (sessions 4 and 6). In these two FreeMode sessions, he was not able to stay focused for more than 70 trials. As Table 1 reveals, he had the most severe deficits in communication. In practice, this means that he was mostly unable to communicate his intended action in the FreeMode. As a result, labels of the trials were not available and a data-driven evaluation of his BCI control in the FreeMode was impossible.
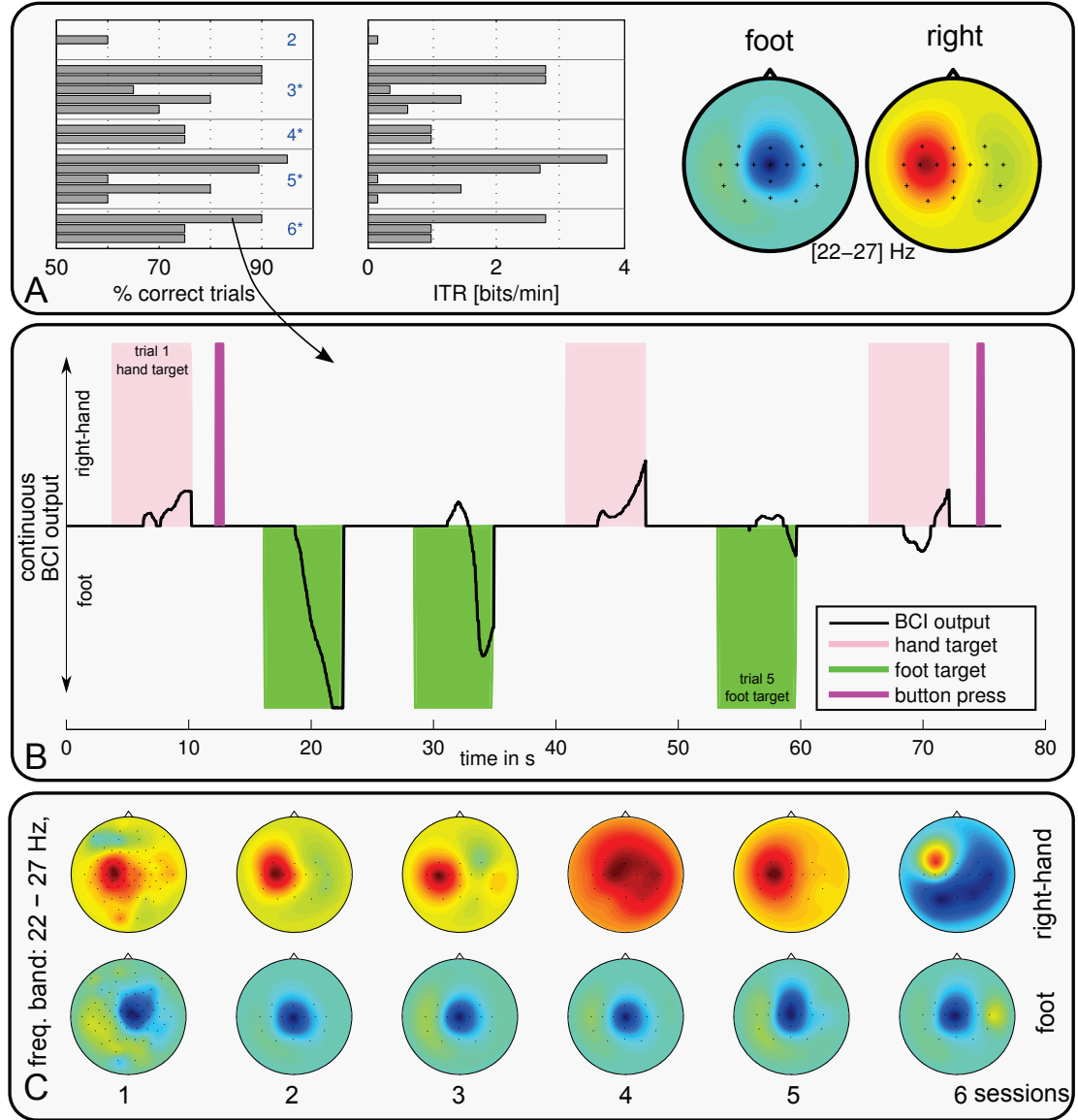
Figure 5.5: *BCI performance and scalp patterns of patient 4. Online binary accuracies, estimated bit rates (left, middle) of the CopyTask, and CSP patterns (right) averaged across all sessions are depicted in the top row (A). Each bar represents one block of at least 20 consecutive trials. Middle row (B) relates the continuous online BCI output to the residual muscle control (button press) for a representative time segment. Colored areas mark trial periods where the patient was asked to initiate a motor action. The excerpt shown was extracted from session 6, revealing that the BCI can detect the users intention far before a muscle contraction can be initiated. The lower row (C) depicts the motor related patterns in the β band for each session individually.*

5.4 Discussion

Four end-users with severe motor restrictions, who heavily depended on AT for communication and interaction in their daily life, agreed to participate in this study. Two of them were impaired in their communication ability to an extent, that no available AT would enable a reliable and – given their physical state – high speed solution. For these two specific patients, a BCI-based solution for control and communication would indeed introduce a novel communication quality. The BCI could enable independent communication and thus represent an added value compared to the AT presently used.

During the course of six BCI sessions, we found that three out of the four subjects could gain significant BCI control using motor imagery. For the most severely impaired patient (patient 4), we found evidence that the BCI outperformed his existing communication solution with AT in terms of accuracy and information transfer – being discussed in a following section.

The chosen end-user environment posed severe limitations in terms of user availability, their concentration span and the communication quality with their standard AT. We responded to these challenges with a flexible BCI framework, enabling us to tailor three major components of the study to the individual needs of the patient: (1) details of the experimental MI paradigm, (2) the form of data processing and type of exploited brain signals, and (3) the software application, which the user interacted with. Many of the internal modules of the BCI system could flexibly be exchanged and such changes remained invisible to the patients. The result was an "out of the box" BCI system, which was adapting itself to the features and needs of each user. Thus, our BCI system was generic and adaptive to meet the extensive requirements of such a pragmatic patient study.

Reducing the Number of Sessions using Machine Learning

With our study we could show, that end-users are able to gain significant online BCI control within six sessions or less. Compared to other end-user studies (Kübler et al., 2005) this is a very low number of sessions. Such a purposeful study design was enabled by the intense combined efforts of those users and the team, consisting of caregivers, psychologists, programmers and data analysts. We thereby followed the principles of user-centered design which implies an iterative process between developers and end-users of a product (see Zickler et al. (2011)). Thus, we used a setup which was flexible enough to adapt to the user's abilities and needs (e.g. choice of MI-classes, temporal constraints or the type of EEG feature such as ERD, β-rebound or LRP). Therefore, the system was designed to accommodate a wide variety of end users. Far from downplaying those individual contributions, the positive effect of advanced machine learning (ML) methods, such as hybrid classifiers with adaptation, should be mentioned. While motor-related BCI tasks are known to require a larger number of user training sessions compared to more salient ERP paradigms (Sellers et al., 2006b; Kübler et al., 2005; Nijboer et al., 2008a), we managed to apply our BCI system successfully within less than 6 sessions in three cases. While for one participant, no BCI control could be established, the remaining three participants gained sufficient online control to play the game relatively early on. (Patient 1: control from session three onwards, Patient 4: control from session four, and Patient 2: control from session five on.)

One crucial step for bringing BCIs closer to clinical application is to reduce the calibration time that is needed to establish a reliable BCI control – see also Section 2.7. In a comparable study with locked-in patients by Kübler et al. (2005), machine learning methods were not applied. Reliable performance was achieved only after a substantial number of sessions.

Patient 4

The case of patient 4 deserves special attention. While displaying severely impaired communication abilities, his level of BCI control was en par with very good unimpaired BCI users performing motor imagery.

This is presumably the most exciting finding of the current study, given that practically the full spectrum of AT solutions had been tested for this patient over the past nine years by AT experts. It should be noted that also ERP based paradigms were tested with patient 4 after the presented MI study. Discriminant ERP components could neither be found for a visual multi-class paradigm (MatrixSpeller (Farwell and Donchin, 1988)) nor for an auditory ERP paradigm (Höhne et al., 2012). The only applicable AT solution (the pinch-grip button press) provided a limited one-class signal with low accuracy and high temporal variability. Nevertheless, the BCI-controlled signal was relatively robust (with up to ∼90% accuracy) and available after 7 seconds at the latest.

Evaluating the speed and accuracy of his BCI control, we found evidence that the BCI could outperform his existing communication solution with AT in terms of accuracy and information transfer: during the online CopyTask, patient 4 accomplished commands which were presented visually through the software interface. Interestingly, he used *the same* (attempted) motor command for the right hand BCI class (i.e. the thumb movement) as for a real button press. Thus, a comparison of temporal dynamics and reliability of his BCI-responses with his button-press responses revealed interesting insights, as shown in Fig. 5.5B.

Contrary to the CopyTask mode, we could not show that patient 4 gained reliable control during the FreeMode. Even though the exact reason for this problem could not be clarified given the limited amount of data available for patient 4, the following – potentially accumulating – causes can be speculated: (1) identification problem, (2) attention problems and fatigue, (3) mental workload (4) self-initiation of actions. Section A.4.3 discusses all mentioned aspects in further detail.

Critical Assessment

The presented study involved only four patients with highly individual physical conditions and residual communication abilities. While the BCI experience with each individual patient can highlight important shortcomings of the current state-of-the-art BCI technology, any form of generalization is troublesome based on the data of four patients only. Therefore, it is necessary to conduct follow-up studies with a highly controlled study protocol and a larger number of patients – i.e. 20 or more. However, the recruitment of such patients might be challenging – especially if only those individuals are considered that could directly increase their communication abilities with the BCI.

In our study, we mostly report online BCI performance which was assessed in the "CopyTask" scenario. The Connect-4 application is not suitable for evaluating BCI control in a completely unconstrained setup, as it is troublesome to interpret the outcome of a single trial, if alternative communication is not accessible. Therefore, the evaluation of unconstrained BCI is more appropriate within different applications (such as a spelling application), as each single trial can be interpreted by the experimenter.

5.5 Conclusion

We could show that patients with severe motor impairments – even patients that are locked-in and almost completely locked-in – were able to gain significant control a noninvasive BCI by motor imagery. While applying state-of-the-art machine learning methods, this control was achieved within six or less sessions. The BCI was then used to operate a gaming application.

These findings are encouraging, since providing communication channels for patients' in-need resembles the major goal of the interdisciplinary research field of BCI. Moreover, our study describes one patient (patient 4), whose communication abilities with existing AT were on the same performance level ($\leq$ 2 bits/min) than his BCI control. In a controlled CopyTask framework, we found evidence that the BCI could even outperform his existing AT solution in terms of accuracy, reaction times and information transfer. Thus, we showed for this patient that neuronal pattern detection of an attempted motor execution can indeed be faster than the muscular output. Future studies may evaluate the BCI control in follow-up sessions, also testing spelling applications. Moreover, broader patient groups will be considered in order to further explore and evaluate the clinical usage of BCI.

