# Applications of Machine Learning in Dynamic Settings

-

## From Computational Neuroscience to Natural Language Processing

Stephanie Brandl

# Applications of Machine Learning in Dynamic Settings

-

## From Computational Neuroscience to Natural Language Processing

vorgelegt von
Dipl.-Math.
Stephanie Brandl

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

Doktorin der Naturwissenschaften
– Dr. rer. nat. –

genehmigte Dissertation

Promotionsausschuss:

| | |
|---|---|
| Vorsitzender | Prof. Dr. Volker Markl |
| Gutachter | Prof. Dr. Klaus-Robert Müller |
| Gutachter | Prof. Dr. Benjamin Blankertz |
| Gutachter | Prof. Dr. Alan Akbik |

Tag der wissenschaftlichen Aussprache: 4. Mai 2021

Berlin 2021

# Abstract

The field of machine learning is deeply intertwined with the dynamics that determine technical progress in our lives. It benefits invaluably from the increasing amount of available data and the access to computing power. Simultaneously, it drives those changes by continuously improving methodologies and thus extending the possible applications in our everyday lives. In order to keep up with those changes, machine learning models need to tackle the obstacles that arise in dynamic settings. This thesis contributes solutions to those challenges in the fields of computational neuroscience and natural language processing.

The first contribution researches possible challenges when moving brain-computer interfaces out of the lab and into the real world. We have recorded data from 16 participants conducting a motor imagery task while handling secondary distraction tasks that simulate everyday life situations. We investigate the artifacts that contaminate the data which makes it difficult to successfully classify within the standard pipeline. We propose two approaches that tackle those difficulties and significantly improved classification results.

In the second contribution, we propose a new version of Source Power Co-Modulation: Fourier-SPoC (f-SPoC) extracts brain components maximally correlating with a periodic target function. We apply f-SPoC to the magnetoencephalogram (MEG) from a passive beat listening task to find neural correlates that show rhythmic entrainment with the regular beat. The resulting components show very regular patterns of peaks and troughs in the spectrogram on the group average and even on some individuals.

The third contribution provides two new methods: Word2Vec with Structure Constraint (W2V&Constr) and Word2Vec with Structure Prediction (W2V&Pred) with which we can learn dynamic word embeddings of high quality based on the underlying structure in the dataset (W2V&Constr) and are even able to predict the structure of the dataset (W2V&Pred). We apply both methods to three English datasets of different structures where we achieve significant improvements on both the embedding accuracy and the structure score compared to the baseline and capture interesting insights with respect to outliers and the underlying data structure. Furthermore, we apply W2V&Pred on a German dataset in an explorative experiment to find connections between different authors where the underlying structure is not known a-priori.

# Zusammenfassung

Maschinelles Lernen (ML) steht in enger Wechselwirkung mit den Dynamiken, die den technischen Fortschritt in unserem Leben bestimmen. Einerseits verbessert der Zugang zu fortlaufend steigender Rechenleistung und immer größeren Datenmengen die Genauigkeit der ML-Algorithmen. Andererseits treibt es diese Veränderungen voran, indem Methoden kontinuierlich verbessert werden und sich so die Anwendungsmöglichkeiten in unserem Alltag erweitern. Um mit all diesen Entwicklungen Schritt halten zu können, müssen ML-Modelle die Hürden bewältigen, die in dynamischen Umgebungen auftreten. Diese Dissertation trägt zu Lösungen für diese Herausforderungen in den Bereichen Computational Neuroscience und Natural Language Processing bei.

Der erste Beitrag untersucht mögliche Hürden für den Transfer von Gehirn-Computer-Schnittstellen aus dem Labor in die reale Welt. Dafür wurden Elektroenzephalographie-Daten (EEG) von 16 Teilnehmer:innen aufgenommen, die sich primär motorische Handbewegungen vorstellen, während sie zusätzlich sekundäre Aufgaben bewältigen, die Alltagssituationen simulieren. Artefakte, welche die Daten stören und eine erfolgreiche Klassifizierung mit der Standard-Pipeline erschweren, werden untersucht. Es werden zwei Ansätze präsentiert, die diese Schwierigkeiten überwältigen und die Klassifikationsergebnisse signifikant verbessern.

Im zweiten Beitrag wird eine neue Version von Source Power Co-Modulation (SPoC) vorgestellt: Fourier-SPoC (f-SPoC) extrahiert Gehirnkomponenten, die maximal mit einer periodischen Zielfunktion korrelieren. Wir wenden f-SPoC auf Magnetoenzephalographie-Aufnahmen (MEG) aus einer passiven Beat-Höraufgabe an, um neuronale Korrelate zu finden, die ein rhythmisches Entrainment mit dem regelmäßigen Rhythmus zeigen. Die resultierenden Komponenten zeigen im Gruppendurchschnitt und sogar bei einigen Individuen sehr regelmäßige Muster im Spektrogramm.

Im dritten Beitrag werden zwei neue Methoden präsentiert: Word2Vec with Structure Constraint (W2V&Constr) und Word2Vec with Structure Prediction (W2V&Pred), mit denen wir dynamische Worteinbettungen von hoher Qualität entweder basierend auf der zugrundeliegenden Struktur in den Datensätzen lernen können (W2V&Constr) oder einen Schritt weiter gehen und auch die Struktur des Datensatzes vorhersagen (W2V&Pred). Beide Methoden wurden auf drei beziehungsweise vier verschiedenen Datensätzen mit unterschiedlichen Strukturen ausgewertet und konnten sowohl bei der Genauigkeit der Worteinbettungen als auch bei der Vorhersage der Struktur im Vergleich zu Baseline-Methoden signifikante Verbesserungen erzielen. Darüber hinaus konnten interessante Erkenntnisse in Bezug auf Ausreißer und die zugrunde liegende Datenstruktur gewonnen werden.

# Acknowledgments

This PhD has been a long and sometimes bumpy journey, looking back I am incredibly grateful to the people who have been with me at one point or the other.

Thanks to the entire Machine Learning group, in particular to Alex, Daniel, David, Kristof, Max, Miriam, Oliver, Thomas and Vera, without you this would not have been half the fun it has been.

A huge thank you goes to Klaus for his supervision and for giving me the freedom to pursue the research I am passionate about. Thank you for creating this special atmosphere in the Machine Learning group where collaboration and support are valued more than any competition.

I thank Benjamin for welcoming me to the BBCI group, for his open-ear and open-door policy and for his unique way of supervision and team building that drives the BBCI group. Thanks to the old and new BBCI group, I will truly miss the "big-office atmosphere" and the joint lunches.

Thanks goes also to Andrea, Cecilia and Dominik for their administrative and technical support over the many years, we would be completely lost without you.

Over the years, I have enjoyed working together with a number of people without whom this thesis would not have been possible. Thank you Manon for your support and advice over the time, for the good conversations we had, research-related or not. Good things take time ;) Thank you Sven for letting me in on your wide-ranging knowledge of signal processing and for handing me over the f-SPoC project. Thank you Wojciech and Johannes for your support and supervision on the BCI project. Thank you Shin for the countless conversations on how machine learners do what machine learners usually do, we and the dynamics project crucially benefited from your input. A very special thanks goes to David for the very close collaboration over the last 3.5 years and for his patience and support during my first steps into the NLP world. I also thank the NLP crew for the Tuesday meetings, discussing papers (and more) with you has been fun. I also want to thank the people from MIB in Aarhus, in particular Elvira and Niels for answering the countless questions about MEG and for the fruitful collaboration.

Thanks to Alex, Daniel, David, Mattia, Miriam, Oliver and Thomas for proof-reading and your valuable input on this thesis.

I am deeply grateful for all the fantastic people in my life outside academia, thanks for being there: Markus, Mattia, Mer, Sarah, Tina & many more. You are truly amazing.

Zu guter Letzt, ein riesiges Dankeschön an meine Eltern und Luc, ohne eure Unterstützung über all die Jahre wäre ich heute nicht da wo ich bin. Danke!

# Contents


**1 Introduction** **1**
- 1.1 Challenges in the application of ML in dynamically changing settings . . 1
- 1.2 Contribution and outline of this thesis . . . . . 2
- 1.3 List of Publications . . . . . 3

**2 Fundamentals of Signal Processing in Neuroscience** **7**
- 2.1 Measures of brain activity . . . . . 7
  - 2.1.1 Electroencephalography . . . . . 7
  - 2.1.2 Magnetoencephalography . . . . . 8
- 2.2 EEG/MEG Signals . . . . . 8
- 2.3 Brain-Computer Interfaces . . . . . 9
- 2.4 The General Linear Model . . . . . 9
  - 2.4.1 Forward Model . . . . . 10
  - 2.4.2 Backward Model . . . . . 10
- 2.5 Spatial Filtering . . . . . 11
  - 2.5.1 Common Spatial Patterns . . . . . 11
  - 2.5.2 Spatio-Spectral Decomposition . . . . . 12
  - 2.5.3 Source Power Co-Modulation . . . . . 13
- 2.6 Linear Classification . . . . . 13
  - 2.6.1 Fisher's Linear Discriminant Analysis . . . . . 14
  - 2.6.2 Regularized LDA . . . . . 14
- 2.7 Limitations and Challenges . . . . . 15

**3 Motor Imagery under Distraction: An Evaluation Study** **19**
- 3.1 Motivation . . . . . 19
- 3.2 Methods . . . . . 20
  - 3.2.1 Participants . . . . . 20
  - 3.2.2 Data Acquisition . . . . . 20
  - 3.2.3 Experimental Setup . . . . . 20
  - 3.2.4 Baseline Analysis . . . . . 24
- 3.3 Evaluation of in-lab training . . . . . 26
  - 3.3.1 Classification based on calibration data . . . . . 26
  - 3.3.2 Why poor performance? . . . . . 27
  - 3.3.3 Artifacts . . . . . 28
  - 3.3.4 Feature Shifts . . . . . 36
  - 3.3.5 Non-discriminativity . . . . . 38
- 3.4 Evaluation of new strategies for out-of-lab . . . . . 38
  - 3.4.1 Improvement via artifact reduction . . . . . 39
  - 3.4.2 Improvement via classifier ensemble: ensemble CSP . . . . . 39
  - 3.4.3 Improvement via 2-step classification: 2-step CSP . . . . . 40
- 3.5 Discussion . . . . . 42

**4 Tracking auditory beat-based entrainment in MEG via SPoC** **47**
4.1 Motivation . . . 47
4.2 Methods . . . 48
4.2.1 MEG Recordings . . . 49
4.2.2 Preprocessing . . . 49
4.3 Sensor-Level Analysis . . . 49
4.3.1 Analysis Pipeline . . . 50
4.3.2 Results . . . 51
4.4 Spatial Filtering . . . 52
4.4.1 Fourier-SPoC . . . 52
4.4.2 Results . . . 53
4.5 Effects over time . . . 59
4.5.1 Changes in the temporal domain . . . 59
4.5.2 Changes in the spectral domain . . . 60
4.6 Discussion . . . 63

**5 Fundamentals 2: Natural Language Processing** **67**
5.1 Vectorial representation of words . . . 67
5.1.1 One-Hot Encodings . . . 67
5.1.2 Distributed word embeddings . . . 68
5.1.3 Contextualized word embeddings . . . 69
5.2 Evaluation on analogy tests . . . 69
5.3 Limitations and Challenges . . . 70

**6 Dynamic Word Embeddings with Structure Prediction** **73**
6.1 Motivation . . . 73
6.2 Methods . . . 74
6.2.1 Dynamic Word2Vec (DW2V) . . . 75
6.2.2 Word2Vec with Structure Constraint (W2V&Constr) . . . 75
6.2.3 Word2Vec with Structure Prediction (W2V&Pred) . . . 76
6.2.4 Optimization . . . 77
6.3 Experiments . . . 77
6.3.1 Datasets . . . 77
6.3.2 Preprocessing . . . 79
6.3.3 Evaluation Metrics . . . 79
6.3.4 Baseline Analysis . . . 81
6.3.5 Parameter Selection . . . 82
6.4 Results . . . 82
6.4.1 New York Times . . . 82
6.4.2 WikiFoS . . . 86
6.4.3 WikiPhil . . . 87
6.4.4 DTA . . . 88
6.4.5 Parameter Selection . . . 90
6.5 Discussion . . . 91

**7 Summary & Conclusion** **97**

**Bibliography** **101**

**List of Figures** **113**

**List of Tables** **114**

**A Toy data experiments for W2V&Pred** **115**
A.1 Data Generation . . . . . 115
A.1.1 Scenario 1: 2 clusters in 10 slices . . . . . 116
A.1.2 Scenario 2: 10 clusters in 10 slices . . . . . 116
A.1.3 Scenario 3: 10 chronological clusters in 10 slices . . . . . 116
A.2 Grid search over parameters . . . . . 116
A.3 Results . . . . . 117

# 1

# Introduction

## 1.1 Challenges in the application of machine learning in dynamically changing settings

Our lives are driven by constant changes. Although the life span of one generation is insignificant compared to the world's history, clearly the circumstances we grow up in for instance with respect to technical progress are very different from those of our parents' generation. Furthermore, the subjects that dominate the public conversation and the way we address them continuously change. As one out of many examples, some recently started talking about *climate crisis* instead of *climate change* to emphasize the seriousness of the matter[1]. Our brains never stop processing information. Even when we sleep, which might come closest of what we consider "doing nothing", our brains continue to learn. Those external and internal dynamics are deeply entangled, a prerequisite for change and thus cannot be isolated from another.

The field of machine learning, its application and models both benefit and drive technical progress. Without the access to bigger datasets and more computing power it would be impossible to constantly improve machine learning models. That in turn enables us to expand the possibilities in our everyday lives where we apply machine learning. In order to further extend the application of machine learning, it is crucial to build models that are able to handle dynamics. Throughout this thesis we will investigate both internal and external dynamics on the examples of the human brain and natural language.

Machine learning has had an essential impact on the field of computational neuroscience. A brain-computer interface (BCI) aims to translate brain signals into computer commands to, e.g. control a spelling device or a wheelchair without the need for muscular control (Müller-Putz et al., 2015). Incorporating machine learning-based classification pipelines into BCIs has massively sped up and improved feature selection and classification performance (Blankertz et al., 2002). The question remains as to how machine learning methods can deal with the interference of distractions in everyday life with a BCI, as external stimuli often introduce additional variance in the data that need to be accounted for in the analysis pipeline to maintain a good classification performance.

Machine Learning methods can also be applied to identify neural correlates that are related to specific processes in the brain such as rhythmic entrainment with an external

[1] https://www.theguardian.com/environment/2019/oct/16/guardian-language-changes-climate-environment

auditory beat stimulus. One question of interest then is to explore what is going on when the same regular beat is played over several minutes. In other words, opposed to the previous scenario, here, we control the external disruption in a way that the brain gets used to it or might get bored. Can machine learning methods capture a neural correlate that shows entrainment with the beat and how does this behave over time?

Combining the effects in question over time and context, we will further investigate how we can tackle both in language that evolves amongst others over time and over context, e.g. domain. Are we able to track dynamics in language, i.e. can we extend existing approaches in a way that learns dynamic word representations and predicting the pace of such changes?

## 1.2 Contribution and outline of this thesis

This thesis contributes solutions to the challenges arising in machine learning methods during dynamic changes due to external or internal influences. We tackle three different research questions in the fields of computational neuroscience and natural language processing. Each originates from a different setting with individual obstacles that will be addressed in the following chapters.

**Chapter 3 Motor Imagery under Distraction: An Evaluation Study**

- Setting: Brain-computer interfaces (BCIs) need to be transferred out of the lab and into the real world in order to be applicable in everyday life situations.
- Research Question: *How do artifacts that occur in everyday life situations influence the classification performance of a motor imagery-based BCI and how can we tackle those challenges?*
- Contribution: We recorded a motor imagery-based BCI study that includes five secondary distraction tasks and one control task on top of the primary motor imagery task to mimic out-of-lab scenarios. We investigate respective artifacts and feature shifts and propose new classification pipelines to handle both.

**Chapter 4 Tracking auditory beat-based entrainment in MEG via SPoC**

- Setting: Rhythm and timing are crucial in our daily lives. The ability to detect and process a regular rhythmic beat in an auditory stream of acoustic events is one of the essential mechanisms to our musical, speech and language skills. This ability can be disturbed in neurological as well as developmental disorders with marked consequences to the individuals' life quality. It is therefore of critical importance to first understand how the brain tracks the beat, to then find ways to monitor and support the rehabilitation of dysfunctional timing processes.
- Research Question: *How to extract relevant brain sources that capture entrainment processes with respect to a regular beat?*

- Contribution: We propose a new version of Source Power Co-Modulation (SPoC) that targets co-modulation with periodic data: Fourier-SpoC (f-SPoC). Based on f-SPoC we present a new pipeline to extract maximally relevant components in the source-space with respect to the oscillatory power in the frequency band of interest and, most importantly, the beat-related modulation of that power without the need for any prior anatomical knowledge.

**Chapter 6 Dynamic Word Embeddings with Structure Prediction**

- Setting: Vectorial representation of language, i.e. word embeddings, are crucial in the field of natural language processing and highly dependent on the corpora they are trained on. Static word embeddings - where exactly one vector is trained per word - are not able to capture language shifts over time or different domains.

- Research Question: *How can we train specific word embeddings of good quality on relatively small datasets for different epochs, domains or any other category in a common vector space?*

- Contribution: We propose two new methods: *Word2Vec with Structure Constraint (W2V&Constr)* and *Word2Vec with Structure Prediction (W2V&Pred)* that are able to simultaneously train and align word embeddings for different data slices organized in any given structure that can be represented by an adjacency matrix. With W2V&Pred we can furthermore predict the structure of a dataset for applications where the structure is unknown. We evaluate the methods on four different datasets and show promising results with respect to the quality of word embeddings as well as data structures.

The remaining part of this thesis is organized as follows: Chapter 2 and Chapter 5 summarize basic concepts in computational neuroscience and natural language processing respectively. Chapter 7 concludes this work with a summary.

## 1.3 List of Publications

Publications printed in *italic* were included in this thesis.

**Journal articles (published)**

1. *S Brandl and B Blankertz (2020d). "Motor Imagery Under Distraction—An Open Access BCI Dataset". In:* Frontiers in neuroscience *14, p. 967*

2. *S Brandl, L Frølich, J Höhne, KR Müller and W Samek (2016b). "Brain–Computer Interfacing under Distraction: An Evaluation Study". In:* Journal of Neural Engineering *13.5, p. 056012*

3. I Winkler, S Brandl, F Horn, E Waldburger, C Allefeld and M Tangermann (2014). "Robust artifactual independent component classification for BCI practitioners". In: *Journal of neural engineering* 11.3, p. 035013

### Journal articles (in preparation)

1. *S Brandl, NT Haumann, S Radloff, S Dähne, L Bonetti, P Vuust, E Brattico and M Grube (2021). "Fourier SPoC: A customised machine-learning analysis pipeline for auditory beat-based entrainment in the MEG". In:* BioRxiv

2. *D Lassner, S Brandl, A Baillot and S Nakajima (2021).* Word Embeddings with Structure Prediction
   * shared first authorship

### Conference contributions

1. *S Brandl and D Lassner (2019). "Times Are Changing: Investigating the Pace of Language Change in Diachronic Word Embeddings". In:* Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pp. 146–150*

2. *S Brandl, KR Müller and W Samek (2016a). "Alternative CSP approaches for multimodal distributed BCI data". In:* 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, pp. 003742–003747*

3. *S Brandl, J Höhne, KR Müller and W Samek (2015a). "Bringing BCI into everyday life: Motor imagery in a pseudo realistic environment". In:* 2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, pp. 224–227*

4. S Brandl, KR Müller and W Samek (2015b). "Robust common spatial patterns based on Bhattacharyya distance and Gamma divergence". In: *The 3rd International Winter Conference on Brain-Computer Interface*. IEEE, pp. 1–4

### Preprints

1. S Brandl, D Lassner and M Alber (2020a). "Balancing the composition of word embeddings across heterogenous data sets". In: *arXiv preprint arXiv:2001.04693*

2. S Brandl and D Lassner (2020c). "Early Corona Twitter Dataset". In: *HAL preprint hal-02861167*

### Datasets

1. *S Brandl (2015).* BCI under distraction: Motor imagery in a pseudo realistic environment*. http://dx.doi.org/10.14279/depositonce-9827.2*

2. S Brandl and D Lassner (2020b). *Corona Twitter Dataset: 16 February 2020 - 03 March 2020.* http://dx.doi.org/10.14279/depositonce-10012

# 2

# Fundamentals of Signal Processing in Neuroscience

Understanding the human brain has been challenging researchers in various fields for many decades. Ambitious projects funded by governments or large corporates have additionally risen awareness to the field of neuroscience. We are far from decoding the entire brain and also our functional understanding of it is still very limited. Yet significant progress has been achieved where improvements both in measurement techniques and signal processing play a key role.

In the following section, we will recap basic principles of electro- and magnetoencephalography with information from (Hämäläinen et al., 1993; Mulert et al., 2009; Buzsáki et al., 2012; Da Silva, 2013)

## 2.1 Measures of brain activity

Information flows in form of local electric currents through our neural system where it is processed in billions of neurons within milliseconds. If the activity is synchronised in time and space by a larger population of neurons, electric and magnetic fields large enough are produced so that they can be measured noninvasively outside the scalp. In the following, the basics of electro- and magnetoencephalography will be briefly described to give the reader a basic understanding of the two measurement techniques as both have been applied within the scope of this thesis. It is important to note that, throughout this thesis, we are not interested in directly comparing both methods as they have been applied independently in different research projects.

### 2.1.1 Electroencephalography

Neural activity originating from an assembly of neurons in synchrony produce an electric field as a result of local electric current sources promoting a volume current flow in the whole head which can be quantified by electroencephalography (EEG) on the scalp surface. EEG measures a spatially smeared superposition of distributed cortical activity on the scalp with a high temporal resolution ($<$1 ms). EEG recordings are relatively cheap and the development of mobile hardware over recent years has made it easier to apply mobile EEG where participants can move freely.

### 2.1.2 Magnetoencephalography

Magnetic fields originating from the same local currents caused by neural activity can be measured with arrays of superconducting quantum interference devices (SQUID) placed outside the head that form the magnetoencephalography (MEG). MEG has a similarly high temporal resolution and a slightly better spatial resolution than EEG as cerebrospinal fluid, skull and skin have larger effects on the propagation of the electric than magnetic field. MEG devices are far more expensive than EEG and cannot as easily be used in moving participants.

## 2.2 EEG/MEG Signals

Although measuring devices clearly differ for EEG and MEG, the signal itself when measured simultaneously can originate from the same neural activity. It has been shown that EEG and MEG signals are partially independent thus recording them both can be beneficial (Malmivuo, 2012). Signal processing of recorded EEG/MEG signals works very similar and we will not distinguish between the two for the rest of this chapter. EEG/MEG signals consist of 3 main components (Pfurtscheller et al., 1999):

1. Event-related potentials (ERPs): Time-locked changes in neural activity triggered by a specific stimuli. In MEG we refer to those as event-related fields (ERFs).

2. Neural oscillations: Frequency-specific changes of the ongoing EEG activity. An increase/decrease of power in a specific frequency band is called an event-related synchronisation/desynchronisation (ERS/ERD).

3. Artifacts: Artifacts originating from electrophysiological sources outside the brain (e.g. muscular or ocular) or non-physiological artifacts originating for instance from electromagnetic interference.

Throughout this thesis we will mainly focus on neural oscillations and in parts also on artifacts.

Neural oscillations in combination with specific frequency bands have been tied to cognitive processes such as motor control (Chatrian et al., 1959; Howe et al., 1972; Pfurtscheller, 1992; Pfurtscheller et al., 1994; Pfurtscheller et al., 1997), rhythm perception (Fujioka et al., 2009; Cirelli et al., 2014; Chang et al., 2016), workload (Gevins et al., 2003; Holm et al., 2009) and attention (Harmony et al., 1996; Klimesch, 1999). This makes it both even more interesting and important to investigate neural oscillations and their behaviour to build useful algorithms which can be applied for instance in brain-computer interfaces.

## 2.3 Brain-Computer Interfaces

Several diseases, such as amyotrophic lateral sclerosis (ALS), brainstem stroke, multiple sclerosis, muscular dystrophies and various others, lead to loss of muscular control (Dornhege et al., 2007; Graimann et al., 2010; Wolpaw et al., 2012). In late stage of such diseases, people are affected by the *complete locked-in* syndrome, which means they have lost any voluntary muscular control and are therefore unable to communicate. Their often completely functioning minds are trapped in paralysed bodies.

A brain-computer interface (BCI) provides a non-muscular communication system which therefore gives people new opportunities to communicate. It detects the user's intents by measuring brain signals, extracts relevant features and directly converts them into control signals. So, it is possible to interact with the environment without using any muscular activity.

BCI paradigms differ based for instance on the measurement technique, the desired application and the user's condition. BCIs can be build on EEG, MEG, hemodynamic measurement techniques such as fNIRS and also on a combination of multiple measurement techniques (Fazli et al., 2012; Von Lühmann et al., 2016; Shin et al., 2016).

There are two main types of BCI paradigms. The first is based on event-related potentials which cause rapid changes in the time domain (i.e. negative or positive peaks) usually triggered by an external stimuli which can be visual (Farwell et al., 1988), auditory (Nijboer et al., 2008; Schreuder et al., 2011) or tactile (Brouwer et al., 2010; Ortner et al., 2014). Another paradigm is based on changes in neural oscillations, e.g. over the sensorimotor cortex where ERS/ERD effects can be observed during imagined or conducted motor control (Pfurtscheller et al., 1999). Throughout this thesis we will focus on the latter, especially on motor imagery-based BCIs.

Motor imagery can be described as the imagination of a movement or the preparation to execute a movement without actually executing it. Imagination and execution show significant changes in e.g. the EEG, especially in mu rhythms. Since this does not require any muscular activity, motor imagery can be utilized for non-muscular communication, thus BCI. Power in the mu and beta frequency bands usually decreases during the preparation or execution of a movement (ERD) over the motor cortex which also means that the amplitude attenuates (Pfurtscheller, 1992). After the movement, power in both bands increases again (ERS). The same effects also appear during imagination of movements (Wolpaw et al., 2002).

## 2.4 The General Linear Model

Noninvasive measuring techniques such as EEG and MEG are only able to measure a superposition on a two-dimensional surface of three-dimensionally distributed brain sources. The general linear model is commonly used to explain generation of this two-dimensional EEG/MEG signal (Müller et al., 2003; Parra et al., 2005; Haufe et al.,

2014b).

### 2.4.1 Forward Model

The forward model assumes that EEG/MEG recordings can be decomposed into $K$ components. Each component contains a fixed spatial activation pattern that linearly projects respective sources in the brain to the surface recording.

Here, $\mathbf{x}(t) \in \mathbb{R}^d$ is the d-dimensional recorded time series at time point $t$, $\mathbf{a}_i \in \mathbb{R}^d$ the i-th spatial activation pattern, $s_i(t) \in \mathbb{R}$ the corresponding i-th brain source activity and $\mathbf{n}(t) \in \mathbb{R}^d$ the activity not explained by the activation pattern and therefore considered as noise.

$$\begin{aligned} \mathbf{x}(t) &= \sum_{i=1}^{K} \mathbf{a}_i \cdot s_i(t) + \mathbf{n}(t) \\ &= \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \end{aligned} \tag{2.1}$$

The activation matrix $\mathbf{A} \in \mathbb{R}^{d \times K}$ contains all $K$ activation patterns $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K]$ and $\mathbf{s}(t) = [s_1(t), \ldots, s_K(t)]^\top \in \mathbb{R}^K$ contains $K$ sources at time point $t$.

The actual number of brain sources here is unknown, we can assume that its number is many magnitudes higher than the number of sensors $d$ used to record the signal. This means multiple brain sources are captured in $\mathbf{s}_i$ and each entry in $\mathbf{A}$ contains an electrode-source pair for multiple brain sources.

### 2.4.2 Backward Model

After introducing the forward model in which we map brain sources to the recorded EEG/MEG signal, we also want to model how to recover the components from the recorded signal. We therefore define the inversion of the forward model as

$$\hat{\mathbf{s}}(t) = \mathbf{W}^\top \cdot \mathbf{x}(t) \tag{2.2}$$

and call this the backward model where each column of $\mathbf{W}$ defines one spatial filter and thus the counterpart to the spatial patterns in $\mathbf{A}$. This inversion is in general not unique and therefore further assumptions about $\mathbf{W}$ need to be made before solving the backward model. In the following sections we will discuss multiple algorithms that, with different assumptions, solve the backward model in individual ways.

It needs to be noted that the information contained in $\mathbf{W}$ is not necessarily neurophyisiologically interpretable. We therefore need to apply the filter-pattern transformation to estimate the activation patterns that carry interpretable information (Haufe et al., 2014b).

$$\hat{\mathbf{A}} = \mathbf{C}_x \mathbf{W} (\mathbf{W}^\top \mathbf{C}_x \mathbf{W})^{-1} \tag{2.3}$$

Here, $\mathbf{C}_x$ denotes the covariance matrix of the recorded signal $\mathbf{x}$.

## 2.5 Spatial Filtering

On the basis of the backward model as stated in (2.2), various assumptions can be formulated to compute spatial filters $\mathbf{w}$. In the following, we will revise three popular spatial filtering methods that will be used within this thesis. The methods discussed in this thesis can be formulated either as a Rayleigh coefficient

$$\max_{\mathbf{w}} \frac{\mathbf{w}^\top \mathbf{A}\mathbf{w}}{\mathbf{w}^\top \mathbf{B}\mathbf{w}} \tag{2.4}$$

or as

$$\max_{\mathbf{w}} \mathbf{w}^\top \mathbf{A}\mathbf{w} \qquad \text{subject to } \mathbf{w}^\top \mathbf{B}\mathbf{w} = C \tag{2.5}$$

where $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ are matrices, $C \in \mathbb{R}$ is a constant and $\mathbf{w} \in \mathbb{R}^d$ denotes one spatial filter.

Both can be reformulated as a generalized eigenvalue problem

$$\mathbf{AW} = \mathbf{DBW}$$

with $\mathbf{W} \in \mathbb{R}^{d\times d}$ where the columns of $\mathbf{W}$ contain the generalized eigenvectors of $\mathbf{A}$ and $\mathbf{B}$. $\mathbf{D}$ is a diagonal matrix of eigenvalues. The eigenvector $\mathbf{w}_* \in \mathbf{W}$ corresponding to the largest eigenvalue $\lambda_* \in \mathbf{D}$ solves the optimization problem in (2.5). Eigenvalue problems are well investigated and can be solved with standard linear algebra.

The formulation of $\mathbf{A}$ and $\mathbf{B}$ is what defines the individual spatial filtering method as we will see in the following sections.

### 2.5.1 Common Spatial Patterns

Common Spatial Patterns (CSP) is a well established spatial filtering method commonly used for feature extraction in motor imagery-based BCIs (Ramoser et al., 2000; Blankertz et al., 2008b). It aims to detect synchronisation and desynchronisation processes (see Section 2.2) and computes discriminative spatial filters by maximizing the variance of one class (e.g. left hand imagination) while minimizing the variance of the other class (e.g. right hand imagination), which can be formulated as

$$\max_{\mathbf{w}} \frac{\mathbf{w}^\top(\mathbf{C}_1 - \mathbf{C}_2)\mathbf{w}}{\mathbf{w}^\top(\mathbf{C}_1 + \mathbf{C}_2)\mathbf{w}} \tag{2.6}$$

where $\mathbf{C}_1$ and $\mathbf{C}_2 \in \mathbb{R}^{D\times D}$ are the average estimated covariance matrices of class 1 (e.g. left) and class 2 (e.g. right), $\mathbf{w} \in \mathbb{R}^{D\times 1}$ is the spatial filter.

Calculating its derivative with respect to $\mathbf{w}$ and setting it to zero results in a generalized

eigenvalue problem

$$(\mathbf{C}_1 - \mathbf{C}_2)\mathbf{W} = \lambda(\mathbf{C}_1 + \mathbf{C}_2)\mathbf{W}.$$

The eigenvectors corresponding to the largest and smallest eigenvalue maximize the variance in class 1 and class 2 respectively. Usually up to three spatial filters per class are chosen for classification.

The covariance matrices $\mathbf{C}_1$ and $\mathbf{C}_2$ are usually estimated by a factor of $\mathbf{X}\mathbf{X}^\top$ where $\mathbf{X} \in \mathbb{R}^{d \times N}$ contains the d-dimensional centered data points. This estimation is not very robust with respect to outliers: a single trial with very high variance can have a major influence on the covariance matrix and thus also on the spatial filters. We therefore will use an alternative procedure to sort the eigenvalues as proposed in Blankertz et al. (2008b).

For each eigenvector/spatial filter $\mathbf{w}_i$, a score is computed based on the median of $\mathbf{w}_i$ over all trials relative to the sum of medians in both classes

$$\begin{aligned} \text{score}(\mathbf{w}_i) &= \frac{m_i^{c_i}}{m_i^{c_i} + m_i^{\bar{c}_i}} \\ \text{with } m_i^{c_i} &= \text{median}_{j \in \mathcal{I}_c}(\mathbf{w}_i \mathbf{X}_j \mathbf{X}_j^\top \mathbf{w}_i), \quad c_i \in \{1, 2\}. \end{aligned} \tag{2.7}$$

The eigenvectors are then sorted according to this score and the top k eigenvectors, largest and smallest, are chosen as features.

### 2.5.2 Spatio-Spectral Decomposition

Neural oscillations, as discussed in Section 2.2, have been linked to various cognitive processes which makes it important to extract related signals in order to investigate those oscillations. As oscillations originate from many sources of neural activity and can also overlap with a potentially much stronger background noise, it is difficult to identify distinct sources of oscillatory activity in the raw EEG/MEG.

Spatio-Spectral Decomposition (SSD, Nikulin et al., 2011) separates oscillatory signals from multi-channel recordings by maximizing the power in a frequency band of interest while simultaneously minimizing the power in neighbouring frequency bands. This method is based on the assumption that noise sources cause signals within a broader frequency range, often modelled as $1/f^\alpha$ where $f$ denotes the corresponding frequency.

This leads to the following optimization function

$$\max_{\mathbf{w}} \frac{\mathbf{w}^\top \mathbf{C} \mathbf{w}}{\mathbf{w}^\top \mathbf{C}_\text{N} \mathbf{w}}. \tag{2.8}$$

Here, $\mathbf{C}$ denotes the covariance matrix of the signal band-pass filtered in the frequency band of interest, whereas $\mathbf{C}_\text{N}$ denotes the sum of the covariance matrices of the signal band-pass filtered in the two neighbouring frequency bands. Spatial filters are displayed

as $\mathbf{w}$.

This optimization problem can be formulated as a generalized eigenvalue problem:

$$\mathbf{CW} = \mathbf{DC}_{\mathrm{N}}\mathbf{W}.$$

### 2.5.3 Source Power Co-Modulation

Source Power Co-Modulation (SPoC$_\lambda$, Dähne et al., 2014) computes a set of spatial filters $\mathbf{W}$ which directly optimize the co-modulation between a target variable $z$ and the power time course of a spatially filtered signal $\Phi$ over epochs $e$. Epochs are chosen such that the target variable only has a single value per epoch. Please note that throughout this thesis we refer to SPoC$_\lambda$ as SPoC.

This can then be expressed as

$$\begin{aligned} \max_{\mathbf{w}} \mathrm{Cov}(\Phi(e), z(e)) &= \mathbf{w}^\top \mathbf{C}_z \mathbf{w} && \text{with } \mathbf{C}_z := \langle \mathbf{C}(e) z(e) \rangle \\ \text{s.t. } \mathrm{Var}(\mathbf{w}^\top x(t)) &= \mathbf{w}^\top \mathbf{C}\mathbf{w} \stackrel{!}{=} 1 && \text{with } \mathbf{C} := \langle \mathbf{C}(e) \rangle \end{aligned} \tag{2.9}$$

where $\langle\ \rangle$ denotes the mean over epochs $e$ and $\mathbf{C}(e)$ represents the covariance matrix of the e-th epoch.

