From Supervised to Unsupervised Machine Learning Methods for Brain-Computer Interfaces and Their Application in Language Rehabilitation

DAVID HÜBNER

Dissertation
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
- Dr. rer. nat -

der Technischen Fakultät der
Albert-Ludwigs-Universität Freiburg im Breisgau

DEKANIN:
Prof. Dr. Hannah Bast

PRÜFUNGSAUSSCHUSS:
Prof. Dr. Bernhard Nebel (Vorsitzender)
Albert-Ludwigs-Universität Freiburg

Dr. Michael Tangermann (Erstgutachter und Betreuer)
Albert-Ludwigs-Universität Freiburg

Prof. Dr. Joschka Bödecker (Zweitgutachter)
Albert-Ludwigs-Universität Freiburg

Prof. Dr. Klaus-Robert Müller (Beisitzender)
Technische Universität Berlin
Korea University, Seoul
MPI für Informatik, Saabrücken

DATUM DER MÜNDLICHEN PRÜFUNG:
03.02.2020

ABSTRACT

The principal idea of brain-computer interfaces (BCIs) is that a decoder translates brain signals into messages or control commands by utilizing machine learning (ML) methods. BCIs hold great promise to improve the living conditions of patients by providing a communication channel that is independent of motor control or by providing feedback about the ongoing brain state that can be used in a training scenario.

Applying this neurotechnology is not without difficulties. A common observation is that brain signals strongly differ across patients, but also vary across different sessions of the same patient or even change within a single session. The reasons range from human factors, e. g., differences or changes in anatomical or functional network structures, to non-human factors such as differences in the measurement environment. While these changes clearly challenge the ML model and require a subject- and session-specific decoder, they can also be partly desired, for instance, in cases where BCI-based feedback is used to trigger targeted neuroplasticity. This thesis addresses both aspects: the quest for learning a good decoder that can cope with changing brain signals and the quest for finding new brain state dependent training protocols that can lead to functional improvements.

In my methodical contributions, I demonstrate how unsupervised ML methods can quickly and reliably learn a good decoder even in the complete absence of labeled data, i. e. data where the user's intentions are unknown. This task is substantially more difficult than the traditional supervised ML where labeled data is collected during a calibration session and then used to associate brain signals to certain tasks. In contrast to supervised learning, unsupervised methods allow for continuous learning, adapting to changes in the data distribution and have the prospect of skipping or shortening the calibration session.

I present a new approach called learning from label proportions (LLP) for BCIs based on event-related potentials (ERPs) where the unsupervised ML model exploits the existence of groups in the data that have different proportions of target and non-target stimuli. For some applications, these groups occur naturally, e. g., when the number of items varies across different selection steps such as in a BCI chess application, while the groups can be created by changing the user interface in other applications. Noticeably, LLP is the first unsupervised method in BCI that is guaranteed to converge to the optimal decoder even if no labels are available. When combined with an expectation-maximization algorithm, the resulting classifier shows remarkable performances. In a visual matrix speller, only 3 minutes of unlabeled electroencephalography (EEG) data were necessary until the unsupervised ML approach has learned a reliable decoder. In addition,

classification performances were almost as good as for a supervised decoder that has full access to the labels. The results also showed that the unsupervised approaches even work well on challenging patient data from an auditory ERP paradigm.

On the application side, I present the first successful BCI-based language training for patients with chronic aphasia. Aphasia refers to a language impairment that frequently occurs after brain strokes. The new training was developed together with the University Medical Center Freiburg. In contrast to previous speech and language therapies, ML methods were used to continuously monitor the ongoing brain state of patients using EEG signals in an auditory ERP paradigm. Based on this information, we continuously provide feedback to the patients that should reinforce beneficial brain states. In a pilot study, 10 stroke patients with chronic aphasia underwent 30 hours of high-intensity BCI-based language training. The results are extremely promising: compared to other therapies, our patients showed large improvements in their verbal abilities, and 5 patients were diagnosed as non-aphasic after the training even though their stroke occurred several months to years before the start of our training.

Taken together, these contributions increase the usability of BCI systems and open the door for a completely new application field of BCIs with an enormous potential user group.

ZUSAMMENFASSUNG

Die Hauptidee von Gehirn-Computer-Schnittstellen (englisch: brain-computer interfaces; BCIs) besteht darin, dass ein Programm Gehirnsignale in Nachrichten oder Steuerbefehle unter Verwendung von Methoden des Maschinellen Lernens (ML) umwandelt. BCIs haben das vielversprechende Potenzial, die Lebensbedingungen von Patienten zu verbessern, indem sie einen von der Motorik unabhängigen Kommunikationskanal bereitstellen oder aber Rückmeldungen über den aktuellen Gehirnzustand liefern, was in einem Trainingsszenario verwendet werden kann.

Die Anwendung dieser Neurotechnologie ist nicht ohne Schwierigkeiten. Eine häufige Beobachtung ist, dass Gehirnsignale nicht nur zwischen Patienten stark variieren, sich aber auch zwischen verschiedenen Sitzungen desselben Patienten unterscheiden oder sich sogar innerhalb einer einzelnen Sitzung ändern können. Die Gründe dafür reichen von menschlichen Faktoren, z. B. Unterschiede oder Änderungen in anatomischen oder funktionellen Netzwerkstrukturen, zu nicht-menschlichen Faktoren wie Unterschiede in der Messumgebung. Während diese Änderungen das ML-Modell eindeutig herausfordern und patienten- und sitzungsspezifische Modellparameter erfordern, können sie teilweise auch erwünscht sein, beispielsweise in Fällen, in denen eine BCI-basierte Rückmeldung verwendet wird, um eine gezielte Neuroplastizität herbeizuführen. In dieser Arbeit werden beide Aspekte behandelt: Die Suche nach ML-Methoden, die mit den sich ändernden Gehirnsignalen zurechtkommen, und die Suche nach neuen, vom Gehirnzustand abhängigen Trainingsprotokollen, die zu Funktionsverbesserungen führen können.

In meinen methodischen Beiträgen zeige ich, wie unüberwachte ML-Methoden schnell und zuverlässig gute Modellparameter erlernen können, selbst bei ungelabelten Daten, bei denen die genauen Absichten der Benutzer unklar sind. Diese Lernaufgabe ist wesentlich schwieriger als herkömmliche überwachte Lernverfahren, bei denen gelabelte Daten während einer Kalibrierungssitzung erfasst und dann verwendet werden, um Gehirnsignale mit bestimmten Tätigkeiten zu assozieren. Im Gegensatz zum überwachten Lernen ermöglichen unüberwachte Methoden die Möglichkeit sich an Änderungen in der Datenverteilung anzupassen, kontinuierlich zu lernen und die Kalibrierung zu überspringen oder zu verkürzen.

In meiner Arbeit zeige ich einen neuen Ansatz namens “Lernen von Label Proportionen” (LLP) für BCIs basierend auf ereigniskorrelierten Potenzialen (ERPs), bei dem das unüberwachte ML-Modell verschiedene Gruppen in den Daten nutzt, die unterschiedliche Anteile von Ziel- und Nichtzielstimuli aufweisen. Bei einigen Anwendungen treten diese Gruppen auf natürliche Weise auf, z. B. wenn die Anzahl der Elemente

in verschiedenen Auswahlschritten variiert, wie beispielsweise in einer BCI-Schachanwendung. In anderen Anwendungen können die Gruppen durch das Anpassen der Benutzeroberfläche erzeugt werden. Theoretische Betrachtungen zeigen, dass LLP die erste unüberwachte Methode für BCIs ist, die garantiert zu den optimalen Modellparametern konvergiert, auch wenn keine Labels verfügbar sind. In Kombination mit einem Expectation-Maximization-Algorithmus zeigt das resultierende ML-Modell bemerkenswerte Leistungen. In einem visuellen Matrix-Speller waren nur 3 Minuten ungelabelte Elektroenzephalographie (EEG)-Daten erforderlich, bis der unüberwachte ML-Ansatz verlässliche Parameter gefunden hatte. Außerdem waren die Klassifizierungsergebnisse fast so gut wie für einen überwachten Algorithmus, der vollen Zugriff auf die Labels hatte. Die Ergebnisse zeigten auch, dass die unüberwachten Ansätze sogar in der Lage sind, schwierige Patientendaten aus einem auditorischen ERP-Paradigma gut zu klassifizieren.

Auf der Anwendungsseite präsentiere ich das erste erfolgreiche BCI-basierte Sprachtraining für Patienten mit chronischer Aphasie. Aphasie bezeichnet eine Beeinträchtigung der Sprache, die häufig nach einem Schlaganfall auftritt. Das Training wurde gemeinsam mit dem Universitätsklinikum Freiburg entwickelt. Im Gegensatz zu früheren Therapien wurden ML-Methoden verwendet, um den fortlaufenden Gehirnzustand von Patienten mithilfe von EEG-Signalen in einem auditorischen ERP-Paradigma kontinuierlich zu analysieren. Basierend auf diesen Informationen geben wir den Patienten fortlaufend Rückmeldungen, die es den Patienten ermöglichen, Gehirnzustände herbeizuführen, die sich positiv auf die Sprache auswirken. In einer Pilotstudie absolvierten 10 Patienten mit einer chronischen Sprachstörung nach einem Schlaganfall 30 Stunden lang ein BCI-basiertes Sprachtraining mit hoher Trainingsintensität. Die Ergebnisse sind äußerst vielversprechend: Im Vergleich zu anderen Therapien zeigten unsere Patienten eine sehr deutliche Verbesserung ihrer verbalen Fähigkeiten und 5 Patienten wurden nach dem Training als nicht-aphasisch diagnostiziert, obwohl ihr Schlaganfall schon mehrere Monate bis Jahre vor dem Trainingsbeginn lag.

Zusammengenommen erhöhen diese Beiträge die Benutzerfreundlichkeit und Anwendbarkeit von BCI-Systemen und öffnen die Tür für ein völlig neues Anwendungsfeld von BCIs mit einer großen potenziellen Zielgruppe.

ACKNOWLEDGMENTS

Even though a PhD thesis is sometimes regarded as this long period of solitary thinking in a dark lab room, I can wholeheartedly say that this dissertation would not have been possible without all the collaboration partners and the great support that I have received.

My foremost gratitude goes to my PhD supervisor Michael Tangermann, who did not only introduce me to the exciting research field of BCIs, but who also gave me the opportunity to freely explore my own research interests. I am especially thankful that your door was never shut and that you have spent so much of your valuable time for highly educational discussions and interesting joint workshops and presentations. Learning would have been much more difficult without the labels that you have provided.

I also want to kindly thank Prof. Joschka Bödecker for co-supervising me and for your highly encouraging feedback during our meetings. My gratitude also goes to Prof. Bernhard Nebel and Prof. Klaus-Robert Müller for agreeing on being part of my thesis committee.

The close collaborations with the groups at TU Berlin and TU Ghent were an integral building block for this thesis. Thank you Pieter-Jan Kindermans for your unmatchable ability to generate and communicate ideas. Your dedication is contagious and very inspiring. I would also like to thank Thibault Verhoeven for the highly efficient and enjoyable collaboration.

For the last three and a half years, the brain state decoding lab has not only been a very productive place for me, but also a place where I met many good friends. I would like to thank my fellow PhD students – Andreas, Henrich, Jan, Sebastián – for your constant support, great ideas, intriguing discussions, valuable feedback and the great time before, during and after conferences. You have taught me a lot and you were the key to make the PhD journey a fantastic experience. I would also like to thank all the students that have helped with experiments, software development and data analysis. Special thanks to Simon for always having a helping hand and to Oleksii, who went far beyond what I would expect from a Master's student. I would also like to thank Konstantin for helping me with the first steps in the lab and for being a good friend. My deep appreciation also goes to Albrecht, Anatolii, Atieh, Eva, Natalija, Lala, Laya, Robin, Sarah and Simone for their reliable and enjoyable support.

This thesis would have been impossible without all the subjects that participated in the experiments. I would like to particularly thank all the aphasia patients who displayed a remarkable optimism despite their past experiences. My gratitude also goes to Mariachristina Musso for the close

clinical collaboration and for allowing me to work in this interdiscplinary field between algorithms, measurements and clinical applications.

I also thankfully acknowledge the financial support by BrainLinks-BrainTools Cluster of Excellence funded by the German Research Foundation (DFG), grant number EXC1086, and the extraordinary support by the z-office team. I also acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 39/963-1 FUGG.

I am convinced that our families shape us more than we realize. I want to use this opportunity to explicitly thank my parents, Karin and Ingolf, for being the best possible role model and for your unlimited trust in me. Thank you for endlessly supporting me on every step I took even if my steps have been far away from home. I also want to thank my big brother Simon for being my patient teacher since my childhood.

Finally, I am deeply grateful that coming to Freiburg has allowed me to meet my sunshine Stefanie. The last three years with you have been wonderful and I am looking forward to a bright future with you.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF NOTATIONS

Please note the following naming conventions throughout the thesis. Matrices will always be denoted in bold upper case letters, e. g., $\mathbf{A}$, whereas vectors are denoted in bold lower-case letters, e. g., $\mathbf{w}$. Scalars will be denoted as non-bold characters which can be in lower or upper case, e. g., N or λ.

Mathematical notation:

$(\cdot)^T$	Transpose of a vector or matrix. 19
$\boldsymbol{\mu} \in \mathbb{R}^D$	Class means 20
$\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$	Covariance matrix 20
$\mathbf{w} \in \mathbb{R}^D$	Projection vector 19
$\mathbf{x} \in \mathbb{R}^D$	Feature vector of a single data point 19
$b \in \mathbb{R}$	Bias term 19

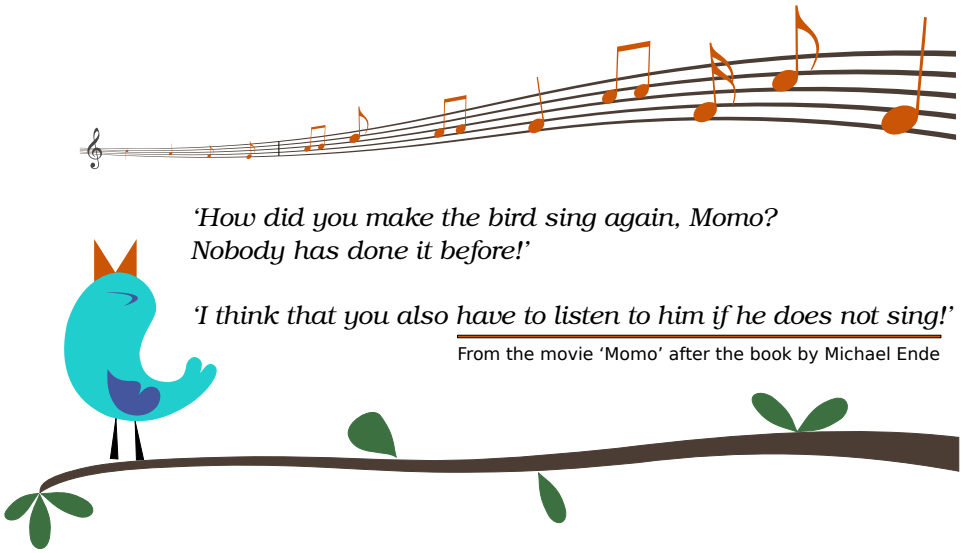
Abbreviations and other terms:

AAT	Aachener Aphasie Test. 90
ALS	Amyotrophic lateral sclerosis 12
AMUSE	Auditory multi-class spatial ERP 88
AUC	Area under the curve 30
BCI	Brain-computer interface 11
CAL	Communication activity log 90
EEG	Electroencephalography 13
EM	Expectation-maximization 43
Epoch	A time-window that is segmented with respect to an event . . . 18
ERP	Event-related potential 15
ICA	Independent component analysis 18
LDA	Linear discriminant analysis 19
LLP	Learning from label proportions 45
MARA	Multiple artifact rejection algorithm 18
MIX	A combination of the LLP and EM mean estimations. 56
ROC	Receiver-operator characteristic 30
SNR	Signal-to-noise ratio 13
SOA	Stimulus onset asynchrony 15
TAP	Test of attentional performance 90
Target	The stimulus that was attended by the user. 15
Trial	The combination of several consecutive single events until the brain-computer interface performs an action. 29

MAIN PART

1

PREFACE



'How did you make the bird sing again, Momo?
Nobody has done it before!'

'I think that you also have to listen to him if he does not sing!'

From the movie 'Momo' after the book by Michael Ende

We live in remarkable times. Most people interact with a computer or smartphone on a daily basis. Using an input device such as a keyboard, mouse or touchpad has become natural for most of us and even children can manage to interact with electronic devices with remarkable speed and precision. Now, what if someone loses the physical ability to speak, point or type due to a disease or injury. Then, suddenly, the person will not only be having difficulties using a computer, but — even worse — sometimes also loses the ability to communicate at all.

To find strategies to overcome this loss, I want you to briefly think about the quote from the beginning of the chapter. Reading this quote, you may ask: "*How can we listen to someone that does not produce any sounds?*". Well, what if we start one step earlier and try to capture the *intention* to produce sounds or to communicate. Maybe brain activity can tell us something about the intended actions although no visible activity has taken place. Then, once we were able to capture a subjects intention, we can translate it into sounds or actions.

Indeed, the idea to communicate via brain signals is not new and its first successful implementation dates back exactly 31 years to Farwell and Donchin [50]. In their ground-breaking work, users could communicate without physical activity in the following way. A computer screen showed all letters from the alphabet in a grid. The user had to concentrate on the letter that they would like to spell. Then, rows and columns of letters were alternatively highlighted by increasing their brightness. Simultaneously,

brain signals were read out using electrodes placed on the head. By comparing a few seconds of recorded data to existing training data of the subject, the computer is then able to infer which letter was attended and displays it on the screen. This enables patients to have a very limited, but extremely useful basic communication, e. g., to express their desires to caretakers or to answer basic questions. It is an example of a brain-computer interface (BCI). A more detailed introduction to BCIs will be given in Chapter 2.

While this first approach enables basic communication, patients will not recover from their disease by using such a BCI. In contrast, the initial quote could also have another interpretation. In the second version, Momo *curated* the bird by carefully paying attention to the bird while it is unable to speak. If we translate this idea to BCIs, then the BCI "listens" to patients by continuously analyzing their brain activity. Once the BCI detects an activation pattern that indicates successful language processing, then the BCI will inform the patient about that state. Otherwise, the patient needs to search for alternative strategies to activate his language network. With that, patients may be able to improve their verbal abilities by repeatedly practicing with a BCI.

Authors contributions

This thesis contributes to both interpretations of the quote, i. e. I have made contributions that are beneficial for general BCIs (e. g., used for communication) and I have substantially contributed to the realization of the first BCI-based language training.

On the theoretical side, a general problem that is encountered in BCIs is that labeled training data needs to be available to calibrate the system to each individual user. Calibration is a process where the user performed a predefined task, e. g., is instructed to always look at a specific letter or to listen to a certain sound. With that, example data is collected where the user's intentions are known (labeled data). A supervised machine learning model can then learn to associate the recorded brain signals with the executed actions which can be used to classify unseen new brain signals. Because brain signals highly vary between users and even differ between different sessions of the same user, frequent (re-)calibrations are normally necessary. This calibration time is effectively lost for the user because no useful outputs can be generated during that period. In addition, brain activity might change between this calibration phase and the real application phase [156] and the data distribution might also change over the course of a session due to human factors (e. g., fatigue, change in motivation, learning) or non-human factors (e. g., changes in the environment).

To overcome these obstacles, it is essential that the decoder is able to learn during the actual usage of the BCI. This is a very challenging task because the user's intentions are unknown in that stage. Hence, the algorithm would need to be able to learn from unlabeled data. This is called **unsupervised learning** and was identified as one of the "*key challenges for BCI deployment*

outside the lab" [129]. Lotte et al. [117] point out that "*there is a need for more robust unsupervised adaptation methods, as the majority of actual BCI applications do not provide labels, and thus can only rely on unsupervised methods*". Millán and colleagues emphasize the importance of unsupervised adaptation and learning for skill-learning in BCIs, as it "*increases the likelihood of providing stable feedback to the user, a necessary condition for people to learn to modulate their brain activity*" [124].

In the thesis, I address this need by introducing and evaluating two novel unsupervised learning methods. The basic idea underlying these approaches is to exploit the relationship between the user interface and the machine learning model. More specifically, the way the user interacts with the computer exerts influence on how the recorded data is structured. In Chapter 3, I demonstrate how the paradigm can be adjusted to meet the requirements of a machine learning approach called learning from label proportions and how this classifier can be combined with an expectation-maximization algorithm. Results from healthy subjects show that the unsupervised learning methods can utilize unlabeled data almost as efficiently as labeled data and can — in practice — replace supervised methods. For a visual speller based an electroencephalography (EEG), an average time of 3 minutes of unsupervised calibration was enough to reach very reliable control. I show that this algorithm can also be naturally applied to other interfaces such as a brain-controlled chess interface. In addition, results from simulations showed that unsupervised learning methods can even work on stroke patient data from an auditory experiment which is generally much more difficult to decode.

On the practical side, this is the first work that shows that BCIs can be used for **language rehabilitation** after a brain stroke. Together with colleagues from the University Medical Center Freiburg, we designed and evaluated the first successful BCI-based language therapy for aphasic patients. Aphasia refers to an impairment of language abilities often caused by a left-hemispheric brain stroke. Our patients predominantly had expressive aphasia, meaning that they had problems to produce language (spoken and written), but comprehension was generally sufficient to comply with the training. In the new training protocol, patients have to detect a target word within a rapid auditory sequence of several words while their brain signals are analyzed. After each trial, patients then receive feedback on whether their brain signals indicate a successful target detection. Providing this information as immediate feedback should allow the patient to strengthen basic processes underlying language function.

Clinical assessments showed that the BCI-supported training has induced significant, strong and lasting improvements not only in language comprehension, but also in language production, writing, reading and everyday communication for every single patient. This is a remarkable finding as all patients were in the chronic phase — meaning that the time from the training begin to the stroke was at least 6 months, but sometimes even several years — and regularly underwent ordinary speech and language therapy before our training which showed only limited effects. In Chapter 4, I will

present this study in more detail. In addition, I will present a side study which demonstrates that the training task might be easier if the subject closes their eyes, see Appendix a.

The following Figure 1.1 gives an overview of the thesis structure and the list of publications is given in Section 1.1.

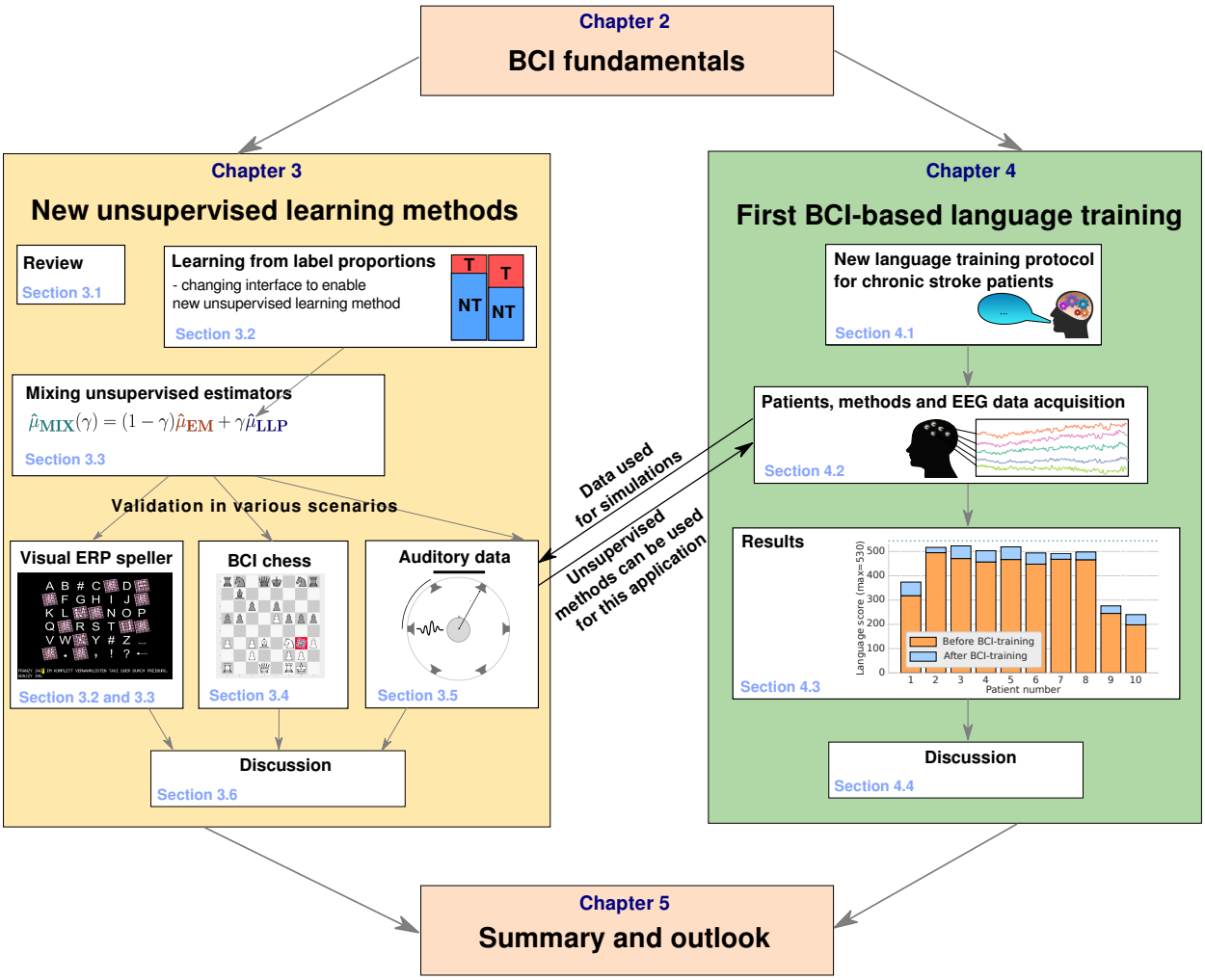


Figure 1.1: **Structure and contributions of this thesis.**

1.1 LIST OF PUBLICATIONS

Journal publications (6):

[1] **Hübner D**, Verhoeven T, Müller K-R, Kindermans P-J and Tangermann M (2018). "*Unsupervised learning for brain-computer interfaces based on event-related potentials: review and online comparison*". IEEE Computational Intelligence Magazine, 14, pp. 66-77.

[2] **Hübner D**, Verhoeven T, Schmid K, Müller K-R, Tangermann M and Kindermans P-J (2017). "*Learning from label proportions in brain-computer interfaces: online unsupervised learning with guarantees*". PLOS ONE. Vol. 12(4), pp. e0175856. Public Library of Science.

[3] **Hübner D**, Schall A and Tangermann M (2019). "*Unsupervised learning in a BCI chess application using learning from label proportions*". Brain-computer interface journal. Taylor&Francis publishing group. **Under review.**

[4] **Hübner D**, Schall A, Prange N and Tangermann M (2018). "*Eyes-closed increases the usability of brain-computer interfaces based on auditory event-related potentials*". Frontiers in Human Neuroscience. Vol. 12:391. doi: 10.3389/fnhum.2018.00391.

[5] Musso M, **Hübner D**, Schwarzkopf S, Weiller C and Tangermann M (2019). A brain-computer interface for language training in chronic post-stroke aphasia patients. Brain. **Under preparation.**

[6] Verhoeven T, **Hübner D**, Tangermann M, Müller K-R, Dambre J and Kindermans P-J (2017). "*Improving zero-training brain-computer interfaces by mixing model estimators*". Journal of Neural Engineering. Vol. 14(3), pp. 036021. IOP Publishing.

Conference contributions (11)

[7] **Hübner D**, Schall A and Tangermann M (2019). "*Two player online brain-controlled chess*". Engineering in Medicine and Biology Society (EMBC), 2019 Annual International Conference of the IEEE. Accepted.

[8] **Hübner D**, Schwarzkopf S, Musso M and Tangermann M (2018). "*BCI-based language training induces changes in ERP responses in chronic post-stroke aphasia patients*". In: Proceedings of the 7th International BCI Meeting 2018. pp. 106-107.

[9] Sosulski J, **Hübner D** and Tangermann M (2018). "*Closed-loop stimulus parameter optimization framework for event-related potential paradigms*". In: Proceedings of the 7th International BCI Meeting 2018. pp. 108-109.

[10] Tangermann M, **Hübner D**, Schwarzkopf S, Weiller C and Musso M (2018). "*Effects on language ability induced by BCI-based training of patients with aphasia*". In: Proceedings of the 7th International BCI Meeting 2018. pp. 110-111. **Second price of the BCI award 2018.**

[11] Musso M, **Hübner D**, Schwarzkopf S, Weiller C and Tangermann M (2018). "*A novel aphasia training based on brain-computer interface*". In: European Stroke Journal, Vol. 3(1S), pp. 205-206.

[12] **Hübner D**, Verhoeven T, Kindermans P-J and Tangermann M (2017). "*Mixing two unsupervised estimators for event-related potential decoding: An online evaluation*". In: Proceedings of the 7th International Brain-Computer Interface Meeting 2017. pp. 198-203. Verlag der Technischen Universität Graz. **Best talk award and nominated for the BCI award 2017.**

[13] **Hübner D**, Kindermans P-J, Verhoeven T and Tangermann M (2017). "*Improving learning from label proportions by reducing the feature dimensionality*". In: Proceedings of the 7th International Brain-Computer Interface Meeting 2017. pp. 186-191. Verlag der Technischen Universität Graz.

[14] **Hübner D** and Tangermann M (2017). "*Challenging the assumption that auditory event-related potentials are independent and identically distributed*". In: Proceedings of the 7th International Brain-Computer Interface Meeting 2017. pp. 192-197. Verlag der Technischen Universität Graz.

[15] **Hübner D**, Verhoeven T, Schmid K, Müller K-R, Tangermann M and Kindermans P-J (2017). "*Learning from label proportions in BCI — A symbiotic design for stimulus presentation and signal decoding*". In: The First Biannual Neuroadaptive Technology Conference. pp. 27-29. **Best talk award.**

[16] Kindermans P-J, **Hübner D**, Verhoeven T, Schmid K, Müller K-R and Tangermann M (2016). "*Making brain-computer interfaces robust, reliable and adaptive with learning from label proportions*". NIPS Workshop.

[17] Musso M, Bambadian A, Denzer S, Umarova R, **Hübner D**; and Tangermann M (2016). "*A novel BCI based rehabilitation approach for aphasia rehabilitation*". In: Proceedings of the Sixth International Brain-Computer Interface Meeting. p. 104.

Book chapters (2)

[18] Kindermans P-J, Verhoeven T, **Hübner D**, Müller K-R and Tangermann M (2018). Chapter 6: "*Alleviating the effects of non-stationarity and improving the efficacy of brain-computer interfaces with unsupervised learning*". In: "*Signal processing and machine learning for brain-machine interfaces*" edited by T. Tanaka and M. Arvaneh. ISBN: 978-1-78561-398-2.

[19] **Hübner, D**, Kindermans, P-J, Verhoeven, T, Müller, K-R and Tangermann M (2018). "*Rethinking BCI paradigm and machine learning algorithm as a symbiosis: zero calibration, guaranteed convergence and high decoding performance.*" In: "*Brain-computer interface research: a state-of-the-art summary 7*" edited by C. Guger, B. Allison, G. Edlinger. Springer. ISBN: 978-3-030-05667-4.

2

FUNDAMENTALS

ABSTRACT

This chapter gives the reader an introduction to the basic tools and methods that will be used throughout this thesis. I will cover important aspects of the basic brain-computer interface loop with a special focus on recording, analyzing and classifying event-related potentials from the electroencephalography, and on application fields.
For more advanced readers, Section 2.4.4 might be of interest. In this section, I will demonstrate how the derivation of the linear discriminant analysis based on the (1) minimization of misclassifications (Bayes classifier), (2) maximization of the Fisher criterion (Fisher linear discriminant analysis) or (3) minimization of the squared residuals with rescaled labels (least squares classifier), will always lead to the same projection vector. One important corollary is that the use of the pooled covariance matrix instead of the class-wise covariance matrix does not change the direction of this projection. This property is of great importance for the unsupervised learning that will be presented in Chapter 3.

2.1 BRAIN-COMPUTER INTERFACES

The term "*brain-computer interface*" (BCI) is almost half a century old. Jacques J. Vidal coined the term in a very prospective paper in 1973 [175]. Although he could not implement a BCI at that time, the basic principle that he had proposed is still in place today. Figure 2.1 shows a schematic overview of the basic operating cycle of a BCI. At least, four components are vital for a BCI.

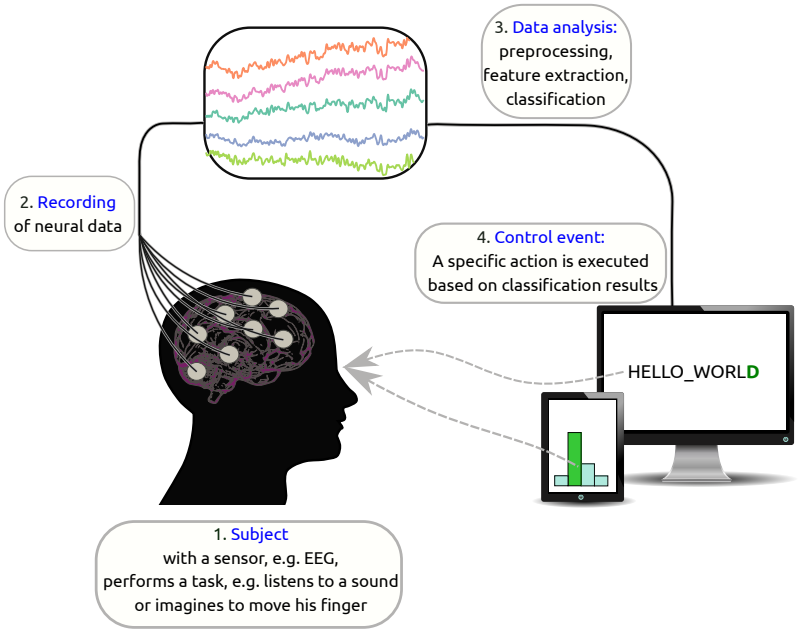


Figure 2.1: **The basic BCI cycle.** A subject's brain signals are recorded while he/she performs a certain task. Based on the neural recordings, a machine learning classifier extracts information about the brain state of the subject. This information is used to control an application (e. g., spelling or as feedback to enable skill learning).

1. **Subject.** First, there needs to be a subject. The subject could be an animal as well as a human. I will focus on human subjects as all studies in this thesis have been conducted with humans. The subject performs a certain task such as to pay attention to specific stimuli or to perform a mental task, e. g., imagining to move his limb or subtracting numbers.

2. **Recording.** During that task, the subject's brain signals are recorded. Most recording techniques either rely on measuring electrical activity or by measuring the hemodynamic activity in the brain. Electrical activity can both be recorded from electrodes placed within the skull (invasive) or on top of the scalp (non-invasive). Hemodynamic, meaning to measure the dynamics of the blood flow, is commonly applied non-invasively by means of functional magnetic resonance imaging (fMRI) or functional near-infrared spectrography (fNIRS). All recordings in this thesis used non-invasive electrical signals from

the electroencephalogram, see Section 2.2, and focus on event-related potentials, see Section 2.3.

3. **Data analysis.** The recorded brain data are processed in the third step. The data analysis generally comprises steps to clean the data from ambient noise and artifacts, and to extract the quantities of interests. Once relevant features are extracted, classification or regression models will then be applied to extract information about the ongoing brain state of the subject. In this step, individualized parameters of the machine learning model are learned for each subject. This is essential because of the high variance across subjects which cannot be captured by a single set of parameters. A more detailed description of the data analysis step is given in Section 2.4.

4. **Control event.** The information about the ongoing brain state can then be used in the final step to change the state of a computer or machine. Various applications are possible. In a control or communication application, this information is used to steer an application such as a spelling device or a wheelchair. Control is then possible if stimuli or mental tasks are associated with actions, e. g., the subject could spell an "A" by paying attention to the highlighting of the letter "A" on the screen, or he/she could steer a wheelchair to the left by imagining to move his left arm. Other application areas are rehabilitation, gaming, basic research, among others, see Section 2.8.

As of April, 15th 2019, more than 7000 publications can be found by searching for "*brain-computer interfaces*" in the PubMed database. Although J. Vidal coined this term already in 1973, the first implementation dates back to 1988 when Farwell and Donchin could realize the first visual speller [50]. In 2001, the first successful communication with patients that had amyotrophic lateral sclerosis (ALS) [103] could be established with a BCI. This — among other factors — led to a steep increase in publications, see Figure 2.2. Vidal wisely remarks in 1973 that "*the long-range implications of systems of that type [meaning BCIs] can only be speculated upon at present*".

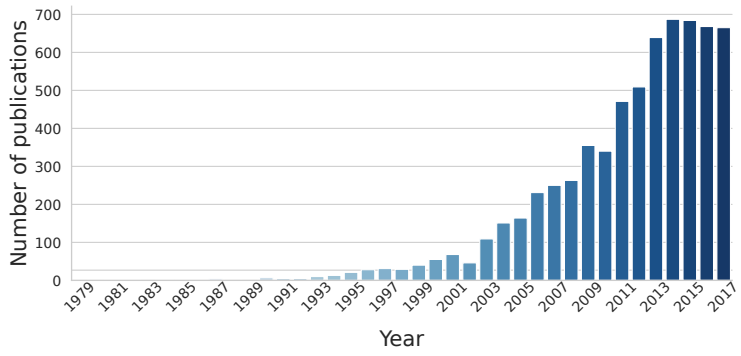


Figure 2.2: **Number of publications in the PubMed database when searching for *"brain-computer interface"*.** The PubMed database was accessed at April, 15th 2019.

2.2 ELECTROENCEPHALOGRAPHY

The majority of BCIs is based on recording neural data with electroencephalography (EEG) [180]. This technique was discovered by Hans Berger in 1924 and coined by him in a publication 1929. Already back then, he realized — based on his own and other research — that every living cell is able to produce electrical currents and that the large synchronous activity of cells can be recorded even with electrodes placed on the scalp [13]. It has now been established that EEG represents mainly the post-synaptic potentials of pyramidal cells close to cortical surfaces [180]. When compared to other recording techniques, EEG has several advantages.

+ EEG is non-invasive, meaning that no permanent damage to the brain is induced.
+ It is possible to get a temporal resolution in the order of milliseconds.
+ Compared to other brain recording methods, it is relatively cheap, easy to use and can be used in a portable setup.

However, there are also some disadvantages.

- EEG has a low spatial resolution.
- The signal-to-noise ratio (SNR), which measures how clearly the signal of interest can be distinguished from the background noise, is relatively poor.
- Measured EEG signals are strongly affected by movement artifacts, especially by eye movements and eye blinks.

2.2.1 *Practical aspects*

Electrodes are predominantly made from Ag-AgCl, because of their property to avoid potential shifts due to electrode polarization. Any measurement not only requires the recording electrode, but also a reference electrode. This electrode should be ideally located electrically far from the reference electrode [180]. Common places for the reference electrode are the ear, mastoid, neck or nose. In this thesis, the reference was always placed on the nose. The second electrode of interest is the ground electrode. This electrode is primarily used to cancel out electric noise in the system (e. g., power line noise or slow drifts) by applying a common mode rejection. The necessity of a ground electrode comes from the fact that a differential amplifier is not only amplifying the differences between the signals of interest, but also the common modes, i.e. signals that are shared between electrodes, although the latter is amplified by a lower factor. The ground can help to reduce the common modes and with that, reduces the overall recording error. In our studies, the ground was always placed on the forehead (position AFz).

To record data with EEG, EEG caps are normally used. These are caps made from a light fabric which have predefined positions for the placement of electrodes. A widely used positioning scheme is the international

10-20 system [86]. This system allows for a placement based on standard landmarks of the skull (i.e. nasion and inion), provides a coverage of the whole head and introduces a unified naming scheme. An example for 64 electrodes is shown in Figure 2.3.

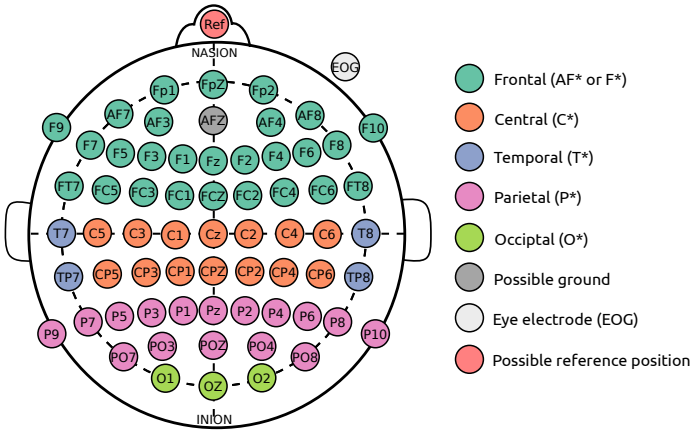


Figure 2.3: **The placement of 64 electrodes according to the 10-20 system.** The position of the ground and reference electrode varies.

To prepare typical wet electrodes, a conducting paste is applied to each electrode. Then, the skin is carefully abraded with cotton tabs such that there is a good contact between the scalp and the electrode. Electrode impedances should be kept on a low level to avoid noise amplification. In all studies in this thesis, impedances were kept below 20 kΩ. Figure 2.4 shows exemplary data of 5 seconds of EEG recordings from 32 channels.

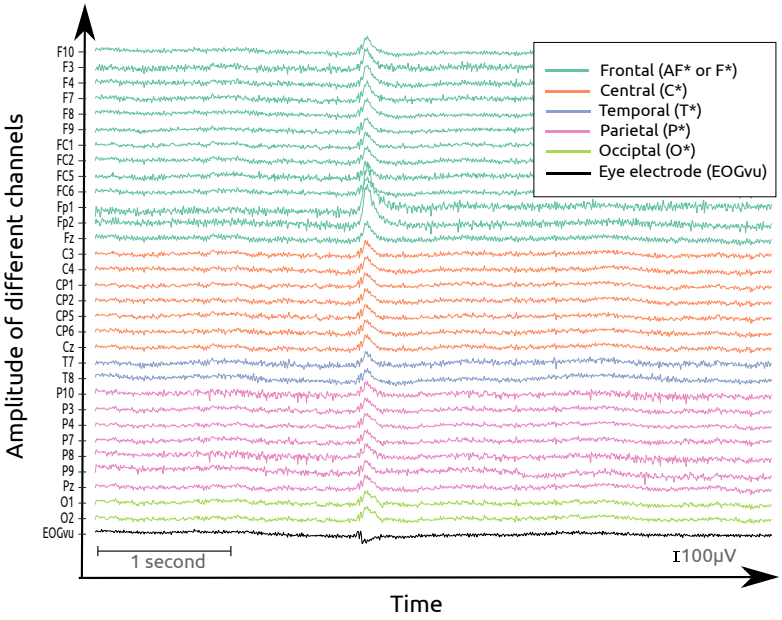


Figure 2.4: **Five seconds of EEG data.** The event at around two seconds comes from an eye blink. For the eye blink, it is interesting to observe that the eye electrode (EOGvu) displays a deflection in the opposite direction because it is located beneath the eye as opposed to the other electrodes. One can also observe a high correlation between channels.

2.3 EVENT-RELATED POTENTIALS

When processing an event, the brain elicits a brain response which is called an event-related potential (ERP). An event can be the perception of a visual or auditory stimulus, observing an error or even the absence of a stimulus when expecting one. The registration of such an ERP can provide valuable information about the ongoing brain processes of the subject. An ERP response consists of a series of voltage deflections. They are named after the sign of the amplitude (P: positive, N: negative) and their peak timing, e. g., P300 (or P3) means that there was a positive deflection after 300 ms. An ERP response is embedded in the ongoing EEG and relatively small compared to the background activity. Hence, the ERP responses over multiple events are normally averaged to obtain a clear estimation of the ERPs.

A commonly used paradigm to elicit ERP responses is the so-called oddball paradigm. In that paradigm, a series of at least two different stimuli (e. g., a high and a low tone) is presented to the user. The user's task is to pay attention to one specific stimulus. Typically, one stimulus is less frequent than others and the subject is instructed to pay attention to the rare stimulus. The attended stimulus is then called a *target* stimuli whereas the other stimuli are called *non-targets*. The delay between the onset of two consecutive stimuli is called stimulus onset asynchrony (SOA) and is one second in the original oddball task.

Figure 2.5 shows typical average target and non-target oddball ERP responses. Two main components are visible. There is an early negativity (N200) which is located around channel Fz and a positive response (P300) peaking around Cz-Pz. In the following, I will provide a more detailed description of these two components.

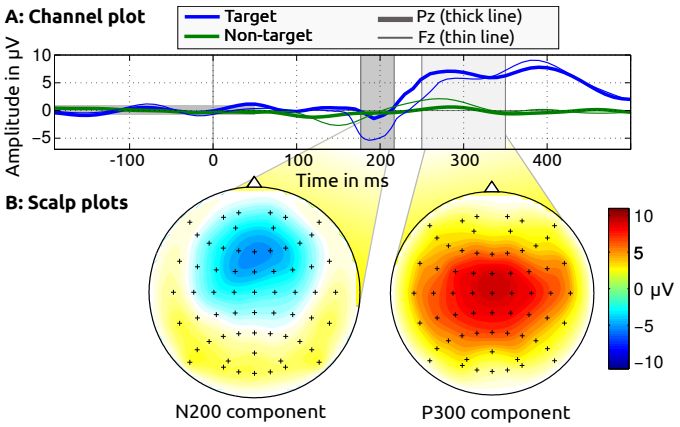


Figure 2.5: **Prototypical oddball ERP response.** Target and non-target ERP responses are depicted for the channels Pz (thick line) and Fz (thin line) where the x-axis depicts the time and the y-axis shows the potential. In the bottom, the activation pattern of the average target responses for the intervals $[130, 170]$ ms and $[250, 350]$ ms are presented.

2.3.1 *P*300

The P300 is probably the most researched brain signal. Yet, Dinteren et al. [43] remark that "*after almost 50 years of intensive research with over 12,000 publications on the P300 it has not been possible to link the P300 to a specific cognitive process.*" They further note that "*presumably, the P300 complex is multifarious, reflecting a culmination of multiple cognitive processes.*"

The P300 has been linked to language processing. A decreased P300 amplitude was observed in children with language impairments [48] and the presence of the P300 component was indicative for the recovery of patients with aphasia (language deficits) [132]. More generally, the shape and latency of the P300 components have been linked to functional abilities. Dinteren et al. [43] find that "*there is evidence that shorter P300 latencies and larger amplitudes are associated with superior information processing.*" These are important findings as they suggest a link between the recorded brain signals and the cognitive capabilities. If the relationship is indeed of causal nature, then a training which reinforces the P300 component may lead to improved functional outcome. This idea was explored in-depth in Chapter 4 where I present the first successful BCI-based language training.

