SAARLAND UNIVERSITY

Faculty of Mathematics and Computer Science
Department of Computer Science
Dissertation

Leveraging EEG-based Speech Imagery Brain-Computer Interfaces

A dissertation submitted towards the degree
Doctor of Engineering (Dr.-Ing.)
of the
Faculty of Mathematics and Computer Science
of Saarland University

submitted by
Maurice Rekrut
Saarbrücken
2023

Dean of the Faculty:
Prof. Dr. Jürgen Steimle

Chair of the Comittee:
Prof. Dr. Anna Maria Feit

Reporters:
First Reviewer: Prof. Dr. Antonio Krüger
Second Reviewer: Dr. Fabien Lotte

Scientific Assistant:
Dr. Michael Feld

Day of the Colloquium:
05/05/2023

Notes on Style:
The majority of the work that is presented in this dissertation was done in collaboration with researchers and students which are mentioned in the acknowledgments and list of publications. Therefore, the scientific plural "we" is used throughout the thesis.

Acknowledgements

First of all, I want to express my deepest gratitude to **Antonio Krüger** for supervising and reviewing this thesis, as well as for providing this fascinating working environment at DFKI and the Ubiquitous Media Technology Lab, in which I was lucky enough to do my research and write this thesis. Antonio always supported me and my research interests, provided valuable feedback and offered me great flexibility and trust throughout the whole process of writing this thesis, thank you!
Special thanks to **Fabien Lotte** for agreeing to be my second reviewer. Your work in the field of Brain-Computer Interfaces inspired me and is cited in several parts of this thesis. It is a great honor, that my work is now verified by your outstanding expertise in the field of BCIs. After mainly meeting virtually over the last years, I am looking forward to an extended collaboration with many interesting BCI research and meetings in person in the future.
Naming all colleagues who have made my work at DFKI in the last years an unforgettable, fascinating and always a bit crazy experience and thus significantly contributed to this work would go beyond the scope of this acknowledgement section. All the names and stories would provide material for another dissertation and I hope you forgive me that I will not name you personally, although you all more than deserve it. Special thanks to my colleagues in the wonderful **Cognitive Assistants** research department and especially to the **Smarthomies** group that I have been assigned to.
Mentioning the Smarthomies group I still have to name one specific person, my group leader **Jan Alexandersson**. Almost 10 years ago I started my work in your group after an initial application and a short but intense interview. Without hesitation you gave me the chance to start my work at DFKI full time, although I was fresh out of university. 10 years, many fascinating research projects and a dissertation later, I hope that you do not regret your decision. Thank you so much for this chance, Tack så mycket!
Finally, my appreciation goes to all study participants, the student research assistants as well as the Bachelor and Master students who selected me as their thesis supervisor and significantly contributed to this work.
Last but most important I want to thank my family and friends.
Starting with my **friends**, thank you so much for your constant support by frequently asking me about the progress of this dissertation thing that I started so many years ago and the many times you pretended to be interested when I actually started talking about the topic. Many restless days and sleepless nights later I can tell you, that this thing came to an end and will not be a valid excuse anymore for postponing invitations to parties, spontaneous drinks or sports sessions, but especially the drinks and parties...
To **my parents**, Ewald and Theresia it is only because of you that I have become who I am. Your constant love, guidance and continuous support, even in difficult times, have made it possible for me to walk this path and follow my dreams of making a career in research. I will never be able to thank you enough, but I hope that this dissertation will compensate you for some of the hard times during my upbringing. I love you.
To **my wife**, Cordula, love of my life, thank you so much for supporting me during the process of writing this thesis and beyond. We have been together for quite some time and whenever I needed you, I could rely on your love and support no matter the situation or how difficult the times, you gave me the strength to face all challenges and struggles in life, thank you.
To **my sons**, Elon and Colin, you make me and us as a family complete, I could not be happier. This thesis is dedicated to you, might it inspire you to follow your dreams, whatever they might be in the future, reach for the stars.

Abstract

Speech Imagery Brain-Computer Interfaces (BCIs) provide an intuitive and flexible way of interaction via brain activity recorded during imagined speech. Imagined speech can be decoded in form of syllables or words and captured even with non-invasive measurement methods as for example the Electroencephalography (EEG). Over the last decade, research in this field has made tremendous progress and prototypical implementations of EEG-based Speech Imagery BCIs are numerous. However, most work is still conducted in controlled laboratory environments with offline classification and does not find its way to real online scenarios.
Within this thesis we identify three main reasons for these circumstances, namely, the mentally and physically exhausting training procedures, insufficient classification accuracies and cumbersome EEG setups with usually high-resolution headsets. We furthermore elaborate on possible solutions to overcome the aforementioned problems and present and evaluate new methods in each of the domains. In detail we introduce two new training concepts for imagined speech BCIs, one based on EEG activity during silently reading and the other recorded during overtly speaking certain words. Insufficient classification accuracies are addressed by introducing the concept of a Semantic Speech Imagery BCI, which classifies the semantic category of an imagined word prior to the word itself to increase the performance of the system. Finally, we investigate on different techniques for electrode reduction in Speech Imagery BCIs and aim at finding a suitable subset of electrodes for EEG-based imagined speech detection, therefore facilitating the cumbersome setups. All of our presented results together with general remarks on experiences and best practice for study setups concerning imagined speech are summarized and supposed to act as guidelines for further research in the field, thereby leveraging Speech Imagery BCIs towards real-world application.

Zusammenfassung

Speech Imagery Brain-Computer Interfaces (BCIs) bieten eine intuitive und flexible Möglichkeit der Interaktion mittels Gehirnaktivität, aufgezeichnet während der bloßen Vorstellung von Sprache. Vorgestellte Sprache kann in Form von Silben oder Wörtern auch mit nicht-invasiven Messmethoden wie der Elektroenzephalographie (EEG) gemessen und entschlüsselt werden. In den letzten zehn Jahren hat die Forschung auf diesem Gebiet enorme Fortschritte gemacht, und es gibt zahlreiche prototypische Implementierungen von EEG-basierten Speech Imagery BCIs. Die meisten Arbeiten werden jedoch immer noch in kontrollierten Laborumgebungen mit Offline-Klassifizierung durchgeführt und finden nicht den Weg in reale Online-Szenarien.
In dieser Arbeit identifizieren wir drei Hauptgründe für diesen Umstand, nämlich die geistig und körperlich anstrengenden Trainingsverfahren, unzureichende Klassifizierungsgenauigkeiten und umständliche EEG-Setups mit meist hochauflösenden Headsets. Darüber hinaus erarbeiten wir mögliche Lösungen zur Überwindung der oben genannten Probleme und präsentieren und evaluieren neue Methoden für jeden dieser Bereiche. Im Einzelnen stellen wir zwei neue Trainingskonzepte für Speech Imagery BCIs vor, von denen eines auf der Messung von EEG-Aktivität während des stillen Lesens und das andere auf der Aktivität während des Aussprechens bestimmter Wörter basiert. Unzureichende Klassifizierungsgenauigkeiten werden durch die Einführung des Konzepts eines Semantic Speech Imagery BCI angegangen, das die semantische Kategorie eines vorgestellten Wortes vor dem Wort selbst klassifiziert, um die Performance des Systems zu erhöhen. Schließlich untersuchen wir verschiedene Techniken zur Elektrodenreduktion bei Speech Imagery BCIs und zielen darauf ab, eine geeignete Teilmenge von Elektroden für die EEG-basierte Erkennung von vorgestellter Sprache zu finden, um so die umständlichen Setups zu erleichtern. Alle unsere Ergebnisse werden zusammen mit allgemeinen Bemerkungen zu Erfahrungen und Best Practices für Studien-Setups bezüglich vorgestellter Sprache zusammengefasst und sollen als Richtlinien für die weitere Forschung auf diesem Gebiet dienen, um so Speech Imagery BCIs für die Anwendung in der realen Welt zu optimieren.

List of Publications

Parts of the work presented in this dissertation, including ideas, applications, studies, results, conclusions and other text passages as well as figures and tables have already been published. The following list provides the reference of the publication, and where it appears in this dissertation:

- Rekrut, M., Jungbluth, T., Alexandersson, J. & Krüger, A. (2021, April). Spinning Icons: Introducing a Novel SSVEP-BCI Paradigm Based on Rotation. In 26th International Conference on Intelligent User Interfaces (IUI) (pp. 234-243). (appears in section 2.1.2)
- Rekrut, M., Fey, A., Nadig, M., Ihl, J. & Krüger, A. (2022, October). Classifying Words in Natural Reading Tasks Based on EEG Activity to Improve Silent Speech BCI Training in a Transfer Approach. In IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering. (appears in section 3.1)
- Rekrut, M., Selim, A. & Krüger, A. (2022, October). Improving Silent Speech BCI Training Procedures Through Transfer from Overt to Silent Speech. In IEEE International Conference on Systems, Man, and Cybernetics. (appears in section 3.2)
- Rekrut, M., Sharma, M., Schmitt, M., Alexandersson, J. & Krüger, A. (2020, February). Decoding semantic categories from eeg activity in object-based decision tasks. In 2020 8th International Winter Conference on Brain-Computer Interface (BCI) (pp. 1-7). IEEE. (appears in section 4.1)
- Rekrut, M., Sharma, M., Schmitt, M., Alexandersson, J. & Krüger, A. (2021, February). Decoding Semantic Categories from EEG Activity in Silent Speech Imagination Tasks. In 2021 9th International Winter Conference on Brain-Computer Interface (BCI) (pp. 1-7). IEEE. (appears in section 4.2)

Bachelor-/Master Theses

The author of this thesis supervised several bachelor's and master's students. The finished works that are relevant in this thesis' context (e.g., having a Brain-Computer Interface, Silent Speech or Speech Imagery focus) and which are in parts presented in the above papers and this dissertation, are listed below:

- Matthias Nadig, 2018, Electroencephalographic Mapping of Semantic Categories in Object-based Decision Tasks. MA [179]
- Mansi Sharma, 2019, Towards Silent Speech BCIs: Decoding Semantic Categories of Imagined Words from EEG Activity. MA [180]
- Tobias Jungbluth, 2020, Development of Meaningful Unobtrusive Stimuli for use in SSVEP BCIs. BA [177]
- Johannes Ihl, 2021, Investigating different methodologies for electrode reduction in Silent Speech Brain Computer Interfaces. BA

- Abdulrahman Mohammed Selim, 2021, EEG-Based Silent Speech Classification Using a Transfer Learning Approach from Overt to Silent Speech. MA [178]
- Andreas Fey, 2022, Towards Automated Training of Silent Speech BCIs in Natural Reading Tasks. MA [176]

Contents

Chapter 1
Introduction

The brain, billions of neurons communicating with each other, transferring information encoded in tiny electrical pulses via dendrites, leading to chemical reactions in the synapses inhibiting or exhibiting connected neurons thereby producing a network of electrical stimulations, which result in a micro-voltage firework inside our scull, commonly referred to as brain activity. The brain coordinates our body, our mind, our conscious self and led to astonishing achievements of mankind ranging from the invention of the wheel and the industrial revolution to the creation of super computers performing millions of processes in parallel, similar to our brain.
Over the last century, researchers have tried to unravel the functioning of the brain, its underlying processes and connectivity, resulting in various models of functional circuits and theories about this biological supercomputer in our head. One of the biggest projects concerned with the topic is the Swiss research initiative Blue Brain Project[1], which started in 2005 and aims at creating biologically detailed digital reconstructions and simulations of the mouse brain. The milestones of the project have been adapted over the years, since the initial goal of transferring those models to the human and provide a digital reconstruction of the human brain within 10 years, had turned out to be too ambitious. However, the project has accomplished various other significant milestones in the meantime, as for example creating a complete atlas of the neurons and glia of the mouse brain in 2018, or the first digital reconstruction of the brain's power source – the Neuro-Glia-Vascular Architecture in 2021. The current final goal of the project is to provide a holistic digital model of the mouse brain scheduled to be finished by 2024-2028. Goals addressing the digital reconstruction of the human brain have been outsourced to the European research initiative "Human Brain Project".
This initiative was started in 2013 and aims at solving various problems that lead to the delay and adjustment of the goals of the Blue Brain Project. According to the head of the project, Prof. Henry Markman, those were mainly caused by the lack of openly available experimental data and standardized labeling and archiving processes of brain data. According to their webpage[2] the vision of the project is "to deepen understanding of human brain structure and function, by building a European research infrastruc-

[1] `https://www.epfl.ch/research/domains/bluebrain/` Last accessed: 01.07.22
[2] `https://www.humanbrainproject.eu/en/` Last accessed: 04.07.22

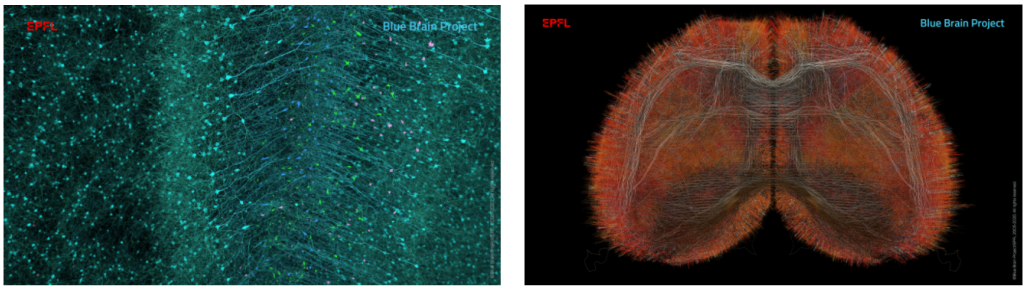

Figure 1.1: Results of the Blue Brain Project. Left: Close-up of synthesized neurons in a digital reconstruction of the cortex. Right: A digitally reconstructed isocortex with a simulation of synthesized white matter. Credit: Copyright © 2005-2020 Blue Brain Project/EPFL. All rights reserved.

ture that harnesses multiple disciplines and computing, and advances science, ICT and medicine, to the benefit of society". The deeper understanding of the brain's function that we have developed over the last years could help solve various problems that have plagued humanity for centuries, such as detecting and decoding the brain processes and structures involved in certain neuro-degenerative diseases, or simulating brain reactions to medication replacing early in-vivo testing in mice or rats.
Furthermore, those insights could help to understand the functional circuits of our brain, which regions or patterns are involved in creating certain emotions, triggering a specific movement or processing language. Those fundamental abilities might be lost due to certain brain damage or disease and the knowledge about how those processes are triggered and conducted by our brain might help to recover those functions in the future. The field of research concerned with this topic is the field of Brain-Computer Interaction.