5.6 Lessons Learned

⋆ *Severely motor impaired patients can learn to control a BCI based on motor imagery within less than six sessions.*

⋆ *Severely motor impaired patients display a wide range of EEG features when performing motor imagery or attempted motor execution. Interestingly, none of the four patients in this study featured a consistent ERD component in the α/μ band.*

⋆ *Neuronal pattern detection of an attempted motor execution can be faster than the muscular output. Thus, a communication pathway based on BCI can potentially outperform the communication with AT which is based on muscular control. This was shown for one patient.*

SUMMARY AND CONCLUSIONS

Technical advances have been enabling researchers to record and analyze an increasing amount of neural data. This poses a demand for appropriate analysis methods and novel experimental paradigms. This dissertation deals with the development and application of machine learning tools for Brain-Computer Interfacing (BCI) and Neurotechnology. The development cycle of BCI systems can be divided into three parts: application with healthy users, methods development and patient studies. This thesis presents contributions to all three parts – see Fig. 6.1.
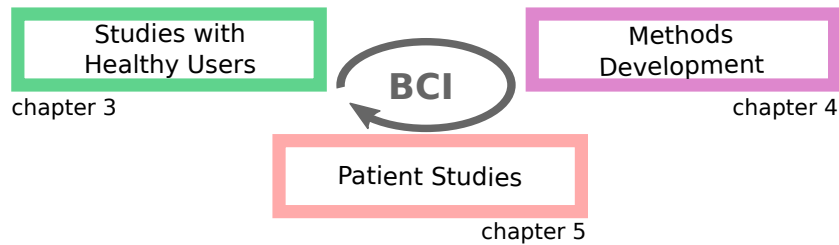
Figure 6.1: *The BCI development cycle.*

Studies with Healthy Users

This thesis contributes alternative communication solutions that can be applied by users with vision impairments. A total number of four ERP studies were conducted with healthy subjects in order to introduce and evaluate two novel auditory BCI paradigms. Moreover the importance of optimized stimulus parameters is investigated.

Two auditory BCI paradigms ("PASS2D" and "CharStreamer") are introduced in online BCI studies. Both paradigms enabled the users to complete their spelling tasks with a high accuracy and competitive speed. It is discussed that auditory BCI paradigm feature a higher workload and complexity than visual paradigms. Striving for the most user-friendly auditory BCI paradigm, this dissertation suggests methods and procedures, that allow to shift complexity from the user to the BCI system.

Firstly, this is demonstrated in the PASS2D paradigm, which implements a predictive text system in a 9-class BCI paradigm with two-dimensional auditory stimuli. With this design, PASS2D is the first auditory ERP speller that enables a letter selection within a single step.

Secondly, the CharStreamer paradigm is presented, featuring several novel characteristics that all contribute to a reduced complexity and workload for the user. As a result, the CharStreamer can be operated with an instruction as simple as "please attend to the character that you want to spell". The stimuli of the CharStreamer comprise 30 spoken sounds of letters and actions, also enabling a selection of a letter with a single step. Usability is further accounted for by an alphabetical stimulus presentation. This resembles a novel concept for BCI paradigms in general, as the user of the CharStreamer can foresee the presentation time of the next target stimulus, in contrast to the random presentation order which is commonly applied in BCI studies. Next to providing a user-friendly auditory spelling paradigm, the findings of the CharStreamer study reveal that a randomized stimulation order is not necessary to elicit class-discriminative ERP components.

This dissertation contributes two offline ERP studies that investigate optimized stimulus parameters for BCI. Both studies underline the importance to carefully choose the stimulation parameters for auditory paradigms.

Firstly, the use of naturalistic stimuli is proposed for auditory BCI paradigms. This is motivated by the idea to utilize the humans' over-trained ability of speech processing. Even though natural stimuli are more complex and less standardized than the artificial tones, they allow for better classification rates and lead to improved subjective ergonomic ratings. In short, it is shown that the auditory BCI becomes "faster" and it is considered "more pleasant" when using these natural stimuli.

Secondly, the importance of the stimulation speed in ERP-based BCIs is investigated. Therefore, a wide range of inter-stimulus intervals are applied in a simple binary oddball paradigm and corresponding ERP responses are analyzed. This analysis reveals that the choice of stimulation speed highly impacts the ergononomics, neurophysiology, as well as the classification accuracy and the resulting BCI performance. Especially the duration of late ERP components such as P300 is affected when changing the stimulation speed. Furthermore, this study provides a quantitative investigation, revealing that the average BCI performance can be increased by $\sim$2 bits/min (i.e. 10%), if the optimal stimulation speed is assessed for each user individually.

Methods Development

As the core methodological contribution, this thesis introduces a novel classification approach – called Relevance Subclass LDA (RSLDA). RSLDA is optimized for binary classification problems in the presence of addition label information, which is motivated by the findings from the studies with healthy users.

The origin and definition of addition label information (i.e. subclasses) is discussed on the basis of several experimental scenarios that are commonly applied in neuroimaging and BCI research. In the example of ERP data arising in BCI experiments, the stimulus identity is shown to be a reasonable subclass, as the differences in the stimulus characteristics can yield distinct ERP signatures.

State-of-the-art classification methods are reviewed and it shown that LDA is unable to exploit such subclass structure in a meaningful manner. However, RSLDA formalizes subclasses as regularization targets and seeks for an optimal exploration of the subclass structure in the data. This is achieved by applying the multi-target mean shrinkage algorithm, which can provide suitable regularization parameters in a highly efficient way. Thereby, RSLDA is shown to improve the classification accuracy for neuroimaging data compared to state-of-the art methods.

RSLDA moreover outputs a regularization profile, which serves as an excellent data-driven tool to visualize and interpret the underlying subclass structure in the data. This is illustrated for several EEG and fMRI datasets, in which numerous characteristics of the experimental design are directly reflected in the regularization profile. Thus, RSLDA not only yields an increased accuracy, it also provides a better understanding and interpretation of the latent structure in the data. Both above-mentioned aspects (performance and interpretability) are highly favorable for numerous classification problems within Neuroscience and beyond.

A Patient Study

The majority of international BCI research is devoted towards establishing brain-computer interfacing as a communication tool for locked-in patients. However, only a small fraction of BCI studies deal with individuals presenting motor restrictions. Even within patient studies, the participants are rarely in need of a BCI, since their residual communication abilities with assistive technology (AT) is higher than the best state-of-the-art BCI could ever provide. Thus, BCI studies with healthy users achieve technological progress, which is rarely tested and applied with patients in need.

This thesis contributes to close that gap with a BCI study with severely motor-impaired individuals. A highly flexible BCI system is presented, that enables to establish BCI control for such patients within a very short time. It is shown that within only six experimental sessions, three out of four patients can gain significant control over the BCI, which is based on motor imagery or attempted motor execution. This success is credited to a highly flexible setup, enabled by the combination of a user-centered design approach and state-of-the art machine learning methods. This allows the exploitation of multiple relevant features contained in the EEG, which leads to substantial BCI control in a short time.

It should be highlighted that the participating patients showed – compared to most other studies – very severe communication deficits. They were dependent on everyday use of AT and two patients were in a locked-in state. This study reveals that severely motor impaired individuals display a wide range of EEG features when performing motor imagery or attempted motor execution. However, they can rapidly learn to gain BCI control, if a suitable BCI system is applied. Moreover, this study shows for the first time, that the neuronal pattern detection of an attempted motor execution can be faster than the muscular output. Thus, a communication pathway based on BCI can potentially outperform the communication with AT which is based on muscular control. This can be considered a significant success, which might stimulate other BCI researchers to apply the recent technological advances with patients in need for a communication solution.

Closing Statement

Machine-Learning tools are one essential component for the recent advances in Neurotechnology and Brain-Computer Interfaces (BCIs). This dissertation presents novel methods as well as five experimental studies with both, healthy users and severely motor-impaired subjects. It is shown that the novel methods and the experimental paradigms contribute to further increase the usability and performance of the BCI technology.

Two auditory BCI paradigms are presented that both enable a fast and intuitive BCI spelling application to a naive user. These paradigms can be regarded as important steps towards non-visual spelling paradigms that can be operated by end-users with instructions as simple as "please attend to the character that you want to spell". Moreover, naturalistic stimuli are proposed to be suitable for auditory BCIs and the impact of the stimulation speed is investigated in detail.

However, shortcomings of state-to-the-art analysis techniques are identified, as the ERP data exhibit an internal subclass structure that is not yet exploited. Therefore, novel machine learning tools are introduced, which improve the classification accuracy for BCI data and also for other kinds of neuroimaging data. In addition, those methods are shown to be data-driven tools that allow an intuitive interpretation of the underlying structure in the data.

One major goal of BCI research is the development of a communication solution for locked-in individuals. Hence, this thesis presents an online BCI study with severely motor-impaired patients. Significant BCI control can be established for such patients within less than six sessions. This is enabled with a highly flexible BCI system which can simultaneously exploit a variety of EEG features which are related to motor imagery. Finally, for the first time a patient is described whose BCI control is superior to any existing communication solution in terms of control accuracy and information transfer rate.

BIBLIOGRAPHY

Acqualagna L, Blankertz B (2013). "Gaze-Independent BCI-Spelling Using Rapid Visual Serial Presentation (RSVP)". In: *Clin Neurophysiol* 124.5, pp. 901–908.

Acqualagna L, Treder MS, Schreuder M, Blankertz B (2010). "A novel brain-computer interface based on the rapid serial visual presentation paradigm". In: *Conf Proc IEEE Eng Med Biol Soc*. Vol. 2010, pp. 2686–2689.

Allison BZ, Pineda JA (2003). "ERPs evoked by different matrix sizes: implications for a brain computer interface (BCI) system". In: *IEEE Trans Neural Syst Rehabil Eng* 11, pp. 110–113.

Allison BZ, Pineda JA (2006). "Effects of SOA and flash pattern manipulations on ERPs, performance, and preference: implications for a BCI system". In: *Int J Psychophysiol* 59.2, pp. 127–140.

An X, Höhne J, Ming D, Blankertz B (2014). "Exploring Combinations of Auditory and Visual Stimuli for Gaze-Independent Brain-Computer Interfaces". In: *PLoS ONE* 9.10, e111070.

Ang KK, Chin ZY, Zhang H, Guan C (2009). "Robust filter bank common spatial pattern (RFBCSP) in motor-imagery-based brain-computer interface". In: *Conf Proc IEEE Eng Med Biol Soc* 2009, pp. 578–581.

Ang KK, Chin ZY, Wang C, Guan C, Zhang H (2012). "Filter Bank Common Spatial Pattern algorithm on BCI Competition IV Datasets 2a and 2b". In: *Frontiers in Neuroscience* 6.00039.

Baillet S, Mosher JC, Leahy RM (2001). "Electromagnetic brain mapping". In: *IEEE Signal Process Mag* 18.6, pp. 14–30.

Bartz D, Müller KR (2013). "Generalizing Analytic Shrinkage for Arbitrary Covariance Structures". In: *Advances in Neural Information Processing Systems 26*. Ed. by C Burges, L Bottou, M Welling, Z Ghahramani, and K Weinberger, pp. 1869–1877.

Bartz D, Höhne J, Müller KR (2014). "Multi-Target Shrinkage". submitted - available on arXiv.

Bell AJ, Sejnowski TJ (1995). "An Information-Maximization Approach to Blind Separation and Blind Deconvolution". In: *Neural Comput* 7.6, pp. 1129–1159.

Bentin S, Deouell LY (2000). "Structural encoding and identification in face processing: ERP evidence for separate mechanisms". In: *Cognitive Neuropsychology* 17.1-3, pp. 35–55.

Berger H (1929). "Über das Elektroenkephalogramm des Menschen". In: *Arch Psychiatr Nervenkr* 87, pp. 527–570.

Bin G, Gao X, Wang Y, Li Y, Hong B, Gao S (2011). "A high-speed BCI based on code modulation VEP". In: *J Neural Eng* 8, p. 025015.

Birbaumer N (2006). "Brain-computer-interface research: coming of age". In: *Clin Neurophysiol* 117, pp. 479–483.

Birbaumer N, Cohen L (2007). "Brain-computer interfaces: communication and restoration of movement in paralysis". In: *J Physiol* 579, pp. 621–636.

Birbaumer N, Ghanayim N, Hinterberger T, Iversen I, Kotchoubey B, Kübler A, Perelmouter J, Taub E, Flor H (1999). "A spelling device for the paralysed". In: *Nature* 398, pp. 297–298.

Birbaumer N, Kübler A, Ghanayim N, Hinterberger T, Perelmouter J, Kaiser J, Iversen I, Kotchoubey B, Neumann N, Flor H (2000). "The Thought translation device (TTD) for Completly Paralyzed Patients". In: *IEEE Trans Rehabil Eng* 8.2, pp. 190–193.