The P300 amplitude and latency is influenced by many factors. It has been found that "*as primary task difficulty is increased, P300 amplitude from the oddball task decreases regardless of modality or the motor requirements of the primary task*" [142]. The task difficulty is commonly increased by decreasing the time between consecutive stimuli (i. e. the target-to-target interval) [142], by using more complex stimuli (e. g., words instead of tones) and other modifications. The P300 has also been linked to factors like motivation [99], age [43] and personality attributes such as introversion/extraversion, arousal, sensation seeking, and compulsivity [142]. Because of these sensitivities, it is recommended to carefully control the general task setting and record character traits and attitudes to allow comparing the P300 across subjects. Please note that despite its name P300, the component may occur with a much longer latency, e. g., it is visible in the whole time interval of 300 ms to 1200 ms when recording word-evoked ERPs from stroke patients with language deficits, see Chapter 4.

2.3.2 *N*200

The other component that will be occurring throughout this thesis, is the N200. Similar to the P300, the N200 is modulated by attention and can be used to discriminate attended from non-attended stimuli. In contrast to the P300, however, this component depends heavily on the modality of the paradigm. While it is located mostly in the anterior-temporal part for auditory stimuli (see Figure 2.5), it is located over the occipital cortex for a visual task. This location can be explained by the position of the corresponding auditory and visual cortex located in the frontotemporal and occipital lobe, respectively. Please note that the timing of the N200 may also vary within 150 ms and 250 ms depending on the subject and task.

Figure 2.6 shows a typical BCI data analysis cycle for an ERP-based paradigm following the very instructional paper by Blankertz et al. [21]. In the previous section, I have covered the fundamental aspects of data acquisition with EEG. All recordings in this thesis were conducted with a sampling frequency of 1 kHz and amplified using BrainAmp amplifiers (BrainProducts). Several other steps are necessary to extract meaningful information from the EEG signals.

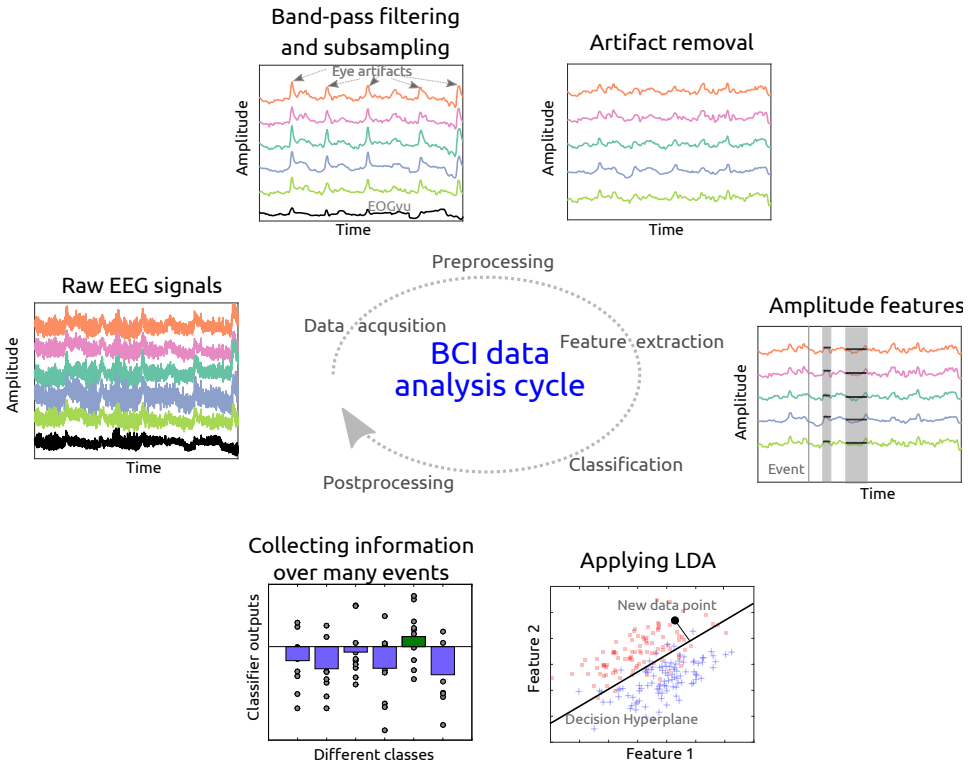


Figure 2.6: **Data analysis pipeline.** The data processing pipeline can roughly be subdivided into five steps. First, raw EEG data is acquired. It is then preprocessed using frequency filtering, subsampling and artifact removal. Afterwards, ERP features are extracted by averaging parts of the signal. Then, a classifier (e. g., a linear discriminant analysis (LDA)) is applied to classify the features as belonging to the target or non-target class. Finally, the classifier outputs are combined to make a prediction for a complete trial about the target class. In a BCI, this loop is usually completed in quasi real-time meaning that a single trial decision is normally obtained within few seconds after the end of a trial.

2.4.1 *Preprocessing*

First, the data is preprocessed. In our recordings, a third-order Chebyshev Type II low-pass filter with a cut-off frequency of around 12 Hz was typically applied. If many artifacts are present, lower frequencies such as 8 Hz

can also be chosen without affecting the classification accuracies, see for instance [49]. Afterward, the data is subsampled to 100 Hz, mainly to reduce the computational and memory requirements. A second Chebyshev Type II filter is then applied which is a high-pass filter with a cut-off frequency of 0.5 Hz. This order of filtering and subsampling is crucial to avoid aliasing effects. For offline analysis, non-causal filters are used, i. e. time-invariant filter that require knowledge about upcoming samples, whereas causal filters are used in online scenarios.

2.4.2 *Artifact rejection*

After basic filtering and subsampling, artifact rejection methods are applied to deal with eye artifacts and other kind of movement artifacts. While some methods work on the continuous data, others require the data to be epoched. An epoch is a time-window that is segmented with respect to an event, e. g., -200 ms to 700 ms around the presentation of a visual stimulus. A subset of the four following artifact removal methods is used throughout this thesis.

- **Regressing out eye artifacts**. Eye movements and eye blinks can be projected out by assuming a stationary eye movement pattern [134]. In this approach, the horizontal eye movement is estimated based on the channels F9 and F10 and the vertical eye movement is approximated by using the channels Fp2 and EOGvu. A regression approach is then used to project out eye movements.

- **MARA.** Alternatively to the first approach, the multiple artifact rejection algorithm (MARA) [178, 179] dissects the continuous EEG signals into single component based on independent component analysis (ICA). The components are then classified with a supervised model that is based on 6303 expert-labeled ICA components as artifactual or non-artifactual. ICA components that were classified as artifacts are then projected out from the data.

- **Variance-criterion.** Electrodes showing only limited variance (less than 0.5 μV in more than 10% of the trials) or too much variance (more than three times the difference between the 90^{th} percentile and the 10^{th} percentile of the variance of all electrodes) were rejected because this indicates that they were faulty or had bad impedances. In addition, high-variance epochs (i. e. epochs that show more than three times the difference between the 90^{th} percentile and the 10^{th} percentile of the variance of all epochs) were also removed because they might have been contaminated by artifacts such as eye blinks.

- **MinMax.** Epochs were marked where the peak-to-peak amplitude in one epoch exceeds a certain threshold in one of the frontal channels (Fp1, Fp2, F7, F8, F9, F10). The threshold may vary depending on the subject with typical values between 60 to 100 μV.

2.4.3 *Feature extraction*

In the next step, the features of interest are extracted from the cleaned data. For that, epochs are first extracted from the data by taking a segment around an event. Then, epochs are baselined meaning that the average amplitude in an interval before the stimulus is subtracted from each epoch. Finally, the average amplitudes are computed for each channel in several intervals after the stimulus. These averages are the typical features that are used for classification in an ERP paradigm. The feature dimensionality D is given by the product of the number of electrodes with the number of intervals. Throughout this thesis, up to 64 electrodes and 10 intervals are used, leading to a maximal feature dimensionality of 640. More formally, our data can be written as

$$(\mathbf{x}_i, y_i)_{i=1}^N$$

where N is the number of epochs, $\mathbf{x}_i \in \mathbb{R}^D$ are the D-dimensional features and $y_i \in \{-1, 1\}$ are the labels representing epoch from the non-target and target class, respectively.

2.4.4 *Classification with linear discriminant analysis*

To decide for a new data point whether it was a target or non-target stimulus, a machine learning classifier is used. The BCI community spent an enormous effort in investigating different classification models [117]. One popular classifier is linear discriminant analysis (LDA) that has shown excellent results and is easy to implement [21]. This is a binary linear classifier that fits a projection vector $\mathbf{w} \in \mathbb{R}^D$ and bias term $b \in \mathbb{R}$ such that the decision hyperplane is given by all points $\mathbf{x}$ with $\mathbf{w}^T\mathbf{x} + b = 0$. For the feature vector of a new data point, $\mathbf{x}_T$, we can then predict its class label by computing

$$y_T = \text{sign}(\mathbf{w}^\text{T}\mathbf{x}_T + b) \tag{2.1}$$

where $(\cdot)^T$ denotes the transpose of a vector or matrix. and $\text{sign}(x) := 1$ if $x \geq 0$ and -1 else.

In this context, the machine learning task is to obtain "good" values for $\mathbf{w}$ and b. Here "good" is placed in quotation marks because the quality of a classifier can only be judged with regard to a certain evaluation metric and it is a priori unknown, which metric can be considered as the most informative one.

Importantly, the LDA classifier is optimal if the data is normally distributed with equal class-wise covariance matrices — an assumption that is met by ERP data [21] after artifacts have been removed. Given this assumption, there are three popular ways to derive the LDA classifier based on the (1) minimization of misclassifications (Bayes optimal solution), (2) maximization of the Fisher criterion or (3) minimization of the sum of squared residuals computed with rescaled labels. In the following, I will

demonstrate that all approaches will lead to the same direction of $\mathbf{w}$, but to different estimations of b. The variety in approaches demonstrates that the LDA classifier is optimal with regard to several different metrics which underlines that it is an excellent choice for classifying ERP signals.

Bayes classifier

In the Bayes approach following [57], the goal is to minimize the 0-1-loss that measures the ratio of misclassified data points. This is equivalent to maximizing the *probability* of assigning the correct class label to a new data point in a two-class problem. Formally, for a two-class problem with class labels y_1 and y_2, our goal is to compute the class with the highest probability $\hat{y}_0$ for a given data point $\mathbf{x}_0$ from a random variable X.

$$\hat{y}_0 = \underset{k \in \{y_1, y_2\}}{\arg\max}\, P(y = k | X = \mathbf{x}_0) \tag{2.2}$$

To solve this problem, we first determine the decision boundary where both classes have equal probabilities. This is obtained by equalizing their posterior probabilities.

$$P(y = y_1 | X = \mathbf{x}) = P(y = y_2 | X = \mathbf{x}) \tag{2.3}$$

We need to apply Bayes theorem and use the normality assumption to solve this equality. First, Bayes theorem is given by

$$P(y = k | X = \mathbf{x}) = \frac{\text{likelihood} \cdot \text{prior}}{\text{marginal}} = \frac{f_k(x) \cdot \pi_k}{P(X = \mathbf{x})}. \tag{2.4}$$

where π_k is the prior probability of class k given by $\pi_k = \frac{N_k}{N}$ where N_k is the number of samples in class k. Following the assumption of normality with class means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^D$ and shared covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{DxD}$, the class-wise density function $f_k(\mathbf{x})$ is given by

$$f_k(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^D \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right). \tag{2.5}$$

where $k = 1, 2$ represents the two classes. Plugging Bayes theorem into Equation 2.3, one can observe that the marginal distribution and N cancel out.

$$f_1(\mathbf{x}) \cdot N_1 = f_2(\mathbf{x}) \cdot N_2 \tag{2.6}$$

Dividing by the right-hand side and applying the logarithm on both sides yields

$$\log f_1(\mathbf{x}) - \log f_2(\mathbf{x}) + \log \frac{N_1}{N_2} = 0. \tag{2.7}$$

Now, one can insert the density function from Equation 2.4 which will make the normalization terms cancel out and give

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) + \log \frac{N_1}{N_2} = 0. \tag{2.8}$$

Using basic algebra, this can be simplified in a final step to

$$\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right)^{\mathrm{T}} \mathbf{x} + \frac{1}{2}\left(-\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2\right) + \log\frac{N_1}{N_2} = 0. \quad (2.9)$$

In this equation, we can now identify our projection $\mathbf{w}$ and bias b as

$$\mathbf{w}_{\text{Bayesian}} := \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \quad (2.10)$$

$$b_{\text{Bayesian}} := \frac{1}{2}\left(-\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2\right) + \log\frac{N_1}{N_2} \quad (2.11)$$

$$= \frac{1}{2}\mathbf{w}^{\mathrm{T}}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \log\frac{N_1}{N_2} \quad (2.12)$$

Based on these quantities, class 2 is predicted if $\mathbf{w}_{Bayesian}^{\mathrm{T}}\mathbf{x} + b_{\text{Bayesian}} \geq 0$ and otherwise class 1 is predicted (although possible, it is easy to see that the opposite assignment will lead to a higher 0-1-loss). In comparison to the following approaches, the Bayes method has the advantage that it cannot only predict the class labels, but actually assigns a probability to it which can serve as a measure of certainty. This might for instance be used as a confidence score in a dynamic stopping method that stops a trial once the BCI is certain about its decision.

Fisher discriminant analysis

Instead of maximizing the probability of correctly assigning the class labels, Fisher proposed the following approach: a good projection $\mathbf{w}$ should achieve two goals at the same time. First, the data points from the same class should be located nearby in the projected subspace. Second, the distance between classes should be large in the projected space. This idea can be formalized with the Fisher criterion.

$$J(\mathbf{w}) := \frac{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{W}}\mathbf{w}} \quad (2.13)$$

In this formula, the nominator comprises the **between-class covariance matrix** $\boldsymbol{\Sigma}_{\mathrm{B}}$ and is defined as

$$\boldsymbol{\Sigma}_{\mathrm{B}} := \sum_{c=1}^{2} N_c(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^{\mathrm{T}} \quad (2.14)$$

where the class-means $\boldsymbol{\mu}_i$ are defined as $\boldsymbol{\mu}_c := \frac{1}{N_c}\sum_{i\in\mathcal{C}_c}\mathbf{x}_i$ and the global mean is defined as $\boldsymbol{\mu} := \frac{i}{N}\sum_{i=1}^{N}\mathbf{x}_i$. The denominator is given by the total **within-class covariance matrix** $\boldsymbol{\Sigma}_{\mathrm{W}}$ and is defined as

$$\boldsymbol{\Sigma}_{\mathrm{W}} := \sum_{c=1}^{2}\sum_{\mathrm{i}\in C_c}(\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^{\mathrm{T}} \quad (2.15)$$

Please note that $\boldsymbol{\Sigma}_{\mathrm{W}}$ is the weighted sum of both class-wise covariances matrices and an unbiased estimator of the aforementioned $\boldsymbol{\Sigma}$.

The goal is to find a $\mathbf{w}$ that maximizes $J(\mathbf{w})$ which means to either obtain a large between-class projection or by getting a small within-class projection following the idea by Fisher. Figure 2.7 depicts two different projections onto a one-dimensional subspace. The first projection (LDA) maps the data on a single dimension such that the classes are nicely separated. In contrast, the second projection (principal component analysis, PCA) finds a projection that maximizes the variance of the one-dimensional subspace. This, however, does not help for classification in this scenario.

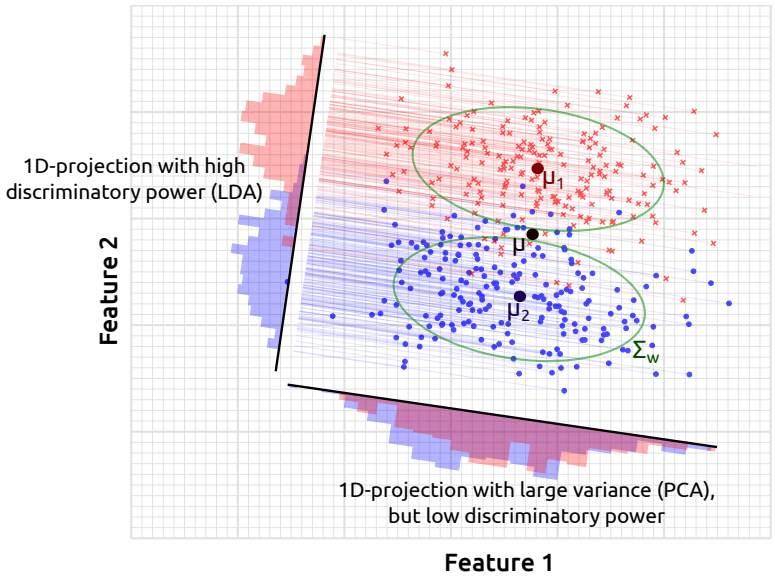


Figure 2.7: **Illustration of LDA and PCA projections.** The data from two classes is projected to a single dimension based on different criteria.

As shown in Bishop's book [17], differentiating $J(\mathbf{w})$ with respect to $\mathbf{w}$ and setting the derivative to zero, one obtains that

$$\mathbf{w}_{\text{Fisher}} \propto \mathbf{\Sigma}_{\text{W}}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \tag{2.16}$$

The projection $\mathbf{w}_{\text{Fisher}}$ is only uniquely determined up to a scaling constant. This constant is not affecting the decision boundary. In practical applications, $\mathbf{w}$ is often rescaled such that the class means are projected to $+1$ and -1, respectively. It is worth noticing that Fisher's LDA does not deliver a formula for computing the bias term b.

For the unsupervised learning part in this thesis, it is essential to show that the class-wise scatter matrix $\mathbf{\Sigma}_{\text{W}}$ can be replaced by a scatter matrix that does not rely on label information. To show this, we first introduce the notion of the **total-scatter matrix** $\mathbf{\Sigma}_{\text{T}}$ which estimates the pooled covariance, i.e. the covariance on the complete data disregarding label information.

$$\mathbf{\Sigma}_{\text{T}} := \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\text{T}} \tag{2.17}$$

Importantly, it holds that

$$\mathbf{\Sigma}_{\text{T}} = \mathbf{\Sigma}_{\text{W}} + \mathbf{\Sigma}_{\text{B}}. \tag{2.18}$$

Proof.

$$
\begin{aligned}
\boldsymbol{\Sigma}_{\mathrm{T}} &= \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \\
&= \sum_{c=1}^{2} \sum_{i \in C_c} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \\
&= \sum_{c=1}^{2} \sum_{i \in C_c} (\mathbf{x}_i - \boldsymbol{\mu}_c + \boldsymbol{\mu}_c - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu}_c + \boldsymbol{\mu}_c - \boldsymbol{\mu})^{\mathrm{T}} \\
&= \sum_{c=1}^{2} \sum_{i \in C_c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^{\mathrm{T}} + \sum_{c=1}^{2} \sum_{i \in C_c} (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^{\mathrm{T}} \\
&\quad + \sum_{c=1}^{2} \sum_{i \in C_c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\boldsymbol{\mu}_c - \boldsymbol{\mu})^{\mathrm{T}} + (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu}_c)^{\mathrm{T}} \\
&= \boldsymbol{\Sigma}_{\mathrm{W}} + \boldsymbol{\Sigma}_{\mathrm{B}} + \\
&\quad \sum_{c=1}^{2} \underbrace{\left[\sum_{i \in C_c} (\mathbf{x}_i - \boldsymbol{\mu}_c) \right]}_{=0} (\boldsymbol{\mu}_c - \boldsymbol{\mu})^{\mathrm{T}} + \sum_{c=1}^{2} (\boldsymbol{\mu}_c - \boldsymbol{\mu}) \underbrace{\left[\sum_{i \in C_c} (\mathbf{x}_i - \boldsymbol{\mu}_c)^{\mathrm{T}} \right]}_{=0} \\
&= \boldsymbol{\Sigma}_{\mathrm{W}} + \boldsymbol{\Sigma}_{\mathrm{B}}
\end{aligned}
$$

□

Second, we will show that for any vector $\mathbf{x} \in \mathbb{R}^D$, it holds that $\boldsymbol{\Sigma}_{\mathrm{B}}\mathbf{x}$ points in the direction of the differences of the class means, i. e.

$$
\boldsymbol{\Sigma}_{\mathrm{B}}\mathbf{x} \propto (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1). \tag{2.19}
$$

Proof. We will make use of $\boldsymbol{\mu} = \frac{N_1}{N}\boldsymbol{\mu}_1 + \frac{N_2}{N}\boldsymbol{\mu}_2$ and first compute

$$
\begin{aligned}
\boldsymbol{\mu}_2 - \boldsymbol{\mu} &= \boldsymbol{\mu}_2 - \left(\frac{N_2}{N}\boldsymbol{\mu}_2 + \frac{N_1}{N}\boldsymbol{\mu}_1 \right) \\
&= \frac{N\boldsymbol{\mu}_2 - N_2\boldsymbol{\mu}_2 - N_1\boldsymbol{\mu}_1}{N} \\
&= \frac{(N - N_2)\boldsymbol{\mu}_2 - N_1\boldsymbol{\mu}_1}{N} \\
&= \frac{N_1\boldsymbol{\mu}_2 - N_1\boldsymbol{\mu}_1}{N} = \frac{N_1}{N}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1).
\end{aligned}
$$

With that and the analogous formula for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}$, the between scatter matrix can be rewritten in the following way.

$$
\begin{aligned}
\boldsymbol{\Sigma}_{\mathrm{B}} &= N_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu})(\boldsymbol{\mu}_1 - \boldsymbol{\mu})^{\mathrm{T}} + N_2(\boldsymbol{\mu}_2 - \boldsymbol{\mu})(\boldsymbol{\mu}_2 - \boldsymbol{\mu})^{\mathrm{T}} \\
&= N_1 \frac{N_2}{N}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \frac{N_2}{N}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}} + N_2 \frac{N_1}{N}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \frac{N_1}{N}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}} \\
&= \frac{N_1 N_2 (N_1 + N_2)}{N^2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}}
\end{aligned}
$$

$$= \frac{N_1 N_2}{N}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}}$$

For any vector $\mathbf{x} \in \mathbb{R}^D$ follows that

$$\boldsymbol{\Sigma}_{\mathrm{B}}\mathbf{x} = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\frac{N_1 N_2}{N}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}}\mathbf{x} = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \cdot \beta \tag{2.20}$$

where $\beta = \frac{N_1 N_2}{N}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}}\mathbf{x} \in \mathbb{R}$. This completes the proof. □

We can now combine the results to show that the projection obtained with the total scatter matrix points in the same direction as the projection obtained with the within-class covariance matrix in Equation 2.16, i. e. we need to show that

$$\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \propto \boldsymbol{\Sigma}_{\mathrm{W}}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1). \tag{2.21}$$

Proof. We define $\mathbf{w} := \boldsymbol{\Sigma}_{\mathrm{W}}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. The objective is to show that there exists a real number $\alpha \in \mathbb{R}$ such that $\mathbf{w} = \alpha \cdot \boldsymbol{\Sigma}_{\mathrm{T}}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. Multiplying both sides with $\boldsymbol{\Sigma}_{\mathrm{T}}$, this is equivalent to showing that $\boldsymbol{\Sigma}_{\mathrm{T}} \cdot \mathbf{w} = \alpha \cdot (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. Equation 2.19 implies that there also exists a $\beta \in \mathbb{R}$ such that $\boldsymbol{\Sigma}_{\mathrm{B}}\mathbf{w} = \beta(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. Taken together, we obtain

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathrm{T}}\mathbf{w} &\overset{\text{Equation 2.18}}{=} (\boldsymbol{\Sigma}_{\mathrm{W}} + \boldsymbol{\Sigma}_{\mathrm{B}})\mathbf{w} \\ &= \boldsymbol{\Sigma}_{\mathrm{W}}\mathbf{w} + \boldsymbol{\Sigma}_{\mathrm{B}}\mathbf{w} \\ &= (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \beta \cdot (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = \alpha \cdot \mathbf{w} \end{aligned}$$

with $\alpha = 1 + \beta$. This finishes the proof and yields a relationship that will be from great importance in the unsupervised learning part. □

Least squares classifier

A third way of deriving LDA is by means of a least square classifier as presented in Bishop's book, chapter 4.1.3 and 4.1.5 [17]. The objective function in this approach is to minimize the squared loss of the data points.

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{N} (\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i)^2 \tag{2.22}$$

Please note that this uses the augmented notation where the bias term is incorporated into the projection as $\mathbf{w} = (b, \mathbf{w}_{old}^{\mathrm{T}})^{\mathrm{T}}$ and into the feature vector as $\mathbf{x} = (1, \mathbf{x}_{old}^{\mathrm{T}})^{\mathrm{T}}$. Differentiating this sum with respect to $\mathbf{w}$ and setting the derivative to zero yields the well-known solution of the least-squares problem.

$$\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}Y \tag{2.23}$$

In this equation, $\mathbf{X} = (\mathbf{x}_1^{\mathrm{T}}, \mathbf{x}_2^{\mathrm{T}}, \ldots \mathbf{x}_N^{\mathrm{T}})^{\mathrm{T}}$ is the feature matrix and $Y = (y_1, y_2, \ldots, y_N)^{\mathrm{T}}$ is the vector containing the labels. In Bishop's book,

it is then shown that choosing the target labels as $\frac{N}{N_1}$ for the positive class and as $-\frac{N}{N_2}$ for the negative class, will lead to a projection $\mathbf{w}$ that again points in the same direction as the previous ones [17].

In summary, LDA can be derived in three different ways. Given the assumption of two normally distributed classes with the same covariance matrix, LDA is optimal in the sense that it maximizes the probability of assigning the correct class labels, minimizes the 0-1-loss, maximizes the Fisher criterion and minimizes the least squares loss for rescaled class labels.

Figure 2.8 shows the implementation of the different approaches. One can observe that the projections are all parallel to each other, however, the bias terms are different. The different bias terms arise from the different optimization criteria (e. g., minimizing squared residuals or minimizing false classifications). In some practical use cases, the bias term is often not relevant when only relative distances from the hyperplane for each class are compared (see Section 2.6 below).

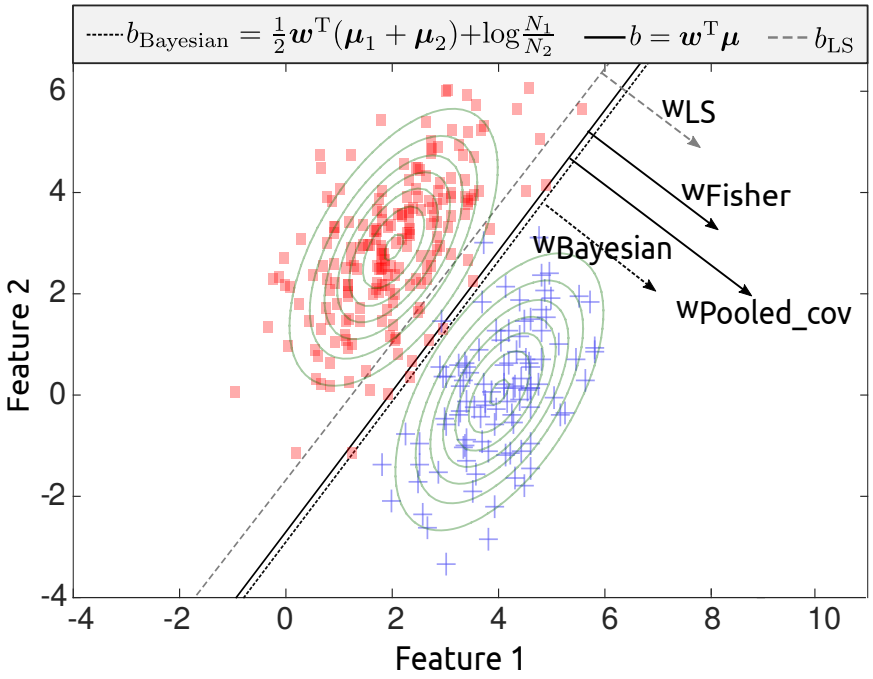


Figure 2.8: **Projection and bias term for different LDA implementations.** While the projections $\mathbf{w}$ always point the same direction, the bias terms may differ depending on the approach and associated loss function.

2.5 SUPERVISED AND UNSUPERVISED LEARNING

A fundamental problem in BCI is that there is a large subject-to-subject variability. Exemplary, I show the average target ERP response in the interval of $[250, 350]$ ms for 20 healthy elderly subjects in an auditory oddball task (one stimulus every second) in Figure 2.9. This should capture the classical P300 component, one of the most widely used and described components in a very basic task.

However, the variation between subjects in amplitude and peak location is astonishing. Only a few subjects actually display the P300 component as it is described in the literature and scalp patterns of different subjects hardly resemble each other. Clearly, the fact that the subjects were elderly (mean age around 60 years) amplified these inter-subject differences. However, it should be convincing without further evidence that patients with a brain injury (e. g., a brain stroke) will display even greater variability.

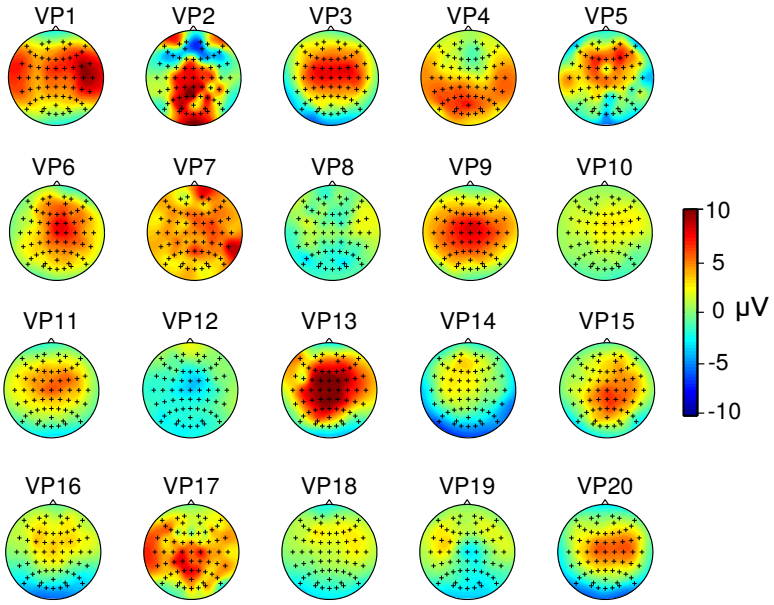


Figure 2.9: **Average oddball target ERP amplitude for 20 elderly subjects.** For each subject, 100 target ERPs were first cleaned using artifact removal methods and then averaged in the interval $[250, 350]$ ms. While the textbook describes a P300 component which has a positive peak around the electrodes Cz and Pz, only a few subjects actually show this P300 component (e. g., VP6, VP9, VP13, VP15) while some subjects display no systematic activity at all (e. g., VP8 or VP18) or a very noisy activation (e. g., VP2 or VP7).

The goal of the training of the machine learning model is to find subject- and session-specific parameters which can be used to classify brain responses. In the previous section, I introduced LDA as a suitable classification model for ERP data. The goal of this section is to understand how the recorded data can be used to achieve good estimations of the projection $\mathbf{w}$ and bias b.

2.5.1 *Supervised learning*

For supervised learning, a calibration session is performed prior to using the BCI. During this period, the subject performs a predefined task, e. g., the user spells a predefined sentence using the BCI. Since the task is known, one has full label information meaning that one always knew which task the subject tried to achieve. This allows to label each data point from the training data as being a target (class 1) or non-target stimulus (class 2).

When looking at the equations to compute the projection and bias term in the previous section, e. g., as given in Equation 2.16, one realizes that only a few quantities need to be known, namely the class-wise means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and the within-scatter-covariance matrix $\boldsymbol{\Sigma}_W$. Based on the labeled data, we can compute these quantities using the sample mean and sample covariance matrix. The hat symbol $\hat{}$ is used to denote estimators.

Sample class means:

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{N_c} \sum_{i \in \mathcal{C}_c} \mathbf{x}_i \tag{2.24}$$

Sample within-class covariance matrix:

$$\hat{\boldsymbol{\Sigma}}_W = \frac{1}{N-1} \sum_{c=1}^{2} \sum_{i \in C_c} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c)^T \tag{2.25}$$

The scaling factor $\frac{1}{N-1}$ for the within-class covariance matrix is necessary to have an unbiased estimator.

2.5.2 *Regularization of the covariance matrix*

In principle, the above estimation of the covariance will converge (in probability) to the true underlying covariance matrix when the number of data points goes to infinity, see Figure 2.10**A**. Hence, it is a consistent estimator. However, there is a systematic error occurring when the feature dimensionality is high and the number of samples is low: the estimated covariance matrix shows overestimated values for the largest eigenvalues and underestimated values for the smallest eigenvalues [110]. This effectively leads to a distortion of the estimated covariance matrix compared to the true covariance matrix. A solution to this problem is to regularize the covariance matrix towards the unity matrix as initially proposed by Ledoit and Wolf [110] and applied to BCI by Blankertz et al. [21].

$$\tilde{\boldsymbol{\Sigma}}(\gamma) := (1-\gamma)\hat{\boldsymbol{\Sigma}} + \gamma \nu \mathbb{I} \tag{2.26}$$

In this formula, $\gamma \in [0,1]$ is a tuning parameter, $\nu := \text{trace}(\hat{\boldsymbol{\Sigma}})/D$ represents the average eigenvalue of $\hat{\boldsymbol{\Sigma}}$ and D is the feature dimensionality. Importantly, Schäfer et al. [151] showed that the shrinkage parameter can be chosen automatically using the Ledoit–Wolf formula [110]. This can substantially improve classification accuracies [21]. An example for this regularization is depicted in Figure 2.10**B**.

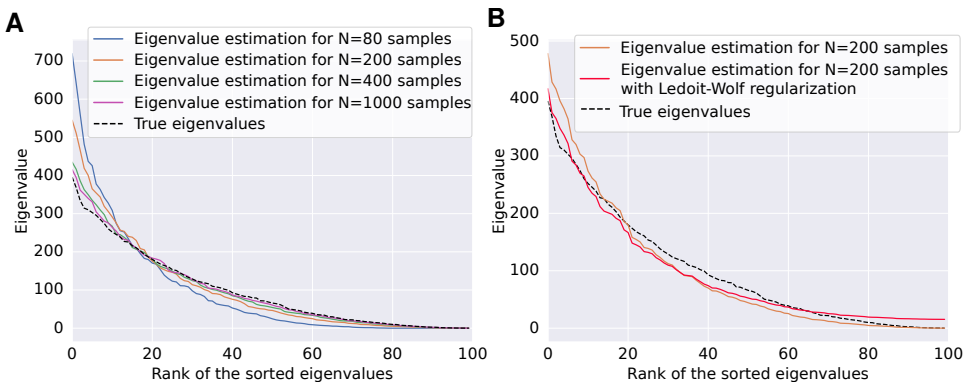


Figure 2.10: **Covariance estimation for different numbers of samples with and without regularization.** The left subplot (**A**) shows the influence of the number of samples on the quality of the covariance estimation. The right subplot (**B**) shows how regularization can improve the quality of the covariance estimation.

2.5.3 *Unsupervised learning*

In the unsupervised learning scenario, one wants to learn the projection and bias without having any label information, i.e. without knowing what the subject tried to accomplish. This is substantially more difficult than supervised learning. Nonetheless, using the results from before, we can partially simplify the problem. In the previous section, it was proven that one can replace the within-scatter covariance matrix by the total-scatter matrix in the computation of $\mathbf{w}$ without changing the direction of $\mathbf{w}$, see Equation 2.21. Importantly, no label information is required to estimate the total-scatter matrix.

Sample pooled/global covariance matrix:

$$\hat{\mathbf{\Sigma}}_{\mathrm{T}} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \tag{2.27}$$

This estimation can be improved by applying the same Ledoit–Wolf regularization as before.

Because the global covariance matrix can be estimated without labels, the unsupervised learning problems boils down to obtaining a good estimate of the class means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. In Chapter 3, two new approaches are introduced that are able to learn these mean values without any calibration phase at all. These classifiers show remarkable properties, namely, (1) a guarantee to converge to the true class means given the assumption of independently and identically distributed (IID) data and (2) a classification performance that — after a short ramp-up — is comparable to supervised classification. In addition, other unsupervised learning approaches will be reviewed.

2.6 POSTPROCESSING

In the previous section, I have explained how LDA can be used to classify a single event as being a target or non-target. In practical applications, the low signal-to-noise ratio of the EEG leads to a high error rate in this classification task. To alleviate this problem, a single trial generally consists of multiple repetitions of each event. This allows for accumulating evidence over a whole trial until one action is performed. Let $V_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b$ for $i = 1 \ldots K$ be the classifier outputs for the K events in one trial. We will sort them according to their classes by writing them as $V_i^C = V_i \cdot \mathbb{1}_C(i)$ where C denotes the set of indices for which class C was presented and $\mathbb{1}_C(i)$ is the indicator function that is 1 if $i \in C$ and 0 else. With these definitions in place, the winner of a trial can be defined as the class that has the largest average classifier outputs.

$$C_{\text{win}} \in \arg\max_C \frac{1}{|C|} \sum_{i=1}^{K} V_i^C \tag{2.28}$$

An example is shown in Figure 2.11. Because V_i is real-valued, two winners should not occur. However, in that rare case, the winning class is randomly selected between the equal classes.

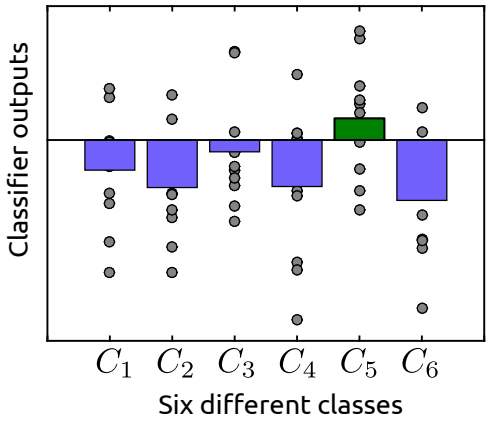


Figure 2.11: **Integrating evidence from one trial during postprocessing.** Classifier outputs (gray dots) are shown for 6 classes (C_1 to C_6). The class-wise averages are depicted in blue for the non-winning classes and in green for the winning class.

To limit the influence of a single event on the class-wise averages, another postprocessing step is typically applied in our data processing pipelines. Rescaling the projection $\mathbf{w}$ by dividing it by the projected difference of the class means $\mathbf{w}^{\mathrm{T}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ will enforce that the target and non-target class means are mapped to $+1$ and -1, respectively. With that, one can simply clip the classifier outputs which exceed a certain threshold, e. g., all classifier values whose absolute value is greater than 5 are set to $+5$ or -5 depending on the sign of the original value. This step can be seen as a simple and robust artifact treatment that is especially valuable in the online scenario where some artifact treatment methods are not directly applicable.

2.7 PERFORMANCE EVALUATION

Evaluating a machine learning model is an essential part to verify its applicability. A wide range of metrics have been previously proposed. They can roughly be subdivided into two categories.

Single-epoch classification. In the first case, it is quantified how well the model is able to distinguish single events (epochs) from one class (e. g., targets) from the stimuli of the other class (e. g., non-targets). Throughout this work, this *single-epoch classification accuracy* will be measured by using the area under the curve (AUC) of the receiver-operator characteristic (ROC) curve. To understand the concept of a ROC curve, two additional terms are introduced. The true positive rate (TPR), also called sensitivity, measures the proportion of instances from the positive class that are correctly detected as such. In contrast, the false positive rate (FPR) measures the proportion of negative data points that are falsely classified as positive examples. The goal is to maximize the TPR while minimizing the FPR.

Recalling the linear classification model from above as given in Equation 2.1, it becomes evident that the bias term is influencing the TPR and FPR, e. g., a very low bias will cause all instances to be classified as negative and thus, leading to a TPR and FPR of 0. While this minimized the FPR, the outcome for the TPR is unsatisfactory. To overcome the difficulty of choosing an appropriate bias term b, the ROC curve iterates over each possible bias term and shows the corresponding FPR and TPR. Five distinct ROC curves are shown in the left part of Figure 2.12. The AUC then measures the area under that curve. A higher AUC indicates better performance where an AUC of 1 represents perfect separability of the classes. The theoretical chance level of the AUC is 0.5, meaning that a random classifier should perform on this level. Systematic AUC values below 0.5 can be improved by simply switching the class allocations of the model. The blue line in Figure 2.12 represents a very good classifier, whereas the yellow line depicts a poor classifier.

The right part of Figure 2.12 shows the class distribution and threshold for three different bias terms. At point P1, both classes are fairly well separable and TPR and FPR were equally optimized. Point P2 is derived from the same distribution (green line) with a different bias. This bias leads to a high TPR (all positive examples are identified as such), but also to a high FPR (negative examples are often falsely identified as positive ones). At point P3, classes heavily overlap leading to a small discriminatory power, and hence, to a small AUC.

Trial-level performance:. Due to the low signal-to-noise ratio, one decision step (a trial) requires the collection of evidence from many single events. The single classification results are normally combined to obtain a prediction for a trial, see the previous Section 2.6 about postprocessing. Instead of quantifying the single epoch classification accuracies as before, the second metric, which is used throughout this paper, is quantifying the trial-wise performance. The trial-wise classification performance is measured by simply dividing the number of correct predictions (e. g., pre-

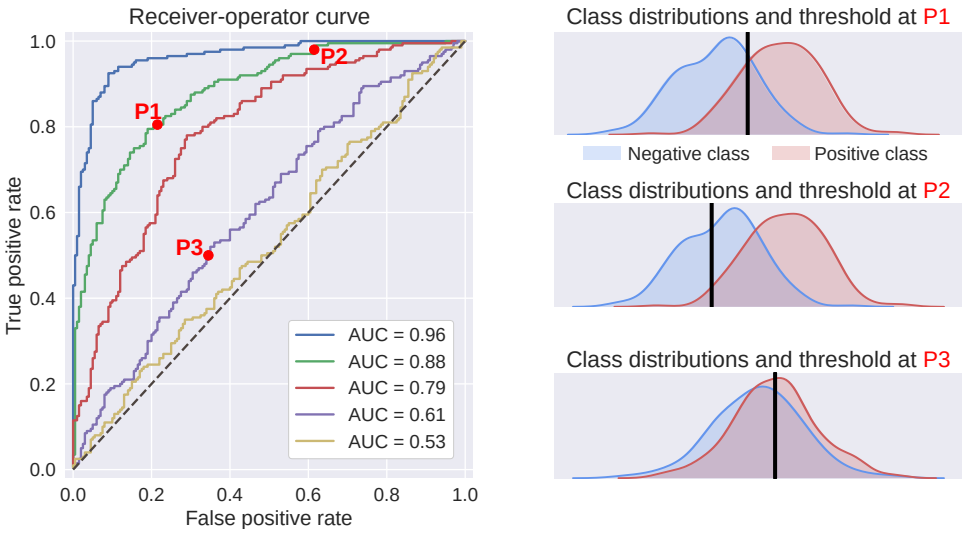


Figure 2.12: **Different receiver-operator curves and their underlying distributions and thresholds.** Plots are based on simulated data. The left plot shows five different ROCs while the right plots show the distribution and bias for three points on the ROC.

dicting the letter that the subject wanted to spell) by the total number of predictions. The theoretical chance level, in this case, is given by 1 over the number of possible choices assuming that each choice is equally likely.

2.8 APPLICATIONS

In the previous sections, I have explained how neural signals can be recorded and analyzed in a BCI. With this pipeline, the computer can infer basic knowledge about the ongoing brain state of a subject. This information can be used in various applications. In their book, Wolpaw and Wolpaw [180] identified five application areas of BCIs. A sixth area (basic research) was added by Brunner et al. in the BNCI Horizon 2020 road map [29]. The areas are:

1. **Replacement.** The traditional use-case of a BCI is to restore function that has been lost due to an injury or a disease. The most prominent example for replacement is the situation where a person lost his/her ability to speak. That person could use a BCI to type words which appear on a screen or are synthesized into sounds [20, 50]. This can enable basic communication, e. g., for patients with ALS [126, 155, 187] which is otherwise not feasible. Another use-case is for patients that lost limb control and use a BCI to operate a motorized wheelchair [112].

2. **Restoration.** BCIs can also restore function, e. g., by replacing lost or faulty neurobiological pathways. An example is the case where a tetraplegic subject, who had no sensory or motor control over his arm, could execute basic reaching and grasping movements through brain-controlled muscle stimulation of his arm [2].

3. **Enhancement.** BCIs can also be used for enhancing the execution of a task by monitoring the ongoing mental state of a subject and adapting the application/device accordingly. For instance, it was demonstrated that a BCI can be used to detect the intention to execute an emergency brake in a driving scenario [22]. This information could potentially be used to actually execute the brake and save valuable time. The detection of fatigue during driving could also potentially be used to inform the driver that he/she should make a break [24].

4. **Supplement.** BCIs can be used in a scenario where the natural neuromuscular output is augmented by an artificial output, e. g., in a scenario where a user controls a third (robotic) arm with a BCI.

5. **Improvement.** The principal goal of BCIs have been medical application. In recent years, many studies have demonstrated how BCIs can be used for motor rehabilitation after stroke. In these studies, users receive immediate sensory feedback (e. g., via functional electrical stimulation [15]) if a motor attempt was decoded. This should reinforce the interaction between efferent and afferent pathways of the brain and has shown medium-to-strong training effects [31].

   Other groups have attempted to use BCI neurotechnology for cognitive assessment and rehabilitation, see the review by Carelli et al. [30]. In this field, however, it is vastly unclear, which neural markers are

sufficiently specific to realize brain-state dependent trainings. The few existing studies mostly target attention in patients with ADHD (with inconclusive results [177]) and general cognitive functions in elderly subjects with a moderate effect [111].

6. **Basic research.** The real-time closed-loop interaction between brain and computer provides a new tool to investigate basic scientific questions. An excellent example of BCIs for basic research is the work by Schultze-Kraft et al. [154]. In their research, subjects had to press a button. Once their intention to press that button was detected in the EEG based on the readiness potentials (RP), subjects should abort their attempt to press that button. This closed-loop setup allowed investigating questions regarding the ability to veto a movement and to link the RP to movements.

2.8.1 *ERP-based applications*

Most results in this thesis focus on ERP-based paradigms. Visual ERP-based BCIs have several desirable features [52, 180]: (a) they require almost no subject training, (b) can be realized with standard hardware, (c) have a high user acceptance, (d) generally need less than 10 minutes to be calibrated [52] and (e) are effective for almost all healthy users [64] and for many patients with ALS [126, 155, 187]. Overall, BCIs based on visual ERPs are widely used, even though faster alternatives exist in terms of information transfer [59]. Examples for faster paradigms are code-modulated visual evoked potentials (c-VEP) [16, 168] and paradigms based on steady state visual evoked potentials (SSVEP) [36]. The c-VEP and SSVEP approaches, however, require a high temporal precision of the visualization hardware and a high level of gaze control. SSVEP stimulation can be perceived as a high workload and — due to its flickering characteristics — may even evoke seizures in epileptic users.