1.1 Brain-Computer Interaction

One possible application resulting from deeper understanding of the brain and its functional areas is the classification of certain patterns of brain activity to use them for interaction. This process is usually referred to as Brain-Computer Interaction and the devices and applications which enable this kind of interaction as Brain-Computer Interfaces (BCIs) [212]. BCIs have been an active topic in research since the 1970s [213] with the purpose to restore lost communication or interaction pathways of disabled people through bypassing the damaged neural circuits with devices measuring and decoding certain patterns in brain activity. Researchers have managed to create prototypical interfaces which enable the control over wheelchairs [151], smarthomes [126] and even drones [114] with this technology but mostly within controlled scenarios in their laboratories. However, this technology has recently made its way from the laboratories of researchers worldwide into the focus of big industrial players, as for example Elon Musk and his company Neuralink[3]. Neuralink has the vision to create an implant consisting of micron-scale threads that are inserted into the brain by minimally invasive precision automated surgery. Each thread contains multiple electrodes and connects them to an implant, which forwards the electrical activity recorded at the brain, to a processing unit where those signals are converted into control commands for computers or mobile

[3] `https://neuralink.com` Last accessed: 04.07.22

devices. On their webpage Neuralink even present concepts, how such a device could be used to create a BCI controlled iOS device (see figure 1.2). An app would allow you to operate keyboard and mouse of the device with the help of brain activity by just thinking about the movement. These technologies could assist many people with neurological disease or injuries to control devices such as wheelchairs, smarthomes or computers, helping them to regain independence.

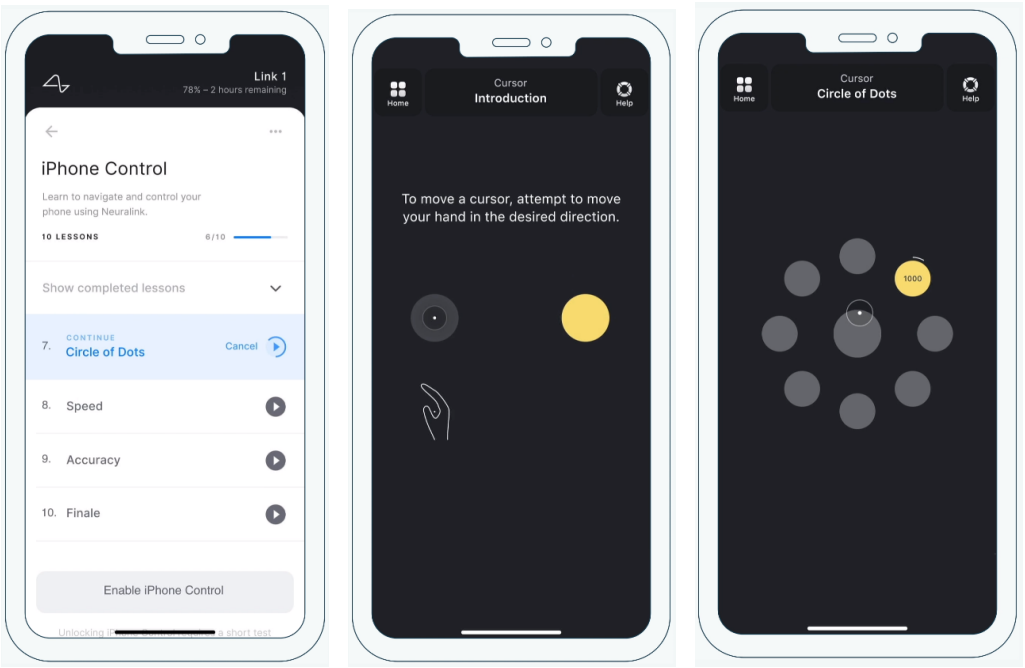

Figure 1.2: Vision of a smartphone control with the Neuralink device. Left: The user can select different training scenarios for different types of interaction. Middle: A training screen for a simple swipe interaction from left to right. Right: a more advanced cursor control with movement in several directions. Screenshots taken from [a] in December 2021.

[a] https://neuralink.com/approach/ Last accessed: 08.07.22

The visions and ambitions of this technology are overwhelming and promise "eventually to expand how we interact with each other, with the world, and with ourselves". To the point of writing this thesis first experiments of implanting the devices into pigs and apes have been successful, but the solution has not been implanted and tested on human beings. Musk is confident that the first prototype will be implanted into a human brain in 2023, without making promises however.
Other solutions have been implanted already for example into the brain of an Australian ALS patient who created the first ever BCI-based tweet on December 23rd 2021 [4]. The person sending this tweet is one of the two participants in a study, in which a novel endovascular Stentrode BCI was implanted in the primary motor cortex [163]. This new type of electrode is inserted through the blood vessels of a patient and functions just like a stent by pressing its tube-like structure towards the outer walls of the vessels ensuring the blood flow in it. The electrodes are embedded into this stent and can measure the brain activity in its surrounding. In the case of this study the method targets the

[4] https://twitter.com/tomoxl/status/1473809025254846467 Last accessed: 05.07.22

motorcortex and enables the control of a PC by imagined movements to trigger mouse clicks. The navigation itself is realized with an eye-tracking device and the BCI is used to trigger the click over a certain position. A promising approach, however invasive and in terms of interaction not fully based on brain activity as it requires the conscious voluntary control of the eyes and an additional screen for stimulus presentation.

Figure 1.3: The worlds first Tweet created with a BCI. Screenshot taken from [a] in December 2021.

[a] https://twitter.com/tomoxl/status/1473809025254846467 Last accessed: 05.07.22

This lack of intuitive interaction is actually a problem of many types of BCIs. Most of them depend on an external stimulus, like a flickering image presented on a screen, or are based on imagined movement which might work well for navigation, but are less suited for communication. We will give a detailed overview on the different techniques and methods used for Brain-Computer Interaction in chapter 2. A more intuitive and natural way of interacting via thoughts, and the focus of this thesis, are imagined or silent speech BCIs.

1.2 Silent Speech

In Human-Computer-Interaction (HCI) silent speech describes the concept of speech communication in the absence of an audible acoustic signal [61]. The fields of application of this technology are numerous. It can be used to restore the ability to communicate for patients with diseases like Amyotrophic lateral sclerosis (ALS) or Locked-In syndrome, for communication in noisy environments, or in general anywhere where overt speech is not applicable. One way of establishing this kind of interface is by measuring muscle activity as presented in [105]. A sensor recording the electrical activity of the facial muscles during speaking can be used to train machine learning algorithms to detect spoken words based on this activity alone. In this case it is sufficient to move the jaw and the lips like speaking the words without producing the actual sound. The device achieved astonishing classification accuracies of 92 % for ten digits from 0 to 9 and was later successfully transferred to a broader vocabulary [214], however it still requires the ability to move muscles. A more intuitive and less obtrusive way to establish a silent speech interface can be achieved by measuring brain activity while the user imagines speaking a word. This form of silent speech is also referred to as imagined speech or

Speech Imagery BCI (SI-BCI) and requires the user to do nothing more than thinking about silently speaking a word without moving any muscle. One can imagine it as the inner monologue of a person like speaking to oneself or reading a text silently.
The research in this field has made significant progress over the last decade, as we will see in more detail in chapter 2, and even a first step towards real world application was done by Facebook when they revealed in 2017, that the company had started the work on building a brain-computer interface that will let users type with their mind [5]. The technology was supposed to work with non-invasive optical imaging techniques, namely functional Near Infrared Spectroscopy (fNIRS), to scan the brain a hundred times per second and detect people speaking silently in their head. They promised an overwhelming typing speed of 100 words per minute using this non-invasive measuring technique. The company's effort resulted in the support of several important studies in the field of imagined speech [149, 138] and the development of a prototype which was however far from being able to classify the promised 100 words a minute. Facebook finally stopped funding the project in 2021. A predictable result, when having a look at the optimistic goal and the current state of the art in non-invasive Speech Imagery BCI research.
Systems developed and tested under controlled conditions inside laboratories are able to distinguish a set of 4 to 5 words with non-invasive brain measures as for example Electroencephalography (EEG). This technique measures the electrical activity of the brain at the scalp surface and is preferably used in BCI research as it is comparably cheap and has been well researched since it was discovered in the 1920s by Hans Berger [16]. The field of EEG-based imagined speech BCIs has made tremendous progress over the last decade. Figure 1.4 illustrates the current state of the art and the impressive possibility of controlling a virtual environment by imagined speech alone [126]. The person in the picture on the left is wearing a 64-channel EEG-headset and can interact with the avatar inside the virtual environment by silently speaking the four Korean words "Help me", "TV", "Weather", "Light". Based on the measured brain activity the system tries to classify the imagined word and according to the detected output the avatar will perform a certain action inside the environment, e.g. looking outside the window and tell the user about the weather conditions. These imagined speech applications with 3 to 4 different words have shown to work with a reliable classification accuracy of around 80 % [92] and illustrate the possibilities of this technology.
On the other hand, the example shown in figure 1.4 perfectly demonstrates the current drawbacks of the technology, which we will elaborate on in more detail in the problem statement.

1.3 Problem Statement

In our opinion, EEG-based imagined speech BCIs offer the most intuitive way of Brain-Computer Interaction. Speaking words silently in the head provides a natural and flexible way of interacting with any kind of system and measuring the brain activity non-invasively at the scalp surface, makes the technology applicable for a broad range of people apart from patients with brain implants. Recent works as presented in [126] illustrate the practical use of the technology in a specific scenario like the presented smarthome control and provide astonishing results, showing the capabilities of the system. At the same time, the scenario as presented in figure 1.4 perfectly illustrates current challenges with the state of the art.

[5] `https://about.fb.com/news/2017/04/f8-2017-day-2/` Last accessed: 04.07.22

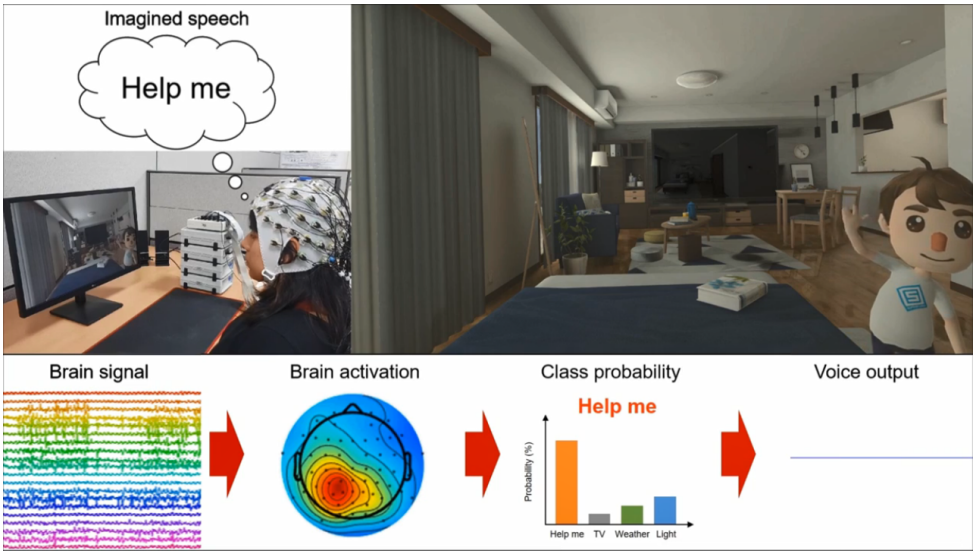

Figure 1.4: Prototype of a Speech Imagery BCI for smarthome control. The avatar can be controlled with 4 command words via imagined speech. Screenshot taken from video presented at [a] in July 2022.

[a] https://www.youtube.com/watch?v=s5SVkbU9yUU&ab_channel=DeepBCI
Last accessed: 10.07.22

First, maybe not obvious from the picture in figure 1.4, such astonishing Speech Imagery BCI applications can only be achieved with a tremendous amount of training. In the case of the smarthome control in [126], the participants had to perform 100 trails of repetition per word. Depending on the number of words to be used and the training procedure, this process can easily take several hours and requires the participant to sit in front of a screen without moving, reducing eye blinks to a minimum and remaining focused for the whole period of time. The words to repeat are usually presented on a screen which turns blank or only contains a fixation cross for the duration of the repetition [92, 93]. This procedure is mentally and physically exhausting and rather boring for the user, which might not only affect the participants mood, but therby also the quality of the recorded data. This problem of cumbersome and exhausting training scenarios for Speech Imagery BCIs is the first issue that shall be addressed within the scope of this thesis.

Second, the control is established with only 4 words. Given the complexity of most smarthomes and spoken human interaction in general, this number of commands is far from being sufficient. This is due to the fact that most multiclass imagined speech applications experience a significant drop in classification accuracy if the number of words to distinguish is increased, which was also the case for the presented smarthome scenario [125]. The screenshot presented in figure 1.4 was taken from a video on the researchers YouTube Channel [6], showing a reduced set of words in their demonstrator. The initial work included 13 words and the recorded data was used for offline [122] and pseudo-online [126] analysis, which we will explain in more detail in chapter 2.5.