Birbaumer N, Murguialday AR, Cohen L (2008). "Brain-computer interface in paralysis". In: *Curr Opin Neurobiol* 21.6, pp. 634–638.

Blankertz B, Curio G, Müller KR (2002). "Classifying Single Trial EEG: Towards Brain Computer Interfacing". In: *Advances in Neural Inf. Proc. Systems (NIPS 01)*. Ed. by TG Diettrich, S Becker, and Z Ghahramani. Vol. 14, pp. 157–164.

Blankertz B, Müller KR, Curio G, Vaughan TM, Schalk G, Wolpaw JR, Schlögl A, Neuper C, Pfurtscheller G, Hinterberger T, Schröder M, Birbaumer N (2004). "The BCI Competition 2003: Progress and Perspectives in Detection and Discrimination of EEG Single Trials". In: *IEEE Trans Biomed Eng* 51.6, pp. 1044–1051.

Blankertz B, Müller KR, Krusienski D, Schalk G, Wolpaw JR, Schlögl A, Pfurtscheller G, R. Millán J, Schröder M, Birbaumer N (2006a). "The BCI Competition III: Validating Alternative Approachs to Actual BCI Problems". In: *IEEE Trans Neural Syst Rehabil Eng* 14.2, pp. 153–159.

Blankertz B, Dornhege G, Lemm S, Krauledat M, Curio G, Müller KR (2006b). "The Berlin Brain-Computer Interface: Machine Learning Based Detection of User Specific Brain States". In: *J Universal Computer Sci* 12.6, pp. 581–607.

Blankertz B, Dornhege G, Krauledat M, Müller KR, Curio G (2007). "The non-invasive Berlin Brain-Computer Interface: Fast Acquisition of Effective Performance in Untrained Subjects". In: *Neuroimage* 37.2, pp. 539–550.

Blankertz B, Kawanabe M, Tomioka R, Hohlefeld F, Nikulin V, Müller KR (2008a). "Invariant Common Spatial Patterns: Alleviating Nonstationarities in Brain-Computer Interfacing". In: *Advances in Neural Information Processing Systems 20*. Ed. by J Platt, D Koller, Y Singer, and S Roweis. Cambridge, MA: MIT Press, pp. 113–120.

Blankertz B, Tomioka R, Lemm S, Kawanabe M, Müller KR (2008b). "Optimizing Spatial Filters for Robust EEG Single-Trial Analysis". In: *IEEE Signal Process Mag* 25.1, pp. 41–56.

Blankertz B, Lemm S, Treder MS, Haufe S, Müller KR (2011). "Single-trial analysis and classification of ERP components – a tutorial". In: *Neuroimage* 56, pp. 814–825.

Brunner P, Joshi S, Briskin S, Wolpaw JR, Bischof H, Schalk G (2010). "Does the "P300" Speller Depend on Eye Gaze?" In: *J Neural Eng* 7, p. 056013.

Bünau P, Meinecke FC, Király F, Müller KR (2009). "Finding Stationary Subspaces in Multivariate Time Series". In: *Physical Review Letters* 103, p. 214101.

Castano-Candamil JS, Höhne J, Castellanos-Dominguez G, Haufe S (2015). "Solving the EEG Inverse Problem based on Space-Time-Frequency Structured Sparsity Constraints". in review.

Clemmensen L, Hastie T, Witten D, Ersbøll B (2011). "Sparse discriminant analysis". In: *Technometrics* 53.4, pp. 406–413.

Cohen RA, Kaplan RF, Zuffante P, Moser DJ, Jenkins MA, Salloway S, Wilkinson H (1999). "Alteration of intention and self-initiated action associated with bilateral anterior cingulotomy". In: *The Journal of Neuropsychiatry and Clinical Neurosciences* 11.4, pp. 444–453.

Cruse D, Chennu S, Chatelle C, Bekinschtein TA, Fernández-Espejo D, Pickard JD, Laureys S, Owen AM (2011). "Bedside detection of awareness in the vegetative state: a cohort study." In: *Lancet* 378.9809, pp. 2088–2094.

Dähne S, Höhne J, Tangermann M (2011a). "Adaptive Classification Improves Control Performance In ERP-Based BCIs". In: *Proceedings of the 5th International BCI Conference*. Graz, pp. 92–95.

Dähne S, Höhne J, Schreuder M, Tangermann M (2011b). "Slow Feature Analysis - A Tool for Extraction of Discriminating Event-Related Potentials in Brain-Computer Interfaces". In: *Artificial Neural Networks and Machine Learning - ICANN 2011*. Ed. by T Honkela, W Duch, M Girolami, and S Kaski. Vol. 6791. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 36–43.

Dähne S, Nikulin VV, Ramírez D, Schreier PJ, Müller KR, Haufe S (2014a). "Finding brain oscillations with power dependencies in neuroimaging data". In: *Neuroimage* 96, pp. 334–348.

Dähne S, Meinecke FC, Haufe S, Höhne J, Tangermann M, Müller KR, Nikulin VV (2014b). "SPoC: a novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters". In: *Neuroimage* 86.0, pp. 111–122.

Daly JJ, Wolpaw JR (2008). "Brain-computer interfaces in neurological rehabilitation". In: *Lancet Neurol* 7, pp. 1032–1043.

De Massari D, Ruf CA, Furdea A, Matuz T, Heiden L, Halder S, Silvoni S, Birbaumer N (2013). "Brain communication in the locked-in state". In: *Brain* 136.6, pp. 1989–2000.

De Vos M, Gandras K, Debener S (2013). "Towards a truly mobile auditory brain–computer interface: Exploring the P300 to take away". In: *Int J Psychophysiol*.

Dornhege G, Blankertz B, Curio G, Müller KR (2003a). "Combining Features for BCI". In: *Advances in Neural Inf. Proc. Systems (NIPS 02)*. Ed. by S Becker, S Thrun, and K Obermayer. Vol. 15, pp. 1115–1122.

Dornhege G, Blankertz B, Curio G (2003b). "Speeding up classification of multi-channel Brain-Computer Interfaces: Common spatial patterns for slow cortical potentials". In: *Proceedings of the 1st International IEEE EMBS Conference on Neural Engineering. Capri 2003*, pp. 591–594.

Dornhege, G, J del R. Millán, T Hinterberger, D McFarland, and KR Müller, eds. (2007). *Toward Brain-Computer Interfacing*. Cambridge, MA: MIT Press.

Duda RO, Hart PE, Stork DG (2001). *Pattern Classification*. 2nd edition. Wiley & Sons.

Dunlop M, Crossan A (2000). "Predictive text entry methods for mobile phones". In: *Personal and Ubiquitous Computing* 4.2-3, pp. 134–143.

Efron B, Morris CN (1977). *Stein's Paradox in Statistics*. WH Freeman.

Farwell L, Donchin E (1988). "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials". In: *Electroencephalogr Clin Neurophysiol* 70, pp. 510–523.

Fazli S, Mehnert J, Steinbrink J, Curio G, Villringer A, Müller KR, Blankertz B (2012). "Enhanced performance by a Hybrid NIRS-EEG Brain Computer Interface". In: *Neuroimage* 59.1, pp. 519–529.

Friederici A, Alter K (2004). "Lateralization of auditory language functions: A dynamic dual pathway model". In: *Brain and Language* 89.2, pp. 267–276.

Furdea A, Halder S, Krusienski DJ, Bross D, Nijboer F, Birbaumer N, Kübler A (2009). "An auditory oddball (P300) spelling system for brain-computer interfaces". In: *Psychophysiology* 46, pp. 617–625.

Gamble ML, Luck SJ (2011). "N2ac: An ERP component associated with the focusing of attention within an auditory scene". In: *Psychophysiology* 48.8, pp. 1057–1068.

Gao S, Wang Y, Gao X, Hong B (2013). "Visual and Auditory Brain-Computer Interfaces". In: *IEEE Trans Biomed Eng*.

Garrett D, Peterson DA, Anderson CW, Thaut MH (2003). "Comparison of linear, nonlinear, and feature selection methods for EEG signal classification". In: *Neural Systems and Rehabilitation Engineering, IEEE Transactions on* 11.2, pp. 141–144.

Geuze J, Farquhar JD, Desain P (2012). "Dense codes at high speeds: varying stimulus properties to improve visual speller performance". In: *J Neural Eng* 9.1, p. 016009.

Gonsalvez C, Polich J (2002). "P300 amplitude is determined by target-to-target interval". In: *Psychophysiology* 39.03, pp. 388–396.

Gonsalvez CJ, Barry RJ, Rushby Ja, Polich J (2007). "Target-to-Target Interval, Intensity, and P300 from an Auditory Single-Stimulus Task". In: *Psychophysiology* 44.2, pp. 245–250.

Green MD, Swets JA (1966). *Signal detection theory and psychophysics*. Huntington, NY: Krieger.

Grosse-Wentrup M, Scholkopf B, Hill J (2011). "Causal influence of gamma oscillations on the sensori-motor rhythm". In: *Neuroimage* 56.2, pp. 837–842.

Grozea C, Voinescu C, Fazli S (2011). "Bristle-sensors - Low-cost Flexible Passive Dry EEG Electrodes for Neurofeedback and BCI Applications". In: *J Neural Eng* 8, p. 025008.

Guo J, Gao S, Hong B (2010). "An Auditory Brain-Computer Interface Using Active Mental Response". In: *IEEE Trans Neural Syst Rehabil Eng* 18.3, pp. 230 –235.

Halder S, Rea M, Andreoni R, Nijboer F, Hammer EM, Kleih SC, Birbaumer N, Kübler A (2010). "An auditory oddball brain-computer interface for binary choices." In: *Clin Neurophysiol* 121.4, pp. 516–523.

Halder S, Agorastos D, Veit R, Hammer EM, Lee S, Varkuti B, Bogdan M, Rosenstiel W, Birbaumer N, Kübler A (2011). "Neural mechanisms of brain-computer interface control". In: *Neuroimage* 55, pp. 1779–1790.

Harmeling S, Dornhege G, Tax D, Meinecke FC, Müller KR (2006). "From outliers to prototypes: ordering data". In: *Neurocomputing* 69.13–15, pp. 1608–1618.

Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, Bießmann F (2014). "On the interpretation of weight vectors of linear models in multivariate neuroimaging". In: *Neuroimage* 87, pp. 96–110.

Hebart MN, Donner TH, Haynes JD (2012). "Human visual and parietal cortex encode visual choices independent of motor plans". In: *Neuroimage* 63.3, pp. 1393–1403.

Hebart MN, Schriever Y, Donner TH, Haynes JD (2014). "The Relationship between Perceptual Decision Variables and Confidence in the Human Brain". In: *Cereb Cortex*, p. 181.

Hill J, Farquhar J, Martens S, Bießmann F, Schölkopf B (2009). "Effects of Stimulus Type and of Error-Correcting Code Design on BCI Speller Performance". In: *Advances in Neural Information Processing Systems* 21, pp. 665–672.

Hill NJ, Lal TN, Bierig K, Birbaumer N, Schölkopf B (2004). "An auditory paradigm for brain-computer interfaces". In: *Adv Neural Inf Process Syst*, pp. 569–576.

Hill N, Schölkopf B (2012). "An online brain–computer interface based on shifting attention to concurrent streams of auditory stimuli". In: *J Neural Eng* 9.2, p. 026011.

Hill N, Lal T, Bierig K, Birbaumer N, Schölkopf B (2005). "An Auditory Paradigm for Brain–Computer Interfaces". In: *Advances in Neural Information Processing Systems*. Ed. by YW Saul L.K. and L Bottou. Vol. 17. Cambridge, MA, USA: MIT Press, pp. 569–576.

Hillyard SA, Hink R, Schwent VL, Picton TW (1973). "Electrical Signs of Selective Attention in the Human Brain". In: *Science* 182.4108, pp. 177–182.