A wide range of applications exists that are based on visual ERPs [59], e. g., for spelling [20, 50], web browsing [12], games [1, 131], browsing and sharing pictures [165], predicting emergency brakes in a driving scenario [22], controlling objects in a virtual environment [10, 63] and artistic expression through painting [126, 187].

In contrast to visual BCIs, auditory BCIs generally suffer from a lower SNR. Nevertheless, there was burst of activity of BCIs relying on auditory stimuli in the last decade [9, 40, 58, 65, 66, 70, 71, 77, 91, 98, 100, 114, 128, 130, 141, 147, 152, 153, 157, 163, 166, 182, 186]. Besides their use-case for communication, there is also an important line of research that utilizes auditory BCI for brain-state assessment of patients with disorders of consciousness [141, 147, 182]. These approaches explore the idea that BCIs can detect residual brain activity even for patients where it is unknown whether they are conscious.

In Chapter 4, I will show a new application for auditory ERPs in which stroke patients use a BCI for training their language abilities.

3

UNSUPERVISED LEARNING FOR ERP-BASED BCIS

The following chapter is mainly taken from the two journal publications in PLOS ONE [80] and IEEE Computational Intelligence Magazine [82] and from a BCI Journal publication that is currently under review [76]. The content is partly copied and partly rewritten and condensed to allow for a more expedient presentation of the material. In addition to the previously published material, I provide a new analysis where the new unsupervised machine learning methods are tested in simulations on challenging data from patients with post-stroke aphasia performing an auditory ERP protocol with word stimuli in Section 3.5.

ABSTRACT

One of the fundamental challenges in BCIs is to tune a brain signal decoder to reliably detect a user's intention. While information about the decoder can partially be transferred between subjects or sessions, optimal decoding performance can only be achieved with new data from the current session. Instead of conducting a time-consuming calibration recording prior to each BCI usage, it is preferable to learn the brain signal decoder from unlabeled data gained from the actual usage of the BCI application. This also has the advantage that an adaptive model can learn to cope with changing distributions in the data over the course of a session and that it can continuously improve when more unlabeled data is recorded.

I present two new unsupervised learning methods for ERP-based BCIs which can learn without label information: learning from label proportions (LLP) is a conceptionally simple approach that relies on the existence of different subgroups in the data with different label proportions. These subgroups naturally exist in some application or can be created in others. They are then utilized to estimate the target and non-target class means which are used in a linear classifier. Given independent and identically distributed (IID) data, it is the first unsupervised classifier for BCIs that is guaranteed to converge to the optimal classifier.

Together with colleagues from TU Ghent and TU Berlin, we then developed a second algorithm (called MIX) which combines the strengths of LLP and an expectation-maximization (EM) algorithm. Two online EEG-studies where healthy volunteers controlled a modified visual speller, confirmed that both novel algorithms work well in practice. Remarkably, the MIX method not only defeats its two unsupervised competitors (LLP and EM), but — after a short ramp-up — even performs on par with a state-of-the-art regularized LDA classifier trained on the same number of data points and with full label access.
As additional verifications, the performance of the new unsupervised methods was tested in two more applications. First, I show that it can work in a BCI chess application without changing the user interface. Second, the unsupervised classifiers were simulated on data from post-stroke aphasia patients performing an auditory ERP paradigm with word stimuli. Although the number of classes is smaller in that task, this data is much more challenging because of a lower signal-to-noise ratio. Simulations show that brain signals are still reliably decodable even without calibration with a purely unsupervised approach.
This research demonstrates that a synergistic design between the user interface and machine learning algorithm opens the door for previously unseen performances. It paves the way for a transition from supervised to unsupervised learning methods in ERP-based BCIs.

Motivation

Many applications in the field of human-device interaction need a calibration phase prior to the actual usage of the application. During calibration, the user is requested to perform a series of predefined tasks in order to collect example data, for which the user's intentions are known. Machine learning methods then use this labeled data to learn the subject-specific brain signal characteristics and predict the user's intentions on new unseen data. This can be used for different applications, see Section 2.8.

Calibration is challenging in BCIs, because the SNR is unfavorable and the subject-to-subject variability is large [180], see for instance Figure 2.9 from the previous chapter which illustrates the large variability even in healthy subjects. Depending on the type of paradigm chosen, the calibration time can differ between minutes [52] to multiple sessions [103]. Even though it was shown that the calibration time can be partly reduced by transferring brain signals from within the same subject [102] or other subjects [53, 54, 87], a rest of subject- and session-specific variation remains to be learned.

To make this problem even more difficult, we can sometimes observe that brain signals change over the course of one single session. An example

is shown in the Figure 3.1 below where I show the P300 amplitude from one aphasia patient (please see Chapter 4 for more details) in channel Cz over the course of a session. One can clearly observe that the P300 amplitudes is (1) showing a high level of variance across sessions and (2) that the P300 amplitudes shows a clear increase from the beginning of a session (run 1) to around run 6 which is approx. 30 minutes into the session. With that, it might even be necessary to (re)-calibrate the classifier within one session. It is almost impossible to adapt to the very fast (and unpredictable) fluctuations in the data distribution (e. g., from run 13 to run 17 in Figure 3.1), but an adaptive classifier should be able to learn these slower trends (e. g., from run 1 to run 7).

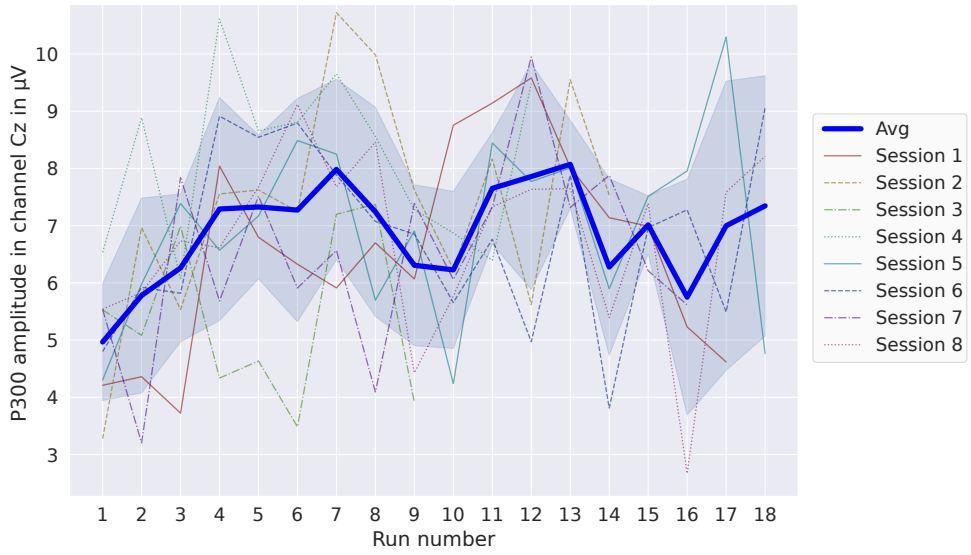


Figure 3.1: **P300 development during different sessions of the same patient.** The thick blue line and shaded area show the average ERP response and standard deviation across sessions, respectively. The thinner lines show the ERPs for single sessions for a total of 8 different sessions.

To tackle these challenges, different learning strategies have been proposed. They can be subdivided into two groups: the first group takes a pre-trained classifier and updates it with unlabeled new data from the current session [23, 38, 39, 118, 133, 176, 185]. We refer to this approach as *unsupervised adaptation*. Algorithms implementing unsupervised adaptation rely on the assumption that suitable training data is available or can be recorded in order to pre-train a classifier. However, for subjects with limited attention span or atypical brain patterns, e. g., stroke survivors, this might not be the case. To overcome this limitation, a second group of algorithms was recently proposed for BCIs. These algorithms can learn the individual brain characteristics from scratch without requiring any labeled data at all [61, 62, 80, 84, 95, 98, 174]. We refer to them as *unsupervised learning* methods. They are a generalization of the first group of algorithms as they can also be initialized with good parameters obtained via transfer learning. See Figure 3.2 for an illustration of the difference between the two groups.

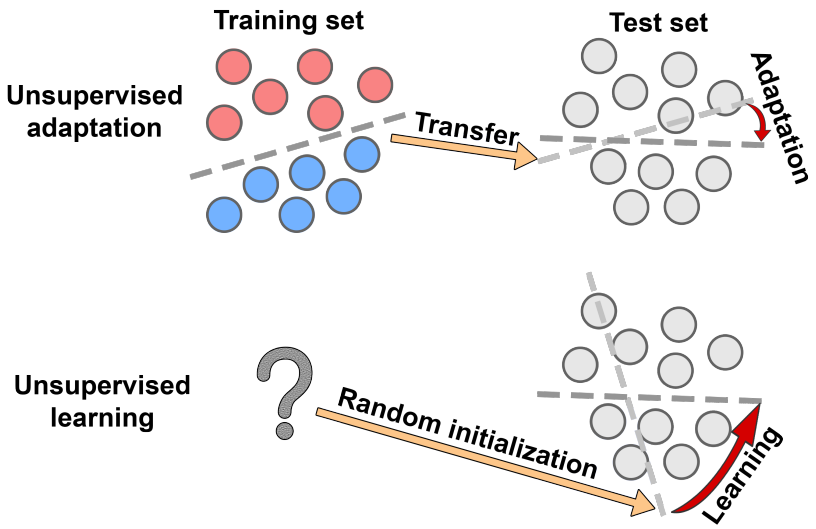


Figure 3.2: **Unsupervised learning and unsupervised adaptation.** Red and blue dots indicate historic labeled training data from two classes. Grey dots depict unlabeled data. Dashed lines indicate classification models. The general goal is to find a model which separates the two classes as good as possible. Label information is necessary only in the adaptation scenario. For transferring the classification model, only a slight adaptation may be necessary while the unsupervised learning algorithm has to learn the model from a random initialization. Figure taken from [82].

Both approaches are able to update their decoding model during the actual usage of the BCI application and hence, are able to adapt to changing brain signals over time. To accomplish this, the ML method is required to learn from unlabeled data, i. e. when the user's intentions are unknown. Unsupervised learning bears the potential of exploiting large unlabeled data sets to find common brain patterns — a key ingredient for developing true plug & play BCI systems.

A different line of work explores strategies to adapt the policy of the interaction between user and computer instead of adapting the brain signal decoder [33, 34, 83, 85, 149, 184]. These *policy adaptation* approaches generally rely on the detection of error-related potentials, i. e. signals that reflect the observation of an error, in order to infer the correct or intended actions of the user.

ERP-based BCIs generally facilitate unsupervised learning.

In the previous chapter, I have introduced the basics of ERPs in Section 2.3 and their BCI applications in Section 2.8.1. Importantly, visual ERP-based BCIs often have the advantage that the stimulus presentation mode leads to a special structure of the collected brain signal data, which can be exploited by unsupervised learning methods.

For instance, in the case of the well-known P300 speller by Farwell and Donchin [50], the user can select to spell between 36 symbols which are arranged in a 6×6 grid by focusing his attention on the target letter,

see Figure 3.3**A**. Rows and columns are then highlighted in alternating order. A complete highlighting round of 12 events is called an *iteration*. Typically, a trial consists of multiple iterations to uniquely determine the attended character. This highlighting scheme is inducing constraints on the data, e. g., exactly one row and one column of the symbol grid will contain the selected letter while five rows and columns do not contain it. Also, knowing the selected symbol uniquely determines each event as being a target or non-target symbol, see Figure 3.3**B**. These and more constraints allow for efficient learning from unlabeled data in ERP-based BCIs, something which is not yet sufficiently explored in the oscillatory domain, see for instance [115]. In the following sections, I will first review other approaches to learn from unlabeled data before presenting two new approaches that heavily rely on the rich structure of ERP-based BCIs.

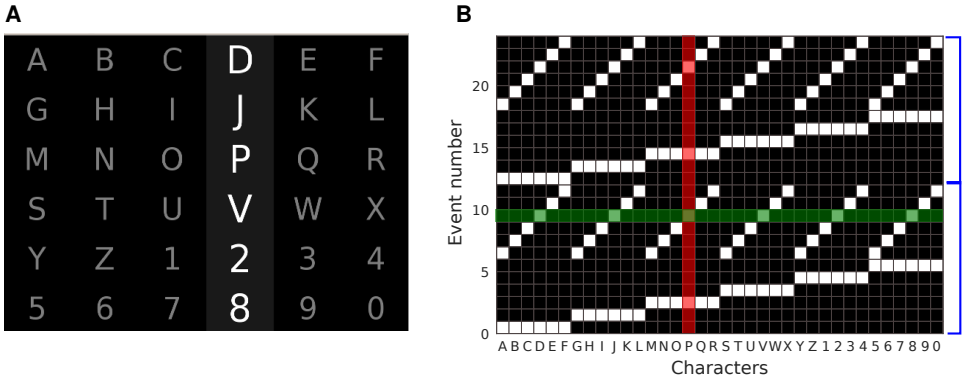


Figure 3.3: **Visual spelling matrix and flash groups of a row-column speller.** The left subplot (**A**) shows the spelling interface of the classical speller by Farwell and Donchin [50]. The right subplot (**B**) shows the flash groups per event where white squares indicate characters that were highlighted for a certain event and black squares indicate those which are not highlighted. The green horizontal bar reflects the highlighting event of the left subplot. Knowing the target letter 'P' uniquely determines all target events (see vertical red bar). The blue bracket on the right shows one full iteration.

3.1 REVIEW OF PREVIOUS APPROACHES

Different attempts have been undertaken to accomplish learning from unlabeled data in ERP-based BCIs. A review of examples from the group of *unsupervised adaptation* techniques (sometimes also referred to as semi-supervised [115] methods) is presented before discussing *unsupervised learning* approaches. It should be emphasized that all unsupervised methods can be used for an ordinary visual P300 speller unless specified otherwise. The ML model is hidden from the user such that the interaction between user and computer remains the same except for the decoding quality of the control signals.

3.1.1 *Unsupervised adaptation for ERP protocols*

Unsupervised adaptation always relies on a classifier that has been pre-trained on supervised data from the same or other subjects. For transferring it to a novel user or to the next session, the pre-trained classifier is then adapted using unlabeled data gained during the usage of the BCI application. An overview of currently published methods is given in the top part of Table 3.1.

Table 3.1: **Overview of unsupervised adaptation and unsupervised learning methods for ERP-based BCIs.**

Unsupervised adaptation
1) Naïve labeling with adaptation based on predicted labels: *Lu* 2009 *[118]; Kindermans 2011 [96].*
2) Co-training two classifiers based on predicted labels: *Panicker 2010 [133].*
3) Usage of error-related potentials as label information: *Zeyl 2016 [185].*
4) Pooled mean & covariance adaptation disregarding labels: *Vidaurre 2011 [176]; in ERP: Dähne 2011 [38].*
5) Alternatively estimating CSP and Riemannian classifier: *Barachant 2014 [5]; in MEG: Bolagh 2016 [23].*
Unsupervised learning
1) Exploiting task constraints and error-related potentials: *Grizou 2014 [61, 62]; Iturrate 2015 [84].*
2) Utilize data constraints with expectation-maximization (EM): *Kindermans 2012 [95]; Kindermans 2014 [98].*
New unsupervised learning approaches (presented in this thesis)
1) Modify paradigm to learn from label proportions (LLP): *Hübner 2017 [80].* See Section 3.2.
2) MIX: Combine the mean estimations from EM and LLP: *Verhoeven 2017 [174]; Hübner 2018 [82]. See Section 3.3.*

Naïve labeling

Lu et. al [118] proposed an approach in which a subject-independent classifier is first trained on historic data and then used to predict the labels for newly recorded ERP signals. Assuming that these predictions are correct, the model is then retrained with the new data to obtain an updated classifier. Obtained labels are called “naïve" as it is uncertain whether they are correct or not. To measure the degree of uncertainty, Lu et al. introduced a confidence score that is measuring how consistently the labels were predicted during the spelling of one letter. Only when a high consistency is observed, they trusted an estimated label. Otherwise, the unlabeled data was discarded. While their approach worked well in an offline study using a visual spelling paradigm with 10 healthy subjects, it can be expected to have severe problems when the initial accuracy is close to chance level, e. g., in patient data or in auditory ERP data with a low SNR. In this case, the instability of the labeling can cause runaway errors [105]. The self-labeling approach was also used by Kindermans et al. [96] for a class re-weighted version of the Ridge regression and was shown to outperform a non-adaptive classifier on the BCI Competition III data set [18].

Two-classifier co-training approach

Panicker et al. [133] extended the idea of naïve labeling using two classifiers — Fisher linear discriminant analysis and Bayesian linear discriminant analysis — which co-train each other. To do so, both classifiers are first initialized on a labeled training data set. Then both classifiers determine the labels for a chunk of unseen and unlabeled data points. These points with corresponding estimated labels are then added to the current training data set of the *other* classifier and both classifiers are retrained. This procedure is repeated until convergence or until the improvements (measured by a confidence score) are minimal. The authors evaluated this approach using an offline visual ERP speller study with data from five healthy subjects. For this relatively small number of subjects, it was found that the co-training approach outperforms the naïve labeling strategy of a single classifier in most situations, however, runaway errors may still occur.

Pooled mean and covariance adaptation

Vidaurre et al. [176] suggested an unsupervised adaptation method of an LDA classifier. Vidaurre et al. also utilized the equivalence of the two LDA formulations based on the local and global covariance matrix, see Equation 2.21. They proposed an adaptation scheme which adapts either only the common class means or both, the class means and the global covariance matrix in an unsupervised fashion. This approach was shown to outperform a fixed supervised classifier on motor imagery data both in simulations and online. It can readily be applied to ERP data as demonstrated in [38].

Adaptation based on error-related potentials

When the user perceives a mistake, e. g., when an incorrectly spelled letter was shown to the user, a time-locked error-related potential (ErrP) can be observed. These ErrPs can be decoded with an accuracy of around 80% [35, 55, 113] and — depending on the application — may be useful to automatically correct detected errors [122]. Initially proposed for code-modulated visual evoked potentials, Spüler et al. [160] proposed to ignore the data, if an ErrP is detected after showing the predicted character since the true class label is unknown and the estimated class label is suspected to be wrong. Other groups used ErrPs to adapt the policy of a virtual or real robot in order to achieve a certain goal [33, 34, 83, 85, 149, 184]. In Section 3.1.2.1, we review an approach that can jointly learn to decode ErrPs and to adapt its policy to control a device.

Recently, Zeyl et al. [185] compared an adaptation of the decoder based on (a) ErrPs, (b) a naïve-labeling approach based on target confidence and (c) a hybrid approach which combines (a) and (b) in a visual ERP speller. The problem with exploiting ErrPs in the context of the classical visual ERP speller is that feedback signals are only shown at the end of each trial, and hence, ErrPs are harvested rarely compared to the number of presented stimulus events. To alleviate this mismatch, Zeyl and colleagues proposed to show both the row and the column selection as two separate decisions to the user to collect ErrPs more frequently. Interestingly, an offline analysis and a simulated online experiment with 11 healthy subjects showed that the naïve-labeling approach performed best, with the hybrid approach close behind and the pure ErrP approach significantly worse. This indicates that additional information from the ErrPs could not contribute in improving the adaptation in this specific experimental scenario.

Alternatively training a spatial filter and Riemannian classifier

Barachant and colleagues proposed an information theoretical framework which allows measuring distances between trial covariance matrices based on concepts of the Riemannian geometry [6]. The use of this representation and Riemannian distance has the advantage of being invariant under affine transformations which would not be the case in the original Euclidean space. Supervised classifiers operating on Riemannian distances have been successfully applied to ERP signals [5]. Although mentioned as an option, unsupervised adaptation was not implemented in their work on EEG-based ERP data [5], but it was implemented successfully on magnetoencephalography (MEG) data by authors around Bolagh from the same group [23]. Again, the premise is that labeled historic data from earlier subjects is available which is used to obtain an initial estimate of the novel unlabeled data.

An iterative two-step procedure for estimating these labels is at the core of their approach. It makes use of a widely-used spatial filtering method, common spatial patterns (CSP) [19]. As this algorithm requires labels, which are not available in an unsupervised adaptation approach,

the current label estimates are used in every iteration of the procedure. The first step involves to replace the original trials by new "super trials". These are formed by CSP-filtered original trials, enlarged by the two CSP-filtered class means. Super trials are then used to calculate the so-called feature covariance matrices (one per trial). The second step takes place in Riemannian space, where distances between these novel feature covariance matrices and mean covariance matrices can be computed. A Riemannian classifier based on labels of the last iteration (or on labels of historic data in case of the first iteration) is used to update the label estimate of each trial. These two steps are repeated until convergence.

This approach won the open "DecMeg2014" Kaggle competition. It could easily be transferred to EEG-based ERP data.

3.1.2 *Unsupervised learning for ERP protocols*

I now address the second group of classifiers, *unsupervised learning approaches*, which can learn the model parameters without requiring any labeled data at all, not even historical data. Compared to the approaches described in the previous section, this type of learning is substantially more difficult as no initialization or prior information of the parameters are available.

Assuming a two-class problem with high SNR, one could imagine an obvious approach: applying a clustering algorithm would allow splitting the data into two groups, e. g., by assuming a Gaussian distribution of each class. One could then further identify the two clusters as target and non-target classes by utilizing structure imposed by the experimental design. In case of ERP paradigms, fewer data points can be expected in the cluster formed by target points compared to the non-target cluster.

However, given the low SNR in ERP-based EEG recordings, this obvious approach would require an enormous number of data points. Practically, it is not feasible. Instead, unsupervised learning methods need to exploit the data constraints provided by the ERP application as good as possible. Only this information allows them to solve the classification task despite the low SNR and missing labels.

Two algorithms that implement unsupervised learning are reviewed: (1) an approach combining task constraints with ErrPs and (2) the probabilistic expectation-maximization algorithm. An overview is shown in the middle part of Table 3.1.

Exploiting task constraints and error-related potentials

The calibration-free approach by Grizou [61, 62] and Iturrate [84] is able to simultaneously calibrate the system while the user controls the BCI by making intelligent use of the given task constraints and ErrPs. The authors demonstrated the feasibility of the approach on a virtual 5×5 grid where the user should move a cursor to a goal position [84]. Users achieve control by monitoring the moving cursor and passively assess whether it moves

in the right or wrong position. In the latter case, an error-related potential is automatically elicited by the user. Detecting those ErrPs would allow a BCI controller to determine the goal position. Now, the learning task is to simultaneously infer the unknown goal position as well as to train an ErrP decoder. This chicken-and-egg problem is solved by utilizing the observation that each of the 25 possible goal positions should lead to a different sequence of elicited ErrPs, thus providing only 25 possible ways to label the ErrP data. Their algorithm then assigns a higher likelihood to data sets that are most consistent, where consistency is measured as the separability between the two classes (correct or incorrect direction). The goal position desired by the user is the one associated with the most consistent data set, which can, in turn, be used to update the parameters of the ErrP classifier. An online study with eight healthy subjects showed that this method allowed users to correctly navigate the cursor to more goals compared to a scenario where a supervised adaptation was conducted prior to the experiment and with the same total experiment time. Although their navigation problem is formulated in a grid shape, this technique was not yet applied to any ERP-based spelling paradigm, see [34].

Expectation-maximization

The approach by Kindermans and colleagues [95] also simplifies the overall learning task by trying to infer the latent variable (selected symbol) of a matrix speller rather than solving the more complicated problem to decide for each stimulus whether it was a target or non-target. This reduces the number of possible configurations from an exponentially growing number in the latter case, to a limited one — 36 possible letters in the case of the original visual ERP speller — per trial, see Figure 3.3**B** from before. Importantly, the number of possible configuration only depends on the grid size, and does not change when more iterations are recorded to spell one character. With this constraint in mind, Kindermans et al. proposed to use a version of Bayesian least square regression [17, 95] which assumes that the feature vectors can be linearly projected onto two one-dimensional Gaussian distributions (one for targets and one for non-targets), which share the same within-class variance. This original approach directly computed the projection vector $\mathbf{w}$ and the within-class variance β.

The learning task is tackled by utilizing an expectation-maximization (EM) algorithm which alternatively estimates the probabilities of the latent variables — which letter was selected by the subject — during the expectation step [E-step] and optimizes the parameters given these probabilities in the maximization step [M-step]. The EM procedure is repeated until convergence. In comparison to a direct optimization of the model parameters, the iterative procedure of the EM algorithm has fewer parameters that are optimized in each step which solves the intractability of a direct optimization. EM can be seen as a mathematically rigorous version of the naïve labeling approach from before and realizes the maximum likelihood estimator.

In the work by Kindermans *et al.* the projection vector $\boldsymbol{w}$ is directly estimated with EM, thereby automatically optimizing the regularization parameter λ for the pooled covariance [95]. An online study with 10 young healthy users showed that the EM algorithm can successfully decode auditory ERP signals from scratch without any label information [98]. Given a sufficient amount of data, the EM approach can compete with a supervised classifier. In cases when non-stationarities occur in the data [98], the EM has the potential to outperform a non-adaptive supervised classifier. Nevertheless, during the first spelled symbols performance is not satisfactory and highly dependent on parameter initialization. This so-called warm-up period is the main claim against the EM method as it is discouraging for the user and as such shows the same disadvantages as a calibration procedure. It was demonstrated that transfer learning can alleviate these problems [97].

3.2 NEW APPROACH 1: LEARNING FROM LABEL PROPORTIONS

In Section 2.5.3 in the previous chapter, I have shown that the unsupervised learning problem boils down to estimating the target and non-target class means, $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_N$, respectively. These quantities and the pooled covariance matrix, which can be computed without label information, are sufficient to obtain the optimal projection $\mathbf{w}$ for ERP classification. This reformulation of the unsupervised learning problem is one of the keys to enable the following approach.

Please note that the bias term b is mostly ignored in this thesis, because in most ERP applications only the relative and not the absolute scores of the classifier outputs between the classes are compared, see Section 2.6. In the case of relative scores, the bias term cancels out. Also, the bias is not relevant for the AUC performance because the AUC is computed by iterating over all possible bias terms. In all experiments, we simply set $b = 0$.

I will now explain how learning from label proportions (LLP) [145] — a simple, yet very powerful idea — can be used to estimate the class means. It is applicable in cases where the data contains groups that have different label proportions. The idea will be presented on an intuitive level with Figure 3.4 before formally introducing LLP. Most of the section is taken from the PLOS ONE journal publication [80].

To enable LLP in a visual ERP speller, two new paradigm modifications are necessary. First, the normal highlighting matrix is extended by adding "#" symbols. These symbols should not be spelled by the user and as such, are non-targets by definition. Second, two sequences are interleaved per trial (see Figure 3.4**A**) resulting in two subgroups in the data. Events from S1 highlight only ordinary symbols while events from sequence S2 also highlight "#" symbols. This means that the sequences S1 and S2 are both composed of some target and some non-target events, but S2 has a higher non-target ratio compared to S1 (see Figure 3.4**B**). Hence, we can write the average response of S1 ($\boldsymbol{\mu}_1$) and of S2 ($\boldsymbol{\mu}_2$) as a function of the target and non-target class means (see Figure 3.4**C**) with different label proportions which represent the coefficients in the linear system. The exact proportions are not yet important, they can be chosen by constructing the sequences and will be explained below. It is important, however, that they are different and known for both sequences and that it is possible to estimate $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ without requiring any label information by simply averaging all events from the corresponding sequence. The final step to obtain the target and non-target class means is then obvious: all one has to do is to solve the linear system of two equations and plug in the estimates of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ (see Figure 3.4**D**).

3.2.1 *Methods*

The LLP idea will now be formalized. Consider a two-class problem and G groups of data where each group is a mixture of two classes (e. g., targets

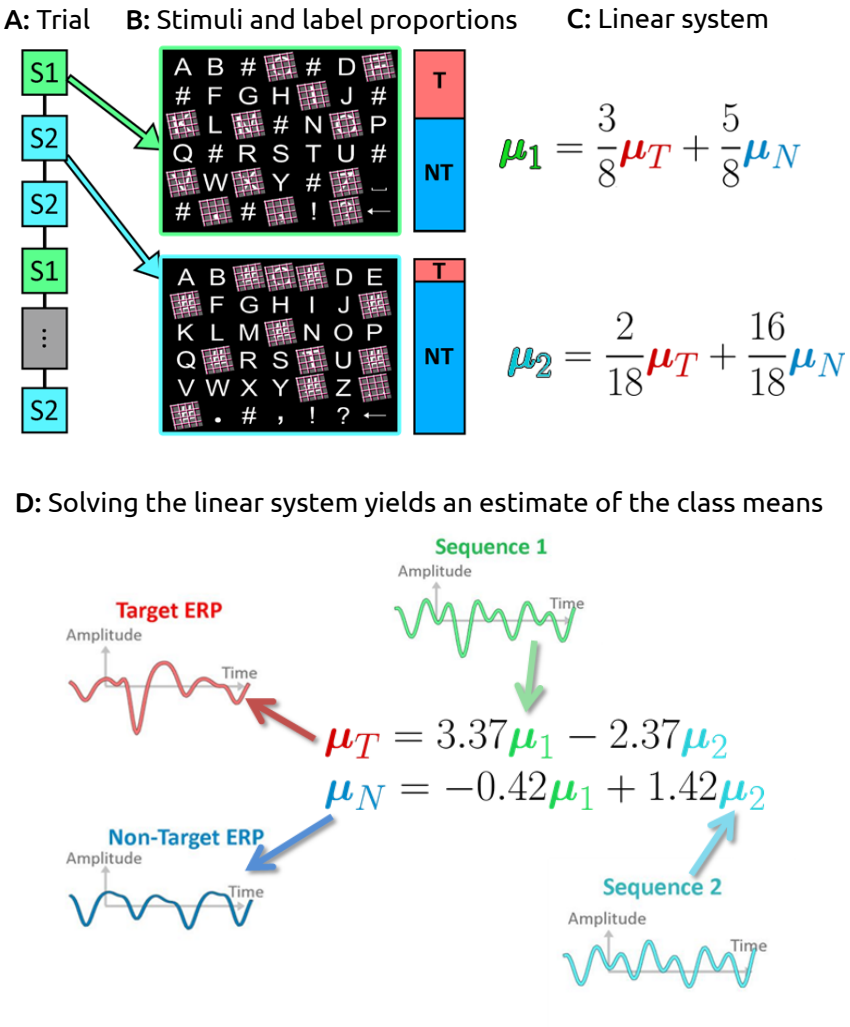


Figure 3.4: **Basic principle of learning from label proportions.** See the description in the text for an explanation. Figure adapted from [79].

and non-targets) with known mixture ratios contained in the matrix $\mathbf{\Pi}$. The means of the feature vectors in the groups $\mu_1, \mu_2, \ldots, \mu_G$ can then be expressed as a function of the class means μ_T, μ_N as follows.

$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_G \end{bmatrix} = \mathbf{\Pi} \begin{bmatrix} \mu_T \\ \mu_N \end{bmatrix}, \quad \mathbf{\Pi} := \begin{bmatrix} \pi_T^1 & \pi_N^1 \\ \vdots & \vdots \\ \pi_T^G & \pi_N^G \end{bmatrix} \tag{3.1}$$

To obtain an empirical estimate of the group means $\mu_1, \mu_2, \ldots, \mu_G$, we do not need label information. These quantities can then be used to approximate the class means by using the pseudoinverse of $\mathbf{\Pi}$, given by $\mathbf{\Pi}^{-1} := (\mathbf{\Pi}^T\mathbf{\Pi})^{-1}\mathbf{\Pi}^T$. Using this notation, the class means are computed as

$$\begin{bmatrix} \mu_T \\ \mu_N \end{bmatrix} = \mathbf{\Pi}^{-1} \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_G \end{bmatrix}. \tag{3.2}$$

Hence, by solving the resulting system of linear equations, one can get an estimation $\hat{\boldsymbol{\mu}}_T, \hat{\boldsymbol{\mu}}_N$ of the true class means $\boldsymbol{\mu}_T, \boldsymbol{\mu}_N$. The implicit *homogeneity assumption* in this formulation is that $\boldsymbol{\mu}_T, \boldsymbol{\mu}_N$ are the same for each group, i. e. the feature distributions for target and non-target samples are independent of the sequence. I will later present data that shows that this assumption is justified. Additionally, there also exists a version of this algorithm using a manifold regularizer that performs better than this version under a violation of the homogeneity assumption [136].

So far, we have obtained an estimation of the target and non-target class means. From this, a classifier can be obtained by multiplying the inverse of the estimated pooled (global) covariance matrix (see Equation 2.27) with the difference of the class means. As shown in Equation 2.21, the resulting orientation of the hyperplane is identical to the one obtained by the regular (supervised) LDA approach.

Guaranteed convergence. One important property of LLP is that it is guaranteed to converge to the optimal classifier. The proof is simple and relies on the central limit theorem. Because of the equivalence given in Equation 2.21, it is sufficient to show that the estimated class means converge to the true class means. Let x_k^i denote the i-th feature of the k-th feature vector $\mathbf{x}_k$. If we assume that each feature instance $x_1^i, x_2^i, \ldots, x_N^i$ is drawn independently from an identical distribution (IID) with finite expected value μ^i and variance σ_i, then the central limit theorem states that the sample average $\hat{\mu}^i$ is normally distributed for large N.

$$\hat{\mu}^i := \frac{x_1^i + \ldots + x_N^i}{N} \sim \mathcal{N}(\mu^i, \frac{\sigma_i^2}{N}) \tag{3.3}$$

This implies that the estimators $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_G$ converge to the true means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G$. After solving the linear system, we therefore have an estimation of the class-wise means which is *guaranteed to converge* for $N \to \infty$. Hence, we have an unsupervised classifier that, under the assumption of IID and homogeneity, is guaranteed to deliver the optimal classifier. As explained in Section 2.4.4, this optimality is with regard to the 0-1-loss (Bayes classifier), Fisher criterion and least squares with rescaled labels.

3.2.2 *Paradigm modification*

In the introduction to LLP, I have stated that the exact label proportions can be chosen by constructing the sequences. I will now explain this process. To generate two different sequences (S1, S2) in the data ($G = 2$), the stimulus presentation paradigm by Verhoeven et al. [173] was used. This paradigm is flexible in the sense that it can generate sequences with a desired mixture ratio of target and non-target stimuli and it can highlight other subsets than just rows and columns. At the same time, it uses a heuristic to increase the SNR in the stimulus responses by avoiding the two most common spelling errors: adjacency distraction (i. e. neighboring fields are highlighted) and double flashes (i. e. fields are highlighted twice in a row).

Stimuli from S1 highlight each character exactly 3 times for every 8 stimuli. This means that no matter which character the user is focusing on, we obtain 3 targets and 5 non-targets in this train of 8 stimuli. However, the decoding method and sequence generator are of course unaware about the exact target positions, i. e. where the 3 targets are located within these 8 stimuli. Similarly, sequence 2 contains trains of 18 highlighting events where each character is only highlighted twice. This leads to a ratio of 2 targets and 16 non-targets out of 18 stimuli in the second sequence. With that, it should have become evident how the numbers in Figure 3.4**C** and **D** arise.

A few additional measures were taken to comply with the assumption that ERP responses are distributed identically and homogeneously within each group. First of all, it is known that the response upon a stimulus event is influenced by its brightness and thus, by the number of symbols highlighted within that stimulus event [90]. With that, it is clear that the number of highlighted symbols should be the same for both sequences. Enforcing the above target and non-target ratios would not be possible in the original spelling matrix as S1 would highlight more symbols than S2 leading to different brightness levels. Hence, the original spelling was extended by adding 10 "#" symbols. These symbols should not be attended and therefore, serve as permanent non-targets. They are solely highlighted within sequence 2 to guarantee that the same number of symbols (12) is highlighted in both sequences and to ensure that the subject does not realize whether the current event is from S1 or S2.

Adding these symbols resulted in the necessity to also increase the overall spelling matrix. Hence, an additional column was added to increase the total number of symbols resulting in a 6 x 7 grid and 42 entries. In addition, to the 10 "#" symbols, the spelling matrix also contained all letters from the alphabet were included plus the symbols "␣" "." "," "!" "?" and "←.

The second precaution is to have sequences from both groups randomly interleaved within a trial. This avoids violating the homogeneity assumption, e. g., non-stationarity in the feature distribution within one trial or a modulation of the P300 amplitude because of differences in the target-to-target interval [60, 77]. This random interleaving leads to a second effect which is crucial for LLP. When averaging all events from one sequence, the average response to the previous ($t = -250\,\text{ms}$) and to the upcoming events ($t = 250\,\text{ms}$, $t = 500\,\text{ms}$, ... cancel out. This is important because otherwise the LLP would reconstruct a periodic signal. A more careful analysis of this aspect is given in the BCI chess Section 3.4.1.2 below.

3.2.3 *Study details*

To verify the applicability of LLP for BCIs, an online EEG study was conducted. Overall, 13 healthy subjects (P1-P13, 5 female, average age: 26 years, std: 1.5 years) were recruited. Only one subject (S2) had prior EEG experience. The EEG study was approved by the Ethics Committee of the University Medical Center Freiburg. Following the principles of the

Declaration of Helsinki, written informed consent was obtained from the subjects prior to participation. One session took about 3 hours (including EEG set-up and washing the hair), and participants were compensated with 8 Euros per hour.

The subjects were asked to spell the sentence: "FRANZY JAGT IM KOMPLETT VERWAHRLOSTEN TAXI QUER DURCH FREIBURG". The sentence was chosen because it contains each letter used in German at least once. Each subject spelled this sentence three times. The SOA was 250 ms (corresponding to 15 frames on the LCD screen utilized) while the stimulus duration was 100 ms (corresponding to 6 frames on the LCD screen utilized). For each character (trial), 68 highlighting events occurred and a total of 63 characters were spelled three times. This resulted in a total of $68 \cdot 63 \cdot 3 = 12852$ EEG epochs per subject. Spelling one character took around 25 s including 4 s for cueing the current symbol, 17 s for highlighting and 4 s to provide feedback to the user. Assuming a perfect decoding, these timing constraints would allow for a maximum spelling speed of 2.4 characters per minute. Figure 3.5 shows the complete experimental structure.

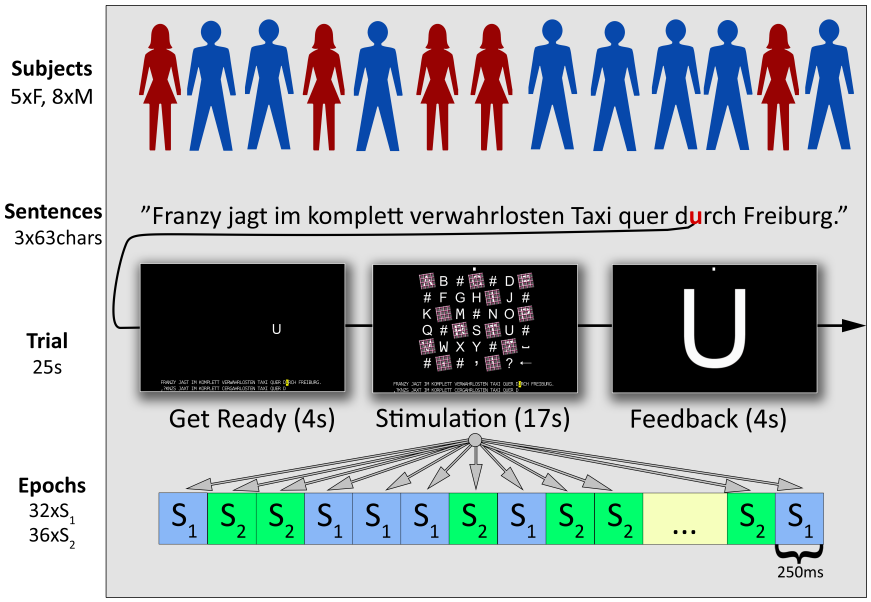


Figure 3.5: **Experimental structure of the LLP study.** Thirteen subjects were asked to write a sentence three times. A trial denotes the process of writing one letter. Each trial consists of 68 events where 32 events were part of sequence 1 and 36 events were part of sequence 2 (see text). Figure adapted from [78].

Subjects were placed in a chair at 80 cm distance from a 24-inch flat screen. EEG signals from 31 passive Ag/AgCl electrodes (EasyCap) were recorded, which were placed approximately equidistantly according to the extended 10–20 system, and whose impedances were kept below 20 kΩ. Please see Chapter 2 in the last chapter for a more detailed explanation about the EEG fundamentals. The signals were registered by multichannel EEG amplifiers (BrainAmp DC, Brain Products) at a sampling rate of 1 kHz. To control for vertical ocular movements and eye blinks, we recorded with

an EOG electrode placed below the right eye and referenced against the EEG channel Fp2 above the eye. The data of all 13 subjects is freely available online at http://doi.org/10.5281/zenodo.192684.

Data processing. The data processing closely follows the pipeline introduced in Section 2.4 in the previous chapter. To process the data in the online experiment and during offline re-analysis, the BBCI Toolbox was used [20]. In both cases, the collected data was bandpass filtered with a third order Chebyshev Type II filter between 0.5 and 8 Hz and downsampled to 100 Hz. Epochs were windowed to $[-200, 700]$ ms relative to the stimulus onset and corrected for baseline shifts observed in the interval $[-200, 0]$ ms. After dismissing channels Fp1 and Fp2, features describing the elicited transient potentials were extracted from the remaining 29 EEG channels. Per channel, the mean amplitudes of six intervals ($[50, 120]$, $[121, 200]$, $[201, 280]$, $[281, 380]$, $[381, 530]$ and $[531, 700]$ ms) were computed, resulting in a representation of each epoch by $6 \cdot 29 = 174$ features.

Classification. At the end of each trial, the LLP algorithm was applied on the complete set of observed responses in order to estimate the class means $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_N$ as explained in the previous Section 3.2.1. Additionally, the pooled (global) covariance matrix $\boldsymbol{\Sigma}_T$ on the combined data of both classes was estimated using shrinkage-regularization, see Equation 2.27 for the pooled sample covariance and the Section 2.5.2 about regularization. Based on the reconstructed class means and the pooled covariance matrix, the projection vector $\mathbf{w}$ was then computed as $\mathbf{w} := \boldsymbol{\Sigma}_{\mathrm{T}}^{-1}(\boldsymbol{\mu}_T - \boldsymbol{\mu}_N)$.

To select a symbol in each trial, the class with the highest average classification output was chosen, see Section 2.6. Please note that this decision does not depend on the bias term, because each symbol is highlighted the same number of times (except "#") and the same bias is summed up for each symbol. Thus, the relative ordering between the classes is not affected by the bias. Visual blanks ("#") were excluded from being chosen as selected symbols.

The classifier was reset and started from scratch for each of the three spellings of the sentence "FRANZY JAGT ..." in the online experiment. After collecting the data of a new character, the classifier was retrained. Label information (target / non-target role of characters) were used exclusively to evaluate the performance during offline analysis, but never to train the LLP classifier or generate the sequences during online use.

3.2.4 *Results*

Before investigating LLP, we first inspected the class-wise visual ERP responses to assess the quality of the data of the online study. They are provided as grand average responses in Figure 3.6. The rhythmic characteristic of the non-target responses generally reflects the SOA of 250 ms. We found a strong early negative ERP upon target stimuli over the occipital lobe (hereafter called N150) at around 150 ms for almost all subjects with an average amplitude of around $-8\,\mu$V. Please note that this is the same component that has been described in the previous chapter as

N200 in Section 2.3.2. For non-target stimuli, the N150 was very reduced. The late positivity of targets (hereafter called P300) in the central electrodes is rather late and weak with an average peak time around 400 ms and an average amplitude of only around 2 μV. Table 3.2 lists the amplitudes and peak latencies per subject observed for channels O1 (for the N150) and Cz (for the P300). By training a supervised shrinkage-LDA on this data set in an offline analysis and calculating the binary target vs. non-target classification accuracy based on a 5-fold chronological cross-validation, we obtained an average AUC of 97.5 % which indicates a very good SNR of the data set.

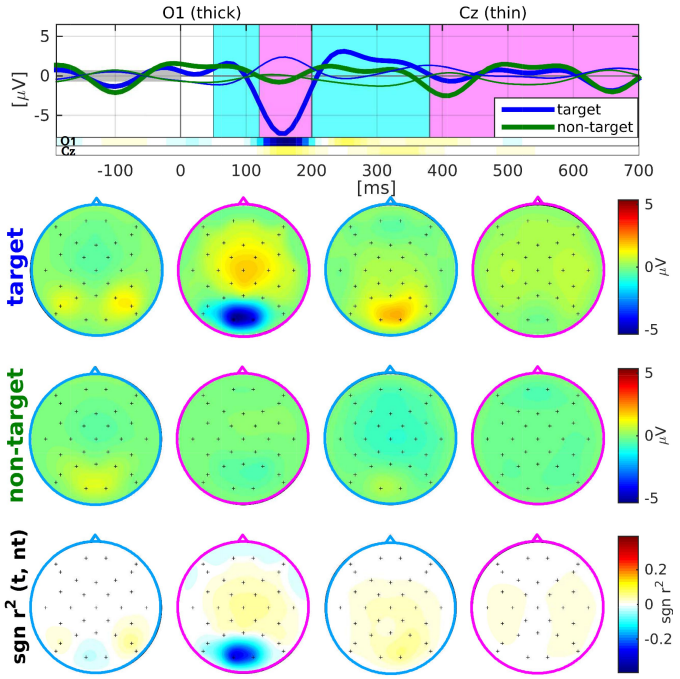


Figure 3.6: **Grand average (N = 13) visual ERP response of the LLP study.** Top row: average responses evoked by visual target (blue) and non-target (green) stimuli in the occipital channel O1 (thick) and the central channel Cz (thin) during the online experiment. Prior to averaging, a baseline correction was performed based on data within the interval $[-200, 0]$ ms. The signed R^2 values for channels O1 and Cz over time are provided by two horizontal color bars. Their scale is identical to the scale of the plots in the bottom row of scalp plots. Middle rows: scalp plots visualizing the spatial distribution of mean target and non-target responses within four selected time intervals: $[50, 120]$, $[120, 200]$, $[201, 380]$ and $[381, 700]$ ms relative to stimulus onset. Bottom row: scalp plots with signed R^2 values indicate spatial areas with high class-discriminative information. Figure taken from [80].

Reconstructed Means

After verifying that the data quality is good, we investigated if LLP could correctly reconstruct the mean target and non-target ERP responses, when

Table 3.2: **Overview of neurophysiological features and supervised classification performance.** The amplitude and latency of peak amplitudes were derived after epoch-wise baseline removal and class-wise averaging of epochs. Values reported for N150 were determined as the minimum of channel O1 of the interval $[100, 200]$ ms, while the late positivity (P300) was derived as the maximum of channel Cz in the interval $[250, 500]$ ms. The last column lists the AUC values estimated via cross-validation from a supervised classifier (see text).