This can be reformulated as a generalized eigenvalue problem

$$\mathbf{C}_z\mathbf{W} = \mathbf{DCW}.$$

where the eigenvalue $\lambda \in \mathbf{D}$ can be interpreted as the covariance between $\Phi$ and $z$.

## 2.6 Linear Classification

Spatial filters are often used in motor imagery-based BCIs to extract relevant features and separate left from right hand trials with a linear classifier such as Fisher's Linear Discriminant Analysis.

Binary linear classifiers in general are defined by

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} - \beta = 0 \tag{2.10}$$

where $\mathbf{x} \in \mathbb{R}^d$ denotes the input feature, $\mathbf{w} \in \mathbb{R}^d$ defines the direction orthogonal to the separating hyperplane and $\beta$ the offset. The separating hyperplane is formulated as

$$\mathbf{w}^\top \mathbf{x} + \beta \stackrel{!}{=} 0. \tag{2.11}$$

A label can then be assigned to a new data point $\mathbf{x}^*$ depending on $\mathrm{sign}(f(x^*))$, i.e. on which side of the hyperplane $\mathbf{x}^*$ is located.

The formulation of $\mathbf{w}$ and $\beta$ defines the individual algorithm.

### 2.6.1 Fisher's Linear Discriminant Analysis

The Fisher criterion (Fisher, 1936) aims to maximize the variance between two classes while simultaneously minimizing the variance within both classes which defines $\mathbf{w}$ and $\beta$ as

$$\mathbf{w} = \hat{\mathbf{C}}^{-1}(\mu_2 - \mu_1) \tag{2.12}$$

$$\beta = \frac{1}{2}\mathbf{w}^\top(\mu_1 + \mu_2) \tag{2.13}$$

where $\hat{\mathbf{C}} = 0.5(\hat{\mathbf{C}}_1 + \hat{\mathbf{C}}_2)$ defines the pooled covariance matrix of both classes and $\mu_c$ denotes the respective class mean.

### 2.6.2 Regularized LDA

Fisher's Linear Discriminant Analysis which we will here just call Linear Discriminant Analysis (LDA) is based on three assumptions

1. Input features of both classes are Gaussian distributed
2. Both classes follow a Gaussian distribution with the same covariance matrix
3. True class distributions and thus mean and covariance matrix are known

The latter can never be fulfilled in real-world applications whereas it is commonly assumed that the other two assumptions generally hold for EEG/MEG data.

It is thus necessary to estimate mean and covariance matrix. The mean is usually estimated by an average of all data points $\mathbf{x}_i \in \mathbb{R}^d$

$$\hat{\mu}_c = \frac{1}{N_c}\sum_{i \in \mathcal{C}_c} \mathbf{x}_i. \tag{2.14}$$

The standard unbiased estimator for the covariance matrix is

$$\hat{\mathbf{C}}_c = \frac{1}{N_c - 1}\sum_{i=1}^{N_c}(\mathbf{x}_i - \hat{\mu}_c)(\mathbf{x}_i - \hat{\mu}_c)^\top. \tag{2.15}$$

This can become imprecise especially for high dimensional input features where only a few samples are available as the number of unknown parameters is quadratic in the number of feature dimension. This leads to systematic errors as large eigenvalues of $\hat{\mathbf{C}}$ are estimated too large and small eigenvalues are estimated too small (Blankertz et al., 2011).

Shrinking the covariance matrix towards a more spherical matrix compensates for this

and is defined as

$$\tilde{\mathbf{C}}(\gamma) = (1-\gamma)\hat{\mathbf{C}} + \gamma\nu\mathbf{I} \tag{2.16}$$

where the shrinkage parameter $\gamma \in [0,1]$ controls the degree of shrinkage (Friedman, 1989). For $\gamma = 0$, $\tilde{\mathbf{C}}$ corresponds to $\hat{\mathbf{C}}$, so no shrinkage is applied and $\gamma = 1$ results in a perfectly spherical covariance matrix scaled by the average eigenvalue $\nu$ of $\hat{\mathbf{C}}$.

LDA combined with a shrinked covariance matrix is a form of Regularized LDA (RLDA). The shrinkage parameter $\gamma$ can either be determined via cross-validation or computed analytically as proposed in Schäfer et al. (2005)

$$\begin{aligned} \gamma^* &= \frac{n}{n-1}\frac{\sum_{i,j=1}^{d}\mathrm{Var}_k(z_{ij}(k))}{\sum_{i\neq j}s_{ij}^2 + \sum_i (s_{ii}-\nu)^2} \\ \text{with} \quad z_{ij}(k) &= ((x_k)_i - (\hat{\mu})_i)((x_k)_j - (\hat{\mu})_j) \end{aligned} \tag{2.17}$$

where $(x_k)_i, (\hat{\mu})_i$ denote the i-th entry of data point $\mathbf{x}_k \in \mathbb{R}^d$ and $\hat{\mu} \in \mathbb{R}^d$ respectively. The variable $s_{ij}$ is the entry in the i-th row and j-th column of $\hat{\mathbf{C}}$.

This analytic solution does not always lead to the best classification accuracy but it is easier to implement and computationally cheaper than cross-validating $\gamma$ (Blankertz et al., 2011). This is particularly useful when applied in online systems where a classifier is retrained after each trial.

## 2.7 Limitations and Challenges

As brain signals are highly individual and sensitive, it can be challenging to build a reliable signal processing pipeline that is applicable to multiple participants. Throughout this thesis we will come across some of the key challenges in computational EEG/MEG analysis which we will investigate in-depth and propose various solutions to overcome them.

**Challenge 1: Nonstationarities due to artifacts**
Signals recorded from the brain are often contaminated by noise that can originate from outside and inside the human body and therefore need to be dealt with accordingly. We distinguish between non-physiological and physiological artifacts. Non-physiological artifacts originate from outside the body such as 50/60 Hz power line noise or changes in electrode impedances. They can mostly be mitigated by appropriate filtering and monitoring of the impedances throughout the recording. Physiological artifacts are much harder to deal with as they originate from the human body such as rhythmic interference by heart beats or respiration but especially ocular and muscular artifacts with their very high amplitudes can significantly disturb signals of the underlying cognitive process (Fatourechi et al., 2007).

Various artifact rejection methods have been proposed in the past. Those methods aim to identify the signal of interest and remove the contaminated noisy parts in EEG

(Winkler et al., 2014; Frølich et al., 2015b; Winkler et al., 2015), MEG (Mantini et al., 2008) and fNIRS (Von Lühmann et al., 2019). Other approaches robustify methods in a way that limits the influence of artifacts (Samek et al., 2012; Samek et al., 2013; Brandl et al., 2015b). We will investigate how artifacts influence classification accuracies in a motor imagery-based BCI and how different algorithmic approaches can overcome the influence of those artifacts to reach reliable classification.

**Challenge 2: Nonstationarities due to dynamic shifts**
Brain signals typically change over time, even if a person is following a specific paradigm you might detect dynamic shifts within the same session. The reason for this might be that the participant adapts to the paradigm, gets tired or bored and thus less focused which also causes a power increase or shift in the alpha band (8-12 Hz) (Benwell et al., 2019). If then a classifier is trained on a calibration phase which usually takes place in the beginning of the session, it might become difficult to classify data towards the end of the experiment.

Constantly adapting or recalibrating the classifier helps to build a classifier that can handle dynamic shifts (Sugiyama et al., 2007; Vidaurre et al., 2010; Vidovic et al., 2015). Other approaches focus on invariant features (Blankertz et al., 2008a) or adapt the data space to consider shifts in data distribution (Arvaneh et al., 2013a). All those approaches are based on a classification problem where the ground truth can be included in finding invariant features or building an invariant classifier. In unsupervised problems dynamic shifts occur nonetheless such that one needs to take them into account and adapt for instance representations accordingly.

**Challenge 3: Artificial lab environments**
Most recordings take place in very controlled lab environments: participants sit still, are asked not to move and distractions are reduced as much as possible. This makes it easier to calibrate for instance a BCI system as signals are less noisy but also makes it almost impossible to apply the BCI in everyday life where people might move around and are exposed to multiple sources of distraction or noise. Another problem is the limitation that comes with applying a recording device. If one wants to record brain signals in daily life it is important for hardware to be mobile and also affordable.

Recently, various studies have been recorded in classrooms, cars or realistic tugboat simulators (Ko et al., 2017; Brouwer et al., 2017; Blankertz et al., 2010b; Miklody et al., 2017). Mobile BCI hardware even allow participants to move freely during the recording (Lotte et al., 2009; Castermans et al., 2011; De Vos et al., 2014; Wriessnegger et al., 2017; Von Lühmann et al., 2019). Though we argue that it is also important to systematically investigate everyday life situations to bring BCIs closer to the daily life.

# 3

# Motor Imagery under Distraction: An Evaluation Study

## 3.1 Motivation

Over the years, various improvements in BCI research have been presented. Integrating machine learning algorithms has caused substantial reduction in calibration time (Blankertz et al., 2007; Krauledat et al., 2008; Fazli et al., 2009; Lemm et al., 2011) which has crucially enhanced the usability of BCIs. Also, novel approaches in robust feature extraction (Samek et al., 2012; Arvaneh et al., 2013b; Samek et al., 2013; Kawanabe et al., 2014; Brandl et al., 2015b), artifact detection (Fatourechi et al., 2007; Schlögl et al., 2007; Winkler et al., 2014) and adaptive methods (Sykacek et al., 2004; Shenoy et al., 2006; Vidaurre et al., 2011) have led to more reliable BCIs.

Those approaches already work well in controlled lab environments which are highly artificial and significantly differ from everyday life situations, where people have to handle various visual, auditory or other cognitive distractions. In order to fulfil its main purpose, to provide a non-muscular communication pathway, BCI research has recently begun to go beyond lab environments.

As summarized in the previous chapter, multiple studies have been carried out recently in dynamic settings. However, a systematic investigation of dynamic changes caused by internal or external influences is crucial to build a reliable BCI for the everyday life. For instance, power changes in relevant frequency bands in the EEG (internal) could heavily influence classification outcome especially in motor imagery-based BCIs as they rely on event-related (de)synchronisations (see Section 2.3). An increase in movement artifacts (external) might also dominate a BCI system when calibrated on artifact-free data. Those are only two typical examples that need to be considered when building reliable BCIs.

So far there has not been a BCI study where such an investigation takes place. We have therefore recorded a motor imagery-based BCI study (N=16) under 5 types of distractions that mimic out-of-lab environments and a control task where no distraction was added. The secondary tasks include watching a flickering video, searching the room for a specific number, listening to news, closing the eyes and vibro-tactile stimulation. We consider this study as a first step, necessary in getting closer to everyday life BCIs.

Throughout this chapter, we investigate standard machine learning techniques, commonly used in BCI research, in this pseudo-realistic scenarios. We identify three major

challenges leading to poor performance in those out-of-lab scenarios

1. Artifact contamination
2. Feature shifting
3. High cognitive workload

and discuss several new approaches which overcome those challenges.

This chapter is structured as follows: In Section 3.2, we will explain the dataset and show results of a baseline analysis. In Section 3.3 we will show the typical analysis pipeline for training data recorded in a controlled environment and investigate possible reasons for a low classification accuracy before we propose new strategies in Section 3.4. The chapter is concluded with a discussion in Section 3.5. This chapter follows our own work as published in Brandl et al. (2016b) and Brandl et al. (2020d).

## 3.2 Methods

### 3.2.1 Participants

16 participants (6 female, average age 26.3 years ± 1.9 years) volunteered to participate in this study. Three volunteers had previously participated in another BCI experiment. All instructions were given in German requiring basic language skills. Volunteers were reimbursed for their participation in the study except for 3 employees of the TU Berlin Machine Learning Group. All participants were instructed on the experimental procedures prior to signing an informed consent. This study was conducted according to the declaration of Helsinki and was approved by the Ethics Committee of the Charite-Universitätsmedizin Berlin (approval number: EA4/012/12).

### 3.2.2 Data Acquisition

EEG signals were recorded with a Fast'n Easy Cap (EasyCap GmbH) with 63 wet Ag/AgCl electrodes placed at symmetrical positions according to the international 10–10 system (Chatrian et al., 1985) and referenced to the nose. We used two 32-channel amplifiers (BrainAmp, BrainProducts) to amplify the signals, which were sampled at 1000 Hz. Data was recorded in the period of 15 April - 18 July 2014 at TU Berlin and raw data without any preprocessing was made publicly available[1].

### 3.2.3 Experimental Setup

During the experiment, the participants were sitting in a comfortable armchair at a distance of 1 m in front of a 24" computer screen. Auditory instructions were given via headphones.

[1] https://depositonce.tu-berlin.de/handle/11303/10934.2

Each experimental session lasted about three hours including preparation and about 90 min of signal recording. Before the main experiment, we recorded 8 trials in which participants had to alternately keep their eyes open or closed for 15 s.

The main experiment was divided into 7 runs à 10 min with 72 trials per run. One trial lasted 4.5 s and was defined by one motor imagery task with an additional secondary task except for the *clean* condition and the first run which served as a calibration phase. The subsequent runs included 3 blocks à 4 trials (2 left and 2 right) of each secondary tasks (72 trials per run) with auditory feedback given after each trial. The blocks were presented in a random order to minimize sequence effects.

### Primary Task

At the beginning of each trial, instructions for left or right hand motor imagination were given over headphones (*links* and *rechts* as the instructions were in German). This was the primary task in this study. At the end of the trial the participant received a *stop* command followed by a break of 2.5 s, after which the next trial started.

Participants were asked to choose one haptic hand movement. Several strategies for motor imagery were presented to the participants to choose from. The majority chose to imagine squeezing a soft ball - other strategies involved opening a water tap, piano playing or using a salt shaker.

Auditory online feedback was given in the 6 runs after the calibration to keep the motivation up. The online feedback was trained on the calibration data and based on Laplacian filters of the C3 and C4 electrodes (McFarland et al., 1997) and regularized linear discriminant analysis (RLDA; Friedman, 1989). For this, EEG data was downsampled to 100 Hz, Laplacian filters of C3 and C4 were calculated and the data was band-pass filtered in the ranges 9-13 Hz and 18-26 Hz with a Butterworth filter of order 5. Data was then cut into epochs of 750-3500 ms and an RLDA classifier was trained on the logarithm of variances as features. During the feedback phase, EEG data was downsampled and band-pass filtered as before, projected on the Laplacian filters and the trained classifier applied on the log-variance features. Furthermore, we applied pooled-mean adaptation to continue training the classifier during the feedback phase (Vidaurre et al., 2010). Classification averaged across all participants reached an accuracy of $57,05\%$. Auditory feedback was given after the *stop* command as *decision left* (*Entscheidung links*) or *decision right* (*Entscheidung rechts*) during the 2.5 seconds break. Online classification was performed with the BBCI toolbox in MATLAB[2].

### Secondary Tasks

We simulated a pseudo-realistic environment by adding 6 secondary tasks on top of the primary motor imagery task to the experimental setup (see Table 3.1 for a summary). They were selected to cover different types of distractions in an out-of-lab scenario.

[2] https://github.com/bbci/bbci_public

| | **Task** | **Expected Artifact** |
|---|---|---|
| | Motor imagery + ... | |
| *Clean* | ... no distraction | None (control task) |
| *Eyes-closed* | ... closed eyes | Increased $\alpha$-rhythm |
| *News* | ... listening to auditory news sequences | Activation of auditory cortex |
| *Numbers* | ... searching the room for a particular letter-number combination on the surrounding walls (left, front, right) | Distraction + muscular artifacts |
| *Flicker* | ... watching a video with a flicker in gray shades at a frequency of 10 Hz | SSVEP |
| *Stimulation* | ... vibtrotactile stimulation on both forearms with carrier frequencies of 50 and 100 Hz, modulated at 9, 10 and 11 Hz | SSVSEP |

Table 3.1: Description of secondary distraction tasks.

1. *Clean*
   This condition served as a control task where no additional distraction was added.

2. *Eyes-Closed*
   Participants were asked to close their eyes before the motor imagery trial started and to keep them closed until the trial finished. Here, we expected a power increase in the alpha band (8-12 Hz) due to the closed eyes to overlap with the motor task related $\mu$-rhythm (8-13 Hz). This task was also the primary reason for providing all instructions and feedback auditorily instead of visually.

3. *News*
   Short sequences of a public newscast (*Tagesschau*) were played over the headphones with current news (January/February 2014) and news from 1994. Each sequence was only played once in each experiment. We expected the participants to be cognitively distracted and the auditory cortex to be activated during the motor imagery task which might influence the motor imagery performance.
   During the experiment, we did not assess active listening of the participants.

4. *Numbers*
   For this task, 26 sheets of paper with a randomly mixed letter-number combination were set up on the wall in front of the participants and also on the left and right side of the room. This implies that participants needed to turn their head in order to see the sheets. For each trial a new window appeared on the screen asking the participants to search the room for a particular letter to match with a stated number and to read it out loud. Each combination was shown 2-3 times to

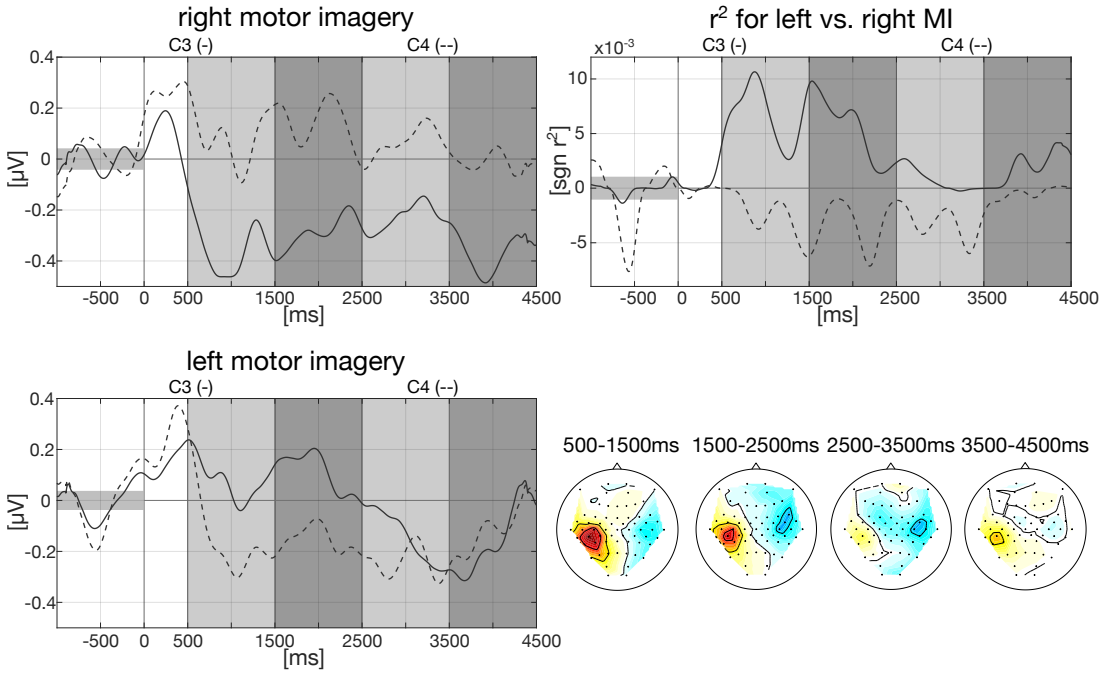


Figure 3.1: Baseline analysis for calibration data. *Left:* Group-average envelopes of electrodes C3 (solid) and C4 (dashed) for right and left motor imagery trials. *Right:* Group-average signed $r^2$ values evolving over time within trials as times series for C3 and C4 (upper) and scalp patterns for 4 different time intervals (lower).

all participants. We counted how often the letters were found. Out of 72 trials, 59.7 combinations were successfully found on average. This task was expected to cause both a high cognitive distraction and additional muscular artifacts.

5. *Flicker*
   A flickering stimulus with alternating gray shades at a frequency of 10 Hz was presented on the screen. We included this task to analyse the influence of the *steady state visually evoked potential* (SSVEP; Morgan et al., 1996).

6. *Stimulation*
   We placed two coin vibration motors with a diameter of 3 cm on the insides of both forearms, one over each wrist and the other just below the elbows. To investigate the interference of *steady state vibration somatosensory evoked potential* (SSVSEP; Tobimatsu et al., 1999; Brouwer et al., 2010) on the motor imagery task, vibro-tactile stimulation was carried out with carrier frequencies of 50 and 100 Hz, each modulated at 9, 10 and 11 Hz.

### 3.2.4 Baseline Analysis

We show group-level results of event-related synchronisation and desynchronisation (ERS/ERD, see Figure 3.1) which can be observed during motor imagination and execution (see Section 2.3). Data analysis was also performed with the BBCI toolbox for MATLAB[2].

Data from the calibration session was band-pass filtered in the frequency band of 9-13 Hz with a 3rd order zero-phase Butterworth filter and cut into epochs for each participant individually, starting 1000 ms prior to trial onset until 4500 ms after trial onset. The envelope was then calculated on the group average based on the Hilbert transformation with a moving average window of 200 ms. Baseline correction was applied, i.e. the average EEG amplitude in the interval of 1000 ms prior to trial onset was subtracted. The resulted smoothed envelope is presented in Figure 3.1 for the electrodes C3 and C4. Here, we clearly see desynchronisation effects in C3 for right hand motor imagery and C4 for left hand motor imagery starting around 300 ms after trial onset.

We further calculated signed biserial correlation coefficients ($r^2$) on the smoothed group-average envelope to determine which EEG channels show the most discriminative information for left and right hand motor imagery. Results can be examined in Figure 3.1 where the scalp patterns of both left and right motor cortex carry relevant class information especially in the beginning of the trial which matches findings in the literature (Pfurtscheller, 1992). Above the scalp patterns, we show the time course over an average of all epochs of the $r^2$-values for C3 and C4. Here, we can see that on average 500-2000 ms after trial onset the two channels carry import information to separate right and left motor imagery as indicated by $r^2$.

In Figure 3.2, we show power spectra for all secondary tasks. For each participant, power spectra were averaged across trials and normalized channel-wise. We then extracted the power spectra for the channels O1 and O2, averaged over the two channels and again across participants. Alpha peaks clearly differ for *eyes-closed* and *numbers* compared to *clean*. For the *eyes-closed* task, we see the expected alpha peak in the range of 8-12 Hz (Berger, 1929). For the *numbers* task there is no clear alpha peak visible in the occipital channels which is in line with the expected suppression of the visual alpha rhythm during visual search. Power spectrum for the *flicker* task shows a small sharp peak between 9-11 Hz which is very close to the frequency of the flickering video and another even smaller peak at 20 Hz which represents the second harmonic of the flicker frequency. The *news* and *stimulation* task do not show clear differences compared to *clean*.

We also show spatial distribution for different frequency bands within the alpha band based on the peaks in the power spectrum. For *eyes-closed* and *flicker* we see a clear activation over the occipital and parietal cortex whereas there is no clear pattern visible for the *numbers* task. Again, patterns for the *news* and the *stimulation* task look very similar to the pattern of the *clean* task.

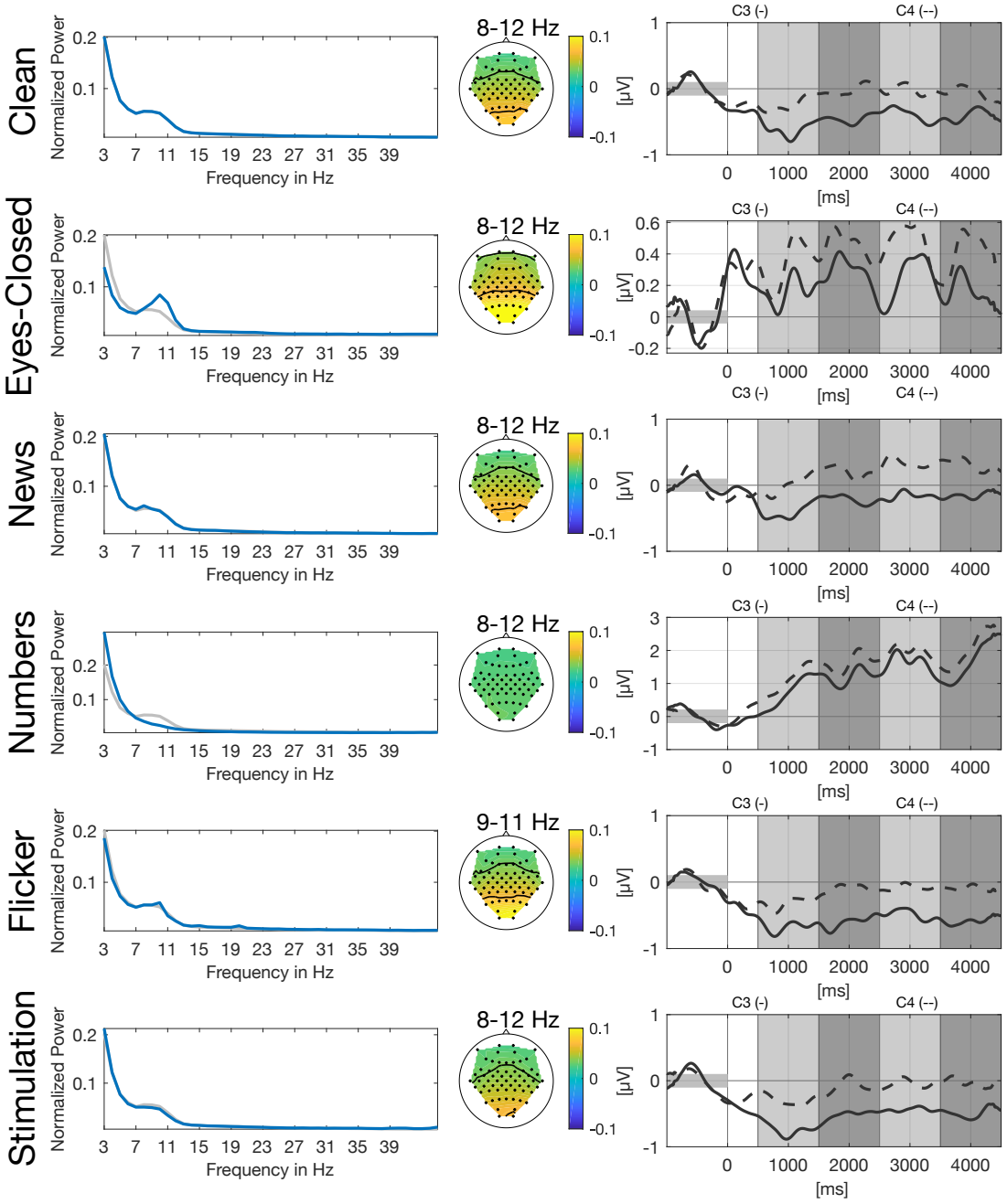


Figure 3.2: Spectral, spatial and temporal information for all secondary tasks. *Left:* Normalized power spectrum averaged across participants and electrodes O1 and O2, in blue for the respective secondary task and in grey for the clean condition. *Center:* Spatial distribution of power spectral values averaged across participants for different frequency bands. *Right:* Group-average envelopes in 9-13 Hz of electrodes C3 (solid) and C4 (dashed) for right hand motor imagery.

Similar to Figure 3.1, we show envelopes of channels C3 and C4 for right hand motor imagery. The modulation of the sensorimotor rhythm is still visible in all conditions as a stronger ERD in C3 compared to C4. However, the effect is obscured by the different artifacts. The disturbances are smallest in the *news, flicker* and the *stimulation* tasks due to the stationary nature of the artifacts. For the *flicker* task we still see a clear difference between both channels, whereas channels are already closer for *eyes-closed* and still even closer for the *numbers* task.

## 3.3 Evaluation of in-lab training

In this section, we present first classification results based on the calibration data. All classification approaches throughout this chapter are based on CSP (see Section 2.5.1) and RLDA (see Section 2.6.2). First, we trained the classifier on the calibration data without secondary tasks and tested it on all remaining data. Translating this to our systematic framework, we trained *in-lab* and tested in the *out-of-lab* setting. Since the resulting classification rates lead to the assumption of poor out-of-lab BCI performance, we further investigated possible reasons for this outcome.

### 3.3.1 Classification based on calibration data

In Table 3.2, average classification rates for all participants are summarized. For this analysis, individual frequency bands between 8 and 30 Hz and time intervals between 250 and 4500 ms after stimulus onset were selected for each participant as described in Blankertz et al. (2008b). Data was then band-pass filtered in the selected frequency band with a 3rd order zero-phase Butterworth filter and cut into epochs. Six CSP filters were extracted, three per class based on the 'ratio-of-median' score as described in Section 2.5.1. The logarithm of the variance of the CSP-filtered signal was then used as features and fed into an RLDA classifier. Overall classification averaged across all participants reached an accuracy of 61.81%.

Average classification accuracies vary much between participant's [between 50% and 94.44%] but also within a participants recording [e.g. *pp10:* 48.61% – 87.5%], so not everyone achieved classification rates significantly higher than chance level. Applying binomial tests ($\alpha = 5\%$) led to thresholds of 54.17% and 61.11% for the overall experiment (432 trials) and for individual secondary task (72 trials), respectively. For classification accuracies above or equal to these thresholds, we assume actual BCI control. The topic of significant BCI control taking into account the number of trials has been addressed in Müller-Putz et al. (2008).

We grouped the 16 participants in three categories. The first category contains the three participants which gained significant BCI control in all secondary tasks. The second category contains 7 participants that reached the overall threshold of at least 54.17%. The last category contains the remaining 6 participants below this threshold which means that we cannot assume any BCI control.

| | Band(Hz) | Ival(ms) | **Overall** | **Clean** | **Eyes** | **News** | **Num** | **Flick** | **Stim** | # |
|---|---|---|---|---|---|---|---|---|---|---|
| **14** | [8 -13.5] | 1046-4196 | 94.44 | 95.83 | **97.22** | 95.83 | 86.11 | 95.83 | 95.83 | |
| **6** | [10.5 -13.5] | 1036-4423 | 87.01 | **95.83** | 86.11 | 83.33 | 72.22 | 94.44 | 90.14 | 1 |
| **5** | [10.5 -14.5] | 251-3324 | 84.03 | 91.67 | 79.17 | 86.11 | 76.39 | **93.06** | 77.78 | |
| **10** | [10.5 -14] | 875-4026 | 71.30 | **87.50** | 80.56 | 77.78 | 48.61 | 75.00 | 58.33 | |
| **16** | [10.5 -14] | 360-3833 | 59.49 | **72.22** | 56.94 | 54.17 | 47.22 | 62.50 | 63.89 | |
| **3** | [10.5 -14.5] | 1288-4798 | 57.87 | **65.28** | 56.94 | 52.78 | 51.39 | 59.72 | 61.11 | |
| **8** | [19.5 -25.5] | 2528-5000 | 56.63 | 48.61 | 63.38 | 55.56 | 58.33 | 50.00 | **63.89** | 2 |
| **7** | [21.5 -26.5] | 788-4320 | 56.48 | 61.11 | 45.83 | **62.50** | 52.78 | 54.17 | **62.50** | |
| **13** | [8 -23] | 282-3238 | 55.79 | **65.28** | 47.22 | 59.72 | 55.56 | 51.39 | 55.56 | |
| **15** | [8 -13.5] | 251-2217 | 55.32 | **58.33** | 56.94 | 55.56 | 54.17 | 54.17 | 52.78 | |
| **4** | [8 -12.5] | 1855-5000 | 53.82 | 58.33 | 50.00 | 50.70 | 52.78 | **62.50** | 48.61 | |
| **9** | [19.5 -24.5] | 2427-4575 | 52.79 | 52.11 | 44.44 | **59.72** | 52.11 | 52.78 | 55.56 | |
| **1** | [14.5 -20.5] | 372-3828 | 52.08 | **58.33** | 52.78 | 48.61 | 47.22 | 55.56 | 50.00 | |
| **11** | [8 -13] | 1506-5000 | 51.39 | 54.17 | 51.39 | 45.83 | 50.00 | **56.94** | 50.00 | 3 |
| **2** | [11.5 -14] | 251-1801 | 50.46 | **54.17** | **54.17** | 43.06 | 47.22 | 52.78 | 51.39 | |
| **12** | [16.5 -19.5] | 2395-5000 | 50.00 | **54.17** | 50.00 | 51.39 | 51.39 | **54.17** | 38.89 | |
| ∅ | | | 61.81 | **67.06** | 60.82 | 61.42 | 56.47 | 64.06 | 61.02 | |

Numbers in right-most column refer to the following categories:
1: BCI control during all distraction tasks (at least 61.11% in each task)
2: BCI control during most distraction tasks (at least 54.17% in overall experiment)
3: no BCI control

Table 3.2: Mean classification accuracies for all secondary tasks and all participants when trained on calibration data. One row represents one participant, the first column shows the participant codes. The second and third columns contain the individually selected frequency band and time interval respectively. Highest (bold) and lowest (red) accuracies are highlighted for each participant. Participants are grouped in 3 groups according to classification accuracy. Accuracies for single tasks (all tasks) are significantly better than chance level if higher or equal than 61.11% (54.17%).

Out of the 16 participants, 6 did not reach the threshold of 61.11% in their best distraction task (*pp15, pp9, pp1, pp11, pp2, pp12*). Especially in the *numbers* task with its dual challenge of contaminated data and increased cognitive load, most participants only reached classification accuracies around 50%. Also for the *eyes-closed* task, only 5 participants reached the threshold of 61.11%. Whereas most users gained their highest classification rates as expected in the *clean* task.

### 3.3.2 Why poor performance?

For most participants we could not reach the significance threshold in at least one task. We therefore need to find out how to improve the overall performance. We found several possible explanations. One is that the distraction tasks influence the

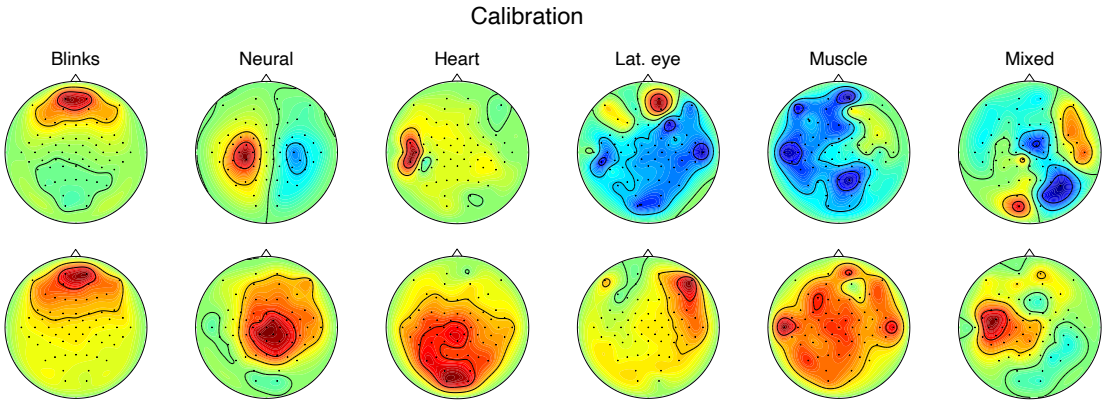


Figure 3.3: Exemplary scalp maps of independent components (IC). Scalp maps were sampled from the calibration data from all participants for each IC class after classification with IC_MARC.

EEG recordings in a way that leads to major feature shifts between calibration data (without secondary tasks) and testing data (with secondary tasks). Some of the tasks might cause too many artifacts. These could contaminate data in a way that makes it impossible to identify actual neural activity. It is also worth considering that some tasks are more cognitively demanding than others. This could increase the difficulty to fully concentrate on the motor imagery task. This leads to three possible explanations for the poor BCI performance which we will tackle in the following sections.