[6] https://www.youtube.com/watch?v=s5SVkbU9yUU&ab_channel=DeepBCI
Last accessed: 10.07.22

The pseudo-online approach resulted in an average classification accuracy of 46.54 % [126]. Astonishing results considering the theoretic chance level of 7.7 %, disillusioning however, if one considers a real-world application in which the user wants to call for help and in 50% of the time the avatar would misinterpret this request and e.g. look outside the window to tell the user the weather forecast. The currently insufficient number of distinguishable words and the resulting drop in classification accuracy if this number is increased, is the second problem that shall be addressed with this thesis.

Lastly, having a look at the setup of the system illustrate in figure 1.4, it is obviously a quite bulky and complex apparatus required for recording the EEG signals. The 64 electrodes are wired to the stack of biosignal amplifiers next to the participant, making it a stationary setup. Furthermore the system is gel-based which requires the experimenter to insert gel in each and every electrode before the measurement can start to improve the skin-to-electrode contact, which is a quite time-consuming process. However, this principle of recording as many electrodes as possible is common in imagined speech research [92, 93, 110, 124, 191, 235], as it covers the scalp of the user in the best possible way and therefore provides data from most parts of the cortex. This bulky setup with numerous electrodes is the third problem that shall be addressed in this thesis to improve usability of EEG-based imagine speech systems by reducing the number of electrodes needed.

Summarizing the above mentioned problem statement, we currently see the following challenges in the field of EEG-based imagined speech BCIs:

Nr 1: **Training:** Exhausting and inefficient training scenarios.

Nr 2: **Classification:** Insufficient accuracy and number of words distinguishable.

Nr 3: **Usability:** Cumbersome setups inappropriate for real-world application.

1.4 Motivation

The motivation and goal of this work is to contribute to the field of imagined speech BCIs and work towards the real-world applicability of those systems. In order to do so, we will address the before mentioned problems and develop new methods and solutions that shall help to overcome the challenges in each of those three different categories.
The problem of cumbersome training scenarios has been present in BCI research ever since. Not only in imagined speech, but also in imagined movement for example, the participants need to record tremendous amounts of data in order to train a BCI. Those training sessions usually require the participants to sit still and reproduce a specific brain pattern linked to a certain thought over and over again. Beside being boring and cumbersome, the output that the participant produces during the training cannot be verified, as imagined speech training involves silent repetitions of a word without any audible output. The experimenter will therefore never know if the participants actually repeated the words correctly or if they maybe produced a different word by accident. We want to overcome this problem by proposing two novel methods of training an imagined speech BCI based on transfer learning and use EEG data recorded during overtly speaking words and silently reading them inside a text. The concept of training a Speech Imagery BCI on data recorded during reading is based on the idea, that the principle of imagined speech, namely, producing a word silently in the head, is similar

to silently reading a text. Overt speech on the other hand is rather the opposite of silent or imagined speech but research shows, that similar brain regions are involved in both processes [160, 30] which lets us conclude, that a transfer of the EEG data from overtly speaking words to silently speaking them should be possible. Summing up our thoughts on the problem of exhausting and inefficient training scenarios in imagined speech BCIs, we want to solve them by providing transfer learning approaches and train imagined speech classifiers based on EEG data recorded during reading and spoken speech.
As possible solution to the second problem, concerning the classification and insufficient number of words, we plan to include semantic processing into the classification process. The words in our language can be divided into different categories according to their meaning. Studies have shown, that the brain processes the words with different semantic meaning in different areas of the cortex and researchers have even managed to create semantic maps of those locations [88]. An additional layer in the classification process which detects the semantic category prior to the actual word classification could have the potential to increase the overall classification accuracy of the system and increase the number of words distinguishable. Within this work we will try to include semantic category classification of imagined words into an imagined speech BCI and show the feasibility of a Semantic Silent Speech BCI in order to increase the number of words distinguishable and the overall classification accuracy of the system.
Concerning the usability of imagined speech BCIs and the cumbersome setups inappropriate for real-world use, we plan to have a look at electrode reduction. Reducing the number of electrodes has been a topic in BCI research before, but in the field of Speech Imagery BCIs this issue has not been consistently addressed. Given the extensive 64+ channel EEG headsets currently used, one can imagine that such setups make those systems rather unusable in real world applications. Our idea to address this problem therefore is, to systematically remove electrodes from imagined speech datasets and try to find a certain subset of electrodes which provides comparable, or in the best case even better classification accuracies, as achieved with the initial set of electrodes. Hence, we could enable to establish an imagine speech BCI with a reduced number of electrodes and thereby reducing setup times and costs. We furthermore plan to investigate on the most important electrode positions in relation to their position on the cortex and their importance in contribution to the detection of imagined speech.

Summing up our motivation, in this thesis we want to provide novel concepts and methods to improve training procedures, classification and the usability of EEG-based imagined speech BCIs, to work towards the real-world applicability of this technology. In order to do so we will address the problems mentioned in section 1.3 by achieving the following goals:

Nr 1: Train an imagined speech classifier during interaction, e.g. while silently reading a text or speaking.

Nr 2: Include semantic categories of words into the classification process to increase the number of words distinguishable and the overall classification accuracy of imagined speech BCIs.

Nr 3: Reduce the number of electrodes needed for imagined speech detection and find a minimum subset of electrodes required for imagined speech BCIs.

1.5 Research Questions

In order to achieve the goals mentioned in the previous section, we want to answer the following research questions with this thesis:

RQ1 Can we automatically train SI-BCIs based on other modalities?

RQ 1.1 Can EEG activity recorded while reading certain words be used to train a classifier to detect those words during imagined speech?

RQ 1.2 Can EEG activity recorded while speaking certain words be used to train a classifier to detect those words during imagined speech?

RQ2 Can semantic category detection of words be used to improve the decision making process in SI-BCI?

RQ 2.1 Can semantic categories be classified from EEG activity during imagined speech production?

RQ 2.2 Can semantic classification increase classification accuracies in EEG-based imagined speech BCIs?

RQ3 Can we make SI-BCIs more user-friendly by reducing the required number of electrodes?

RQ 3.1 Is there a single best minimal set of electrodes for imagined speech BCIs?

RQ 3.2 Can we determine certain electrode positions related to good imagined speech classification accuracies?

RQ 1.1 and **RQ 1.2** will address the problem of currently exhausting and inefficient training scenarios in EEG-based imagined speech BCIs. **RQ 2.1** aims at training an imagined speech classifier on EEG-data recorded during reading while in **RQ 2.2** we plan to train the system on EEG-data recorded during speaking certain words. Both approaches will help to make training procedures less tedious and exhausting and enable a more productive way of training during interacting with the system. As both concepts have to the best of our knowledge not been addressed in literature before, we will provide a first implementation and answer those basic research questions for both of the approaches.
RQ 2.1 and **RQ 2.2** target the problem of insufficient classification accuracies and number of words currently distinguishable with EEG-based imagined speech BCIs. As mentioned in section 1.4, we want to include a classification of the semantic category of the word prior to the actual word classification under the hypothesis, that we can classify the semantic category of an imagined word. In a first step we need to evaluate the feasibility of this approach in EEG-based SI-BCIs in general and answer **RQ 1.1**. If this hypothesis can be confirmed we can proceed to testing the developed system and compare it to a standard classification approach in **RQ 1.2**.
By answering **RQ 3.1** and **RQ 3.2** we want to improve the usability of EEG-based imagined speech BCIs which are currently implemented by collecting data from numerous electrodes. In **RQ 3.1** we want to investigate on the possibility of electrode reduction by systematically reducing the number of electrodes and at the same time retaining a sufficient classification accuracy. Our goal is to provide a recommendation on a minimal number of electrodes required for reliable imagined speech detection from EEG data.

Based on these results we plan to identify relevant electrode positions inside these reduced subsets which contribute most to the classification process in imagined speech BCIs by answering **RQ 3.2**. Answering those two questions will provide insights in how far the currently high-resolution setups are required to provide appropriate classification results and in how far setups can be reduced in the future to facilitate the use of this technology in real-world applications.
The studies conducted in this work will help to answer those questions and provide major contributions to the research field of EEG-based imagined speech BCIs. In the following we will elaborate on those concrete contributions in detail.

1.6 Contributions to the Field

By answering the research questions in the previous section we will provide several contributions to the field of EEG-based imagined speech BCIs. Figure 1.5 gives an overview of those contributions which can be categorized according to the problems they are addressing.
The problem of exhausting and inefficient training scenarios will be addressed by developing training procedures based on spoken interaction and reading.
Training an imagined speech BCI while reading can improve the training conditions of imagined speech BCIs in many ways. Beside the fact that reading a word inside a text is much more comfortable and less exhausting as compared to repeating a word silently in front of a screen, the text in which the target words are hidden could be used to transfer knowledge about the application to the user. It would allow a bidirectional information transfer from the text to the users, e.g. in form of a manual, and at the same time from the users to the system, which learns to interpret their brain activity while processing certain commands. (RQ 1.1).
Similar revolutionary is the concept of training a Speech Imagery BCI with spoken speech. The training appears more natural and again less exhausting than repeating words silently in front of the screen. An additional advantage is the possibility of using the approach in an actual interaction scenario with spoken speech. The user could interact with a system with overt speech while an imagined speech classifier is trained in the background, enabling him to use the system in future interaction with the help of imagined speech as well. Hence, while using the system it would be possible to train an imagined speech classifier simultaneously and switch interaction paradigms depending on the situation (RQ 1.2).
Both of those novel training scenarios have, to the best of our knowledge, not been implemented in imagined speech BCIs before and have the potential to revolutionize training procedures of BCI applications in general beyond imagined or silent speech.
Concerning the problem of insufficient classification accuracy and number of words in imagined speech BCIs, we will develop a new technique of integrating semantic classification into the imagined speech detection process. The suggested approach of classifying semantic categories prior to actual words will have an impact on the way how we design imagined and silent speech BCIs and create a whole new type, so called Semantic Silent Speech BCIs. A successful classification of the semantic category prior to the word itself should allow an increase in the number of words classifiable in one application, as it would add additional features to the classification process and therefore, in the best case, improve classification results. (RQ 2.1 & RQ 2.2).
Reducing the number of electrodes of a Speech Imagery BCI will contribute to the usability of the systems since most of the current work is done in controlled environments in

laboratories with a large number of electrodes (64-128), which is time-consuming in setup and uncomfortable to wear for the participant or later user of the system. In real-world applications one would aim for a reduced number of electrodes which is the focus of RQ 3.1 and RQ 3.2. The successful transfer to a reduced number of electrodes will deliver insights on the most relevant electrode positions for imagined speech detection based on EEG and a minimal subset required to detect imagined speech reliably. This will make the setup less time consuming and will be a first step towards the applicability of the technology in real-world scenarios. Beside the targeted real-world application, study setup times can be reduced drastically with reducing the number of electrodes which will contribute to facilitating studies and increase scientific output in the field.
All the before mentioned contributions with regard to the research questions will overall contribute to improving usability of imagined speech BCIs. The results of the studies in the different fields presented in this thesis can act as guidelines for the design and implementation of imagined speech BCIs and therefore pave the way for real-world application in the future.
The implementation of a real-world online imagined speech BCI application is however **not** part of this thesis.

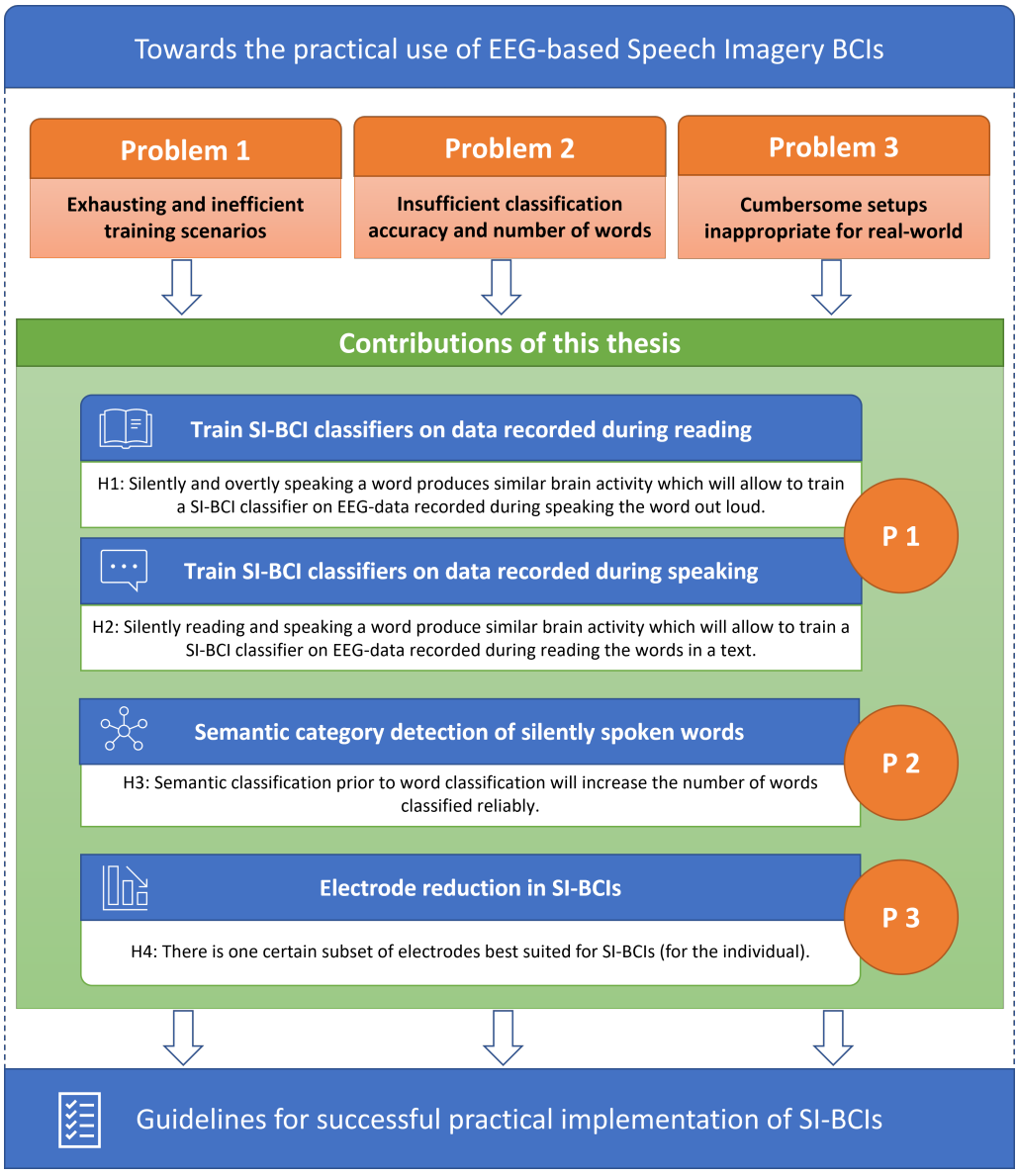

Figure 1.5: Overview of the contributions of this work in relation to the defined problems and hypothesis.