	N150 (O1)		**P300 (Cz)**		
Subject	Ampl. (μV)	Lat. (ms)	Ampl. (μV)	Lat. (ms)	**AUC (%)**
P1	-9.76	150	2.72	340	98.85
P2	-11.11	150	1.48	400	98.73
P3	-5.63	170	1.94	500	98.06
P4	-9.48	160	-0.25	500	99.82
P5	-7.59	160	1.15	410	97.05
P6	-12.17	170	0.65	470	97.12
P7	-7.79	150	1.13	450	99.92
P8	-3.57	180	3.87	360	91.69
P9	-13.25	140	0.11	380	99.56
P10	-12.01	140	3.67	380	99.72
P11	-2.93	180	1.31	300	89.18
P12	-4.35	150	3.49	370	98.89
P13	-4.10	160	3.57	370	98.45
Mean	**-7.98**	**158.46**	**1.91**	**402.31**	**97.46**

the full amount of data (all 12852 epochs) are available. The ERP plots for subject S6 and four intervals are given in Figure 3.7. It compares the target and non-target ERP means estimated by LLP (Figure 3.7**A**) with the true class means (Figure 3.7**B**). We observe, that the reconstructed class means capture the characteristics of the original means almost flawlessly.

It is also of interest, how the class means estimated by LLP evolve using a growing amount of data. As an example the target mean for subject P6 is provided in Figure 3.7**C**. Using epochs that correspond to $1, 3, 7, 14, 28, 42$ and 63 symbols, the mean target pattern in the interval $[120, 200]$ ms stabilizes towards the supervised true mean. While the negative potential over occipital channels undergoes a linear development from strong to weak intensity, the activity in frontal and central channels reveals jumps between negative and positive potentials specifically during the first 10 symbols until finally converging towards the ground truth.

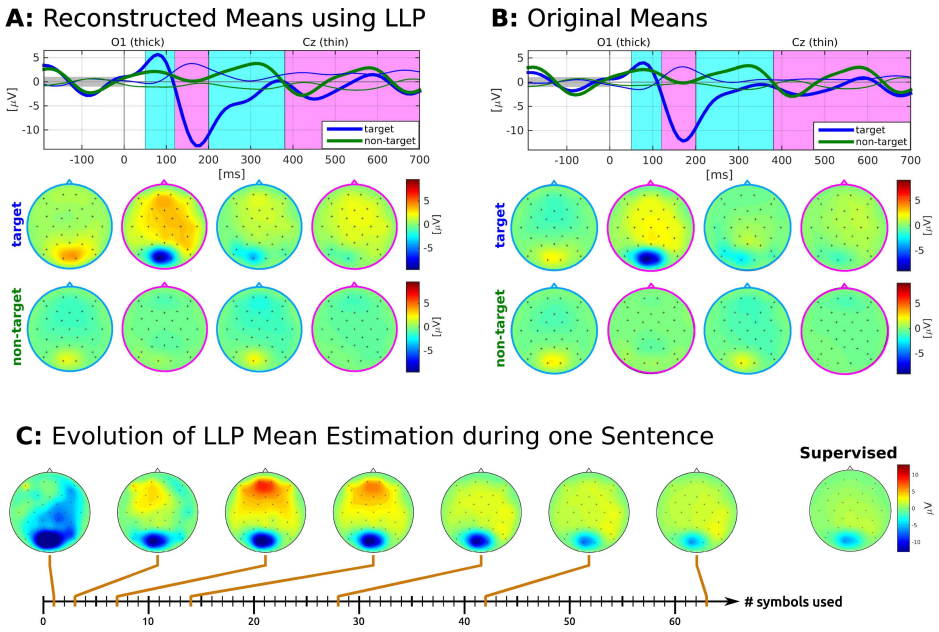


Figure 3.7: **Original and LLP-reconstructed average ERP responses.** For the prototypical subject P6, subplot **A** displays the reconstructed class-wise means using LLP and subplot **B** shows the average ERPs using the (supervised) sample means. Subplot **C** shows the LLP target estimations in $[120, 200]$ ms for different numbers of training points. Figure taken from [80].

Online Spelling Performance

Having found that LLP is able to successfully reconstruct the class means, LLP should be able to correctly decode the attended symbol. We will now present the results from the online study. Figure 3.8 shows the character-wise online spelling performance with LLP for all 13 subjects including the grand average. In total, 84.5 % of all characters were spelled correctly (chance level = 3 %). After a ramp-up phase of around 7 characters (which corresponds to 3 minutes wall clock time), this accuracy reaches 90.2 % correct characters on the remaining characters on average. In general, the algorithm worked well for all subjects except for P11. The reason for P11's low performance could be determined as an overall low SNR. It is evident also when looking at the supervised performance values provided in Table 3.2 and by the lack of class-discriminative N150. In the next section, we will analyze whether a violation of the homogeneity assumption could be the reason for the poor performance.

Homogeneity

To test the homogeneity assumptions of LLP, i.e. that both sequences have the same average target and non-target ERP responses, we visually inspected the responses for both sequences and each subject with the goal to detect systematic differences in the ERP amplitudes and latencies between

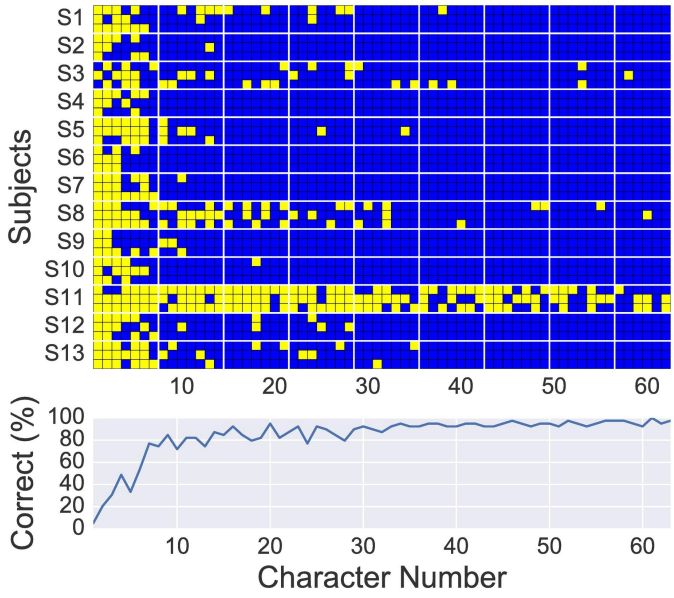


Figure 3.8: **LLP online spelling performance.** Top: each row represents a single spelling of the test sentence "FRANZY JAGT ...", with yellow squares indicating incorrectly spelled characters and blue squares indicating correctly spelled characters. Bottom: the averaged spelling accuracy across sentences and subjects is shown for each character. Figure taken from [80].

the two sequences. Figure 3.9 shows the ERP plots from subject P11 for both sequences. Even though small differences can be observed, the ERP responses generally look extremely similar, and we could not detect any systematic differences by visual inspection.

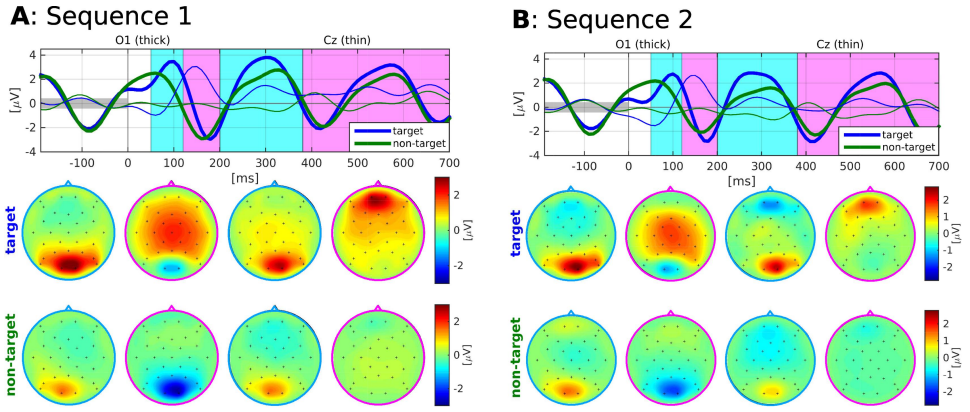


Figure 3.9: **Sequence-wise average target and non-target ERPs** for subject P11. Subplot **A** shows the average response for sequence 1 and subplot **B** displays the average for sequence 2. See the caption of Figure 3.6 for a detailed description of the plots. Figure taken from [80].

We also performed a leave-one-out bootstrapping test comparing the similarity of a sample from sequence 1 to the average ERP responses of both sequences. The idea of this test is to compute the similarity of a sample from S1 to the average ERP response from S1 and to the average

ERP response from S2. The similarity values allow testing whether the null hypothesis holds that target and non-target responses follow the same distribution for both sequences. After applying the same preprocessing steps as mentioned before, we iterated over each target (non-target) epoch of S1. The average target (non-target) ERP responses for both sequences were computed where the selected epoch was excluded when calculating the average of S1. In the next step, the squared distance (L_2 - norm) between the selected epoch and the previously computed averages was calculated in the interval $[0, 700]$ ms using all channels. A two-sided paired T-test was finally conducted to check, whether the distances to S1 differ significantly from the distances to S2. This procedure was done separately for each class (target / non-target) and subject, yielding a total of $2 \cdot 13 = 26$ tests. The significance level was corrected by dividing by 13 accounting for testing on 13 independent subject.

One significant difference for the corrected significance level ($\alpha := 0.05/13$) was found, namely the differences in target ERP responses for P4. However, the good spelling performance in Figure 3.8 suggests that this violation is not critically harming LLP. No violation of the homogeneity assumption was found for the worst subject P11.

In summary, this section has introduced the learning from label proportions idea to the BCI community. Results from an online study provide strong empirical evidence that LLP is readily applicable. With that, we have opened the door for a new class of unsupervised learning algorithms.

AUTHOR'S CONTRIBUTION

This work was a joint work with Pieter-Jan Kindermans, Konstantin Schmid, Thibault Verhoeven, Klaus-Robert Müller and Michael Tangermann. The LLP idea originated from a meeting with Pieter-Jan Kindermans. I took the leading role in formulating the idea, in implementing and testing the idea in the online study and in the data analysis, result visualization and paper writing. A full list of contributions can be found at the end of the published paper [80].

3.3 NEW APPROACH 2: MIXING MODEL ESTIMATORS

In the last section, I have derived LLP as a simple, yet powerful unsupervised machine learning algorithm by modifying the paradigm according to the needs of the classifier. LLP is the first unsupervised classifiers with guaranteed convergence for ERP-based BCIs. It shows a quick learning behavior in the beginning, but has a low convergence rate.

In contrast, the EM-algorithm for a Gaussian mixture model has a better convergence rate, but it relies on a good initialization. Without a good initialization, EM performs poorly. Observing these complementary strengths naturally leads to the idea of combining both algorithms. In a collaboration with Thibault Verhoeven, we have derived a theoretical framework to obtain an analytical combination of the two algorithms [174]. A short summary of the method will be given in the following section. This framework was then tested in an online study [82]. The text and Figures are largely taken from the latter publication.

3.3.1 *Methods*

This method relies on the same classification framework as before. Again, the goal is to find a projection $\mathbf{w}$ which only depends on the (inverse) global covariance matrix and the difference of the class means, see Section 2.5.3. The global covariance matrix can be estimated without label information. To integrate both classifiers, a new method needs to derived that is able to combine the mean estimations of the target and non-target classes.

Analytical combination of LLP and EM

The LLP algorithm directly calculates the class-wise means. Importantly, the EM-algorithm can also be formulated in a way that yields mean estimations, see [174] or [172] for a detailed derivation. We define the new (mixed) mean estimation $\hat{\boldsymbol{\mu}}_{MIX}$ as a linear combination of the EM mean $\hat{\boldsymbol{\mu}}_{EM}$ and the LLP mean $\hat{\boldsymbol{\mu}}_{LLP}$

$$\hat{\boldsymbol{\mu}}_{MIX}(\gamma) = (1-\gamma)\hat{\boldsymbol{\mu}}_{EM} + \gamma\hat{\boldsymbol{\mu}}_{LLP} \tag{3.4}$$

where $\gamma \in [0, 1]$ is the mixing coefficient and regulates the influence of each individual mean.

Optimal mixing coefficient

Inspired by the concept of mean shrinkage for supervised classification [72], the optimal mixing coefficient γ^* is obtained as the value that minimizes the expected mean squared error between the estimated value $\hat{\boldsymbol{\mu}}$ and the unknown true parameter value $\boldsymbol{\mu}$:

$$\gamma^* = \arg\min_{\gamma} E\left[\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{MIX}(\gamma)\|^2\right]. \tag{3.5}$$

Given the approximation that the EM algorithm gives an unbiased estimation of the mean ($\hat{\boldsymbol{\mu}}_{EM} = \boldsymbol{\mu}$), Verhoeven and colleagues showed that this approach leads to an analytical formulation for the optimal mixture coefficient γ^* [174]:

$$\gamma^* = \frac{1}{2}\left(\frac{\sum_{d=1}^{D} Var\left[\hat{\boldsymbol{\mu}}_{EM,d}\right] - \sum_{d=1}^{D} Var\left[\hat{\boldsymbol{\mu}}_{LLP,d}\right]}{\|\hat{\boldsymbol{\mu}}_{EM} - \hat{\boldsymbol{\mu}}_{LLP}\|^2} + 1\right). \tag{3.6}$$

Here, d runs over the features, and $Var\left[\hat{\boldsymbol{\mu}}_{(\cdot),d}\right]$ denotes the variance (over different realization of the data) of the estimator for the d^{th} entry of the estimated mean $\hat{\boldsymbol{\mu}}_{(\cdot)}$. This variance is a measure of the uncertainty of the estimated value. The higher the uncertainty on the output of the LLP method, the higher the weight given to the output of the EM method and vice versa.

To estimate the variance in LLP, one can derive a closed-form solution which only depends on the mixing matrix and data variance. For the EM, no closed-form solution exists. Additionally, only one realization of the data is observable in practical applications, and simulating other realization is time-consuming and inaccurate. To overcome these limitations, one can utilize that the EM-estimator converges asymptotically to a Gaussian distribution where the variance can be computed based on the data [174].

Study details

In a single experimental session per subject, we compared the LLP, EM and newly-derived MIX classifier. Twelve healthy volunteers (8 female, 4 male, mean age: 26 yrs, age range: $19-31$ yrs) performed a copy-spelling session using a visual BCI speller. Two of the subjects (S2, S8) had prior EEG experience. All participants gave written informed consent prior to participation and the ethics committee of the University Medical Center Freiburg approved the study. A session took about 3 hours (including the EEG set-up and washing the hair), and participants were compensated with 8 Euros per hour.

The setup is very similar to the previous study that has been described in Section 3.2. Therefore, the description of this experiment is restricted to the essential points and main differences to the previous study.

Within a single session, a subject was asked to spell the beginning of a sentence in each of three blocks. The text consists of the 35 symbols "FRANZY JAGT IM TAXI DURCH DAS ". Each block, one of the three decoding algorithms (EM, LLP, MIX, see Section 3.1.2) was used in order to estimate the attended symbol. The order of the blocks was pseudo-randomized over subjects, such that each possible order of the three decoding algorithms was used twice. This randomization should reduce systematic biases by order effects or temporal effects, e. g., due to fatigue or task-learning. An overview of the experimental structure is given in Figure 3.10**B**.

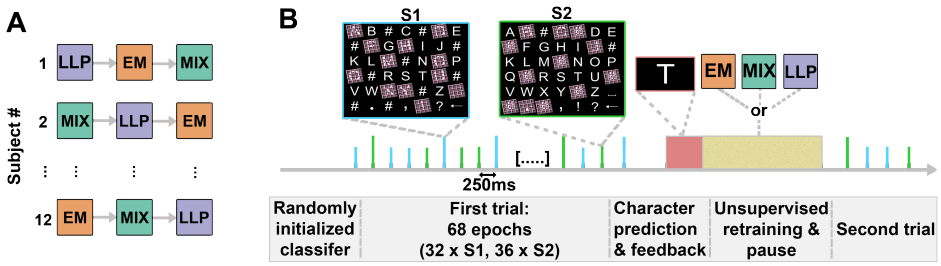


Figure 3.10: **Structure of the online experiment. A:** each subject performed three experimental blocks. Each block used a different unsupervised classifier (expectation-maximization (EM), learning from label proportions (LLP) or a mixing of the two (MIX)). **B:** at the start of a block, the corresponding classifier was initialized randomly. The speller was modified to allow the application of LLP by introducing visual blanks "#" and two sequences (S1,S2), see the description of LLP. Attended- (target events) and not attended stimuli (non-target events) are indicated by shorter and longer bars, respectively. This label information, however, remained unknown to the classifiers. After each trial, the attended character was predicted and the classifiers were updated. Figure taken from [82].

Implementation

Since the EM algorithm relies on a good (random) initialization, Kindermans et al. [98] proposed to initialize five classifier pairs in parallel, thus, increasing the chance of having a good random initialization. After each trial, the classifier with the highest log likelihood was selected as the active classifier, while the other classifiers were also updated. Pairs were used because the unsupervised EM classifier can also learn to solve the inverse problem (meaning that non-target and target are swapped). As shown by Kinderman et al., a higher log likelihood correlates with better AUC and selection performance [95]. Hence, always the classifier (and its negative) of each pair with the highest log likelihood were kept in the following.

3.3.2 *Results*

Figure 3.11**A** shows the target vs. non-target classification accuracies for each subject and each of the three unsupervised learning method and Figure 3.11**B** shows the grand averages over the 12 subjects. While LLP reliably improves in the beginning but only shows slow learning over time, the EM algorithm performs more dichotomous. Depending on the random initialization, the classifier can either find the projection very early (S7) or only relatively late (S6, S9). The MIX method performs best for almost all subjects and is able to consistently reach a high decoding accuracy with an average of around 80 % after data of around seven characters has been recorded. We would like to emphasize that seven characters correspond to only 168 s of unsupervised training time or 476 unlabeled epochs. This small amount of data suffices to reliable estimate attended characters (see

Figure 3.11**C**). The characteristic behaviors of the three classifiers also transfer to the spelling accuracy depicted in Figure 3.11**C**.

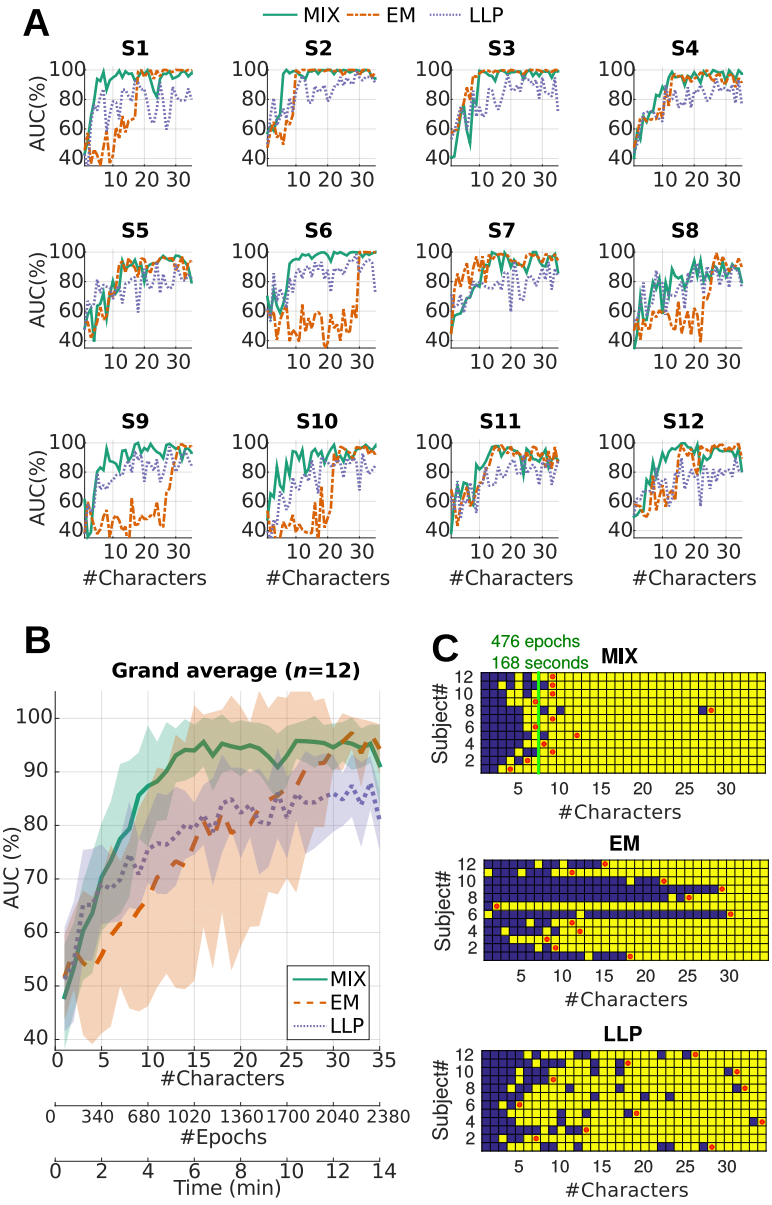


Figure 3.11: **MIX, EM and LLP online performance. A:** target vs. non-target classification accuracies for each subject. The AUC was computed on the latest (unseen) character. **B:** grand average classification accuracy. Shaded area depicts the mean ± standard deviation across subjects. **C:** overview of correctly and incorrectly spelled characters for all 12 subjects. Blue squares denote incorrectly spelled characters while yellow squares indicate correctly spelled characters. For each subject and method, the red circles indicate the first time point of perfect control, where all post-hoc reanalyzed characters and all upcoming letters are correctly decoded. Post-hoc reanalyzed characters are obtained by reapplying the current (improved) classifier to the data from all trials up to the current one. **EM**: expectation-maximization, **LLP**: learning from label proportions, **MIX**: combination of the two methods. Figure modified from [82].

In the following, we want to have a closer look at the mixing coefficients γ_{pos} and γ_{neg} for the positive and negative class, respectively. They are shown for the twelve subjects in Figure 3.12. Three observations can be made. First, interestingly, the mixture coefficients of the poor-performing subject (S8) appear to be different from the rest. Second, non-target mixture coefficients are higher than the target mixture values. This is because there are more non-target examples than target examples in the data which leads to a higher confidence in the LLP mean estimation for the non-targets. Third, the mixture coefficient is not going down to zero which is surprising because the EM estimator should have the lowest variance as it is the maximum likelihood estimator. Thibault Verhoeven argues that this behavior is because the variance does not decrease fast enough compared to the norm in the nominator of the mixture formula in Equation 3.6, see [174]. Having found that the MIX method is outperforming the two competing

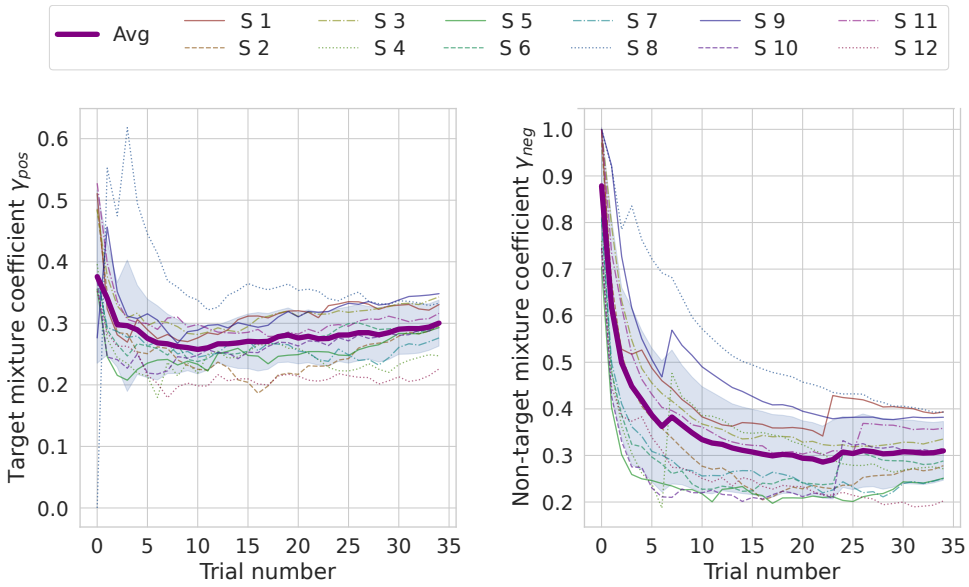


Figure 3.12: **Mixture coefficients for target and non-target means.** The left plot shows the mixture ratio γ for the targets while the right plots shows them for the non-target in dependence of the number of trials (i. e. the number of spelled characters).

unsupervised learning methods by a large margin in the online study, the question remains how well it competes with a supervised classifier. We compared the unsupervised MIX performance with supervised shrinkage-regularized LDA classifier [21] which is a highly competitive supervised classifier in the field of BCIs [117]. As no supervised classifier was used in the online experiments, we could realize such a comparison only in a post-hoc offline re-analysis of the data. In this offline re-analysis, both classifiers were trained on the first $N-1$ characters and tested on the N^{th} character. Figure 3.13 shows the results.

We tested the null hypothesis that both single epoch classification accuracies come from the same distribution. The non-parametric Wilcoxon signed-rank test showed that significant differences exist only for the first

9 characters (for $p = 0.05$). This is convincing evidence that the unsupervised MIX method can utilize data that is unlabeled almost as efficient as a supervised method with full label access.

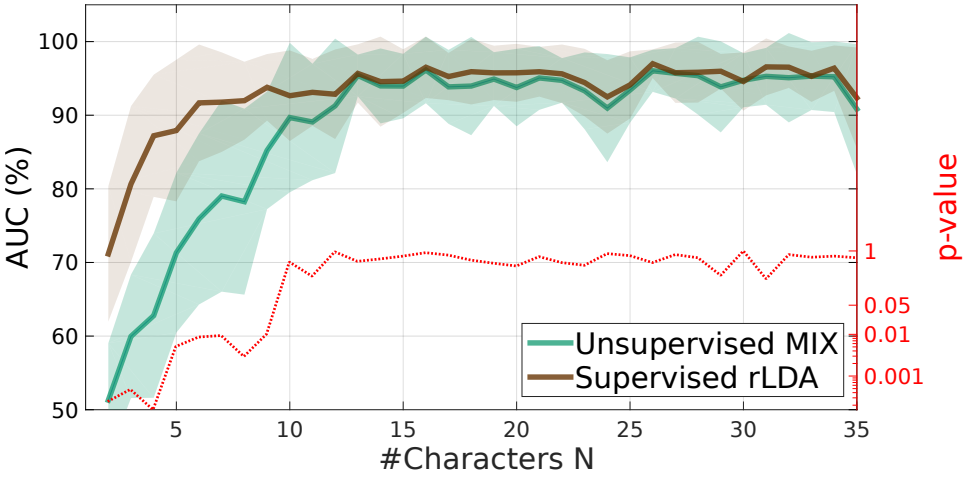


Figure 3.13: **Comparison of the unsupervised MIX method with a supervised regularized LDA classifier.** Both classifiers were trained on the first $N-1$ characters and tested on the N^{th} character. The thick lines depict the grand average over 12 subjects while the shaded area shows the standard deviation across subjects. The red dotted line shows the p-value of a Wilcoxon rank sum test comparing the supervised and unsupervised performance for character N. Figure taken from [82].

AUTHOR'S CONTRIBUTION

This work was again joint work with researchers from TU Berlin, TU Ghent and the University of Freiburg. Thibault Verhoeven took the leading role in deriving the MIX classifier. I took the leading role in testing the new method in the online study, in performing the EEG data analysis, in visualizing the results and in the writing process of the IEEE publication including the review from Section 3.1.

3.4 UNSUPERVISED LEARNING IN A BCI CHESS APPLICATION

In the previous two sections, I have shown how the LLP and MIX method can be derived and that they show extremely promising results in a modified visual speller. One limitation is that additional symbols (the "#" symbols) needed to be introduced which effectively do not transfer any information, and hence, lead to a slightly inefficient interface. Additional programming effort is also needed when implementing these extra symbols. In this section, I want to explore the idea that the LLP and MIX method can also be applied to an interface that does not have additional explicit "non-target symbols". This will be tested in a recently presented two-player online chess application [75] where both players make all their moves using a visual ERP-based BCI system. The following text and Figures are largely taken from a publication that is currently under review [76].

The control over the chess application is realized with a visual highlighting scheme that highlights candidate chess pieces and fields in a two-step process. First, all possible chess pieces are highlighted one by one, and the user is instructed to focus on the piece that they want to move, see Figure 3.14**A**. The goal of the highlighting is to elicit distinguishable transient ERPs in the subjects that can be recorded with EEG and used to identify the selected piece. The same visual highlighting pattern (trichromatic grid overlay) as before was used. Once the piece was selected, the possible fields are highlighted one by one and the user can select them by paying attention to the one field where the player wants to move his piece, see Figure 3.14**B**.

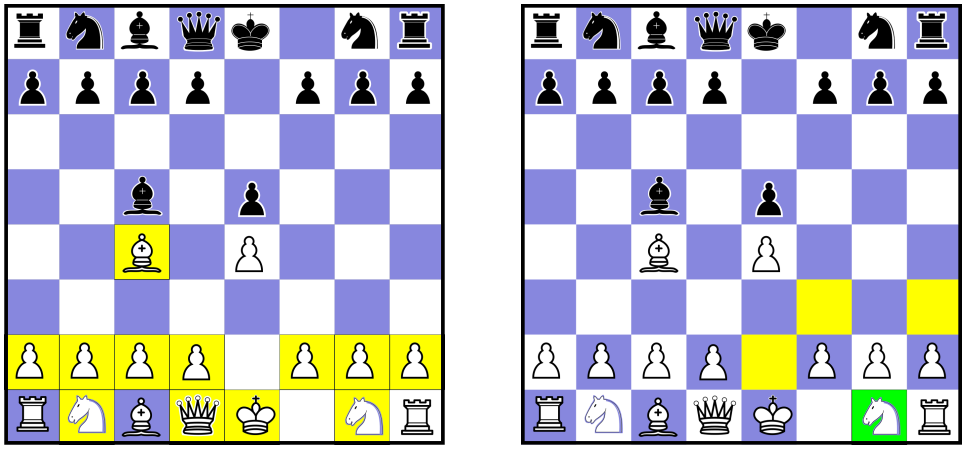


Figure 3.14: **Example move of a BCI-based chess game, broken down into two steps. A:** in the first step, the white player can select any of the 12 pieces marked by a yellow background. **B:** after the knight had been selected (marked by green background), the user can move it in the second step to any of the 3 fields marked by a yellow background. During online gaming, the 12 / 3 options are highlighted one by one to elicit time-locked ERP responses in the player's brain. Recording them with EEG and decoding them with machine learning methods, the BCI can then infer the desired piece and its next position.

Now I claim that this is a very well suited application to make use of unsupervised learning via LLP. If we have a closer look at the two selection steps above, we can easily observe that the number of highlighted fields differs between the steps. For instance, in the game situation depicted in Figure 3.14**A**, 12 figures could be selected. Assuming that the figures are highlighted one by one, this yields a target to non-target ratio of 1 : 11. In contrast, only 3 fields can be selected in the second step (see Figure 3.14**B**), yielding a target to non-target ratio of 1 : 2. With other words, the ERP response averaged over all stimuli in the first step will resemble a non-target response more closely compared to the average response in the second step, which will look more similar to a target response. Over multiple trials, we can then define a (sub)group as the collection of all EEG epochs that come from trials with the same number of highlighted pieces or fields. To give an example, the epochs from Figure 3.14**A** fall into a group with sequence length 12. In summary, we have identified different groups in the data that display different label proportions. This is the central prerequisite for LLP and it was achieved without a paradigm modification.

Even though the central prerequisite for LLP is fulfilled, we face three new challenges when applying LLP in this interface compared to the modified visual speller. First, we will not only have two different groups in the data, but as many groups as there are different numbers of chess pieces/fields that can be selected in one step. This number theoretically ranges from a single possible move, e. g., if a player needs to respond to a check, to as many as 28 possible moves if a queen can freely move on the board. Second, a different number of epochs will be collected over time for each of the groups because some chess situations are more likely than others, see Figure 3.15.

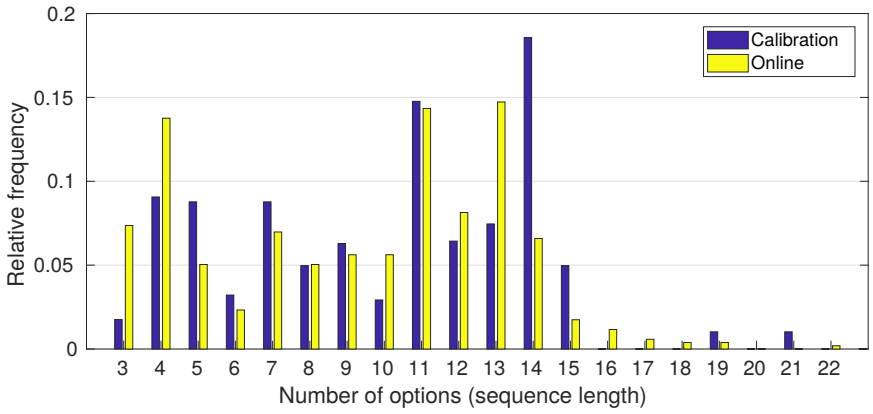


Figure 3.15: **Relative frequency of different sequence lengths.** The x-axis denotes the number of options (pieces or fields) that can be selected during one step. Please note that there is always one additional option to request more thinking time and that the program automatically executes the move if only one move is possible. During calibration, users were presented with predefined real chess positions.

Third, the label proportions are not mixed within a single trial anymore, but they are mixed across trials. This is in stark contrast to the scenario of the visual speller application, where two groups with different label proportions were present during the spelling of a single character. In the BCI chess application, different label proportions only exist when we have collected data of multiple and different selection steps. I will explain in more depth below why this entails consequences for the experimental paradigm.

3.4.1 *Methods*

Weighted least squares regression

Consider the LLP scenario from before as shown in Section 3.2.1. We faced a two-class classification problem (target vs non-target) with G groups of data where each group is a mixture of these two classes with known mixture ratios contained in the matrix $\mathbf{\Pi}$. The means of the feature vectors in the groups $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_G$ can then be expressed as a function of the class means $\boldsymbol{\mu}_T, \boldsymbol{\mu}_N$ as follows.

$$\begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_G \end{bmatrix} = \mathbf{\Pi} \begin{bmatrix} \boldsymbol{\mu}_T \\ \boldsymbol{\mu}_N \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_G \end{bmatrix}, \quad \mathbf{\Pi} := \begin{bmatrix} \pi_+^1 & \pi_-^1 \\ \vdots & \vdots \\ \pi_+^G & \pi_-^G \end{bmatrix}, \quad \boldsymbol{\epsilon}_i \sim N(0, \sigma^2) \tag{3.7}$$

The G different groups in the data correspond to subsets of the chess data where each subset is given by a certain number of possible fields/pieces that can be selected. The mixture ratios contained in $\mathbf{\Pi}$ indicate the proportions of targets and non-targets in each of these groups. We can obtain an estimate for the group averages $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_G$ by simply averaging all responses contained in the respective group. This averaging, again, does not require label information. The term $\boldsymbol{\epsilon}$ is a vector containing random variables that describe the errors we make during that averaging and which is mostly influenced by the SNR of the data and the number of data points. If the errors ϵ_i are independent and normally distributed with constant variance σ^2 and an expected value of 0 for all $i \in \{1 \ldots G\}$, then this is the ordinary least squares (OLS) problem where the solution was presented in Equation 3.2:

$$\begin{bmatrix} \boldsymbol{\mu}_T \\ \boldsymbol{\mu}_N \end{bmatrix}_{OLS} = (\mathbf{\Pi}^T \mathbf{\Pi})^{-1} \mathbf{\Pi}^T \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_G \end{bmatrix}. \tag{3.8}$$

However, we face the following problem. The number of samples per group differs, see Figure 3.15. Remember that a group is defined by the number of possible fields/pieces that can be selected. The critical observation is now, that the number of samples influences the quality of the mean estimation for the different groups, resulting in better group mean

estimates for some groups than for others. Hence, we need to consider the number of samples per group. To incorporate this, we adjust the noise model by replacing the error vector $\boldsymbol{\epsilon}$ from before by an error vector $\tilde{\boldsymbol{\epsilon}}$:

$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_G \end{bmatrix} = \mathbf{\Pi} \begin{bmatrix} \mu_T \\ \mu_N \end{bmatrix} + \tilde{\boldsymbol{\epsilon}}, \quad \tilde{\boldsymbol{\epsilon}} := \begin{bmatrix} \tilde{\epsilon}_1 \\ \vdots \\ \tilde{\epsilon}_G \end{bmatrix}. \tag{3.9}$$

In Equation 3.7, all ϵ_i had the same variance. As they can differ now, we face a **heteroscedastic** problem for which, however, we can estimate the error terms. From the central limit theorem, it is known that the variance of a mean estimation based on independent and identically distributed random variables is proportional to $\frac{1}{N}$, where N is the number of available samples:

$$\mathrm{Var}(\tilde{\epsilon}_i) \propto \frac{1}{\#\text{samples in group i}}. \tag{3.10}$$

Simply said, more data per group will reduce the variance (noise) of the corresponding group mean estimate, and we should trust these group averages more than those obtained for less frequent groups. Accordingly, we define the weighting matrix $\mathbf{W}$ whose diagonal entries indicate the relative importance of each group. Please note that any scaling of this matrix (e. g., multiplication with a factor such as the total number of samples) will become irrelevant later on.

$$\mathbf{W} := \begin{pmatrix} \#\text{samples group 1} & 0 & \dots & 0 \\ 0 & \#\text{samples group 2} & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \#\text{samples group G} \end{pmatrix} \tag{3.11}$$

With these definitions in place, we can retrieve a model with constant variance by computing $\frac{1}{\sigma_i}\tilde{\epsilon}_i$. Based on that, we can see that our new error term $\tilde{\boldsymbol{\epsilon}}$ is given by

$$\tilde{\boldsymbol{\epsilon}} = \mathbf{W}^{-\frac{1}{2}}\boldsymbol{\epsilon} \cdot C \tag{3.12}$$

where $C \in \mathbb{R}$ is an (irrelevant) scaling constant. This holds because we can recover the relationship from Equation 3.10:

$$\begin{aligned} \mathrm{Var}(\tilde{\epsilon}_i) &\propto \mathrm{Var}((\mathbf{W}^{-\frac{1}{2}}\boldsymbol{\epsilon})_i) && (3.13)\\ &= \left(w_{ii}^{-\frac{1}{2}}\right)^2 \cdot \mathrm{Var}(\epsilon_i) && (3.14)\\ &= \frac{1}{\#\text{samples in group } i} \cdot \sigma^2 && (3.15) \end{aligned}$$

Please note, that neither the expected value nor the independence of the new error term is affected by this transformation. With this, we can rewrite Equation 3.9 by multiplying both sides with $\mathbf{W}^{\frac{1}{2}}$ and obtain

$$\mathbf{W}^{\frac{1}{2}} \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_G \end{bmatrix} = \mathbf{W}^{\frac{1}{2}} \boldsymbol{\Pi} \begin{bmatrix} \boldsymbol{\mu}_T \\ \boldsymbol{\mu}_N \end{bmatrix} + \mathbf{W}^{\frac{1}{2}} \tilde{\boldsymbol{\epsilon}} = \mathbf{W}^{\frac{1}{2}} \boldsymbol{\Pi} \begin{bmatrix} \boldsymbol{\mu}_T \\ \boldsymbol{\mu}_N \end{bmatrix} + \boldsymbol{\epsilon}. \tag{3.16}$$

We have now found a new linear system where the error terms are independent and identically distributed according to $N(0, \sigma^2)$. It can be solved by minimizing the least squares of the error term. It is easy to see that we can formulate the minimization problem as the weighted least squares (WLS) regression problem.

$$\begin{bmatrix} \boldsymbol{\mu}_T \\ \boldsymbol{\mu}_N \end{bmatrix}_{WLS} = \arg\min_{\mu_+, \mu_-} \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \tag{3.17}$$

$$= \arg\min_{\mu_+, \mu_-} \tilde{\boldsymbol{\epsilon}}^T \mathbf{W} \tilde{\boldsymbol{\epsilon}}. \tag{3.18}$$

Multiplying this optimization problem with a constant will not change the outcome which justifies why the constant C is irrelevant. The solution is given by the following analytical expression [125].

$$\begin{bmatrix} \boldsymbol{\mu}_T \\ \boldsymbol{\mu}_N \end{bmatrix}_{WLS} = (\boldsymbol{\Pi}^T \mathbf{W} \boldsymbol{\Pi})^{-1} \boldsymbol{\Pi}^T \mathbf{W} \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_G \end{bmatrix} \tag{3.19}$$

It delivers an optimal estimation of the target and non-target class means, $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_N$, respectively. This estimator has the same properties (guaranteed convergence to true class means for independent and identically distributed data points; variance decreases with $1/N$) as the one used before. In fact, it is a generalization of the previous approach where the number of samples was equal for all groups. With that, we have successfully tackled the challenges that arise from an unequal number of data points per group, and from a large number of different groups.

Randomized SOAs

The third challenge might not be obvious at first glance. For LLP, we need to average the ERP responses for each group. Computing the group-wise averages for a fixed SOA of 200 ms yields periodic responses, see Figure 3.16. Why is that? Please notice that ERP plots are normally aligned such that a target event or a non-target event was presented at $t = 0$ ms. In our case, we do not have that class label information and just compute the average per group. For each of the time points $t = 0$ ms, $t = 200$ ms, $t = 400$ ms, ..., we thus have a certain fraction of target and non-target events. That fraction is constant within each group. For this reason, we

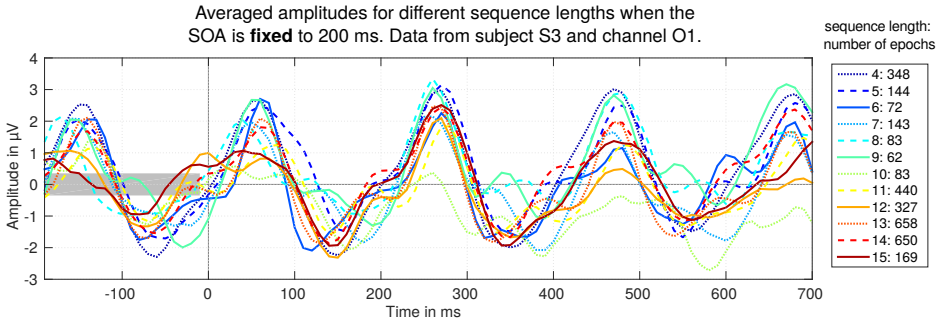


Figure 3.16: **Average ERP responses for different sequence lengths and a fixed SOA of 200 ms.** The data is grouped for all chess positions with the same sequence lengths, i.e. those positions where the same number of fields/pieces could be selected by the user. It can be observed that the average responses for shorter sequences show stronger amplitudes, as they are more similar to a (pure) target response. Similarly, longer sequences rather resemble non-target responses. The fixed SOA of 200 ms creates periodic ERP responses that obstruct the use of LLP.

have the same average ERP response with regards to these different time points and obtain periodic group-wise average ERP responses.

LLP estimates the class means by computing a weighted sum of these average responses where the same weights are applied for each time point. This is clearly a problem because it means that any LLP reconstruction based on these group estimations will also lead to a periodic estimation of the true ERP responses. Hence, LLP will deliver a bad approximation of the real ERPs.

A simple trick can solve this problem. Instead of using fixed SOAs, we uniformly sample the SOA from the interval [100, 300] ms. The interval is chosen such that its range covers a whole period of 200 ms with an average SOA of 200 ms. With that modification, the ERP responses to stimuli that occur before or after the stimulus at $t = 0$ are canceled out when averaging all responses. The result is shown in Figure 3.17. It depicts the averaged ERP responses per group for the same subjects, but using variable SOAs.

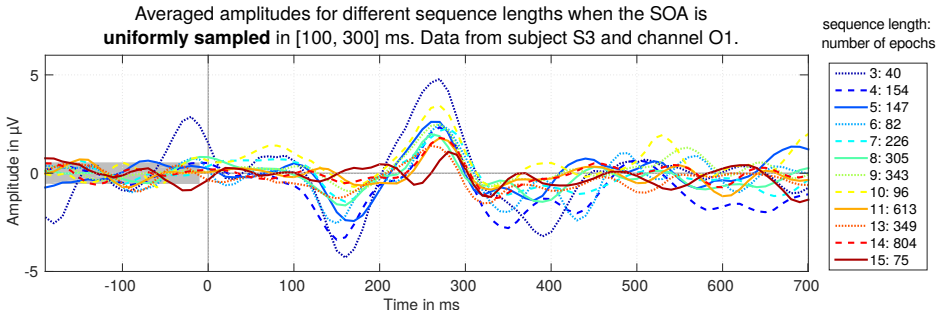


Figure 3.17: **Average ERP responses for different sequence lengths and a variable SOA.** The SOA is randomly sampled between 100 ms and 300 ms which has the effect that earlier and later ERP responses cancel out and a single ERP relative to t=0 ms survives the averaging.

For the sake of comparing the two conditions, we randomly selected each trial to be with fixed or with variable SOA in the following study protocol. In the result section, we will show that this simple but essential SOA modification only has a minor impact upon (supervised) classification performances.

Study protocol

We recruited six healthy male subjects (S1-S6) with a mean age of 28 years (range: $25 - 31$ years), who participated in the chess gaming study after providing written informed consent. The ethics committee of the University Medical Center Freiburg had approved the study. Two subjects (S3, S4) had prior experience with BCIs. Brain activity was recorded with EEG using 31 passive Ag/AgCl electrodes (EasyCap) from which we only used the nine channels O1, O2, Cz, Pz, P4, P3, C3, C4, Fz. This reduction in feature dimensionality had been found beneficial for the unsupervised methods [78]. The signals were processed as described before in Section 2.4 using the same intervals as in Section 3.2.3.

Two players always competed against each other in pairs. Both players were controlling their pieces with a BCI chess application that was implemented based on an open-source Java chess application. Communication between both computers was realized over a free Telnet chess server [1]. The actual experiment consisted of three stages (Figure 3.18**A**): a calibration phase where labeled data was collected, an online phase where players controlled a BCI chess application in free play using individually trained supervised classifiers, and a final test phase to collect further labeled data.

During the online phase, both players made moves in an alternating order. After the opponent has made a move, the interface starts the highlighting sequence of the next move after a predefined thinking time. We have set the pause between moves to around 15 s and to 5 s between the piece and field selection, respectively, to allow players to think about their next move. In the case that the player had only one choice (e. g., only one piece could be moved or a piece could only be moved to one field), the application automatically executed that move. Each field was highlighted exactly 5 times. There were additional options to request more thinking time and to revert the last selection.