1. artifact contamination
2. major feature shifts between distraction tasks
3. participants are too distracted to focus on the motor imagery task, so the data is not separable

Apart from those three reasons there is also the probability of BCI illiteracy which is the lack of ability to control a BCI. This phenomenon affects an estimated proportion of $15\% - 30\%$ of the participants and has already been discussed in broad variety in the literature (Blankertz et al., 2010a; Grosse-Wentrup et al., 2011; Hammer et al., 2012; Marchesotti et al., 2016). In the *clean* task which represents a classical motor imagery task, 8 out of 16 participants reached the significance threshold of 61.11% which means, there is no significant BCI control for half of the participants. This proportion lies much higher than the $15\% - 30\%$ range and could be explained by the fast change of scenery as the secondary task changed every 4 trials. We will investigate the subject of non-separability of the data in Section 3.3.5.

### 3.3.3 Artifacts

Since different artifact types influence data in different ways, the impact of their contamination and the methods for their removal also differ. We aimed to quantify the

extent of contamination in each secondary task to investigate differences that might explain classification performances in the distinct tasks. Additionally, we investigated whether the removal of groups of artifacts could improve classification. We performed these analyses by decomposing the band-pass filtered and epoched calibration data (see Section 3.3.1) with Independent Component Analysis (ICA) and classifying the resulting independent components as different artifact types. We used the implementation of Extended Infomax in EEGLab (Delorme et al., 2004) to perform the ICA and an automatic classifier of independent components of EEG data, IC_MARC (Frølich et al., 2015a), to classify ICs. IC_MARC assigns a probability to each independent component of representing neural activity, eye blinks, heart beat artifacts, lateral eye movements, or muscle contractions. In addition to these five well defined classes, a class referred to as "mixed" is also included. This class contains artifact types other than those already mentioned and independent components that include several types of activity or noise. We assigned independent components to the class for which the highest probability was predicted by IC_MARC.

Figure 3.3 shows two randomly selected examples of each class from the calibration data from all participants. The samples from the classes blink and neural are good examples of what we would expect in these classes. The top sample from heart is not clearly identifiable, activation in this class can usually be seen on the left temporal/parieto-temporal cortex gradually diminishing towards the right temporal cortex but is not necessarily restricted to only two electrodes. The bottom sample rather looks like occipital neural activity. The bottom sample from lateral eye movements could result from eye movements, activation in the right temporal cortex makes it difficult to confidently decide. The top sample seems to be a mixture of activity resulting from lateral eye movements and other components. Both samples from the muscle class look like a mixture of different components that are not clearly identifiable similar to both samples from the mixed class where this behaviour is expected. For a fuller picture one usually looks not exclusively at topographies as they only carry spatial information but also at temporal and spectral information. Frequency spectra however are here not clearly interpretable as selected frequency bands (see Table 3.2) are very narrow such that power spectra from different ICs look very much alike within participants. This could also be one reason that made it difficult to separate ICs and classify them reliably. We grouped these artifact classes into five groups:

1. Muscular (muscle artifacts).

2. Ocular (blinks and lateral eye movements).

3. Non-neural (blinks, heart beats, lateral eye movements, muscle, and mixed artifacts).

4. Muscular and mixed (muscle and mixed artifacts).

5. Non-mixed artifacts (blinks, heart beats, lateral eye movements, and muscle artifacts).

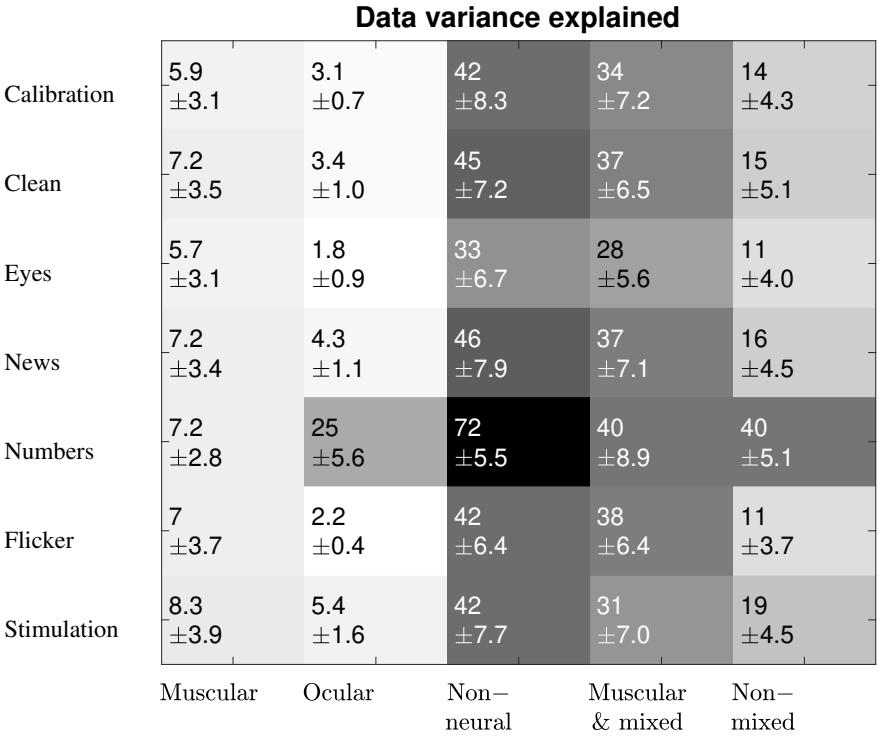


Figure 3.4: Percentage of data variance explained by each artifact group in each condition. The numbers represent the mean over participants plus/minus the standard deviation.

Figure 3.4 shows the percentage of data variance explained by each artifactual group for all secondary tasks. Displayed values represent mean and standard deviation over all participants. The percentage of variance explained by the artifact groups is quite similar for all tasks, except for the *numbers* task. There, the ocular artifacts, non-neural independent components, and non-mixed artifacts explain more than twice as much variance as in the other tasks. We might expect that a BCI system would perform similarly across the secondary tasks whose artifact distributions are similar. If this is the case, one classifier should work well across the *not-numbers* tasks (*clean, eyes-closed, news, flicker, stimulation*). The *numbers* task on the other hand requires a separate classifier. The artifact distribution of unseen data might be informative enough to distinguish between these two cases in order to select the appropriate classifier. This would make the 2-step approach introduced later a suitable method for out-of-lab BCI systems.

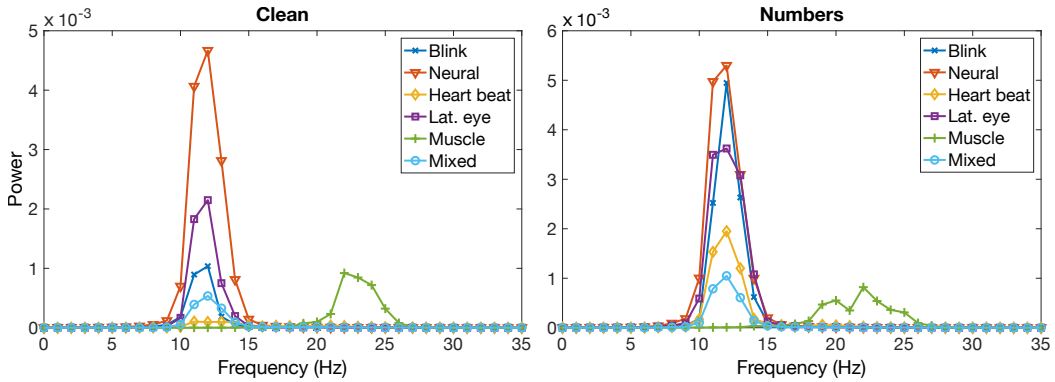


Figure 3.5: Median power spectra of independent components from the six classes (blinks, neural, heart beats, lateral eye, muscle, and mixed) for *clean* and *numbers*.

Figure 3.5 shows the median power of each independent component class as a function of frequency for the independent components from the secondary tasks *clean* and *numbers*. The power spectra were calculated for each epoch individually, after that the median was first taken over epochs for each participant, and then over participants. Since data was individually band-pass filtered during preprocessing, the spectra are flat below 8 Hz and above 26 Hz. It is reassuring that the neural independent components' power peaks at around 9-15 Hz for both tasks since this band contains the motor-imagery related mu-band. The low amplitudes of the blink and lateral eye independent components' spectra relative to neural components in *clean* is not surprising since for these artifacts, activity typically lies in frequency bands lower than 8 Hz. Power for eye blinks usually peaks at 3 Hz and drops off before 10.5 Hz while lateral eye movements exhibit most power at frequencies below 6 Hz (Niedermeyer et al., 2005, p. 1237; Gasser et al., 1985). Frequency information of heart beats mainly lies between 15-32 Hz (Jiang et al., 2007). The power seen at lower frequencies than 15 Hz for heart beats, especially in *numbers*, indicates that independent components classified as heart beat artifacts probably also contain other types of activity. The high peaks for ocular ICs in *numbers* could indicate

that data has been misclassified or ocular ICs have not been clearly separated during the ICA and contain other activity. Muscle artifacts are active at high frequencies, from about 20-300 Hz (Muthukumaraswamy, 2013) which we can see in Figure 3.5. Since mixed artifacts may contain many types of activity, we do not have any expectations for how their power spectrum may look.

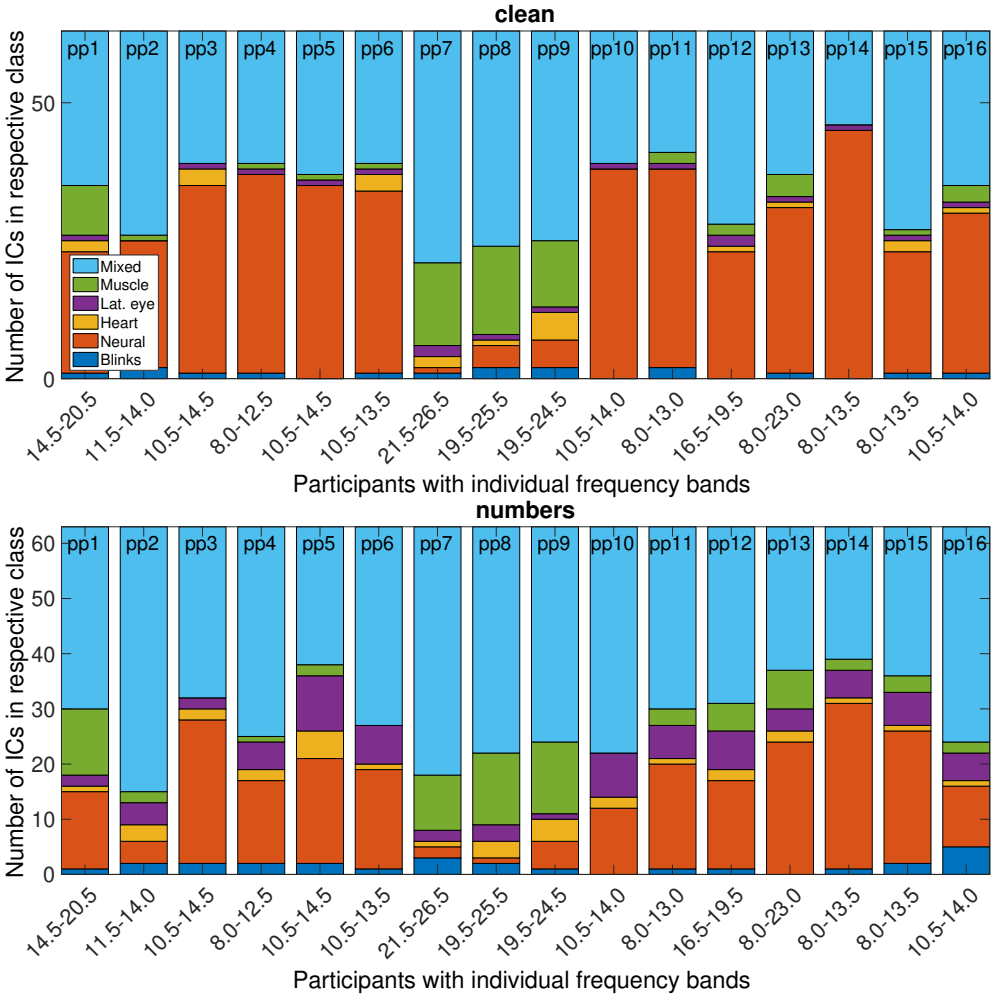


Figure 3.6: Distribution of 63 classified ICs for each participant for the secondary tasks *clean* and *numbers*. Each bar represents one participant. Individual frequency bands from Table 3.2 are displayed below each bar.

To have a closer look at the results of ICA and IC_MARC we will look at the distributions of ICs in both tasks *clean* and *numbers* for all 16 participants, see Figure 3.6. Classification of ICs in *clean* show a high variance in the distribution of different artifact classes, the number of neural components vary between 1 (*pp 7*) and 45 (*pp14*) out of 63. It is not surprising that *pp1* and *pp7-9* have more ICs classified as muscular activity since their individual frequency bands lie in the range of typical muscular activity (see Table 3.2).

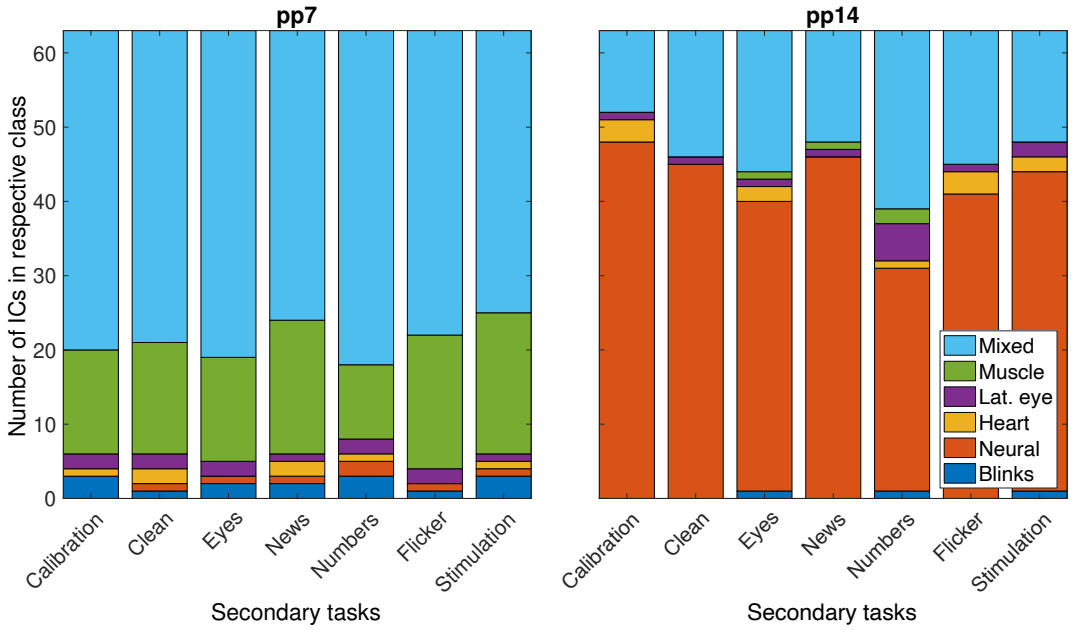


Figure 3.7: Distribution of 63 classified ICs for all secondary tasks and calibration for participants *pp7* and *pp14*. One bar represents one secondary task.

There is also a group of participants (*pp2, pp7-9, pp12, pp15*) with a majority of mixed ICs (32 or more ICs classified as mixed) which indicates that artifacts couldn't be clearly identified as the ICA failed to properly separate the data into independent components. The distribution for *numbers* shows even more mixed ICs and also more ICs classified as lateral eye movements, on the other hand the number of neural ICs has decreased drastically such that all participants show more ICs classified as artifacts than neural activity.

We select *pp7* and *pp14* as two exemplary participants and look at their distributions over classified ICs over all tasks (see Figure 3.7). While *pp7* shows almost no neural ICs and a clear majority of mixed ICs over all tasks, *pp14* contains mostly neural ICs, even 30 (out of 63) in *numbers*. The topographies of the ICs from the calibration data for both participants (see Figures 3.8, 3.9) underline those results. The topographies of *pp7* only show very few ICs with clear neural activity, they mostly seem to be muscular activity and/or a superposition of several components. This indicates that the ICA had difficulties separating EEG data. For *pp14*, separation with ICA seems to have been more successful as neural activity mostly looks like what we would expect: central, occipital and parieto-occipital activity over the motor cortex.

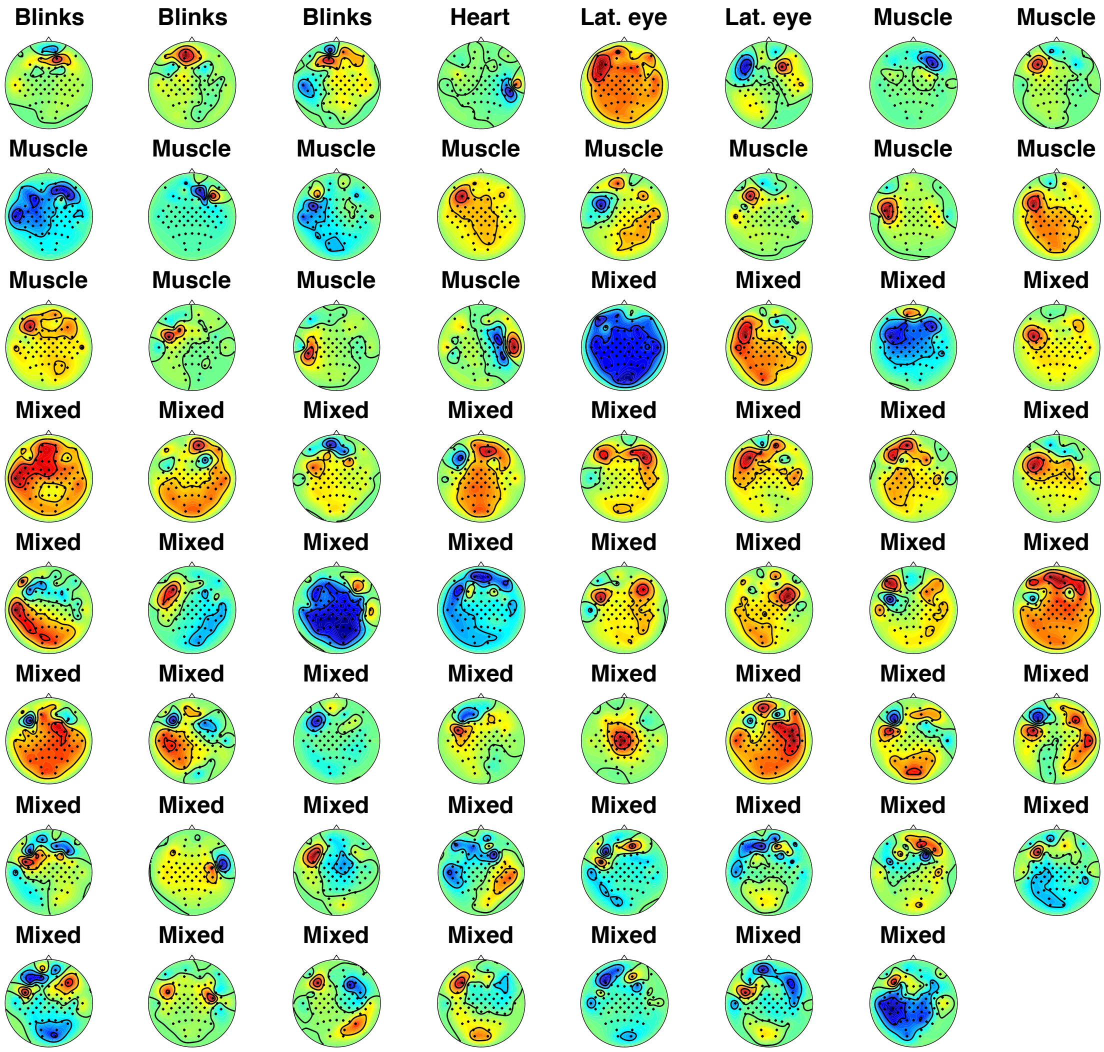


Figure 3.8: Activation patterns from all ICs of the calibration data for participant *pp7* with respective labels displayed above. ICs are ordered according to their label.

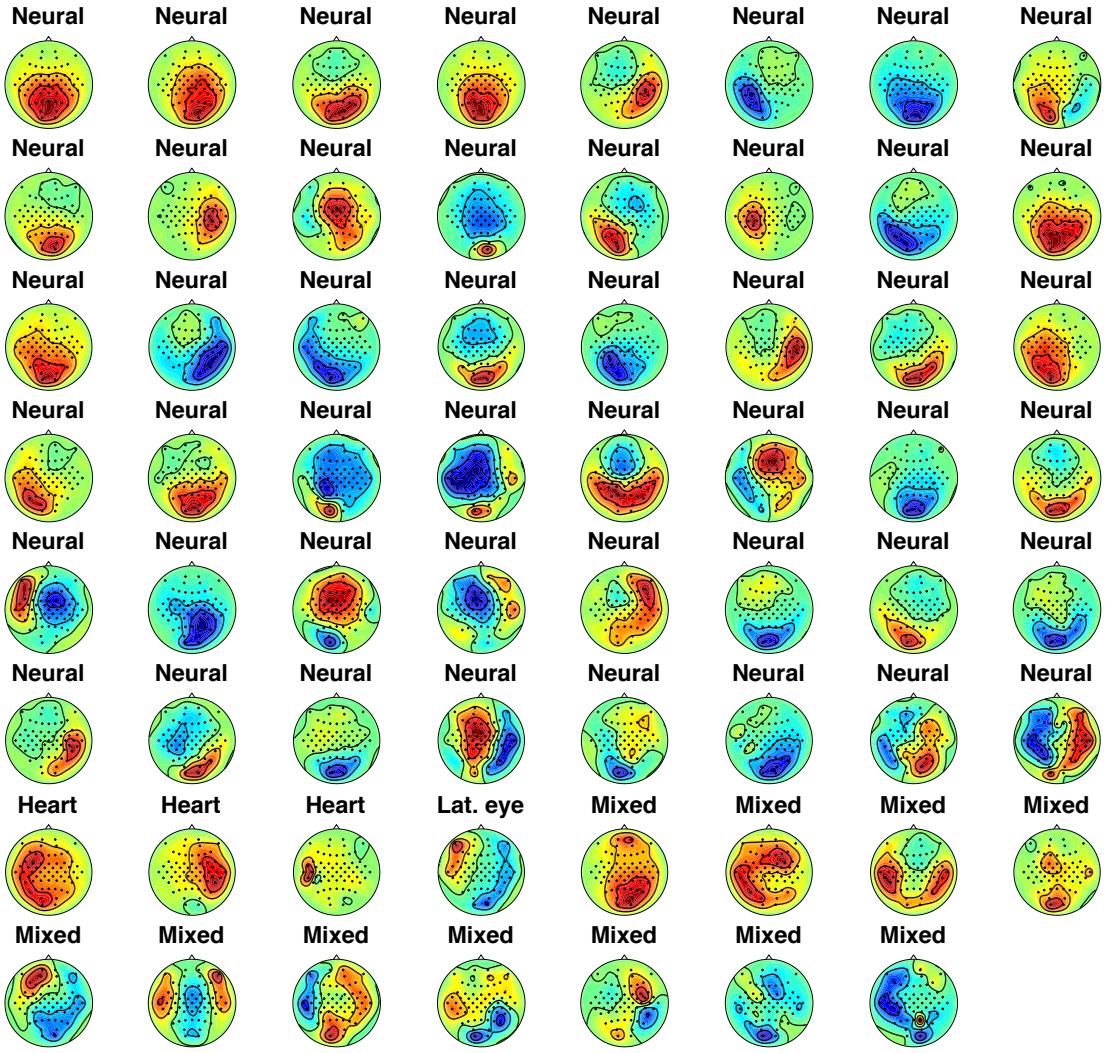


Figure 3.9: Activation patterns from all ICs of the calibration data for participant *pp14* with respective labels displayed above. ICs are ordered according to their label.

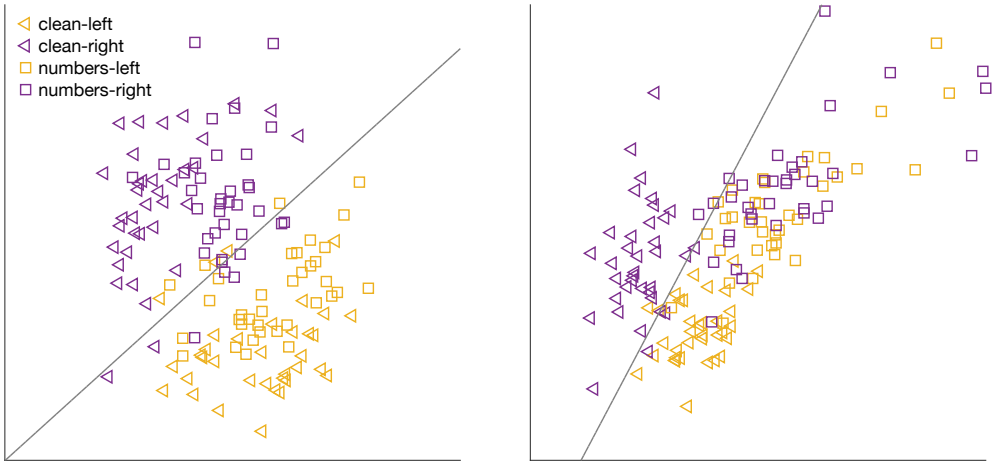


Figure 3.10: Features of participants *pp6 (left)* and *pp16, (right)* of the classifications between left and right hand motor imagery (two best CSP filters) where CSP was only trained on *clean* and tested on *numbers*.

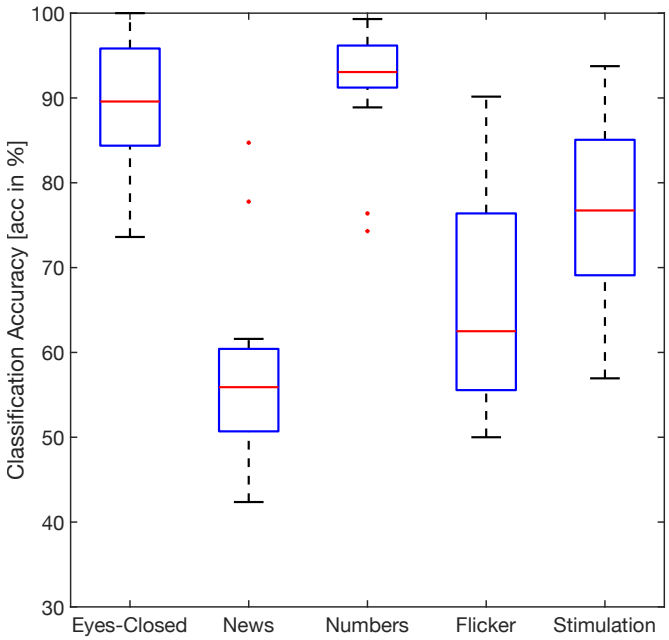


Figure 3.11: Classification accuracies of distraction tasks against *clean* averaged across all 16 participants for *left* and *right* combined.

### 3.3.4 Feature Shifts

In Figure 3.10, we plotted train and test features for the *numbers* task (2 best CSP filters) for participants *pp6* and *pp16* respectively with corresponding decision boundaries. Please note that this does not show the actual hyperplane which is six-dimensional.

Since we trained both tasks on calibration data without secondary tasks, this plot shows how differently data shifts between training (no secondary tasks) and testing (with secondary tasks). For participant *pp6* testing features are still separable while for participant *pp16* test set features shift in a way that makes it impossible to separate them with the trained classifier. The corresponding classification rates (72.22% and 47.22%) support that finding.

With one CSP filter per task, we also performed a 6-fold cross-validation to classify each secondary tasks against *clean* for both hands. Average classification rates over all 16 participants are visualized in the form of boxplots in Figure 3.11. While classification rates for *news* against *clean* are mostly around chance level, *stimulation, eyes-closed* or *numbers* against *clean* achieved median accuracies between 76% and 93%.

This means, we can indeed assume major feature shifts in the data, especially in the *eyes-closed*, *numbers* and *stimulation* tasks which might cause the difficulties in the standard classification pipeline. Before including this finding into the classification process, we first want to investigate whether trials are separable into left and right hand imaginations.

| pp | Overall | Clean | Eyes | News | Numbers | Flicker | Stimulation |
|---|---|---|---|---|---|---|---|
| 14 | **96.30** | **100.00** | 97.22 | **100.00** | 86.11 | **100.00** | 94.44 |
| 6 | **90.85** | 93.06 | 86.11 | **94.44** | **87.50** | 93.06 | **90.91** |
| 5 | 83.33 | **93.06** | **80.56** | 84.72 | 75.00 | 93.06 | 73.61 |
| 10 | 70.60 | 77.78 | 80.56 | **80.56** | **59.72** | 62.50 | **62.50** |
| 16 | **70.60** | **75.00** | **73.61** | **70.83** | **55.56** | **83.33** | **65.28** |
| 3 | **64.35** | **69.44** | **68.06** | **63.89** | **66.67** | 58.33 | 59.72 |
| 8 | 51.96 | **50.00** | 57.58 | 55.56 | 48.61 | 48.61 | 51.39 |
| 7 | **57.64** | 48.61 | **62.50** | 58.33 | **65.28** | 54.17 | 56.94 |
| 13 | **58.80** | 65.28 | **68.06** | 55.56 | 52.78 | 47.22 | **63.89** |
| 15 | **56.02** | 50.00 | **62.50** | **56.94** | **55.56** | 54.17 | **56.94** |
| **4** | **61.55** | 56.94 | **58.33** | **65.15** | **56.94** | **70.83** | **61.11** |
| 9 | **53.54** | **56.06** | **61.11** | 47.22 | **65.15** | 40.28 | 51.39 |
| 1 | 51.62 | 48.61 | 51.39 | 48.61 | **54.17** | 52.78 | **54.17** |
| 11 | 50.23 | 45.83 | **56.94** | **48.61** | 48.61 | 43.06 | **58.33** |
| 2 | 46.99 | **55.56** | 43.06 | **45.83** | 47.22 | 43.06 | 47.22 |
| 12 | 49.07 | **58.33** | 40.28 | 50.00 | **55.56** | 41.67 | **48.61** |
| ∅ | **63.34** | 65.22 | **65.49** | **64.14** | **61.28** | 61.63 | **62.28** |

Table 3.3: Mean classification accuracies for 6 separate classifiers (one per secondary task). One row represents one participant, the first column shows the participant codes and the second column contains the averaged accuracies over all secondary tasks. The results that improved compared to Table 3.2 are highlighted in bold. Participants assigned to a higher (lower) group compared to Table 3.2 are highlighted in bold (red).

### 3.3.5 Non-discriminativity

In order to find out whether the data is separable, we computed one classifier for each secondary task and only tested on the same secondary task on which we trained. We therefore conducted a 6-fold cross-validation on each secondary task (60 training trials, 12 testing trials, preprocessing and feature extraction as in Section 3.3.1). Average classification rates are displayed in Table 3.3.

In comparison with the results from Table 3.2 where we computed one classifier for all tasks (see Section 3.3.1), the overall classification rates could be improved for most participants, the average accuracy over all tasks and participants increased from 61.81% to 63.34%. Yet, the overall improvement tested with a Wilcoxon signed rank test is not significant ($\alpha = 5\%$). While the overall classification rate for the *clean* and *flicker* tasks decreased, the performance for both the *numbers* task and the *eyes-closed* task improved by almost 5%. This improvement is not surprising considering the detection of feature shifts in both tasks in Section 3.3.4.

Several participants (*pp1, pp2, pp11, pp12*) could still not reach the threshold of 61.11% in any task which indicates that their data is not separable into left and right hand motor imagination. Whereas, for participants *pp16* and *pp4*, classification rates improved by around 11% and 8%, respectively.

## 3.4 Evaluation of new strategies for out-of-lab

In the last section, we found that artifacts highly contaminate the data, especially in the *numbers* task. We also discovered that test features crucially shift with respect to the calibration data for some participants. However, when we computed task-specific classifiers, we could separate left from right hand motor imagination for most participants. Since standard machine learning methods might fail in case of feature shifts and artifact contamination, we will further investigate other methods such as artifact removal. In the following section, we propose three strategies to improve classification.

1. Artifact removal. Since we discovered in Section 3.3.3 that data is highly contaminated with artifacts, we removed the artifacts before classification.

2. Classifier ensemble. Instead of dividing the dataset into the different secondary tasks, we applied all 6 classifiers from Section 3.3.5 and average the classifiers output. We will call this approach *ensemble CSP*

3. 2-step classification. We first identified the type of distraction task during the motor imaginary task before applying the respective classifier. This approach will be called *2-step CSP*

| | Muscle | Ocular | Non-neural | Muscle &mixed | Non-mixed | None |
|---|---|---|---|---|---|---|
| **Clean** | 66.36 | 64.46 | **68.59** | 65.32 | 66.20 | 67.06 |
| **Eyes** | 60.72 | 60.47 | **62.71** | **61.39** | **60.90** | 60.82 |
| **News** | **62.02** | **62.63** | **64.80** | **63.11** | **62.63** | 61.42 |
| **Numbers** | **56.72** | 56.38 | **61.43** | **57.55** | **56.81** | 56.47 |
| **Flicker** | 63.63 | **64.50** | **65.38** | **64.17** | **65.28** | 64.06 |
| **Stimulation** | 59.45 | 59.54 | **63.34** | **61.65** | 60.15 | 61.02 |
| ∅ | 61.49 | 61.33 | **64.38** | **62.20** | **61.99** | 61.81 |

Table 3.4: Average classification accuracies for each secondary task for classification after artifact removal. One column represents a different group of artifacts, the last column shows classification without artifact removal. Those are the same results as the last row in Table 3.2. The results that improved compared to Table 3.2 are highlighted in bold.

### 3.4.1 Improvement via artifact reduction

Table 3.4 shows the average classification rates over participants for all secondary tasks when each artifact group is removed from the data. To remove artifact groups, the independent components from each artifact group were extracted from both the calibration and test data. When testing whether the overall performance per participant differs from the baseline accuracy with a one-sided Wilcoxon signed rank test, no p-value is below 0.05 (not corrected for multiple hypothesis tests).

Although removing artifacts does not cause any significant differences in performance, some qualitative trends are interesting. It shows that the removal of non-neural artifacts shows the highest increase in classification accuracy for all secondary tasks. Removing muscular and ocular artifacts both result in a decrease in classification performance for 4 secondary tasks compared to not removing any artifacts. The performance in the *news* task is improved by the removal of each artifact group. Yet, the improvements are not statistically significant. These results are consistent with previous investigations in which removing artifacts did not improve BCI performance significantly (Frølich et al., 2015b).

### 3.4.2 Improvement via classifier ensemble: ensemble CSP

Instead of choosing a task-specific classifier for each trial, we propose an ensemble approach, where we applied all six classifiers to all trials and averaged over the output to determine whether the user performed left or right hand motor imagination. We therefore also conducted a 6-fold cross-validation on each secondary task. Since we recorded 72 trials per secondary task, we trained on 60 trials and tested on 12 trials, equivalent to Section 3.3.5.

| pp14 | pp6 | pp5 | pp10 | pp16 | pp3 | pp8 | pp7 | pp13 |
|---|---|---|---|---|---|---|---|---|
| **96.99** | **95.45** | **87.50** | **74.07** | **76.85** | **70.14** | 52.53 | **59.95** | **62.96** |
| **pp15** | **pp4** | pp9 | **pp1** | pp11 | **pp2** | pp12 | ∅ | |
| **56.25** | **62.12** | 50.76 | **54.40** | **53.01** | **56.25** | **53.01** | 66.39 | |

Table 3.5: Mean classification for ensemble CSP. The first and third row show the participant codes. The overall accuracy that improved compared to Table 3.2 are highlighted in bold. Participants assigned to a higher (lower) group compared to Table 3.2 are highlighted in bold (red).