Hinterberger T, Birbaumer N, Flor H (2005). "Assessment of cognitive function and communication ability in a completely locked-in patient." In: *Neurology* 64.7, pp. 1307–1308.

Hodge VJ, Austin J (2004). "A survey of outlier detection methodologies". In: *Artificial Intelligence Review* 22.2, pp. 85–126.

Höhne J, Tangermann M (2011a). "Natural stimuli for auditory BCI". In: *Neurosc Let*. Vol. 500. Supplement 1, e11.

Höhne J, Tangermann M (2011b). "Stimulation Speed Boosts Auditory BCI Performance". In: *Proc. 5th Int. BCI Conf. Graz*. Ed. by GR Müller-Putz, R Scherer, M Billinger, A Kreilinger, V Kaiser, and C Neuper. Graz: Verlag der Technischen Universität Graz, pp. 16–20.

Höhne J, Tangermann M (2012). "How stimulation speed affects Event-Related Potentials and BCI performance". In: *Conf Proc IEEE Eng Med Biol Soc*. Vol. 2012. IEEE, pp. 1802–1805.

Höhne J, Tangermann M (2014). "Towards User-Friendly Spelling with an Auditory Brain-Computer Interface: The CharStreamer Paradigm". In: *PLoS ONE* 9.6, e98322.

Höhne J, Schreuder M, Blankertz B, Tangermann M (2010). "Two-dimensional auditory P300 Speller with predictive text system". In: *Conf Proc IEEE Eng Med Biol Soc*. Vol. 2010, pp. 4185–4188.

Höhne J, Schreuder M, Blankertz B, Tangermann M (2011a). "A novel 9-class auditory ERP paradigm driving a predictive text entry system". In: *Front Neuroscience* 5, p. 99.

Höhne J, Schreuder M, Blankertz B, Müller KR, Tangermann M (2011b). "Novel Paradigms for Auditory ERP Spellers with Spatial Hearing: Two Online Studies". In: *Int J Bioelectromagnetism* 13.2, pp. 96–97.

Höhne J, Krenzlin K, Dähne S, Tangermann M (2012). "Natural Stimuli improve Auditory BCIs with respect to Ergonomics and Performance". In: *J Neural Eng* 9.4, p. 045003.

Höhne J, Bartz D, Müller KR, Blankertz B (2014a). "Analyzing Neuroimaging Data with Subclasses: a Shrinkage Approach". In: in preparation.

Höhne J, Blankertz B, Müller KR, Bartz D (2014b). "Mean shrinkage improves the classification of ERP signals by exploiting additional label information". In: *Proceedings of the 2014 International Workshop on Pattern Recognition in Neuroimaging*. IEEE Computer Society, pp. 1–4.

Höhne J, Holz EM, Staiger-Sälzer P, Müller KR, Kübler A, Tangermann M (2014c). "Motor Imagery for Severely Motor-Impaired Patients: Evidence for Brain-Computer Interfacing as Superior Control Solution". In: *PLoS ONE* 9.8, e104854.

Holz EM, Zickler C, Riccio A, Höhne J, Cincotti F, Tangermann M, Halder S, Mattia D, Kübler A (2013a). "Evaluation of Four Different BCI Prototypes by Severely Motor-Restricted End-Users". In: *Proceedings of the Fifth International Brain-Computer Interface Meeting 2013*. Ed. by J d. R. Millán, S Gao, GR Müller-Putz, JR Wolpaw, and JE Huggins. Verlag der Technischen Universität Graz, pp. 362–363.

Holz EM, Höhne J, Staiger-Sälzer P, Tangermann M, Kübler A (2013b). "Brain-computer interface controlled gaming: Evaluation of usability by severely motor restricted end-users". In: *Artificial Intelligence in Medicine* 59.2. Special Issue: Brain-computer interfacing, pp. 111 –120.

Hyvärinen A, Karhunen J, Oja E (2004). *Independent component analysis*. Vol. 46. John Wiley & Sons.

James W, Stein C (1961). "Estimation with quadratic loss". In: *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 1961, pp. 361–379.

Jin J, Allison BZ, Brunner C, Wang B, Wang X, Zhang J, Neuper C, Pfurtscheller G (2010). "P300 Chinese input system based on Bayesian LDA." In: *Biomed Tech (Berl)* 55.1, pp. 5–18.

Kandel ER, Schwartz JH, Jessell TM et al. (2000). *Principles of neural science*. Vol. 4. McGraw-Hill New York.

Kanoh S, Miyamoto K, Yoshinobu T (2008). "A brain-computer interface (BCI) system based on auditory stream segregation". In: *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of theIEEE*. IEEE. Vancouver BC, pp. 642–645.

Käthner I, Ruf CA, Pasqualotto E, Braun C, Birbaumer N, Halder S (2012). "A portable auditory P300 brain–computer interface with directional cues". In: *Clin Neurophysiol*.

Kaufmann T, Schulz S, Grünzinger C, Kübler A (2011). "Flashing characters with famous faces improves ERP-based brain–computer interface performance". In: *J Neural Eng* 8, p. 056016.

Kaufmann T, Schulz SM, Köblitz A, Renner G, Wessig C, Kübler A (2012). "Face stimuli effectively prevent brain–computer interface inefficiency in patients with neurodegenerative disease". In: *Clin Neurophysiol* 124.5, pp. 893–900.

Kaufmann T, Holz EM, Kübler A (2013). "Comparison of tactile, auditory and visual modality for brain-computer interface use: A case study with a patient in the locked-in state". In: *Front Neuroscience* 7.129.

Kawanabe M, Samek W, Müller KR, Vidaurre C (2014). "Robust Common Spatial filters with a Maxmin Approach". In: *Neural Computation* 26.2, pp. 1–28.

Kim DW, Hwang HJ, Lim JH, Lee YH, Jung KY, Im CH (2011). "Classification of selective attention to auditory stimuli: Toward vision-free brain-computer interfacing". In: *J Neurosci Methods* 197.1, pp. 180 –185.

Kindermans PJ, Verstraeten D, Schrauwen B (2012). "A bayesian model for exploiting application constraints to enable unsupervised training of a P300-based BCI". In: *PloS ONE* 7.4, e33758.

Kindermans PJ, Tangermann M, Müller KR, Schrauwen B (2014). "Integrating dynamic stopping, transfer learning and language models in an adaptive zero-training ERP speller". In: *J Neural Eng* 11.3, p. 035005.

Kleih SC, Nijboer F, Halder S, Kübler A (2010). "Motivation modulates the P300 amplitude during brain-computer interface use". In: *Clin Neurophysiol* 121, pp. 1023–1031.

Klobassa DS, Vaughan TM, Brunner P, Schwartz NE, Wolpaw JR, Neuper C, Sellers EW (2009). "Toward a high-throughput auditory P300-based brain-computer interface". In: *Clin Neurophysiol* 120, pp. 1252–1261.

Kornhuber HH, Deecke L (1965). "Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale". In: *Pflugers Arch* 284, pp. 1–17.

Kriegeskorte N, Goebel R, Bandettini P (2006). "Information-based functional brain mapping". In: *Proc Natl Acad Sci U S A* 103, pp. 3863–3868.

Krusienski DJ, Sellers EW, Cabestaing F, Bayoudh S, McFarland DJ, Vaughan TM, Wolpaw JR (2006). "A comparison of classification techniques for the P300 Speller". eng. In: *J Neural Eng* 3.4, pp. 299–305.

Küber A, Mattia D, Rupp R, Tangermann M (2013). "Facing the challenge: Bringing brain-computer interfaces to end-users". In: *Artificial Intelligence in Medicine* 59.2, pp. 55–60.

Kübler A (2000). *Brain-computer communication - development of a brain-computer interface for locked-in patients on the basis of the psychophysiological self-regulation training of slow cortical potentials (SCP)*. Tübingen: Schwäbische Verlagsgesellschaft.

Kübler A, Birbaumer N (2008). "Brain-computer interfaces and communication in paralysis: extinction of goal directed thinking in completely paralysed patients?" In: *Clin Neurophysiol* 119, pp. 2658–2666.

Kübler A, Nijboer F, Mellinger J, Vaughan TM, Pawelzik H, Schalk G, McFarland DJ, Birbaumer N, Wolpaw JR (2005). "Patients with ALS can use sensorimotor rhythms to operate a brain-computer interface". In: *Neurology* 64.10, pp. 1775–1777.

Kübler A, Furdea A, Halder S, Hammer EM, Nijboer F, Kotchoubey B (2009). "A brain-computer interface controlled auditory event-related potential (p300) spelling system for locked-in patients". In: *Annals of the New York Academy of Sciences* 1157, pp. 90–100.

Kübler A (2013). "Brain-computer interfacing: science fiction has come true". In: *Brain* 136.6, pp. 2001–2004.

Kübler A, Kotchoubey B, Kaiser J, Wolpaw J, Birbaumer N (2001). "Brain-Computer Communication: Unlocking the Locked In". In: *Psychol Bull* 127.3, pp. 358–375.

LaConte S, Strother S, Cherkassky V, Anderson J, Hu X (2005). "Support vector machines for temporal classification of block design fMRI data". In: *NeuroImage* 26, pp. 317–329.

Langers DRM, Dijk P, Backes WH (2005). "Lateralization, connectivity and plasticity in the human central auditory system." In: *Neuroimage* 28.2, pp. 490–499.

Langville AN, Meyer CD (2011). *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press.

Ledoit O, Wolf M (2004). "A well-conditioned estimator for large-dimensional covariance matrices". In: *J Multivar Anal* 88, pp. 365–411.

Leeb R, Sagha H, Chavarriaga R, R Millán J (2010). "Multimodal fusion of muscle and brain signals for a hybrid-BCI". In: *Conf Proc IEEE Eng Med Biol Soc*. IEEE, pp. 4343–4346.

Lemm S, Blankertz B, Curio G, Müller KR (2005). "Spatio-Spectral Filters for Improving Classification of Single Trial EEG". In: *IEEE Trans Biomed Eng* 52.9, pp. 1541–1548.

Lemm S, Blankertz B, Dickhaus T, Müller KR (2011). "Introduction to machine learning for brain imaging". In: *Neuroimage* 56, pp. 387–399.

Lopez-Gordo M, Fernandez E, Romero S, Pelayo F, Prieto A (2012). "An auditory brain–computer interface evoked by natural speech". In: *J Neural Eng* 9.3, p. 036013.

Lotte F, Congedo M, Lécuyer A, Lamarche F, Arnaldi B (2007). "A review of classification algorithms for EEG-based brain-computer interfaces". In: *J Neural Eng* 4, R1–R13.

Lulé D, Häcker S, Ludolph A, Birbaumer N, Kübler A (2008). "Depression and quality of life in patients with amyotrophic lateral sclerosis". In: *Deutsches Ärzteblatt international* 105.23, p. 397.

Lulé D et al. (2013). "Probing command following in patients with disorders of consciousness using a brain-computer interface". In: *Clin Neurophysiol* 124.1, pp. 101 –106.

Mainsah B, Colwell K, Collins L, Throckmorton C (2014). "Utilizing a Language Model to Improve Online Dynamic Data Collection in P300 Spellers". In: *IEEE Trans Neural Syst Rehabil Eng*.

Mak JN, Arbel Y, Minett JW, McCane LM, Yuksel B, Ryan D, Thompson D, Bianchi L, Erdogmus D (2011). "Optimizing the P300-based brain-computer interface: current status, limitations and future directions". In: *J Neural Eng* 8.2.

Mak JN, Wolpaw JR (2009). "Clinical applications of brain-computer interfaces: current state and future prospects". In: *Biomedical Engineering, IEEE Reviews in* 2, pp. 187–199.

Matsumoto Y, Makino S, Mori K, Rutkowski TM (2013). "Classifying P300 responses to vowel stimuli for auditory brain-computer interface". In: *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*. IEEE, pp. 1–5.