Simulated unsupervised classifiers

In addition to the online experiment, we simulated the learning behavior fof different unsupervised learning methods. We considered two scenarios. In the first scenario, classifiers were randomly initialized except for the supervised classifier that always had been trained on 5 minutes ($\approx$ 1000 epochs) of calibration data. In the second scenario, this training data was made available to all classifiers and was used to initialize them. Later on, the unsupervised classifiers sequentially (trial-by-trial) got access to the online data and tried to predict the target or non-target labels for all epochs

1 https://www.freechess.org/

of each new trial before the unsupervised models were updated using the new data. In this phase, the fixed supervised classifier also estimated labels, but was not changed over time. The final classifiers were then tested on the test data set. Please see Figure 3.18**B** for an overview. The target vs non-target AUC was taken as performance measure, see Section 2.7.

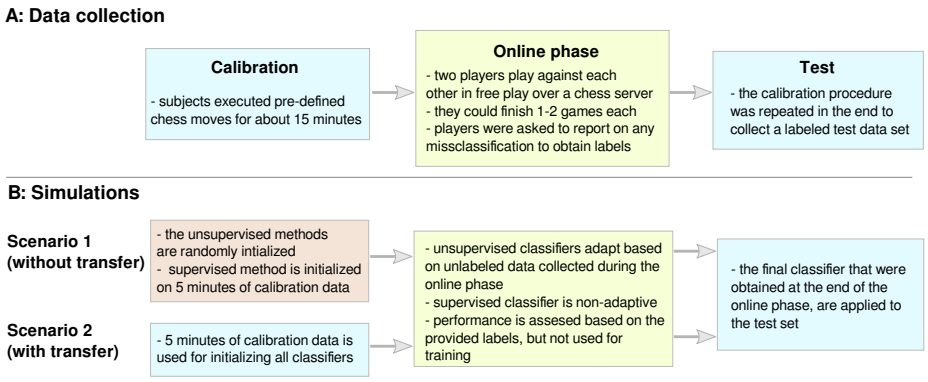


Figure 3.18: **Study protocol of the BCI chess study. A:** EEG data was collected from a BCI chess experiment with three phases: calibration, online phase and test phase. **B:** this data was also used to simulate an online experiment where a supervised (fixed) classifier is compared to three unsupervised classifiers.

We compared the following four classifiers.

1. **Supervised (fixed).** As a supervised baseline, we chose the regularized LDA classifier as explained in Section 2.4.4. The supervised classifier was always trained on the (labeled) calibration data and did not adapt over time.

2. **LLP (unsupervised-adaptive).** The LLP classifier is based on the weighted least squares regression model as given in Equation 3.19.

3. **EM (unsupervised-adaptive).** We used the same EM-algorithm as explained before in Section 3.1.2. In this implementation, we utilized one EM-classifier and its inverse classifier (which has the negative weights) in parallel. As the EM can learn to solve also the inverse problem during its optimization, we always evaluated the classifier which showed the better log likelihood.

4. **MIX (unsupervised-adaptive).** Similar to our previous approach [82, 174], we used a mixture (MIX) of the EM and the newly proposed weighted least square LLP classifier. In the MIX method, the estimation of the class-wise means is proposed to be a linear combination of the mean estimations found with the EM ($\hat{\boldsymbol{\mu}}_{EM}$) and those estimated by the LLP method ($\hat{\boldsymbol{\mu}}_{LLP}$),

$$\hat{\boldsymbol{\mu}}_{MIX}(\gamma) = (1-\gamma)\hat{\boldsymbol{\mu}}_{EM} + \gamma\hat{\boldsymbol{\mu}}_{LLP} \tag{3.20}$$

where $\gamma \in [0,1]$ denotes the mixing coefficient. A higher value of γ gives more weight to the LLP classifier. In the previous Section 3.3, an analytic solution for the mixture coefficient γ was presented that relied

on the variances of the LLP and EM estimator (estimated by using an approximation based on the Fisher information) and on the assumption that the EM estimator is unbiased [174]. While this showed great results for the visual matrix speller, we found that this approach gave suboptimal results in the chess application because the high number of different groups leads to an overestimation of the LLP variance.

In this contribution, we want to argue that the choice of a suitable mixing coefficient γ should be seen as a hyperparameter optimization problem for which it is not clear a-priori which exact choice is optimal. In contrast to the analytical solution, we present a new heuristic as a simple and computationally efficient alternative for finding a good value of γ. The heuristic is based on the following three observations. First, the LLP contribution should be high in the beginning to realize a good initialization. Thibault Verhoeven showed that an initial mixture ratio between $\gamma = 0.5$ and $\gamma = 1$ gives excellent performances during the initial learning phase (see [172], Figure 5.5) in a visual speller. The second observation is that we needed around 500 epochs to reach a very good performances with the MIX classifier in the previous section. One can also observe that the LLP contribution becomes less important after this time point. Putting these three observations together, we chose the mixing coefficients for target and non-target as

$$\gamma(N) = min(1, \frac{50}{N}) \tag{3.21}$$

where N denotes the number of epochs. The parameter γ is bound between 0 and 1 and starts with a high weight for LLP, but quickly reduces the LLP importance to only 10 % after 500 epochs (see Figure 3.19).

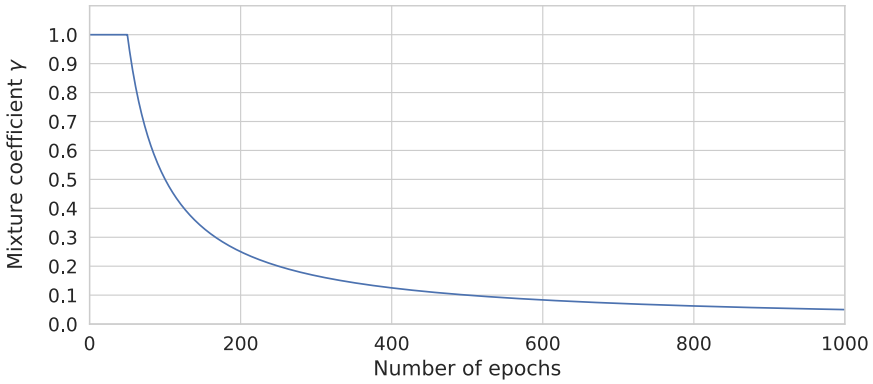


Figure 3.19: **Heuristic mixture coefficient γ plotted against the number of epochs.** A higher value of γ indicates a higher weight for LLP. The heuristic is designed such that LLP can provide a good initialization for the EM classifier and the importance decreases if more data is available.

3.4.2 *Results*

The results of the online chess experiment are shown in Table 3.3. All players were able to achieve meaningful control of the system and completed at least one chess game. We observed that player needed slightly more than half a minute to execute a single move and that an average of 1.5 incorrect moves occurred for a total of around 39 moves per player, yielding an error rate of around 4%. The reasons for the errors are of technical (e. g., failure of the optical sensor, channels in saturation) as well as subject-specific origin (e. g., the player tried to select a piece that could not be moved). Interestingly, the modification to use either a fixed SOA of 200 ms or to use a uniformly drawn SOA from the interval $[100, 300]$ ms did not have a strong impact on the classification accuracies.

Table 3.3: **Results of the BCI-chess experiment.**

	Moves			Supervised target vs non-target accuracy (AUC in %)	
	Number	Incorrect	Time per move (in s)	Fixed SOA	Variable SOA
S1	36	3	42.9	93.74	94.45
S2	36	0	36.3	91.26	92.33
S3	34	1	23.8	97.00	99.31
S4	34	0	27.2	97.37	94.38
S5	47	3	39.5	89.71	88.19
S6	46	2	34.7	95.30	96.21

ERP estimations using learning from label proportions

Next, we examined whether LLP based on weighted least square regression is able to recover the average target and non-target ERP responses without using labels. Figure 3.20 and Figure 3.21 show the true ERP responses and the LLP-reconstructed ERP responses for all subjects, respectively. In both cases, the combination of the labeled calibration and test data was used. One can observe that LLP is able to find the major ERP peaks (N200, P300), although the N200 amplitude is overestimated.

Unsupervised performances for randomly initialized classifiers

In the next analysis, we simulated EM, LLP and MIX in the case that they all start from a random initialization and compared them to a pretrained supervised LDA classifier (scenario 1). In Figure 3.22**B**, the performance in the test set is reported after the unsupervised models have been trained. Remarkably, the MIX method was again able to reach the same performance

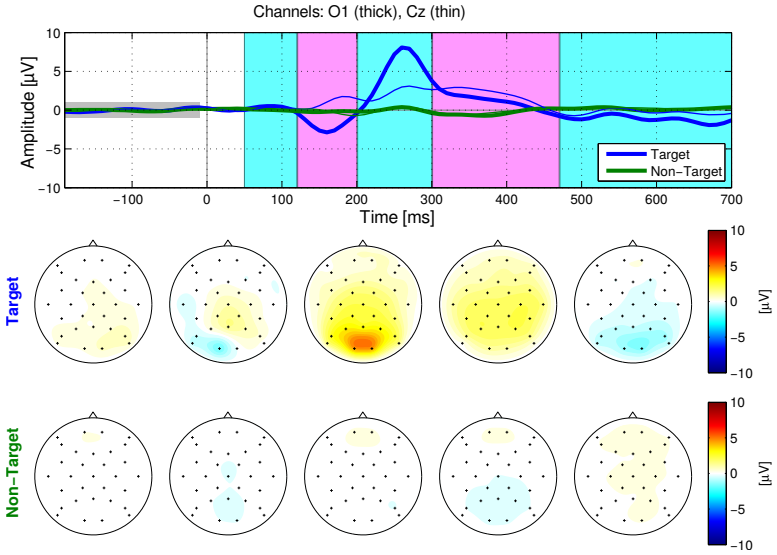


Figure 3.20: **Grand average (N=6) ERP responses given the true labels.** The scalp activity is shown for the five intervals [50, 120], [121, 200], [201, 300], [301, 500], [501, 700] ms.

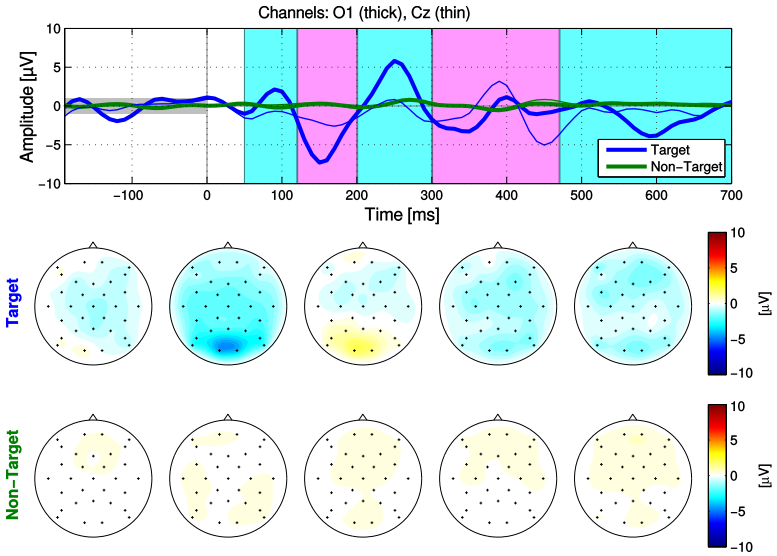


Figure 3.21: **Estimated grand average (N=6) ERP responses using learning from label proportions.** LLP is able to recover the two major components (N200, P300) without requiring labels, but delivers partially noisy estimates.

level as a supervised calibration although no label information was used. The other two unsupervised methods performed worse, with the EM being better than the LLP. When inspecting the learning behavior during the online phase in Figure 3.22**A**, the reasons for this behavior become evident. In agreement with the results from the visual matrix speller, we found the EM algorithm to perform dichotomous: it works well for data of 4 subjects, but fails to recover from a bad initialization for the 2 remaining subjects

(S3, S5). Generally, the LLP performs worse than in earlier studies [80], however, it can still provide sufficient information to the MIX method such that the latter quickly ramps up to the performance level of a supervised classifier for all six subjects.

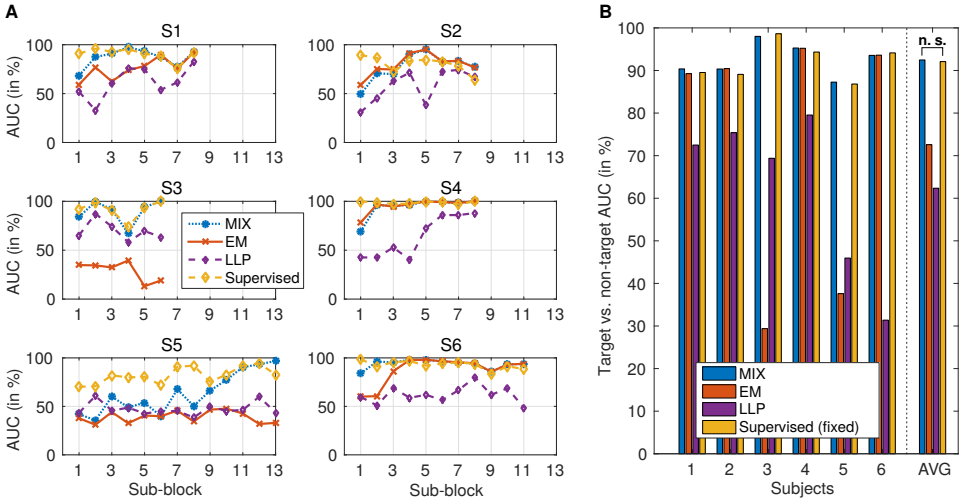


Figure 3.22: **Simulation of unsupervised classifiers starting from a random initialization. A:** The average classification accuracy for single epochs is shown for the three unsupervised adaptive classifiers (LLP, EM, MIX) and a supervised (fixed) classifier for sub-blocks of 5 trials and all six subjects (S1-S6) during the online learning phase. The classifiers started from trial 3 to guarantee that different sequence lengths have occurred which is a prerequisite for LLP. The number of sub-blocks varies across players as they made a different number of moves and sometimes used the options to request additional thinking time or to revert a piece selection. **B:** The single epoch performance of the final classifier (the one that was obtained at the end of the online learning phase) on the test set is shown for all six subjects and averaged (AVG). The differences between the MIX and supervised classifier are not significant (n. s.) on this unseen test data set when tested with a one-sided Wilcoxon signed-rank test (Z=4, p=0.75).

Effect of unsupervised online adaptation on classification performances

In the final analysis, we compared EM, MIX and the supervised classifier in a "fair" comparison where all of them were initialized on the same amount of **labeled** training data (scenario 2) before unsupervised adaptation took over in the following simulated online phase. The initialization had the effect that all classifiers started roughly on the same performance level at the beginning of the online phase, although the unsupervised methods had a slight lead which may be caused by the internal whitening step. When more data comes in, the adaptive unsupervised classifiers further outperform the non-adaptive supervised classifier, see Figure 3.23**B**. On the final test set, the unsupervised classifiers were around $1-2\,\%$ better than the supervised (non-adaptive) LDA (Figure 3.23**A**) which was significant

when tested with a Wilcoxon-signed rank test ($Z = 20$, $p = 0.03$). One can observe that the MIX and EM algorithm converge to a similar solution when LLP's influence has become smaller and smaller. The results of the LLP classifier are not presented here, because LLP should not be used as a standalone classifier if a very good initialization is available.

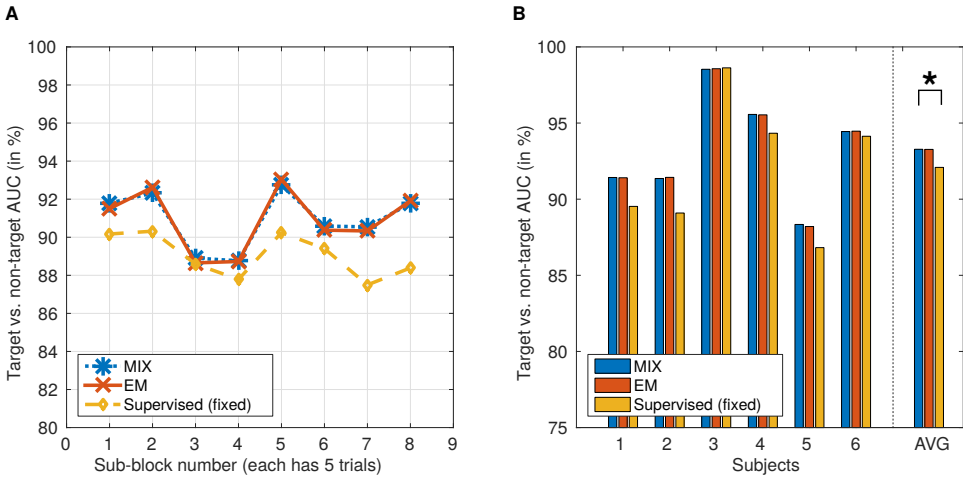


Figure 3.23: **Simulation of unsupervised classifiers starting from a supervised initialization.** All classifiers were initialized on around 5 minutes of labeled calibration data. **A:** The time courses of the single-epoch classification performances (averaged across the 6 subjects) during the simulated online learning process. **B:** The performance on the test set after the unsupervised classifiers have been adapted on the unlabeled online data. The differences in the average accuracy between the MIX and supervised classifier are significant at an α-level of 0.05 as tested by a one-sided Wilcoxon signed-ranks test ($Z = 20$, $p = 0.03$).

These simulations demonstrate that the learning from label principle can successfully be applied in an application without a change of the user interface. The resulting unsupervised MIX classifier showed remarkable performances again. This drastically increases the scope of applications for which the new unsupervised methods can be applied.

AUTHOR'S CONTRIBUTION & ACKNOWLEDGEMENTS

This project was conducted together with Albrecht Schall and Michael Tangermann. Albrecht Schall took the leading role in the implementation of the BCI-based chess application. I took the leading role in deriving the LLP extension, testing this idea in simulations and reporting the results. I want to thank Max Sagebaum for contributing parts of the code, and Andreas Rueckert and Harald Faber for implementing the original chess Java application.

3.5 SIMULATIONS ON PATIENT DATA FROM AN AUDITORY BCI

The last sections have demonstrated that the unsupervised MIX method performs exceptionally well in a modified visual spelling paradigm and in a BCI chess application with performance comparable to supervised methods. In the following, I evaluated the different unsupervised classifiers on the *auditory* ERP data that stems from our new BCI-supported language rehabilitation and will be described in more depth in Chapter 4. This application scenario is much more challenging, because it has a much lower signal-to-noise ratio. In addition, it is more difficult to apply LLP in an auditory ERP paradigm when compared to a visual setup because auditory stimulation protocols are less flexible, e. g., playing more than one sound at a time will quickly confuse the user while multiple letters can be flashed at the same time without problems in a visual paradigm.

The simplest way to apply LLP is by adding auditory blanks to the stimulation sequence. These could be tones that are not linked to control commands or pseudowords that have no meaning. With that, the user should never attend these sounds or tones, and they should always be in the non-target role. Naturally, these non-target examples will allow to learn from negative examples. In addition, this creates two sequences, one with only non-target examples and one with a mixture of target and non-target words. This allows for applying learning from label proportions again. This idea will be evaluated in simulations in this section.

3.5.1 *Methods*

For understanding the results of the simulations, I briefly outline the experiment in which the data was recorded and how the data was analyzed.

Study details

The data was recorded as part of a BCI-supported language rehabilitation for patients with chronic post-stroke aphasia. A detailed description of the underlying motivation, training protocol and more analysis will be given in the following Chapter 4.

In this protocol, patients were seated within a ring of six loudspeakers (AMUSE protocol [152]). Six bisyllabic German words (length=300 ms) were chosen as auditory stimuli. In each trial, one of these 6 words was cued by a sentence. Then, a sequence of these stimuli was played to the subject either via headphones or via 6 loudspeakers with a $1:1$ relation between words and loudspeakers. Per iteration, i. e. every six stimuli, each word was played exactly once. Words were played with a stimulus onset asynchrony of 250 ms or 350 ms depending on the subject's abilities to perform the task. A single trial consisted of at least 42 stimuli and a maximum of 90 stimuli depending on whether the BCI issued an early stopping. EEG signals from 31 passive Ag/AgCl electrodes (EasyCap) were recorded, which were placed according to the 10-20 system. Impedances

were kept below 20 kΩ, and channels were referenced against the nose. The signals were registered by multichannel EEG amplifiers (BrainAmp DC, Brain Products) at a sampling rate of 1 kHz.

Only the data of the first 8 patients was taken in the analysis. We recorded about 30 hours of data which resulted in between 11 and 25 sessions per subject. A total of 117 EEG sessions were recorded and are analyzed in the following. The data is completely labeled because the patients always had to perform a predefined task. In the following, this label information was only used to assess the quality of the classifiers, but not to train them at any point. For each session, only the first 36 trials were used to ensure that the same number of trials is available for each session.

Simulated online analysis

All the results that follow are from posthoc-simulations. In these simulations, an online scenario was simulated by successively providing the recorded data to the unsupervised classification methods as it would be in a real BCI session.

To enable LLP, one or two artificial non-target classes were created in the following two scenarios. In the first scenario, all trials were discarded where the first out of 6 words was the target word. In the second scenario, all trials were discarded where either the first or the second word was the target word. The remaining trials were left untouched such that they also include these two words. This exactly simulates the scenario that there is one or two words in the stimulation sequence that are never attended by the user, and as such, are guaranteed to be non-targets.

Four different unsupervised classifiers were then trained on this data.

1. **LLP.** Learning from label proportions was applied by defining two sequences. The first sequence only consists of non-targets from the one or two words where the target trials were discarded. The second sequence consisted of the remaining four or five words. This leads to a target to non-target ratio of $1:3$ or $1:4$ in the second sequence. With that, LLP can be applied as explained before in Section 3.2.

2. **EM original.** The same EM method as explained in Section 3.1 with five classifiers that are initialized in parallel.

3. **MIX.** This algorithm was also used as described before in Section 3.3.

4. **EM with negative examples.** In this version of the EM classifier, I incorporated the information about negative examples in the expectation step. For each trial, the classifier assigns a probability to each class of how likely it is that this class was the target during that specific trial. In the first scenario with one non-target class, the probability of that class being attended was set to zero. In the second scenario with two non-target classes, the probability for both classes were set to zero. As this approach uses the available label information perfectly, I expected this EM-version to outperform both the MIX and

the original EM version. For this classifier, also five initialization were used in parallel.

All classifiers were evaluated on the same test data. In this data, not only the trials containing the one or two artificial non-target words were removed, but also all their instances in the other trials. This means that in the final test set, only 5 or 4 classes were presented for scenario 1 and scenario 2, respectively. This was done because the user should never select one of the artificial non-target words in a real application.

3.5.2 *Results*

I will now present the results of the simulations. For 2 out of 8 patients, the unsupervised classifiers could not reach a satisfactory performance meaning that the average spelling performance and target vs. non-target AUC were both below 60% after training on the complete data. These two patients were excluded in the following analysis. Figure 3.24 shows the grand average results for the remaining 6 patients. In this Figure, the results for scenario 1 and scenario 2 were averaged.

All classifiers start on chance level and improve over time. A clear ordering is visible: the EM that also learns from negative examples outperforms all other classifiers. It reaches remarkable decoding performances of correctly predicting around 85% of the characters with a binary target vs. non-target accuracy of 75%. The original EM is on the second place while MIX and LLP perform worse. Thirty trials correspond to around 20 - 30 minutes of recording time and to about 1000 – 2000 epochs.

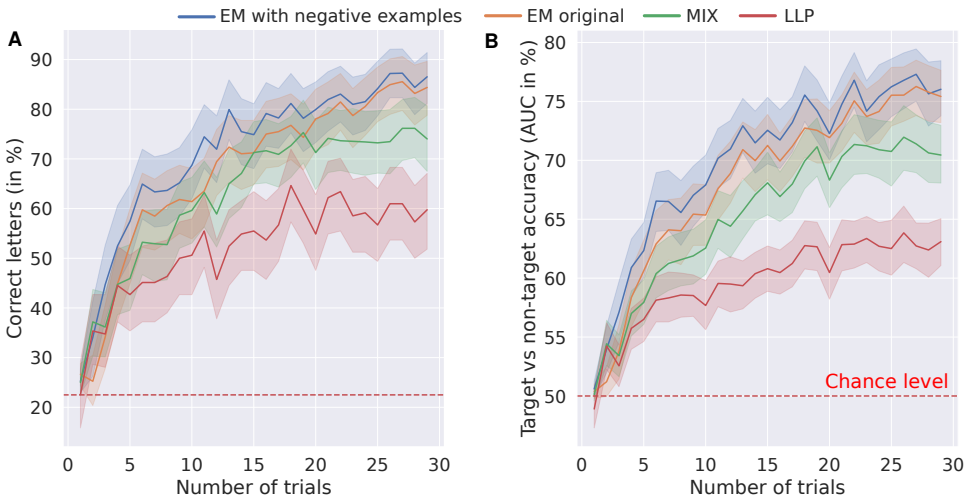


Figure 3.24: **Simulated grand average unsupervised performances for post-stroke aphasia patients performing a challenging auditory ERP task.** Subplot **A** shows the grand average spelling performance for the 4 different classifiers while subplot **B** displays the target vs non-target classification accuracy. The shaded areas in both plots shows the estimated 95% confidence intervals based on 1000 bootstrapping runs where the sessions were randomly resampled in each run.

Next, I investigated how the number of non-target classes influences the relative performance of the 4 classifiers. Therefore, the two scenarios (1: only one artificial non-target class, 2: two artificial non-target classes) were evaluated separately. Figure 3.25 shows the result for only one non-target class (scenario 1). In this Figure, the spelling performance was averaged over all 30 trials meaning that the results incorporate the initial ramp-up phase as well as the final phase where the classifier has improved. The classifier performance can be ordered quite consistently as before.

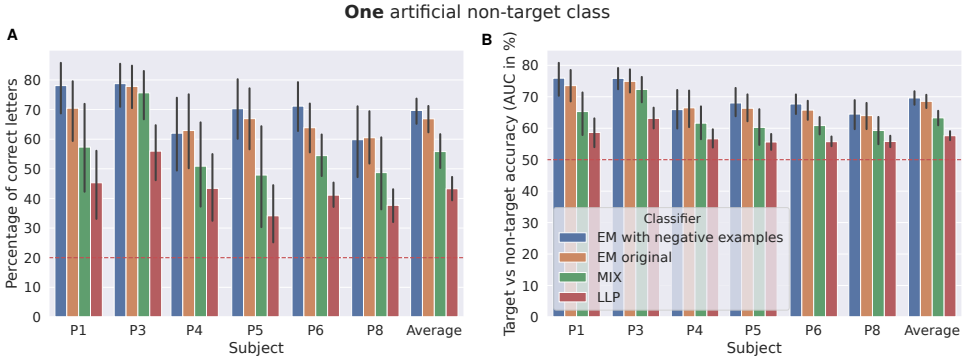


Figure 3.25: **Average unsupervised performances per subject for one artificial non-target class.** The bars depict the standard deviation across sessions and the red dashed lines depict the chance level. Subplot **A** shows the spelling performance while **B** shows the target vs. non-target classification performance.

Interestingly, this ordering slightly changes in scenario 2, see Figure 3.26. In this scenario, the MIX method performs better than the original EM method, but still worse than the EM algorithm which also learns from negative examples. The improved MIX performance is very likely to follow from an improved LLP classification performance.

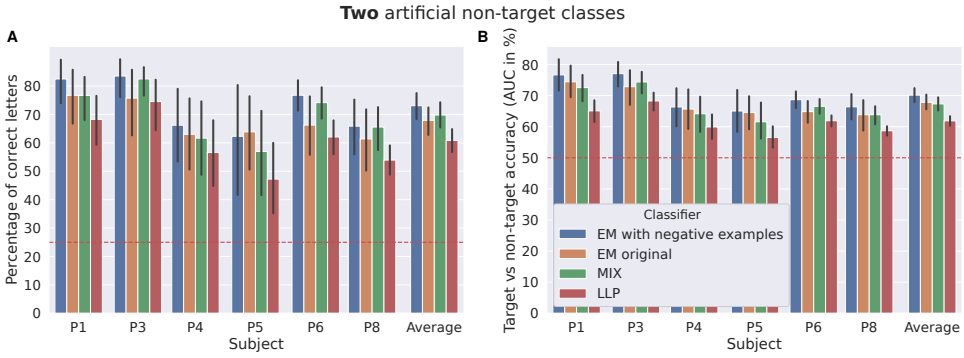


Figure 3.26: **Average performances per subject for two artificial non-target classes.** Subplot **A** shows the spelling performance while subplot **B** shows the target vs. non-target classification performance.

In summary, these findings show that the EM method can profit from negative examples. In addition, these final average target vs non-target performances of over 75% are quite remarkable given that the data comes from an auditory BCI and that no labels were used to train the classifiers.

3.6 DISCUSSION

In this chapter, learning from label proportions has been introduced to the BCI community. Compared to previous unsupervised machine learning approaches in BCI, LLP is a conceptually new approach. Previous studies have successfully improved many aspects of the individual parts of a BCI system, e. g., a lot of effort was put into improving the supervised machine learning models or to improve the SNR by using stimuli with a better saliency and more discriminatory information. Compared to that, the novelty of LLP is that it considers the classifier and paradigm as a holistic system which requires a holistic solution. This general idea to either utilize the existing rich (temporal or spatial) structure in the data or to create additional structure, seems to be a key in the unsupervised learning task to overcome the limited data and bad SNR. I have presented three different case studies where the paradigms were tuned to meet the prerequisites of LLP. This synergistic design leads to previously unseen capabilities.

The resulting LLP classifier has not only low conceptual and computational complexity, but has also a guarantee to recover the correct class means without using explicit label information. To the best of my knowledge, this is the first unsupervised classifier in the BCI field that has this property. Assuming discriminatory information between targets and non-targets, this implies that the user can rely on the system to correctly decode his or her brain signals given enough data from a stationary distribution.

When combining LLP with an expected-maximization algorithm, the resulting MIX shows fantastic performance, especially in the online study with the visual speller. In that online study, the results showed that subjects only required around 3 minutes of unsupervised learning time to obtain reliable control over the BCI system. And even this initial ramp-up time is not lost in some applications, but can be fully used when applying a later (improved) classifier to reanalyze initial trials that possibly were misclassified. When more data was collected, the observed average target vs. non-target classification AUC did not only exceed 90%, but the best unsupervised method was even on par with a supervised classifier that had full label information. This is very strong evidence that the approaches, which have been presented in this thesis, are able to efficiently exploit unlabeled data which is often abundantly available in BCI studies. This opens the door for realizing true plug-and-play systems that can dynamically adapt to the users brain signals over time. Ultimately, this does not only increase the usability of BCI systems, but can also give the opportunity to gain new insights when doing BCI data analysis.

While it was necessary to assign some stimuli into a pure non-target role for the visual speller and auditory ERP decoding, I have demonstrated how a weighted least squares LLP extension and a small paradigm modification (using variable SOAs) can allow for successful unsupervised learning even in a natural user interface, i.e. the chess interface in Section 3.4. This LLP extension dramatically increases the number of possible application scenarios. LLP can be applied to all BCI paradigms that rely on a multi-step

selection process with a different number of items per step. In addition, it can also be applied in scenarios where the number of selectable items changes over the course of the experiment, e. g., in a web browser where the number of objects continuously changes depending on the website visited [12]. In such applications, the only paradigm modification required is to replace the fixed SOA by a variable SOA. As shown empirically, this did not affect the classification accuracies systematically.

3.6.1 *Comparison of different unsupervised methods*

Although convergence is guaranteed for LLP, the convergence rate is rather low. While LLP quickly ramps up, it is not able to find a very good separation of the two classes with limited data. Looking at the convergence proof, the reason for this becomes evident. The variance of the mean estimation decreases with $\frac{1}{N}$. This means that the convergence rate is linear in N. Other algorithms have a better convergence rate. For instance, it was shown that the EM-algorithm has a superlinear (better than linear) convergence rate [121]. Hence, the main practical purpose of the LLP is to provide a good initialization for the EM-algorithm which results in the MIX method.

Compared to a pure EM-based solution, the MIX method shows the largest benefits if (a) the EM initialization is poor (e. g., due to high-dimensionality) and/or if (b) the number of classes is high. The first aspect is due to the fact that the EM algorithm often converges to local extrema and cannot easily recover from that. For that reason, Kindermans et al. [98] proposed to train five parallel classifiers and select the one with the highest data log-likelihood to maximize the chance that one of them is well-initialized. However, the MIX online study showed that even 5 randomly initialized EM classifiers are inferior compared to the MIX method in case of a visual speller. On the other side, a well-initialized EM classifier does not profit from LLP anymore as shown in the BCI chess section.

The EM also works well if the number of classes is rather low. This could nicely be observed when running the simulations based on the auditory ERP data where the EM performances exceeded the MIX performance. This observation can be explained by the following considerations. The EM has to solve the task of assigning a class label to each trial. The number of different options to accomplish this task, is given by C^T where C is the number of different classes and T is the number of different trials. It is easy to see that the number of options explodes with increasing number of classes and trials. Therefore, the task is easier for the EM classifier if the number of classes is small (e. g., $C = 6$ for the auditory paradigm) and if there are only a few, but very long trials (e. g., the length of the trials in the auditory paradigm is up to 90 epochs). In the visual speller on the other hand, the high number of classes ($C = 32$) despite a shorter trial length ($\leq$ 68 epochs) makes the classification problem for the EM very difficult. The chess application has an average of 10 options that can be selected per trial and with that, is somewhat in between. If we compare

this dependency on the number of classes to the LLP classifier, we observe that the LLP is actually agnostic towards the number of classes per trial because LLP is only concerned with the target vs non-target classification. For this reason, the number of classes and trial lengths should not affect the LLP mean estimations given a constant SNR and the same number of data points.

3.6.2 *Unsupervised classification on visual and auditory data*

Compared to previous studies with visual ERPs, the N200 elicited for target stimuli in the visual speller and chess application is very large [89, 169, 170], even when compared to using familiar faces as stimuli [92, 183] or motion onset [73]. It may be caused by two main factors: first, the trichromatic grid overlay is perceived as a very salient stimulus compared to traditional brightness intensifications. The short rotation of the grid may have been beneficial for the saliency as well [73], even though Tangermann et al. [164] found that most of the salience improvement compared to brightness highlighting is caused by the grid effect alone. Second, precise markers based on an optical sensor on the screen were used to determine stimulus onset time points. Compared to an alternative strategy to use markers elicited by the presentation software, jitter and delay caused by the graphics adapter and the LCD screen are eliminated by this approach which leads to better classification performance.

This high SNR in the visual data makes the unsupervised learning problem easier. To see this, imagine an extreme scenario where no noise is in the data. Then the target and non-target class will emerge naturally without the necessity of any advanced clustering method. In contrast, it is much more difficult to extract the class distributions when the data is very noisy, because the distributions are less prevalent. The high SNR in case of the visual speller is one of the reason why the MIX method could reach a comparable level compared to the supervised classifier.

In contrast to that, it is well-known that decoding auditory ERP data is much more challenging than visual ERP data because of the lower SNR [59]. The AMUSE paradigm was developed to mitigate this problem by adding spatial information to the paradigm [152]. In the original publication, 8 loudspeakers and artificial tones were used. In the best condition, young healthy subjects had a target vs non-target classification accuracy (AUC) of around 75%. Remarkably, due to the improved stimuli (bisyllabic words instead of tones) and with user training, some of our post-stroke aphasic patients could reach classification accuracies above that level — *even* when relying *solely* on unsupervised learning and without transfer learning. This is a great advance in terms of usability of auditory BCIs. It also gives hope that future auditory BCIs will have increased SNR which make the unsupervised learning problems even easier.

3.6.3 *The role of the different label proportions*

Learning from label proportions crucially depends on the possibility to include at least two groups with different target to non-target ratios into the BCI paradigm. Without this, it is not directly applicable to standard ERP paradigms. The best performance can be obtained when one group predominantly contains targets and the other group predominantly contains non-targets. In the limit, this would lead to a supervised scenario where one group only contains target events and the other only contains non-target events.

For the visual matrix speller two sequences were selected: one sequence of stimuli dominated by targets and one sequence dominated by non-targets. This specific choice of the sequences and associated mixing matrix reflects a trade-off between classifier quality, spelling matrix size and sequences lengths. Other choices are, of course, also possible. However, it is important to realize that practical limitations come into play when choosing the groups. For instance, enforcing a target ratio of $1/2$ in a visual speller requires a simultaneous highlighting of half of the selectable symbols, which may be undesired from a usability point of view. If another sequence only consists of non-targets (visual blanks) and the number of highlighted symbols needs to be matched, this would require that half the amount of selectable characters are to be added as visual blanks. This would drastically increase the number of items on the screen. Additionally, if many symbols are highlighted simultaneously, then the number of epochs required to obtain unique decodability of a character increases [173].

In the auditory ERP paradigm, it is more difficult to create groups with different target to non-target ratio because auditory paradigms provide less flexibility. It is, for instance, more difficult for the user when two stimuli are played at the same time compared to multiple visual stimuli. Also, the total number of stimuli is often much less than in visual spellers. With these limitations in place, two strategies could be used to harvest different label proportions. The first strategy is the same as in the BCI chess application from Section 3.4 where the user interface comprises a multi-step selection process, which has a varying number of items per step, to execute an action. The second strategy is to add non-target stimuli by using pseudo-words or other stimuli that the user should always ignore. The feasibility of the latter approach was demonstrated in Section 3.5.

When adding blanks stimuli (auditory or visual ones) to the paradigm, another limitation is a reduction of the spelling speed because some of the highlighting time is used for an event that is not associated with any control command. To overcome this limitation, one could consider a strategy where the LLP initially learns on the extended paradigm with blank stimuli and then, once it reached a satisfying performance level, switches back to an ordinary unmodified paradigm.

In the BCI chess game, the different groups in the data were given by the application itself. With that, one has no flexibility over the choice of the target to non-target ratio in those groups, but at the same time, no

additional symbols/fields needed to be added to the interface. The only modification that was required to unlock LLP/MIX was to replace the fixed SOA by a uniformly drawn SOA. Interestingly, most subjects did not even notice the difference and showed similar performance for both SOAs.

3.6.4 *An adaptive version with a limited time horizon*

It is known that the EEG feature distribution can change over the course of a session [98, 156, 176] and thus, violates the IID assumption, see e. g., Figure 3.1. The reason for these changes could be human factors (fatigue, motivation and learning) [88, 181] or non-human factors (drying gel leading to changing impedances, changed environmental conditions, among others). Compared to supervised methods, unsupervised methods have the distinct advantage that they can continuously learn when more unlabeled data comes. This bears the potential to adapt to changes in the data distribution over the course of a session. When realizing such an adaptive version, one needs to find a strategy to "forget" older data.

At least two questions quickly arise, namely how exactly older data should be discarded and which time frame should be considered to be from importance. Regarding the first question, many options are possible. The easiest approach would be to simply use a sliding window of a predefined length and to discard all data that is outside this window. In another approach, Vidaurre et al. [176] proposed that the updated BCI model parameters θ_{t+1} should be a weighted linear combination of the previous model parameter θ_t and the estimates of the model parameters based on the current samples $\hat{\theta}_{t+1}$:

$$\theta_{t+1} = (1 - \eta) \cdot \theta_t + \eta \hat{\theta}_{t+1}. \tag{3.22}$$

This effectively realizes an exponential decay where η regularizes the rate of the exponential decay.

Even more sophisticated approaches have been proposed without specifically targeting BCIs, but non-stationary data streams in general [104]. The key idea is to use an ensemble of different models where each model has a different forgetting strategy (e. g., exponential decays with different rates, window based, etc.) and with that, each model can adapt to different kind of non-stationary effects. Some models might be better suited to model very quick changes while others are better to model slow drifts. The different models in the ensemble are then weighted according to a score that needs to be computed on the unlabeled data. In our case, a good choice would be the data log-likelihood which measures how well each of the models can fit the data. Models with higher likelihood would then receive greater weight.

Regarding the second question to find suitable time constants (e. g., update rates η, window lengths), I believe that this problem can only be solved empirically. One way of finding the constants is to look at the ramp-up behavior of the unsupervised learning methods to get an upper bound of the data that is necessary to reach a decent performance. Based on the Figure 3.11 and the BCI chess results, around $3 - 5$ minutes are enough

for visual ERP paradigms and a substantially longer period in the range of $20 - 30$ minutes is required for the auditory data. The difference in time scales can again be explained by the difference in SNR. One should select suitable time constants according to these time scales, however, they can, of course, strongly vary depending on the application. The time constants could also be computed in a more thorough approach using cross-validation or simulations based on modified data where non-stationary effects are artificially injected into the data.

3.6.5 *Limitations*

The biggest limitation of the presented unsupervised learning approaches is that — so far — they are mostly restricted to ERP data and are not directly applicable to, e. g., motor imagery data. The reason is that they explicitly utilize the rich structure introduced by the ERP paradigms. For instance, the EM algorithm exploits that one latent variable — the selected symbol — uniquely determines all target and non-targets epochs of a trial. The LLP approach requires groups with different label proportions. In general, future work should go towards jointly adapting the paradigm and classifier by considering the user, interface and decoder as a holistic system.

I mentioned before that mistakes in the initial learning phase can be post-hoc corrected when an updated classifier is available. The prerequisite for this is that subjects continue writing their sentence or keep executing the task although they receive misleading/incorrect feedback. For some subjects, this kind of feedback leads to confusion which in turn can interfere with the use of the application because subjects think that they need to change their strategy, or they do not trust the system to successfully read out their signals. A changed behavior might then lead to changes in the data distribution which might cause instability. One solution to this problem is to not display any feedback until the classifier reached a high certainty (e. g., measured by the data log-likelihood or by the consistency of the predictions) or another solution is to very clearly explain the expected behavior of an unsupervised learning system to the user.

The MIX method is the result of combining two unsupervised learning ideas with complementary strengths [174]. While this combination has proven to be beneficial, other means of receiving information should definitely also be harvested in future approaches. This comprises the usage of a language model and transfer learning [97] as well as the exploitation of error-related potentials [185] to increase the model's capacity. Ultimately, I think that the low SNR in BCI data can only be compensated by aggregating information from different temporal and neuronal sources in combination with a careful exploitation of the underlying data constraints.

SUMMARY

The goal in this chapter was to introduce new unsupervised learning methods for brain-computer interfaces based on event-related potentials. In the beginning, I reviewed different strategies to learn from unlabeled data which showed clear evidence that unsupervised adaptation outperforms non-adaptive supervised classifiers.

I also presented two new approaches that heavily rely on the rich temporal structure in the data and which can learn to decode ERPs without requiring any labeled data. The first approach, learning from label proportions, is the first unsupervised BCI classifier that is guaranteed to converge to the optimal decoder given i.i.d. data points.

The new algorithms were tested in three different application scenarios. First, a visual speller was modified. In an online study, the best unsupervised classifier showed a very quick ramp-up behavior with almost perfect control after only 3 minutes of unsupervised learning time and average target vs non-target classification accuracies of over 90% after only 5 minutes of learning time. Similarly, I could demonstrate in simulations on data from a BCI chess game that such a quick ramp-up is also possible without modifying the user interface in certain applications. Ultimately, simulations based on data from auditory ERP experiments show that a modification in this paradigm also increases decoding performance of unsupervised learning algorithms and that they can even be successfully applied to very challenging patient data.

4

LANGUAGE REHABILITATION WITH A BRAIN-COMPUTER INTERFACE

This chapter is the result of a close cooperation with the University Medical Center Freiburg with the goal to establish a new clinical application of BCIs. Large parts of the text and of the Figures are part of a publication that is currently under preparation [127].

ABSTRACT

Aphasia refers to an impairment of language abilities mainly due to a left-hemispheric stroke. About 20% of all first stroke patients remain with a chronic communicative impairment which has a large impact on their quality of life. For the 610,000 first stroke incidents in the US alone, this translates to approx. 122,000 new chronic aphasic patients every year. For chronic patients, high-intensity language training guided by speech therapists can lead to improvements, but effect sizes are rather limited and generalized training effects are difficult to obtain.
In this chapter, a new BCI-based language training for aphasia is proposed. In this online BCI training, patients were asked to infer a target word based on a spoken sentence with a missing word in the end. They then heard a rapid sequence of words and were asked to detect the appearances of the target word while ignoring non-target words. After each trial, patients received feedback based on how well the attended word – as classified based on auditory ERP responses – did match the target word. We tested the feasibility and effectiveness of the new training protocol in 10 stroke patients with different levels of aphasia. Per patient, we conducted about 30 hours of high-intensity training which typically took 4 to 5 weeks. The primary endpoint for the study was the pre-post comparison in the Aachener Aphasie Test which is a standardized clinical language assessment. In this language test, patients showed strong, generalized, significant and persistent language improvements. Although patients' brain responses were delayed before the training, we found that patients' ERP responses showed a timing comparable to those of 20 normally-aged (healthy) controls after the training, which indicates an improved word processing speed. These findings may open the door for a completely new application field of BCIs with an enormous potential user group.

4.1 INTRODUCTION

Aphasia refers to an impaired ability to understand or produce language, as a result of brain damage. Its leading cause is a left-hemispheric stroke with about $21-38\%$ of stroke patients experiencing aphasia [8, 47]. In Germany alone, there are $196,000$ patients with a first stroke and $66,000$ repeated strokes every year [67] leading to around $40,000-80,000$ new aphasic patients annually. In the US, the number of first stroke incidents is estimated to be around $610,000$ [11]. Spontaneous functional recovery from post-stroke aphasia is often observed, however, it is estimated that 20% of patients survive with persistent communicative impairments [42].

Aphasia has a large negative impact on the quality of life [68] as it often results in a loss of independence and reduces the likelihood of returning to work [44]. In a survey with over $60,000$ patients in long-term care, a study found that the presence of aphasia shows the largest negative relationship to the quality of life, ahead of cancer, Alzheimer's disease, Huntington's chorea and quadriplegia [108]. Compared to stroke patients without language deficits, the mortality rate in aphasic patients is twice as high [109] and the incidence of major depressions is three times as high [93].

Patients with aphasia spontaneously recover to a large extent within the first 6 months, but only minimal further improvements are reported in the chronic phase thereafter [14]. It was shown that speech and language therapy (SLT) can help in the chronic phase, especially for functional communication, reading and writing when compared to a non-aphasia therapy program [27, 167]. However, general effect sizes of SLT are small or moderate at most [27, 28, 167]. In addition, a major limitation of current SLT is that complete recovery is often not achieved [27, 42, 137] and that severely affected patients with a global aphasia (incidence rate of about 2.5% [150]) show a bad recovery and therapy-resistance [27, 167].