Average classification rates for all 16 participants can be found in Table 3.5. Compared to the results from Table 3.2, we could significantly improve classification accuracies by 4.6% (one-sided Wilcoxon signed rank test, $\alpha = 5\%$).

Ensemble methods benefit from accuracy and diversity of individual classifiers. Because of the feature shifts we detected in Section 3.3.4, we can assume that data of the different secondary tasks does not arise from the same two Gaussian distributions (one per class). Hence corresponding classifiers are indeed diverse. Although separate classifiers do not classify all participants significantly above chance level, most of them are accurate on average. Combining those two findings explains why ensemble CSP improves accuracy for almost all participants.

### 3.4.3 Improvement via 2-step classification: 2-step CSP

Calculating one classifier for each distraction task indeed yielded higher classification rates but if we think about applying this concept to real world situations, we might not have that much prior knowledge about the scenarios the BCI is used in. Therefore, we propose a 2-step classification approach which combines classifying the respective secondary task before separating *left* from *right* hand MI and call this 2-step CSP.

In Section 3.3.4 we identified three secondary tasks with major feature shifts: *eyes-closed, numbers* and *stimulation*. Since features can shift in different dimensions we have to treat those tasks separately. The *numbers* task reaches the lowest accuracy when applying the original CSP approach (see Table 3.2). Therefore, in the first step of our 2-step classification approach, we separated the *numbers* task from the *not-numbers* task (*clean, eyes-closed, news, flicker, stimulation*). After categorizing a trial to one of these groups, we applied one of two *left* vs. *right* classifiers (one for *numbers* and one for *not-numbers*) to decide whether this trial consisted of a *left* or *right* hand motor imagination. For this approach, we only considered the 6 runs including secondary tasks and conducted a 6-fold cross-validation. This means that for the first step we used 360 training trials and 72 testing trials. In the second step we used 300 training trials for *not-numbers* and 60 training trials for *numbers*. The number of testing trials in the second step depends on the accuracy in the first step and can theoretically vary between 0 and 72 for both classes. For instance, when all trials were classified as *not-*

| | | 1st step | 2nd step | |
|---|---|---|---|---|
| **pp** | **Overall** | **Cond** | **Not-Num** | **Numbers** |
| 14 | **97.22** | 99.31 | 98.90 | 88.41 |
| 6 | **94.13** | 96.71 | 95.53 | 86.76 |
| 5 | **88.89** | 99.31 | 91.09 | 78.08 |
| 10 | **77.08** | 98.38 | 80.22 | 61.64 |
| 16 | **75.46** | 94.68 | 82.23 | 46.99 |
| 3 | **73.15** | 98.38 | 73.70 | 70.15 |
| 8 | 56.10 | 93.19 | 57.91 | 47.22 |
| 7 | **63.66** | 88.89 | 64.86 | 58.54 |
| 13 | **61.81** | 98.61 | 62.71 | 57.14 |
| 15 | **59.49** | 99.07 | 60.28 | 55.56 |
| **4** | **68.78** | 98.12 | 72.07 | 51.47 |
| 9 | 52.58 | 97.89 | 51.10 | 60.94 |
| **1** | **54.63** | 93.29 | 54.84 | 53.85 |
| 11 | **53.01** | 98.61 | 54.75 | 44.59 |
| 2 | **53.47** | 97.92 | 54.34 | 49.33 |
| 12 | 47.92 | 94.21 | 47.91 | 47.95 |
| ∅ | **67.34** | 96.66 | 68.90 | 59.91 |

Table 3.6: Mean classification accuracy for two-step classifier. Each row represents one participant, the first column shows the participant codes. In the *overall* column, a weighted average accuracy of the second step is displayed. Column *Cond* shows the accuracy for the first step (no weighting) where data is classified in *not-numbers* and *numbers*. The last two columns show the respective accuracy for both classes. The overall accuracies that improved compared to Table 3.2 are highlighted in bold. Participants assigned to a higher group compared to Table 3.2 are highlighted in bold.

*numbers* in the first step, all trials will also be further classified by the *left* vs. *right* classifier for *not-numbers*.

Results are summarized in Table 3.6 where the average classification rates for both steps are listed. The overall classification rate is the weighted average performance of the *numbers* and *not-numbers* task in step 2. The weighting compensates for the different number of trials in both tasks.

As already discussed in Section 3.3.4, the different secondary tasks are easily separable, results of the first step in Table 3.6 show that classification rates are mostly between 93% and 100%. Except for *pp7* where we are only able to classify 89% of the tasks correctly. Since only 12 out of 72 trials in each run belong to the *numbers* task (16.7%), this result leads to the conclusion that we were not able to separate the tasks for this participant. Overall classification accuracies significantly improved by 5.5% (one-sided Wilcoxon signed rank test, $\alpha = 5\%$) compared to the original CSP results in Table

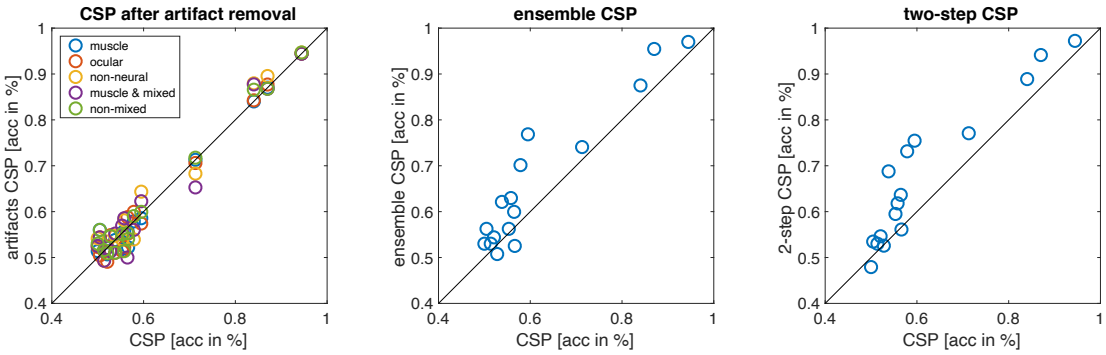


Figure 3.12: Accuracies of the three new approaches compared to the accuracy of CSP trained on clean. One circle represents the average accuracy of one participant. Data points above the diagonal have improved in the respective CSP approach compared to the original CSP analysis pipeline.

3.2 and for most participants (except *pp8, pp9, pp12*) we reached higher classification rates with this 2-step approach. One reason for this may also be the amount of training data we used to train the *not-numbers* classifier. Whereas we used between 60 and 72 training trials for the previous approaches, here we could use up to 300 training trials.

In Figure 3.12, classification results for all 3 new approaches (artifact removal, ensemble and 2-step approach) are displayed in comparison to the original CSP approach (see Table 3.2). Ensemble CSP and 2-step CSP show clear improvements in classification accuracy for participants that previously reached accuracies of at least 60%.

## 3.5 Discussion

We recorded a motor imagery-based BCI study with 16 participants where different distraction scenarios were added as secondary tasks to simulate a semi-realistic environment and systematically investigated the influence of those noise sources on the motor imagery performance. Those distraction tasks included watching a *flicker*ing video, listening to *news* sequences, vibro-tactile *stimulation* on the forearms, searching the room for letter-*number* combinations and performing the motor imagery task with *eyes-closed*.

We have presented group-averages that show typical ERD/ERS effects especially during the first half of the trial over the motor cortex, typical phenomena according to the literature. Power spectra for occipital channels and spatial patterns show expected differences for different frequency bands in the alpha range for 3 of the 5 secondary tasks.

### Main findings

The standard CSP+RLDA classification pipeline clearly shows that classification accuracy decreases in the distraction tasks. Results between participants varied a lot and

only 10 out of 16 participants reached an overall accuracy significantly above chance level. Especially classification accuracies in the *numbers* task were much lower compared to the *clean, news* and *flicker* tasks. We investigated three possible reasons for this poor BCI performance

1. *Artifact contamination.* We found a high amount of artifacts in the data, particularly for the *numbers* task, where 72% of data variance can be explained by non-neural artifacts.

2. *Feature shifts.* We detected major feature shifts between data without and data including distraction tasks (e.g. *eyes-closed* and *numbers*) by classifying each distraction task versus *clean* data. This is not surprising after seeing differences in power spectra for those tasks compared to *clean*.

3. *Non-separability.* We trained one separate classifier for each secondary task. Since we could improve classification accuracies for most participants, we can conclude that, applying the correct classifier, left and right hand motor imagination is indeed separable for most participants. Thinking about real-world scenarios, the problem however is, that we do not always know which task a user is carrying out while controlling the BCI.

We therefore proposed three different approaches to improve performance:

1. *Artifact removal.* After decomposing calibration data with ICA, we classified the ICs with IC_MARC (Frølich et al., 2015a) into neural data and different artifact types. After removing five different groups of artifacts, we used RLDA based on CSP for classification. Results are displayed in Table 3.4.

2. *Ensemble CSP.* For each of the six secondary tasks, we trained one CSP+RLDA classifier. We applied all six classifiers to the data and used the averaged output as a classification decision, Table 3.5 shows the results. We chose this approach to build a more robust classifier with respect to feature shifts because of different artifact types in the training data.

3. *2-step CSP.* We classified in two steps. During the first step, we separated the noisiest secondary task (*numbers*) from the remaining 5 tasks (*not-numbers*, i.e. four remaining distraction tasks plus *clean*). In the second step we used two different classifiers, one for *numbers* and one for the remaining *not-numbers* tasks to separate the data into *left* and *right* hand MI, results can be found in Table 3.6. By applying this 2-step approach, we could improve classification accuracy for the numbers task which was most affected by artifacts.

The artifact removal did not lead to significant improvement (see Figure 3.12). One reason for that might be that ICA could not separate all participants' data into meaningful independent components (see Figure 3.8) which then makes the application of IC_MARC and the following classification after artifact removal difficult. For ensemble CSP and 2-step CSP, we could however significantly improve accuracy results. For both

approaches, classification performances improved particularly for those participants already reaching significant BCI control in the original CSP approach (#1 and #2 in Table 3.2). It is worth mentioning again that the 2-step approach clearly benefits from the increased number of training trials compared to the original and the ensemble CSP approach.

### Limitations and Outlook

After first steps have been made to leave the controlled lab environment, this study systematically and quantitatively analyses how different scenarios influence BCI performance. The challenges we identified, especially the ones of dual character in the *numbers* task need to be considered for future studies and are worth being further analysed.

Although two of the proposed approaches significantly improved classification accuracy we should take into account that when applying them to everyday life, additional challenges might occur. Whereas in our study we exactly know under which secondary task a participant is performing the experiment, it is not always clear in which situation users find themselves when controlling a BCI in their everyday life environment. This issue might be tackled with a large amount of training data in various possible real-world scenarios. Recordings over a larger time span might also uncover additional challenges as users get tired over time, electrodes can shift and the environment could change.

Furthermore, we also have to consider the number of EEG channels we used. Preparation of 64 electrodes for gel-based EEG cap is time-consuming and challenging, particularly outside of the lab since electrodes dry out over time and users need to wash their hair afterwards. Reducing the number of channels and considering dry and mobile EEG systems would also further improve usability of real-world applications.

We also have to point out that most of the participants were for the first time confronted with a BCI system. Imagining a movement is abstract and some participants may improve their performance by engaging in more feedback training before going "out-of-lab".

Future studies could also explore whether harvesting a database of significantly larger numbers of participants and tasks may allow an invariant and user independent decoding (Fazli et al., 2015), e.g. using deep neural networks.

# 4

# Tracking auditory beat-based entrainment in MEG via Source Power Co-Modulation

## 4.1 Motivation

Rhythm and timing are crucial in our daily lives. Sequencing information at the level of a few hundred milliseconds in particular plays a major role for both the perception and production of music, speech, and movement. This relevant time scale here corresponds to that of our ability to "feel the beat", that is from approximately 200 to 800 ms (London et al., 2004). The ability to detect and process a regular rhythmic beat in an auditory stream of acoustic events is one of the essential mechanisms to our musical, speech and language skills (e.g. Bekius et al., 2016). This ability can be disturbed in neurological as well as developmental disorders with marked consequences to the individuals' life quality, in terms of the perception and or production of music, speech, or bodily movement (e.g. Grahn et al., 2009; Huss et al., 2011; Ladányi et al., 2020). It is therefore of critical importance to first understand how the brain tracks the beat, to then find ways to monitor and support the rehabilitation of dysfunctional timing processes in individuals with those disorders.

In terms of neural mechanisms, one that has been emerging from recent literature is that of coupled oscillations, i.e. the entrainment of the power of a faster oscillation with the phase of the slower (e.g. Large et al., 2015). More specifically, the timing system has been promoted to feature the entrainment of neural oscillations in the beta range (13-30 Hz) with the periodicity of the beat in musical rhythms (Iversen et al., 2009).

In attempts to get to the bottom of the structural and functional neural correlates that underlie our entrainment with and "feeling of the beat", previous MEG and EEG work both demonstrated clear effects in oscillatory activity. In their seminal 2012 MEG study, Fujioka et al. (2012) demonstrated a clear periodic modulation in the beta band power, with the beat frequency in the MEG. The data showed one peak and one trough per beat cycle within the beta band and the same pattern could be seen for an isochronous (i.e. perfectly regular) beat for three different tempi of 390, 585, and 780 ms. These activity patterns were extracted as the result of an anatomically guided analysis and appeared most prominently in the auditory cortex of the temporal lobe. Further support for the beta band "driving the brain on beats" came from related EEG work (Chang et al., 2016) with converging conclusions (Teki, 2014).

We here present an alternative, customised, and machine-learning-based analysis pipeline that enables us to extract in a data-driven way the maximally relevant brain activity in source space. Therefore, we propose a modified version of the Source Power Co-Modulation algorithm (SPoC, see Chapter 2.5.3) to correlate with a regular target function modulated by the beat frequency: Fourier-SPoC (f-SPoC). This allows us to seek components in source space that maximally vary their activity with the periodicity of the beat (here, 400 ms) The advantages of this pipeline are that no a priori input about brain structures is required, neither in terms of pre-defined anatomical brain regions of interest nor in terms of the individual's brain. Our pipeline also entails Spatio-Spectral Decomposition (SSD, see Chapter 2.5.2). The application of SSD prior to f-SPoC optimizes the input for f-SPoC with respect to the activity in the band of interest, initially. In order to test the efficacy of our analysis pipeline, we employ a most sophisticated sensor-space analysis, based on a state-of-the-art cluster-based permutation analysis (CBPA; Maris et al., 2007). To compare the results between the two, we employ a wavelet-based time-frequency analysis (TFA) followed by a fast-Fourier transformation (FFT, Nussbaumer, 1981) to calculate the depth and specificity of the periodic modulation of activity as two quantifiable outcome measures.

Results for both approaches show periodic patterns in a combined alpha-beta band whereas results from f-SPoC show stronger power changes for the respective peaks and troughs and also clearer results on the individual participant level. We also tested our pipeline on resting state data where no stimulus is present. A more detailed analysis on the temporal level shows changes in the activity between the beginning and the end of the 5-minute passive beat listening task both in the frequency and time domain.

The final goal of this analysis is to provide a data-driven assumption-free brain marker of neuronal entrainment to an external beat.

The rest of this chapter is structured as follows: In Section 4.2, we describe the MEG recordings and the preprocessing pipeline. In Section 4.3 and 4.4 we describe the sensor-level and f-SPoC analysis respectively together with the corresponding results. In Section 4.5 we will investigate the temporal dynamics during the experiment before we round the chapter up with a detailed discussion in Section 4.6.

This chapter follows our own work from Brandl et al. (2021) which has been published as a preprint and is currently in preparation for journal submission.

## 4.2 Methods

In order to investigate on neural oscillations that are expected to entrain to a regular auditory input (Fujioka et al., 2012), we analysed the MEG of participants hearing a metronome-like steady repetition of a tone.

### 4.2.1 MEG Recordings

MEG data was recorded with an Elekta Neuromag TRIUX system (Elekta Neuromag, Helsinki, Finland) equipped with 306 channels at a sampling rate of 1000 Hz. During the recording, participants (N=28) were performing a passive listening task while watching a silent movie. The auditory stimuli consisted of a continuous sequence of 300 Hz sine tones with a duration of 100 ms (with 10 ms fade-in and -out) presented with an isochronous inter-onset interval of 400 ms for a total duration of 5 min. For Section 4.4.2, we will also consider 5 min of resting state data which has been recorded in the very beginning of the experiment. The present study was conducted on a subset of a larger dataset acquired at Aarhus University (Bonetti et al., 2020) and approved by the Ethics Committee of the Central Denmark Region (De Videnskabsetiske Komiteer for Region Midtjylland) (ref. 1-10-72-411-17), carried out according to the Declaration of Helsinki (World Medical Association, 2001).

### 4.2.2 Preprocessing

The raw MEG sensor data (204 planar gradiometers and 102 magnetometers) was preprocessed by the Elekta Neuromag MaxFilter(TM) software version 2.2.15 (Taulu et al., 2006; Taulu et al., 2009) for attenuating the interference originating from outside the scalp by applying temporal signal space separation (tSSS). The signals were corrected for head movements with a Maxfilter. The default tSSS parameters were applied with inside expansion order of 8, outside expansion order of 3, optimization of both inside and outside bases, default subspace correlation limit of 0.98 and default raw data buffer length of 10 seconds. Subsequent data preprocessing steps were performed by using the FieldTrip Toolbox (2018-11-12) for Matlab (Oostenveld et al., 2011). For better comparison with EEG and other MEG systems, we only considered the 102 axial magnetometer sensors and left out the 204 planar gradiometers. Data has been low- and high-pass filtered at 500 Hz and 0.5 Hz, respectively, with a finite impulse response filter. To remove artifacts introduced by heart beat and eye movements, independent component analysis has been applied (Makeig et al., 1996) and up to 2 out of 64 components per participant (one for heart beat and one for eye movements) were removed after visual inspection. Data was then cut into epochs of two beat periods, i.e. 800 ms, and 1000 ms before and after as a buffer for the wavelet analysis and aligned 1000 ms prior to tone onset.

Finally, we split the data into train and test set where every third tone was taken out and added to the test split and excluded from the analysis steps in Sections 4.3.1 and 4.4.1.

## 4.3 Sensor-Level Analysis

We chose cluster-based permutation analysis as a data-driven method to identify significant sensors from which we will compute the time-frequency response across participants and trials to detect beat-based (de)synchronisation. CBPA serves as baseline

method and comparison for our machine learning-based spatial filtering approach with f-SPoC (see Section 4.4).

### 4.3.1 Analysis Pipeline

#### Cluster-based Permutation Analysis

One-sample cluster-based permutation tests can detect statistically significant power modulations across participants in MEG or EEG (Maris et al., 2007). This is a data-driven statistical method that is useful in finding statistically significant clusters of power (de)synchronisation spanning coherent triplets in (time, frequency, and sensors) in case the relevant pairs are unknown. The permutation version of the one-sample t-test iteratively performs t-tests where random signs are added to the observed sample to estimate a null-distribution of t-values. The p-value is then defined as the ratio of t-values in the null-distribution exceeding the observed t-value obtained without adding random signs (Basu, 2011). For the one-sample test, it is further assumed that the population values are symmetrically distributed, which is a reasonable assumption given that we are searching for periodic power modulations. The cluster-based version of the permutation test includes an initial step, where neighbouring significant points in time, frequency and sensor location are merged into larger clusters, thereby offering a solution to the multiple comparison problem (Maris et al., 2007).

We applied the cluster-based permutation test carrying out 1000 randomizations. For the initial cluster statistic we used the default channel neighbour configuration for the Neuromag Triux MEG system implemented in FieldTrip, the maximum sum t-statistic, and a significance threshold of p=0.01. For the cluster-based permutation statistics we used a significance threshold of p=0.025. We considered all 102 axial magnetometer sensors, 20 logarithmically scaled frequency bins between 5-35 Hz and 100 linearly scaled time points in a time window of 400 ms corresponding to one beat period. We thus found 3 desynchronisation [negative] and 2 synchronisation [positive] clusters in time, frequency, and sensor location that are significant across participants.

Furthermore, we found 78 significant sensors across all participants in the cluster-based permutation analysis. For each sensor we therefore averaged over a 2-dimensional binary array of all frequency bins and time points ($29 \times 100$) which indicates if a corresponding sensor at an individual time point and frequency bin is significant. Each sensor with an average value larger than 0, i.e. each sensor that is significant in at least one (frequency, time) combination, is considered as significant. The average values, i.e. the average percentage of significance for each sensor, is displayed in the scalp plot in Figure 4.1. We selected those 78 significant sensors and applied spectral analysis as explained in detail in the following section.

#### Spectral Analysis

Time-frequency analysis (TFA) was applied to the spatially filtered signal. We applied wavelet analysis to individual time windows (onset-aligned) of 2-beat periods for all

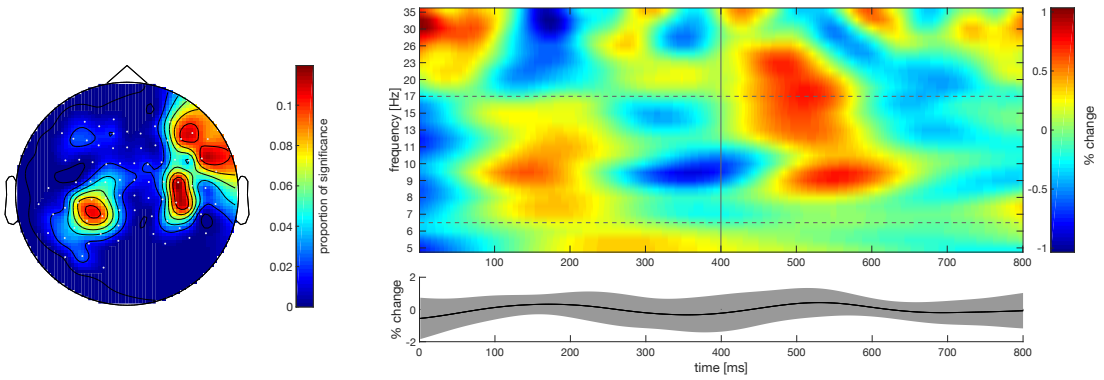


Figure 4.1: Results of cluster-based permutation analysis (CBPA) *Left:* The topography shows the average proportion of significance for each sensor averaged over participants and frequency bins. The 78 sensors that proved to be significant across all participants in the CBPA are highlighted by a white dot. *Right:* Time frequency response (TFR, *upper*) averaged across participants, time windows and significant sensors shows percentage change in power. The *lower* subplot shows averaged TFR across the frequency band of 6-18 Hz periodic activity can be seen with a peak around 150 ms after tone onset and a through around 350 ms after tone onset. The grey area shows standard deviation across participants.

78 significant sensors in a frequency band of 5-35 Hz, individually for each participant (Morlet wavelet, window width of 7). We computed the spectral amplitude after removing the buffer window, averaged over trials and detrended. We then calculated the percentage change of power by subtracting and dividing by the mean of the entire trial length for each frequency bin individually. Finally, we averaged over sensors and participants.

### 4.3.2 Results

In Figure 4.1, we show results of the cluster-based permutation analysis and the corresponding time frequency response (TFR). The topography on the left shows the average percentage of significance for each sensor. The significant clusters of power changes were overlapping in channel, time and frequency extended in the time span of $4 - 360$ ms after tone onset and within frequencies of 6-30 Hz.

On the upper right we visualized the time frequency response in 5-35 Hz averaged over all 78 significant sensors and participants. We see a clear regular event-related synchronisation/desynchronisation (ERS/ERD, Pfurtscheller et al., 1999) pattern in the frequency band of 6-18 Hz with a synchronisation starting around 50 ms after each stimulus onset and a desynchronisation starting 280 ms after stimulus onset up to 50 ms after the next stimulus onset with up to $\pm 1\%$ power change.
On the lower right we further averaged over the frequency band of 6-18 Hz, i.e. an

average envelope where we see a regular ERS/ERD even clearer with a positive peak around 150 ms after stimulus onset and a negative peak around 350 ms after stimulus onset.

## 4.4 Spatial Filtering

To extract an amplitude-modulated oscillatory component, we applied spatial filtering methods to the preprocessed MEG data. Spatio-Spectral Decomposition as introduced in Section 2.5.2 computes spatial filters that aim to improve signal-to-noise-ratio in a frequency band of interest. In our pipeline this serves as a preprocessing step to find a lower dimensional representation on which we afterwards apply a novel version of Source Power Co-Modulation (SPoC, see Section 2.5.3) to extract components that maximally correlate with the beat frequency. It is important to note that we do not need any anatomical knowledge of the individual brain to extract those components. Both methods are purely mathematical and based on inverting the forward model (Haufe et al., 2014b). This makes it easy to apply them to multidimensional brain signals such as EEG or MEG.

For the special case of a regular target function, we propose Fourier-SPoC, an extension to SPoC.

### 4.4.1 Fourier-SPoC

As we are interested in finding regular synchronisation and desynchronisation effects within a specific frequency band, we need to choose a periodic target function for the optimization of the SPoC-filters. We chose a sine function as the target function $z(e)$ for Fourier-SPoC (f-SPoC) modulated with the beat frequency of 2.5 Hz as we assume a rhythmic entrainment with the beat. First, we bandpass-filtered the SSD-transformed data with a passband in the range of the inner SSD band with a 3rd order zero-phase Butterworth filter. Please note that the frequency of the sine function and the inner SSD-band do not need to overlap. We then applied SPoC to the bandpass-filtered data and optimized over an evenly spaced grid of 10 different phases individually for each participant. Since we are looking for non-phase locked activity we do not know in which phase of the sine function the entrainment is happening. This means, we computed one set of SPoC filters for each of the 10 phases and selected the sine function that maximally correlates with the envelope of the SPoC-filtered signal. We selected the first and second SPoC filter for the respective phase, i.e. the ones with the largest eigenvalue and project the SSD-filtered signal $X_{\text{SSD}}$ to this two-dimensional SPoC subspace.

$$W_{\text{SPoC}}^{\top} \cdot X_{\text{SSD}} = X_{\text{SPoC}} \tag{4.1}$$

where $W_{\text{SPoC}} \in \mathbb{R}^{15\times 2}$ contains the first and second SPoC filter, $X_{\text{SSD}}$ the SSD transformed MEG signal as described in Section 2.5.2 and $X_{\text{SPoC}} \in \mathbb{R}^{2\times \text{N}}$ the projected signal.

SPoC filters are optimized for each participant on the training data and afterwards applied on the testing data according to Section 4.2.2. Spectral analysis is then applied to the SPoC-filtered time course $X_{\mathrm{SPoC}}$ as described in Section 4.3.1 with the difference that we do not average over sensors but look at the two SPoC components individually.

Please note that we will use SPoC and f-SPoC interchangeably throughout the rest of this chapter. For all analyses though, we only apply f-SPoC.

### 4.4.2 Results

#### alpha(8-12 Hz), low beta(15-20 Hz), high beta(20-25 Hz)

We applied the analysis pipeline to the recorded MEG data of each participant individually. Based on previous work, we selected three different frequency bands in the range of the alpha and beta band which have been reported to show entrainment: alpha (8-12 Hz), low beta (15-20 Hz, Fujioka et al., 2009; Cirelli et al., 2014; Chang et al., 2016) and high beta (20-25 Hz, Cirelli et al., 2014). This means, we fed the respective frequency band as the frequency band of interest (inner SSD band) into the SSD algorithm and applied f-SPoC afterwards.

In Figure 4.2 we show the time frequency responses (TFR) for the first and second SPoC-components for each frequency band. For all three frequency bands we show a group average TFR in the range of 5-35 Hz and below the average envelope over the band of interest.

Clearest ERS/ERD effects are visible in the alpha band. Here, we see a clear synchronisation in the first SPoC component especially in the frequency band of 9-17 Hz with a peak around 150 ms after stimulus onset followed by a clear desynchronisation about 350 ms after stimulus onset until 20 ms after the next stimulus with up to $\pm 3\%$ power change. The envelope below the TFR shows an average over the frequency band of 8-12 Hz over all participants and reveals repetitive ERS/ERD effects after stimulus onset.

For low beta (15-20 Hz) we see a similar pattern with a synchronisation around 100 ms after stimulus onset and a desynchronisation around 300 ms after stimulus onset with up to $\pm 2.5\%$ power change in the second component.

For high beta (20-25 Hz, lower left) the patterns are not clearly visible within 20-25 Hz but are still prominent in 10-17 Hz.

Since the results for the alpha band show a clear ERS/ERD ranging further than the band of interest, in a next step, we combine frequency bands of alpha and low beta.

It is important to note that SSD aims to minimise the power in the outer frequency band while maximising the power in the inner frequency band. This means for ERS/ERD effects present in both, the frequency band of interest and the neighbouring frequencies might not be as clear in the resulting signal in case they are tried to be suppressed

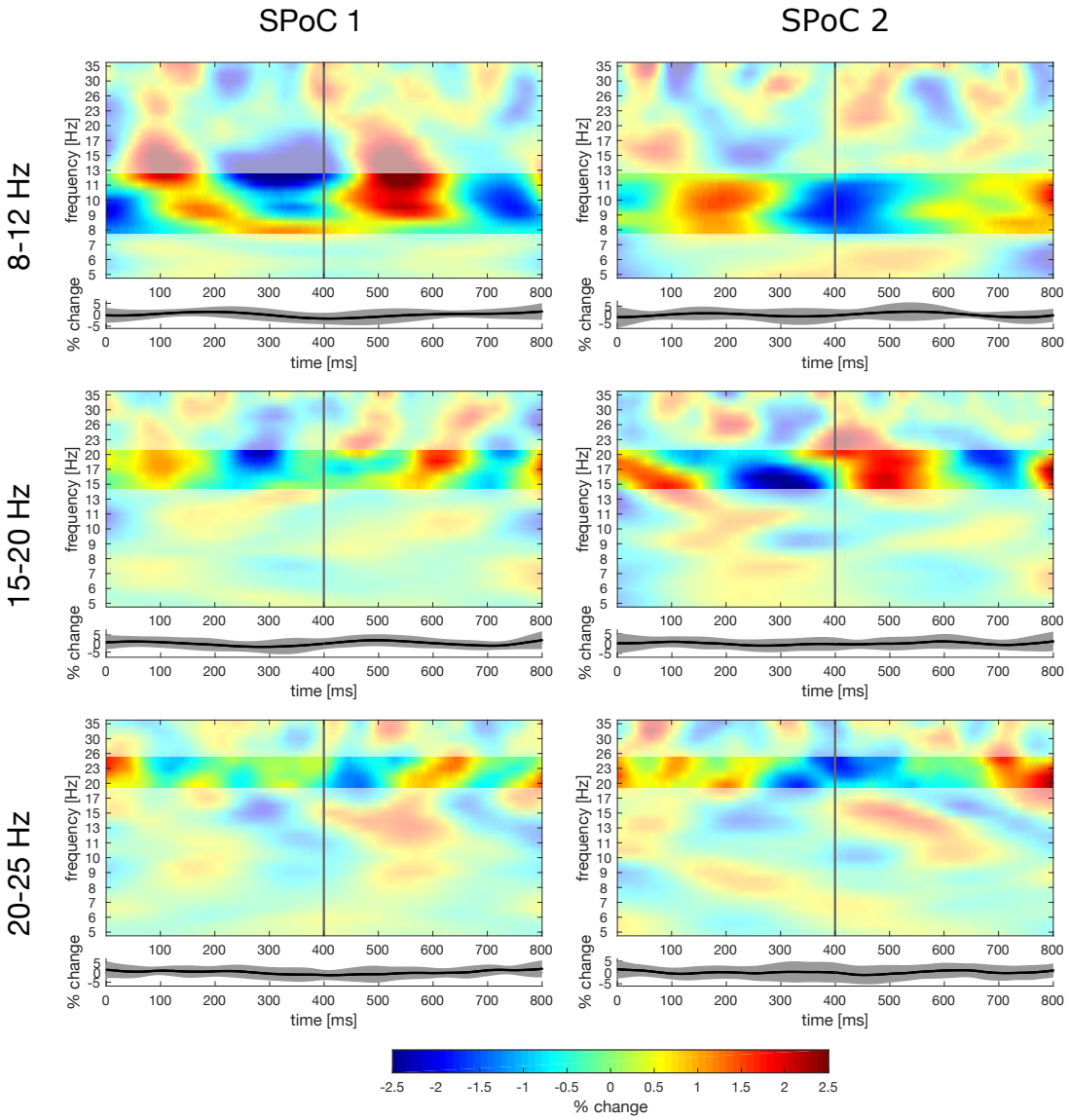


Figure 4.2: Time frequency response (TFR) of first and second SPoC component shows percentage change in power for different frequency bands averaged over all participants. Upper plots show TFR for 5-35 Hz with band of interest highlighted, lower plots show the average envelope over band of interest with the standard deviation across participants displayed as a grey area around the solid black line. Periodic peaks and troughs are most visible in the range of 8-20 Hz (upper left, center right and last row).

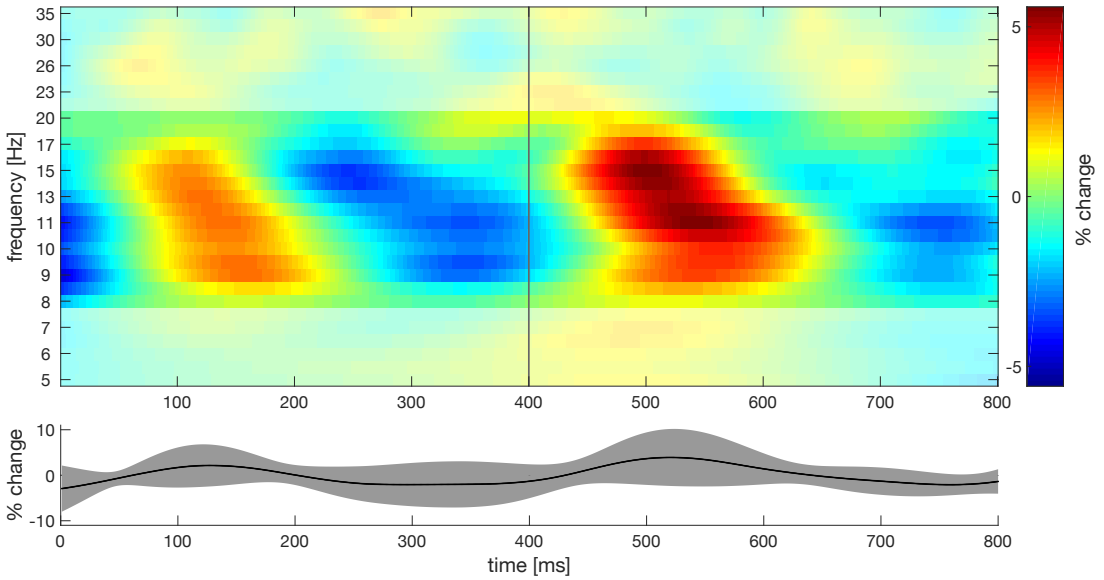


Figure 4.3: TFR of first SPoC component for the frequency band of 8-20 Hz. Lower plot shows the average envelope over the band of interest. Changes of up to 10% in power are visible after beat onset at 120 and 520 ms.

and increased at the same time. Furthermore, SPoC is based on a generalized eigenvalue decomposition which is known to be prone to outliers. This can cause the first eigenvector/component to capture artifacts with higher variance instead of relevant information. This might here be the case for low beta where see clearer patterns in the second component.

#### combined alpha-beta (8-20 Hz)

The resulting TFR for the combined frequency band of alpha and low beta in Figure 4.3 shows very clear synchronisation and desynchronisation effects in the frequency range of 8-20 Hz with up to ±5% power change. The envelope below shows an even more consistent and regular course than the one in Figure 4.2.