Millan J, Galan F, Vanhooydonck D, Lew E, Philips J, Nuttin M (2009). "Asynchronous non-invasive brain-actuated control of an intelligent wheelchair". In: *Conf Proc IEEE Eng Med Biol Soc*, pp. 3361–3364.

Misaki M, Kim Y, Bandettini PA, Kriegeskorte N (2010). "Comparison of multivariate classifiers and response normalizations for pattern-information fMRI". In: *Neuroimage* 53.1, pp. 103–118.

Müller KR, Mika S, Rätsch G, Tsuda K, Schölkopf B (2001). "An Introduction to Kernel-based Learning Algorithms". In: *IEEE Neural Networks* 12.2, pp. 181–201.

Müller KR, Blankertz B (2006). "Toward noninvasive Brain-Computer Interfaces". In: *IEEE Signal Process Mag* 23.5, pp. 125–128.

Müller KR, Anderson CW, Birch GE (2003). "Linear and Non-Linear Methods for Brain-Computer Interfaces". In: *IEEE Trans Neural Syst Rehabil Eng* 11.2, pp. 165–169.

Müller KR, Tangermann M, Dornhege G, Krauledat M, Curio G, Blankertz B (2008). "Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring". In: *J Neurosci Methods* 167.1, pp. 82–90.

Murguialday AR, Hill J, Bensch M, Martens S, Halder S, Nijboer F, Schölkopf B, Birbaumer N, Gharabaghi A (2011). "Transition from the locked in to the completely locked-in state: A physiological analysis". In: *Clin Neurophysiol* 122.5, pp. 925–933.

Näätänen R, Gaillard AWK, Mäntysalo S (1978). "Early Selective-Attention Effect on Evoked Potential Reinterpreted". In: *Acta Psychol (Amst)* 42.4, pp. 313–329.

Näätänen R, Gaillard AWK, Varey CA (1981). "Attention Effects on Auditory EPs as a Function of Inter-Stimulus Interval". In: *Biol Psychol* 13, pp. 173–187.

Näätänen R, Paavilainen P, Rinne T, Alho K (2007). "The Mismatch Negativity (MMN) in Basic Research of Central Auditory Processing: A Review". In: *Clin Neurophysiol* 118.12, pp. 2544–2590.

Nambu I, Ebisawa M, Kogure M, Yano S, Hokari H, Wada Y (2013). "Estimating the Intended Sound Direction of the User: Toward an Auditory Brain-Computer Interface Using Out-of-Head Sound Localization". In: *PLoS ONE* 8.2, e57174.

Neumann N, Kübler A (2003). "Training locked-in patients: a challenge for the use of brain-computer interfaces". In: *IEEE Trans Neural Syst Rehabil Eng* 11, pp. 169–172.

Neuper C, Müller G, Kübler A, Birbaumer N, Pfurtscheller G (2003). "Clinical application of an EEG-based brain-computer interface: A case study in a patient with severe motor impairment". In: *Clin Neurophysiol* 114.3, pp. 399–409.

Nicolas-Alonso LF, Gomez-Gil J (2012). "Brain computer interfaces, a review". In: *Sensors* 12.2, pp. 1211–1279.

Nijboer F et al. (2008a). "A P300-based brain-computer interface for people with amyotrophic lateral sclerosis". In: *Clin Neurophysiol* 119, pp. 1909–1916.

Nijboer F, Furdea A, Gunst I, Mellinger J, McFarland D, Birbaumer N, Kübler A (2008b). "An auditory brain-computer interface (BCI)". In: *J Neurosci Methods* 167, pp. 43–50.

Pereira F, Mitchell T, Botvinick M (2009). "Machine learning classifiers and fMRI: A tutorial overview". In: *Neuroimage* 45.1, S199 –S209.

Pfurtscheller G, Allison BZ, Bauernfeind G, Brunner C, Escalante TS, Scherer R, Zander TO, Mueller-Putz G, Neuper C, Birbaumer N (2010). "The hybrid BCI". In: *Front Neuroscience* 4, p. 42.

Picton TW (1992). "The P300 Wave of the Human Event-Related Potential". In: *J Clin Neurophysiol* 9.4, pp. 456–479.

Polich J (1989). "Frequency, intensity, and duration as determinants of P300 from auditory stimuli". In: *J Clin Neurophysiol* 6.3, p. 277.

Polich J (2007). "Updating P300: an integrative theory of P3a and P3b". In: *Clin Neurophysiol* 118, pp. 2128–2148.

Polich J, Ellerson PC, Cohen J (1996). "P300, stimulus intensity, modality, and probability". In: *Int J Psychophysiol* 23, pp. 55–62.

Popescu F, Fazli S, Badower Y, Blankertz B, Müller KR (2007). "Single Trial Classification of Motor Imagination Using 6 Dry EEG Electrodes". In: *PLoS ONE* 2.7.

Pritchard WS, Shappell SA, Brandt ME (1991). "Psychophysiology of N200/N400: A Review and Classification Scheme". In: *Adv Psychophysiol* 4, pp. 43–106.

Quek M, Höhne J, Murray-Smith R, Tangermann M (2012). "Designing future BCIs: Beyond the bit rate". In: *Towards Practical Brain-Computer Interfaces*. Ed. by BZ Allison, S Dunne, R Leeb, J del R. Millán, and A Nijholt. Berlin Heidelberg: Springer, pp. 173–196.

R. Millán J et al. (2010). "Combining Brain-Computer Interfaces and Assistive Technologies: State-of-the-Art and Challenges". In: *Frontiers in Neuroprosthetics* 4.

Ramoser H, Müller-Gerking J, Pfurtscheller G (2000). *Optimal spatial filtering of single trial EEG during imagined hand movement*. version of 1998.

Riccio A, Mattia D, Simione L, Olivetti M, Cincotti F (2012). "Eye Gaze Independent Brain Computer Interfaces for Communication". In: *J Neural Eng* 9, p. 045001.

Röttger S, Schröger E, Grube M, Grimm S, Rübsamen R (2007). "Mismatch negativity on the cone of confusion". In: *Neurosci Lett* 414.2, pp. 178 –182.

Sajda P, Gerson A, Müller KR, Blankertz B, Parra L (2003). "A Data Analysis Competition to Evaluate Machine Learning Algorithms for use in Brain-Computer Interfaces". In: *IEEE Trans Neural Syst Rehabil Eng* 11.2, pp. 184–185.

Samek W, Vidaurre C, Müller KR, Kawanabe M (2012). "Stationary Common Spatial Patterns for Brain-Computer Interfacing". In: *Journal of Neural Engineering* 9.2, p. 026013.

Samek W, Kawanabe M, Müller KR (2014). "Divergence-based Framework for Common Spatial Patterns Algorithms". In: *Biomedical Engineering, IEEE Reviews in* 7, pp. 50–72.

Sannelli C, Vidaurre C, Müller KR, Blankertz B (2011). "Common Spatial Pattern Patches - an Optimized Filter Ensemble for Adaptive Brain-Computer Interfaces". In: *J Neural Eng* 8.2, 025012 (7pp).

Schaefer RS, Vlek RJ, Desain P (2010). "Decomposing rhythm processing: electroencephalography of perceived and self-imposed rhythmic patterns". In: *Psychol Res*. in press.

Schaeff S, Treder MS, Venthur B, Blankertz B (2012). "Exploring motion VEPs for gaze-independent communication". In: *J Neural Eng* 9.4, p. 045006.

Schlögl A, Kronegg J, Huggins J, Mason SG (2007). "Evaluation Criteria for BCI Research". In: *Towards Brain-Computer Interfacing*. Ed. by G Dornhege, J del R. Millán, T Hinterberger, D McFarland, and KR Müller. Cambridge, MA: MIT press, pp. 297–312.

Schreuder M (2014). "Towards Efficient Auditory BCI Through Optimized Paradigms and Methods". PhD thesis. Berlin Institute of Technology.

Schreuder M, Tangermann M, Blankertz B (2009). "Initial results of a high-speed spatial auditory BCI". In: *Int J Bioelectromagnetism* 11.2, pp. 105–109.

Schreuder M, Blankertz B, Tangermann M (2010). "A New Auditory Multi-class Brain-Computer Interface Paradigm: Spatial Hearing as an Informative Cue". In: *PLoS ONE* 5.4, e9813.

Schreuder M, Rost T, Tangermann M (2011a). "Listen, you are writing! Speeding up online spelling with a dynamic auditory BCI". In: *Front Neuroscience* 5.112.

Schreuder M, Höhne J, Treder MS, Blankertz B, Tangermann M (2011b). "Performance Optimization of ERP-Based BCIs Using Dynamic Stopping". In: *Conf Proc IEEE Eng Med Biol Soc*, pp. 4580–4583.

Schreuder M, Thurlings ME, Brouwer AM, Erp JB, Tangermann M (2012). "Exploring the use of direct feedback in ERP-based BCI". In: *Conf Proc IEEE Eng Med Biol Soc*. Vol. 2012, pp. 6707–6710.

Schreuder M, Höhne J, Blankertz B, Haufe S, Dickhaus T, Tangermann M (2013a). "Optimizing ERP Based BCI - a Systematic Evaluation of Dynamic Stopping Methods". In: *J Neural Eng* 10.3, p. 036025.

Schreuder M, Riccio A, Risetti M, Dähne S, Ramsey A, Williamson J, Mattia D, Tangermann M (2013b). "User-Centered Design in BCI - a Case Study". In: *Artificial Intelligence in Medicine* 59.2, pp. 71–80.

Sellers EW, Donchin E (2006). "A P300-based brain-computer interface: initial tests by ALS patients". In: *Clin Neurophysiol* 117, pp. 538–548.

Sellers EW (2013). "New horizons in brain-computer interface research". In: *Clin Neurophysiol* 124, pp. 2–4.

Sellers E, Krusienski D, McFarland D, Vaughan T, Wolpaw J (2006a). "A P300 event-related potential brain-computer interface (BCI): the effects of matrix size and inter stimulus interval on performance". In: *Biol Psychol* 73, pp. 242–252.

Sellers E, Kübler A, Donchin E (2006b). "Brain-computer interface research at the University of South Florida Cognitive Psychophysiology Laboratory: the P300 Speller". In: *IEEE Trans Neural Syst Rehabil Eng* 14, pp. 221–224.

Shannon CE (1949). "Communication in the presence of noise". In: *Proceedings of the IRE* 37.1, pp. 10–21.

Speier W, Arnold C, Lu J, Taira RK, Pouratian N (2012). "Natural language processing with dynamic classification improves P300 speller accuracy and bit rate". In: *J Neural Eng* 9.1, p. 016004.

Sutton S, Braren M, Zubin J, John E (1965). "Evoked-potential correlates of stimulus uncertainty". In: *Science* 150, pp. 1187–1188.

Tangermann M, Höhne J, Schreuder M, Sagebaum M, Blankertz B, Ramsay A, Murray-Smith R (2011a). "Data Driven Neuroergonomic Optimization of BCI Stimuli". In: *Proc. 5th Int. BCI Conf. Graz*. Graz, pp. 160–163.

Tangermann M, Schreuder M, Dähne S, Höhne J, Regler S, Ramsay A, Quek M, Williamson J, Murray-Smith R (2011b). "Optimized Stimulation Events for a Visual ERP BCI". In: *Int J Bioelectromagnetism* 13.3, pp. 119–120.

Tangermann M, Höhne J, Stecher H, Schreuder M (2012a). "No Surprise — Fixed Sequence Event-Related Potentials for Brain-Computer Interfaces". In: *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, pp. 2501–2504.

Tangermann M et al. (2012b). "Review of the BCI Competition IV". In: *Front Neuroscience* 6.55.

Tomioka R, Müller KR (2010). "A regularized discriminative framework for EEG analysis with application to brain-computer interface". In: *Neuroimage* 49 (1), pp. 415–432.

Townsend G, Shanahan J, Ryan DB, Sellers EW (2012). "A general P300 brain-computer interface presentation paradigm based on performance guided constraints". In: *Neurosci Lett*.