In the past decade, new approaches were developed that use non-invasive brain stimulation techniques to modulate cortical excitability via repetitive transcranial magnetic stimulation (rTMS) or transcranial direct current stimulation (tDCS) in combination with SLT. While no clear benefit was found for tDCS in the latest Cochrane review [46], a positive training effect was found for rTMS in a recent meta-review [148], although most participating patients were in the subacute stage and long-term effects are still unknown.

Due to the aging population, increasing survival rates after initial strokes [56] and the limited success in preventing stroke incidences [45], stroke-related costs are going to increase in the future [135]. The combination of high costs, the severe negative impact on the quality of life, a high aphasia incidence rate and the modest successes of language therapies make aphasia a top ten research priority for life after stroke [143]. This calls for new evidence-based effective interventions for post-stroke aphasia.

In this study, we present the results of a novel therapeutic approach for rehabilitation of aphasia based on an ERP-based BCI.

The idea to use a BCI for rehabilitation is not new. Especially the usage of BCIs in motor rehabilitation has shown encouraging results. The principal idea of BCI-based motor rehabilitation is that immediate sensory feedback via functional electrical stimulation [32, 138], a virtual avatar [139] or a robotic/orthotic device [146] is triggered, if a movement intent (e. g., of the hand) was detected by the BCI. Such closed-loop training protocols aim at reinforcing the interaction between efferent and afferent pathways of the brain. Compared to traditional physiotherapy, they have the advantage that the movement intention can be detected even if its execution is infeasible. Moreover, afferent feedback provided by the BCI is time-locked to the movement attempt and not mediated by an external person (the physiotherapist). The results are promising: a meta-analysis for upper limb rehabilitation found a medium to large effect size [32] and another study provided compelling evidence that the BCI is indeed the key for the rehabilitation effect [15].

In this contribution, we explore the possibility of using a BCI for aphasia rehabilitation. We postulate that a BCI approach may elicit great therapeutic effects for aphasia recovery when continuously providing an attention-constrained and causal feedback about the ongoing language-related activity in the brain. Despite the enthusiasm for BCIs in the last years, no successful BCI-based language training could yet be established. The main difficulty is that it is vastly unclear, which neural markers are sufficiently language-specific and capture a broad range of language functions at the same time, to realize a brain-state dependent closed-loop training. Decoding the intended speech, which would be the direct analogy to BCI-based motor rehabilitation, is not possible with state-of-the-art non-invasive BCIs due to the limited spatial resolution. While some language-specific potential responses can be measured in the EEG, e. g., the N400 component, which is assumed to be related to semantic processing, our concern is that reinforcing this specific feature may not lead to a sufficiently generalized training effect. Other known components (e. g., the N200 or P300) are not language-specific, and therefore may not be sufficiently definite to realize a successful training. In addition, we also face the problem that the signal-to-noise ratio of the recorded components is rather poor, making it very difficult to reinforce specific components in isolation.

In auditory BCI paradigms, simple two-tone oddball paradigms were used originally where the user was instructed to attend a high tone while ignoring a low tone [162]. From that, several improvements have been made. It was found that the task can be made easier by adding spatial information which was realized by playing each sound from a distinct loudspeaker [152]. This paradigm called Auditory MUlti-Class Spatial ERP (AMUSE) is also used in this contribution. It was also found that the stimulus onset asynchrony (SOA), meaning the time that passes between the onset of two consecutive stimuli, can be reduced from original values of 1 second or longer to SOAs as short as a few hundred milliseconds, e. g., [69]. A final step was to move away from the simplistic artificial low and

high tones to more naturalistic sounds, such as animal sounds [9, 65, 157], spoken syllables [71] or words [120, 155, 163].

With these improvements in place, we designed a fast auditory BCI where bisyllabic word stimuli were played from different loudspeakers and the patients were instructed to attend a cued target word while ignoring other non-target words. This task was designed such that it should elicit language-related ERPs that, however, are not related to a *specific* aspect of language processing. By giving feedback on how well patients accomplish this task, we expected to reinforce language processes that lead to discriminative target vs non-target ERPs. We hypothesized that this may lead to a generalized language improvement.

However, as none of the previous auditory BCIs was tested in aphasic patients, it was an open empirical question, if the single-trial decoding of ERPs from rapid word stimuli in aphasia patients is feasible at all. Another open question is if the reinforcement of task-specific word ERPs would rather lead to an improvement of general attention and working memory — which clearly are needed for the task — or if it would lead to a specific improvement of language competences.

In summary, this study examined the following research questions.

1. Is the new BCI-based training protocol feasible for patients with chronic aphasia?
2. Does the training lead to a generalized, persistent language improvement?
3. How is the training affecting word ERPs?
4. Is the training specifically training language competences or is it a rather general attention training?

4.2 METHODS AND SUBJECTS

4.2.1 *Patients*

We recruited 10 patients with chronic aphasia (>6 months) after a left-hemispheric brain stroke. Most patients were mildly affected by aphasia, but we also included 2 patients with global aphasia. In order to participate in the study, patients fulfilled the following criteria.

Inclusion criteria: informed consent; aged between $18 - 80$; presence of aphasia in AAT; right-handed; first-ever ischemic stroke; time onset of stroke at least 6 months ago; German as mother language; sufficient cognitive functions to comply with study requirements.

Exclusion criteria: hemorrhagic stroke; other structural brain or skull lesions (tumor, trauma) in MRI; severe cerebral microangiopathy; high cerebral artery stenosis; implanted medical devices or intracranial ferromagnetic objects; cognitive impairment and other medical, neurological or psychiatric disorders interfering with participation; the patient is early bilingual or professional musician; hearing loss or loss of vision; severe adverse skin reaction caused by EEG recordings.

A detailed description of the patients is given in Table 4.1.

4.2.2 *Study protocol and endpoints*

Patients that have been found eligible for the training in the screening phase, followed the study protocol described in Figure 4.1. They started with a familiarization phase in which the training task was practiced. Afterward, and at multiple time points throughout the training, language and neuropsychological abilities were assessed.

As a primary endpoint, we used the Aachener Aphasie Test (AAT) [74], a standardized language test in German. We assessed this endpoint five times for each patient (at first presentation (which was already in the chronic phase), pre-training, mid-training, post-training, follow-up after 3 months). The AAT contains six subtests: (1) Token Test which is a complex test addressing auditory comprehension, working memory and semantic understanding, (2) repetition of words, (3) written language, (4) naming objects, (5) comprehension and (6) spontaneous speech. The last aspect is measured on a different scale compared to the other subtests and hence, is always analyzed individually.

As secondary endpoints concerning verbal abilities, we took a picture naming test based on 233 items of the Snodgrass & Vanderwart (S&V) picture set [159] and assessed functional everyday communication by using the communication activity log (CAL) [144] before and after the training. To assess changes in cognitive abilities, we used the digit span test [4], the block-tapping test which tests working memory [94], a word fluency test (Regensburger Wortflüssigkeits-Test [3] and the TAP [188] before and after the training.

General information				Stroke-related information								Aphasia-related information			Others
Patient	Sex	Age	Ed. age	Stroke etiology	Stroke risk factors	Stroke severity (mRS) at To, T1, T2	Infarct volume (ml)	MCA stroke location	Additional stroke location	Months post-stroke at training start	Hemi-paresis (severity)	Aphasia severity	AAT-based aphasia subtype	Speech apraxia severity	Comorbidity
1	m	76	11	CE, LAA	H,AF,D	3/2/2	113	F,T,P,In		10		moderate	Broca		
2	m	58	17	LAA	H	4/4/2	13	F,P,In,NC	ACA,AChA	18	severe	minimal	anomic	mild	
3	m	71	23	LAA	H,CHD	4/3/2	43	F,T,P	ACA	36	mild	mild	anomic		epilepsy, MM
4	m	70	11	EO	H	4/4/3	47	F,T,P,In,		9	severe	mild	Broca		prostate cancer
5	m	60	12	LAA	H,HL,N	5/4/2	68	F,T,P,In,NC		27		mild	anomic		
6	m	43	19	LAA	H,HL	3/3/2	125	F,T,P,In,NC	PBZ	10	severe	mild	Broca	moderate	epilepsy
7	w	54	23	ICA-D		5/4/2	100	F,T,In,NC	ACA	8	mild	mild	anomic		depression
8	m	61	17	ICA-D	H,HL	3/2/1	87	F,T,In		149		mild	anomic		
9	m	38	12	CE		5/3/3	217	F,T,P,In		21		severe	Broca	mild	
10	m	53	12	CE	H,HL	5/5/3	145	F,T,P,In		12	severe	severe	global	mild	depression

Table 4.1: **Overview of patient-specific information.**
Ed. age refers to the educational age, i. e. the number of years in school and in higher education.
Stroke etiology of (ischemic) stroke subtype: cardioembolism (CE), large-artery atherosclerosis (LAA), small-vessel occlusion stroke (SVD), internal carotid artery dissection (ICA-D), EO=embolic undetermined etiology.
Risk factors: atrial fibrillation (AF), coronary heart disease (CHD), diabetes (D), hypertension (H), hyperlipidaemia (HL), nicotine (N).
Stroke severity was assessed with the modified Rankin Scale (mRS) at admission (To) / discharge (T1) / before training (T2).
Location of stroke: all patients had an infarct of middle cerebral artery (MCA). Within the MCA, the infarct involved areas from frontal (F), temporal (T), Insula (In), parietal (P), nucleus caudatus/ thalamus (NC) regions.
In some patients, also the anterior cerebral artery (ACA), anterior choroidal artery (AChA) or posterior border zone (PBZ) were affected.
AAT-subtype. Anomic: mild form of aphasia with difficulties to name objects; Broca: partial loss of the ability to produce language (spoken and written); global: most severe form of aphasia heavily affecting comprehension and production.
Apraxia of speech refers to a disorder which affects an individual's ability to translate conscious speech plans into motor plans.
Comorbidity: MM=multiples myeloma.

In addition, two EEG-sessions without feedback were also conducted before the training to tune stimulation parameters (e. g., adjust the difficulty level according to the patient and train a supervised classifier) and a single session was conducted after the training to be able to compare the training-induced changes in the EEG. We also conducted several resting-state fMRIs but please note that their evaluation is not part of this work.

The training protocol was designed to be in high-intensity, i. e. patients should train 4 times per week over the course of $4-5$ weeks. In terms of training duration, our goal was to reach 30 hours of effective training time which is comparable to other language trainings in the chronic phase [28].

The exact arrangements of the tests and the time course is displayed in Figure 4.1**a**.

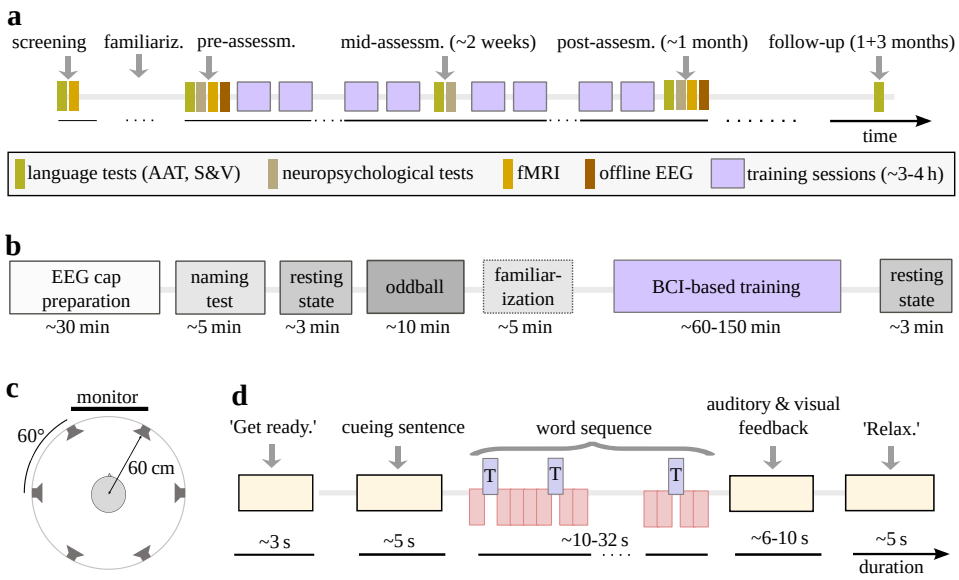


Figure 4.1: **Study protocol for BCI-based language training. a:** overview of the different clinical testings and training sessions that each patient undergoes. AAT = Aachener Aphasie Test, S&V = Snodgrass & Vanderwart naming test, **b:** structure of a single training session. **c:** setup of the AMUSE protocol: a subject is placed in the center of a ring of 6 loudspeakers [152]. **d:** time course of a single trial of the training task. During the word sequence, there are non-target words (red rectangles) and a few (rare) target words (blue rectangles).

4.2.3 *Structure of a single session*

The time course of a single training session is depicted in Figure 4.1**b**. Brain activity was recorded and amplified by a multichannel EEG amplifier (BrainAmp DC, Brain Products) with 63 passive Ag/AgCl electrodes (EasyCap) during the offline EEG-sessions before and after the training and with 31 passive electrodes during the online training sessions. The channels were placed according to the $10-20$-system referenced against the nose and grounded at channel AFz. The sampling rate was $1\,kHz$. Impedances were always kept below $20\,k\Omega$. Eye signals were recorded

by electrooculography (EOG) with an electrode below the right eye of a subject.

At the beginning of each session, patients named around $25 - 30$ pictures which neither had an overlap with pictures used in the Snodgrass & Vanderwart nor with the AAT naming test. This was done to assess the daily naming performance but is not evaluated in this work. Afterward, a resting state recording with one minute of open and closed eyes was recorded. Then, patients underwent two non-verbal ordinary oddball runs with a high or low tone played every second. Each run took about 5 minutes and contained 50 targets (high tones) and 250 non-targets (low tones).

4.2.4 *Training task and feedback*

As a training task, we chose a target vs non-target detection task based on rapid bisyllabic word stimuli. This is a modified version of the AMUSE protocol [152], see Figure 4.1**c**, where subjects are placed in a ring of 6 loudspeakers. An individual training trial (Figure 4.1**d**) started with a familiar cueing sentence with a missing word in the end, e. g., "*the toner cartridge is already in the . . .* ". The missing word (hereafter called target) in that case is "*printer*". The patients should remember the target word, but do not need to speak it. Afterward, they hear a rapid sequence of words which contains the target word and 5 other (non-target) words. Each word is played from a single distinct loudspeaker. The subject has the task to attend/recognize all playbacks of the target word while ignoring the other words. Words (length = 300 ms) were played with an SOA of 250 ms or 350 ms depending on the patient and a trial consisted of a maximum of 90 words (15 targets, 75 non-targets) that were played in a pseudo-randomized order. Each word was once in the target role for every run, i. e. within 6 trials. If patients performed very well, we also changed from 6 loudspeakers to headphones which removes the spatial information to make the task more challenging.

At the end of a trial, patients receive feedback based on how well target stimuli could be discriminated from non-target stimuli based on the recorded brain responses. This was done by using a regularized supervised LDA classifier that is based on amplitude features of the ERP responses. For each event, this linear classifier yields a real-valued output that can be understood as a likelihood that this event was a target (for positive outputs) or non-target (for negative outputs). These single classification output are aggregated over the course of a session (see Figure 4.2). Please find more information about the BCI classification in Chapter 2 with details about the ERP features (Section 2.4.3), LDA classifier (Section 2.4.4), supervised learning (Section 2.5.1), covariance regularization (Section 2.5.2) and the aggregating of information over a trial (Section 2.6).

Patients received four different levels of graded feedback at the end of the trial.

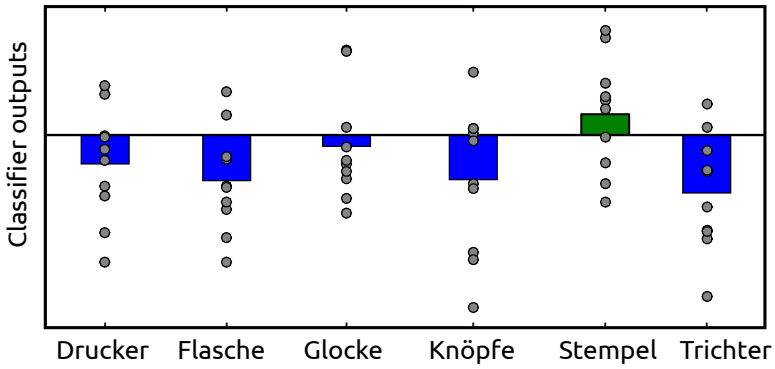


Figure 4.2: **Example of the classifier outputs of a single trial.** Patients hear 6 German different words (x-axis). Each time they hear a word, the LDA classifier outputs a value (the single grey dots) that indicates the class membership: a high value indicates that the word was more likely to be a target while a low value indicates that the word was more likely to be a non-target. The goal of the patients is to generate high values for the target word (green) while the non-target words (blue) should obtain low values.

1. **Neutral feedback.** If our classifier could not successfully decode the target word, but predicted that a different word was attended, then we gave neutral auditory feedback and displayed a bar chart on the screen similar to Figure 4.2 which indicates which other word was most dominant.

2. **Positive feedback.** If we could successfully detect the target word (e. g., shown in Figure 4.2), we gave positive auditory feedback and also provided the bar chart to indicate the best runner-up word.

3. **Very positive feedback.** We used a dynamic stopping routine [153] that stops the trial if the classifier displayed a significant difference between the target word and the best non-target word as assessed by a one-sided Welch's t-test. In this case, the classifier is very confident that the patient executed the task correctly. As a feedback, we then showed a smiley and gave very positive auditory feedback.

4. **Exceptional feedback.** The early stopping routine was activated after 7 repetitions of the target word. If the classifier could detect significant differences already at this early stage, we showed a cheering animation with a probability of 33%. This was only triggered at most a few times per session.

4.2.5 *Transfer learning and supervised adaptation*

The training protocol is special from a BCI point of view because the data is completely labeled, i. e. for each trial we know the target word. This information can be used to initialize the classifier via a session-to-session transfer learning approach by training a supervised classifier based on the

data of the two offline sessions (in the first online session) or based on the previous online session (in all other online sessions).

However, we observe that the data distribution may also change over the course of a session (e. g., Figure 3.1 in the previous chapter) which could, among other reasons, be caused by a changed task-solving strategy. To account for these changes within one session and to be able to adapt the classifier to new task-solving strategies, we used a supervised mean and (unsupervised) covariance adaptation based on the work by Vidaurre et al. [176]. In their approach, the class-wise means $\boldsymbol{\mu}_i(t)$ with $i \in \{\text{Target}, \text{Non-target}\}$ and inverse of the pooled covariance matrix $\boldsymbol{\Sigma}(t)^{-1}$ at time point t are updated after each epoch by using the mean and covariance at time point $t-1$ and the data of the current epoch $\mathbf{x}(t)$. Please note that the mean $\boldsymbol{\mu}_i(t)$ is only updated if the current epoch belongs to class i.

$$\boldsymbol{\mu}_i(t) = (1-\eta_m) \cdot \boldsymbol{\mu}_i(t-1) + \eta_m \cdot \mathbf{x}(t) \tag{4.1}$$

$$\boldsymbol{\Sigma}(t)^{-1} = \frac{1}{1-\eta_C} \cdot \left(\boldsymbol{\Sigma}(t-1)^{-1} - \eta_C \cdot \frac{\mathbf{v}(t)\mathbf{v}(t)^T}{1-\eta_C + \eta_C \cdot \mathbf{x}(t)^T \mathbf{v}(t)} \right) \tag{4.2}$$

We used that $\mathbf{v}(t) := \boldsymbol{\Sigma}(t-1)^{-1} \cdot \mathbf{x}(t)$. This describes an efficient exponential update rule where the update coefficients $\eta_m = 0.001$ and $\eta_C = 0.005$ were determined by an initial grid search. The advantage of this update scheme is that only the current means and covariance matrices need to be stored and no costly matrix inversion needs to be executed for each step.

4.2.6 *Performing the task with eyes-closed*

During training trials, patients are instructed to avoid eye movements and eye blinks. This should minimize the number of artifacts in the EEG, however, this secondary task can lead to exhaustion and subjects may not succeed in suppressing eye movements. This is especially challenging for patients, e. g., one patient could not control the BCI with eyes-open due to the occurrence of too many eye artifacts. Therefore, we investigated the option to control an auditory BCI with eyes-closed. In 12 healthy young subjects, we found that the number of eye artifacts was actually not reduced in the eyes-closed condition, but subjects expressed a significant general preference towards the eyes-closed condition and were also less tensed in that condition while having classification accuracies similar to eyes-open. Please find the complete study in the Appendix a. Based on these results, it seems evident that patients could also profit from performing the task with eyes-closed and for this reason, we allowed single patients to perform the task with eyes-closed which is unusual in BCIs.

4.2.7 *Healthy controls*

In addition to the patient study, we also did a single EEG session with normally-aged controls (NACs). Data of 20 elderly participants (10 female,

10 male, $M_{age} = 60.20$ years, $SD = 8.04$, range: $48 - 74$ years) was recorded and analyzed. NACs performed the same paradigm as the patients except for the following few modifications.

- They did not receive feedback at the end of the trial. Data was only analyzed offline.
- All NACs used an SOA of 250 ms and words were played either from 6 loudspeakers, one loudspeaker, or headphones in a pseudo-randomized order. For each condition, we recorded 36 trials with 90 words each (leading to $36 \cdot 90 = 3240$ epochs). In the following, we only analyze the condition with the 6 loudspeakers.
- They performed only a single session and not a complete training.

4.2.8 *Statistical evaluation*

For all endpoints, we calculated the effect size as the standardized mean difference (SMD) as recommended by the Cochrane research group [27] as follows.

$$\text{effect size} = \frac{\text{difference in mean outcome between groups}}{\text{standard deviation of outcome among participants}} \tag{4.3}$$

More specifically, let $\mathbf{z}_{\text{pre}}$ and $\mathbf{z}_{\text{post}}$ denote the test results before and after the training for N patients, respectively. The numerator is always computed as the difference of the means of both group. In the denominator, the standard deviation needs to be calculated. We follow the Cochrane research group [27] which recommend using Hedges' g_s as the effect size.

$$g_s = \frac{\bar{\mathbf{z}}_{\text{post}} - \bar{\mathbf{z}}_{\text{pre}}}{\sqrt{\frac{SD(\mathbf{z}_{\text{pre}})^2 + SD(\mathbf{z}_{\text{post}})^2}{2}}} \cdot \left(1 - \frac{3}{8 \cdot N - 9}\right) \tag{4.4}$$

The first term corresponds to Cohen's d_s while the second factor is a normalization constant that corrects for biases in d_s for small samples sizes ($N < 20$), see [107]. Please note that the difference between d_s and g_s is rather small (around 5% for $N = 10$). Please also note that the standard deviation does not need to be estimated for the T-transformed AAT scores because they have been normalized to a standard deviation of 10. For this reason, we calculate the effect size d^* for the T-transformed AAT scores as follows.

$$d^* = \frac{\bar{\mathbf{z}}_{\text{post}} - \bar{\mathbf{z}}_{\text{pre}}}{10} \tag{4.5}$$

Empirically, our standard deviation was very close to 10, too.

To assess significance, we used two-sided paired t-tests for normally distributed quantities and exact Wilcoxon-signed rank sum tests otherwise. All tests are specified within the result section.

4.3 RESULTS

4.3.1 *Primary endpoint (AAT)*

Our primary outcome for the training success is the AAT score, a standardized clinical German language test with 6 subtests. Following common practice, raw AAT scores were transformed into T-scores that are normally distributed with mean 50 and a standard deviation of 10. The results for all subtests are displayed in Table 4.2 and Figure 4.3 shows the results for 5 subtests (spontaneous speech is measured on a different scale and cannot be transformed to T-transformed scores). A significant and consistent training effect can be observed for all AAT subtests, even after applying a Benjamini-Hochberg correction to control the false discovery rate, see Figure 4.3**a** and Figure 4.3**b**.

Table 4.2: **Training effects from baseline to after 30 hours of high-intensity BCI-based training for the primary endpoint.** Raw p-values are reported. Effect sizes are calculated as the mean difference divided by the population standard deviation (which is taken as 10 for the T-scores in d^*, see methods) and computed as Hedges' g_s for spontaneous speech.

Test name	N	Pre-training	Post-training	p-value	Effect size
Aachener Aphasie Test		Mean (SD)	Mean (SD)	Paired t-test	d^*
Token Test [T-scores]	10	58.9 (13.25)	63.3 (11.40)	**0.0023**	0.44
Repetition [T-scores]	10	56.9 (8.33)	60.7 (9.31)	**0.023**	0.38
Written language [T-scores]	10	55.4 (7.46)	62.0 (10.87)	**0.0033**	0.66
Naming test [T-scores]	10	56.5 (8.28)	67.5 (13.48)	**0.0019**	1.1
Comprehension [T-scores]	10	59.1 (8.13)	64.8 (13.26)	**0.0232**	0.57
AAT: spontaneous speech		Median (range)	Median (range)	Wilcoxon signed-rank sum test	Hedges' g_s
All spont. speech subtests [sum]	10	24 (14-29)	26 (16-30)	**0.0072**	0.52

The naming ability showed the largest improvements with an effect size of above 1. Across all five categories (except spontaneous speech), the average effect size is SMD = 0.63 (SD across subject $\pm$0.36).

Before our intervention, most patients had regular conventional speech and language therapy (cSLT). For all patients except P3, we have an additional measurement point at first consultation (in the chronic phase) which was an average of 169 days ($\pm$149 days) before training start (see Table 4.3 and Figure 4.3**c**). None of the improvements during that period were significant at a α-level of 0.05. This was the case despite that all except one patient (P8) underwent language therapy at least twice per week resulting in an average of 30 hours ($\pm$32 hours) of SLT before training start when assuming an average of 40 weeks of training per year.

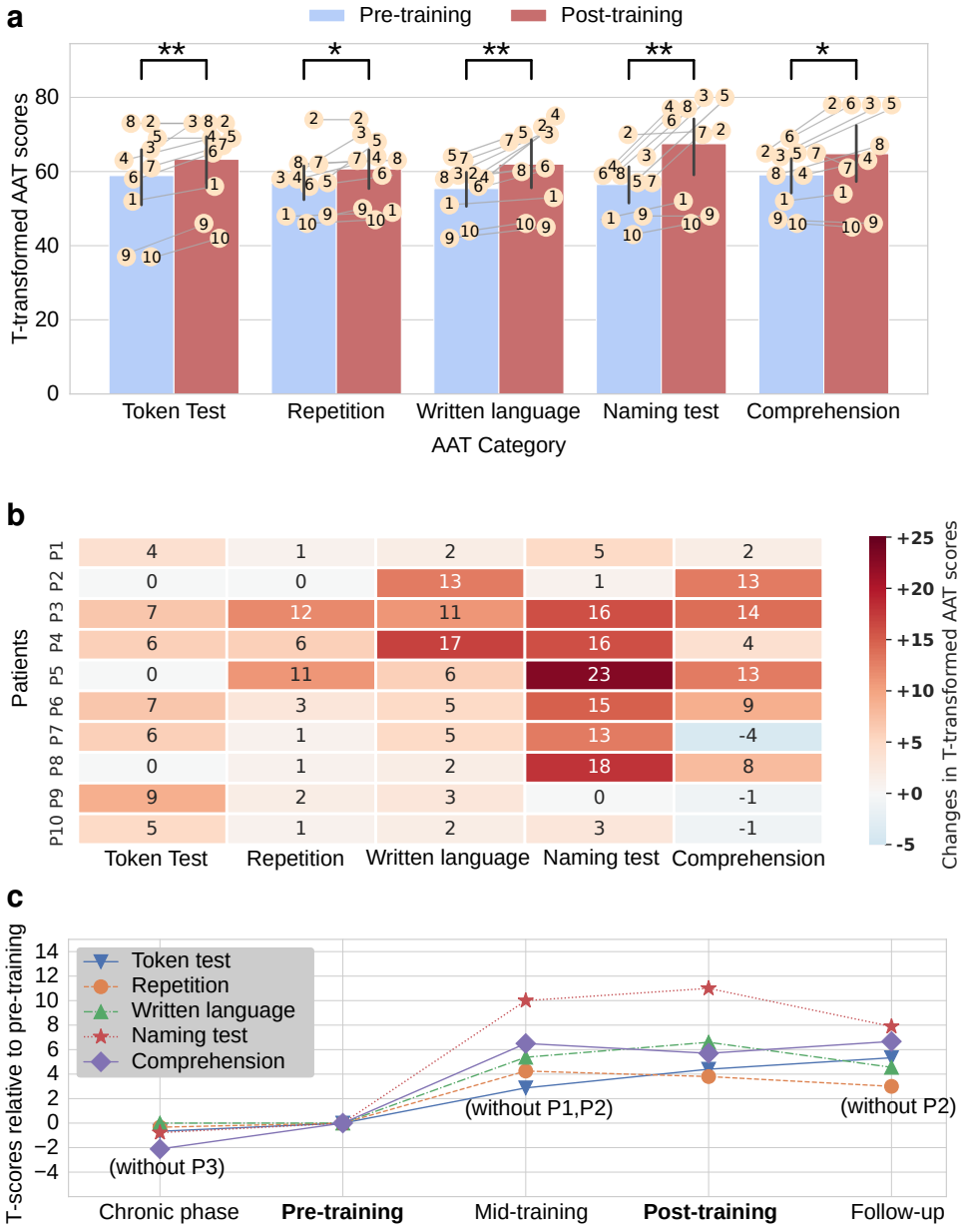


Figure 4.3: **Changes in language functions measured by the Aachener Aphasia Test.** Subplot **a** shows individual (dots) and groupwise changes (bars with standard deviation) of the language functions measured by the T-transformed AAT-scores. The AAT is normalized such that 10 T-transformed AAT points correspond to one standard deviation. Numbers denote individual patients. Significance was assessed by two-sided paired t-tests with Benjamini-Hochberg correction: * marks $p < 0.05$ and ** marks $p < 0.01$. Subplot **b** shows the improvements in T-scores for each subject and AAT category. Subplot **c** shows the average language performance on the group level for five different time points relative to the pre-training performance. Missing data points are annotated and were excluded from the computation of the averages and the statistical tests.

	cSLT between first AAT in chronic phase and BCI-training start		BCI-training		Total AAT points and severity of aphasia according to AAT					
						Pre-training		Post-training		
Patient	Times per week (*45 min)	Duration in days	Number of sessions	Duration in hours	AAT points before cSLT	AAT points	Severity	AAT points	Severity	AAT points at follow-up
1	3	103	15	35.3	321	318	medium	374	medium	374
2	2	356	11	24.2	490	495	mild	517	no aphasia	x
3	x	x	14	30.0	x	471	mild	523	no aphasia	518
4	2	48	17	29.7	470	457	mild	503	no aphasia	475
5	3	57	11	29.3	448	467	mild	519	no aphasia	508
6	3	423	13	30.0	430	448	mild	494	mild	473
7	3	64	25	30.2	446	468	mild	492	no aphasia	504
8	0	300	15	29.5	473	466	mild	498	mild	503
9	4	48	15	30.4	243	245	severe	276	severe	291
10	3	117	13	30.0	181	198	severe	240	severe	254
Avg (SD)	2.6 (1.1)	168 (149)	14.9 (4.0)	29.9 (2.6)	389 (113)	403 (108)		444 (107)		433 (101)

Table 4.3: **Summary of training and aphasia-specific patient data.** The sum of the AAT points for all subtests (except spontaneous speech) are reported. A total of 530 points can be achieved. The classification of aphasia severity is according to the AAT. Abbreviations: 'x' = missing values, cSLT = conventional speech and language therapy.

A follow-up assessment at 3 months after the training showed that these improvements remain relatively stable, although small fluctuations occur (see Figure 4.3**c**). Patient P2 was not available for a follow-up due to an accident that was not related to the training. When comparing the follow-up to the pre-training performance with Benjamini-Hochberg correction with a two-sided paired t-test, then the improvements were still highly significant for the Token Test ($t(8) = 4.208$, $p = 0.001$), written language ($t(8) = 3.957$, $p = 0.001$), naming test ($t(8) = 4.318$, $p = 0.006$) and significant for comprehension ($t(8) = 2.731$, $p = 0.03$). The changes in repetition was not significant anymore ($t(8) = 2.736$, $p = 0.052$).

Another view on the data is in terms of the total number of raw AAT points which allows judging the current language ability of each patient based on a single value. Comparing post- vs pre-training raw AAT scores, an average of 49% of the maximal possible change (MPC) could be realized (see Figure 4.4). Especially, mildly-affected patients showed remarkable improvements when considering their maximum possible improvement. According to the AAT criteria, 5 out of 10 patients were not aphasic anymore post-training (see Table 4.3).

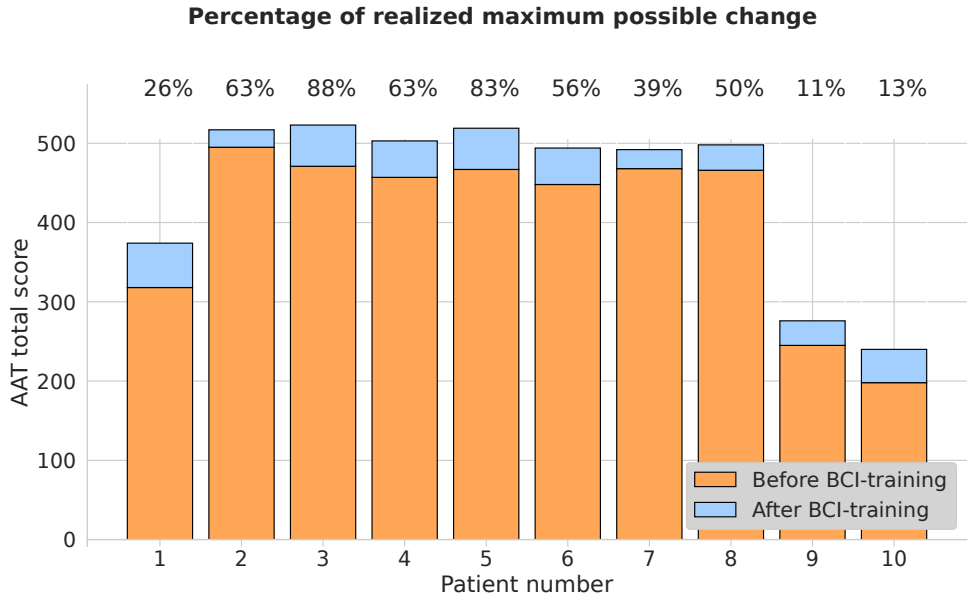


Figure 4.4: **Realized maximal possible change per patient.** The numbers indicate the percentage of recovery that could be achieved considering that the maximal number of AAT points is 530.

4.3.2 *Secondary endpoints*

In addition to the AAT, we assessed various other secondary endpoints. Table 4.4 shows the endpoints regarding naming ability, functional communication and cognitive abilities.

Table 4.4: **Secondary endpoints of the BCI-based language training.** Uncorrected p-values are reported. **Abbreviations**: S&V = Snodgrass & Vanderwart naming test, see [158], Wilcoxon test = Wilcoxon signed-rank test, CAL = communication activity log.

Test name	N	Pre-training	Post-training	p-value	Effect size
S&V naming test		Mean (SD)	Mean (SD)	Paired t-test	Hedges' g_s
Correct Words [in %]	9	53.1 (22.06)	59.5 (20.86)	**0.042**	0.28
Functional communication (CAL)		Mean (SD)	Mean (SD)	Paired t-test	Hedges' g_s
Quantitative [sum]	10	28.9 (11.10)	34.3 (11.45)	**0.0003**	0.46
Qualitative [sum]	10	76.9 (25.45)	84.4 (24.48)	**0.0002**	0.29
Cognitive tests		Median (range)	Median (range)	Wilcoxon test	Hedges' g_s
Digit span [total count]	9	8 (0-12)	9 (0-13)	0.8867	-0.03
Go/NoGo [number of errors]	10	4 (1-31)	4 (0-36)	0.726	-0.01
Go/NoGo [ms]	9	548 (417-944)	596 (461-700)	0.7344	-0.06
Alertness without signal [ms]	10	248 (201-358)	282 (218-483)	**0.0371**	0.59
Alertness with signal [ms]	10	263 (188-374)	287 (218-366)	0.5071	0.27

4.3.3 *Naming results*

The extensive Snodgrass & Vanderwart [158] was used to get a detailed clinical assessment of the patient's ability to name pictures. We used a subset of 233 images from the corpus of 260 images. The following Table 4.5 shows the percentage of correctly named words at four different time points during the course of the training. In the direct pre-post comparison, the percentage of images that could correctly be named improved from 53% before the training to around 60% during post-training and follow-up assessment. A paired t-test indicates that this is a significant improvement, see Table 4.4.

Table 4.5: **Percentage of correctly named words based on the Snodgrass & Vanderwart naming test.** The test comprises 233 images that should be named by the patients. 'x' denotes missing values.

Patient	Pre-training	Mid-training	Post-training	Follow-up (3 months)
P1	9.44%	18.03%	16.31%	19.31%
P2	73.82%	x	73.94%	x
P3	59.23%	62.66%	58.80%	58.80%
P4	39.83%	61.37%	65.95%	69.53%
P5	59.23%	63.95%	64.22%	67.81%
P6	70.31%	71.98%	72.10%	76.77%
P7	68.70%	71.12%	75.86%	75.97%
P8	67.81%	69.96%	75.00%	75.54%
P9	29.61%	30.04%	33.48%	36.05%
P10	x	x	x	x
AVG	**53.11%**	**56.14%**	**59.52%**	**59.97%**
SD	**22.06%**	**20.46%**	**20.86%**	**21.25%**

4.3.4 *Functional communication*

We also asked patients to report on the quality and quantity of language use in the real-world using the communicative activity log (CAL) questionnaire [144]. As their partner assisted in the completion process, we required that the part of CAL which is normally completed by an external person should not be the partner, but e. g., by a speech therapist. Unfortunately, we did not receive a sufficient amount of these external ratings and thus, only report on the self-rated part. A two-tailed paired t-test shows that the self-reported changes are highly significant for the quality and quantity of language use with a consistent improvement for all patients (Figure 4.5).

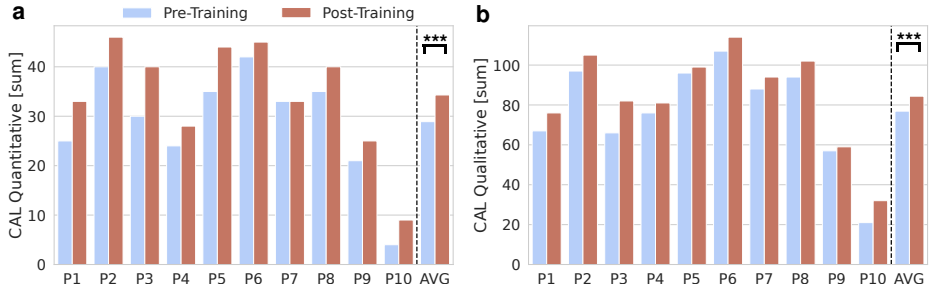


Figure 4.5: **Self-reported everyday communication measured by the communication activity log (CAL).** Subplot **a** shows the quantity and subplot **b** shows the quality of language use in everyday situations as reported by the patients. Significance was assessed with a two-sided paired t-test where *** indicates $p < 0.001$.

4.3.5 *Cognitive tests*

Before and after the training, patients underwent a series of cognitive tests regarding their working memory and attention, among others. We statistically evaluated a subset of five quantities.

1. Working memory as measured by the digit span test [4] which requires patients to repeat a sequence of numbers (either in the same order or in a reversed order).

2-3. Selective attention as measured by a visual Go/NoGo-task (pressing a button only if a cross appears on the screen, but not if a plus appears) in terms of reaction time and the number of errors (part of the TAP [188]).

4-5. Alertness as measured by the median reaction time to a visual stimulus with and without a prior (auditory) warning signal (part of the TAP).

Pathological performance is defined as a performance below the 15^{th} percentile rank. Before the training, we found that patients show pathological performances for all categories with 4/10 patients showed pathological

performance for alertness, while up to 9/10 patients showed it in the working memory task. Comparing post-pre training performances, a two-sided Wilcoxon-signed rank test showed a significant increase ($p < 0.05$) for the reaction time in the alertness test without warning signal. All other categories had p-values > 0.05. After correcting for multiple testing, none of the changes were significant anymore (see Figure 4.6).

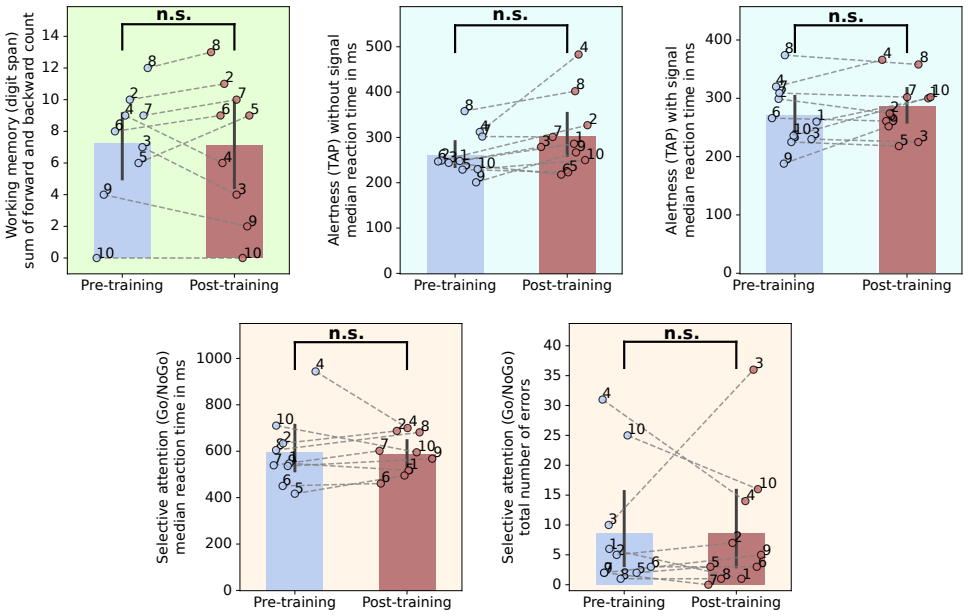


Figure 4.6: **Cognitive test results for tasks regarding working memory, alertness and attention.** Single points indicate individual patients. Please see the text for an explanation of the different tests. The p-values have been corrected for multiple testing with Bonferroni-Holm correction. **n.s.** = not significant, **TAP** = test of attentional performance.

4.3.6 *Word-induced ERP responses*

We also assessed the ERP responses to word stimuli before and after the training and for 20 normally-aged controls which underwent a single BCI session (see methods for more information). The data provides clear evidence that the training has led to changed target ERP responses (Figure 4.7). Four out of six categories showed significant changes after p-values were corrected with Benjamini-Hochberg to control the false discovery rate (see Figure 4.7**c**). Specifically, patients showed a significant increase in P300 peak amplitude in channel Cz after the training (two-tailed paired t-test, $t(9) = 3.35$, $p = 0.023$), an earlier onset of the P300 in Cz (Wilcoxon signed-rank test, $Z = 0$, $p = 0.023$, two ties), and an increase in target vs non-target classification accuracy (Wilcoxon signed-rank test, $Z = 54.0$, $p = 0.023$). On the other hand, N200 peak amplitude in channel Fz did not significantly change (two-tailed paired t-test, $t(9) = 1.25$, $p = 0.24$) and the peak latency did not change for the P300 in Cz (Wilcoxon signed-rank test,

$Z = 11.5$, $p = 0.15$) and for the N200 in Fz (Wilcoxon signed-rank test, $Z = 10.0$, $p = 0.17$).

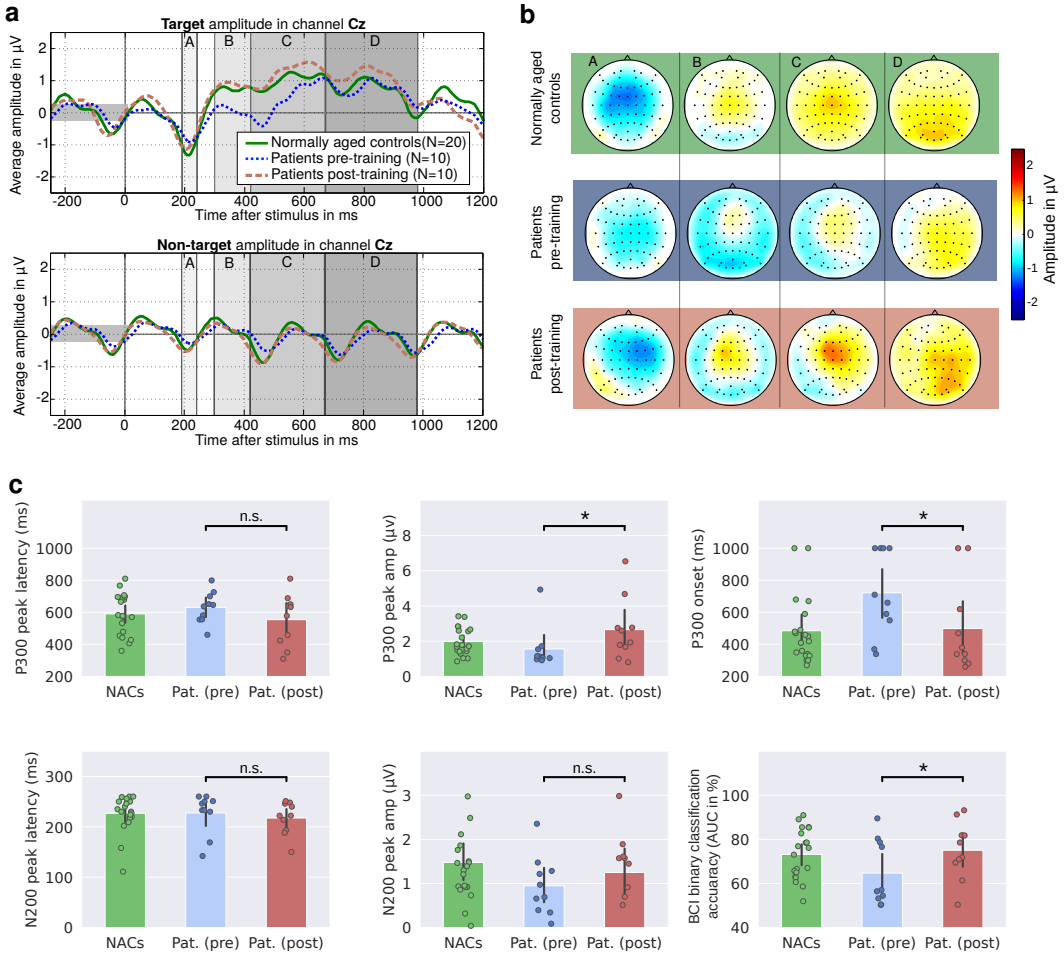


Figure 4.7: **Target ERP responses for patients (pre- and post-training) and for 20 normally-aged controls (NACs).** All analysis is based on ERP responses where the words were played with an SOA of 250 ms and from 6 loudspeakers. Subplot **a** shows the average target ERP responses for channel Cz and Fz. Subplot **b** visualizes the spatial distribution of mean target responses within four selected time intervals (in ms relative to stimulus onset): A: [191, 240], B: [301, 420], C: [421, 670], D: [671, 800]. Subplot **c** shows the average (bars with standard deviation) and individual values (dots) of the three groups for six different metrics. P300 peak onset was defined as the first time point where a significant difference between targets and non-targets could be observed. It was set to 1000 ms if no such difference could be observed. Peak amplitude and latencies were determined in a 10-time bootstrapping with 80% of the data each time. Classification accuracies were determined in 5-fold chronological crossvalidation. Abbreviations: **n.s.** = not significant, **AUC** = area under the receiver operator curve, * corresponds to p<0.05.