1.7 Thesis Outline

We continue in the following with presenting the necessary background and an overview on related work in the research field of imagined speech BCIs as well as BCIs in general in chapter 2. After providing this fundamental background knowledge and the current state-of-the-art, we will present our own works starting with the improvement on training procedures in Speech Imagery BCIs in chapter 3. In chapter 4 we will continue with our studies conducted on semantic classification in imagined speech, to improve classification accuracy. Chapter 5 will provide details on our works on electrode reduction in Speech Imagery BCIs and the attempts to improve usability of their setups. We will summarize and discuss our results of the work in the general discussion in chapter 6. This chapter will not only contain the results of our three main works in chapters 3, 4 and 5, but also a more high-level discussion on EEG-based Speech Imagery BCIs in general, as well as recommendations concerning study setups and implementations for future application. Additionally, we will discuss ethical concerns which might arise with the topic presented in this thesis. Finally, we will conclude by providing a summary on the presented work, our contributions, the limitations and future work in chapter 7.

Chapter 2
Background and Related Work

In this chapter we present background on Brain-Computer Interfaces (BCIs), Silent Speech and the combination of both topics, Speech Imagery BCIs (SI-BCIs). We will cover the most relevant definitions and related work in the field and provide the required links to the main topic of the thesis, EEG-based SI-BCIs. We will highlight the most recent and relevant work in this field in order to support our motivation and claims as mentioned in the problem statement in chapter 1.

2.1 Brain-Computer Interfaces

A Brain-Computer Interface describes a method of communication based on neural activity generated by the brain and is independent of its normal output pathways of peripheral nerves and muscles [212].
This definition from 2005 has come a long way since Richard Catons first observations of electrical current between the cerebral cortex and the scalp in 1875 [29] and its first recording by Hans Berger in 1929 [16]. Berger used surface electrodes to record the electrical activity at the scalp and defined this process as Electroencephalography (EEG). The discoveries of Caton and Berger have paved the way for brain activity measurements into the field of electrophysiology and made it a distinguished tool in clinical diagnosis and neurophysiological research [187, 203]. Since then it has become an important means for monitoring patients with stroke [225], epileptic seizures [158] or diagnosis of brain tumors [90]. However, it took almost 50 years after the invention of the EEG until Jacques Vidal coined the term Brain-Computer Interface in his publication "Toward direct brain-computer communication" from 1973 [213]. In his work he describes the possibility of evoking potentials in the EEG as response to external stimuli, as for example a brief illumination of the visual field or a tap on the forearm, and the use of those evoked potentials in man-machine communication. He describes a pilot project on direct Brain-Computer Communication in which he mentions the term Brain-Computer Interface for the first time. The aim of the project was to create an infrastructure consisting of several biosignal measures including EEG for data acquisition and a variety of output devices to present feedback to the participants in a shielded environment, thereby

creating a controlled setting for first experiments in the field of direct brain-computer communication. Within his work Vidal discusses a variety of possible applications of the presented setup and the technology in general, from concrete plans for experiments to identify features in the evoked responses of the EEG signals, but also possible applications of the technology like controlling prosthetic devices or even spaceships.
50 years later again, we are far from controlling a spaceship with this technology, however, research and engineering have made progress and so have the methods for measuring brain activity. The electrical activity of the brain measured by EEG is not the only method to establish BCIs anymore and the definition from 2005 [212] referring to neural activity measured by whatever means, becomes more accurate.

2.1.1 Measuring Brain Activity

The early works of Vidal and the majority of other BCI applications and research is concerned with the measurement of the electrical activity of the brain by EEG and so is our work. Throughout this thesis we will however be referring to other methods of measuring brain activity as they have also been used to establish BCIs or because they have been used to investigate basic principles which are important for our work. Therefore, we will present a brief explanation of several different methods for measuring brain activity, which will be mentioned in the course of this thesis, to provide a basic understanding of the underlying principles.

Functional Magnetic Resonance Imaging (fMRI)

Functional magnetic resonance imaging (fMRI) detects changes in blood flow by measuring the blood-oxygen-level dependent (BOLD) contrast [132]. It makes use of the fact, that hemoglobin responds differently to magnetic fields depending on whether it has bound oxygen molecules or not. These changes in oxygen level can therefore be measured in the response to the magnetic field produced by an MRI scanner and give evidence on oxygen consumption of the different areas of the brain and therefore the activity in those areas. In contrast to the electrical activity of the brain measured by EEG, the fMRI measures the hemodynamic response related to energy use in the brain [28]. It has a high spatial resolution and can conclude on activity in areas as small as a square millimeter, whereby the temporal resolution is rather poor with around 1 to 2 seconds [87]. Furthermore, the devices are bulky, stationary and require the participant to be lying inside a narrow tube, making them unsuitable for mobile real-world application in a BCI setting. However, they are frequently used to investigate basic principles of functional circuits of our brain as they deliver a precise view on the local sources of activity and have significantly contributed to our understanding of the functioning of the brain over the last decade.

Magnetoencephalography (MEG)

The Magnetoencephalography (MEG) measures magnetic fields which result from the electric activity of the brain. Due to the small electric currents that are produced by the neurons, the resulting magnetic fields are equally tiny and the measurement environment as well as the sensors need appropriate shielding against environmental influences [78]. The EMG has a high spatial resolution and in comparison to the EEG it is not affected by volume conduction, which means that the measurable magnetic field on the scalp surface

is not distorted as for example the electrical signals of the EEG after traveling through brain liquids, bones and skin [164]. The biggest drawback of the technology is the bulky setup, similar to the fMRI, since the most commonly used sensors or magnetometers are superconducting quantum interference devices (SQUIDs) which need special shielding and cooling with liquid helium. However, there exists first work on more portable and user friendly sensors for future mobile applications [25, 164]. Figure 2.1 shows a conventional MEG on the left and the newly developed solution on the right. The approach works with optically pumped magnetometers (OPMs) instead of SQUIDs, which do not need special cooling. Within their paper the researchers could show that their system measures up well against conventional MEG and is not affected by head movement, giving it one important advantage over established portable brain measurement methods as for example the EEG. However, one major drawback remains, which is the sensibility to environmental artifacts like the magnetic field of the earth or any other electrical device. Therefore, the aspect of portability and mobility will be limited to movement in a shielded room. Nevertheless, with further development this technology has the potential to become a game changer in BCI applications within the next ten to twenty years.

Figure 2.1: A conventional MEG device on the left and the newly developed portable solution on the right, as presented in [25]. The conventional SQUID based solution is comparably bulky like the fMRI while the new solution based on OPMs allows the user to move and e.g. hold a cup.

Electrocorticography (ECoG)

The Electrocorticography (ECoG) is an invasive method using electrode grids either placed on the surface of the brain or partially inserted in it, for recording neural activity from its surface, in particular from the cortex [211]. Those grids can contain hundreds of tiny electrodes and are mostly placed at certain functional areas as for example the motor cortex. The technology provides insights on the activity of those regions, not only for

BCI applications but also during surgeries as feedback about possible functional changes of the area of interest [118]. Due to the fact that the electrodes are placed directly at the surface of the brain, this technology provides the most valuable and clean signals for BCI applications without distortions caused by the electrical currents spreading through the scull and skin before reaching the surface electrodes of the EEG [80]. The disadvantages on the other hand are obvious, as it usually requires to open the scull to implant the electrodes, involving risks for the patient. But even minimal invasive approaches of inserting stent electrodes through blood vessels as described in [163] can cause blood clotting or rejection of the implant. For this reason, this technology is currently solely used for people who have to undergo brain surgery anyways, and for whom limited life circumstances justify such a risky procedure.
First online BCI-studies with ECoG were conducted in 2004 [129] when Leuthardt et al. demonstrated the control of a simple computer-cursor by measuring changes in cortical rhythms induced by imagined or real movements like opening a hand or speaking a word. The study was conducted with four participants and with a brief training period of 3 - 24 minutes they achieved success rates of 74 - 100% in a binary task.
Over the last decades, numerous groups from all over the world have started to work on ECoG-based BCIs, especially in the field of motor imagery and silent speech BCIs as we will see in section 2.4. Due to the risks for the user arising with invasiveness however, the majority of BCI applications is focusing on non-invasive approaches, as for example EEG or fNIRS.

Functional near-infrared spectroscopy (fNIRS)

Functional near-infrared spectroscopy (fNIRS) is an optical brain measuring method, which uses near-infrared light to detect changes in oxygenated blood flow. Similar to fMRI, fNIRS measures hemodynamic responses resulting from neural activity. In comparison to fMRI, fNIRS does not measure changes in the magnetic field but the changes in absorption of near infrared light. The sensors are placed on the scalp surface and emit near-infrared light into the scull. By measuring the amount of its reflection, the system can conclude on the blood volume and oxygenation resulting from the activity of the brain [59]. The technology is non-invasive, has a good temporal resolution and has recently been made portable, which makes such devices applicable for BCI solutions [45, 155]. However, the spatial resolution is rather low and the optical method allows only shallow penetration depth. Furthermore, hemodynamic changes come with some latency since the oxygenated blood is delivered to areas with increased activity by demand. This process can take time and the measured signal about the blood flow is just a by-product of what BCIs actually want to measure, the activity of the neurons of the brain. This is one of the reasons why methods which measure the electrical activity, the actual output of the information transfer in our brain, are still preferably used for the implementation of BCIs.

Electroencephalography (EEG)

The Electroencephalography (EEG) measures the electrical activity of the brain at the scalp surface. More precisely, according to the volume conduction model, it measures the sum activity of the apical dendrites of pyramidal neurons, which are the most prevalent cells in the cerebral cortex [38, 171], resulting in the measurable surface potential on the scalp.

The concrete underlying physiological causes and processes are a matter of discussion in the scientific community since the first recording in the 1920s by Hans Berger [16] and are not fully understood until today [38, 26].

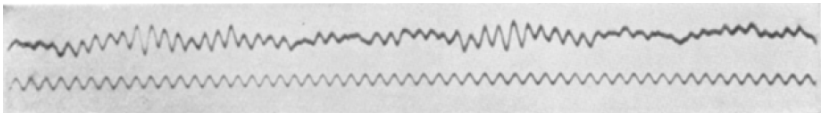

Figure 2.2: The first ever recorded EEG by Hans Berger as reported in [16]. The top row represents the EEG signal, the bottom row is a sinus signal for reference.

One commonly accepted approach towards an understanding of the anatomical sources is the neural field model [26, 40]. This model treats the cortex as continuous sheet and represents time course and spatial coupling of brain activity by differential equations and integrals [26]. It models the large scale behavior of groups of neurons, which basically represents the result measurable at the scalp surface with electrodes of a square centimeter, the overlapping electrical fields produced by ten-thousands of neurons at this measuring point. Figure 2.3 left illustrates the creation and measurement of the EEG signal. The synaptic currents of the pyramidal neurons in the cerebral cortex, which lies right under the skull, generate tiny electrical fields, which are measured by the electrode at the scalp surface. Only if thousands of cells contribute their small voltage, the signal is strong enough to reach the surface electrodes.

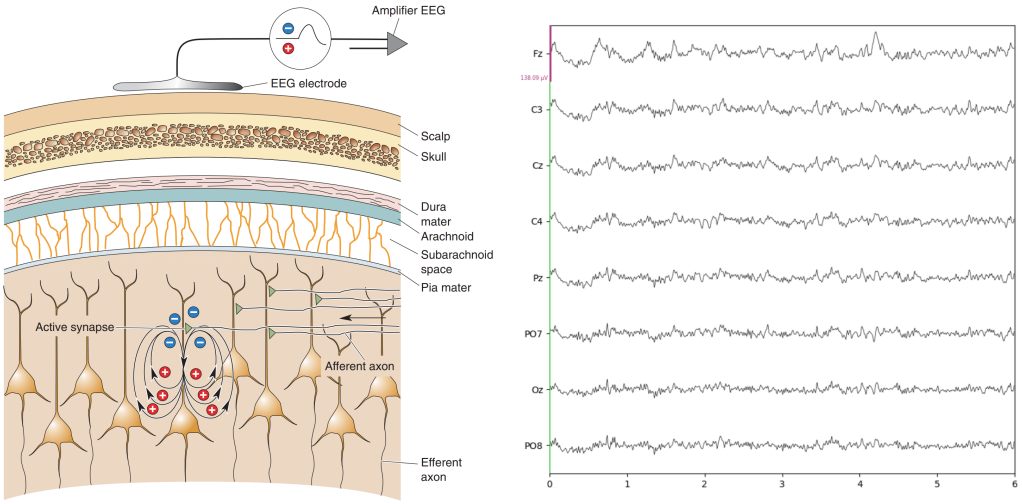

Figure 2.3: Left, measurement of the EEG as presented in [15]. The synaptic currents of the pyramidal cells generate tiny electrical fields, which are measured by the electrode at the scalp surface. Right, the output of 8 recorded EEG electrodes during one of our experiments.