Treder MS, Purwins H, Miklody D, Sturm I, Blankertz B (2014). "Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification". In: *J Neural Eng* 11, p. 026009.

Treder MS, Blankertz B (2010). "(C)overt attention and visual speller design in an ERP-based brain-computer interface". In: *Behav Brain Funct* 6, p. 28.

Treder MS, Schmidt NM, Blankertz B (2011). "Gaze-independent brain-computer interfaces based on covert attention and feature attention". In: *J Neural Eng* 8.6, p. 066003.

Vapnik V (1995). *The nature of statistical learning theory*. New York: Springer Verlag.

Venthur B, Scholler S, Williamson J, Dähne S, Treder MS, Kramarek MT, Müller KR, Blankertz B (2010). "Pyff – A Pythonic Framework for Feedback Applications and Stimulus Presentation in Neuroscience". In: *Front Neuroscience* 4, p. 179.

Venthur B, Dähne S, Höhne J, Heller H, Blankertz B (2014). "Wyrm: A Brain-Computer Interface Toolbox in Python". In: *Neuroinformatics* in review.

Vidaurre C, Kawanabe M, Bünau P, Blankertz B, Müller KR (2011a). "Toward Unsupervised Adaptation of LDA for Brain-Computer Interfaces". In: *IEEE Trans Biomed Eng* 58.3, pp. 587 –597.

Vidaurre C, Sannelli C, Müller KR, Blankertz B (2011b). "Co-adaptive calibration to improve BCI efficiency". In: *J Neural Eng* 8.2, 025009 (8pp).

Vidaurre C, Sannelli C, Müller KR, Blankertz B (2011c). "Machine-Learning Based Co-adaptive Calibration". In: *Neural Comput* 23.3, pp. 791–816.

Volosyak I, Valbuena D, Malechka T, Peuscher J, Gräser A (2010). "Brain–computer interface using water-based electrodes". In: *J Neural Eng* 7.6, p. 066007.

Waal M, Severens M, Geuze J, Desain P (2012). "Introducing the tactile speller: an ERP-based brain–computer interface for communication". In: *J Neural Eng* 9.4, p. 045002.

Wills S, MacKay D (2006). "DASHER–an efficient writing system for brain-computer interfaces?" In: *IEEE Trans Neural Syst Rehabil Eng* 14, pp. 244–246.

Winkler I, Haufe S, Tangermann M (2011). "Automatic Classification of Artifactual ICA-Components for Artifact Removal in EEG Signals". In: *Behav Brain Funct* 7.1, p. 30.

Wolpaw, JR and EW Wolpaw, eds. (2012). *Brain-computer interfaces : principles and practice*. ISBN-13: 978-0195388855. Oxford University press.

Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM (2002a). "Brain-computer interfaces for communication and control". In: *Clin Neurophysiol* 113.6, pp. 767–791.

Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM (2002b). "Brain-computer interfaces for communication and control". In: *Clin Neurophysiol* 113, pp. 767–791.

Wolpaw J (2007). "Brain-computer interfaces as new brain output pathways". In: *J Physiol* 579, pp. 613–619.

Xu H, Zhang D, Ouyang M, Hong B (2013). "Employing an active mental task to enhance the performance of auditory attention-based brain–computer interfaces". In: *Clin Neurophysiol* 124.1. epub was in 2012, thus the abbreviation code XuZhsaOuyHon12 instead of XuZhsaOuyHon13, pp. 83–90.

Zickler C, Riccio A, Leotta F, Hillian-Tress S, Halder S, Holz E, Staiger-Sälzer P, Hoogerwerf E, Desideri L, Mattia D, Kübler A (2011). "A Brain-Computer Interface as Input Channel for a Standard Assistive Technology Software". In: *Clinical EEG and Neuroscience* 24.4, p. 222.

Ziehe A, Müller KR (1998). "TDSEP – an efficient algorithm for blind separation using time structure". In: *Proc. of the 8th International Conference on Artificial Neural Networks, ICANN'98*. Ed. by L Niklasson, M Bodén, and T Ziemke. Perspectives in Neural Computing. Berlin: Springer Verlag, pp. 675 –680.
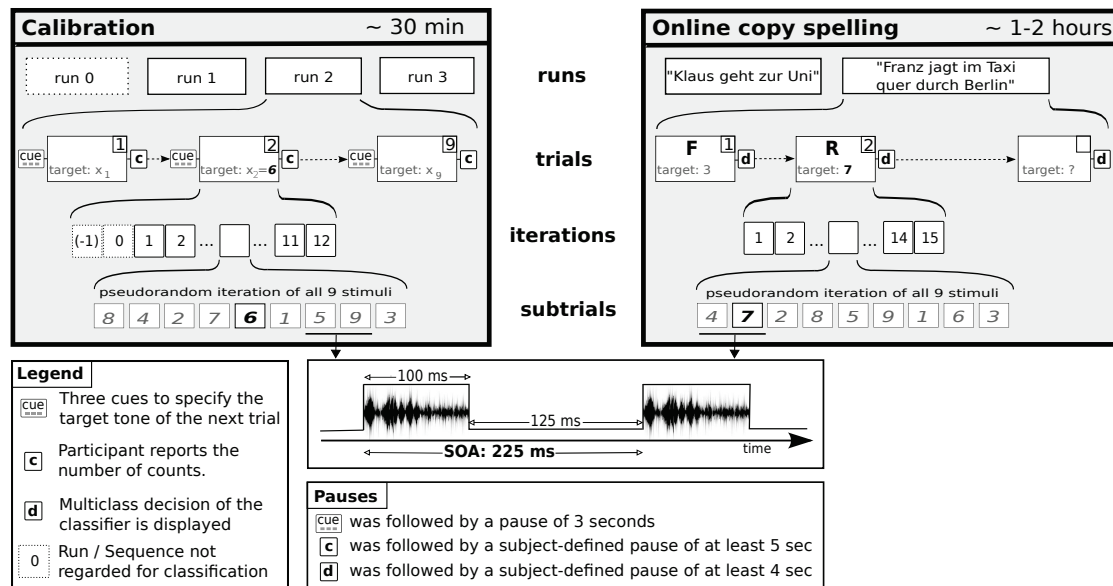
List of Figures

LIST OF TABLES

INDEX

APPENDIX

A.1 Supplementary Material to Experiment 1 (the PASS2D Study)

A.1.1 Study Design.

Figure A.1: *Visualization of the experimental design of the PASS2D experiment.*

A.1.2 Subject-specific Data and Spelling Performance for each Subject.

Table A.1: *Subject-specific data and different measures of spelling performance. Results based on offline data have a gray background and binary classification accuracy is given in classwised balanced accuracy. Each subject spelled the same two sentences: a short (18 char) sentence [a] and a long (36 char) sentence [b]. Spelling performance is quantified by the time [min], required to spell the complete sentence. Individual differences arise due to false selections. For participants VPnv VPoc and VPoe, the second sentence was canceled or not even started. One spelling run [x] was stopped after the 45th trial.*

	VPnv	VPnw	VPnx	VPny	VPnz	VPmg	VPoa	VPob	VPoc	VPod	VPja	VPoe	Avg
sex	m	m	w	w	m	m	m	m	m	m	m	w	
age	26	21	25	23	34	23	23	24	24	25	29	24	25.1
binary cl. acc.	77.0	74.8	52.7	75.0	74.4	60.2	72.2	78.6	79.6	70.0	82.0	73.2	72.5
|diff| in counts	23	13	18	35	5	33	16	6	14	10	1	26	16.7
# selections[a]	29	31		23	29		38	26		28	23		28.4
# false sel.[a]	2	4		0	2		5	1		3	0		2.1
time (min) [a]	25.1	23.0		15.4	21.7		26.9	18.1		19.1	17.9		20.9
# selections[b]		63		53	97		51	49	45[x]	61	49	48	57.3
# false sel.[b]		7		4	27		5	3	1[x]	8	3	4	6.9
time (min) [b]		47.1		36.5	76.7		36.8	39.5	30.9[x]	48.9	36.2	38.6	43.5

A.1.3 Behavioral Data

Table A.1 shows the results of the counting task during the calibration phase. Row '|diff| in counts' contains the sum of the absolute differences between the correct and the reported number of target presentations for each trial. A variation ranging from one to 35 was observed, which also indicates a varying ability across subjects to discriminate between the stimuli. It can be seen that this behavioral data is not directly linked to the spelling performance: some subjects (*VPny*, *VPoe*) have poor behavioral results - i.e. inaccurate counting - but perform well in the online spelling. On the other hand, those subjects with a bad spelling performance didn't have particularly bad behavioral results (see *VPmg* and *VPnx*). Based on these results, the behavioral data alone could not be used as a predictor for the spelling performance in the online phase.

A.1.4 Confusion Matrix for Multiclass Selections.

Table A.2: *Confusion matrix for multiclass selection. The diagonal elements (correct decisions) are marked in bold. Column 'Acc' provides the specific accuracy for each key.*

Target	Selected									Acc
	1	2	3	4	5	6	7	8	9	
1	**137**	13	3	0	0	1	1	0	0	0.89
2	1	**55**	1	0	0	0	0	6	1	0.86
3	2	4	**84**	0	0	0	1	1	2	0.89
4	3	0	1	**153**	5	2	1	1	1	0.92
5	0	0	3	0	**44**	2	2	1	2	0.82
6	0	0	2	1	4	**35**	0	0	1	0.81
7	0	0	0	0	0	0	**61**	2	2	0.94
8	0	0	0	1	0	0	0	**69**	3	0.95
9	0	0	0	0	0	0	0	2	**26**	0.93
										0.89

A.2 Supplementary Material to Experiment 3 (CharStreamer)

Temporal Distribution of class-discriminative Information
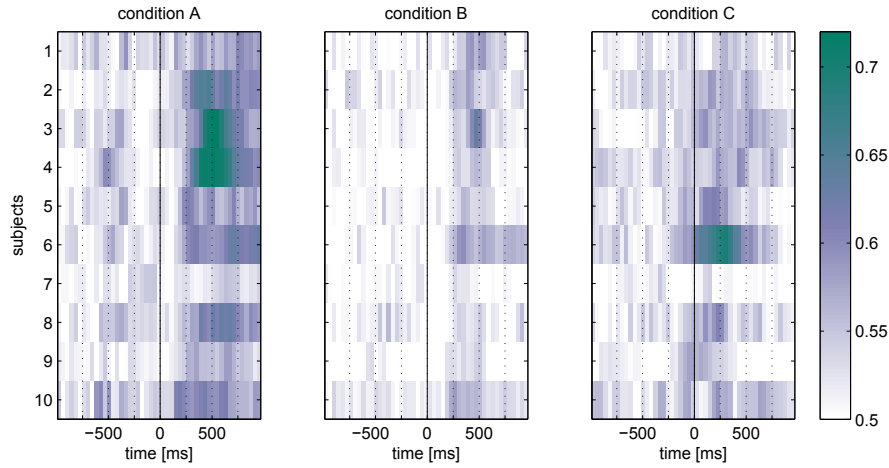
Figure A.2: *Comparison of class-discriminative information contained in the time structure of one epoch. For each subject and condition, one row depicts the estimated sliding binary classification accuracy (SBCA) of a window of 60 ms width. It was estimated by cross-validation with a 0.5 chance-level.*