Compared to healthy controls, patients showed a more right-lateralized brain activity before and after the training (see Figure 4.7**b**) which is probably caused by their left-hemispheric brain stroke. The timing of the

P300 onset appears to become more similar to that of healthy controls after the training (see Figure 4.7**a**).

4.3.7 *Non-verbal oddball ERP responses*

In addition to the word ERPs, we also conducted an ordinary oddball task where the patients needed to pay attention to a rare high-pitched target tone while ignoring low-pitched non-target tones with a SOA of 1 second (see methods for more details). The data was analyzed similarly to the word-ERPs. We found that from the six quantities (N200/P300 amplitude and latency, P300 onset, BCI classification accuracy), only the BCI classification accuracy showed a significant change at the α-level of 0.05 where the performance after the training was lower than before the training. After correcting for multiple testing, none of the changes was significant anymore (see Figure 4.8 for more information).

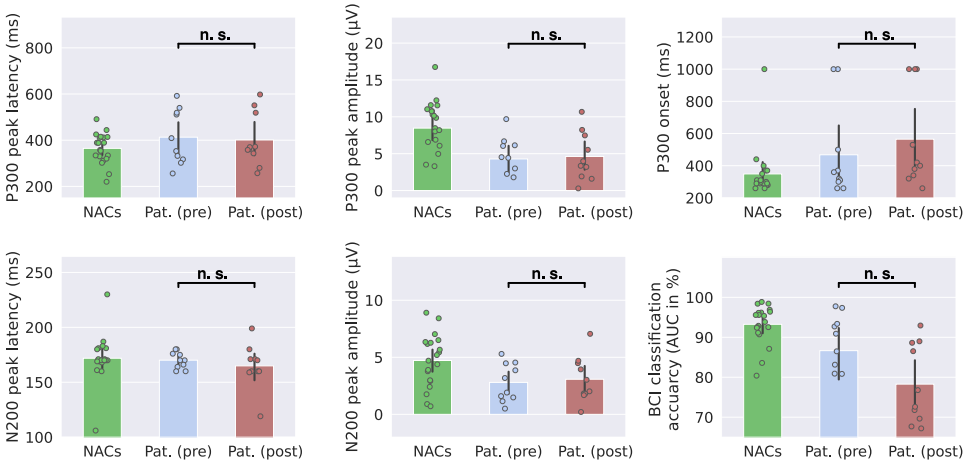


Figure 4.8: **Statistical analysis of the pre-post ERP differences in a non-verbal oddball task.** Please see the description of Figure 4.7**c** for more information.

4.4 DISCUSSION

This is the first study that shows that a BCI-mediated feedback, which informs the patient about the ongoing brain state, can successfully be used for language training in patients with chronic post-stroke aphasia. We did not only show the general feasibility, but our medium-to-strong effect sizes demonstrate that this training approach is highly competitive compared to existing speech and language therapies. Many aspects of the training are novel.

4.4.1 *Feasibility*

Regarding research question 1, it was an open question whether a sufficient target vs. non-target word ERP decoding quality in single trial could be achieved for aphasic patients. Compared to healthy subjects, patient data shows more missed stimuli, larger delays of word ERP responses and frequent movement and eye artifacts. This resulted in a lower SNR that clearly challenged the BCI decoding system but could be mitigated by a rigorous combination of algorithmic improvements, among them ICA-based artifact removal [178], session-to-session transfer learning to reduce the need for calibration [87, 97], and automatic regularization of the classifier to cope with high feature dimensionality and low SNR [21]. As shown in Figure 4.7**c**, some patients started around chance level, but patients showed a significant training-induced increase in BCI performance with an average performance of around 77% after the training. This value is comparable to healthy controls.

We think that the choice to use continuous adaptation of the decoding classifier [176] was important, too. While it generally allows compensating non-stationarity effects, it may have been pivotal for embracing novel task-solving strategies developed by patients over time. Patients may explore different strategies (e. g., repeating the target word in inner speech, visually imagining the target word, or focusing more on the spatial or phonological cues) which can lead to different brain signals depending on the strategy. Because this explorations occur in short time scales within one session, an adaptive classifier is better suited to model and reward strategies that lead to more discriminant ERPs.

4.4.2 *Training effect on language abilities*

Regarding research question 2, i. e. how the training affects language competences, we observed a significant and generalized training effect for all linguistic competences (including functional communication, spontaneous speech, object naming, among others) when comparing pre- to post-training assessments, see Table 4.2 and Table 4.4. A strong training effect is reflected by the observations that 5 out of 10 patients were classified as non-aphasic based on the AAT after the training (see Table 4.3) and that almost all pa-

tients improved in almost every subtest (see Figure 4.3**b**. The training even showed an improvement in patients (P9, P10) who had a severe aphasia which is generally much harder to obtain.

A well-standardized and widely used language test as primary endpoint allowed us to compare our results with preceding studies assessing the rehabilitation of aphasia. Compared to SLTs reported in literature, our average effect size of SMD = 0.63 in the primary endpoint (AAT) is high. The recent Cochrane review [27] analyzed the effect size for 27 randomized studies that compared one form of SLT versus no SLT. These studies were not limited to patients in the chronic phase, but also included subacute patients, whose improvement may have been dominated by spontaneous recovery. On average, the SMD was 0.28 for functional communication and 0.06 to 0.41 for other categories (written language, naming, comprehension) assessed by the AAT.

A more detailed comparison of our results to those of selected studies which also had focused on patients with varying severity of chronic aphasia, and that also had administered high-intensity training for around 30 hours, confirms that our training effect is relatively large. The recent high-profile study by the FCET2EC group [28] with 78 patients in the treatment group showed an effect size of SMD = 0.23 for verbal communication compared to a treatment-deferral group [1]. Other studies using the AAT as primary endpoint with a similar number of patients as in our study [119, 123, 144, 161] reported average AAT improvements of around 2-5 T-transformed AAT points which translates to effect sizes of SMD = 0.2-0.5.

The observed training effects were still significant at a 3 months follow-up for all AAT categories except for repetition. During the time period between the end of the training and the follow-up assessment, patients were not allowed to start another high-intensity training. For ethical reasons and to minimize the number of dropouts, we decided to conduct the follow-up at 3 months, and not after 6 months.

Interestingly, our main training effect could be observed already at the mid-training assessment which took place after around 15 hours of training. This indicates that a significant improvement of language competences can be reached quickly, and that a shorter training duration might also be an option, especially for mildly affected patients. It is, however, possible that a shorter training may lead to a weaker consolidation effect.

Importantly, the positive effect induced by our training was not restricted to the competences directly trained by the BCI task (e. g., comprehension), but the improvements generalized to all linguistic competences tested by AAT, the S&V naming test, as well as to functional communication (Figure 4.5). On an individual level, mildly affected patients improved

1 Please note that the authors reported the effect size of 0.57 in their publication, which had been computed based on the standard deviation of the *differences* and not, as recommended by the Cochrane Study Group, based on the standard deviation of the *population* [27]. It is known that the former leads to higher estimated effect sizes in a within-subject analysis [107]. To establish a fair comparison, we have thus recomputed the effect size based on the population standard deviation using the data published by Breitenstein and colleagues.

mainly in production aspects, while more severely affected patients mainly improved in the Token Test, see Figure 4.3. It is unclear, if prolonged training durations of more than 30 hours could induce stronger language production improvements in the latter group, too.

4.4.3 *Training-induced ERP changes and training efficiency*

Regarding research question 3, how ERPs are affected by the training, we observed a significant increase in P300 peak amplitude, an earlier onset of the P300 together with significantly improved BCI performances after the training. This underlines a relationship between the presence of the P300 and recovery from aphasia [132], a reduced P300 in children with language deficits [48] and increased P300 amplitude and latency reduction observed for easier training tasks [101]. Given that our patients performed the same task in pre- and post-evaluations, we conclude that the task has become easier for them.

According to a widely acknowledged hypothesis, the P300 reflects a context-update [142]. This theory says that the P300 is the result of a matching process where the linguistic representation of the target word in the working memory (context) is compared to the perceived word. If the perceived word does not match the context (e. g., a non-target word), then it does not provide new information regarding the representation of the target word and thus, does not elicit a strong P300. On the other hand, if the perceived word matches the target word, then the context is updated with the new information (e. g., how exactly the word sounded like) which is reflected by a P300. This explains why targets elicit a strong P300 while non-targets do only elicit a weak P300.

The question remains whether the training-induced increase in P300 amplitudes is explained by an improved ability to maintain the context or whether the updating process has become more efficient. Although patients had severe deficits in the verbal working memory before the training, post-training assessments showed no improvement regarding that ability. This provides evidence that indeed the updating process has become more efficient due to the training. It is also assumed that the P300 latency corresponds to the stimulus evaluation time [106, 142]. We argue therefore that our BCI training approach has induced a faster stimulus evaluation and has individually strengthened one or more aspects related to language integration. This improvement together with the observation that language comprehension and language production are heavily interwoven [140] might explain the observed improvements in various language competences.

It is difficult to pinpoint the exact reason for the high training efficiency. We think that it was important to incorporate the latest design recommendations for quick user learning in BCI applications [116]. Three major points should be noticed here. First, we gave informative feedback by not only showing the success or failure per trial, but by providing graded feedback with additional information, e. g., about runner-up words and

the confidence of the classifier. Second, we adjusted the task difficulty considering the patients' abilities to maintain a high training pressure while not making the task too difficult. This was realized by changing the speed of the word presentation (we chose relatively fast SOAs) and by changing the stimulus presentation from 6 loudspeakers to headphones if the task became too easy with the spatial information. Finally, we tried to maximize the training intensity because randomized controlled trials have found that high-intensity training is superior compared to a lower training intensity when controlling for the total training duration [27, 167]. This was realized by three different methods: (a) we maximized the number of trials executed per session by making use of a dynamic stopping procedure in each trial [153], (b) we did not limit the length of individual training sessions (but respected the patients' stamina, of course) and (c) we scheduled around 4 trainings per week to quickly accumulate to 30 hours of effective training. Unfortunately, it is not possible for us to disentangle which exact design decision was the most important without conducting further studies.

4.4.4 *Language-specificity of the training*

The final research question 4 was in how far the training improves language functions or whether it rather improves general attention. Alternatively to the context-updating theory, it has been proposed that P300 amplitude specifically reflects the activation of an event-categorization network that is controlled by the joint operation of attention and working memory [101]. Before the training, our stroke patients showed — concomitant to aphasia — deficits in many other higher-level cognitive functions. These deficits are in accordance with results from the literature [37]. The lack of significant improvements in these tasks (see Figure 4.6) as well as the absence of significance changes during the non-linguistic oddball task (Figure 4.8) despite an increase in P300 word amplitude contradicts the hypothesis that changes in the P300 can directly be linked to changes in the working memory or attention. We conclude that (1) our approach does not primarily train general attention or working memory and (2) that the observed language improvements cannot be explained by improved attention or working memory. Overall, we come to the conclusion that the context-updating hypothesis, which involves closing the loop between bottom-up processing and top-down prediction, delivers a better explanation of our results.

4.4.5 *Limitations and future work*

The BCI system was designed such that it reacts to any information in the ERPs that is discriminative between target and non-target ERPs. As explained earlier, this approach was chosen to improve the SNR and might be beneficial to strengthen a broad range of language functions. It is based on the assumption that ERPs, which are induced by a language task, are

language-related. However, there is a limitation to the approach, namely, that the classifier would also react to other components that are not related to language-processing. The two most prominent unwanted instances are eye artifacts and motor evoked potentials due to motor execution (or imagery) which are time-locked to target stimuli. If patients, for instance, blinked whenever they heard the target stimuli, then the resulting brain signals would also be class-discriminative and would have been rewarded by the classifier. In this case, we would essentially teach patients to blink at the right moment.

To prevent this, we carefully examined the patient and their brain signals during the training, employed artifact-removal methods, and analyzed the origin of the class-discriminative information in post-session analyzes including an evaluation in the frequency spectrum. For this analysis, it is from great value that the relevant features can easily be visualized for a linear classifier. However, for patient P2, we were not successful. The patient started to perform motor imagery upon perceiving the target stimuli which manifested itself in event-related (de)synchronization effects over the motor cortices. To avoid that an irrelevant component (or in the worst case, a harmful component) is reinforced, we decided to stop the training after 24 hours for that patient. Future work should address this problem by implementing advanced monitoring software that can detect unwanted components early on and by further refining the signal processing pipeline with the goal to restrict the reinforced signals to a pool of components that are known to be language-related.

The second limitation of the study is that we did not have an explicit control group. In our study, patients served as their own control group to some extent. For all but one patient, we could monitor the progress before starting our training when they underwent ordinary speech and language therapy $2-3$ times per week. Compared to that, our training shows a strong effect size. However, without a control condition, we cannot make a conclusive statement whether the brain state dependent feedback is indeed the key to the observed training success. Hence, a randomized controlled trial should be conducted in the future where the control group trains in a setup that is as similar to the BCI-based training as possible, but which does not rely on the analysis of the brain signals.

A candidate for such a control group would be a button-press based feedback. In this scenario, patients are instructed to press a button upon hearing the target word. They will then receive feedback based on their button press accuracies and timings. All other factors should be made comparable (e. g., patients should also wear an EEG cap). With that, it is possible to control for confounding factors of the therapy success, e. g., the stimulus repetition, interactions with the examiners (social support and social stimulation can support the recovery of language to a similar level as conventional SLT [25]) and repetitions of the language tests, to ultimately come to a conclusion in how far the BCI-based feedback is key to the training success.

AUTHOR'S CONTRIBUTION

This project was not possible without the close collaboration with the University Medical Center Freiburg (UMC Freiburg). Together with Mariacristina Musso (UMC Freiburg) and Michael Tangermann, we designed the training task and study protocol. I took the leading role in the technical implementation of the BCI-based online training, data analysis and visualization, and statistical testing. I also conducted or supervised the majority of the training sessions.

SUMMARY

In this chapter, I presented results from the first successful BCI-based language training for patients with post-stroke aphasia. I showed how the application of state-of-the-art machine learning methods allows patients to perform a challenging auditory ERP task. The clinical language assessments of 10 chronic stroke patients provided compelling evidence that the BCI-based training leads to strong and generalized verbal improvements. In addition, evoked brain responses showed larger amplitudes and lower latencies after the training, which suggests an improved language processing. Cognitive tests showed no statistical differences for verbal memory and attention-related tests which indicates that the training is specifically improving language functions. While future work is necessary to get a more detailed understanding of the exact training mechanism, I am convinced that this is the first step towards a new clinically-relevant application field for BCIs with many potential users.

5

SUMMARY AND OUTLOOK

This thesis has brought two major improvements to the BCI community. First, I have presented new unsupervised machine learning methods for ERP-based BCIs that are able to quickly and reliably learn a brain signal decoder from completely unlabeled data in Chapter 3. For the first time, it was possible to derive a completely unsupervised classifier that has the favorable theoretical property of *guaranteed convergence* under the assumption that the data points are independent and identically distributed. In addition, the learning efficiency of a combined unsupervised classifier was very high when compared to other unsupervised classifiers. Indeed, in some scenarios, the best new method could utilize unlabeled data almost as efficiently as a supervised algorithm with complete label access. This is a big step towards self-calibrating BCIs where the decoder is able to constantly extract meaningful information from the brain signals without the need for frequent time-consuming (re-)calibrations.

The second contribution was to introduce the first BCI-supported language training for patients with language deficits as explained in Chapter 4. The principal idea of this new approach is that patients receive meaningful feedback about their performance in a language task based on their ongoing brain signals, with the goal to reinforce strategies or brain states that are beneficial for the patients' language abilities. Together with clinical partners, we could not only show the feasibility of this approach, but we also observed that the training led to significant and long-lasting beneficial training effects. In contrast to language therapies with comparable patients and intensity, the training-induced language effects were strong and generalized to all verbal abilities. This makes the new BCI-based language training a promising alternative to existing language therapies and opens the door for a completely new application field of BCIs with an enormous potential user group.

In the following, I want to discuss how both contributions can be combined.

It should first be noted that the current BCI-based language training is completely supervised. This is the case because we instruct the patients to attend a certain (predefined) target word in every trial. This target word is known by the classifier and hence, we can use this information to adapt our classifier. In general, if label information is available *and* reliable, then unsupervised methods are always inferior compared to supervised methods. On the other hand, if label information is unreliable, then unsupervised methods can indeed provide a benefit. The scenario of unreliable labels is actually observed in the BCI-based training because some patients have

problems remembering the target word and start paying attention to a different word during a trial. While a supervised classifier learns misleading associations between brain signals and the task in this case, unsupervised methods can cluster the data without requiring the true labels. This allows unsupervised methods to cope with certain types of label noise as long as the user still performs any of the proposed tasks (e. g., listens to one specific word). In future work, one should investigate how much label noise can still be tolerated by a supervised classifier and when exactly it is beneficial to switch to an unsupervised classifier.

The introduction of robust unsupervised learning methods also provides an opportunity to move away from the completely supervised language training to one, where patients can train with a more flexible training task. This should make the training more motivating and engaging for the patients, and with that, ultimately more efficient and enjoyable. I envision a training task where the users freely select their desired targets (e. g., certain words). When successful, this triggers an event and users get access to new options. Various events are imaginable. For instance, a successful target word selection could trigger a new segment of a story that can again be continued based on a new set of word stimuli. With that, users could freely explore their desired story based on personal preferences. For other patients, it might be interesting to explore these words that they do not know well and then receive additional information about that word upon successful selection. Another idea is that patients receive a reward based on their choice (e. g., some words give more points than others based on difficulty) and with that, patients can select the difficulty-level more freely based on their current ability, ambition and confidence level.

While it is very difficult to foresee the future development of BCIs, I think that the quests for new application fields and for increased usability are two of the main challenges in the field. With my theoretical and practical contributions, I could make a significant step towards realizing robust and efficient algorithms that can be used to steer neurotechnological applications with real benefits for the user.

APPENDIX

a

AN AUDITORY BRAIN-COMPUTER INTERFACE WITH EYES-CLOSED.

The following text and Figures are mostly taken from the journal publication in Frontiers in Human Neuroscience [81].

ABSTRACT

Recent research and the previous section have demonstrated how brain-computer interfaces (BCI) based on auditory stimuli can be used for communication and rehabilitation. In these applications, users are commonly instructed to avoid eye movements while keeping their eyes open. This secondary task can lead to exhaustion and subjects may not succeed in suppressing eye movements. In this work, the option to use a BCI with eyes-closed was investigated. Twelve healthy subjects participated in a single electroencephalography (EEG) session where they were listening to a rapid stream of bisyllabic words while alternatively having their eyes open or closed. In addition, different usability aspects for the two conditions were assessed with a questionnaire. The analysis shows that eyes-closed does not reduce the number of eye artifacts and that event-related potential responses and classification accuracies are comparable between both conditions. Importantly, we found that subjects expressed a significant general preference towards the eyes-closed condition and were also less tensed in that condition. Furthermore, switching between eyes-closed and eyes-open and vice versa is possible without a severe drop in classification accuracy. Also, a patient which could not control the BCI with open eyes could control the BCI when closing his eyes. These findings suggest that eyes-closed should be considered as a viable alternative in auditory BCIs that might be especially useful for subjects with limited control over their eye movements.

A.1 INTRODUCTION

A problem that is typically encountered when recording brain activity by means of EEG is the occurrence of artifacts. Although most subjects have fewer problems to suppress body movements, *eye artifacts* such as blinks or eye movements are hard to eliminate during the measurement and their associated EEG signals are much stronger than the brain signals of interest. This is especially challenging for subjects wearing contact lenses (leading to dry eyes) and for theoften elderlypatients. Blinking rates were shown to be influenced by the workload [171] which is often quite high in BCI experiments.

In our BCI-based language training that was described in Chapter 4, we experienced one case where a chronic stroke patient was unable to voluntarily reduce the number of eye blinks. The subject was persistently blinking about once every second. This led to an extremely deteriorated quality of the EEG recordings. Different methods have been developed to alleviate the effects of eye artifacts using linear regression methods [134] or independent component analysis (ICA) decompositions [51, 178, 179], but they still lead to a significant data loss and cannot perfectly separate eye artifacts from underlying brain activity.

Additionally, the unnatural instruction to avoid eye blinks for a prolonged period constitutes for an unwanted secondary task that is distracting the subject from the main task and typically involves a substantial level of stress. This can have the undesired consequence that a training based on EEG signals is less efficient due to the split of cognitive resources to the main training task and to the secondary task of avoiding eye blinks. In an extreme scenario, subjects may spend so much attention on suppressing eye artifacts, that they are unable to perform the main task.

The difficulty of avoiding eye movements over a long period leads to the question if the number of eye artifacts could be reduced and the measurement can be made more comfortable for the test subject by having the subject *close their eyes while collecting the data*. This idea is feasible in auditory BCIs since visual input is not needed during a trial. In this study, we will compare two conditions: eyes-closed (**EC**) and eyes-open (**EO**). While many studies have shown that EC leads to an increase in occipital alpha as well as a changed topology and activity in different frequency bands compared to EO (see [7]), the existing literatureto our knowledgelacks an analysis of the EC condition for event-related potentials (ERPs) in the fast paradigms that are used for BCIs. For the slower conditions, a lot of data exists. A recent meta-review found that latency and amplitude of the P300 were not significantly different between EO and EC in the standard oddball task with an SOA of1 second and tones as stimuli [43]. Remarkably, several hundred subjects were included in this meta-analysis for each condition ($N_{EO} = 555$, $N_{EC} = 998$) where the data was collected from several studies (16 studies used EO and 23 studies used EC).

However, results from this meta-analysis are not directly transferable to BCIs as (a) the SOA between two stimuli in the meta-review (1 sec-

ond) is much longer than in recent BCIs (typically SOAs vary between $250 - 550$ ms) and because (b) tone stimuli lead to different ERP responses compared to natural (more complex) sounds (animals sounds or words) that are used in modern BCIs [9, 65, 71, 157, 163, 166]. In addition, questions regarding (c) the number of eye artifacts and (d) user comfort and usability were not investigated.

Another relevant research question is whether a system trained on data recorded with EO could be applied when the subject has their eyes closed and the other way around. If this is the case, subjects could switch between conditions within one session. This could be expected to improve the overall comfort of the subject during the measurement and decrease the stress level.

In summary, this study should investigate four main hypotheses.

H1: EC leads to fewer eye artifacts than EO.

H2: The achieved target vs non-target classification accuracies do not differ significantly between EO and EC.

H3: The measuring process is overall more comfortable for the subjects for EC than for EO.

H4: A system trained on data recorded in one condition can be applied in the other condition without a substantial loss in classification accuracy.

A.2 MATERIAL AND METHODS

In a within-subject design, we compared the EEG signals and usability aspects for the conditions EC and EO in an auditory BCI paradigm using words as stimuli and a fast SOA of 250 ms. The raw EEG data sets for this study can be found in the Zenodo Database[1].

A.2.1 *Participants*

Twelve healthy volunteers (11 subjects between 22-29 (mean = 25.2 years, SD = 2.04 years), and one subject (S7) aged 76, 5 female in total) were recruited for the experiment. All twelve subjects reported having normal hearing. Following the Declaration of Helsinki, approval for this study was obtained by the ethics committee of the University Medical Center Freiburg and all participants gave written informed consent prior to participation. A session took about 3.5 hours (including the EEG set-up and washing the hair).

A.2.2 *Experimental structure and stimuli*

Subjects were asked to be seated comfortably on a chair, facing a computer monitor. Six loudspeakers were centered in 60-degree steps, at ear height around the subjects head, with a radius of approximately 60 cm (see Figure a.1**A**). The auditory stimuli were presented from the six loudspeakers according to the AMUSE (Auditory MUlticlass Spatial ERP) paradigm [152].

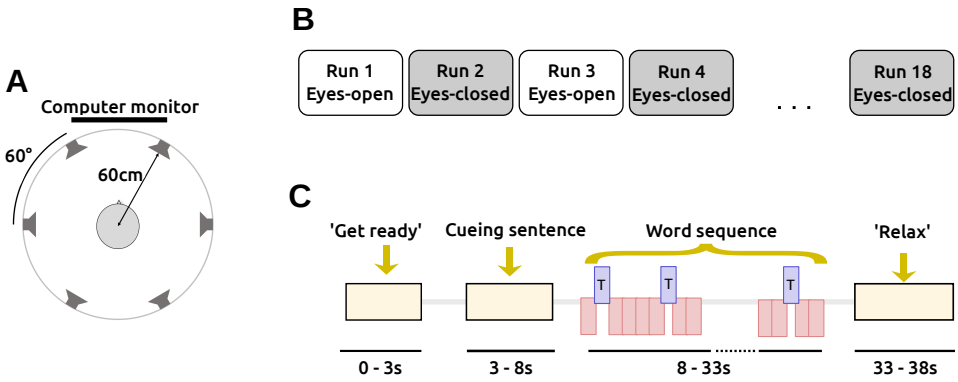


Figure a.1: **Structure and design of the eyes-open/closed study. A:** AMUSE setup in a top view. Six loudspeakers are spatially centered around the subjects head. Figure adapted from [152]. **B:** a session consisted of 18 runs alternating between eyes-open and eyes-closed. Each run consists of 6 trials. **C:** a trial comprises 4 distinct stages. The timings (in seconds) indicate the beginnings of each stage. During the word sequence, targets (T; blue) and non-targets (red) are interleaved and played with a fast SOA of 250 ms. Figure taken from [81].

1 DOI: http://doi.org/10.5281/zenodo.1298606

A session consisted of a total number of 18 runs, each contained 6 trials. Runs where subjects had their eyes closed were followed by runs where subjects had their eyes open and vice versa (see Figure a.1**B**) to alleviate effects of non-stationaritiy in the EEG signals. The current condition was indicated to the user on the screen. To prevent systematic errors, the condition used in the first run alternated between participants.

In each trial, one out of six bisyllabic words (length = 300 ms) were cued by a sentence as target stimuli before presenting a sequence of word stimuli (SOA = 250 ms), see also Figure a.1**C**. In a familiarization phase before the EEG recording, these sentence-word mappings were practiced with the subjects. During the sequence, each speaker played a different distinct word 15 times, resulting in a class-wise ratio of 1 : 5, with 15 target and 75 non-target stimuli per trial. Per condition (EO/EC), 9 runs were recorded. As each of them contains 6 trials, our experiment resulted in 54 trials per condition. Multiplying these 54 trials per condition with the number of targets per trials (15) and the number of non-targets per trial (75) results in a total of 810 targets and 4050 non-targets per subject and condition (EO / EC), respectively. In a run, each of the six stimuli was chosen exactly once as a target, while the other stimuli served as non-targets. We pseudo-randomized the ordering in which the stimuli were presented and in which the targets were selected. The mapping from stimulus to loudspeaker was also performed pseudo-randomized.

A.2.3 *Data acquisition and processing*

The study consisted of the EEG recordings during the AMUSE paradigm and the subjective ratings mainly after the EEG measurements.

For both conditions (EO / EC), we assessed several subjective ratings after the session in a questionnaire. Subjects were asked to rate their ergonomic experience during the EEG recordings for eight items regarding motivation, concentration, fatigue, eye movement suppression, eye blink suppression, stimulus discrimination, exhaustion, and difficulty of the task on a 5-point Likert scale. We also asked the subject which condition they preferred overall (EO / EC / undecided). We further used the self-assessment manikin (SAM) [26], which is a non-verbal pictorial assessment technique, to assess valence from 1 (negative) to 9 (positive), and arousal from 1 (calm) to 9 (excited). In addition, we asked the subjects to indicate their general fatigue before and after the EEG measurement on a 5-point Likert scale.

EEG activity was recorded and amplified by a multichannel EEG amplifier (BrainAmp DC, Brain Products) and with 63 passive Ag/AgCl electrodes (EasyCap). The channels were placed according to the 10-20-system, referenced against the nose and grounded at channel AFz. Electrode impedances were kept below 15 kΩ. Eye signals were recorded by electrooculography (EOG) with an electrode below the right eye of a subject (the channel associated with this electrode is hereafter called EOGvu). The signal was sampled at a rate of 1 kHz.

In addition to the EEG and EOG channels, pulse (on an index finger) and respiration (diaphragmatic breathing) were recorded, but not further analyzed.

EEG data preprocessing

The offline analysis of the EEG data was performed using the BBCI toolbox [20]. The data was bandpass filtered in $[0.5 8]$ Hz using a Chebyshev Type II filter and downsampled to 100 Hz. EEG signals were then epoched between -200 ms and 1200 ms relative to the stimulus onset. A baseline correction was then performed based on data within the interval $[-200, 50]$ ms.

We marked those epochs where the difference of the highest and lowest value in one epoch exceeded 60 μV in one of the frontal channels (Fp1, Fp2, F7, F8, F9, F10) to capture eye- or other muscular artifacts. We call this step *Minmax*_60. The percentage of epochs that gets flagged by this procedure (and by additional steps that will be described below) is reported in the result section. In total, we applied three different preprocessing pipelines in addition to the steps mentioned before:

PP1: Only the above steps were applied (*Minmax*_60).

PP2: Before applying *Minmax*_60, we regressed out eye artifacts and applied the variance criterion. Please see, Section 2.4.2 for more information.

PP3: Multiple Artifact Rejection Algorithm, short MARA [178, 179], an ICA-based supervised machine learning algorithm to reject eye components, was applied before *Minmax*_60. Please also see Section 2.4.2 for more information.

A.2.4 *Classification*

Per EEG channel, the amplitudes were averaged in eight intervals: $[100, 190]$, $[191, 300]$, $[301, 450]$, $[451, 560]$, $[561, 700]$, $[701, 850]$, $[851, 1000]$ and $[1001, 1200]$ ms. These intervals have shown good classification results in Chapter 4. They had been handcrafted to capture the time intervals with the highest discriminatory power for typical subjects. We fixed them in the study design before recording the data to avoid a potential overfitting to the obtained classification accuracies. The selected interval boundaries were the same for the two conditions (EO / EC) to guarantee a fair comparison. For visualization (Figure a.5), we manually picked those intervals that show the most discriminatory time intervals after computing the grand averages.

This led to a 504-dimensional feature vector (= 63 channels $\cdot$ 8 intervals) per epoch. The classification between target and non-target stimuli was performed using the LDA classifier with shrinkage-regularized covariance matrix [21], see Section 2.4.4 and Section 2.5.2, respectively. If not specified further, we applied a five-fold chronological cross-validation for estimating the classification accuracies.

A.3 RESULTS

Hypothesis 1 (Eye artifacts). In order to test whether the EC condition leads to fewer artifacts than the EO condition, we applied three different preprocessing pipelines (PP1-PP3) to the data as explained in the method section. The results are shown in Figure a.2**A**. By visual inspection, one can observe that the number of artifacts is higher for the EC condition. A Wilcoxon signed rank test over the percentage of artifact trials for each participant for EO and EC shows that the number of artifacts is significantly higher for EC when only *Minmax*_60 is applied (**PP1**: $W = 3$, $p = 0.0024$), but not for the other two preprocessing condition (**PP2**: $W = 29, p = 0.5$; **PP3**: $W = 9$, $p = 0.037$) when applying the Bonferroni-Holm correction (uncorrected p-values are reported). Hence, the hypothesis that there are less artifact trials in the EC condition could not be confirmed. Given the very consistent results, it is unlikely that more subjects will deliver different results. Instead, the data suggests the opposite, namely, that more eye artifacts exist with EC compared to EO.

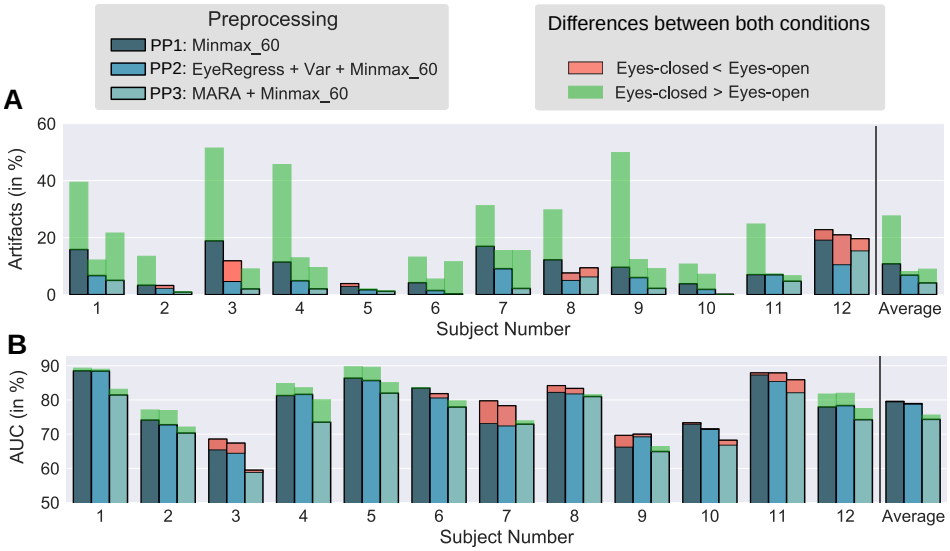


Figure a.2: **Number of artifacts and classification accuracies for different preprocessing methods. A:** the relative number of artifacts obtained by *Minmax*_60 (and the variance criterion in case of PP2) for all subjects. **B:** cross-validated classification accuracy for all subjects. The solid blue-ish bars depict the smaller value for the two conditions (EO/EC). The red or green bars indicate the value of that condition which led to a higher outcome. Figure taken from [81].

Hypothesis 2 (Accuracy). We examined whether the accuracies differ between EO and EC. Depending on the preprocessing and condition, the grand average performance was around $75 - 80\%$ (see Figure a.2**B**). The Wilcoxon signed-rank test was used to test the null hypothesis that the accuracies are the same for both conditions. We found that for all three preprocessing pipelines, there was no significant different between the two conditions (**PP1**: $W = 38, p = 0.9$, **PP2**: $W = 40, p = 0.9$, **PP3**: $W = 17, p =$

0.1). It may be the case that a clear trend evolves in the case of measuring a larger number of subjects. However, the small difference between the two groups (the absolute difference between the average performances is less than 1.5 % classification accuracy for all three preprocessing pipelines in our data) and the non-significant result from the meta-review concerning the oddball ERP responses for several hundreds of subjects, convinces us that the effect of the condition on classification accuracy is rather limited.

Hypothesis 3 (Usability). In order to determine whether the measuring process is more comfortable for subjects in the EC condition than in the EO condition, we statistically evaluated a subset of five questions that the participants have answered in the questionnaire.

1. How much did you struggle with fatigue in the different conditions?
2. How easy was it to avoid eye movements in the different conditions?

3-4. How was your mood during the different conditions in terms of valence (negative vs. positive) and arousal (calm vs. tensed)?

5. Overall, which condition did you prefer?

We limited the statistical evaluation to these five questions to reduce the number of multiple comparisons, but report the results for all categories of the questionnaire (see Figure a.3). For the five statistical tests, we corrected the resulting p-values with the Bonferroni-Holm correction. A paired t-test was applied as it was shown to have the same statistical power as a signed Wilcoxon signed-rank test in case of a 5-point Likert scale, see [41].

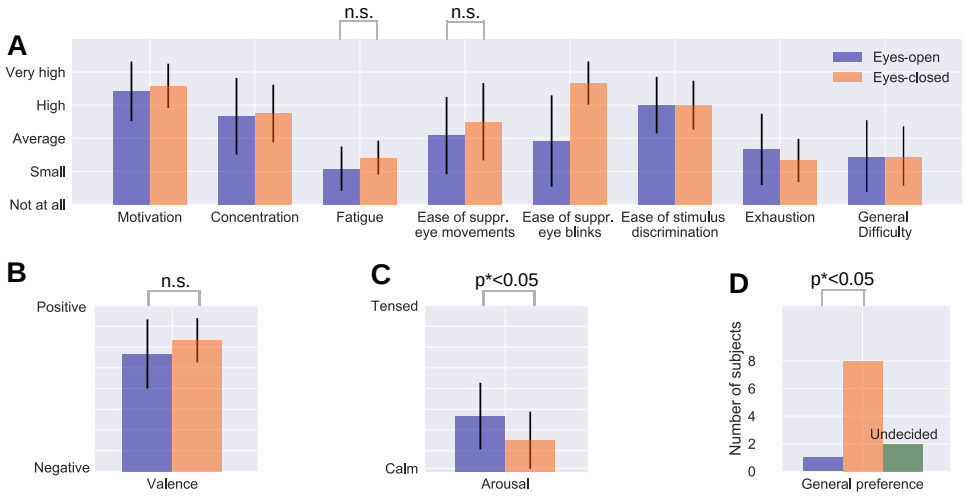


Figure a.3: **Questionnaire results regarding usability.** The mean values and standard deviation of the 12 subjects are shown for each category (**A-D**). p^* indicates Bonferroni-Holm corrected p-values, n.s. means 'not significant'. Figure taken from [81].

We found no significant differences for fatigue ($t(11) = 1.77, p = 0.10$) and the ease of suppressing eye movements ($t(11) = 1.16, p = 0.27$), see Figure a.3**A**. For valence, results suggest that EC was perceived as more

positive ($t(11) = 2.35, p = 0.039$ (uncorrected)), but this was not significant after Bonferroni-Holm correction (Figure a.3**B**). Significant effects were found for arousal ($t(11) = -3.92, p = 0.002$ (uncorrected)) showing that participants were calmer with EC and also for the general preference (Figure a.3**C**). Nine out of twelve subjects preferred EC, only one subject preferred EO and two subjects were undecided. A one-sided binomial test yields $p = 0.006$ (uncorrected), hence we can reject the null hypothesis that both conditions have the same comfort level (Figure a.3**D**).

Hypothesis 4 (Transferability). We investigated whether a system trained on data recorded with EO could be applied in runs with EC and vice versa. Therefore, we ran a post-hoc offline simulation consisting of two parts. The first part describes the influence of the training set size only, while no transfer learning between conditions was applied. For each subject, we utilized data of the first 18 trials of a condition (EO / EC) to draw an increasing number of randomly chosen trials. Then each of these sets was used to train a shrinkage-regularized LDA classifier. The performance of each classifier was then tested on another randomly selected (but unseen) trial from the same condition and subject. This procedure was repeated many times and with different seeds for the random selection of training and testing data. The average over these repetitions delivered a reliable performance estimate for growing sizes of training data sets. The grand average results are shown in Figure a.4**A** (left to the red dashed line). Both conditions performed very similarly during this part.

In the second part we investigated the effects of transfer learning, i.e. the switching between conditions after 18 trials and continued application on the remaining 36 trials (remember that we had 54 trials per condition in total). Four different transitions were simulated offline: two transitions with a change of conditions (EO $\rightarrow$ EC, EC $\rightarrow$ EO) andto allow for comparisonsanother two without a change of conditions (EO $\rightarrow$ EO, EC $\rightarrow$ EC). In each of the four scenarios, we took the LDA classifier that was trained on the first 18 EO or EC trials (depending on the condition before the transition). Afterward, we tested the classifier on a randomly drawn trial of the condition after the transition. This trial was then added to the training data and the LDA classifier was retrained on the slightly enlarged training data. As a result of changed conditions, the target vs. non-target accuracy initially dropped around $3 - 4\,\%$ (from ~74 % to ~71 %), while no drop was observed when conditions were maintained (see Figure a.4**A**, right to the red dashed line). Collecting and including more data from the condition after the transition, the performance differences between change and no change rapidly decreased until they were not distinguishable anymore after 30 new trials (see Figure a.4**B**). In both phases, we applied the aforementioned randomization procedure with 20 different seeds to obtain reliable results.

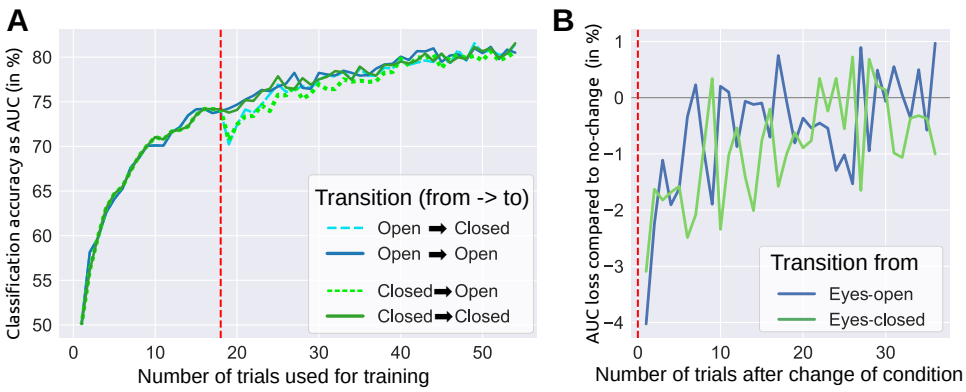


Figure a.4: **Influence of changing from eyes-closed to eyes-open and vice versa.** **A:** a switch of conditions was simulated after 18 trials (dashed red line) yielding a small reduction in target vs. non-target classification accuracy (measured by AUC). All classifiers were continuously retrained after each trial (see text). **B:** this subplot shows the loss in accuracy when changing from one condition to the other. Figure taken from [81].

A.3.1 *ERP analysis*

In addition to the four main hypotheses, we also investigated the shapes, amplitudes and latencies of the ERP responses for both conditions. Figure a.5 shows the grand average ERP responses after processing the data with pipeline PP2 (although noisy channels were not removed when computing the grand average). The most relevant features (in a linear discriminatory sense) can be inferred from the signed R^2 plots in the bottom row of Figure a.5**A** and Figure a.5**B**. Two main components are visible for EO and EC: An early negativity with a peak location around FCz and a peak latency of around 200 ms ('N200') and a later positivity ('P300') in the parietal area. To quantitatively describe these components, we computed the peak amplitudes and latencies for each subject. The results are presented in Table a.1.

The most striking difference between EO and EC is that the late parietal positivity (P300) appears to be earlier in the EC condition compared to the EO condition, see Figure a.5**C**. A two-sided paired t-test for the four quantities (N200 amplitude and latency and P300 amplitude and latency) showed no significant differences between the experimental conditions after Bonferroni-Holm correction, although the P300 latency differs strongly (uncorrected T-test, $t(11) = 2.96, p = 0.013$).

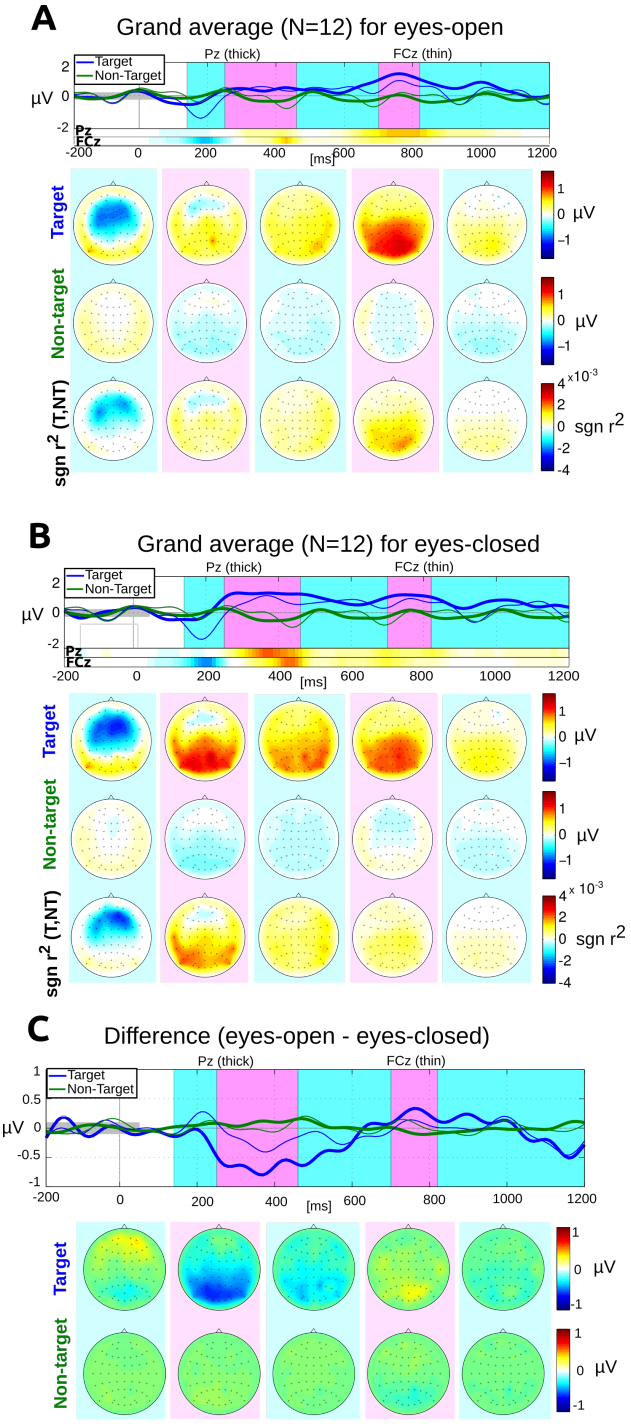


Figure a.5: **Grand average ERP responses for eyes-open, eyes-closed and their differences.** Top rows: average responses evoked by target (blue) and non-target (green) stimuli in the central channel Cz (thick) and the parietal-occiptal channel POz (thin). The signed R^2 values for these two channels are provided by two horizontal color bars. Their scale is identical to the scale of the plots in the bottom row of scalp plots. Target / Non-target rows: scalp plots visualizing the spatial distribution of mean target and non-target responses within five selected time intervals: $[140, 250]$, $[251, 460]$, $[461, 700]$, $[701, 820]$ and $[821, 1200]$ ms relative to stimulus onset. Bottom row: scalp plots with signed R^2 values indicate spatial areas with high class-discriminative information. Figure taken from [81].

Table a.1: **Overview of peak latencies (in ms) and amplitudes (in μV) for the 12 subjects.** The N200 peaks were computed using channel FCz in the interval $[140, 280]$ ms. P300 peaks were calculated from channel Pz in the interval $[300, 900]$ ms. A bootstrapping approach was used to improve the reliability of the peak estimates in which peaks were estimated 10 times on subsets containing randomly-drawn 80% of the data and then averaged across subsets.