In Figure 4.4, we show individual results of three exemplary participants for f-SPoC optimised over 8-20 Hz and CBPA which are also part of the group averages in Figure 4.1 and Figure 4.3. Individuals were handpicked to show the range of results on the individual level. SPoC and CBPA for the first participant both show periodic patterns while the former shows a much clearer and stronger pattern than the sensor-level analysis. SPoC analysis for the second participant still shows a very clear pattern while CBPA results in a rather unclear pattern. SPoC and CBPA for the third participant both do not show periodic patterns.

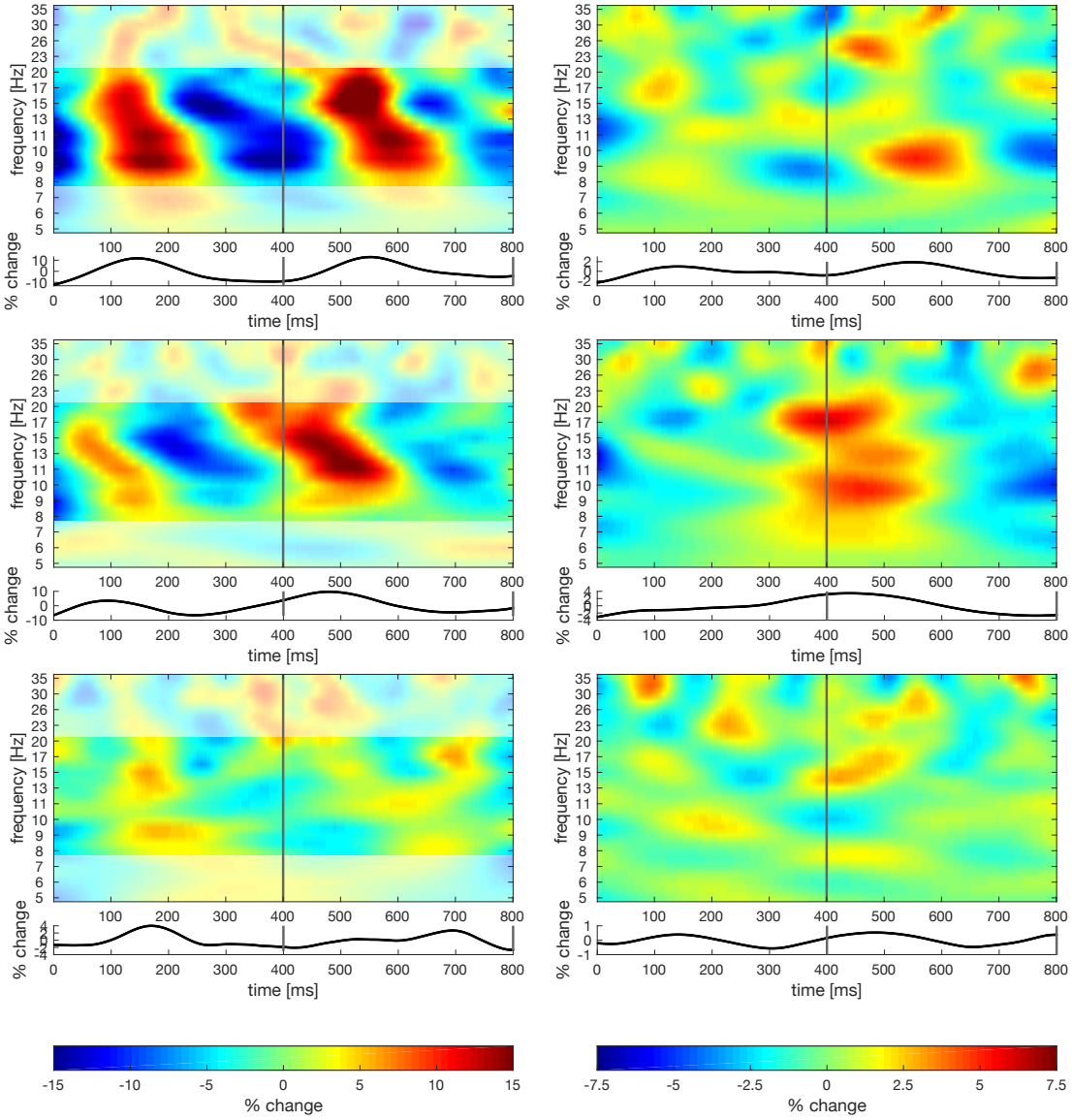


Figure 4.4: TFR of first SPoC component (left) and CBPA (right) shows percentage change in power for three individual participants (one per row) in the range of -15/15 % and -7.5/7.5 % for SPoC and CBPA respectively. Upper plots show TFR for 5-35 Hz with band of interest (8-20 Hz) highlighted in SPoC, lower plots show the envelope over 8-20 Hz (SPoC) and 6-18 Hz (CBPA).

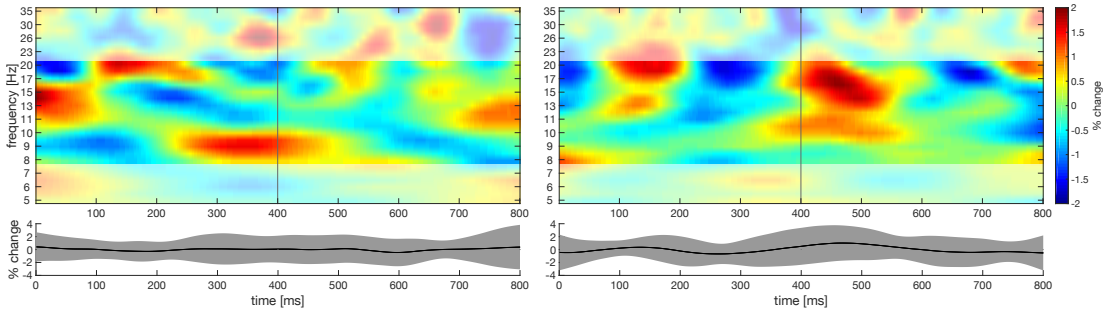


Figure 4.5: TFR of first SPoC component on resting state for 8-20 Hz, the first tone onset was chosen arbitrarily and data epoched accordingly. *Left:* F-SPoC was trained on beat listening and applied to resting state with the same SPoC components as in Figure 4.3 *Right:* F-SPoC was trained on resting state data and applied on resting state data.

### SPoC on resting state data

As a further comparison and sanity check, we also applied the SPoC component from Figure 4.3 on the resting state test data where stimulus onset was once chosen arbitrarily and following epochs were cut accordingly. Please note that resting state data has been split into train and test set the same way as the beat listening data (see Section 4.2.2).

In order to verify that the f-SPoC component from the previous section trained on beat listening for 8-20 Hz does not overfit, we apply the same component to the resting state data. In other words, we want to check whether the repetitive patterns we see in Figure 4.3 are meaningful and will disappear as soon as we apply the component to data without any regular activity.

Furthermore to prove that f-SPoC only picks up regular activity where regularity is present, we trained and applied f-SPoC for 8-20 Hz on the resting state data. In both cases, we prove the *specificity* of f-SPoC, i.e. it does not show regular patterns when there should not be any entrainment.

TFR of both analysis is shown in Figure 4.5. Whereas the TFR with the beat-trained f-SPoC component does not show any periodic pattern, the TFR of the resting state component shows some troughs and peaks. However, this does not seem to repeat in the second "trial" and is therefore not modulating with the beat frequency.

Please note that the transformation to power change which is calculated on the entire epoch automatically evens out positive and negative activity which could amplify the alternating peaks and troughs here.

### CBPA vs. f-SPoC

Time frequency results from CBPA (Figure 4.1) and f-SPoC (Figure 4.3) both show rhythmic and regular activity in a combined alpha-beta band whereas CBPA shows

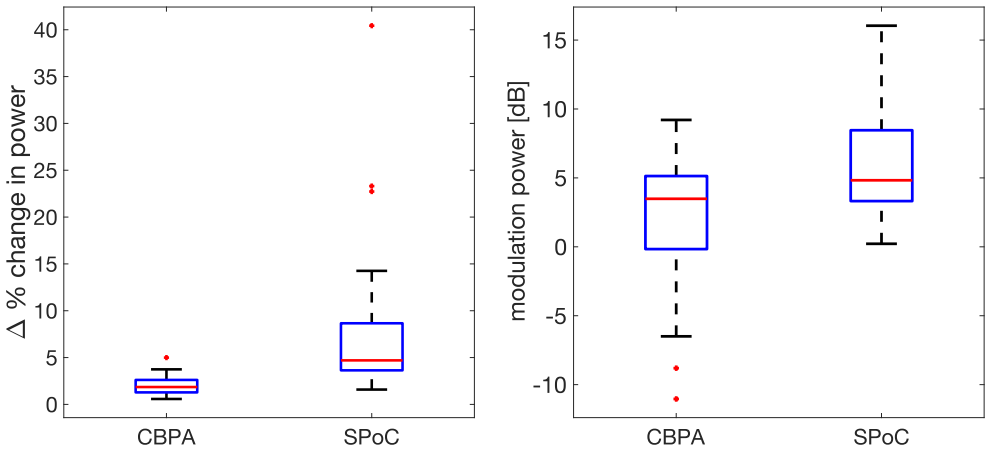


Figure 4.6: Comparative quantification of modulation at the beat frequency for CBPA and SPoC as boxplots. *Left:* Peak-to-peak difference between event-related synchronisation and desynchronisation (ERS/ERD) shows individual modulation depth. *Right:* Fast Fourier Transformation of the time frequency response at the beat frequency shows individual modulation power. Boxes show upper and lower quartiles, red line indicates the median.

strongest activity within 6-18 Hz and f-SPoC within 8-20 Hz. F-SPoC patterns show a higher percentage change in power ±5% vs. ±1%. Peaks and troughs (ERS/ERD) show their maximum amplitude around 100 ms and 350 ms after stimulus onset for both analyses

We quantify the event-related synchronisation and desynchronisation for each participant by the difference between the maximal positive amplitude and the maximal negative amplitude in the average envelope of 8-20 Hz (SPoC) and 6-18 Hz (78 significant sensors in CBPA). We here consider the first SPoC component and an average across the 78 significant sensors as show in Figure 4.3. Values for each participant are plotted in Figure 4.6 (*left*). A higher value here means more power change across the respective frequency band over the average epoch. Results for SPoC are significantly higher (one-sided Wilcoxon signed rank test with $p < 10^{-4}$) than the ones for CBPA.

We measured modulation power by applying FFT to each participant's TFR of the first SPoC component in 8-20 Hz (Figure 4.3). After averaging the result over 8-20 Hz, we calculated local peaks in the FFT of the beat frequency (2.5 Hz) as the mean difference in power to its two neighbouring sample points. Those "local peaks" give us a measurement for modulation power, i.e. a higher local peak means relatively more modulation power at the beat frequency. We also calculated local peaks for the TFR of the 78 significant sensors of the CBPA analysis in Figure 4.1 and averaged over 6-18 Hz. Results are shown in Figure 4.6 (*right*) as boxplots where one values is assigned per participant for SPoC and CBPA. Again, local peaks for SPoC are significantly higher (one-sided Wilcoxon signed rank test with $p = 0.001$) than for CBPA which means that TFR for SPoC shows significantly higher relative modulation power in the beat

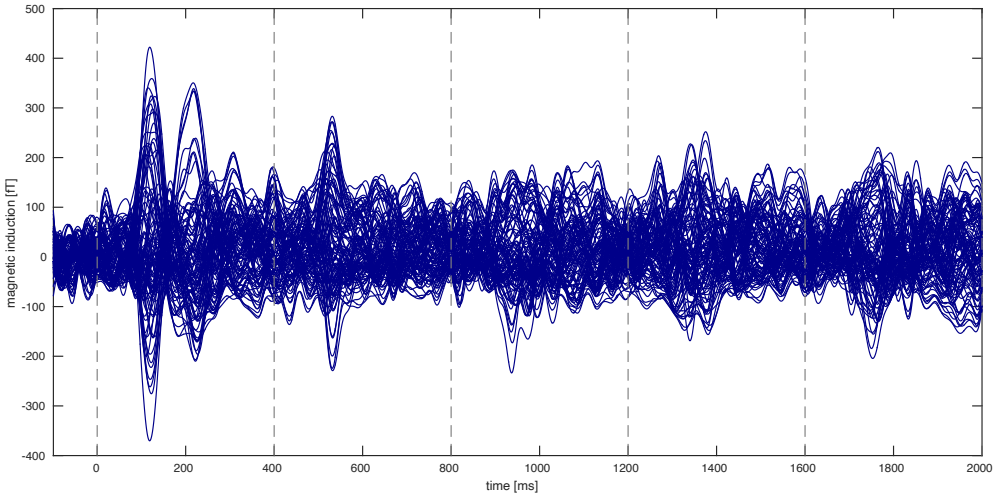


Figure 4.7: Event-related fields for the first 5 trials for all sensors averaged across participants. Tone onsets are marked by dashed lines

frequency.

## 4.5 Effects over time

Considering the monotony of the experiment, 5 minutes of passive beat listening with a continuously regular beat while watching a silent movie, we are also interested in how the MEG behaves in the course of the experiment both in the temporal and the spectral domain.

### 4.5.1 Changes in the temporal domain

It is widely known and well researched that event-related response in MEG and EEG decrease when stimuli are repeated. This is being used as an advantage, for instance in the oddball paradigm where rare stimuli can be distinguished from repetitive stimuli and therefore be used, e.g. in spelling paradigms (Farwell et al., 1988).

The authors in Hämäläinen et al. (1993) describe the auditory evoked field, i.e. the typical evoked response in the MEG following an auditory stimulus, as a small peak 50 ms after stimulus onset (P50m), a more prominent peak around 100 ms (N1m) and another deflection at 200 ms (P2m) after stimulus onset. $P$ and $N$ correspond to the typical polarity of the peaks in the electric potential (EEG) and the $m$ stands for magnetic.

Figure 4.7 shows the signal of all sensors of the first five trials averaged across all 28 participants. The signal was band-pass filtered between 1 and 30 Hz (Butterworth filter of order 5) and baseline corrected -100 to 0 ms relative to tone onset. In the first and second trial, there is clear peak at around 100 ms after tone onset visible which

we will assume to be the N1m. The following trials still show a more or less clear negative peak at about 100 ms, especially in trial 3 and 5. The first trial even shows a second component which could be identified as a P2m around 200 ms after tone onset which is not visible in the other trials. A similar effect of a disappearing P2m after the second trial has been reported in Recasens et al. (2015). It needs to be noted that event-related fields (ERFs) usually are shown as averages across many repetitions to only extract the phase-locked components. As ERFs differ in latency and amplitude between individuals it is not surprising that the signal here is very noisy and difficult to interpret. Since the beat listening experiment does not include any breaks we only have one *first* tone and thus cannot average across trials to show the habituation effects after the first stimulus. Figure 4.7 can so only be in part compared to Rosburg et al. (2010) where an average across trials of the first 5 repeated stimuli for one participant is shown.

There is an ongoing debate in the literature whether the amplitude decrement with repeating stimuli is coming from an habituation effect or a refractory period (Budd et al., 1998) and how this depends on the inter-onset-interval (Rosburg et al., 2010), number of stimuli (Fruhstorfer et al., 1970), attention (Öhman et al., 1972; Meltzer et al., 2015) or dishabituation effects in terms of deviant stimuli (Woods et al., 1986). Furthermore, there exist the two effects of repetition suppression and repetition enhancement while the former reflects a decreased neural response due to stimulus repetition and the latter an increased response associated to stimulus recognition and prediction (Todorovic et al., 2012; Recasens et al., 2015).

Also, it is not clear if all components tend to decrease or if there are also effects that lead to higher amplitudes. For instance, the authors in Costa-Faidella et al. (2011) find an increase in the P50 and P2 component and a decrease in N1 in the EEG with a repeated stimulus while the authors of Recasens et al. (2015) could not reproduce this result with a similar experimental paradigm for the MEG.

In our case we see a clear decrease in what we assume to be the N1m already after the first trial, as expected. To localize the auditory field and look at the respective brain regions one would need to conduct more advanced source localization methods such as beamforming (Van Veen et al., 1988) which is beyond the scope of this thesis.

### 4.5.2 Changes in the spectral domain

Figure 4.8 shows windows of TFR from the CBPA and the f-SPoC analyses, averaged across participants and 100 trials each. For better comparison between CBPA and the test set of SPoC, the first window begins with trial 101 and continues with a sliding window of 50 epochs, i.e. neighbouring windows overlap by 50%, to show the evolvement of the activation over the 5 minute beat listening experiment. Please note that we look at the average of only one beat cycle here.

Trough and peak following the tone onset is visible in most but not all windows of the CBPA analysis. The first window shows clear ERD/ERS in the frequency band of

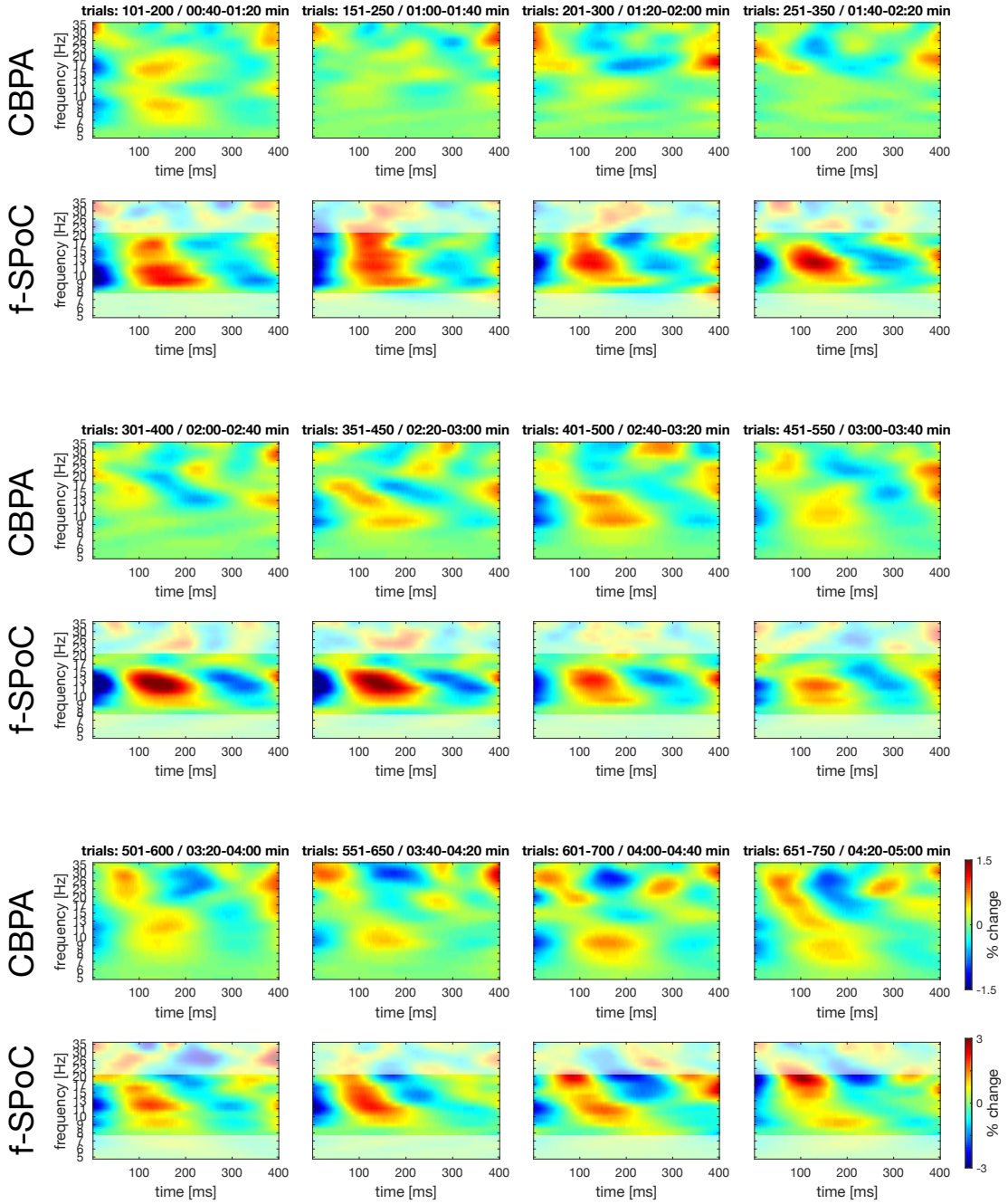


Figure 4.8: TFR over time of CBPA and first SPoC component. Each window shows the average over 100 epochs with 50% overlap between neighbouring windows. SPoC was trained on first 100 epochs for 8-20 Hz.

7-20 Hz while the following windows show no visible effect. Later in the experiment, around minute 2:20 the pattern is again visible in a more narrow frequency band of around 9-15 Hz before towards the end of the experiment the frequency band seems to become even more narrow and only shows activity in the alpha band (7-11 Hz) and then ends with an again broader activation in 7-15 Hz.

In comparison, we trained SPoC for each participant on 8-20 Hz on the first 100 epochs of the experiment which corresponds to 40 seconds and apply it to the remaining epochs. The windows in Figure 4.8 are aligned in time. The ERD within 100 ms after stimulus onset is still visible in all epochs followed by an ERS between 100-200 ms as we have seen before in Figure 4.3. Interestingly in the beginning, the activity is visible in both the alpha and beta band across 8-26 Hz but soon vanishes in frequencies above 15 Hz and then again reappears towards the end of the experiment where alpha activity gets weaker. Also, the overall power change within a beat cycle is clearly stronger in the beginning, gets weaker in the middle of the experiment and comes back stronger in the final epochs.

We clearly see a change in power across different frequency bands over time in both methods. Especially in the f-SPoC analysis, patterns are still visible even if trained only on 100 trials. Though this is not the case when we train f-SPoC on trials from the middle or the end of the experiment. Changes however are not aligned and differ between analyses. Whereas in CBPA the activation is not visible in the beginning and then evolves from being present in a combined alpha-beta band to only being visible in the alpha band, the f-SPoC patterns show a clear alpha-beta activation in the beginning, an alpha activation in the middle and only beta activation in the end of the experiment. This makes it difficult to interpret the results as furthermore no comparable analysis of such an experiment can be found in the literature. An increase in alpha activity has been linked to different forms of meditation (Lee et al., 2018), boredom (Raffaelli et al., 2018) and also mind wandering (Compton et al., 2019) e.g. when participants were "drawn into the music" (Markovic et al., 2017). On the other side, an increase in beta activity is associated with attention and negatively correlated with boredom. The ratio of beta power/alpha power has therefore also been used as a measure of engagement (Raffaelli et al., 2018). In general activation seems to be higher in passive vs. active listening (Jäncke et al., 2018).

The results in Figure 4.8 do not show enough evidence to conclude what exactly causes the shifts in power across frequency bands. It needs to be further investigated if effects as mind wandering or boredom and a shift in attention cause those changes and where in the brain they are originating as CBPA and f-SPoC, both not optimized for those changes, might focus on different parts of the brain which would explain the differences in the spectral analysis.

## 4.6 Discussion

Understanding how the brain processes rhythm is crucial to support those with dysfunctional timing processes. In this chapter, we have proposed and discussed a new version of Source-Power Co-Modulation, *Fourier-SPoC (f-SPoC)*. F-SPoC optimizes the co-modulation between a signal and a sine function modulated by a given beat frequency to find spatial components that show rhythmic entrainment with that beat. The main advantage of our approach is that no a-priori anatomical knowledge of the brain is required for the application.

Together with Spatio-Spectral Decomposition (SSD), we apply f-SPoC to MEG data from a passive beat listening experiment where participants were listening to regular tones with an inter-onset-interval of 400 ms thus a beat frequency of 2.5 Hz.

### Main Findings

We compared the results from the SSD+f-SPoC pipeline to a sensor-level analysis for which we performed a cluster-based permutation analysis (CBPA) to detect significant sensors. Time frequency results averaged over the 78 selected significant sensors show a periodic pattern, especially across 6-18 Hz with a percentage power change of ±1% (see Figure 4.1).

Based on previous work, we first applied f-SPoC on frequency bands of alpha, low and high beta. Group averages of the first and second SPoC component show clear periodic patterns for alpha and low beta with a percentage power change of ±2.5%. Interestingly, periodic patterns across 13-17 Hz are even clearly visible when SPoC is optimized on alpha (8-12 Hz , see Figure 4.2). We therefore combined alpha and low beta and optimized SPoC for 8-20 Hz which resulted in even clearer periodic patterns with a power change of ±5%, on group average (see Figure 4.3) which is much stronger than CBPA. For both methods, f-SPoC and CBPA, highest troughs and peaks can be seen at around 100-150 and 350 ms after stimulus onset which indicates that they are both picking up the same entrainment processes. For some participants, patterns are even visible on the individual-level (see Figure 4.4).

In order to verify that results are reliable, meaningful and not a consequence of overfitting, we also applied the same SPoC component (optimized on 8-20 Hz) to the resting state data. TFR there indeed shows no periodic pattern and is also much weaker in power change (±2%). When SPoC is optimized and applied on resting state data for 8-20 Hz, TFR also does not show periodic patterns as expected. We therefore assume that f-SPoC-results on the beat listening data contains meaningful patterns caused by rhythmic entrainment with the isochronous auditory stimulus.

Comparing individual modulation depth (peak-to-peak differences in ERS/ERD) and modulation power (FFT of TFR at the beat frequency) for CBPA and f-SPoC, results in significantly higher values for f-SPoC. A stronger modulation depth means that the difference in power between the highest peak (ERS) and the strongest trough (ERD)

is higher for f-SPoC than for CBPA. Higher modulation power at the beat frequency signifies a higher correlation with the beat which is what we were optimizing with f-SPoC. Thus, f-SPoC shows clearer and stronger periodic patterns in the combined alpha and beta power than selected sensors from the CBPA.

We further investigated the dynamics across the 5-minutes beat listening experiment to find out what changes inside the brain in the course of this rather monotone paradigm. Event-related fields show a clear decrease in amplitude already after the first and second trial as expected according to the literature. Optimizing f-SPoC on the first 100 trials ($\sim$ 40 seconds) and applying it to the remaining data shows that the frequency range where periodic activity in f-SPoC is visible changes from a combined alpha and beta band to only alpha ($\sim$ minute 3) and only beta towards the end of the experiment. As CBPA shows a different behaviour it needs be further investigated on how the oscillations evolve over time. It would particularly be interesting to use shorter trains of trials with repeated trains in order to investigate the attention of the participants and a possible correlation between entrainment and the alpha/beta ratio in f-SPoC.

### Limitations and Outlook

Most important hyperparamters in the SSD+f-SPoC pipeline are the number of SSD and SPoC components (Haufe et al., 2014a) and the frequency band of interest. As we see in Figure 4.2, a dominant frequency band can still be visible even if SSD and f-SPoC were not optimized on it, however this might not always be the case. When activity is present in both the inner and the outer band, relevant activity can also be suppressed in both and thus not be visible in the frequency band of interest so that prior knowledge about that might be necessary in finding the respective components. Furthermore, both methods are based on eigendecompostions and thus sensitive to outliers. The first components can therefore also show artifacts and relevant activity is more prominent in the following components as we have seen in Figure 4.2.

Even though anatomical knowledge about the brain itself is not necessary to apply this pipeline, it remains crucial to carefully evaluate the results.

Temporal dynamics have shown to influence power and distribution in the frequency space which needs to be accounted for in both the experimental paradigm and the analysis. Also, individuals might not show clear patterns, it is therefore important to record a large enough group of individuals to get meaningful results, otherwise one might miss interesting findings.

Future research could focus on a more advanced approach to select the correct phase of the target function, searching over a grid might not be the optimal solution, also different periodic target functions with a shape closer to the ERD/ERS effects in the entrainment could be investigated. So far, we have only applied f-SPoC to a particular steady beat of 2.5 Hz. Future studies could also experiment with different beat frequencies and explore more complex rhythms.

This approach is by no means restricted to MEG or brain data in general. It could easily be applied to EEG, therefore no conceptional changes would be necessary. It might also be interesting to apply this pipeline to other periodic neurophysiological signals, such as the heart beat, to find correlating components.

# 5

# Fundamentals 2: Natural Language Processing

Natural language processing (NLP) can be broadly defined as the processing of human language with computers. In today's research, NLP combines the fields of linguistics and machine learning in order to investigate natural languages on a large scale. With NLP, we train algorithms to understand and analyse natural language to solve tasks such as question answering, entity recognition, relation extraction or machine translation. Meaningful representation of language is therefore crucial to successfully apply NLP methods and it plays an important role in current research. In this chapter, we will discuss different representations of words and what kind of challenges one might encounter.

## 5.1 Vectorial representation of words

In order to process language on a word-based level, individual words are usually translated into a vectorial representation within a predefined vocabulary. How this translation process takes place depends on the algorithm and the application, as does the definition of the vocabulary or the dimensionality of the vectors. Those vectorial representations are mostly referred to as *word embeddings* and are used as input vectors, e.g. in machine learning models but also enable linguistic investigation in the humanities and other domains. In the following, we will discuss two key algorithms on how to derive word embeddings from text.

### 5.1.1 One-Hot Encodings

We consider a vocabulary $\mathcal{V}$ where each word is identified by an index $i$. One-hot encodings are then stored in unit vectors of length $V$, the number of entries in the vocabulary, where all entries are 0 except for the $i$-th entry which is 1. Words can thus be stored in a matrix $U \in \mathbb{R}^{V \times N}$ where $N$ is the number of words that occur in a document and can be used as input vectors, for instance in neural networks or other mathematical models. This representation is easy to calculate but also very limited when investigating semantic or syntactic similarities within natural languages.

For example, the cosine similarity between two vectors $u_i, u_j$

$$cos(u_i, u_j) = \frac{u_i^\top u_j}{\|u_i\| \, \|u_j\|} \in [-1, 1] \tag{5.1}$$

measures the angle between $u_i$ and $u_j$ and is thus often used as a similarity measure where values close to 1 signify high similarity. When we apply this measure to one-hot encodings, this would always result in 0, unless $u_i = u_j$

Furthermore, the sparse representations cause large matrices when entire documents are stored and is thus computationally costly.

One-hot encodings are therefore of limited use when investigating semantic or syntactic relationships in natural languages. Distributed representations as we will discuss in the following section can be stored in smaller matrices and carry semantic and syntactic information which makes them more suitable for many applications in the field of NLP.

### 5.1.2 Distributed word embeddings

Distributed representations as introduced by Bengio et al. (2003) depend on local context defined by the words surrounding a target word. This is based on the assumption that words that occur in similar contexts also share semantic and grammatical properties. We will briefly recap GloVe which, next to e.g. word2vec (Mikolov et al., 2013b) depicts one algorithm that trains distributed word embeddings.

The authors of GloVe which stands for **Glo**bal **Ve**ctors (Pennington et al., 2014) claim to combine the advantages of global matrix factorization (Deerwester et al., 1990) and the mentioned local context windows to learn vectorial representations. Thus, word vectors are trained based on the global co-occurrence matrix that contains word-word co-occurrence counts over the entire corpus. The following optimization function is therefore minimised

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^\top \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2 \tag{5.2}$$

with

$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise.} \end{cases} \tag{5.3}$$

Here, $V$ is again the length of the considered vocabulary and $X_{ij} \in X \in \mathbb{R}^{V \times V}$ denote the entries of the deterministic co-occurrence matrix, i.e. the number of times word $j$ occurs in the context of word $i$. The vectors $w_i \in \mathbb{R}^d$ and $\tilde{w}_j \in \mathbb{R}^d$ represent the d-dimensional word and context vector of words $i$ and $j$ respectively and are parameters we optimise during training. Biases for both words are added for reason of symmetry as $b_i$ and $\tilde{b}_j$. The weighting function $f(x)$ is introduced to down-weight rare occurrences where the authors suggest to set $x_{\max} = 100$ and $\alpha = 0.75$. For a symmetric co-occurrence matrix, $w_i$ and $\tilde{w}_i$ only differ in terms of random initialization and can therefore be summed up afterwards to get a more stable word vector.

Word embeddings often serve as preprocessing or pretraining for downstream tasks such

as text classification, language understanding or question answering. Distributed word embeddings such as the ones trained by GloVe can also be used for linguistic analyses as they highly depend on the training corpus. For instance, the authors of Garg et al. (2018) investigate ethnic and gender stereotypes by training word embeddings on specific corpora and evaluate which words are considered most similar to different ethnic groups or genders.

### 5.1.3 Contextualized word embeddings

Large machine learning models such as Bidirectional Encoder Representations from Transformers (BERT, Devlin et al., 2019) go one step further and build multiple instances of each word based on its local context. Those models are trained on huge training corpora with millions of parameters which then will be finetuned on a specific task.

In this work however, we will focus on approaches where we are able to learn embeddings from scratch also on smaller, specific datasets to mitigate the influence from large unknown training datasets.

## 5.2 Evaluation on analogy tests

The quality of word embeddings is often measured by a set of semantic and syntactic questions as for instance proposed by Mikolov et al. (2013a). Distributed word embeddings as described in the previous section are expected to contain syntactic and semantic similarities in their vectorial representation which are often evaluated based on word similarities, also called analogies. For example, the word embedding of *Paris*, the capital of France, is assumed to be close to both embeddings *France* and *Abuja*, the capital of Nigeria.

A possible semantic analogy would thus be that *Paris* corresponds to *France* as *Abuja* to *Nigeria*.

$$Paris : France :: Abuja : Nigeria$$

This can be validated by simple vector arithmetic where we consider the word embeddings $u_i$ of all 4 words and check whether

$$u_{Paris} - u_{France} + u_{Nigeria} \approx u_{Abuja.}$$

An example for a syntactic analogy is

$$amazing : amazingly :: happy : happily$$

which is evaluated in the same way as the example above.

## 5.3 Limitations and Challenges

**Challenge 1: Dynamic changes over time**

Word embeddings as introduced here are highly dependent on the corpus they are trained on which can cause biases for example with respect to historic context. Depending on the application, one might aim to include this bias in their analysis, e.g. when investigating historic language shifts between different corpora. In other applications, this bias might be neither intended nor detected. For instance, when word embeddings were trained on corpora where all doctors are female and all nurses are male, an algorithm might prefer a female over a male applicant to fill a position for a physician only based on their gender.

Language and the usage of words as well as their meaning evolve over time and depend on their context. For instance, the word *virus* in the Oxford Learners Dictionary is defined as[1]

1. a living thing, too small to be seen without a microscope, that causes infectious disease in people, animals, and plants
2. *(informal)* a disease caused by a virus
3. instructions that are hidden within a computer program and are designed to cause faults or destroy data

It can thus be used in the context of computers but also in the context of biology. Furthermore there have been and probably will be different infectious diseases caused by viruses and in terms of the COVID-19 pandemic, we currently most probably mean a different virus as we have in 2009 during the H1N1/09 pandemic.

Those shifts in time and domain need to be considered when training word embeddings. In case of a single embedding, a word like *virus* might contain multiple meanings which cannot be separated within one word vector. Recently several methods have been proposed to train individual word embeddings for different epochs or domains to detect semantic and syntactic changes, see e.g. Azarbonyad et al. (2017) and Kutuzov et al. (2018). For most approaches however embeddings need to be trained and aligned afterwards which might induce additional noise as we will discuss in the following section.

**Challenge 2: Alignment** In order to be able to compare multiple instances of word embeddings, embedding matrices need to be transformed in one common vector space, they need to be aligned. One solution would be to apply the orthogonal Procrustes problem (Schönemann, 1966)

$$\text{argmin}_R \left\| RU - U' \right\|_F \text{ subject to } R^\top R = I$$

[1] https://www.oxfordlearnersdictionaries.com/definition/american_english/virus

where $U$ and $U'$ are two embedding matrices and $R$ the matrix that maps $U$ to $U'$. This can be solved by singular value decomposition of $U'U^\top$ (Zhang, 2000).