A.2.1 ERP Responses of individual Subjects
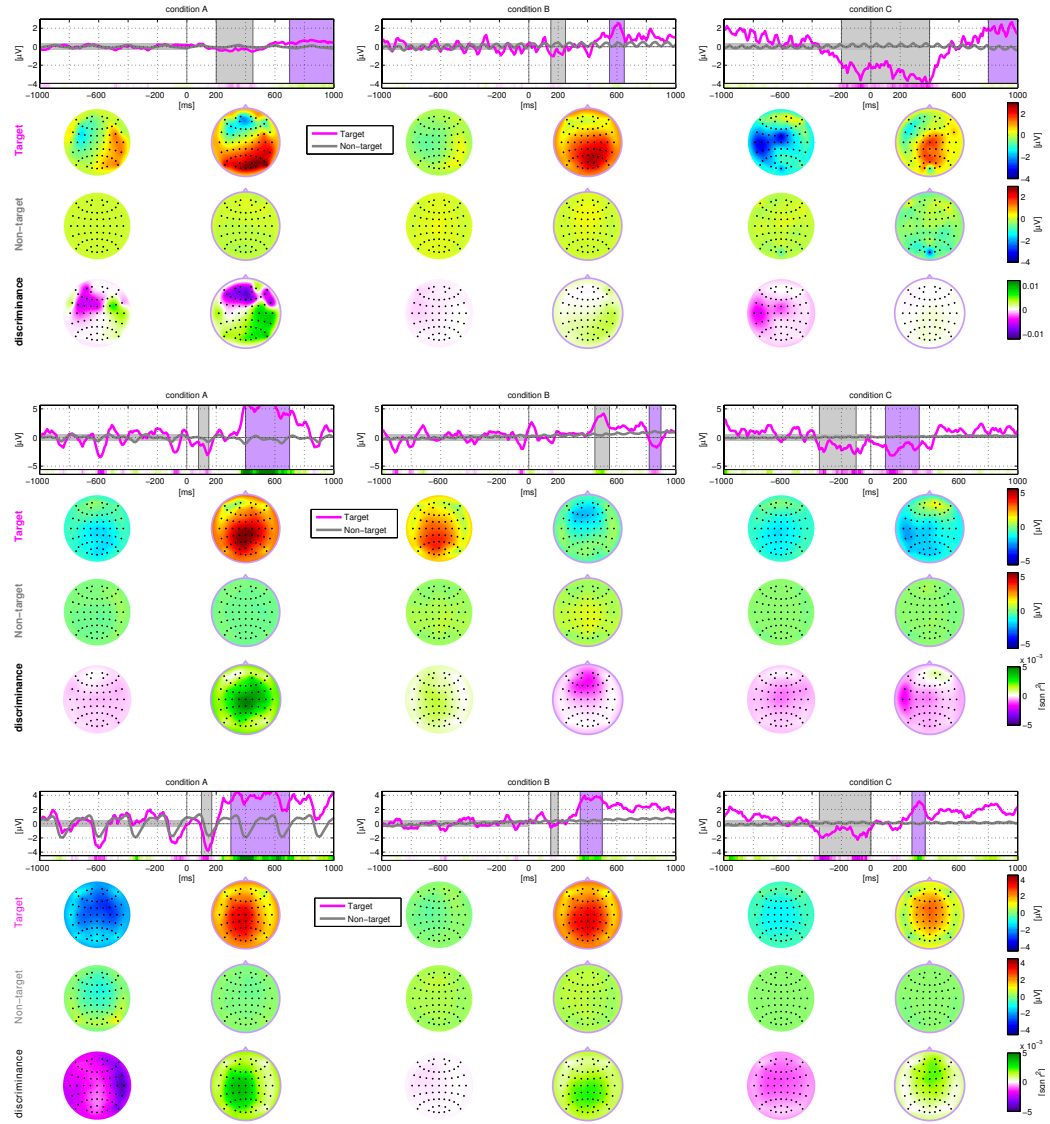
Figure A.3: *ERPs for all three conditions for subject 6, 8 and 10.*

A.3 Supplementary Material to Chapter 4

A.3.1 Classification Accuracy for each Subject and Method

Table A.3: *Binary classification accuracy (estimated by cross-validation) for each subject and classification method.*

Subject	Relevance Subclass LDA	STS Subclass LDA	Sample Subclass LDA	Global LDA	xvalSTS
AMUSE sbj01	87.46	87.42	82.69	87.05	87.10
AMUSE sbj02	76.67	76.45	72.14	74.77	76.13
AMUSE sbj03	67.65	67.49	65.90	64.63	67.35
AMUSE sbj04	81.83	81.75	76.90	81.18	81.50
AMUSE sbj05	90.26	90.10	85.24	89.58	89.76
AMUSE sbj06	94.89	94.91	92.35	94.78	94.79
AMUSE sbj07	79.52	79.31	75.20	78.01	79.20
AMUSE sbj08	88.20	88.08	82.35	88.02	87.63
AMUSE sbj09	93.91	93.94	90.75	93.57	93.72
AMUSE sbj10	75.37	75.56	70.08	74.96	75.27
AMUSE sbj11	89.77	89.72	85.15	89.08	89.31
AMUSE sbj12	86.01	85.88	80.25	85.73	85.53
AMUSE sbj13	89.40	89.17	84.09	89.08	88.84
AMUSE sbj14	81.43	81.30	74.68	80.80	80.80
AMUSE sbj15	88.58	88.38	85.51	87.92	88.20
AMUSE sbj16	70.67	70.65	66.66	69.83	70.61
AMUSE sbj17	82.88	82.69	79.53	80.81	82.42
AMUSE sbj18	71.94	71.72	68.67	70.61	72.05
AMUSE sbj19	85.99	85.00	84.58	80.58	85.69
AMUSE sbj20	71.06	70.39	70.90	66.11	71.57
AMUSE sbj21	84.08	83.68	77.08	83.49	83.45
PASS2D sbj01	86.43	86.37	77.52	86.34	86.00
PASS2D sbj02	80.91	81.20	69.60	81.33	79.48
PASS2D sbj03	60.48	59.37	59.13	58.37	60.31
PASS2D sbj04	79.49	79.41	68.84	79.88	78.50
PASS2D sbj05	83.47	83.72	71.26	83.77	82.43
PASS2D sbj06	65.94	66.24	60.12	66.60	66.32
PASS2D sbj07	82.29	82.42	70.76	82.40	81.37
PASS2D sbj08	87.72	87.79	75.80	88.01	86.20
PASS2D sbj09	90.54	90.53	81.63	90.36	89.64
PASS2D sbj10	68.61	68.56	61.42	68.89	65.91
PASS2D sbj11	90.83	91.11	79.89	90.97	89.59
PASS2D sbj12	85.15	85.01	76.40	85.03	84.34
CenterSpeller sbj01	95.22	95.18	90.63	94.73	94.29
CenterSpeller sbj02	89.00	89.08	82.78	88.92	88.07
CenterSpeller sbj03	89.41	89.26	84.00	88.90	88.70
CenterSpeller sbj04	93.00	93.05	89.00	92.46	92.43
CenterSpeller sbj05	89.40	89.38	83.57	89.56	88.96
CenterSpeller sbj06	88.39	88.33	84.60	87.58	88.19
CenterSpeller sbj07	96.48	96.45	92.41	96.34	96.08
CenterSpeller sbj08	95.42	95.37	93.44	94.95	95.17
CenterSpeller sbj09	93.65	93.75	89.27	93.09	93.34
CenterSpeller sbj10	97.66	97.63	95.82	96.95	97.40
CenterSpeller sbj11	83.90	83.95	79.28	83.02	82.82
CenterSpeller sbj12	93.03	92.95	88.63	92.47	92.48
CenterSpeller sbj13	96.82	96.88	93.69	96.62	96.57
MVEP sbj01	73.01	73.15	70.12	73.22	73.32
MVEP sbj02	81.38	81.75	76.38	81.80	81.47
MVEP sbj03	78.82	78.81	74.04	79.70	79.79
MVEP sbj04	85.89	85.78	81.50	86.07	85.87
MVEP sbj05	88.41	87.94	85.37	87.27	88.16
MVEP sbj06	75.36	74.73	69.66	74.98	74.68

MVEP sbj07	84.18	84.04	79.32	83.29	83.92
MVEP sbj08	84.78	84.82	81.92	83.48	85.28
MVEP sbj09	82.76	82.75	77.00	82.53	82.27
MVEP sbj10	78.32	78.36	72.74	77.62	77.82
MVEP sbj11	83.92	83.76	79.14	83.50	83.66
MVEP sbj12	76.14	76.26	71.45	76.49	75.99
MVEP sbj13	77.51	77.51	74.07	76.60	77.62
MVEP sbj14	79.90	79.14	75.52	79.60	79.42
MVEP sbj15	91.06	91.25	88.15	90.68	91.19
MVEP sbj16	68.37	67.93	64.97	68.50	68.53
RSVP sbj01	87.20	88.12	46.48	88.22	77.80
RSVP sbj02	84.37	84.69	44.55	84.20	73.46
RSVP sbj03	90.50	90.23	53.39	88.83	78.81
RSVP sbj04	90.17	91.48	51.47	91.04	81.83
RSVP sbj05	85.82	87.12	41.84	86.28	74.55
RSVP sbj06	92.67	92.94	56.52	92.39	83.19
RSVP sbj07	84.14	84.84	46.94	84.29	73.06
RSVP sbj08	88.14	88.79	45.43	87.50	76.14
RSVP sbj09	90.28	91.08	51.37	89.93	78.06
RSVP sbj10	85.39	86.19	39.34	85.81	70.58
RSVP sbj11	90.26	90.16	44.21	89.57	76.81
RSVP sbj12	93.81	93.51	58.77	92.52	84.04

# A.4 Supplementary Material to Experiment 5 (the Patient Study)

A.4.1 Investigating the Session-to-Session Transfer

During CopyTask, an adaptive classifier was applied, which was trained on data from preceding sessions. Thus, while each trial was continuously classified, Fig. 5.4 and Fig. 5.5B show the online selection accuracy as a bar for each block. It should be noted that the type of feature used for classification was changed between (sometimes also within) sessions. This happened especially within the first three BCI-sessions, since the experimenters could not be sure which feature would suit best for each patient. Fig. A.4 pictures theses modifications over time, which resembles the closed-loop design cycle following a user-centered design. One example for such a transition can be found in patient 4 between session 2 and session 3: in session 2 (the first feedback session), a classifier in the μ band was used, resulting in a poor BCI control. After reanalyzing the data of session 1 and 2, a new classifier was generated for session 3. The new classifier evaluated an ERD in the beta-band, leading to a considerably increased offline accuracy (estimated with cross-validation) and also the online performance all following sessions was increased considerably.