The measured voltages are represented as oscillating signals over time, as shown in figure 2.3 right. These signals consist of certain frequency patterns often referred to as EEG rhythms [15]. Those rhythms are known to correlate with particular mental states as for example the level of attentiveness, sleepiness or cognitive load. Brain rhythms can vary between 0.05 and 500 Hz and can be categorized by their frequency range. The five main frequency ranges or bands are delta, theta, alpha, beta and gamma. Delta rhythms

are slow, below 4 Hz, but often large in amplitude and an indicator for deep sleep. Theta rhythms range from 4 to 7 Hz and are often experienced during sleep but also in waking states. Alpha rhythms are associated with quiet, waking states, and are situated in the range of 8 to 13 Hz. Beta rhythms are dominant in a normal wake state ranging from 15 to 30 Hz while gamma rhythms are relatively fast and a sign of an activated or attentive cortex ranging from 30 to 90 Hz [15].
Figure 2.4 shows a 4 channel EEG recording during a normal wake state with eyes closed in the first 4 seconds. We can see the alpha rhythms most dominantly present due to the closed eyes and a relaxed state of the participant. Between seconds 5 and 6 we can detect an eye blink artifact indicating the opening of the eyes and the following more dominant beta rhythms after second 6, when the participant remained eyes open [15]. A fascinating feature of those rhythms is that they are similar across most humans and can therefore be used to establish BCIs, as presented in section 2.1.2.

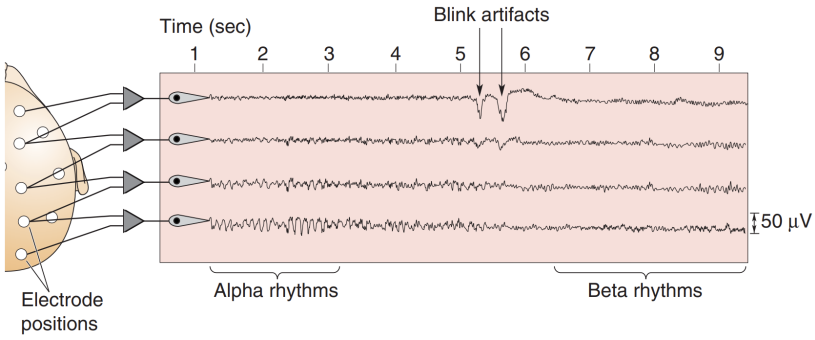

Figure 2.4: Recording of a 4 channel EEG for a subject in a wake state as presented in [15]. In the first 4 seconds, the eyes are closed producing mainly alpha rhythms. After 5 seconds the subject opens the eyes, indicated by blink artifacts, and remains with eyes open producing beta rhythms.

Although well established and researched over the last decades, the EEG comes with several drawbacks in comparison to other measurement methods for brain activity, one of them being the poor spatial resolution. A clear conclusion on the activity of a single neuron can certainly not be made from the measured signal but also not on the scale of several square millimeters of the cortex, which is possible for example with the fMRI. The spatial resolution of the EEG is defined by the number of surface electrodes used for recording which usually ranges between 32 and 64 channels, clearly depending on the use case, but rarely exceeding an amount of 256 electrodes [209]. Yamazaki et al. reported a spatial resolution of approximately 25 mm^2 with a 256 electrode dense array EEG [227]. In order to cover the scalp in a best possible way to obtain spatial information and ensure reproducibility of their experiments, researchers usually use standardized ways to position electrodes. One of the most commonly used methods is the international 10-20 system [159]. According to this system adjacent electrodes are placed in distances of 10 and 20 % from nasion to inion as illustrated in figure 2.5. Each position is identified by a letter and a number. The letter describes the area of the brain lobe, from pre-frontal (Fp) to occipital (O). Numbers are assigned related to the hemisphere of the brain, even numbers for the right and odd numbers for the left hemisphere. In the case of higher resolution systems, the intermediate sites are halfway filled by additional e.g 10% divisions, depending on the number of electrodes. This system is then referred to as

10-10 system accordingly [2]. Those systems are well established and commonly used, not even in high- but also low-resolution setups or setups targeting specific brain regions. However, even those high-resolution setups, which are only limited by electrode size and available surface on the scalp, do still measure electrical fields that travel through the cerebrospinal fluid, the bone of the scull and several layers of skin, before they reach the electrodes (see figure 2.3 left), and are therefore highly influenced by the activity of broad and distant brain regions spreading towards the surface [38]. Furthermore, this transmission through fluid and bones leads to distortions of the signal, which places high demands on filtering and signal processing techniques in order to clean the data. Additionally, the signal is prone to artifacts induced by the environment as for example electrical sources in the surrounding, movement of the head or electrodes and in general electrical signals with higher amplitude, even those produced by the human body itself like for example muscle movement of the eyes or the neck. What we receive with the EEG in the end is a highly sensitive signal in the microvolt range representing the activity of a group of neurons in the human cortex, strongly affected by environmental influences.
In order to overcome the sensitivity of the signal and improve conductivity, most high resolution devices work with water-based gel added in between the electrode and the scalp to improve signal quality. Although still being the gold standard, technology has made progress over the last decade and dry electrode solutions are capable of delivering acceptable signal quality especially in combination with new techniques for active shielding of electrodes and connecting leads [101].
Despite these drawbacks, which moreover are addressed by improved algorithms for signal processing and filtering, the advantages of this technique for BCI applications outweigh them in most cases.
First, in contrast to the poor spatial resolution of other methods, the EEG provides an excellent temporal resolution in the millisecond domain, especially compared to functional neuroimaging, such as the fMRI or fNIRS [200]. Linked to this better temporal resolution is the fact, that the EEG provides a more direct measure of brain activity, since it records the actual electrical activity of the neurons in comparison to the hemodynamic response measured by fMRI and fNIRS.

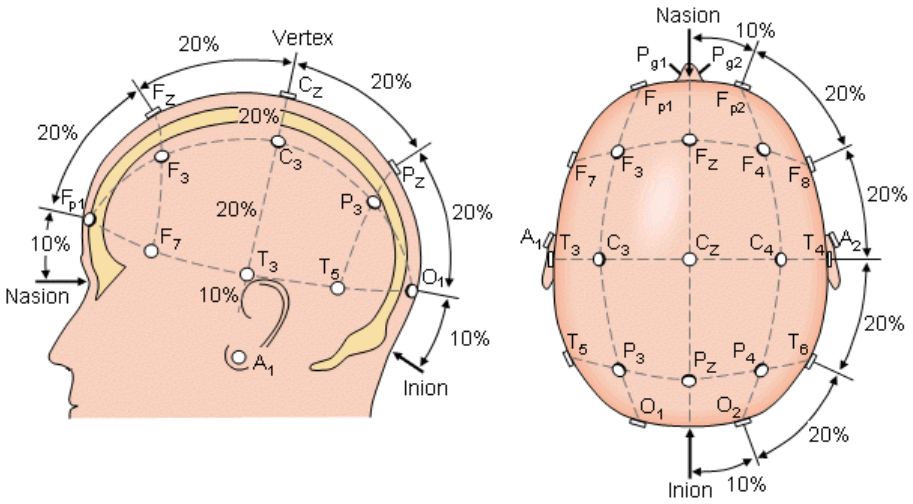

Figure 2.5: Electrode placement according to the international 10-20 system taken from [195]. The electrodes are placed in 10 and 20% steps from nasion to inion.

Second, the technology measures brain activity non-invasively, in comparison to the ECoG for example, which makes it more user-friendly, moreover supported by its flexibility and variety in hardware, since it can be integrated into various forms and shapes, for example headbands [12], earphones [107] or VR-Headsets [97].
Finally, EEG devices are comparably cheap and well researched since their discovery in the 1920s, which still makes this technology the most used measure for brain activity in the implementation of BCIs.

All of the aforementioned methods of measuring brain activity, aside from fMRI, are frequently used in literature to establish Brain-Computer Interaction. Those methods make it possible to categorize the works and applications according to their measurement methods into invasive methods (ECoG) and non-invasive methods(fNIRS, fMIR and EEG). Another way of categorizing BCIs is according to the different concepts of interaction which we present in the following.

2.1.2 Interaction Concepts

Brain-Computer Interfaces can be realized with different measurement means but also with different interaction principles. Until now we have reported on imagined movement and imagined speech, there are however a variety of other principles and ways of interaction, which allow a categorization of the different types of BCIs according to their interaction principle. A basic categorization can be achieved by grouping those different applications into exogenous and endogenous systems [168]. Exogenous BCIs rely on external stimuli to evoke a measurable brain response, for example a certain visual or auditory cue, which is then used for interface control. Endogenous BCIs measure brain patterns evoked by the user without any external dependencies as for example speech or motor movement imagination. These two groups can then be further divide into synchronous and asynchronous systems [168]. Asynchronous systems let the user interact freely independent from any kind of predefined time frames for interaction whereas synchronous systems use cues to synchronize the data analysis and restrict the users interaction to predefined time windows.
This categorization into exogenous, endogenous, synchronous and asynchronous systems is still widely used and accepted in BCI research [168, 120], however, it excludes a certain type of application, namely BCIs for mental state detection. Those BCIs try to detect the mental state of a user and adapt the interface or environment based on it. This method is neither willingly triggered by the user (endogenous) nor evoked by an external stimulus (exogenous) but rather passively evoked based on changes of the users cognitive conditions. Zander et al. tried to overcome those limitations in their work of 2009 [232] by introducing a categorization of BCIs into three types, namely active, reactive and passive BCIs, with the intention to shift the existing categorization from the user's perspective to the application side. In our opinion, these three types allow for a best possible categorization of the variety of different BCI applications which is why we will explain them in more detail in this section by giving a basic definition as well as concrete examples for each of them. In doing so, we aim to provide a broad overview on the field of BCIs to allow the reader to put our approach and achievements into perspective.

Passive BCIs

Passive BCIs describe the process of measuring the users brain activity continuously and provide controls or adjustments e.g. of their environment based on this activity, without the purpose of voluntary control [233]. The most common use case is to measure a user's mental state and refer on e.g. the cognitive workload in order to adjust the complexity of the user interface and information flow. Aricò et al. [11] used the mental workload recorded with an EEG device to adapt the UI of an Air Traffic Management (ATM) Simulator. Twelve Air Traffic Controller (ATCO) students participated in an experiment in which they were asked to perform ATM scenarios with different difficulty levels. The complexity of the task could be modulated according to the number of aircrafts to control, number and type of clearances required over time and the number or trajectory of other interfering flights. The system included several adaptive automation components which could be triggered based on the mental state measurement of the participant and allowed to adjust the interface in various ways as for example to adapt situation awareness monitoring by removing alerts or reducing visual load by removing non relevant aircraft for the sector and highlighting points of interest. A 9 channel wet electrode EEG headset was used to measure the brain activity mainly in the frontal region of the cortex of the participants. The Power Spectral Density (PSD) of the time signal was calculated by using frontal theta and parietal alpha bands for epochs of two seconds and the resulting values forwarded to a Linear Discriminant Analysis (LDA) classifier previously trained in a baseline measurement. A comparison of the calculated values with the task complexity showed, that the system was able to differentiate workload levels related to different difficulty tasks. Furthermore, the system successfully triggered the adaption of the interface mostly when the workload of the operator was high, therefore preventing overload and underload situations. Finally, the results demonstrated a significant reduction of the subjective experienced workload level and a significant increasing of task performance execution.
Another passive BCI paradigm are so called error potentials, that occur in the users brain activity as soon as they observe a behavior of the system which does not match their expectations or intention. It can be used to adjust a systems classification process and correct false decisions on the fly. Kim et al. [111] implemented a system which used EEG-based error potentials as feedback to enable the learning of adaptive behaviors of a robot during interaction. Seven subjects were asked to teach a robot gestures and the robot adapted its decisions online, based on the measured error potentials of the user. As those signals occur in the EEG of the user several hundreds of milliseconds after error detection, this method allows a rapid correction of the output of the machine, based on the user's reaction. Kim et al. were able to realize this kind of teaching process, which they call intrinsic interactive reinforcement learning, successfully in a real interaction with a robot in which subjects could freely chose gestures and the robot learned the mapping between gestures and actions during interaction. Within this scenario the error detection based on EEG signals recorded with a 64 channel amplifier using wet electrodes, achieved an accuracy of 90% and could be used for reinforcement learning in this setting.
Although passive BCIs show promising application scenarios and reliable precision in conditions in the laboratory, as illustrated in the presented studies, the technology has not yet found its way into a broader use in the real-world. One reason might be usability aspects and reduced user comfort of the EEG devices currently integrated in such systems, but probably also the need for high precision especially in scenarios targeting occupational safety or tasks involving high risks, as for example in flight control.