Although this alignment works well in many cases there is always the risk of inducing additional numerical noise on the word embeddings. It might be beneficial to exchange information between word embeddings during the training so this way, embeddings would be aligned more naturally in the process and smaller datasets could benefit from the information exchange. So far, three different approaches for such dynamic word embeddings have been proposed, all of them investigate only temporal changes. In Chapter 6, we will discuss them and extend this approach with a new method, *Word2Vec with Structure Constraint*, so that language shifts with respect to other data structures can be investigated. Furthermore, we will propose *Word2Vec with Structure Prediction*, which additionally predicts the underlying structure of a dataset while simultaneously training and aligning word embeddings for each data slice.

**Challenge 3: Low quality of word embeddings trained on small datasets** Another disadvantage when training word embeddings individually and align them afterwards is a decrease in embedding quality as we will show in Chapter 6. Multiple embeddings trained within one framework can benefit from information being exchanged between sub-corpora and thus might show a higher embedding accuracy e.g. in word similarity tests compared to embeddings trained individually for each sub-corpora. We will therefore compare embeddings trained by *Word2Vec with Structure Constraint* and *Word2Vec with Structure Prediction* with embeddings trained by GloVe.

# 6

# Dynamic Word Embeddings with Structure Prediction

## 6.1 Motivation

Static word embeddings as discussed in Section 5.3, cannot reflect semantic changes as occur between different domains, time periods or regional differences (e.g. South African vs. American English).

Various approaches to track, detect and quantify semantic shifts in text have been proposed where typically first individual static embeddings for different time stamps are trained and then aligned afterwards (see e.g. Kulkarni et al., 2015; Hamilton et al., 2016). The authors of Kutuzov et al. (2018) identify three communities where investigating those shifts is interesting: computational linguistics, information retrieval and political science. We also want to refer to Tahmasebi et al. (2018) for another very detailed survey where most of the mentioned articles and many more have been discussed and set in context.

Although most research focuses on temporal change (Kim et al., 2014; Kulkarni et al., 2015; Hamilton et al., 2016; Zhang et al., 2016; Marjanen et al., 2019) others have also considered more general approaches (Azarbonyad et al., 2017; Gonen et al., 2020) or focused on specific domains such as social media (Zeng et al., 2017; Shoemark et al., 2019).

With the rise of larger language models such as Bidirectional Encoder Representations from Transformers (BERT, Devlin et al., 2019) and with that contextualized embeddings, a part of the research question has shifted towards detecting language change in contextualized word embeddings, (see e.g. Jawahar et al., 2019; Hofmann et al., 2020). Since BERT trains one word vector for each context it occurs, in order to consider dynamic word embeddings over a specific category it is necessary to first extract relevant word embeddings out of the many.

At least three different algorithms have been introduced to train dynamic word embeddings in a common vector space which means embeddings do not need to be aligned afterwards. The authors of Bamler et al. (2017) propose a Bayesian approach that extends the skip-gram model (Mikolov et al., 2013a) to learn dynamic word embeddings that evolve over time. Their method is, as is the skip-gram model, based on matrix factorization while Rudolph et al. (2018) analyse dynamic changes in word embeddings based on exponential family embeddings, a probabilistic framework that generalizes the

concept of word embeddings to other types of data (Rudolph et al., 2016). Both publications focus on word-level changes within and between text corpora spanning from the 19th century until today. The authors of Yao et al. (2018) propose Dynamic Word2Vec where word embeddings for each year of the New York Times corpus are trained based on individual positive point-wise information matrices and aligned simultaneously.

We argue that apart from diachronic word embeddings there is a need to train dynamic word embeddings that not only capture temporal shifts in language but for instance also semantic shifts between domains or regional differences. It is therefore important that those embeddings can be trained on small datasets and are not influenced by huge training datasets as in BERT.

In this chapter, we will propose two generalizations of Dynamic Word2Vec: *Word2Vec with Structure Constraint (W2V&Constr)* and *Word2Vec with Structure Prediction (W2V&Pred)*. With W2V&Constr, we are able train dynamic word embeddings based on a neighbourhood graph that is fed into the model. Word embeddings are aligned in the process and smaller datasets profit from the information exchanged between datasets. With W2V&Pred, we go one step further and besides training dynamic word embeddings, we also predict the structure within the dataset which is especially useful when no informed decision can be made regarding the neighbourhood graph. We will apply both methods to the *New York Times* dataset, two *Wikipedia* datasets that we assume to be structured as a tree based on their categories and data from the *Deutsches Textarchiv*, where we predict the connections between different German authors from the 18th and 19th century. We can show that the quality of word embeddings improves significantly compared to static embeddings trained by GloVe and that W2V&Pred can give interesting insights into data structures also with respect to outliers.

With our methods we are thus able to detect language change on word level, train domain-specific word embeddings in a common vector space and gain interesting insights into the structure of datasets as we will show in this chapter.

The remaining part of the chapter is organized as follows: In Section 6.2, we will recap Dynamic Word2Vec and present W2V&Constr and W2V&Pred followed by Section 6.3 where we describe datasets, experiments and evaluation metrics. Results will be presented in Section 6.4 before we conclude this chapter with a discussion in Section 6.5.

This chapter follows our own work from Brandl et al. (2019) which is currently in preparation for journal submission Lassner et al. (2021).

## 6.2 Methods

We will first discuss Dynamic Word2Vec and then extend this approach with two new models: *Word2Vec with Structure Constraint* and *Word2Vec with Structure Prediction*.

### 6.2.1 Dynamic Word2Vec (DW2V)

Dynamic Word2Vec (DW2V) proposed by Yao et al. (2018) trains diachronic word embeddings while regularizing the embedding matrix and adding a neighbourhood constraint to keep previous and following neighbours close. The optimization function for a specific time point $t$ is thus defined as

$$\min_{U_t} \frac{1}{2} \left\| Y_t - U_t U_t^\top \right\|_F^2 \tag{6.1}$$

$$+\frac{\tau}{2} \left( \|U_{t-1} - U_t\|_F^2 + \|U_t - U_{t+1}\|_F^2 \right) \tag{6.2}$$

$$+\frac{\lambda}{2} \|U_t\|_F^2 \tag{6.3}$$

where $U_t \in \mathbb{R}^{V \times D}$ contains $D$-dimensional embeddings for all words of the vocabulary of length $V$ and $Y_t \in \mathbb{R}^{V \times V}$ represents the PPMI matrix at time point $t$.

The positive pointwise mutual information (PPMI) between a word $w$ and a context word $c$ in a corpus $D$ has been proposed by Levy et al. (2014) and is defined as

$$\begin{aligned} \text{PPMI}(w, c) &= \max(\text{PMI}(w, c), 0) \\ \text{with } \text{PMI}(w, c) &= \log \left( \frac{\#(w, c) \cdot |D|}{\#(w) \cdot \#(c)} \right) \end{aligned}$$

where $\#w$ is the number of times a word occurs in a document and $\#(w, c)$ the number of times the words $w$ and $c$ co-occur within a given context window. $|D|$ is the total number of words in a document. The PPMI matrix thus contains the logarithmic and normalized co-occurrence counts for a fixed context window where negative or very unlikely co occurrences with $\#(w) \cdot \#(c) \ggg \#(w, c)$ are set to 0.

### 6.2.2 Word2Vec with Structure Constraint (W2V&Constr)

We will extend DW2V by reformulating (6.2)

$$\begin{aligned} &\|U_{t-1} - U_t\|_F^2 + \|U_t - U_{t+1}\|_F^2 \\ &= \sum_{t' \neq t}^{T} w_{t,t'} \left( \|U_t - U_{t'}\|_F^2 \right) \\ \text{with } w_{t,t'} &= \mathbb{1}_{\{|t-t'|=1\}}. \end{aligned} \tag{6.4}$$

This gives us the opportunity to generalize DW2V for different neighbourhood structures by including any matrix in the form of $w^{\text{fix}} \in \mathbb{R}^{T \times T}$ where $T$ is the total number of data slices. Instead of a chronological sequence, $w^{\text{fix}}$ could represent any weighted, undirected graph as the underlying structure of the dataset. We therefore define the pairwise distance matrix $D$ where $D_{t,t'} = \|U_t - U_{t'}\|_F^2$.

Furthermore, we adapt the regularization term (6.3) and regularize for the pairwise

Euclidean distances in $D$. This way, we make sure that the distances between different embedding matrices stay close and do not jump between extreme values.

Together, this yields

$$\min_{U_t} F_t = \min_{U_t} \frac{1}{2} \left\| Y_t - U_t U_t^\top \right\|_F^2 \tag{6.5}$$

$$+\tau \sum_{t' \neq t}^{T} w_{t,t'}^{\text{fix}} \cdot D_{t,t'} \tag{6.6}$$

$$+\lambda \left\| D \right\|_F \tag{6.7}$$

$$\text{with } D_{t,t'} = \|U_t - U_{t'}\|_F^2$$

We call this method *Word2Vec with Structure Constraint (W2V&Constr)* and with this generalization of DW2V, we can apply W2V&Constr to any kind of dataset where we can make an informed assumption about the underlying structure to train word embeddings while simultaneously aligning them.

For datasets where we do not know the structure, we want to propose *Word2Vec with Structure Prediction*.

### 6.2.3 Word2Vec with Structure Prediction (W2V&Pred)

*Word2Vec with Structure Prediction (W2V&Pred)*, requires no assumption about the underlying structure of the dataset but learns the structure together with the corresponding word embeddings in an unsupervised manner. We thus further generalize W2V&Constr by also learning $w$ with the following optimization function

$$\begin{aligned} \min_{U_t} F_t = \min_{U_t} & \frac{1}{2} \left\| Y_t - U_t U_t^\top \right\|_F^2 \\ & +\tau \sum_{t' \neq t}^{T} w_{t,t'}^{*} \cdot D_{t,t'} \\ & +\lambda \left\| D \right\|_F \end{aligned} \tag{6.8}$$

$$\begin{aligned} \text{s.t. } D_{t,t'} &= \|U_t - U_{t'}\|_F^2 \\ w_{t,t'} &\leftarrow \frac{1}{D_{t,t'}} \text{ which will be normalized with} \\ w_{t,t'}^{*} &\leftarrow \frac{w_{t,t'}}{\sum_{t''} w_{t,t''} + \sum_{t''} w_{t'',t'}} \end{aligned} \tag{6.9}$$

With the neighbourhood constraint in (6.8), we minimize the pairwise distance between embedding matrices and also transfer information between them while aligning matrices at the same time. The matrix $w$ serves as a weighting term that strengthens connections between close embeddings.

### 6.2.4 Optimization

The PPMI matrix $Y \in \mathbb{R}^{T \times V \times V}$ for all slices can become very large and thus memory consuming, that is why we did not optimize over all slices in one step but updated after each slice $t$. We define one epoch as complete when $U_t$ has been updated for all $t$. We applied gradient descent with Adam (Kingma et al., 2014) with default values for $\beta$ and a learning rate of 0.1. The learning rate has been reduced after 100 epochs to 0.05 and after 500 epochs to 0.01 with a total number of 1000 epochs. Both models have been implemented in PyTorch.

## 6.3 Experiments

We ran experiments on three different datasets for W2V&Constr and W2V&Pred with an additional explorative experiment for W2V&Pred. In the following, we will describe all datasets, the preprocessing procedure, evaluation metrics, the baseline analyses and the parameter selection. Throughout this chapter, we will use adjacency matrices to define distances between data slices. For instance, the distance between a node and a sub-node in a tree would be 1 while the distance between two sub-nodes belonging to the same node is 2 (see middle and right panel in Figure 6.1).

Furthermore, we conducted toy data experiments for three different scenarios to further evaluate W2V&Pred, see Appendix A

### 6.3.1 Datasets

#### New York Times

The *New York Times* dataset[1] (NYT) contains 100.945 news articles including metadata such as headlines, lead texts and paragraphs. Articles have been published online and offline between January 1990 and June 2016. We group the dataset by years and define the years 1990-1998 as the train set and 1999-2016 as the test set.

In Figure 6.1 (left panel), we show the structure of the New York Times dataset in form of an adjacency matrix where we only consider the previous and following years as neighbours

#### Wikipedia Field of Science (WikiFoS)

We selected categories of the OECD's list of Fields of Science and Technology[2] and downloaded the corresponding articles from the English Wikipedia. The resulting dataset Wikipedia Field of Science (WikiFoS) contains four clusters with one main category and three subcategories each, in total 226.386 unique articles. Articles that appear in multiple categories have been randomly assigned to only one category. We assume

[1] https://sites.google.com/site/zijunyaorutgers/
[2] http://www.oecd.org/science/inno/38235147.pdf

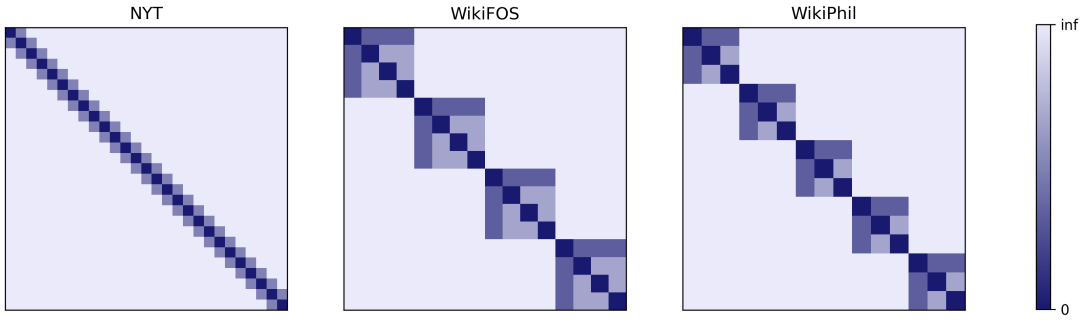


Figure 6.1: True adjacency matrices for NYT, WikiFoS and WikiPhil. On the x and y axis are the corresponding years/categories. The values indicate distance thus pairs with small values are considered as close. For NYT, we are only considering the previous and following year. For WikiFoS and WikiPhil, we take into account the corresponding cluster where main categories are closer to subcategories than subcategories to each other.

| Category WikiFoS | nArticles |
|---|---|
| **Natural Sciences** | 8472 |
| Chemistry | 19120 |
| Computer Science | 11217 |
| Biology | 11034 |
| **Engineering & Technology** | 20009 |
| Civil Engineering | 17884 |
| Electrical & Electronic Engineering | 6825 |
| Mechanical Engineering | 4931 |
| **Social Sciences** | 17600 |
| Business & Economics | 14760 |
| Law | 13290 |
| Psychology | 5748 |
| **Humanities** | 14860 |
| Literature & Languages | 24765 |
| History & Archaeology | 16442 |
| Religion & Philosophy & Ethics | 19429 |

| Category WikiPhil | nArticles |
|---|---|
| **Logic** | 3437 |
| Concepts in Logic | 1419 |
| History of Logic | 75 |
| **Aesthetics** | 7324 |
| Philosophers of Art | 35 |
| Literary Criticism | 3803 |
| **Ethics** | 5804 |
| Moral Philosophers | 152 |
| Social Philosophy | 3847 |
| **Epistemology** | 3209 |
| Epistemologists | 357 |
| Cognition | 8526 |
| **Metaphysics** | 1810 |
| Ontology | 781 |
| Philosophy of Mind | 1024 |

Table 6.1: Categories in both Wikipedia datasets and the number of articles. *Left:* Main categories are marked in bold, the following 3 define the corresponding subcategories. Fields joined by & originate from multiple categories in Wikipedia but were joined for this dataset. *Right:* All categories are subcategories of *Philosophy*, the first category in each group of 3 (in bold) is a direct (first-order) subcategory of *Philosophy*, the following 2 are second-order subcategories.

the underlying structure of this dataset to be arranged as a tree where each node corresponds to one field of science. The nodes in the first level contain the four main categories followed by the corresponding subcategories. We split this dataset randomly in train and test set by sampling 1/3 of the articles in each category as the train set and the remaining 2/3 as the test set.

#### Wikipedia Philosophy (WikiPhil)

We selected five first-order subcategories of *Philosophy* (WikiPhil) with two second-order subcategories each and downloaded the articles from the English Wikipedia, 41.603 unique articles in total. As before, duplicates were assigned randomly to one category and articles were sampled randomly so that 1/3 of the articles of each category define the train set and the remaining articles define the test set. The underlying structure is also assumed to be a tree with five nodes on the first and ten nodes on the second level, two nodes following each first-level node.

#### Deutsches Textarchiv (DTA)

We extracted fiction texts (*Belletristik*) from the lemmatized version of DTA (Deutsches Textarchiv), containing the 20 most represented authors in 1770-1900, 10.219.336 lemmas in total. We applied W2V&Pred in an exploratory manner to predict the connections between those authors without assuming or knowing any structure. This serves as a real-world application which might be of interest for the Digital Humanities community.

### 6.3.2 Preprocessing

We lemmatized all tokens, i.e. assigned their base forms, in each dataset with the python package spacy[3] except for DTA where we downloaded already lemmatized data. We then grouped the data accordingly, by years (NYT), authors (DTA) and categories (WikiPhil, WikiFoS) and define individual vocabularies for each dataset. Vocabularies consist of the 20.000 most frequent words used in each dataset where trend words that occur in less than 3 slices were filtered out. Furthermore, we filtered out stop words, i.e. the 100 most frequent words used. We set the symmetric context window to 5, i.e. five words before and after a specific word are defined as its context and considered for the PPMI matrix.

### 6.3.3 Evaluation Metrics

We evaluated the quality of the word embedding and the structure with different evaluation metrics that we will describe in the following sections.

[3] https://spacy.io

### Analogies

We evaluated the word embedding on two different test sets of questions. The first set contains general analogies of 4 words as described in Section 5.2 and published by Mikolov et al. (2013a) where we applied simple vector calculations on the three words and then checked which word in our vocabulary comes closest to the resulting vector. We therefore considered a set of $n$ next neighbours of the resulting vector to verify whether the fourth word of the analogy is contained. Since those analogies are not domain-specific and quite general, we call them *global analogies*.

The authors of Yao et al. (2018) proposed a test set with temporal analogies where a year and a term is given, for instance the president of the United States at that specific year, and the counterpart for another year needs to be found.

$$\textit{1992 : Clinton :: 2014 : Obama}$$

We call those analogies *temporal analogies* and apply them to NYT.

### Recall@k

Let $D \in \mathbb{R}^{T \times T}$ be the true adjacency matrix, as shown in Figure 6.1 and $\hat{D}$ the distance matrix of the learned embeddings. We evaluated the structure with recall@k defined as

$$\text{recall@k} = \frac{1}{T} \sum_{t}^{T} \text{recall@k}_t \tag{6.10}$$

$$\text{where} \quad \text{recall@k}_t = \frac{\text{TP}_t(k)}{\text{TP}_t(k) + \text{FN}_t(k)} \tag{6.11}$$

$$\text{TP}_t(k) = \sum_{t'}^{T} \mathbb{1}_{\{D_{tt'} \in D_{t:k}\}} \cdot \mathbb{1}_{\{\hat{D}_{tt'} \in \hat{D}_{t:k}\}} \tag{6.12}$$

$$\text{FN}_t(k) = \sum_{t'}^{T} \mathbb{1}_{\{D_{tt'} \in D_{t:k}\}} \cdot \mathbb{1}_{\{\hat{D}_{tt'} \notin \hat{D}_{t:k}\}} \tag{6.13}$$

Here, $\mathbb{1}$ represents the indicator function, $D_{tt'}$ and $\hat{D}_{tt'}$ the elements of the distance matrices and $D_{t:k}, \hat{D}_{t:k}$ the first k elements of the sorted rows in $D, \hat{D}$. Thus, (6.12) counts the true positives, the correctly found nearest neighbours whereas (6.13) gives back the false negatives, the nearest neighbours that have not been identified.

We chose $k = 2$ for the NYT dataset, considering only the following and previous year in the dataset. For WikiPhil and WikiFoS, we chose $k = 2$ and $k = 3$, considering only the members of the corresponding subtree.

### 6.3.4 Baseline Analysis

We trained embeddings with GloVe (Pennington et al., 2014) in order to compare them to embeddings trained by W2V&Constr and W2V&Pred. Furthermore, we computed Burrows' Delta based on word occurrences in comparison to the structure scores based on embedding matrices. We also applied DW2V to the New York Times dataset.

#### GloVe

We compared the learned embeddings by W2V&Pred and W2V&Constr with static embeddings trained with GloVe (Pennington et al., 2014, see Section 5.1.2). We therefore set $x_{\max} = 100$ and $\alpha = 0.75$ as recommended by the authors. We chose a symmetric window size of 5 as we did for our methods. One set of word embeddings as well as individual embeddings for each slice were trained on all datasets. The minimum occurrence for a word to be considered was set to 10 (individual) and 25 (entire dataset) respectively. For the individual embeddings we calculated the accuracy on the temporal analogies, recall@k and Burrows' delta. For the embeddings trained on the entire dataset, we calculated the accuracy on the global analogies. We therefore used the same vocabulary as for W2V&Constr/W2V&Pred. Words that have not been trained on the individual slice because of too few occurrences were replaced by the embedding learned on the entire dataset.

#### Burrows' Delta

Burrows' Delta is a measure usually used for authorship attribution (Burrows, 2002; Argamon, 2008) and can be defined by

$$\Delta(D, D') = \left\| z(f_i(D)) - z(f_i(D')) \right\|_2$$
$$\text{where } z(f_i(D)) = \frac{f_i(D) - \mu_i}{\sigma_i}.$$

Here, $D$ and $D'$ are two documents, $f_i(D)$ is the number of occurrences of word $i$ in document $D$ and $z(f_i(D))$ the z-transformation where $\mu_i$ and $\sigma_i$ are mean and standard deviation of the occurrence with respect to a reference corpus.

Burrows' Delta measures how similar documents are based on word occurrences. We used it here as a baseline method to compute a neighbourhood graph on which we then computed recall@k.

For our experiments, we defined the entire train or test set as the reference corpus and one document as the data from one slice, i.e. one year, category or author.

Parameters for the minimum and maximum occurrence of words were selected on the train set based on the best recall@k score.

### 6.3.5 Parameter Selection

Hyperparameters $\lambda$ and $\tau$ for DW2V, W2V&Constr and W2V&Pred were selected on the train set of each dataset based on the best accuracy on the global analogy tests. More specifically, for $n = 1, 5$ and 10 accuracies of all runs were scaled for each model to $[0, 1]$ and averaged, this score defines the embedding accuracy. The selected parameters were then applied on the test set.

For W2V&Constr, we define $w^{\text{fix}}$, see (6.6), as

$$w^{\text{fix}}_{t,t'} = \begin{cases} 0 & \text{for } t = t' \\ 1 - \frac{d_{t,t'}}{\max(D)} & \text{else} \end{cases}$$

where $d_{t,t'} \in D$ are the elements of the true adjacency matrix and $\max(D)$ is the maximum element in $D$.

## 6.4 Results

For all methods and English datasets except for GloVe, embeddings have been averaged across slices and then evaluated on the global analogy tests. For GloVe, we additionally trained one set of embeddings on the entire dataset and applied global analogy tests. Hyperparameters for the test runs were selected based on the best embedding score in the train runs for DW2V, W2V&Constr and W2V&Pred. All test runs were repeated 10 times and accuracies averaged afterwards except for Burrows' Delta which is a deterministic method and therefore has only been computed once. We selected the best method for each dataset separately for the embedding score (marked as **bold***) to compare them to all other methods and test for significance with a Wilcoxon signed rank test ($\alpha = 0.05$). For the structure score, we only compare W2V&Pred to GloVe as all other methods had prior knowledge about the structure. The results that were not significantly outperformed are also marked in **bold**, see Tables 6.2, 6.4. Accuracies within the text are rounded to the first decimal for better readability.

### 6.4.1 New York Times

In Table 6.2, we show results for all NYT experiments, including GloVe and DW2V. We can see that for the global analogies, all methods, except GloVe achieve similar results with differences of $< 2\%$ for the highest and lowest accuracies. Here, DW2V and W2V&Constr achieve comparable results with accuracies of $10.5 - 43.3\%$. DW2V significantly outperforms GloVe and W2V&Pred for all $n$ and W2V&Constr for $n = 1$.

Temporal analogy tests were evaluated on the individual embeddings where W2V&Pred significantly outperforms all other methods with accuracies of $41.9 - 69.7\%$ by a margin of around 1% followed by DW2V($39.5 - 67.4\%$) and W2V&Constr ($38.2 - 65\%$). GloVe again only reaches accuracies of $6.7 - 18\%$.

| method | $\lambda$ | $\tau$ | global analogy tests | | | temporal analogy tests | | | recall@k |
|---|---|---|---|---|---|---|---|---|---|
| | | | n=1 | n=5 | n=10 | n=1 | n=5 | n=10 | |
| GloVe | | | 3.54 | 18.83 | 26.04 | 6.70 | 14.03 | 17.96 | 64.17 |
| DW2V | 0.5 | 65536 | **11.27*** | **32.88*** | **42.97*** | 39.47 | 61.94 | 67.35 | 100 |
| W2V&C | 2048 | 2 | 10.54 | **32.64** | **43.31** | 38.23 | 57.29 | 64.96 | 69.72 |
| W2V&P | 512 | 2048 | 10.28 | 31.66 | 41.88 | **41.87*** | **64.60*** | **69.67*** | **81.67*** |
| Burrows' | | | | | | | | | 55.56 |

Table 6.2: Results for the NYT dataset on DW2V, W2V&Constr and W2V&Pred. All experiments have been repeated 10 times, displayed results show averages, except for Burrows' Delta which is deterministic and has therefore only been computed once. Signed rank tests with $\alpha = 0.05$ have been carried out to test for significance compared to the best method (marked in **bold***), here DW2V for global analogies, W2V&Pred for both structure score and temporal analogies. Accuracies for which the improvement have not been significant are marked in bold. For the structure score, we only compared W2V&Pred to GloVe as all other methods had prior information on the structure.

Burrows' Delta reaches a structure accuracy of 55.6%. W2V&Pred and GloVe which denote the only methods without any prior knowledge about the structure score, achieve accuracies of 81.7% and 64.2%.

Figure 6.2 shows a two-component spectral embedding (Belkin et al., 2002) of the final distance matrix for NYT (test run W2V&Pred). The circular structure shows very nicely that our method finds the chronological order except for 2006 which seems to be distant to all other years.

In order to show the exemplary temporal evolution of one word, we calculated the pairwise distance for the word *blackberry* between years and selected the years with the largest distance: 2000 and 2011. In Figure 6.3, we show the top-5 neighbours for both years and the cosine similarity between each of the 10 neighbouring words and *blackberry* for each year. The graphic shows the change in meaning. At the beginning *blackberry* was closer to other fruits such as raspberry and apple and then it shifted toward technical devices such as *phone* and *smartphone*. After 2013, the similarity between technical terms dropped and at the same time grew again between *blackberry* and fruits. Blackberry besides the fruit is also the name of a Canadian company that sells technical devices such as smartphones. In 2002, the company released their first convergent smartphone[4] which explains the peak in Figure 6.3. Their smartphones became more and more popular until sales numbers dropped after 2011[5] which is also visible in the figure. Interestingly, the word *apple* stays relatively close during the entire time period as the word vector also reflects both meanings, the fruit and the company.

[4] https://www.businessinsider.com/blackberry-smartphone-rise-fall-mobile-failure-innovate-2019-11

[5] https://www.businessinsider.com/blackberry-phone-sales-decline-chart-2016-9

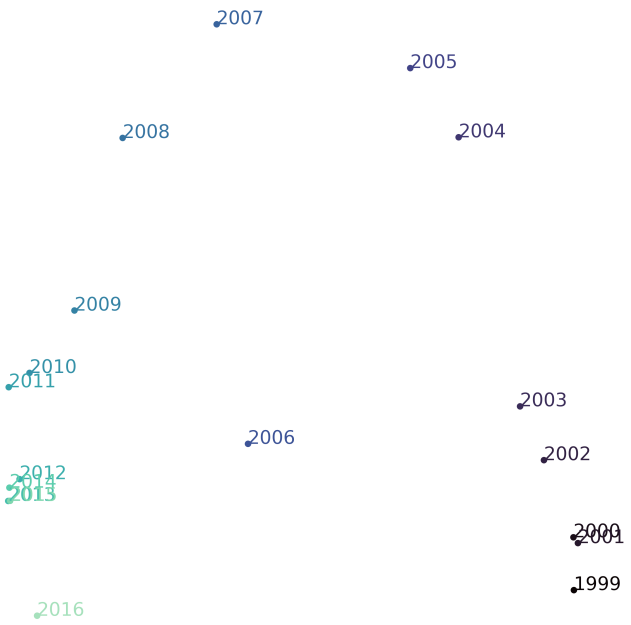


Figure 6.2: Laplacian eigenmaps of the distance matrix of the trained word embeddings for NYT. Points are coloured respective to the chronological order.

In Table 6.3 we displayed the 5 nearest words to *mayor* during epochs of 4 years. We therefore averaged embeddings for each epoch and computed words with the highest cosine similarity. Rudolph Giuliani who was the mayor of New York City (NYC) between 1994 and 2001[6] is present in all time periods. Bill de Blasio the current mayor of NYC since 2014, a former member of the NYC Council (2002-2009) and a former Public Advocate of NYC (2010-2013)[7] first appears in the top-5 in 2007-2010 and is the closest neighbour in 2011-2014 and 2015-2016 which matches his term as mayor.

| **1999-2002** | **2003-2006** | **2007-2010** | **2011-2014** | **2015-2016** |
|---|---|---|---|---|
| rudolph | suozzi | daley | blasio | blasio |
| giuliani | rudolph | giuliani | rudolph | rudolph |
| vallone | giuliani | rudolph | daley | giuliani |
| daley | daley | dinkins | giuliani | cuomo |
| dinkins | councilman | blasio | cuomo | dinkins |

Table 6.3: Five nearest Neighbors to ***mayor*** in different epochs from the NYT dataset. Embeddings were averaged across 4 years. Then the 5 closest words w.r.t. the cosine similarity were extracted. Words that appear only once are underlined.

[6] https://www1.nyc.gov/site/dcas/about/green-book-mayors-of-the-city-of-new-york.page

[7] https://www1.nyc.gov/office-of-the-mayor/bio.page

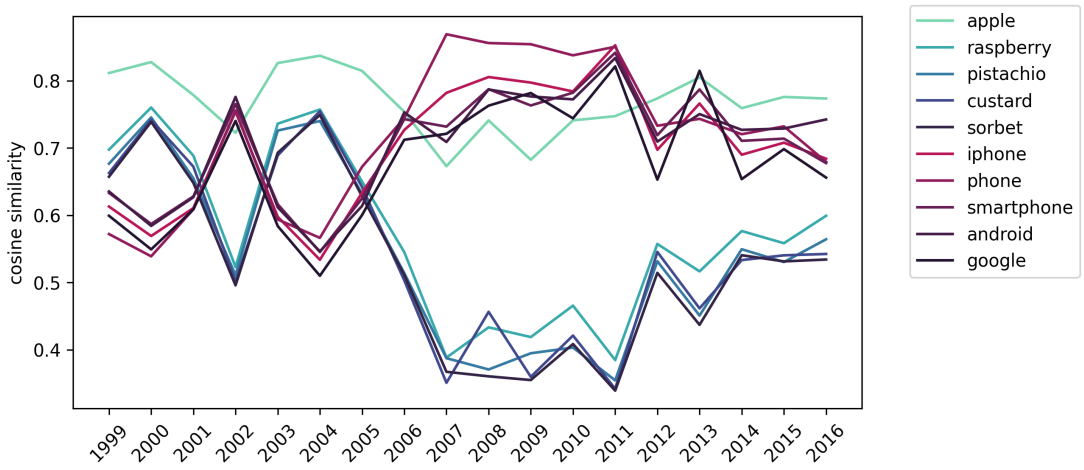


Figure 6.3: Evolution of the word *blackberry* in NYT. Nearest neighbours of the word *blackberry* have been selected in 2000 (blueish) and 2011 (reddish). Cosine similarity between all nearest Neighbors and *blackberry* are plotted over time showing the shift in dominance between fruit and smartphone brand. The word *apple* represents also both, the fruit and the company.

Peter Vallone ran for office of mayor in NYC in 2001[8], Richard M. Daley was mayor of Chicago (1989-2011)[9], David Dinkins was mayor of NYC in 1986-1993 and Thomas Suozzi was mayor of Glen Cove, NY (1994-2001) and has been a US Representative for the 3rd district in NYC since 2017[10]. Andrew Cuomo has been governor of New York since 2011[11]. We thus see a clear relation between epochs and next neighbours.

| dataset | method | $\lambda$ | $\tau$ | global analogy tests | | | recall@k |
|---|---|---|---|---|---|---|---|
| | | | | n=1 | n=5 | n=10 | |
| WikiFos | Glove | | | **12.17** | 43.48 | 52.30 | 46.04 |
| | W2V&C | 512 | 256 | **11.92*** | **46.12*** | **57.06*** | 79.17 |
| | W2V&P | 512 | 1024 | **11.82** | **45.73** | **56.40** | **62.50*** |
| | Burrows' | | | | | | 22.92 |
| WikiPhil | Glove | | | **4.76** | 28.90 | 37.13 | 20.00 |
| | W2V&C | 4096 | 128 | 0.42 | 10.37 | 15.02 | 28.67 |
| | W2V&P | 512 | 512 | **4.37*** | **32.00*** | **41.75*** | **23.33*** |
| | Burrows' | | | | | | 6.67 |
| | restructured | 128 | 64 | 5.96 | 36.21 | 46.15 | 50.00 |

Table 6.4: Results for both Wikipedia datasets. See Table 6.2 for detailed explanation.

[8] https://www.nytimes.com/2001/06/01/nyregion/vallone-run-for-mayor-is-official.html
[9] https://www.chipublib.org/mayor-richard-m-daley-biography/
[10] https://suozzi.house.gov/about/biography
[11] https://www.ny.gov/agencies

| **Natural Sciences** | **Engineering &Technology** | **Social Sciences** | **Humanities** | **GloVe** |
|---|---|---|---|---|
| generator | generator | powerful | powerful | control |
| PV | inverter | control | control | supply |
| thermoelectric | alternator | wield | counterbalance | capacity |
| inverter | converter | drive | drive | system |
| converter | electric | generator | supreme | internal |

Table 6.5: Five nearest neighbours to ***power*** in different field of sciences from the WikiFoS dataset. Embeddings were averaged across the respective main category and its three subcategories. Then the 5 closest words w.r.t. the cosine similarity were extracted. On the right column the nearest neighbor to the embedding trained by GloVe on the entire dataset are shown.

### 6.4.2 WikiFoS

In Table 6.4 we show results for WikiFoS. W2V&Constr reaches the best embedding scores ($11.9 - 57.1\%$) closely followed by W2V&Pred ($11.8 - 56.4\%$). GloVe achieved an accuracy of $12.2 - 52.3\%$.

W2V&Pred reached a structure score of 62.5% which is significantly higher than GloVe with 46%. Burrows' Delta only reached an accuracy of 22.9%.

In Table 6.5 we show the 5 nearest neighbouring words of the word *power* for the 4 main categories in WikiFoS and for the embedding trained with GloVe on the entire dataset. For WikiFoS , we averaged the embeddings within the main category and then selected the words with the largest cosine similarity to *power*. The results suggest that in *Natural Sciences* and *Engineering & Technology* the word is used in a physical context for instance in combination with generators which is the closest word in both categories. In *Social Sciences* and *Humanities*, the nearest word is *powerful* followed by *control* which indicates that here it refers to "the ability to control something or someone"[12]. The remaining neighbours such as *counterbalance* and *generator* show that here the word also contains information about *power* in the physical context. The nearest neighbours to the GloVe embedding implicate a more general meaning.