Fig. A.5 shows depicts the spatial distribution of class discriminative information for each patient across all sessions as scalp maps.
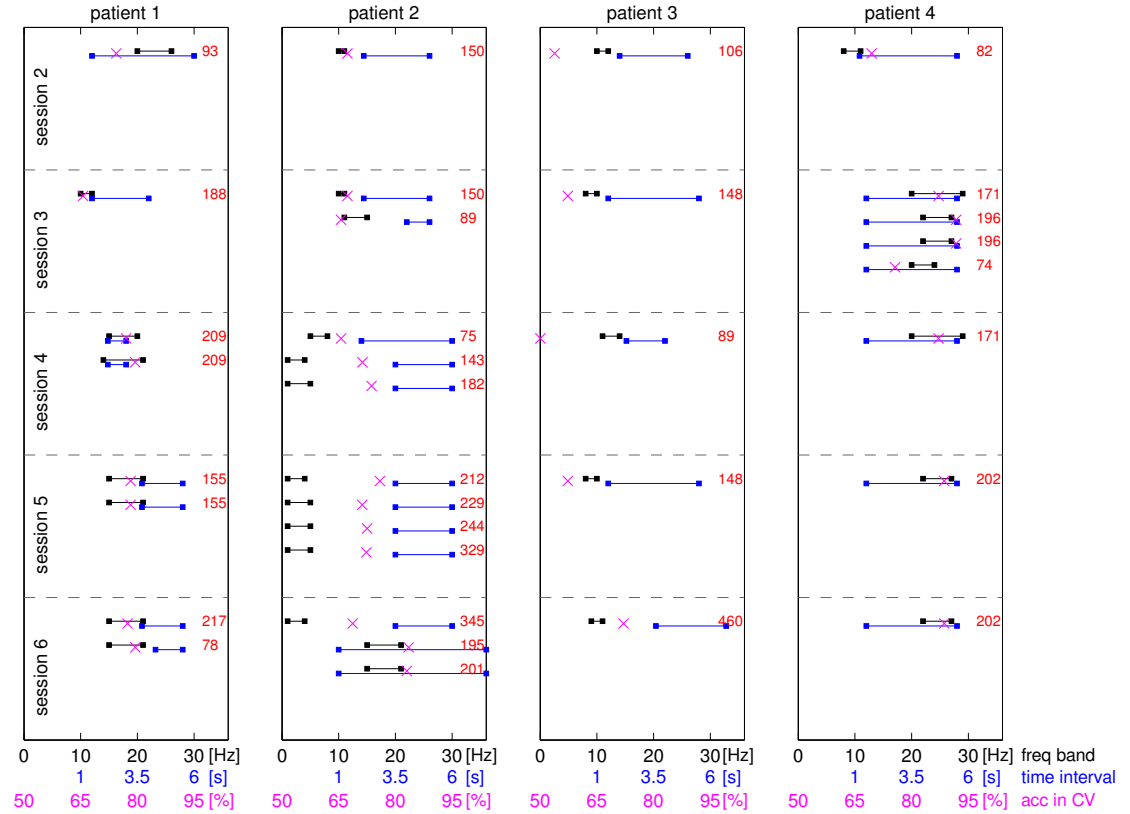
Figure A.4: *Description of the different classifiers used within for online BCI. Across and within sessions, the classifier was retrained on varying subsets of the data and different features. One classifier is described by the set of two neighboring lines (back and blue), a cross in magenta and the number in red. The black lines mark the chosen frequency band, the blue lines mark the time interval used to train and apply the classifier. The cross marks the accuracy of the classifier, estimated with cross-validation on training data. The number in red specifies the number of trails which were used to train the classifier. Note that beginning with the 6th session, the trial length for patient 1 was shortened to 3.5 seconds - resulting in a classification interval after the end of the trail (β rebound). For all other patients the trial length was 5-7 seconds.*
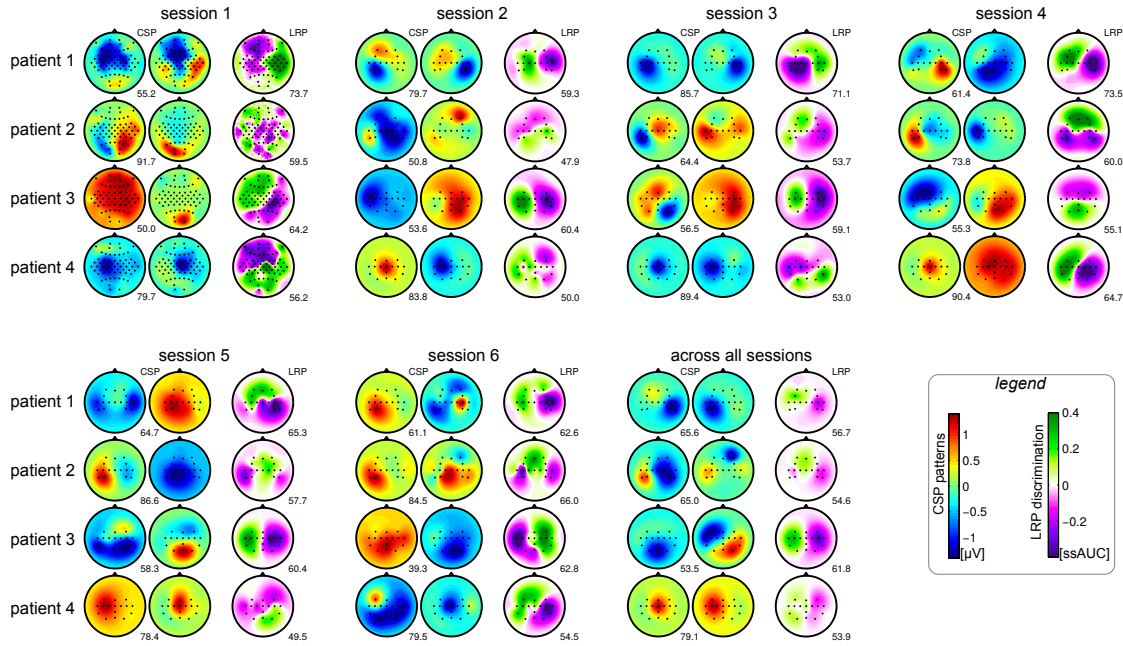
Figure A.5: *Class discriminant information for each patient across sessions. For each session, the spatial pattern of the most (left) and second-most (middle) discriminant CSP filter is depicted. Therefore, the same frequency band as well as the same time intervals were chosen for one subject and all sessions. The same parameters were used to generate Fig. 5.3. The right scalpplot visualizes class discrimination of the LRP feature. The classification accuracy of the spectral (CSP-based) classifier and the LRP classifier is printed next the scalpplots. This classification accuracy is estimated with a 5 fold cross-validation and gives a quantification of how separable the data was in the corresponding session. In the online scenario, a different classifier was used which was trained on more trails from preceding sessions. Note that the sign of the scalpmaps is arbitrary, thus red and blue (as well as their corresponding graduations) are exchangeable. Note that two colorbars (for CSP patterns and LRP discrimination) are given in the legend. The abbreviation "ssAUC" stands for a signed and scaled modification of the area under the curve (AUC), see Section 2.4.5.*

A.4.2 BCI Performance in the FreeMode
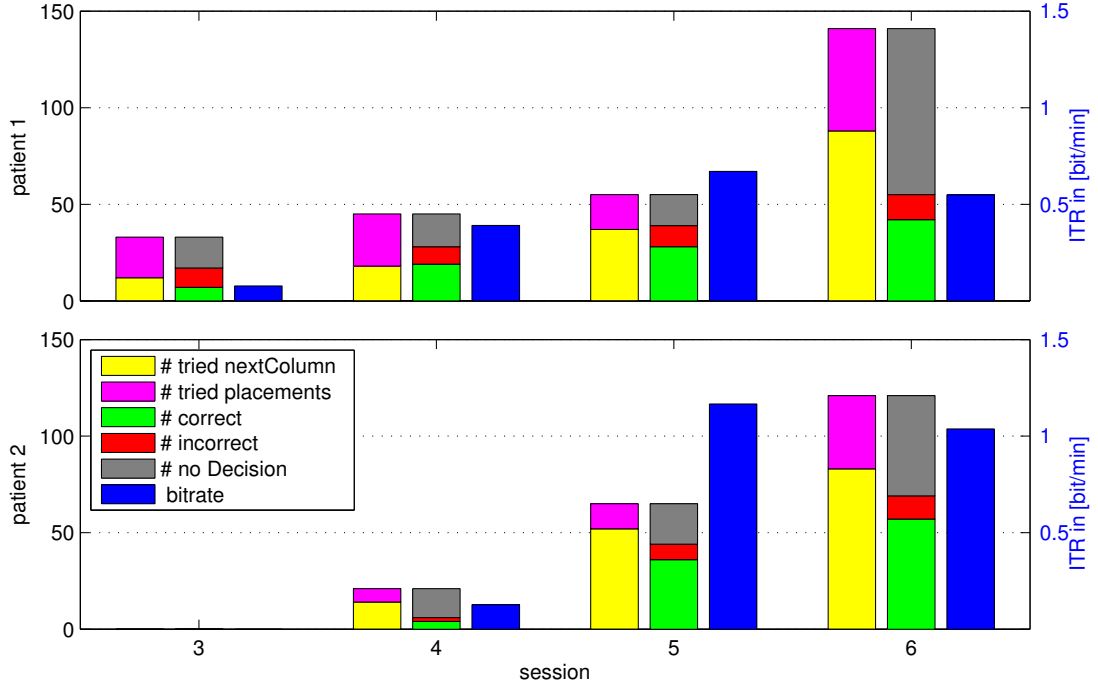
Figure A.6: *BCI performance in the FreeMode. Patient 1 and patient 2 could communicate their intentions with AT. Their comments were used as labels for trials in the FreeMode. Note that the scaling of the bitrate is on the right axis. The patients did not enter the FreeMode in session 3 and session 4.*

A.4.3 Discussion of the Performance of Patient 4 in the FreeMode

Although patient 4 was the best-performing subject in the CopyTask, we were not able to show that patient 4 gained reliable control during the FreeMode. This section discussed some potential reasons in more detail.

Although being unlikely, it may still be possible that patient 4 had (at least partial) control over the BCI and we were simply not able to identify this control during FreeMode. Numerous actions in the FreeMode gaming phase of patient 4 seemed inappropriate to us, but we cannot exclude that he was willing to do these actions. This "identification problem" can only occur in situations, when the BCI user has no other means of reliable communication: the only statement we could assess from patient 4 was, that a number of actions in the FreeMode phase were not intended. We thus assume that the control was insufficient and BCI accuracy dropped with the transition from the CopyTask to the FreeMode.

Fatigue can be expected to accumulate over the course of a session. It should be stressed that for patients in such a severe condition as patient 4, every type of communication might be exhausting. Due to muscle fatigue, conventional AT (button press) may even be more tiring than communication

through BCI. Several short periods of sleep (detected through visual inspection of the patient) had been present even during the online CopyTask blocks, and parts of the FreeMode phase might have taken place while he was not focused or even asleep. In CopyTask blocks, the failure of full experimental blocks could typically be avoided by longer breaks between runs, or by interruption of a block and its repetition at a later moment. As the FreeMode runs were recorded at the end of a session, and as the amount of recorded data was highly limited, the strategy of longer breaks and repetitions would have been neither practical nor effective for the patient during the FreeMode.

The presence of a Bereitschaftspotential (BP) could indicate if the user was not asleep/fatigue, but instead preparing to execute commands. For healthy users, the BP is known to precede self-initiated motor actions (Kornhuber and Deecke, 1965), be it executed or imagined ones. In a post-hoc analysis of the EEG data of patient 4, we thus looked for BP activity. If it was detectable, it could serve to distinguish effortful, but unsuccessful trials from trials, where the user did not even prepare for or even attempt to execute a command. Despite the slow trial timing, the BP was neither present in the EEG recording of the successful CopyTask of patient 4, nor during the FreeMode. Thus, the BP could not provide further insights into the mental state of patient 4 during the FreeMode. Given the currently available data, it seems not possible to validate the hypothesis of attention problems and fatigue in a data-driven approach.

Compared to the CopyTask mode, the mental workload is probably higher in self-initiated gaming and might have been too high for patient 4. To check this, we have tested the FreeMode ability separately during two additional video-taped sessions without EEG. Here the patient played a physical Connect-4 game against a caregiver. While the caregiver would indicate the current column for several ($\sim$ 13) seconds, patient 4 had the option to communicate his intent to place a coin by pressing the button. The analysis of the obtained video material is difficult, as the true labels of intended decisions are again not known. Partially, reaction times were very long (and obvious target columns were missed), partially seemingly good decisions were communicated within 5-10 s. As patient 4 revealed both, reasonable and very unreasonable selections given the current game situation, no clear result could be obtained from the analysis of the video material. But we can state that the physical control (with AT) of the game was on a similar level than the BCI-controlled game in FreeMode.

Patient 4 may have a reduced ability for initiating an action without an external cue or request, while executing a given "command" in the CopyTask did not impose a problem. Self-initiated action is known to be impaired in patients with lesions to the anterior cingulate cortex (Cohen et al., 1999). Problems with self-initiating behavior are supported by his neurologist and by reports from his caregivers. The latter can – at some times – get answers from patient 4, but they report, that he would not start a request or communication himself.

In an alternative theory, Birbaumer et al. (2008) stated that patients in completely locked-in condition may lose the ability of goal directed behavior. This theory is independent of any fronto-lateral lesion but it is based on the lack of communication for a long time. Due to his weak but surely existing residual communication ability, this theory might however not directly be applicable for patient 4.