Reactive BCIs

Reactive BCIs are the type of BCIs which are best researched and most often applied in real-world applications already. They rely on an external stimulus which evokes a measurable response in the brain activity of a user after perceiving this stimulus [233]. Although this stimulus can be of various types (auditory, somatosensory, visual), visual stimulation is the most frequently used paradigm in reactive BCIs. Flickering stimuli produce a peak in the user's brain activity which can be used to conclude on e.g. a certain target the user was looking at. Those stimuli produce responses in brain activity as for example the P300 potential, which is the third positive peak in the EEG that occurs approximately 300 ms after the onset of a stimulus standing out from other stimuli (oddball-paradigm) [204]. Knowing this delay, one can elicit stimuli in a certain order and calculate back on the one which produced this potential in the user's EEG.
Another well-researched paradigm in reactive BCIs is the so called Steady-State-Visual-Evoked-Potential (SSVEP). This potential can be found as a peak in the frequency spectrum of the user's EEG activity of the occipital region, the area related to visual processing, when looking at a stimulus flickering in a certain frequency. If the user focuses a stimulus flickering with 15 Hz, a peak can be found in the frequency spectrum at 15 Hz at the electrodes placed around the primary visual cortex. In this case, SSVEP BCIs can be seen as sophisticated eye-tracking systems which do not depend on additional hardware (beside the EEG device) and the number of distinguishable commands is only limited by the visual field of the user, e.g. the screen size to present those flicker stimuli on. Various applications have been realized with this technology as for example a wheelchair [151] and a hand orthosis control [162] but mainly brain-speller or -typer applications [229, 36, 154]. Chen et al. demonstrated the potential of this technology for typing with brain activity in their work of 2015 [35]. They presented a joint frequency-phase modulation method together with a user-specific decoding algorithm which achieved astonishing spelling rates of up to 60 characters per minute on a 5×8 stimulation matrix resembling an alphanumerical keyboard.
A disadvantage of this method however is the use of the flickering stimuli which are tiring to look at and cause visual stress and fatigue [139]. Some research is therefore trying to disguise this flickering in movement [172]. In one of our recent works we introduced the concept of spinning icons to evoke SSVEPs [177]. The icons are rotating in a certain frequency around their vertical axis and are supposed to appear more natural and be less stressing for the human eye. Furthermore, this concept is not bound to any kind of abstract motion based pattern but rather supposed to work with any type of icon or image. The newly designed stimuli were evaluated in an application-oriented scenario and compared to standard SSVEP and state-of-the-art movement-based SSVEP stimuli regarding the classification accuracy and subjective experienced visual fatigue. The results showed that the newly created stimuli performed equally well and partially even better in terms of classification accuracy and were rated throughout better concerning visual fatigue by the study participants. This work therefore laid the foundation for more comfortable SSVEP-BCIs which can be used with basically every icon or UI element spinning around their vertical axis.
Another disadvantage of reactive BCIs is that they usually require additional equipment, since the stimuli need to be presented to the user with some sort of hardware, e.g. a display or a flickering LED which makes them to some extend stationary and cumbersome to use. Furthermore, the interaction is realized as a reaction to an interactive surface mimicking eye-tracking approaches, what makes the overall number of interaction commands in theory limitless, the commands which are distinguishable simultaneously however are still limited to the available space of the presenter hardware,

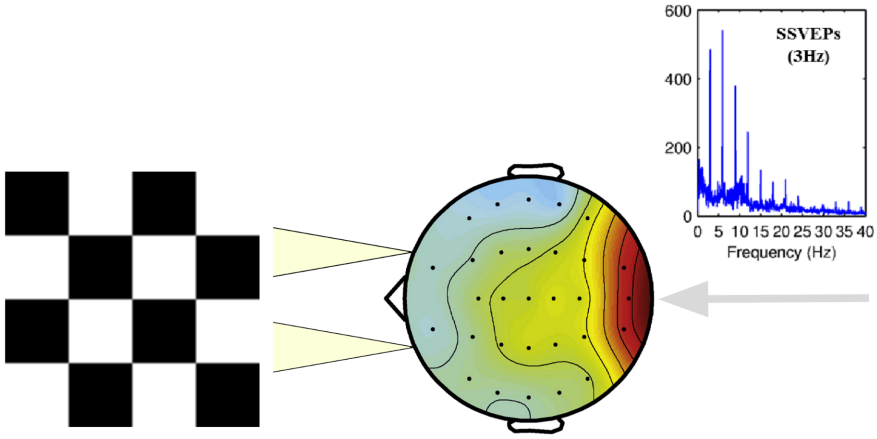

Figure 2.6: An example of SSVEP being evoked. The checkerboard pattern on the left oscillates with 3 Hz. By viewing the checkerboard pattern, SSVEP occurs in the occipital region of the head, coloured in red. The resulting potential can then be measured and displayed. A power spectrum of SSVEP can bee seen on the upper right with a peak at the fundamental frequency (3Hz) followed by the harmonics (6 Hz,9 Hz, etc.). The head and power spectrum were taken from [234] and [65]

e.g. a screen. In terms of UI design one is always bound to the trade-off between using as many controls as possible risking to overwhelm the user with too many flickering stimuli at once and reducing the number of stimuli which requires to switch to a level- or step-wise interaction process. This makes those kind of BCIs to some extend unnatural during interaction and doesn't do justice to the possibilities and expectations of most people thinking of interacting via brain activity. A more natural and intuitive way of brain computer interaction are so called active BCIs.

Active BCIs

Active BCIs probably most closely resemble the basic idea of Brain-Computer Interaction that occurs to a person with no prior knowledge on the subject. In this case, a computer uses algorithms to find patterns in human brain activity, consciously and willingly provoked by the user, in order to conclude on a certain command, trained prior to the classification process [232]. Basically, a person tries to "think" in a certain way, to evoke a specific pattern of brain activity, which the computer has learned to distinguish in previous training sessions. One of the most commonly used patterns are patterns resulting from motor cortex activity in so called Motor Imagery BCIs (MI-BCIs). Those BCIs make use of brain activity related to the control of the motor apparatus of the human body [24]. Motor imagery is a well-established BCI paradigm and uses low-frequency oscillations (< 35Hz) in the EEG, so called Sensory Motor Rhythms (SMR) [20, 92]. This activity does not occur independently but is the result of many cognitive functions in the brain patterns and has to be trained in order for a machine to detect those patterns. In the training session the users repeatedly think about moving e.g. their right hand and the computer records the brain activity during this process. In the later application phase, the users try to replicate this thought by thinking about moving their right hand again and in the best case the computer manages to classify this pattern correctly and triggers a command linked to it. Similar to the previously mentioned

reactive SSVEP BCIs, MI-BCIs have been used to establish various applications as for example controls for wheelchairs [231, 181], robotic arms [218], video games [44, 24], or drones [114, 120]. Guger et al. wanted to find out [76] how many people are able to operate a MI-BCI outside of a controlled setting in the lab and conducted a study at an exposition, where they invited people from the audience to try a BCI application with imagined right-hand versus foot movement. Ninety-nine healthy people participated in the experiment which consisted of two sessions. In the first session the subjects were advised to perform 40 trials of imagined hand and foot movement without any feedback. Based on the recorded data over the motorcortex a subject-specific classifier was trained and used in the second session in which the participants had to control a horizontal bar on a computer screen in another 40 trails. Figure 2.7 illustrates the study setup. EEG was recorded from only two channels at C3 and CZ placed according to the international 10-20-system. This method splits the surface of the scalp into parts of 10 and 20%, measured from nasion to inion [144]. With those two electrodes and a task time of 20 - 30 min on average, roughly 93% of the participants managed to achieve classification accuracy above 60% after two sessions of training with maximum accuracies of up to 90%.

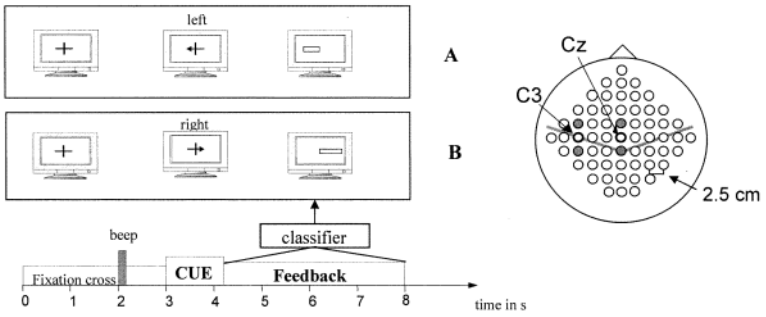

Figure 2.7: Study setup of Guger et al. [76]. The left side (A and B) illustrate the timing of the experimental paradigm for right hand and feet movement imagery. The right side, shows the electrode positions for EEG measurements from a top view on the participants head.

Although Guger et al. showed with their results, that a large population can perform a BCI operation based on motor imagery, further research indicated, that between 10 and 30% of the population is not able to control a MI-BCI [20]. A more reliable and robust active BCI paradigm is based on mental arithmetic. Performing mental calculations produces significant changes in spectral power bands of brain activity [58] and can be used to trigger simple binary interaction by detecting mental arithmetic versus non mental arithmetic in EEG activity. The advantages of this paradigm are, that it is easy to understand and to reproduce in comparison to motor imagery, as those calculations can be clearly defined and follow a certain rule set, while imagining moving an arm or hand for example, can have a variety of different types of executions, which might be performed slightly differently between individuals and sesseions. Furthermore, arithmetic operations can be adjusted to the skill of the user and are easier to validate given the users output. Bojorges-Valdez et al. [23] evaluated the performance of a continuous detection of mental computation versus resting episodes in order to simulate the conditions for an online BCI. Participants had to compute basic arithmetic operations using

positive integers while the EEG was measured with a 32 channel EEG headset. For the analysis the channels were later reduced to four electrodes tailored to the best set for the individual. Bojorges-Valdez et al. reported better classification accuracies than those of state-of-the-art MI based BCIs. Their results also showed no real effect associated with practice of the paradigm by the user, suggesting that subjects would achieve a good control of the BCI from the first operating session. However, the paradigm was not tested under real but simulated online conditions and the degrees of freedom are rather limited with the binary classification of the two mental states, arithmetic calculation vs no arithmetic calculation.

In summary, active BCIs can be described as the most intuitive and probably natural way of Brain-Computer Interaction. They can be used to establish continuous direct interaction produced intentionally by the user and do not require any additional equipment beside the headset and a PC of course for data analysis. On the other hand they either require extensive training sessions (MI-BCIs) or can only distinguish a limited number of commands (mental arithmetic).

Summary

The here presented categorization of BCIs into different types of interaction gives an overview over the broad field of application of the technology and the variety of different studies and research done over the last decades. Most of those methods are well researched and established principles which have proven to be feasible even in real-world applications. However, there are several drawbacks with some of those approaches. Reactive BCIs require additional setup for stimulus presentation and are in their way of interaction, as defined by their name, rather reactive, which makes it cumbersome to establish an intuitive interaction based on brain activity. The same holds for passive BCIs. Detecting changes in brain activity is promising and holds great potential, it is however not interactive by nature and the basic principles behind it on mental state detection needs further research.
Active BCIs are by their definition what most people might probably imagine when thinking about the interaction with a system by brain activity. Producing certain patterns of brain activity and training a machine to recognize these patterns when intentionally reproducing them, appears to be the most intuitive way of Brain-Computer Interaction. Most work in this field has been done on Motor Imagery, but as mentioned in the previous paragraph, this interaction principle has limitations as well. It is perfectly suited for navigational tasks, but if it comes to more complex interaction like controlling a smarthome or communicating with a robot, interaction based on imagined movement might become rather abstract, as these cognitive tasks usually have no direct association with the communicative intent. A more natural and flexible way of establishing BCIs can be realized by imagined or silent speech.

2.2 Silent Speech

Silent Speech, often also referred to as covert speech, is a comprehensive field of research with a variety of different paradigms, measuring methods and application scenarios. Freitas et al. give the following definition in their book "An Introduction to Silent Speech Interfaces" [61]:

> *Silent speech interface (SSI) systems allow human–computer interaction (HCI) through speech in the absence of an acoustic signal.*

This rather broad definition of the topic allows for the inclusion of a variety of different methods and research not necessarily based on brain activity measures, as the only restriction in this case is the absence of any acoustic signal. Also methods for automated lip reading or sign language classification are per definition silent speech procedures if the participants only move their lips or hands without producing a sound. According to the definition of Freitas et al., Silent Speech Interfaces sense and collect data from the human speech production process as illustrated in figure 2.8.

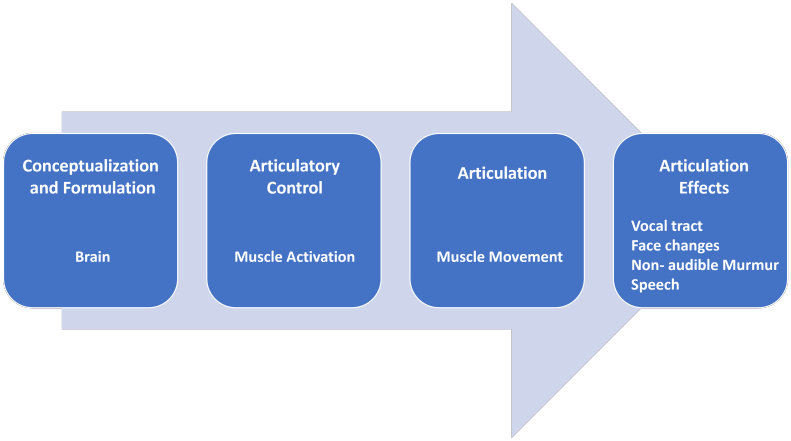

Figure 2.8: Speech production process adapted from [61] visualizing the four stages of speech production and the contributing sources or measurable effects. From the conceptualization in the brain over the articulatory control by muscle activation to the actual muscle movement in the articulation stage and finally the effects of articulation as for example face changes or speech output.