	Eyes-open				**Eyes-closed**			
	N200 (FCz)		P300 (Pz)		N200 (FCz)		P300 (Pz)	
	Lat.	Ampl.	Lat.	Ampl.	Lat.	Ampl.	Lat.	Ampl.
S1	197	-2.42	792	2.41	206	-1.70	310	2.32
S2	230	-0.65	808	1.53	203	-0.87	428	3.73
S3	150	-0.23	740	0.80	150	-0.50	805	0.81
S4	195	-0.82	796	1.99	197	-0.41	346	2.93
S5	237	-1.85	390	1.81	244	-1.91	437	1.96
S6	169	-0.41	704	1.19	179	-1.04	561	0.89
S7	239	-2.36	615	2.09	250	-2.08	543	2.01
S8	220	-1.27	840	1.86	212	-1.60	512	1.64
S9	195	-2.02	755	1.94	221	-1.46	687	1.91
S10	213	-1.28	359	1.84	223	-1.37	339	1.25
S11	190	-3.53	742	1.59	190	-2.92	570	1.81
S12	179	-1.48	743	1.47	195	-3.23	742	2.11
Mean	201	-1.53	690	1.71	206	-1.59	523	1.95
SD	27.6	0.96	158	0.43	27.4	0.87	161	0.82

A.4 DISCUSSION AND CONCLUSION

The goal of this study was to compare the EC and EO condition in a fast auditory BCI paradigm. In brief, our results show that EC leads to comparable signals (with slightly more eye artifacts) while clearly being preferred by the users. Although we have investigated a limited number of subjects only, we observed significant effects which indicate a strong influence of the condition on usability. In the introduction, we mentioned a stroke patient that could not avoid very frequent eye blinks. We instructed this patient to proceed with EC. Afterward, he could successfully control an auditory BCI although he reported to sometimes 'drift away', i.e. to lose focus.

These important findings can have a direct impact on the usability of auditory BCIs. It suggests that subjects should either start with EC right from the beginning or, even better, subjects should simply have the choice to use their preferred condition (EC/EO). This strategy could mitigate major difficulties that are faced when working with subjects that have limited control over their eye movements. In addition, we could show that a transition from one condition to another leads only to a small loss in classification accuracy that quickly diminishes when the classifier is retrained on new data. Especially during longer sessions, we think that this small sacrifice of classification accuracy justifies the improved user comfort.

To understand why condition EC led to an increased number of eye artifacts, we have conducted an additional analysis where we computed the number of artifacts for the two bipolar channels EOGh and EOGv (see preprocessing pipeline PP2). These channels should mainly capture horizontal and vertical eye movements, respectively. The analysis shows that eye artifacts in the EC condition originate from vertical as well as from horizontal eye movements with a similar proportion. We believe that the increased number of eye artifacts in the EC condition comes from the absence of a fixation cross. With that, it is rather difficult to not move the eyes and subjects involuntarily produce small saccades. Interestingly, this point has not been reported by the subjects in the questionnaire. Although not significant, they reported that they perceived it as easier to suppress eye movements in the EC condition.

Although not significant, we observed that the P300 peak latency is much larger for the EO condition compared to the EC condition. To explain this observation, we hypothesize that the EO condition has a higher task demand due to the need to simultaneously process visual and auditory input whereas no visual input needs to be processed in the EC condition. This may lead to higher overall workload in the EO condition and thus, explain increased P300 latencies.

We designed the protocol is such a way that EC and EO runs are alternating. The idea behind this design was to reduce the effect of any non-stationarities that occur over the course of a longer session due to human factors (user learning, changed user strategies, fatigue), medication or external factors (drying gel, changed cap position) changing the ERP

responses [156]. On the one side, we believe that this design actually led to an underestimation of the severity of eye movements in the EO condition due to the frequent runs where subjects had their eyes closed. One subject remarked that 'it would have been difficult to leave the eyes open without the runs where I had my eyes closed'. On the other side, fatigue might become a more severe problem when longer sessions with EC are conducted. We think that the EC strategy should be further tested in real application scenarios to identify possible shortcomings.

A possible limitation with our questionnaire regarding the subjective ratings is that the answers for each item were ordered from unfavorable to favorable, for the question 'How motivating were the different conditions for you?' the possible answers were sorted from 'not at all motivating' to 'very motivating'. This same ordering for all questions might increase the effect of participants trying to answer consistently. Ordering the possible answers for each item randomly might help to avoid this issue.

Taken together, this is the first study that systematically compares the eyes-closed and eyes-open condition for an auditory BCI. We found that the eyes-closed condition should be considered as a viable alternative to increase the user comfort. In addition, we encourage other scientists and BCI practitioners to test the eyes-closed condition for subjects that fail to control a BCI due to frequent eye movements.

AUTHOR'S CONTRIBUTION

The comparison between open and closed eyes was joint work with Albrecht Schall, Natalie Prange and Michael Tangermann. My contribution was to design the study together with Michael Tangermann. In addition, I collected and analyzed the data together with Natalie Prange and Albrecht Schall and wrote the initial draft of the paper.

BIBLIOGRAPHY

[1] M. Ahn, M. Lee, J. Choi, and S. C. Jun. "A review of brain-computer interface games and an opinion survey from researchers, developers and users." In: *Sensors* 14.8 (Aug. 2014), pp. 14601–14633.

[2] A. B. Ajiboye, F. R. Willett, D. R. Young, et al. "Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration." In: *The Lancet* 389.10081 (2017), pp. 1821–1830.

[3] S. Aschenbrenner, O. Tucha, and K.-W. Lange. "Regensburg word fluency test." In: *Göttingen: Hogrefe* (2000).

[4] A. Baddeley. "Working memory." In: *Science* 255.5044 (1992), pp. 556–559.

[5] A. Barachant and M. Congedo. "A plug&play P300 BCI using information geometry." In: *arXiv preprint:1409.0107* (Aug. 2014).

[6] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. "Multi-class brain computer interface classification by Riemannian geometry." In: *IEEE Transactions on Biomedical Engineering* 59.4 (Apr. 2012), pp. 920–928.

[7] R. J. Barry, A. R. Clarke, S. J. Johnstone, C. A. Magee, and J. A. Rushby. "EEG differences between eyes-closed and eyes-open resting conditions." In: *Clinical Neurophysiology* 118.12 (2007), pp. 2765–2773.

[8] A. Basso, M. Forbes, and F. Boller. "Rehabilitation of aphasia." In: *Handb. Clin. Neurol.* 110 (2013), pp. 325–334.

[9] E. Baykara, C.-A. Ruf, C. Fioravanti, et al. "Effects of training and motivation on auditory P300 brain–computer interface performance." In: *Clinical Neurophysiology* 127.1 (2016), pp. 379–387.

[10] J. D. Bayliss. "Use of the evoked potential P3 component for control in a virtual apartment." In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 11.2 (June 2003), pp. 113–116.

[11] E. J. Benjamin, M. J. Blaha, S. E. Chiuve, et al. "Heart disease and stroke statistics — 2017 update: a report from the American Heart Association." In: *Circulation* 135.10 (Mar. 2017), e146–e603.

[12] M. Bensch, A. A. Karim, J. Mellinger, et al. "Nessi: An EEG-controlled web browser for severlely paralyzed patients." In: *Computational Intelligence and Neuroscience* 2007 (Sept. 2007), e71863. ISSN: 1687-5265. DOI: 10.1155/2007/71863.

[13] H. Berger. "Über das Elektrenkephalogramm des Menschen." In: *Journal für Psychologie und Neurologie* (1930).

[14] M. L. Berthier. "Poststroke aphasia: epidemiology, pathophysiology and treatment." In: *Drugs Aging* 22.2 (2005), pp. 163–182.

[15] A. Biasiucci, R. Leeb, I. Iturrate, et al. "Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke." In: *Nature communications* 9.1 (2018), p. 2421.

[16] G. Bin, X. Gao, Y. Wang, Y. Li, B. Hong, and S. Gao. "A high-speed BCI based on code modulation VEP." In: *Journal of Neural Engineering* 8.2 (Mar. 2011), p. 025015.

[17] C. M. Bishop. *Pattern Recognition and Machine Learning*. Vol. 128. New York: Springer, 2006, pp. 1–58.

[18] B. Blankertz, K.-R. Müller, D. J. Krusienski, et al. "The BCI competition III: validating alternative approaches to actual BCI problems." In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 14.2 (June 2006), pp. 153–159.

[19] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller. "Optimizing spatial filters for robust EEG single-trial analysis." In: *IEEE Signal Processing Magazine* 25.1 (Jan. 2008), pp. 41–56.

[20] B. Blankertz, M. Tangermann, C. Vidaurre, et al. "The Berlin brain-computer interface: Non-medical uses of BCI technology." In: *Frontiers in Neuroscience* 4 (Dec. 2010), p. 198. ISSN: 1662-453X. DOI: 10.3389/fnins.2010.00198.

[21] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller. "Single-trial analysis and classification of ERP components, a tutorial." In: *NeuroImage* 56.2 (May 2011), pp. 814 –825.

[22] B. Blankertz, L. Acqualagna, S. Dähne, et al. "The Berlin brain-computer interface: progress beyond communication and control." In: *Frontiers in Neuroscience* 10.530 (Nov. 2016), pp. 1–24.

[23] S. N. G. Bolagh, M. B. Shamsollahi, C. Jutten, and M. Congedo. "Unsupervised cross-subject BCI learning and classification using Riemannian geometry." In: *24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2016)*. Bruges, Belgium, 2016.

[24] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni. "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness." In: *Neuroscience Biobehavioral Review* 44 (2014), pp. 58–75.

[25] A. Bowen, A. Hesketh, E. Patchick, et al. *Clinical effectiveness, cost-effectiveness and service users' perceptions of early, well-resourced communication therapy following a stroke: a randomised controlled trial (the ACT NoW Study)*. NIHR Journals Library, 2012.

[26] M. M. Bradley and P. J. Lang. "Measuring emotion: The self-assessment manikin and the semantic differential." In: *Journal of Behavior Therapy and Experimental Psychiatry* 25.1 (1994), pp. 49–59.

[27] M. C. Brady, H. Kelly, J. Godwin, P. Enderby, and P. Campbell. "Speech and language therapy for aphasia following stroke." In: *Cochrane database of systematic reviews* 6 (2016).

[28] C. Breitenstein, T. Grewe, A. Flöel, et al. "Intensive speech and language therapy in patients with chronic aphasia after stroke: a randomised, open-label, blinded-endpoint, controlled trial in a health-care setting." In: *Lancet* 389.10078 (Apr. 2017), pp. 1528–1538.

[29] C. Brunner, N. Birbaumer, B. Blankertz, et al. "BNCI Horizon 2020: towards a roadmap for the BCI community." In: *Brain-computer interfaces* 2.1 (2015), pp. 1–10.

[30] L. Carelli, F. Solca, A. Faini, et al. "Brain-computer Interface for clinical purposes: cognitive assessment and rehabilitation." In: *BioMed research international* 2017 (2017).

[31] M. A. Cervera, S. R. Soekadar, J. Ushiba, et al. "Brain-Computer Interfaces for Post-Stroke Motor Rehabilitation: A Meta-Analysis." Nov. 2017.

[32] M. A. Cervera, S. R. Soekadar, J. Ushiba, et al. "Brain-computer interfaces for post-stroke motor rehabilitation: a meta-analysis." In: *Annals of clinical and translational neurology* 5.5 (2018), pp. 651–663.

[33] R. Chavarriaga, P. W. Ferrez, and J. d. R. Millán. "To err is human: Learning from error potentials in brain-computer interfaces." In: *Proceedings of the International Conference on Cognitive Neurodynamics (ICCN) 2007*. Shanghai, China: Springer, 2008, pp. 777–782.

[34] R. Chavarriaga, I. Iturrate, and J. d. R. Millán. "Robust, accurate spelling based on error-related potentials." In: *Proceedings of the 6th International Brain-Computer Interface Meeting*. Asilomar, USA: Verlag der Technischen Universität Graz, 2016, p. 15.

[35] R. Chavarriaga, A. Sobolewski, and J. d. R. Millán. "Errare machinale est: The use of error-related potentials in brain-machine interfaces." In: *Frontiers in Neuroscience* 8.208 (July 2014), pp. 1–13.

[36] X. Chen, Y. Wang, M. Nakanishi, X. Gao, T.-P. Jung, and S. Gao. "High-speed spelling with a noninvasive brain–computer interface." In: *Proceedings of the National Academy of Sciences* 112.44 (Sept. 2015), E6058–E6067.

[37] S. C. Christensen and H. H. Wright. "Verbal and non-verbal working memory in aphasia: what three n-back tasks reveal." In: *Aphasiology* 24.6-8 (July 2010), pp. 752–762.

[38] S. Dähne, J. Höhne, and M. Tangermann. "Adaptive classification improves control performance in ERP-based BCIs." In: *Proceedings of the 5th International Brain-Computer Interface Conference*. Graz, Austria, 2011, pp. 92–95.

[39] E. Daucé, T. Proix, and L. Ralaivola. "Reward-based online learning in non-stationary environments: adapting a P300-speller with a 'backspace' key." In: *Neural Networks (IJCNN), 2015 International Joint Conference on Neural Networks*. IEEE. Killarney, Ireland, 2015, pp. 1–8.

[40] M. De Vos, K. Gandras, and S. Debener. "Towards a truly mobile auditory brain–computer interface: Exploring the P300 to take away." In: *International Journal of Psychophysiology* 91.1 (2014), pp. 46–53.

[41] J. C. De Winter and D. Dodou. "Five-point Likert items: t test versus Mann-Whitney-Wilcoxon." In: *Practical Assessment, Research & Evaluation* 15.11 (2010), p. 2.

[42] H. C. Dijkerman, V. A. Wood, and R. L. Hewer. "Long-term outcome after discharge from a stroke rehabilitation unit." In: *J. R. Coll. Physicians Lond.* 30.6 (Nov. 1996), pp. 538–546.

[43] R. van Dinteren, M. Arns, M. L. Jongsma, and R. P. Kessels. "P300 development across the lifespan: a systematic review and meta-analysis." In: *PLOS ONE* 9.2 (2014), e87347.

[44] T. Doucet, F Muller, C Verdun-Esquer, X Debelleix, and P Brochard. "Returning to work after a stroke: A retrospective study at the Physical and Rehabilitation Medicine Center "La Tour de Gassies"." In: *Annals of physical and rehabilitation medicine* 55.2 (2012), pp. 112–127.

[45] Economist Intelligence Unit. "Preventing stroke: uneven progress." In: *A global policy research programme* (2017).

[46] B. Elsner, J. Kugler, M. Pohl, and J. Mehrholz. "Transcranial direct current stimulation (tDCS) for improving aphasia in patients with aphasia after stroke." In: *Cochrane Database Syst. Rev.* 5 (May 2015), p. CD009760.

[47] S. T. Engelter, M Gostynski, S Papa, M Frei, and B. et al. "Epidemiology of aphasia attributable to first ischemic stroke." In: *Am Heart Assoc* (2006).

[48] J. L. Evans, C. Selinger, and S. D. Pollak. "P300 as a measure of processing capacity in auditory and visual domains in specific language impairment." In: *Brain Research* 1389 (2011), pp. 93–102.

[49] J. Farquhar and N. J. Hill. "Interactions between pre-processing and classification methods for event-related-potential classification." In: *Neuroinformatics* 11.2 (2013), pp. 175–192.

[50] L. A. Farwell and E. Donchin. "Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials." In: *Electroencephalography and Clinical Neurophysiology* 70.6 (Dec. 1988), pp. 510–523.

[51] M. Fatourechi, A. Bashashati, R. K. Ward, and G. E. Birch. "EMG and EOG artifacts in brain computer interface systems: A survey." In: *Clinical Neurophysiology* 118.3 (2007), pp. 480–494.

[52] R. Fazel-Rezai, B. Z. Allison, C. Guger, E. W. Sellers, S. C. Kleih, and A. Kübler. "P300 brain computer interface: current challenges and emerging trends." In: *Frontiers in Neuroengineering* 5.14 (July 2012), pp. 1–14.

[53] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K.-R. Müller, and C. Grozea. "Subject-independent mental state classification in single trials." In: *Neural networks: The Official Journal of the International Neural Network Society* 22.9 (Nov. 2009), pp. 1305–1312.

[54] S. Fazli, S. Dähne, W. Samek, F. Bießmann, and K.-R. Müller. "Learning from more than one data source: Data fusion techniques for sensorimotor rhythm-based brain-computer interfaces." In: *Proceedings of the IEEE* 103.6 (May 2015), pp. 891–906.

[55] P. W. Ferrez and J. d. R. Millán. "Error-related EEG potentials generated during simulated brain–computer interaction." In: *IEEE Transactions on Biomedical Engineering* 55.3 (Feb. 2008), pp. 923–929.

[56] A. Fisher, J. Martin, W. Srikusalanukul, and M. Davis. "Trends in stroke survival incidence rates in older Australians in the new millennium and forecasts into the future." In: *J. Stroke Cerebrovasc. Dis.* 23.4 (Apr. 2014), pp. 759–770.

[57] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, NY, USA: 2001.

[58] A Furdea, S Halder, D. Krusienski, et al. "An auditory oddball (P300) spelling system for brain-computer interfaces." In: *Psychophysiology* 46.3 (Jan. 2009), pp. 617–625.

[59] S. Gao, Y. Wang, X. Gao, and B. Hong. "Visual and auditory brain–computer interfaces." In: *IEEE Transactions on Biomedical Engineering* 61.5 (May 2014), pp. 1436–1447.

[60] C. J. Gonsalvez and J. Polich. "P300 amplitude is determined by target-to-target interval." In: *Psychophysiology* 39.3 (2002), pp. 388–396.

[61] J. Grizou, I. Iturrate, L. Montesano, P.-Y. Oudeyer, and M. Lopes. "Calibration-free BCI based control." In: *Twenty-Eighth AAAI Conference on Artificial Intelligence*. Quebec, Canada, 2014, pp. 1213–1220.

[62] J. Grizou, I. Iturrate, L. Montesano, P.-Y. Oudeyer, and M. Lopes. "Interactive learning from unlabeled instructions." In: *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*. Québec, Canada, 2014, pp. 1–8.

[63] C. Groenegress, C. Holzner, C. Guger, and M. Slater. "Effects of P300-based BCI use on reported presence in a virtual environment." In: *Presence: Teleoperators and Virtual Environments* 19.1 (Feb. 2010), pp. 1–11.

[64] C. Guger, S. Daban, E. Sellers, et al. "How many people are able to control a P300-based brain–computer interface (BCI)?" In: *Neuroscience Letters* 462.1 (Sept. 2009), pp. 94–98.

[65] S. Halder, I. Käthner, and A. Kübler. "Training leads to increased auditory brain–computer interface performance of end-users with motor impairments." In: *Clinical Neurophysiology* 127.2 (2016), pp. 1288–1296.

[66] S. Halder, M. Rea, R. Andreoni, et al. "An auditory oddball brain–computer interface for binary choices." In: *Clinical Neurophysiology* 121.4 (2010), pp. 516–523.

[67] P. U. Heuschmann, O Busse, M Wagner, et al. "Schlaganfallhäufigkeit und Versorgung von Schlaganfallpatienten in Deutschland." In: *Aktuelle Neurol.* 37.07 (Oct. 2010), pp. 333–340.

[68] K. Hilari. "The impact of stroke: are people with aphasia different to those without?" In: *Disabil. Rehabil.* 33.3 (2011), pp. 211–218.

[69] J. Höhne and M. Tangermann. "How stimulation speed affects event-related potentials and BCI performance." In: *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE. 2012, pp. 1802–1805.

[70] J. Höhne, M. Schreuder, B. Blankertz, and M. Tangermann. "A novel 9-class auditory ERP paradigm driving a predictive text entry system." In: *Frontiers in Neuroscience* 5.99 (Aug. 2011), pp. 1–10.

[71] J. Höhne, K. Krenzlin, S. Dähne, and M. Tangermann. "Natural stimuli improve auditory BCIs with respect to ergonomics and performance." In: *Journal of Neural Engineering* 9.4 (2012), p. 045003.

[72] J. Höhne, D. Bartz, M. N. Hebart, K.-R. Müller, and B. Blankertz. "Analyzing neuroimaging data with subclasses: a shrinkage approach." In: *NeuroImage* 124 (2016), pp. 740–751.

[73] B. Hong, F. Guo, T. Liu, X. Gao, and S. Gao. "N200-speller using motion-onset visual response." In: *Clinical Neurophysiology* 120.9 (Sept. 2009), pp. 1658 –1666. ISSN: 1388-2457. DOI: http://dx.doi.org/10.1016/j.clinph.2009.06.026.

[74] W. Huber, K. Poeck, D. Weniger, and K. Willmes. *Aachener Aphasie Test (AAT)*. Verlag für Psychologie Hogrefe, 1983.

[75] D. Hübner, A. Schall, and M. Tangermann. "Two player online brain-controlled chess." In: *Engineering in Medicine and Biology Society (EMBC), 2019 Annual International Conference of the IEEE*. Under review. IEEE. 2019.

[76] D. Hübner, A. Schall, and M. Tangermann. "Unsupervised learning in a BCI chess application using learning from label proportions." In: *Brain-Computer Interfaces Journal* (2019). **Under review.**

[77] D. Hübner and M. Tangermann. "Challenging the assumption that auditory event-related potentials are independent and identically distributed." In: *Proceedings of the 7th International Brain-Computer Interface Meeting 2017: From Vision to Reality*. Graz, Austria.: Verlag der TU Graz, Sept. 2017, pp. 192–197.

[78] D. Hübner, P.-J. Kindermans, T. Verhoeven, and M. Tangermann. "Improving learning from label proportions by reducing the feature dimensionality." In: *Proceedings of the 7th International Brain-Computer Interface Meeting 2017: From Vision to Reality*. Ed. by G. R. Müller-Putz, D. Steyrl, S. C. Wriessnegger, and S. R. Verlag der Technischen Universität Graz, 2017, pp. 186–191. ISBN: 978-3-85125-533-1. DOI: 10.3217/978-3-85125-533-1-35.

[79] D. Hübner, T. Verhoeven, K. Schmid, K.-R. Müller, M. Tangermann, and P.-J. Kindermans. "Learning from label proportions in BCI — a symbiotic design for stimulus presentation and signal decoding." In: *The First Biannual Neuroadaptive Technology Conference*. 2017, pp. 31–33.

[80] D. Hübner, T. Verhoeven, K. Schmid, K.-R. Müller, M. Tangermann, and P.-J. Kindermans. "Learning from label proportions in brain-computer interfaces: online unsupervised learning with guarantees." In: *PLOS ONE* 12.4 (2017), e0175856.

[81] D. Hübner, A. Schall, N. Prange, and M. Tangermann. "Eyes-closed increases the usability of brain-computer interfaces based on auditory event-related potentials." In: *Frontiers in Human Neuroscience* 12 (2018).

[82] D. Hübner, T. Verhoeven, K.-R. Müller, P.-J. Kindermans, and M. Tangermann. "Unsupervised learning for brain-computer interfaces based on event-related potentials: review and online comparison." In: *IEEE Computational Intelligence Magazine* 13.2 (2018), pp. 66–77.

[83] I. Iturrate, L. Montesano, and J. Minguez. "Robot reinforcement learning using EEG-based reward signals." In: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. Anchorage, USA, 2010, pp. 4822–4829.

[84] I. Iturrate, J. Grizou, J. Omedes, P.-Y. Oudeyer, M. Lopes, and L. Montesano. "Exploiting task constraints for self-calibrated brain-machine interface control using error-related potentials." In: *PLOS ONE* 10.7 (July 2015), e0131491.

[85] I. Iturrate, R. Chavarriaga, L. Montesano, J. Minguez, and J. d. R. Millán. "Teaching brain-machine interfaces as an alternative paradigm to neuroprosthetics control." In: *Scientific reports* 5.13893 (Sept. 2015), pp. 1–10.

[86] H. H. Jasper. "The ten-twenty electrode system of the International Federation." In: *Electroencephalogr. Clin. Neurophysiol.* 10 (1958), pp. 370–375.

[87] V. Jayaram, M. Alamgir, Y. Altun, B. Schölkopf, and M. Grosse-Wentrup. "Transfer learning in brain-computer interfaces." In: *IEEE Computational Intelligence Magazine* 11.1 (Jan. 2016), pp. 20–31.

[88] C. Jeunet, B. N'Kaoua, and F. Lotte. "Advances in user-training for mental-imagery-based BCI control: psychological and cognitive factors and their neural correlates." In: *Progress in Brain Research* 228 (June 2016), pp. 3–35.

[89] J. Jin, B. Z. Allison, X. Wang, and C. Neuper. "A combined brain–computer interface based on P300 potentials and motion-onset visual evoked potentials." In: *Journal of Neuroscience Methods* 205.2 (Apr. 2012), pp. 265–276.

[90] S. Johannes, T. F. Münte, H. J. Heinze, and G. R. Mangun. "Luminance and spatial attention effects on early visual processing." In: *Cognitive Brain Research* 2.3 (July 1995), pp. 189–205.

[91] I. Käthner, C. A. Ruf, E. Pasqualotto, C. Braun, N. Birbaumer, and S. Halder. "A portable auditory P300 brain–computer interface with directional cues." In: *Clinical neurophysiology* 124.2 (2013), pp. 327–338.

[92] T. Kaufmann, S. Schulz, C. Grünzinger, and A. Kübler. "Flashing characters with famous faces improves ERP-based brain–computer interface performance." In: *Journal of neural engineering* 8.5 (2011), p. 056016.

[93] M.-L. Kauhanen, J. Korpelainen, P Hiltunen, et al. "Aphasia, depression, and non-verbal cognitive impairment in ischaemic stroke." In: *Cerebrovascular Diseases* 10.6 (2000), pp. 455–461.

[94] R. P. Kessels, M. J. Van Zandvoort, A. Postma, L. J. Kappelle, and E. H. De Haan. "The Corsi block-tapping task: standardization and normative data." In: *Applied neuropsychology* 7.4 (2000), pp. 252–258.

[95] P.-J. Kindermans, D. Verstraeten, and B. Schrauwen. "A Bayesian model for exploiting application constraints to enable unsupervised training of a P300-based BCI." In: *PLOS ONE* 7.4 (Apr. 2012), e33758.

[96] P.-J. Kindermans, D. Verstraeten, P. Buteneers, and B. Schrauwen. "How do you like your P300 speller: adaptive, accurate and simple?" In: *5th International Brain-Computer Interface Conference (BCI-2011)*. Graz, Austria, 2011, pp. 96–99.

[97] P.-J. Kindermans, M. Tangermann, K.-R. Müller, and B. Schrauwen. "Integrating dynamic stopping, transfer learning and language models in an adaptive zero-training ERP speller." In: *Journal of Neural Engineering* 11.3 (May 2014), p. 035005.

[98] P.-J. Kindermans, M. Schreuder, B. Schrauwen, K.-R. Müller, and M. Tangermann. "True zero-training brain-computer interfacing – an online study." In: *PLOS ONE* 9.7 (July 2014), e102504.

[99] S. Kleih, F Nijboer, S Halder, and A Kübler. "Motivation modulates the P300 amplitude during brain–computer interface use." In: *Clinical Neurophysiology* 121.7 (2010), pp. 1023–1031.

[100] D. S. Klobassa, T. M. Vaughan, P. Brunner, et al. "Toward a high-throughput auditory P300-based brain–computer interface." In: *Clinical neurophysiology* 120.7 (2009), pp. 1252–1261.

[101] A. Kok. "On the utility of P3 amplitude as a measure of processing capacity." In: *Psychophysiology* 38.3 (2001), pp. 557–577.

[102] M. Krauledat, M. Tangermann, B. Blankertz, and K.-R. Müller. "Towards zero training for brain-computer interfacing." In: *PLOS ONE* 3.8 (Aug. 2008), e2967.

[103] A. Kübler, N. Neumann, J. Kaiser, B. Kotchoubey, T. Hinterberger, and N. P. Birbaumer. "Brain-computer communication: self-regulation of slow cortical potentials for verbal communication." In: *Archives of Physical Medicine and Rehabilitation* 82.11 (Nov. 2001), pp. 1533–1539.

[104] L. I. Kuncheva. "Classifier ensembles for changing environments." In: *International Workshop on Multiple Classifier Systems*. Springer. 2004, pp. 1–15.

[105] L. I. Kuncheva, C. J. Whitaker, and A Narasimhamurthy. "A case-study on naïve labelling for the nearest mean and the linear discriminant classifiers." In: *Pattern Recognition* 41.10 (Oct. 2008), pp. 3010–3020.

[106] M. Kutas, G. McCarthy, and E. Donchin. "Augmenting mental chronometry: the P300 as a measure of stimulus evaluation time." In: *Science* 197.4305 (1977), pp. 792–795.

[107] D. Lakens. "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs." In: *Frontiers in psychology* 4 (2013), p. 863.

[108] J. M. C. Lam and W. P. Wodchis. "The relationship of 60 disease diagnoses and 15 conditions to preference-based health-related quality of life in Ontario hospital-based long-term care residents." In: *Med. Care* 48.4 (Apr. 2010), pp. 380–387.

[109] A. Laska, A Hellblom, V Murray, T Kahan, and M Von Arbin. "Aphasia in acute stroke and relation to outcome." In: *Journal of internal medicine* 249.5 (2001), pp. 413–422.

[110] O. Ledoit and M. Wolf. "A well-conditioned estimator for large-dimensional covariance matrices." In: *Journal of Multivariate Analysis* 88.2 (Feb. 2004), pp. 365–411.

[111] T.-S. Lee, S. Y. Quek, S. J. A. Goh, et al. "A pilot randomized controlled trial using EEG-based brain–computer interface training for a Chinese-speaking group of healthy elderly." In: *Clinical Interventions in Aging* 10 (2015), p. 217.

[112] R. Leeb, D. Friedman, G. R. Müller-Putz, R. Scherer, M. Slater, and G. Pfurtscheller. "Self-paced (asynchronous) BCI control of a wheelchair in virtual environments: a case study with a tetraplegic." In: *Computational Intelligence and Neuroscience* 2007.79642 (July 2007), pp. 1–8.

[113] A. Llera, M. A. van Gerven, V. Gómez, O. Jensen, and H. J. Kappen. "On the use of interaction error potentials for adaptive brain computer interfaces." In: *Neural Networks* 24.10 (Dec. 2011), pp. 1120–1127.

[114] M. Lopez-Gordo, E Fernandez, S Romero, F Pelayo, and A. Prieto. "An auditory brain–computer interface evoked by natural speech." In: *Journal of Neural Engineering* 9.3 (2012), p. 036013.

[115] F. Lotte. "Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces." In: *Proceedings of the IEEE* 103.6 (June 2015), pp. 871–890.

[116] F. Lotte, F. Larrue, and C. Mühl. "Flaws in current human training protocols for spontaneous brain-computer interfaces: lessons learned from instructional design." In: *Frontiers in Human Neuroscience* 7.568 (Sept. 2013), pp. 1–11.

[117] F. Lotte, L. Bougrain, A. Cichocki, et al. "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update." In: *Journal of Neural Engineering* 15.3 (2018), p. 031005.

[118] S. Lu, C. Guan, and H. Zhang. "Unsupervised brain computer interface based on intersubject information and online adaptation." In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 17.2 (Feb. 2009), pp. 135–145.

[119] G. Lucchese, F. Pulvermüller, B. Stahl, F. R. Dreyer, and B. Mohr. "Therapy-induced neuroplasticity of language in chronic post stroke aphasia: a mismatch negativity study of agrammatical and meaningful/less mini-constructions." In: *Front. Hum. Neurosci.* 10 (2016), p. 669.

[120] D. Lulé, Q. Noirhomme, S. C. Kleih, et al. "Probing command following in patients with disorders of consciousness using a brain–computer interface." In: *Clinical Neurophysiology* 124.1 (2013), pp. 101–106.

[121] J. Ma, L. Xu, and M. I. Jordan. "Asymptotic convergence rate of the EM algorithm for Gaussian mixtures." In: *Neural Computation* 12.12 (2000), pp. 2881–2907.

[122] P. Margaux, M. Emmanuel, D. Sébastien, B. Olivier, and M. Jérémie. "Objective and subjective evaluation of online error correction during P300-based spelling." In: *Advances in Human-Computer Interaction* 2012.4 (Jan. 2012), pp. 1–13.

[123] M. Meinzer, D. Djundja, G. Barthel, T. Elbert, and B. Rockstroh. "Long-term stability of improved language functions in chronic aphasia after constraint-induced aphasia therapy." In: *Stroke* 36.7 (July 2005), pp. 1462–1466.

[124] J. d. R. Millán, R. Rupp, G. R. Müller-Putz, et al. "Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges." In: *Frontiers in Neuroscience* 4.161 (Mar. 2010), pp. 1–15.

[125] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to linear regression analysis*. Vol. 821. John Wiley & Sons, 2012.

[126] J. I. Münßinger, S. Halder, S. C. Kleih, et al. "Brain painting: first evaluation of a new brain–computer interface application with ALS-patients and healthy volunteers." In: *Frontiers in Neuroscience* 4.182 (Nov. 2010), pp. 1–11.

[127] M. Musso, D. Hübner, S. Schwarzkopf, C. Weiller, and M. Tangermann. "A brain-computer interface for language training in chronic post-stroke aphasia patients." In: *Nature communications.* (2019). **Under preparation.**

[128] I. Nambu, M. Ebisawa, M. Kogure, S. Yano, H. Hokari, and Y. Wada. "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 (2013), e57174.

[129] L. F. Nicolas-Alonso and J. Gomez-Gil. "Brain computer interfaces, a review." In: *Sensors* 12.2 (Jan. 2012), pp. 1211–1279.

[130] F. Nijboer, A. Furdea, I. Gunst, et al. "An auditory brain–computer interface (BCI)." In: *Journal of Neuroscience Methods* 167.1 (2008), pp. 43–50.

[131] A. Nijholt, D. P.-O. Bos, and B. Reuderink. "Turning shortcomings into challenges: brain–computer interfaces for games." In: *Entertainment Computing* 1.2 (Apr. 2009), pp. 85–94.

[132] G. Nolfe, A. Cobianchi, L. Mossuto-Agatiello, and S. Giaquinto. "The role of P300 in the recovery of post-stroke global aphasia." In: *European Journal of Neurology* 13.4 (2006), pp. 377–384.

[133] R. C. Panicker, S. Puthusserypady, and Y. Sun. "Adaptation in P300 brain–computer interfaces: a two-classifier cotraining approach." In: *IEEE Transactions on Biomedical Engineering* 57.12 (July 2010), pp. 2927–2935.

[134] L. C. Parra, C. D. Spence, A. D. Gerson, and P. Sajda. "Recipes for the linear analysis of EEG." In: *NeuroImage* 28.2 (Nov. 2005), pp. 326–341.

[135] A. Patel, V. Berdunov, D. King, Z. Quayyum, R. Wittenberg, and M. Knapp. "Current, future and avoidable costs of stroke in the UK. Executive summary Part 2: societal costs of stroke in the next 20 years and potential returns from increased spending on research." In: *Association S (ed). London* (2017).

[136] G. Patrini, R. Nock, T. Caetano, and P. Rivera. "(Almost) No Label No Cry." In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Curran Associates, Inc., 2014, pp. 190–198.

[137] P. M. Pedersen, H. Stig Jørgensen, H. Nakayama, H. O. Raaschou, and T. S. Olsen. "Aphasia in acute stroke: incidence, determinants, and recovery." In: *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 38.4 (1995), pp. 659–666.

[138] G. Pfurtscheller, G. R. Müller, J. Pfurtscheller, H. J. Gerner, and R. Rupp. "'Thought' – control of functional electrical stimulation to restore hand grasp in a patient with tetraplegia." In: *Neurosci. Lett.* 351.1 (Nov. 2003), pp. 33–36.

[139] F. Pichiorri, G. Morone, M. Petti, et al. "Brain–computer interface boosts motor imagery practice during stroke recovery." In: *Annals of neurology* 77.5 (2015), pp. 851–865.

[140] M. J. Pickering and S. Garrod. "An integrated theory of language production and comprehension." In: *Behavioral and Brain Sciences* 36.4 (2013), pp. 329–347.

[141] C. Pokorny, D. S. Klobassa, G. Pichler, et al. "The auditory P300-based single-switch brain–computer interface: paradigm transition from healthy subjects to minimally conscious patients." In: *Artificial Intelligence in Medicine* 59.2 (2013), pp. 81–90.

[142] J. Polich. "Updating P300: an integrative theory of P3a and P3b." In: *Clinical Neurophysiology* 118.10 (Oct. 2007), pp. 2128–2148.

[143] A. Pollock, B. St George, M. Fenton, and L. Firkins. "Top ten research priorities relating to life after stroke." In: *Lancet Neurolology* 11.3 (Mar. 2012), p. 209.

[144] F. Pulvermüller, B. Neininger, T. Elbert, et al. "Constraint-induced therapy of chronic aphasia after stroke." In: *Stroke* 32.7 (2001), pp. 1621–1626.

[145] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le. "Estimating labels from label proportions." In: *Journal of Machine Learning Research* 10 (Oct. 2009), pp. 2349–2374.

[146] A. Ramos-Murguialday, D. Broetz, M. Rea, et al. "Brain–machine interface in chronic stroke rehabilitation: a controlled study." In: *Annals of neurology* 74.1 (2013), pp. 100–108.

[147] R. G. Real, S. Veser, H. Erlbeck, et al. "Information processing in patients in vegetative and minimally conscious states." In: *Clinical Neurophysiology* 127.2 (2016), pp. 1395–1402.

[148] C.-L. Ren, G.-F. Zhang, N. Xia, et al. "Effect of low-frequency rTMS on aphasia in stroke patients: a meta-analysis of randomized controlled trials." In: *PLoS One* 9.7 (July 2014), e102557.

[149] A. F. Salazar-Gomez, J. DelPreto, S. Gil, F. H. Guenther, and D. Rus. "Correcting robot mistakes in real time using EEG signals." In: *IEEE International Conference on Robotics and Automation*. 2017.

[150] M. T. Sarno and E. Levita. "Some observations on the nature of recovery in global aphasia after stroke." In: *Brain and Language* 13.1 (1981), pp. 1–12.

[151] J. Schäfer, K. Strimmer, et al. "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics." In: *Statistical Applications in Genetics and Molecular Biology* 4.1 (Nov. 2005), p. 32.

[152] M. Schreuder, B. Blankertz, and M. Tangermann. "A new auditory multi-class brain-computer interface paradigm: Spatial hearing as an informative cue." In: *PLOS ONE* 5.4 (Apr. 2010), e9813.

[153] M. Schreuder, T. Rost, and M. Tangermann. "Listen, you are writing! Speeding up online spelling with a dynamic auditory BCI." In: *Frontiers in Neuroscience* 5 (2011), p. 112.

[154] M. Schultze-Kraft, D. Birman, M. Rusconi, et al. "The point of no return in vetoing self-initiated movements." In: *Proceedings of the National Academy of Sciences* 113.4 (2016), pp. 1080–1085.

[155] E. W. Sellers and E. Donchin. "A P300–based brain–computer interface: Initial tests by ALS patients." In: *Clinical Neuropsychology* 117.3 (Mar. 2006), pp. 538–548.

[156] P. Shenoy, M. Krauledat, B. Blankertz, R. P. Rao, and K.-R. Müller. "Towards adaptive classification for BCI." In: *Journal of Neural Engineering* 3.1 (Mar. 2006), R13–R23.

[157] N. Simon, I. Käthner, C. A. Ruf, E. Pasqualotto, A. Kübler, and S. Halder. "An auditory multiclass brain-computer interface with natural stimuli: Usability evaluation with healthy participants and a motor impaired end user." In: *Frontiers in Human Neuroscience* 8.1039 (Jan. 2015), pp. 1–14.

[158] J. G. Snodgrass and M Vanderwart. "A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity." In: *J. Exp. Psychol. Hum. Learn.* 6.2 (Mar. 1980), pp. 174–215.

[159] J. G. Snodgrass and M. Vanderwart. "A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity." In: *Journal of experimental psychology: Human learning and memory* 6.2 (1980), p. 174.

[160] M. Spüler, W. Rosenstiel, and M. Bogdan. "Online adaptation of a c-VEP brain-computer interface (BCI) based on error-related potentials and unsupervised learning." In: *PLOS ONE* 7.12 (Dec. 2012), e51077.

[161] B. Stahl, B. Mohr, F. R. Dreyer, G. Lucchese, and F. Pulvermüller. "Using language for social interaction: Communication mechanisms promote recovery from chronic non-fluent aphasia." In: *Cortex* 85 (Dec. 2016), pp. 90–99.

[162] S. Sutton, M. Braren, J. Zubin, and E. John. "Evoked-potential correlates of stimulus uncertainty." In: *Science* 150.3700 (1965), pp. 1187–1188.

[163] M. Tangermann, N. Schnorr, and M. Musso. "Towards Aphasia Rehabilitation with BCI." In: *Proceedings of the 6th International Brain-Computer Interface Conference*. Graz, Austria: Verlag der TU Graz, 2014, pp. 65–68.

[164] M. Tangermann, J. Höhne, M. Schreuder, et al. "Data driven neuroergonomic optimization of BCI stimuli." In: *5th International Brain-Computer Interface Conference 2011*. Verlag der Technischen Universität Graz, Sept. 2011, pp. 160–163.

[165] M. Tangermann, M. Schreuder, S. Dähne, et al. "Optimized stimulation events for a visual ERP BCI." In: *Int. J. Bioelectromagn* 13.3 (Nov. 2011), pp. 119–120.

[166] M. Tangermann, D. Hübner, S. Schwarzkopf, C. Weiller, and M. Musso. "Effects on language ability induced by BCI-Based Training of Patients with Aphasia." In: *Proceedings of the Seventh International Brain-Computer Interface Meeting,* Pacific Grove, USA: Verlag der TU Graz, May 2018, p. 110.

[167] R. Teasell, A. Cotoi, J. Chow, et al. *Evidence-based review of stroke rehabilitation (18th edition)*. 2018.

[168] J. Thielen, P. v. d. Broek, J. Farquhar, and P. Desain. "Broad-Band visually evoked potentials: re(con)volution in brain-computer interfacing." In: *PLOS ONE* 10.7 (July 2015), e0133797. ISSN: 1932-6203.

[169] G. Townsend, B. LaPallo, C. Boulay, et al. "A novel P300-based brain–computer interface stimulus presentation paradigm: moving beyond rows and columns." In: *Clinical neurophysiology* 121.7 (2010), pp. 1109–1120.

[170] M. S. Treder and B. Blankertz. "(C)overt attention and visual speller design in an ERP-based brain-computer interface." In: *Behavioral and Brain Functions* 6.28 (May 2010), pp. 1–13.

[171] K. F. Van Orden, W. Limbert, S. Makeig, and T.-P. Jung. "Eye activity correlates of workload during a visuospatial memory task." In: *Human factors* 43.1 (2001), pp. 111–121.

[172] T. Verhoeven. "Brain-computer interfaces with machine learning: a symbiotic approach." PhD thesis. Ghent University, 2017.

[173] T. Verhoeven, P. Buteneers, J. Wiersema, J. Dambre, and P.-J. Kindermans. "Towards a symbiotic brain–computer interface: Exploring the application–decoder interaction." In: *Journal of Neural Engineering* 12.6 (Nov. 2015), p. 066027.

[174] T. Verhoeven, D. Hübner, M. Tangermann, K.-R. Müller, J. Dambre, and P.-J. Kindermans. "Improving zero-training brain-computer interfaces by mixing model estimators." In: *Journal of Neural Engineering* 14.3 (Apr. 2017), p. 036021.

[175] J. J. Vidal. "Toward direct brain-computer communication." In: *Annual review of Biophysics and Bioengineering* 2.1 (1973), pp. 157–180.

[176] C. Vidaurre, M. Kawanabe, P. von Bünau, B. Blankertz, and K.-R. Müller. "Toward unsupervised adaptation of LDA for brain–computer interfaces." In: *IEEE Transactions on Biomedical Engineering* 58.3 (Nov. 2010), pp. 587–597.

[177] M. A. Vollebregt, M. van Dongen-Boomsma, J. K. Buitelaar, and D. Slaats-Willemse. "Does EEG-neurofeedback improve neurocognitive functioning in children with attention-deficit/hyperactivity disorder? A systematic review and a double-blind placebo-controlled study." In: *Journal of Child Psychology and Psychiatry* 55.5 (2014), pp. 460–472.

[178] I. Winkler, S. Haufe, and M. Tangermann. "Automatic classification of artifactual ICA-components for artifact removal in EEG signals." In: *Behavioral and Brain Functions* 7.1 (2011), p. 30.

[179] I. Winkler, S. Brandl, F. Horn, E. Waldburger, C. Allefeld, and M. Tangermann. "Robust artifactual independent component classification for BCI practitioners." In: *Journal of Neural Engineering* 11.3 (2014), p. 035013.

[180] J. Wolpaw and E. W. Wolpaw. *Brain-computer interfaces: principles and practice*. New York: Oxford University Press, 2012.

[181] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. "Brain–computer interfaces for communication and control." In: *Clinical Neurophysiology* 113.6 (June 2002), pp. 767–791.

[182] J. Xiao, Q. Xie, Y. He, et al. "An auditory BCI system for assisting CRS-R behavioral assessment in patients with disorders of consciousness." In: *Scientific reports* 6 (2016), p. 32917.

[183] S.-K. Yeom, S. Fazli, K.-R. Müller, and S.-W. Lee. "An efficient ERP-based brain-computer interface using random set presentation and face familiarity." In: *PLOS ONE* 9.11 (Nov. 2014), e111157.

[184] T. O. Zander, L. R. Krol, N. P. Birbaumer, and K. Gramann. "Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity." In: *Proceedings of the National Academy of Sciences* 113.52 (2016), pp. 14898–14903.

[185] T. Zeyl, E. Yin, M. Keightley, and T. Chau. "Partially supervised P300 speller adaptation for eventual stimulus timing optimization: Target confidence is superior to error-related potential score as an uncertain label." In: *Journal of neural engineering* 13.2 (Feb. 2016), p. 026008.

[186] S. Zhou, B. Z. Allison, A. Kübler, A. Cichocki, X. Wang, and J. Jin. "Effects of background music on objective and subjective performance measures in an auditory BCI." In: *Frontiers in Computational Neuroscience* 10 (2016), p. 105.

[187] C. Zickler, S. Halder, S. C. Kleih, C. Herbert, and A. Kübler. "Brain painting: Usability testing according to the user-centered design in end users with severe motor paralysis." In: *Artificial Intelligence in Medicine* 59.2 (Oct. 2013), pp. 99–110.

[188] P. Zimmermann and B. Fimm. "A test battery for attentional performance." In: *Applied neuropsychology of attention. Theory, diagnosis and rehabilitation* (2002), pp. 110–151.