The left panel of Figure 6.4 displays a dendrogram for WikiFoS calculated on the resulting distance matrix of W2V&Pred. It shows two larger clusters where one contains all categories of *Humanities* and *Social Sciences* except for Psychology which is included in the second cluster together with all categories of *Engineering & Technology* and *Natural Sciences*. The first big cluster can be separated in *Humanities* and its subcategories next to *Social Sciences* and two subcategories. The second cluster also shows reasonable neighbourhoods as for instance *Computer Science* and *Electrical & Electronic Engineering* as well as *Biology* and *Chemistry* are paired together. Although the first pair does not belong to the same main category it makes sense that their word

[12] https://www.oxfordlearnersdictionaries.com/definition/english/power_1

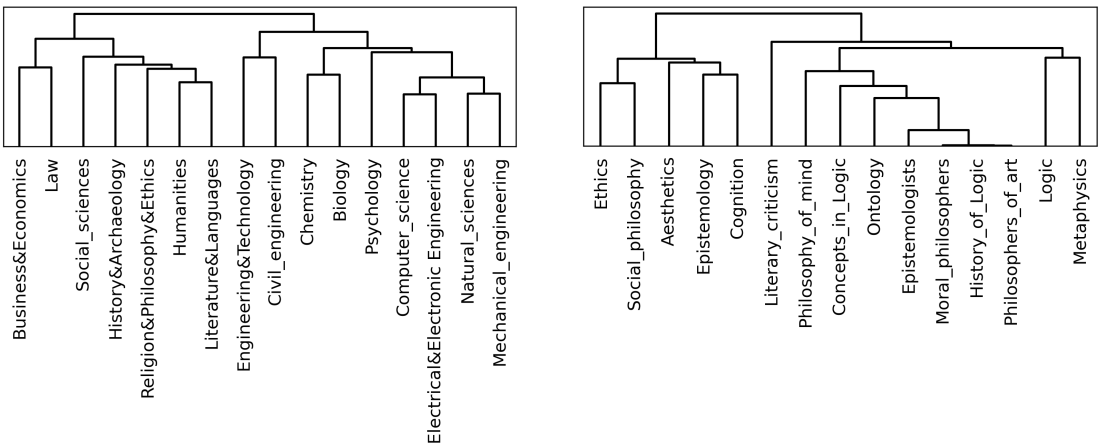


Figure 6.4: Dendrograms for WikiFoS (left) and WikiPhil (right) of test runs on W2V&Pred. U-shaped links indicate closeness between categories and hierarchical clusters are built by merging two legs of the U-link. The length of the legs indicate the distance between the categories or child clusters.

vectors are close as there are many parallels between *Computer Science* and *Electrical & Electronic Engineering*.

### 6.4.3 WikiPhil

For WikiPhil, results in Table 6.4 show that W2V&Pred reached the highest embedding accuracy for $n = 5, 10$ with 32% and 41.8% followed by GloVe ($4.8 - 37.1\%$) whereas the accuracy of W2V&Constr drops to $0.4 - 15\%$.

The structure score evaluated on W2V&Pred is significantly higher than the one evaluated on GloVe: 23.33% vs. 20%. Burrows' Delta reached a structure accuracy of only 6.7%.

The dendrogram for WikiPhil (right panel in Figure 6.4) groups *Ethics* and *Social Philosophy* as well as *Cognition* and *Epistemology* together. Both pairs also belong to the same cluster in the assumed structure. What is more interesting is the grouping of *Epistemologists*, *Moral Philosophers*, *History of Logic* and *Philosophers of Art* which mostly contain biographical articles, furthermore, they belong to four different clusters in the structure we assumed before.

This result combined with the clear decrease in embedding accuracy for W2V&Constr indicates that the biographical structure within articles leads to more similarity between word embeddings than the subject they belong to. We therefore rearranged the assumed structure and reran the W2V&Constr experiments.

We rearranged the structure for W2V&Constr, such that we build a new cluster out

of *Epistemologists*, *Moral Philosophers*, *History of Logic* and *Philosophers of Art* and repeated the experiments for W2V&Constr. The last row Table 6.4 (*restructured*) shows the embedding accuracies. In comparison to the previous run where we assumed the original structure, we could clearly improve the embedding accuracy and have even outperformed the highest accuracies of Table 6.4 by almost 5% for $n = 10$.

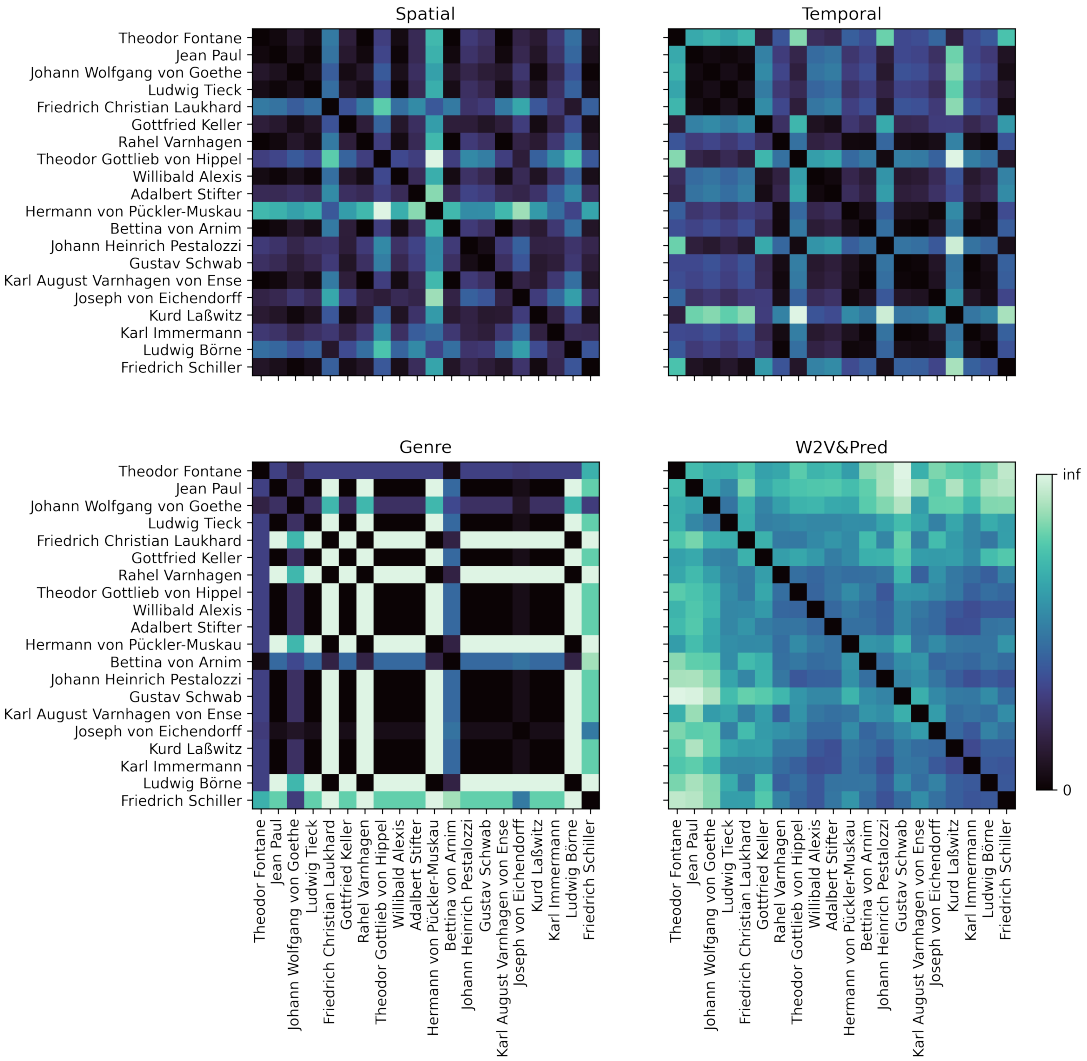


Figure 6.5: Spatial, temporal and genre distance matrices between authors. Lower right plot shows the distance matrix from W2V&Pred.

### 6.4.4 DTA

The interest in applying quantitative and statistical methods to large datasets in the field of literary studies where not all texts can be analysed by humans, give rise to interdisciplinary studies, especially in the field of digital humanities.

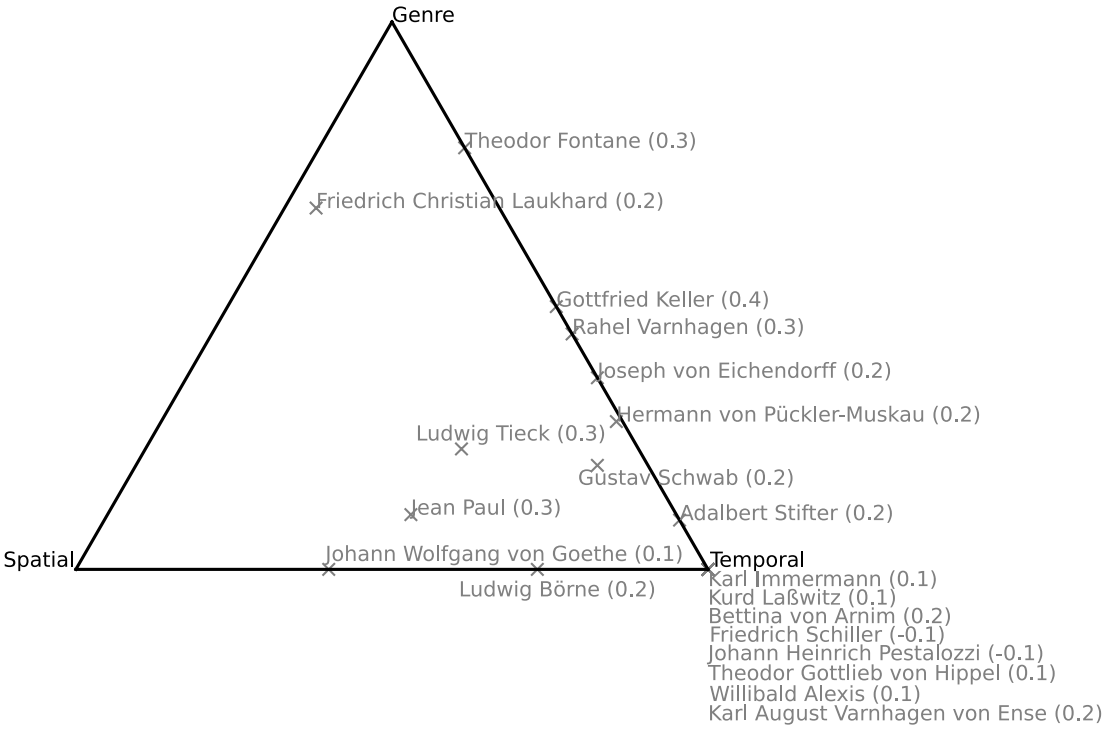


Figure 6.6: Spatio-temporal-genre correlation between authors. For each author we correlated a weighted spatio-temporal-genre representation with the respective values from W2V&Pred, and displayed the highest correlation in a barycentric coordinates triangle.

In an explorative experiment, we applied W2V&Pred to texts of the 20 most represented German authors between 1770 and 1900 within the Deutsches Textarchiv (DTA). We used the parameters $\lambda = 512, \tau = 1024$, the same parameter combination as for W2V&Pred on WikiFoS.

We compare the resulting distance matrix with temporal, spatial and genre information for each author. Therefore, we extracted the year and place of publication for each piece as well as one out of three genre categories (ego document, verse and fiction) and averaged across all pieces per author. Pairwise distance matrices between authors were calculated with absolute (temporal), geodesic (spatial) and cosine distances (genre vectors) and are displayed in Figure 6.5.

For each author, we computed correlation coefficients between their corresponding weighted spatio-temporal-genre vector and their W2V&Pred-vector, i.e. we want to find out how distances to other authors in the W2V&Pred matrix can be explained by a spatio-temporal-genre context. We here left out the values on the diagonal (the author-to-author distance), to not artificially increase the correlation. Highest correlation values are shown in barycentric coordinates in Figure 6.6. Most authors are located in the temporal context, only a few show their highest correlation in the spa-

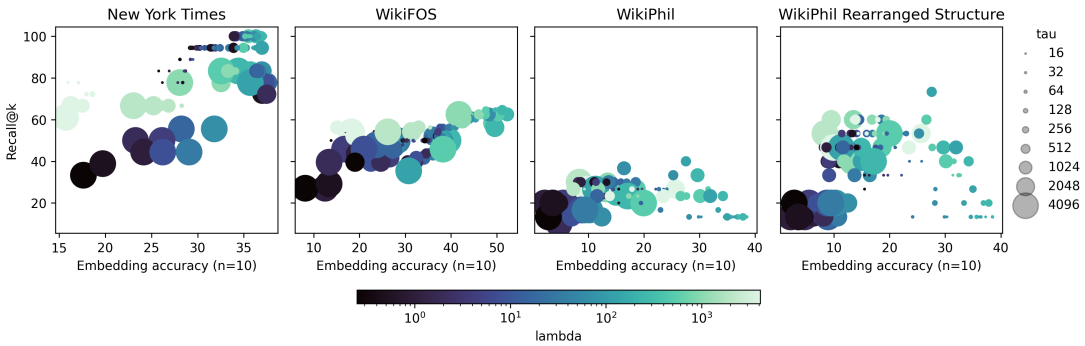


Figure 6.7: Evaluation for W2V&Pred on train sets for various parameters based on embedding accuracy (n=10) and recall@k. Size and color of the dots indicate the magnitude of the parameters tau and lambda respectively. Parameters with the best embedding accuracy for $n = 1, 5, 10$ were selected for W2V&Constr, WEN and DW2V.

tial or genre dimension. Five authors show a tendency towards the spatial dimension. Goethe, Paul and Tieck have all worked in the center of today's Germany. Börne and Laukhard both lived in France. This could explain their tendency towards the spatial dimension. Fontane and Pückler both wrote travel stories which presumably are similar in style, this might explain the influence of genre on their connections to other authors. Fontane's closest neighbour in the W2V&Pred matrix is also Pückler.

Location, time and genre are in most cases not sufficient to explain an author's oeuvre, especially since not all their pieces are contained in the dataset. Further and a more in-depth analysis would be needed. For instance, the influence of French language and culture on the writing of Börne and Laukhard deserves further attention. For Fontane, the unique mixture of poems, ego-documents and fiction might be the reason that he shows less similarities with other authors in the W2V&Pred matrix.

We argue that W2V&Pred still can be helpful in finding new connections in the field of digital humanities, especially for large datasets and when domain-experts are included. Similarities between spatio-temporal-genre features and the linguistic representation found by W2V&Pred can help identifying new connections between authors that have not been considered in traditional literary analyses.

### 6.4.5 Parameter Selection

In Figure 6.7, we show the evaluation of various parameter combinations $(\lambda, \tau)$ for W2V&Pred on the train sets of NYT, WikiFoS and WikiPhil. For $\lambda$, we applied log-scaled parameters in the range of $[2^{-2} - 2^{12}]$ and for $\tau$ in the range of $[2^4 - 2^{12}]$. For NYT and WikiFoS, we see a significant positive Pearson correlation coefficient of 0.58 and 0.65, respectively (with $\alpha = 0.05$ for all significance tests). In other words, parameters that achieve a high embedding accuracy, will probably also achieve a high structure

score. For WikiPhil, we get a significant negative Pearson correlation of $-0.19$ when we evaluate on the original structure. This indicates that we achieve either a good structure score or a high embedding accuracy. With the rearranged structure, we could improve the recall@k score. The Pearson correlation with $-0.14$ is still negative but not significant anymore ($p = 0.11$).

For all datasets we see that low values for $\lambda$, especially combined with large values for $\tau$ perform poorly on both metrics. In contrast, the combination of $\tau = 512, 1024$ and $\lambda = 512$ results in good embeddings and a high recall@k. For all methods, we started the parameter search with the reported ranges for $\lambda$ and $\tau$ and expanded those in case the best scores lay at the margin.

## 6.5 Discussion

We proposed two new methods, *Word2Vec with Structure Constraint (W2V&Constr)* and *Word2Vec with Structure Prediction (W2V&Pred)*, that generalize Dynamic Word2 Vec (Yao et al., 2018) such that dynamic word embeddings for different data slices can be simultaneously aligned and trained.

With W2V&Constr, we aim to train specific embeddings of high quality for each data slice (e.g. data slice $\widehat{=}$ domain) of datasets where the true structure between those data slices is known. Compared to training individual embeddings, e.g. with GloVe, embeddings trained by W2V&Constr benefit from the information exchange between data slices.

W2V&Pred goes one step further and predicts the structure of a dataset while word embeddings are trained for each data slice, i.e. no information about the structure is fed into the algorithm.

We applied all methods on three English datasets: chronological data containing New York Times articles between 1990-2016, articles from Wikipedia for different fields of science and Wikipedia articles from subcategories of Philosophy. For both Wikipedia datasets, we assumed a tree-like structure according to the level of subcategories the articles appeared in. Furthermore, we applied W2V&Pred to German texts from *Deutsches Textarchiv (DTA)* in an explorative experiment, where we do not know the underlying structure and do not finetune for hyperparameters.

We compared the results to two different baseline methods: static word embeddings trained by GloVe (Pennington et al., 2014) and the structure score for Burrows' Delta, a method used for authorship attribution which is based on word occurrences (Burrows, 2002). We could significantly outperform both methods with respect to embedding and structure accuracy on all English datasets (Wilcoxon signed rank test, $\alpha = 0.05$).

Finetuning W2V&Pred on the English datasets shows a significant correlation between the best embedding accuracy (n=10) and the best structure score (recall@k) of 0.58 and 0.65 for WikiFoS and NYT. This means, for those datasets there is a clear tendency

that a high embedding accuracy also yields a good structure score.

In the following, we will briefly summarize and discuss the results.

#### Embedding accuracy

We computed embedding accuracies on global analogy tests for all three English datasets. We could outperform GloVe on all three datasets for $n = 5$ and $n = 10$ (see Tables 6.2, 6.4). For $n = 1$, GloVe achieved higher results than W2V&Pred and W2V&Constr on WikiFoS with a very small margin of 0.25%.

For NYT, we reached much higher embedding accuracies compared to the ones trained by GloVe on the entire dataset. When looking into the analogies that were missed by GloVe, it seems that embeddings trained by GloVe are more influenced by current events, e.g. some countries are more associated to conflicts or their governments, states are closer to senators or universities. Averaging across years, what we do for the embeddings trained by W2V&Constr and W2V&Pred, might even out that effect and therefore improve the general embedding accuracy.

The performance of W2V&Pred on the temporal analogies which combine structure and word embedding quality shows that we here could not only replicate the results of Yao et al. (2018) but also significantly improve them by a margin of $2 - 3\%$.

For the Wikipedia datasets, we could not evaluate the embeddings on domain-specific test sets, we therefore looked into some exemplary words that carry different meanings across domains. For instance, in Table 6.5 we see that in both *Natural Sciences* and *Engineering & Technology* the word *power* contains the meaning in a physical context whereas in *Humanities* and *Social Sciences*, the word is used in a political context. Many similar examples can be found, e.g., charge:electrical-legal, performance:quality-acting, resistance:physical-social, race:championship-ethnicity.

#### Structure prediction with W2V&Pred

We evaluated the structure accuracy with recall@k as described in Section 6.3.3, results are also displayed in Tables 6.2, 6.4. We here only compare W2V&Pred with both baseline methods GloVe and Burrows' Delta since structure information is already fed into the training of W2V&Constr. For all three English datasets we could significantly outperform GloVe. W2V&Pred also achieved a much higher structure accuracy than Burrows' Delta which we cannot include in the significance tests over repetitions as the method is deterministic and can only be computed once. W2V&Pred is thus able to successfully predict the structure on a variety of datasets.

Furthermore, W2V&Pred can also be used to correct an assumed structure as we have seen for WikiPhil. After applying W2V&Constr, we saw a clear decrease in embedding accuracy along with a low recall@k score (see Table 6.4). The structure predicted by W2V&Pred resulted in a different grouping than the one we assumed. Interestingly,

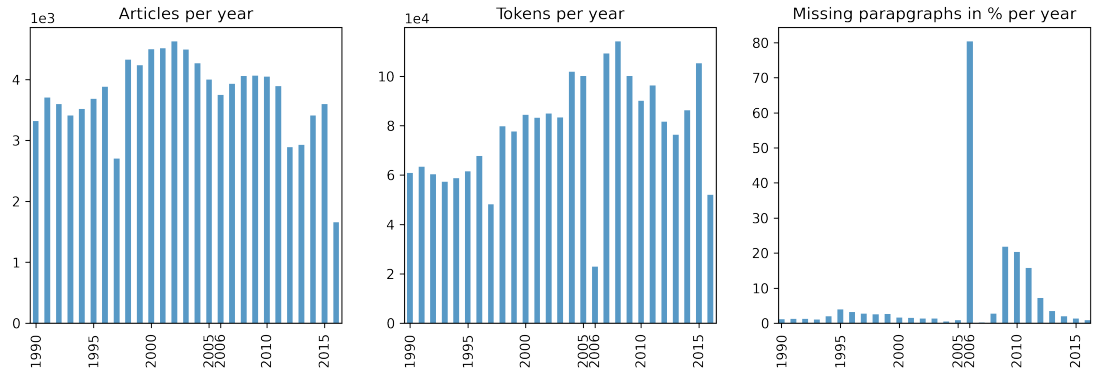


Figure 6.8: Number of articles, tokens and % missing paragraphs per year in the New York Times dataset. The original dataset (1990-2005) has been extended which might explain fewer tokens in 2006 due to missing paragraphs while the number of articles in 2006 does not crucially differ from other years.

texts that are mainly influenced by biographical articles form a cluster independent from the category they belong to. Restructuring the graph shows a clear increase in embedding and structure accuracy.

Also for the NYT dataset, we gained interesting insights after applying W2V&Pred. In Figure 6.2, we detected the year 2006 as an outlier. The reason for this might be that the New York Times dataset is an extension of an earlier dataset which only contained articles between 1990 and 2005. This might be the reason that the year 2006 only contains a fraction of the number of tokens compared to the other years. This can be seen in Figure 6.8 where we show the number of articles, the number of tokens and the percentage of missing paragraphs in all articles over years. The number of articles in 2006 is comparable to the other years but we found that many paragraphs and thus the majority of tokens are missing for 2006.

### W2V&Constr vs. W2V&Pred

With W2V&Constr, we feed a given structure into the method to train domain-specific word embeddings for each data slice based on that structure. This implies an information exchange during the training process and that way, we can improve the domain-specific embedding accuracy compared to individually trained embeddings which we showed by comparing our results to embeddings trained with GloVe (see results on temporal analogies GloVe in Tables 6.2).

W2V&Pred goes one step further and also predicts the underlying structure of the dataset while learning embeddings for each slice. This way, we can predict connections between data slices as we have done in the DTA experiment. The predicted structure with W2V&Pred outperformed both baseline methods GloVe and Burrows' Delta on all three English datasets.

In cases where the structure we assumed in W2V&Constr is not ideal, as we have seen for WikiPhil, we can also reevaluate this structure with W2V&Pred. For all three English datasets W2V&Constr outperforms W2V&Pred with respect to the embedding accuracy. The reason for this might be that W2V&Constr contains less degrees of freedom, i.e. the structure is given which is not the case for W2V&Pred. As we have seen in Figure 6.7, a good structure score very likely also leads to a high embedding accuracy. In other words, if the correct structure is already fed into the algorithm, W2V&Constr might benefit by converging to a better solution regarding embedding accuracy.

#### Limitations and Outlook

So far, W2V&Constr and W2V&Pred have been applied for a fixed vocabulary size and embedding dimension. Further experiments should be conducted to investigate the impact of the two parameters and whether an automatic approach in selecting both could improve the outcome.

A more in-depth evaluation of word embeddings on domain specific test sets or additional tasks could give more insights into the methods and how they could be improved. Recall@k is a valuable criteria to evaluate structure accuracy but it is often difficult to select accurate values for $k$ which crucially influence the result.

Expert knowledge is still necessary for the evaluation of explorative experiments, as we have seen when applying W2V&Pred to DTA. Expanding the evaluation metrics could also be helpful here. Though we strongly argue to include domain-experts, so far algorithms can hardly replace their knowledge and experience in the field.

Future research could also focus on an automatic approach to interpret the outcome of W2V&Constr and W2V&Pred. At the moment, detailed follow-up analyses are necessary to find out how the structure is predicted, e.g. what are the relevant words that change between sub-corpora.

# 7

# Summary & Conclusion

The application of machine learning models in dynamic contexts comes along with various challenges that need to be carefully investigated and addressed when building new models. In this thesis, we selected three different settings in the fields of computational neuroscience and natural language processing and investigated respective challenges, proposed novel methods and new approaches and thereby gained valuable insights.

Language and neuroscience both take place in a spatio-temporal dimension where dynamics over time and space crucially influence the data. Thus, similar question arise in the two fields as we have seen throughout this thesis, e.g. in language shifts over years in newspaper articles and changing patterns in the time frequency domain over the course of a 5-minute MEG experiment. Detecting and analysing those dynamics play a key role in both applications though tackling the relevant challenges still leads to different solutions.

**Chapter 3:** We presented a motor imagery-based BCI study where on top of the primary task, the imagination of left and right hand movements, we added secondary distraction tasks to mimic everyday life situations. Those distraction tasks included watching a flickering video, vibro-tactile stimulation on the forearms, closing eyes, searching the room for a letter-number combination and listening to the news. With those additional tasks we aimed to either distract the participants or disturb the signals by additional artifacts, for some tasks even both. Results with the standard CSP+RLDA classification pipeline to separate left from right hand movements did not achieve satisfying classification accuracies. We therefore investigated the influence of artifacts and feature shifts and proposed new approaches to tackle both. The first approach, ensemble CSP, trains one classifier for each distraction task and decides by a majority voting. We also proposed two-step CSP where we first separated the noisiest task and then applied individual classifiers in the second step. Both methods could significantly improve the results.
*Outlook:* As the tasks we included in this study are far from reflecting the entirety disturbance one experiences in the real world, a next step should be to further explore the influence of such realistic scenarios on the BCI performance and address the question on how to build reliable methods despite such distractions. Furthermore, we used a non-mobile EEG device with 64 gel-based electrodes in our study where preparation is time-consuming, electrodes dry out over time and participants cannot freely move. It would be highly beneficial to conduct this or a similar study with mobile devices.

**Chapter 4**: We proposed and applied a new version of SPoC: Fourier-SPoC (f-SPoC) optimizes the co-modulation between a signal and a periodic target function, here a sine

function, to extract spatial components in e.g. brain recordings. Together with Spatio-Spectral Decomposition (SSD), we apply f-SPoC to MEG data from a 5-minutes passive beat listening experiment in order to extract maximally relevant spatial components that correlate with the beat frequency. We compared the results from this approach to a sensor-level analysis for which we performed a cluster-based permutation analysis (CBPA). Both pipelines show periodic patterns in the time frequency response in the combined frequency range of alpha and low beta on group average. However, results on SSD+f-SPoC show a significant advantage in terms of modulation depth (peak-to-peak differences in ERS/ERD) and modulation power (FFT of TFR at the beat frequency) compared to CBPA. We further looked into the dynamics across the rather monotonous 5-minutes beat listening paradigm and found variation in the dominance of alpha and beta bands over the course of the experiment.
*Outlook:* This pipeline is by no means restricted to MEG or the sinusoidal target function and could easily be applied to EEG or other periodic neurophysiological signals such as the heart beat. Drawbacks of this pipeline are the search for the correct phase of the target function as well as finding the appropriate frequency band of interest for the SSD. For the latter an informed guess could be helpful but as we have seen, this should be reevaluated in case the results do not match the domain expert's expectations. The application to other beat frequencies or beats with small deviations in timing could be interesting in comparison to our findings. The variation in the time frequency response we found over time might be worth considering for future studies where repetitive trials over several minutes would be necessary in order to make a reliable statement.

**Chapter 6**: We presented two new methods: Word2Vec with Structure Constraint (W2V&Constr) and Word2Vec with Structure Prediction (W2V&Pred) train specific word embeddings for different domains or time stamps while capturing the dynamics between data slices. W2V&Constr learns and aligns dynamic word embeddings based on the underlying structure in the dataset while W2V&Pred goes one step further by also predicting the structure of a dataset. We evaluated both methods on three different English datasets with chronological and also tree-like structures. Furthermore, we compared our method to two baseline methods, namely GloVe and Burrows' Delta, with respect to word embedding accuracy and structure score. Our results show significant improvements for all three datasets compared to the baselines and furthermore captured interesting characteristics of the datasets. We could show that our methods are able to train high-quality word embeddings with and without prior structure and can even improve the prior structure when applied in combination. We further applied W2V&Pred on a German dataset from the *Deutsches Textarchiv* categorised by different authors in an explorative manner where we do not know the underlying structure of the data, i.e. the connections between authors.
*Outlook:* With a more in-depth evaluation of the word embeddings on specific test sets and additional tasks, we could gain more insights into the method and how it could be improved. Outliers and wrong assumptions about the data structure have led to interesting insights into the datasets but also need to be taken into consideration when applying both methods. Applying W2V&Pred without any expertise about the data structure might further lead to poor interpretation, at this stage it is therefore advisable

to consult an expert in the respective domain.

In conclusion, in this thesis we investigated three different applications of machine learning methods in dynamic settings. We could successfully classify a motor imagery-based BCI experiment in a highly artifact-contaminated environment, provide a new analysis pipeline to extract relevant brain components that show rhythmic entrainment with a regular beat and propose two new methods to learn dynamic word embeddings of high quality and additionally predict the underlying structure of the dataset. With that we contribute but do not want to limit our findings to the fields of computational neuroscience and natural language processing.

# Bibliography


Argamon S (2008). "Interpreting Burrows's Delta: Geometric and probabilistic foundations". In: *Literary and Linguistic Computing* 23.2, pp. 131–147.

Arvaneh M, Guan C, Ang KK and Quek C (2013a). "EEG data space adaptation to reduce intersession nonstationarity in brain-computer interface". In: *Neural computation* 25.8, pp. 2146–2171.

Arvaneh M, Guan C, Ang KK and Quek C (2013b). "Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain–computer interface". In: *IEEE transactions on neural networks and learning systems* 24.4, pp. 610–619.

Azarbonyad H, Dehghani M, Beelen K, Arkut A, Marx M and Kamps J (2017). "Words are malleable: Computing semantic shifts in political and media discourse". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1509–1518.

Bamler R and Mandt S (2017). "Dynamic word embeddings". In: *International conference on Machine learning*. PMLR, pp. 380–389.

Basu D (2011). "Randomization analysis of experimental data: the Fisher randomization test". In: *Selected Works of Debabrata Basu*. Springer, pp. 305–325.

Bekius A, Cope TE and Grube M (2016). "The beat to read: a cross-lingual link between rhythmic regularity perception and reading skill". In: *Frontiers in human neuroscience* 10, p. 425.

Belkin M and Niyogi P (2002). "Laplacian eigenmaps and spectral techniques for embedding and clustering". In: *Advances in neural information processing systems*, pp. 585–591.

Bengio Y, Ducharme R, Vincent P and Jauvin C (2003). "A neural probabilistic language model". In: *Journal of machine learning research* 3.Feb, pp. 1137–1155.

Benwell CS, London RE, Tagliabue CF, Veniero D, Gross J, Keitel C and Thut G (2019). "Frequency and power of human alpha oscillations drift systematically with time-on-task". In: *NeuroImage* 192, pp. 101–114.

Berger H (1929). "Über das Elektroenkephalogramm des Menschen". In: *Archiv für Psychiatrie und Nervenkrankheiten* 87.1, pp. 527–570.

Blankertz B, Curio G and Müller KR (2002). "Classifying single trial EEG: Towards brain computer interfacing". In: *Advances in neural information processing systems* 1, pp. 157–164.

Blankertz B, Dornhege G, Krauledat M, Müller KR and Curio G (2007). "The non-invasive Berlin brain–computer interface: fast acquisition of effective performance in untrained subjects". In: *NeuroImage* 37.2, pp. 539–550.

Blankertz B, Kawanabe M, Tomioka R, Hohlefeld F, Müller KR and Nikulin VV (2008a). "Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing". In: *Advances in neural information processing systems*, pp. 113–120.

Blankertz B, Tomioka R, Lemm S, Kawanabe M and Müller KR (2008b). "Optimizing spatial filters for robust EEG single-trial analysis". In: *IEEE Signal processing magazine* 25.1, pp. 41–56.

Blankertz B, Sannelli C, Halder S, Hammer EM, Kübler A, Müller KR, Curio G and Dickhaus T (2010a). "Neurophysiological predictor of SMR-based BCI performance". In: *NeuroImage* 51.4, pp. 1303–1309.

Blankertz B, Tangermann M, Vidaurre C, Fazli S, Sannelli C, Haufe S, Maeder C, Ramsey LE, Sturm I, Curio G et al. (2010b). "The Berlin brain–computer interface: non-medical uses of BCI technology". In: *Frontiers in neuroscience* 4, p. 198.

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

Bonetti L, Brattico E, Carlomagno F, Cabral J, Stevner A, Deco G, Whybrow PC, Pearce M, Pantazis D and Kringelbach ML (2020). "Spatiotemporal brain dynamics during recognition of the music of Johann Sebastian Bach". In: *BioRxiv*.

Brandl S (2015). *BCI under distraction: Motor imagery in a pseudo realistic environment*. http://dx.doi.org/10.14279/depositonce-9827.2.

Brandl S, Höhne J, Müller KR and Samek W (2015a). "Bringing BCI into everyday life: Motor imagery in a pseudo realistic environment". In: *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, pp. 224–227.

Brandl S, Müller KR and Samek W (2015b). "Robust common spatial patterns based on Bhattacharyya distance and Gamma divergence". In: *The 3rd International Winter Conference on Brain-Computer Interface*. IEEE, pp. 1–4.

Brandl S, Müller KR and Samek W (2016a). "Alternative CSP approaches for multimodal distributed BCI data". In: *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, pp. 003742–003747.

Brandl S, Frølich L, Höhne J, Müller KR and Samek W (2016b). "Brain–Computer Interfacing under Distraction: An Evaluation Study". In: *Journal of Neural Engineering* 13.5, p. 056012.

Brandl S and Lassner D (2019). "Times Are Changing: Investigating the Pace of Language Change in Diachronic Word Embeddings". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pp. 146–150.

Brandl S, Lassner D and Alber M (2020a). "Balancing the composition of word embeddings across heterogenous data sets". In: *arXiv preprint arXiv:2001.04693*.

Brandl S and Lassner D (2020b). *Corona Twitter Dataset: 16 February 2020 - 03 March 2020*. http://dx.doi.org/10.14279/depositonce-10012.

Brandl S and Lassner D (2020c). "Early Corona Twitter Dataset". In: *HAL preprint hal-02861167*.

Brandl S and Blankertz B (2020d). "Motor Imagery Under Distraction—An Open Access BCI Dataset". In: *Frontiers in neuroscience* 14, p. 967.

Brandl S, Haumann NT, Radloff S, Dähne S, Bonetti L, Vuust P, Brattico E and Grube M (2021). "Fourier SPoC: A customised machine-learning analysis pipeline for auditory beat-based entrainment in the MEG". In: *BioRxiv*.

Brouwer AM and Van Erp JB (2010). “A tactile P300 brain-computer interface”. In: *Frontiers in neuroscience* 4, p. 19.

Brouwer AM, Snelting A, Jaswa M, Flascher O, Krol L and Zander T (2017). “Physiological effects of adaptive cruise control behaviour in real driving”. In: *Proceedings of the 2017 ACM Workshop on An Application-oriented Approach to BCI out of the laboratory*, pp. 15–19.

Budd TW, Barry RJ, Gordon E, Rennie C and Michie PT (1998). “Decrement of the N1 auditory event-related potential with stimulus repetition: habituation vs. refractoriness”. In: *International Journal of Psychophysiology* 31.1, pp. 51–68.

Burrows J (2002). “‘Delta’: a measure of stylistic difference and a guide to likely authorship”. In: *Literary and linguistic computing* 17.3, pp. 267–287.