This process covers different stages of speech production according to the contributing source and measurable effect, starting from early conceptualization and formulation in the brain, over articulatory control by muscle activation to the actual muscle movement during articulation and finally the overtly detectable output of articulation effects as face changes or actual auditory output by speech. In conclusion, Silent Speech Interfaces (SSI) sense and collect data from the different stages of the human speech production process and extend it by integrating biometric signals aside from acoustic voice, measured by sensing devices such as electromyography, electroencephalography, vision, or other types of sensors to decode and further process speech. Given the different stages of the speech production process as illustrated in figure 2.8, there are various possibilities to position sensors for silent speech detection.

Starting with the articulation effects, Wand et al.[216] used video material of participants speaking without acoustic vocalization and applied deep learning algorithms on the data. With 51 different words to classify, the method achieved an astonishing best word accuracy of 79.6%. Externally placed cameras or sensors to decode language from movement of the lips however binds those solutions to stationary setups making them immobile.
The study of Wand et al. clearly targets articulation effects according to the speech production process (figure 2.8), namely the movement of the lips during speech. As speech is considered the most complex motor task performed by humans [190] it offers a great potential for silent speech detection based on the measurement of the movement of the facial muscles and therefore in the stage of articulation.
A recently published work of Wagner et al. showed a concept of radar for silent speech detection [215]. The developed solution measures the changes in the transmission spectra during speech between three antennas, located on both cheeks and the chin. With a vocabulary of 50 German words recorded for two individual speakers they reported 99.17% and 88.87% for a speaker-dependent multi-session and inter-session accuracy, respectively, when applying a long-short term memory network. Especially the inter-session accuracy seems promising and shows the potential of the technology for non-invasive silent speech interfaces. However, the speakers in the study had to vocalize the words and speak them out loud. Further tests on simply moving the facial muscles without any audible output still need to be conducted.
In their work of 2011 [217] Wand and Schulz presented a surface Electromyogramm (EMG) based silent speech approach using a phoneme acoustic model. The EMG measures the electric activation potentials of the human articulatory muscles by surface electrodes and can be used to create silent speech Interfaces, since the EMG signal is available even when there is no audible signal. Wand and Schulz trained a system on 280 utterances of 7 different sessions of two speakers which produced those utterances silently and achieved an average word error rate of 21.93% on a testing set of 108 words also silently spoken by the participants, just by moving the facial muscles. Within their work they could furthermore show, that an extension of the vocabulary of up to 2100 words was possible, which makes this technology almost suitable for spontaneous conversation. The obvious drawback of those silent speech implementations from the stages of articulation and articulation effects is, that they require the user to explicitly move the mouth and are therefore not discrete.
One promising approach targeting the earlier stage of articulatory control and muscle activation independent of visible movement of the facial muscles was developed by Kapur et al. with the AlterEgo device [105]. It measures the electrical activity of the facial muscles while speaking internally by training machine learning classifiers on the recorded data. The device is a jaw mounted electrode grid (see figure 2.9) measuring the Electromyogramm, hence the electrical activity produced by the muscles while speaking. Most fascinating however, this device does not require distinct but rather minimal movements of the speech apparatus while speaking the words silently without opening the mouth using movements of the tongue and activation potentials of the surrounding facial muscles.
The final 7 electrodes and their positions on the skin were selected from an intial 30-point grid spatially covering the cheeks on both sides around the mouth (see figure 2.9). The recorded data is bandpass filtered between 1.3 and 50 Hz and an artifact removal is applied with Independent Component Analysis (ICA). A Convolutional Neural Network (CNN) is used to classify the cleaned signal and to predict words spoken. Karpur et al. evaluated the word accuracy of the solution with 10 participants (6 female) between 19 and 31 years old in an offline experiment. Each participant was shown a total of 10 digits, from 0 to 9, randomly sequenced on a computer screen and were instructed

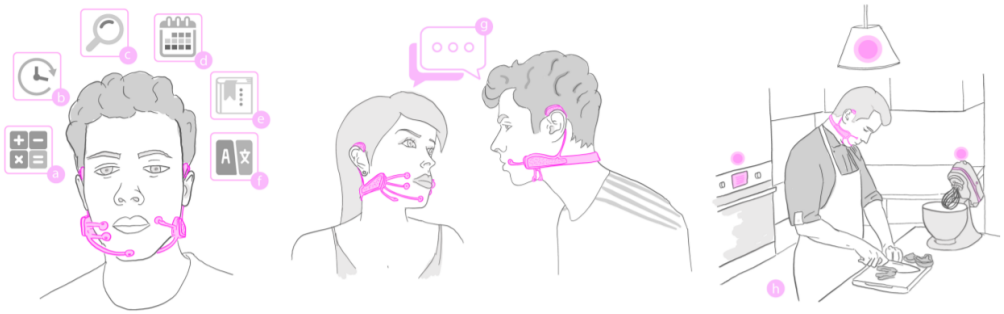

Figure 2.9: Concept of the AlterEgo solution [105]. The device has electrodes placed around the mouth, measuring the electrical activity of the facial muscles enabling silent speech for information retrieval (left), a dialogue situation (middle) and in a smarthome scenario (right).

to read those numbers silently to themselves without moving any facial muscles. The system was trained on the data of the individual and a 80/20 random split was applied for training and testing. Average classification accuracies for all users of 92.01% were achieved demonstrating the potential of the developed solution, in an offline scenario however.
In their follow-up work [106] Kapur et al. transferred their approach from healthy subjects to patients suffering from Multiple Sclerosis (MS). They collected data of 3 MS patients with dsyphonia while speaking 15 sentences silently by articulating the words in the mouth without producing any sound. The classification algorithms were again trained on the data of the individual and achieved a mean overall test accuracy of 82% with an astonishing mean information transfer rate of 203.73 bits per minute. Kapur et al. successfully demonstrated how their technology can be used in the future to help patients with speech disorders by silent speech measured with EMG surface electrodes.

The drawbacks of all of the here presented approaches are however, that they require at least a minimal capability of facial muscles movement which might not be given in the case of patients suffering from amyotrophic lateral sclerosis (ALS) or Locked-In syndrome. Furthermore, most of those methods produce a visible output, as for example in the case of lip reading or EMG based silent speech detection, visible for possible bystanders raising privacy issues and might not be appropriate in some situations.

A more unobtrusive and natural way of realizing SSIs and also the definition that we refer to in our work is imagined speech as internal vocalization or subvocalization [13]. This procedure can best be described as the inner voice of a person, similar to silently speaking in the head when reading a book for example, usually excluding any movement of the facial muscles. Although Kapur et al. claim that their AlterEgo solution works without any explicit muscle movement, the detection is still based on subtle movements of internal speech articulators, resulting from the measurement of the electrical acitvity of muscles via EMG. In order to be completely independent of any muscle movement of the speech apparatus, other measures than EMG need to be applied, as for example brain activity. Measuring the production of speech directly at the source where it is produced, according to the model of the speech production process (see figure 2.8) in the stage of conceptualization and formulation, provides a very early and clear measure, independent of any muscle movement. This approach can be described as Speech

Imagery Brain-Computer Interface (SI-BCI). In order to establish such a SI-BCI, we need a basic understanding on how speech production in the brain works, which is the topic of the next section.

2.3 Speech and the Brain

The study of speech and its representation in brain activity or specific brain regions has occupied researchers for centuries. Among the most popular is probably French neurologist Paul Broca who studied aphasias, or language disorders, in the second half of the 19th centruy already. He focused on a type of disorder in which patients were capable of understanding language but could not construct grammatically correct sentences. Post-mortem examination of these patients revealed invariable lesions in the left frontal area of the brain, later named after him as Broca area, which lead him to the conclusion, that speech production is located in the left hemisphere [10].
Another aphasia named after German neuropahtologist Carl Wernicke, describes a pathology in which people have problems with language understanding. Those people speak freely and may use grammatically correct sentences but the content of speech is unintelligible or incoherent usually manifested by the use of random words or wrong substitution of words, as for example "television" instead of "telephone". Wernicke discovered lesions in the superior temporal gyrus of the patients placed among the sensory, visual, auditory and somatosensory areas of the cortex (see figure 2.10). This region was later named after him as Wernicke area [222].
Wernicke was able to support the findings of Broca concerning the dominance of the left hemisphere of the brain in speech production, which was further confirmed with improving medical imaging techniques in large scale studies. Literature suggests that language function is predominantly restricted to the left hemisphere in over 90% of right-handers and as many as 60-70% of left-handers [31].

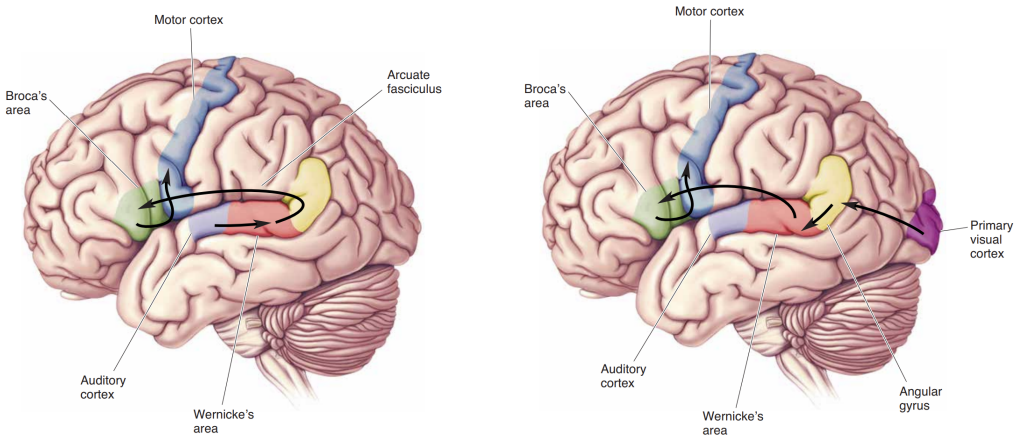

Figure 2.10: The Wernicke-Geschwind Model as presented in [15]. The arrows indicate the pathways of spoken language comprehension (left) and written language comprehension (right).

In contrast to Broca's area, which is responsible for speech production, Wernicke's area is involved in understanding language. The collaboration of these two areas was described in an early neurological model of language by Norman Geschwind in the 1960s as the Wernicke-Geschwind model [66]. The model mainly consists of two pathways for spoken and written word comprehension.
Auditory input is received via the ear and in a first stage processed in the primary auditory cortex. From there the information is forwarded to Wernicke's area which tries to make a meaning out of it. In a second step the information is send to Broca's area which passes the information to the motor cortex to produce speech based output as response to the received auditory input (see figure 2.10 left).
Information extracted from written words arrives in the occipital region of the brain at the primary visual cortex and is then forwarded to the angular gyrus, which translates the written words into the corresponding auditory signal and sends it to Wernicke's area (see figure 2.10 right). A motor response to this audirory stimulus is once more triggered by Broca's area which receives information from Wernicke's area and forwards this information to the motor cortex, which produces an output if necessary.
This early and very simple model can however not comply with the complex neural organization of speech in the brain as we know about it today, and is therefore considered obsolete [208]. In the meantime we know that speech is not strictly limited to the left or the dominant hemisphere, but also involves the right or non-dominat hemisphere [94]. Current research even questions established concepts as the role of Wernicke's area in speech perception in general and its function of word comprehension, rather showing evidence that it supports retrieval of phonological forms, which are used for speech output [18]. Tremblayab and Dick even go one step further and claim that "there is no consistent definition of Broca's and Wernicke's areas, and the terms should no longer be used."[208].
A more recent model of speech processing in the brain is the dual-stream model formalized by Hickok and Poeppel [83]. The model suggests that there are two possible sources of predictive coding in speech perception, the motor speech system and the lexical-conceptual system [81]. The model explains the functional connectivity of certain brain regions during aurally presented speech, comprising a ventral and dorsal stream, similar to the dual-route theory of the visual pathway by Goofale and Milner [70].
Figure 2.11 illustrates the model and the different streams, with the ventral route being responsible for speech comprehension, involving structures in the superior and middle portions of the temporal lobe. The dorsal stream is involved in translating acoustic speech signals into articulatory representations, essential for speech production, and involves structures in the posterior planum temporale region and posterior frontal lobe [83]. In contrast to Broca's and Wernicke's early work on lesion mapping resulting in the belief that speech processing is mainly left hemisphere dependent, a wide range of evidence suggests that the ventral stream is bilaterally organized while the dorsal stream, is strongly left-dominant [81].
Although the model of Hickock and Poeppel had some shortcomings concerning scalability and the inclusion of syntactic processes, the basic model is widely accepted and further developed to apply to the majority of linguistic and neuroanatomic dependencies [82]. As illustrated figure 2.11), the dual-stream model involves a widely distributed conceptual network for speech processing, supporting assumptions that brain activity responsible for those processes is classifiable as distributed spatial activity patterns. This idea of classifiable patterns related to speech, together with the fact that spoken and imagined speech trigger similar brain regions, as discussed in more detail in chapter 3.2, drives the research on Speech Imagery Brain-Computer Interfaces.