Buzsáki G, Anastassiou CA and Koch C (2012). “The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes”. In: *Nature reviews neuroscience* 13.6, pp. 407–420.

Castermans T, Duvinage M, Petieau M, Hoellinger T, De Saedeleer C, Seetharaman K, Bengoetxea A, Cheron G and Dutoit T (2011). “Optimizing the performances of a P300-based brain–computer interface in ambulatory conditions”. In: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 1.4, pp. 566–577.

Chang A, Bosnyak DJ and Trainor LJ (2016). “Unpredicted pitch modulates beta oscillatory power during rhythmic entrainment to a tone sequence”. In: *Frontiers in psychology* 7, p. 327.

Chatrian G, Lettich E and Nelson P (1985). “Ten percent electrode system for topographic studies of spontaneous and evoked EEG activities”. In: *American Journal of EEG technology* 25.2, pp. 83–92.

Chatrian GE, Petersen MC and Lazarte JA (1959). “The blocking of the rolandic wicket rhythm and some central changes related to movement”. In: *Electroencephalography and Clinical Neurophysiology* 11.3, pp. 497–510.

Cirelli LK, Bosnyak D, Manning FC, Spinelli C, Marie C, Fujioka T, Ghahremani A and Trainor LJ (2014). “Beat-induced fluctuations in auditory cortical beta-band activity: using EEG to measure age-related changes”. In: *Frontiers in psychology* 5, p. 742.

Compton RJ, Gearinger D and Wild H (2019). “The wandering mind oscillates: EEG alpha power is enhanced during moments of mind-wandering”. In: *Cognitive, Affective, & Behavioral Neuroscience* 19.5, pp. 1184–1191.

Costa-Faidella J, Baldeweg T, Grimm S and Escera C (2011). “Interactions between “what” and “when” in the auditory system: temporal predictability enhances repetition suppression”. In: *Journal of Neuroscience* 31.50, pp. 18590–18597.

Da Silva FL (2013). “EEG and MEG: relevance to neuroscience”. In: *Neuron* 80.5, pp. 1112–1128.

Dähne S, Meinecke FC, Haufe S, Höhne J, Tangermann M, Müller KR and Nikulin VV (2014). “SPoC: a novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters”. In: *NeuroImage* 86, pp. 111–122.

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

Deerwester S, Dumais ST, Furnas GW, Landauer TK and Harshman R (1990). "Indexing by latent semantic analysis". In: *Journal of the American society for information science* 41.6, p. 391.

Delorme A and Makeig S (2004). "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis". In: *Journal of neuroscience methods* 134.1, pp. 9–21.

Devlin J, Chang MW, Lee K and Toutanova K (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186.

Dornhege G, Millán JdR, Hinterberger T, McFarland DJ, Müller KR et al. (2007). *Toward brain-computer interfacing*. Vol. 63. MIT press Cambridge, MA.

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

Fatourechi M, Bashashati A, Ward RK and Birch GE (2007). "EMG and EOG artifacts in brain computer interface systems: A survey". In: *Clinical neurophysiology* 118.3, pp. 480–494.

Fazli S, Popescu F, Danóczy M, Blankertz B, Müller KR and Grozea C (2009). "Subject-independent mental state classification in single trials". In: *Neural networks* 22.9, pp. 1305–1312.

Fazli S, Mehnert J, Steinbrink J, Curio G, Villringer A, Müller KR and Blankertz B (2012). "Enhanced performance by a hybrid NIRS–EEG brain computer interface". In: *NeuroImage* 59.1, pp. 519–529.

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

Fisher RA (1936). "The use of multiple measurements in taxonomic problems". In: *Annals of eugenics* 7.2, pp. 179–188.

Friedman JH (1989). "Regularized discriminant analysis". In: *Journal of the American statistical association* 84.405, pp. 165–175.

Frølich L, Andersen TS and Mørup M (2015a). "Classification of independent components of EEG into multiple artifact classes". In: *Psychophysiology* 52.1, pp. 32–45.

Frølich L, Winkler I, Müller KR and Samek W (2015b). "Investigating effects of different artefact types on motor imagery BCI". In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 1942–1945.

Fruhstorfer H, Soveri P and Järvilehto T (1970). "Short-term habituation of the auditory evoked response in man". In: *Electroencephalography and Clinical Neurophysiology* 28.2, pp. 153–161.

Fujioka T, Trainor L, Large E and Ross B (2009). "Beta and gamma rhythms in human auditory cortex during musical beat processing". In: *Annals of the New York Academy of Sciences* 1169.1, pp. 89–92.

Fujioka T, Trainor LJ, Large EW and Ross B (2012). “Internalized timing of isochronous sounds is represented in neuromagnetic beta oscillations”. In: *Journal of Neuroscience* 32.5, pp. 1791–1802.

Garg N, Schiebinger L, Jurafsky D and Zou J (2018). “Word embeddings quantify 100 years of gender and ethnic stereotypes”. In: *Proceedings of the National Academy of Sciences* 115.16, E3635–E3644.

Gasser T, Sroka L and Möcks J (1985). “The transfer of EOG activity into the EEG for eyes open and closed”. In: *Electroencephalography and Clinical Neurophysiology* 61.2, pp. 181–193.

Gevins A and Smith ME (2003). “Neurophysiological measures of cognitive workload during human-computer interaction”. In: *Theoretical Issues in Ergonomics Science* 4.1-2, pp. 113–131.

Gonen H, Jawahar G, Seddah D and Goldberg Y (2020). “Simple, interpretable and stable method for detecting words with usage change across corpora”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 538–555.

Grahn JA and Brett M (2009). “Impairment of beat-based rhythm discrimination in Parkinson’s disease”. In: *Cortex* 45.1, pp. 54–61.

Graimann B, Allison BZ and Pfurtscheller G (2010). *Brain-computer interfaces: Revolutionizing human-computer interaction*. Springer.

Grosse-Wentrup M, Schölkopf B and Hill J (2011). “Causal influence of gamma oscillations on the sensorimotor rhythm”. In: *NeuroImage* 56.2, pp. 837–842.

Hämäläinen M, Hari R, Ilmoniemi RJ, Knuutila J and Lounasmaa OV (1993). “Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain”. In: *Reviews of modern Physics* 65.2, p. 413.

Hamilton WL, Leskovec J and Jurafsky D (2016). “Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1489–1501.

Hammer EM, Halder S, Blankertz B, Sannelli C, Dickhaus T, Kleih S, Müller KR and Kübler A (2012). “Psychological predictors of SMR-BCI performance”. In: *Biological psychology* 89.1, pp. 80–86.

Harmony T, Fernández T, Silva J, Bernal J, Díaz-Comas L, Reyes A, Marosi E, Rodríguez M and Rodríguez M (1996). “EEG delta activity: an indicator of attention to internal processing during performance of mental tasks”. In: *International Journal of Psychophysiology* 24.1-2, pp. 161–171.

Haufe S, Dähne S and Nikulin VV (2014a). “Dimensionality reduction for the analysis of brain oscillations”. In: *NeuroImage* 101, pp. 583–597.

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

Hofmann V, Pierrehumbert JB and Schütze H (2020). “Dynamic Contextualized Word Embeddings”. In: *arXiv preprint arXiv:2010.12684*.

Holm A, Lukander K, Korpela J, Sallinen M and Müller KM (2009). “Estimating Brain Load from the EEG”. In: *The Scientific World Journal* 9, pp. 639–651.

Howe RC and Sterman M (1972). "Cortical-subcortical EEG correlates of suppressed motor behavior during sleep and waking in the cat". In: *Electroencephalography and Clinical Neurophysiology* 32.6, pp. 681–695.

Huss M, Verney JP, Fosker T, Mead N and Goswami U (2011). "Music, rhythm, rise time perception and developmental dyslexia: perception of musical meter predicts reading and phonology". In: *Cortex* 47.6, pp. 674–689.

Iversen J, Repp B and Patel A (2009). "Top-down control of rhythm perception modulates early auditory responses". In: *Annals of the New York Academy of Sciences* 1169.1, pp. 58–73.

Jäncke L, Leipold S and Burkhard A (2018). "The neural underpinnings of music listening under different attention conditions". In: *Neuroreport* 29.7, pp. 594–604.

Jawahar G and Seddah D (2019). "Contextualized Diachronic Word Representations". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pp. 35–47.

Jiang JA, Chao CF, Chiu MJ, Lee RG, Tseng CL and Lin R (2007). "An automatic analysis method for detecting and eliminating ECG artifacts in EEG". In: *Computers in biology and medicine* 37.11, pp. 1660–1671.

Kawanabe M, Samek W, Müller KR and Vidaurre C (2014). "Robust common spatial filters with a maxmin approach". In: *Neural computation* 26.2, pp. 349–376.

Kim Y, Chiu YI, Hanaki K, Hegde D and Petrov S (2014). "Temporal Analysis of Language through Neural Language Models". In: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pp. 61–65.

Kingma DP and Ba J (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Klimesch W (1999). "EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis". In: *Brain research reviews* 29.2-3, pp. 169–195.

Ko LW, Komarov O, Hairston WD, Jung TP and Lin CT (2017). "Sustained attention in real classroom settings: An EEG study". In: *Frontiers in human neuroscience* 11, p. 388.

Krauledat M, Tangermann M, Blankertz B and Müller KR (2008). "Towards zero training for brain-computer interfacing". In: *PloS one* 3.8, e2967.

Kulkarni V, Al-Rfou R, Perozzi B and Skiena S (2015). "Statistically significant detection of linguistic change". In: *Proceedings of the 24th International Conference on World Wide Web*, pp. 625–635.

Kutuzov A, Øvrelid L, Szymanski T and Velldal E (2018). "Diachronic word embeddings and semantic shifts: a survey". In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1384–1397.

Ladányi E, Persici V, Fiveash A, Tillmann B and Gordon RL (2020). "Is atypical rhythm a risk factor for developmental speech and language disorders?" In: *Wiley Interdisciplinary Reviews: Cognitive Science*, e1528.

Large EW, Herrera JA and Velasco MJ (2015). "Neural networks for beat perception in musical rhythm". In: *Frontiers in systems neuroscience* 9, p. 159.

Lassner D, Brandl S, Baillot A and Nakajima S (2021). *Word Embeddings with Structure Prediction.* (In preparation).

Lee DJ, Kulubya E, Goldin P, Goodarzi A and Girgis F (2018). "Review of the neural oscillations underlying meditation". In: *Frontiers in neuroscience* 12, p. 178.

Lemm S, Blankertz B, Dickhaus T and Müller KR (2011). "Introduction to machine learning for brain imaging". In: *NeuroImage* 56.2, pp. 387–399.

Levy O and Goldberg Y (2014). "Neural word embedding as implicit matrix factorization". In: *Advances in neural information processing systems*, pp. 2177–2185.

London J and Time HI (2004). *Psychological Aspects of Musical Meter*.

Lotte F, Fujisawa J, Touyama H, Ito R, Hirose M and Lécuyer A (2009). "Towards ambulatory brain-computer interfaces: A pilot study with P300 signals". In: *Proceedings of the International Conference on Advances in Computer Enterntainment Technology*, pp. 336–339.

Makeig S, Bell AJ, Jung TP and Sejnowski TJ (1996). "Independent component analysis of electroencephalographic data". In: *Advances in neural information processing systems*, pp. 145–151.

Malmivuo J (2012). "Comparison of the properties of EEG and MEG in detecting the electric activity of the brain". In: *Brain topography* 25.1, pp. 1–19.

Mantini D, Franciotti R, Romani GL and Pizzella V (2008). "Improving MEG source localizations: an automated method for complete artifact removal based on independent component analysis". In: *NeuroImage* 40.1, pp. 160–173.

Marchesotti S, Bassolino M, Serino A, Bleuler H and Blanke O (2016). "Quantifying the role of motor imagery in brain-machine interfaces". In: *Scientific reports* 6, p. 24076.

Maris E and Oostenveld R (2007). "Nonparametric statistical testing of EEG-and MEG-data". In: *Journal of neuroscience methods* 164.1, pp. 177–190.

Marjanen J, Pivovarova L, Zosa E and Kurunmäki J (2019). "Clustering ideological terms in historical newspaper data with diachronic word embeddings". In: *5th International Workshop on Computational History, HistoInformatics 2019*. CEUR-WS.

Markovic A, Kühnis J and Jäncke L (2017). "Task context influences brain activation during music listening". In: *Frontiers in human neuroscience* 11, p. 342.

McFarland DJ, McCane LM, David SV and Wolpaw JR (1997). "Spatial filter selection for EEG-based communication". In: *Electroencephalography and Clinical Neurophysiology* 103.3, pp. 386–394.

Meltzer B, Reichenbach CS, Braiman C, Schiff ND, Hudspeth A and Reichenbach T (2015). "The steady-state response of the cerebral cortex to the beat of music reflects both the comprehension of music and attention". In: *Frontiers in human neuroscience* 9, p. 436.

Miklody D, Uitterhoeve WM, van Heel D, Klinkenberg K and Blankertz B (2017). "Maritime cognitive workload assessment". In: *10.1007/978-3-319-57753-1*.

Mikolov T, Sutskever I, Chen K, Corrado GS and Dean J (2013a). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*, pp. 3111–3119.

Mikolov T, Chen K, Corrado G and Dean J (2013b). "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781*.

Morgan S, Hansen J and Hillyard S (1996). "Selective attention to stimulus location modulates the steady-state visual evoked potential". In: *Proceedings of the National Academy of Sciences* 93.10, pp. 4770–4774.

Mulert C and Lemieux L (2009). *EEG-fMRI: physiological basis, technique, and applications*. Springer Science & Business Media.

Müller KR, Anderson CW and Birch GE (2003). "Linear and nonlinear methods for brain-computer interfaces". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 11.2, pp. 165–169.

Müller-Putz G, Scherer R, Brunner C, Leeb R and Pfurtscheller G (2008). "Better than random: a closer look on BCI results". In: *International Journal of Bioelectromagnetism* 10.1, pp. 52–55.

Müller-Putz G, Leeb R, Tangermann M, Höhne J, Kübler A, Cincotti F, Mattia D, Rupp R, Müller KR and Millán JdR (2015). "Towards noninvasive hybrid brain–computer interfaces: framework, practice, clinical application, and beyond". In: *Proceedings of the IEEE* 103.6, pp. 926–943.

Muthukumaraswamy S (2013). "High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations". In: *Frontiers in human neuroscience* 7, p. 138.

Niedermeyer E and da Silva FL (2005). *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins.

Nijboer F, Furdea A, Gunst I, Mellinger J, McFarland DJ, Birbaumer N and Kübler A (2008). "An auditory brain–computer interface (BCI)". In: *Journal of neuroscience methods* 167.1, pp. 43–50.

Nikulin VV, Nolte G and Curio G (2011). "A novel method for reliable and fast extraction of neuronal EEG/MEG oscillations on the basis of spatio-spectral decomposition". In: *NeuroImage* 55.4, pp. 1528–1535.

Nussbaumer HJ (1981). "The fast Fourier transform". In: *Fast Fourier Transform and Convolution Algorithms*. Springer, pp. 80–111.

Öhman A and Lader M (1972). "Selective attention and "habituation" of the auditory averaged evoked response in humans". In: *Physiology & behavior* 8.1, pp. 79–85.

Oostenveld R, Fries P, Maris E and Schoffelen JM (2011). "FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data". In: *Computational intelligence and neuroscience* 2011.

Ortner R, Lugo Z, Noirhomme Q, Laureys S and Guger C (2014). "A tactile brain-computer interface for severely disabled patients". In: *2014 IEEE Haptics Symposium (HAPTICS)*. IEEE, pp. 235–237.

Parra LC, Spence CD, Gerson AD and Sajda P (2005). "Recipes for the linear analysis of EEG". In: *NeuroImage* 28.2, pp. 326–341.

Pennington J, Socher R and Manning CD (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Pfurtscheller G (1992). "Event-related synchronization (ERS): an electrophysiological correlate of cortical areas at rest". In: *Electroencephalography and Clinical Neurophysiology* 83.1, pp. 62–69.

Pfurtscheller G and Neuper C (1994). "Event-related synchronization of mu rhythm in the EEG over the cortical hand area in man". In: *Neuroscience letters* 174.1, pp. 93–96.

Pfurtscheller G and Neuper C (1997). “Motor imagery activates primary sensorimotor area in humans”. In: *Neuroscience letters* 239.2-3, pp. 65–68.

Pfurtscheller G and Da Silva FL (1999). “Event-related EEG/MEG synchronization and desynchronization: basic principles”. In: *Clinical neurophysiology* 110.11, pp. 1842–1857.

Raffaelli Q, Mills C and Christoff K (2018). “The knowns and unknowns of boredom: A review of the literature”. In: *Experimental brain research* 236.9, pp. 2451–2462.

Ramoser H, Muller-Gerking J and Pfurtscheller G (2000). “Optimal spatial filtering of single trial EEG during imagined hand movement”. In: *IEEE transactions on rehabilitation engineering* 8.4, pp. 441–446.

Recasens M, Leung S, Grimm S, Nowak R and Escera C (2015). “Repetition suppression and repetition enhancement underlie auditory memory-trace formation in the human brain: an MEG study”. In: *NeuroImage* 108, pp. 75–86.

Rosburg T, Zimmerer K and Huonker R (2010). “Short-term habituation of auditory evoked potential and neuromagnetic field components in dependence of the interstimulus interval”. In: *Experimental brain research* 205.4, pp. 559–570.

Rudolph M, Ruiz F, Mandt S and Blei D (2016). “Exponential family embeddings”. In: *Advances in Neural Information Processing Systems*, pp. 478–486.

Rudolph M and Blei D (2018). “Dynamic embeddings for language evolution”. In: *Proceedings of the 2018 World Wide Web Conference*, pp. 1003–1011.

Samek W, Vidaurre C, Müller KR and Kawanabe M (2012). “Stationary common spatial patterns for brain–computer interfacing”. In: *Journal of neural engineering* 9.2, p. 026013.

Samek W, Kawanabe M and Müller KR (2013). “Divergence-based framework for common spatial patterns algorithms”. In: *IEEE Reviews in Biomedical Engineering* 7, pp. 50–72.

Schäfer J and Strimmer K (2005). “A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics”. In: *Statistical applications in genetics and molecular biology* 4.1.

Schlögl A, Keinrath C, Zimmermann D, Scherer R, Leeb R and Pfurtscheller G (2007). “A fully automated correction method of EOG artifacts in EEG recordings”. In: *Clinical neurophysiology* 118.1, pp. 98–104.

Schönemann PH (1966). “A generalized solution of the orthogonal procrustes problem”. In: *Psychometrika* 31.1, pp. 1–10.

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

Shenoy P, Krauledat M, Blankertz B, Rao RP and Müller KR (2006). “Towards adaptive classification for BCI”. In: *Journal of neural engineering* 3.1, R13.

Shin J, von Lühmann A, Blankertz B, Kim DW, Jeong J, Hwang HJ and Müller KR (2016). “Open access dataset for EEG+ NIRS single-trial classification”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.10, pp. 1735–1745.

Shoemark P, Liza FF, Nguyen D, Hale S and McGillivray B (2019). “Room to Glo: A systematic comparison of semantic change detection approaches with word em-

beddings". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 66–76.

Sugiyama M, Krauledat M and Müller KR (2007). "Covariate shift adaptation by importance weighted cross validation." In: *Journal of Machine Learning Research* 8.5.

Sykacek P, Roberts SJ and Stokes M (2004). "Adaptive BCI based on variational Bayesian Kalman filtering: an empirical evaluation". In: *IEEE Transactions on biomedical engineering* 51.5, pp. 719–727.

Tahmasebi N, Borin L and Jatowt A (2018). "Survey of computational approaches to lexical semantic change". In: *arXiv preprint arXiv:1811.06278*.

Taulu S and Simola J (2006). "Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements". In: *Physics in Medicine & Biology* 51.7, p. 1759.

Taulu S and Hari R (2009). "Removal of magnetoencephalographic artifacts with temporal signal-space separation: demonstration with single-trial auditory-evoked responses". In: *Human brain mapping* 30.5, pp. 1524–1534.

Teki S (2014). "Beta drives brain beats". In: *Frontiers in Systems Neuroscience* 8, p. 155.

Tobimatsu S, Zhang YM and Kato M (1999). "Steady-state vibration somatosensory evoked potentials: physiological characteristics and tuning function". In: *Clinical neurophysiology* 110.11, pp. 1953–1958.

Todorovic A and de Lange FP (2012). "Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields". In: *Journal of Neuroscience* 32.39, pp. 13389–13395.

Van Veen BD and Buckley KM (1988). "Beamforming: A versatile approach to spatial filtering". In: *IEEE assp magazine* 5.2, pp. 4–24.

Vidaurre C, Kawanabe M, von Bünau P, Blankertz B and Müller KR (2010). "Toward unsupervised adaptation of LDA for brain–computer interfaces". In: *IEEE Transactions on Biomedical Engineering* 58.3, pp. 587–597.

Vidaurre C, Sannelli C, Müller KR and Blankertz B (2011). "Co-adaptive calibration to improve BCI efficiency". In: *Journal of neural engineering* 8.2, p. 025009.

Vidovic MMC, Hwang HJ, Amsüss S, Hahne JM, Farina D and Müller KR (2015). "Improving the robustness of myoelectric pattern recognition for upper limb prostheses by covariate shift adaptation". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 24.9, pp. 961–970.

Von Lühmann A, Wabnitz H, Sander T and Müller KR (2016). "M3BA: a mobile, modular, multimodal biosignal acquisition architecture for miniaturized EEG-NIRS-based hybrid BCI and monitoring". In: *IEEE Transactions on Biomedical Engineering* 64.6, pp. 1199–1210.

Von Lühmann A, Boukouvalas Z, Müller KR and Adalı T (2019). "A new blind source separation framework for signal analysis and artifact rejection in functional Near-Infrared Spectroscopy". In: *NeuroImage* 200, pp. 72–88.

Winkler I, Brandl S, Horn F, Waldburger E, Allefeld C and Tangermann M (2014). "Robust artifactual independent component classification for BCI practitioners". In: *Journal of neural engineering* 11.3, p. 035013.

Winkler I, Debener S, Müller KR and Tangermann M (2015). "On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP". In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 4101–4105.

Wolpaw J and Wolpaw EW (2012). *Brain-computer interfaces: principles and practice*. OUP USA.

Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G and Vaughan TM (2002). "Brain–computer interfaces for communication and control". In: *Clinical neurophysiology* 113.6, pp. 767–791.

Woods DL and Elmasian R (1986). "The habituation of event-related potentials to speech sounds and tones". In: *Electroencephalography and Clinical Neurophysiology/ Evoked Potentials Section* 65.6, pp. 447–459.

World Medical Association et al. (2001). "World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects." In: *Bulletin of the World Health Organization* 79.4, p. 373.

Wriessnegger SC, Krumpl G, Pinegger A and Müller-Putz G (2017). "Mobile BCI Technology: NeuroIS Goes Out of the Lab, into the Field". In: *Information Systems and Neuroscience*. Springer, pp. 59–65.

Yao Z, Sun Y, Ding W, Rao N and Xiong H (2018). "Dynamic word embeddings for evolving semantic discovery". In: *Proceedings of the eleventh acm international conference on web search and data mining*, pp. 673–681.

Zeng Z, Yin Y, Song Y and Zhang M (2017). "Socialized Word Embeddings." In: *IJCAI*, pp. 3915–3921.

Zhang Y, Jatowt A, Bhowmick SS and Tanaka K (2016). "The past is not a foreign country: Detecting semantically similar terms across time". In: *IEEE Transactions on Knowledge and Data Engineering* 28.10, pp. 2793–2807.

Zhang Z (2000). "A flexible new technique for camera calibration". In: *IEEE Transactions on pattern analysis and machine intelligence* 22.11, pp. 1330–1334.

# List of Figures


3.1 Baseline analysis for calibration data. . . . . 23
3.2 Spectral, spatial and temporal information for all secondary tasks. . . . 25
3.3 Exemplary scalp maps of independent components. . . . . 28
3.4 Percentage of data variance explained by each artifact group. . . . . 30
3.5 Median power spectra of independent components. . . . . 31
3.6 Distribution of 63 classified ICs for *clean* and *numbers*. . . . . 32
3.7 Distribution of 63 classified ICs for participants *pp7* and *pp14*. . . . . 33
3.8 Activation patterns from all ICs of the calibration data for *pp7*. . . . . 34
3.9 Activation patterns from all ICs of the calibration data for *pp14*. . . . . 35
3.10 Feature shifts between *clean* and *numbers*. . . . . 36
3.11 Classification accuracies of distraction tasks against *clean*. . . . . 36
3.12 Comparison of new approaches vs. original classification pipeline. . . . . 42

4.1 Results of cluster-based permutation analysis (CBPA). . . . . 51
4.2 TFR of first and second SPoC component for various frequency bands. . 54
4.3 TFR of first SPoC component for 8-20 Hz. . . . . 55
4.4 TFR of CBPA and first SPoC component for three individuals. . . . . 56
4.5 TFR of first SPoC component on resting state. . . . . 57
4.6 Quantitative comparison of modulation with the beat CBPA vs. SPoC. . 58
4.7 Event-related fields for the first 5 trials for all sensors. . . . . 59
4.8 TFR over time of CBPA and first SPoC component. . . . . 61

6.1 True adjacency matrices for NYT, WikiFoS and WikiPhil. . . . . 78
6.2 Spectral embedding New York Times . . . . . 84
6.3 Temporal evolution of the word blackberry . . . . . 85
6.4 Dendrograms for WikiFoS and WikiPhil. . . . . 87
6.5 Distance matrices between authors. . . . . 88
6.6 Spatio-temporal-genre correlation between authors. . . . . 89
6.7 Grid search results for all English datasets . . . . . 90
6.8 Statistics on the New York Times dataset. . . . . 93

A.1 Toy train data for different scenarios. . . . . 115
A.2 Grid search results for parameter selection on toy train split. . . . . 117
A.3 Results for toy test runs. . . . . 118

# List of Tables


3.1 Description of secondary distraction tasks. . . . . 22
3.2 Classification accuracies for training on calibration data. . . . . 27
3.3 Classification accuracies for training on secondary tasks. . . . . 37
3.4 Classification accuracies for training after artifact removal. . . . . 39
3.5 Classification accuracies for ensemble CSP. . . . . 40
3.6 Classification accuracies for two-step CSP. . . . . 41

6.1 Categories in both Wikipedia datasets and the number of articles. . . . 78
6.2 Results for the New York Times dataset. . . . . 83
6.3 The word *mayor* in different epochs . . . . . 84
6.4 Results Wikipedia . . . . . 85
6.5 The word *power* in different scientific contexts. . . . . 86

# A

# Toy data experiments for Word2Vec with Structure Prediction

We carried out toy data experiments for Word2Vec with Structure Prediction (W2V&Pred) to further show that the method picks up connections between sub-corpora for various numbers of clusters. In the following, we will explain how data was generated in three different scenarios, how we selected the best parameters and how W2V&Pred performed on the test runs.

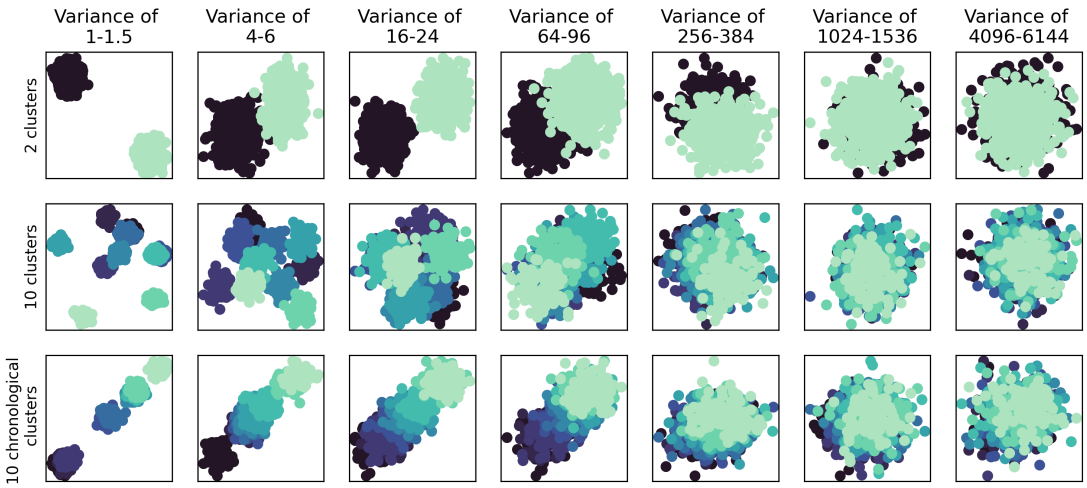


Figure A.1: Different toy data scenarios (rows) with varying variances (columns). Train sets are displayed, for all scenarios data has ben reshaped to 10 slices. Ticks have been remove for better readability, margin for variances are displayed for orientation.

## A.1 Data Generation

Toy data was generated in three different scenarios where the number of clusters and the continuity of the data varied, see Figure A.1. For all scenarios, we generated 2-dimensional data with a vocabulary of length 100 and reshaped the data to embedding matrices of 10 slices, once for the train set and then again for the test set. Corresponding co-occurrence matrices were then defined by the quadratic embedding matrices.

### A.1.1 Scenario 1: 2 clusters in 10 slices

Data was drawn from two multivariate Gaussian distributions. We therefore first sampled 2-dimensional vectors for each mean from a uniform distribution between $-20$ and 20. Two-dimensional covariance matrices for each cluster were constructed as diagonal matrices with variances also drawn from a uniform distribution with different values. We started with a low variance scenario with variances drawn between 1 and 1.5 and gradually increased variances on a logarithmic scale up to $4096 - 6144$ for the high variance scenario (lower bounds: $2^x$, higher bounds: $1.5 \cdot 2^x$). For each cluster we sampled 500 data points which we then reshaped into 10 slices à 100 data points, i.e. number of words in the vocabulary.

This scenario simulates data similar to the Wikipedia dataset which we proposed in Section 6.3.1 and Section 6.3.1 where several categories belong to the same main category, here cluster.

### A.1.2 Scenario 2: 10 clusters in 10 slices

Here, we sampled data from ten multivariate Gaussian distributions in the same way as before. Instead of 500 data points, we sampled 100 data points per cluster as one cluster corresponds to one slice and therefore one sample corresponds to one word per cluster.

We here simulated a rather general scenario where we do not have any prior knowledge about the structure between different clusters, as was the case for the DTA dataset in 6.3.1

### A.1.3 Scenario 3: 10 chronological clusters in 10 slices

As before, we sampled 100 data points from ten multivariate Gaussian distributions. To implement a continuous change between slices, we manipulated the second dimension of the mean to be the same as the first dimension and sorted the clusters according to their mean, i.e. they are all located on the main diagonal in the coordinate system.

This scenario resembles a chronological dataset, as for instance the New York Times dataset we discussed in Section 6.3.1.

## A.2 Grid search over parameters

For each dataset, we conducted a grid search over parameters for $\tau$ and $\lambda$ on the train set. For $\lambda$, we applied log-scaled parameters in the range of $[2^{-2} - 2^{12}]$ and for $\tau$ in the range of $[2^4 - 2^{12}]$, as before in Section 6.4.5. We also used the same optimization parameters as before (see Section 6.2.4). Structure accuracy was calculated on the predicted $Y$ which corresponds to the PPMI matrix and is, in contrast to the embedding matrices, invariant under rotation and scaling. Each run was evaluated with

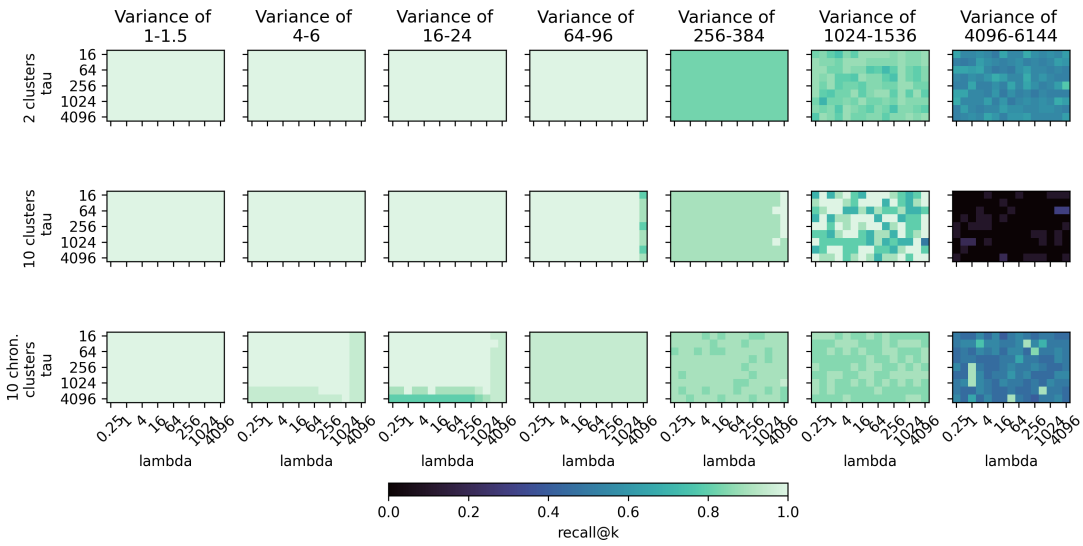


Figure A.2: Grid search results for parameter selection of $\lambda$ (x-axis) and $\tau$ (y-axis) on toy train split for different data scenarios (rows) with varying variances (columns). We evaluated each run on recall@k. For better readability we only displayed every second parameter on the axes.

recall@k, where values for $k$ varied between scenarios ($k = 4, 1, 2$ for the three scenarios respectively). Results are displayed in Figure A.2.

We can clearly see that for the first three variance values, we reached perfect structure accuracy for most parameter combinations in all scenarios. With increasing variance, structure score also decreases as expected. It is also not surprising that the accuracy for the second scenarios shows the lowest values in the high variance case, as we here only consider the nearest neighbor ($k = 1$) which is the most conservative structure evaluation.

## A.3 Results

For each (data, variance) combination, we selected the best parameters (randomly if multiple best parameters were available) and applied them to the corresponding test splits. We repeated each experiment three times. Results can be seen in Figure A.3 where the structure accuracies resemble the results from the grid search.

In this simplified toy data experiment, we could show that W2V&Pred is able to recover structures between different sub-corpora even in a high variance setting. This further highlights the results we saw in Chapter 6.

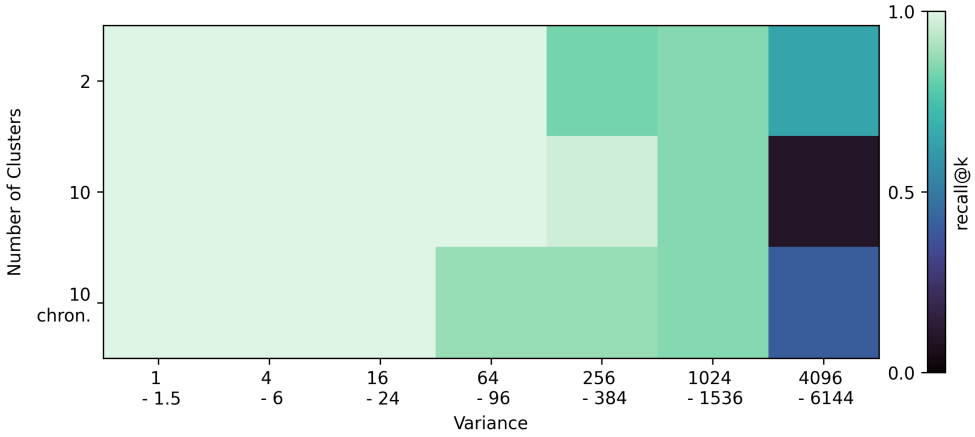


Figure A.3: Results for toy test runs averaged across repetitions.