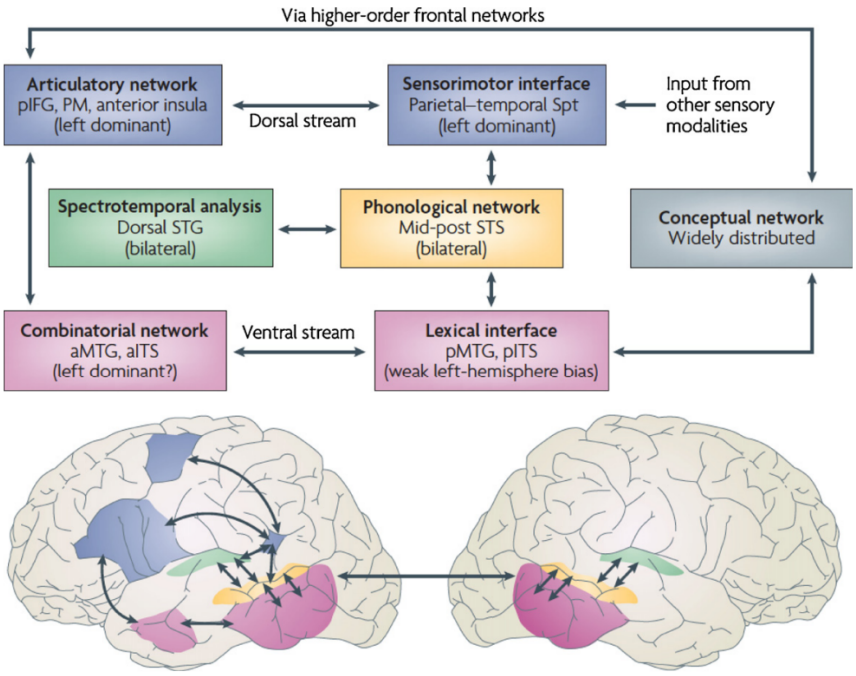

Figure 2.11: Dual-stream model of speech perception as presented in [83].

2.4 Speech Imagery Brain Computer Interfaces

Speech Imagery Brain-Computer Interfaces (SI-BCIs) try to classify imagined words based on brain activity alone. In our opinion, imagine speaking words to oneself silently in the head, is one of the most intuitive, flexible and direct forms for establishing Brain Computer Interaction.
Research in this field dates back to the time of Hans Berger and the discovery of the EEG in 1928. It is said, that his initial goal behind the development of the EEG was to enable synthetic telepathy in a kind of mind reading approach [104].
Since back then researchers have been trying to measure and decode the underlying neural correlates of speech and developed brain activity based speech recognition systems in various different ways. However, throughout the literature and studies one can find various names and definitions for speech imagery BCIs and the underlying concepts, as for example covert speech, imagined speech, subvocalization, inner speech or silent speech. Those definitions are interchangeably used in different contexts for differing paradigms, making it difficult to find a clear separation and categorization based on the literature. Actually there is no real common agreement in the community on how to properly categorize and define the different paradigms.
One recent, but certainly not the only example, is a publication on imagined speech classification from EEG signals, which uses the term "Silent Speech" in the title, "Envisioned Speech" in the abstract and "Imagined speech" in the introduction [3]. While Silent Speech can be considered a generic term, envisioned and imagined speech are clearly two different paradigms, which will be explained in more detail in section 2.4.1.
This confusion can frequently be found to different extends in literature [143, 63, 67] and also in our early works we refer to Speech Imagery BCIs as Silent Speech BCIs.

Given the definition of silent speech as speech in the absence of any audible sound (see section 2.2), this can not precisely describe our technique, as one could also measure brain activity while a participant is silently vocalizing certain words. This paradigm would then mainly focus on the motor cortex and the classification of neural patterns of the facial muscles. Using imagined speech as paradigm to establish a silent speech BCI, would on the other hand make the definition valid again.
However, due to the possible confusion and mix-up with muscle activity based silent speech interfaces, we decided to change our wording to Speech Imagery BCI (SI-BCI), inspired by imagined movement being called Motor Imagery BCI (MI-BCI).
This confusion beyond the circles of the BCI community and its extent is also obvious when having a look at the first sentence of the Wikipedia articles of imagined speech, subvocalization and silent speech interfaces.
Imagined speech is defined in this article as:

"Imagined speech (also called silent speech, covert speech, inner speech, or, in the original Latin terminology used by clinicians, endophasia) is thinking in the form of sound – "hearing" one's own voice silently to oneself, without the intentional movement of any extremities such as the lips, tongue, or hands." [7]

while subvocalization is presented as:

"Subvocalization, or silent speech, is the internal speech typically made when reading; it provides the sound of the word as it is read." [8]

and silent speech interfaces as:

"Silent speech interface is a device that allows speech communication without using the sound made when people vocalize their speech sounds. As such it is a type of electronic lip reading." [9]

Although we would mostly agree with the definition of the different paradigms, it is obvious that the terminology interchanges between those definitions which makes it hard to distinguish between the different concepts, which is reflected in literature.
In order to shed light on the partially confusing variety of different approaches of SI-BCIs, we will in the following present different techniques categorized based on the acquisition modality, the concept of interaction and the paradigm used to produce imagined speech.

2.4.1 Speech Imagery BCI Paradigms

Speech Imagery BCIs can be established by different means and processes which makes it important to clearly define a certain paradigm to apply for the user. An unsupervised study setup asking ten participants to produce imagined speech without any further advise on the technique would probably lead to ten different individual ways of imagined speech production. Researchers have investigated various different paradigms for producing and classifying imagined speech, which can roughly be categorized into imagined and envisioned speech (see figure 2.12).

[7] `https://en.wikipedia.org/wiki/Imagined_speech` Last accessed: 07.07.22
[8] `https://en.wikipedia.org/wiki/Subvocalization` Last accessed: 07.07.22
[9] `https://en.wikipedia.org/wiki/Silent_speech_interface` Last accessed: 07.07.22

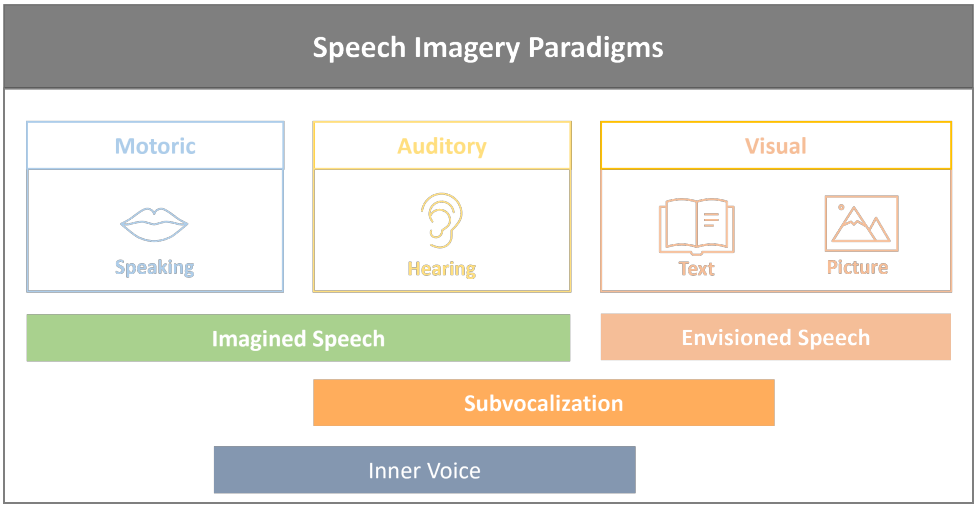

Figure 2.12: Categorization of speech imagery BCI paradigms. The two main categories are envisioned speech, involving visual paradigms, and imagined speech, focusing on motoric and auditory activity. Subvocalization describes a combination of the two, experienced while silently reading. Our paradigm depicts the inner voice, described as similar to reading words to oneself, therefore including parts of motoric, auditory and visual paradigms.

The concept of speech visualization or envisioned speech can be explained by visualizing seeing the written word or picture of a certain object in front of the inner eye. Kumar et al. [116] advised their participants to imagine different objects previously shown on a screen with their eyes closed. Although Kumar et al. did not specifically mention the technique in detail, such an object imagination is usually accompanied with an unconscious pronunciation of the imagined object and can therefore be listed as a speech imagery paradigm. However, the lack of detail on the study design and the discussion arising with it highlights the problem with this technique, that a clear separation from actual object visualization as presented in [123] is not possible. This fact and the problem of providing a proper description to the participants on how to produce envisioned speech correctly and reproducible, lead us the decision to exclude this paradigm from our study setups.
Imagined speech on the other hand is the imagination of speech itself and can be separated into motoric, auditor and approaches combining motoric and auditory speech imagination.
Starting with motoric speech imagination, in this case the user does not visualize the written word or a picture of the object but imagines the actual movement of the facial muscles while speaking a certain word or vowel [50]. DaSalla et al. [47] instructed their participants to perform two tasks targeting different processes of speech production during their study. On a visual cue they were advised to either imagine mouth opening and vocalization of the vowel "a", imagined lip rounding and vocalization of the letter "u", or in a control condition to perform no action at all. Features were extracted with the Common Spatial Pattern (CSP) method. The basic principle behind CSP is to apply a linear transformation to project the multi-channel EEG signal data to a lower-dimensional spatial subspace. The transformation results in the maximization of the variance of one class while minimizing the variance of other class at the same time [4]. Based on

those features, DaSalla et al. performed binary classification, vowel against rest or vowel against vowel. The overall classification accuracies ranged from 68 to 78%, which showed that motor cortex activations associated with imagined speech based on vowels can be classified.
Although this paradigm can be considered intuitive and well reproducible, one could claim that it appears cumbersome and in the worst case exhausting having to imagine the specific movement of facial muscles without actually moving them.
A more unobtrusive paradigm targets the auditory processes during speech perception and production and requires the participant to imagine hearing the word either spoken by oneself or somebody else.
Morooka et al. [148] conducted a study in which three healthy men imagined hearing six sounds, "/a/i/u/e/o/mute,", while the EEG was measured with 8 electrodes over the left hemisphere, targeting speech related areas as for example Broca's area. The vowels were presented auditory and the participants advised to "repeat" the presented sound silently in their head. Different classifiers were trained to perform binary classification of the vowels against each other and Morooka et al. were able to achieve average classification accuracy of 79.7% with a Support Vector Machine (SVM) classifier. From the usability perspective we see a problem with the method presented in the paper, since the user is forced to remember the specific sound the classifier was initially trained with. Varying perception or sound imagination during repetitive use might have an influence on the brain patterns and therefore the classification accuracy and the performance of such systems. We therefore recommend to use written cues instead of auditory ones and advise the participant to imagine hearing them speak the word to themselves, like with their inner speech or voice.
This leads us to our last paradigm, subvocalization which can be considered a combination of auditory and text based visual paradigm where the participants are advised to imagine reading the word to themselves. It takes into account that silently reading a text will also always lead to a certain auditory perception inside the brain [199]. At this point, the borders between the different paradigms start to vanish and from a user perspective it is getting confusing. The difference between imagine hearing oneself speaking a word or silently reading it are barely distinguishable. However, the most important aspect on the paradigm in SI-BCIs and active BCIs in general is reproducibility. Participants need to be able to reproduce the same thought as precise as possible over and over again during training as well as during the actual later interaction. The better they are capable of reproducing it, the better the pattern recognition in the classification process and therefore the performance of the system. Reading the word to oneself is usually a quite stable process and should not vary too much in between recording sessions, however, we see a problem with the visual aspect of the paradigm. At a certain point, users might accidentally switch to envisioned speech by imagine seeing the written word like during reading.
In our study setups we therefore advise our participants to use their inner voice to repeat the presented word. As an additional description we tell them, to produce it just like reading the word to themselves without moving any muscle and producing any audible sound. This definition does not clearly draw a line between auditory imagined speech and subvocalization, however, gives a clear description of the paradigm, making use of techniques that most people are probably aware of and able to reproduce reliably. Although partially including motoric activity, it shifts the focus from envisioned speech and imagined movement of facial muscles to actually freely "speaking from the mind" as most people probably do when practicing presentations or, as previously mentioned, when reading a book.

Our approach as illustrated in figure 2.12 therefore uses a combination of motoric, auditory and visual paradigms, with a focus on the auditory component, to trigger a broad network of brain regions.
Nevertheless, one needs of course to be careful with the degree of freedom given to the participants in imagined speech production. Different paradigms trigger activity in different brain regions and in case of an attempted cross-subject classification, those should be aligned between participants of course. Focusing solely on imagined movement of facial muscles would mainly target the motor cortex and envisioned speech the visual cortex of the brain alone (although partially involving other regions as well) contradicting our goal of recording brain activity from broader distributions of brain areas for imagined speech as described in section 2.3.
In our opinion the description and paradigm of the inner voice is the most intuitive, understandable and best reproducible paradigm for speech imagery and also BCI paradigms in general. Furthermore, most of our studies were concerned with subject-wise classification, training and testing a classifier on a single participants data, due to the novelty of the developed methods. Therefore, all of our studies concerning EEG-based Speech Imagery BCIs, presented in the chapters 3, 4 and 5, make use of the inner voice paradigm as a combination of imagined motoric and auditory speech as well as subvocalization. For the sake of simplicity and due the large overlap, we refer to this paradigm to our participants as inner voice, and throughout this thesis as **imagined speech**.

2.4.2 Speech Imagery BCI Concepts

Beside different paradigms used for Speech Imagery BCIs, one can categorize existing works according to the concept of speech representation that is used. Existing work mainly focuses on three different categories, vowel or syllable, word and sentence imagination. Studies involving silent repetitions of vowels or syllables aim at decoding smaller fractures of words which can later be concatenated to create commands for BCI applications, as we have seen in the last section [47, 148] and will see in several studies in the following section. This concatenation can either involve the classification process, as for example having a classifier trained on vowels or syllables and applying it on brain activity recorded during silently speaking a word, or it can be applied on the interaction side, when for example having the user produce each syllable step-wise to form commands. Those approaches show promising results in terms of classification accuracy, however, an interaction based on syllables or vowels alone contradicts the natural and intuitive interaction concept of SI-BCIs.
The classification of single words holds more potential as it lets the user speak out a whole command at once which allows for more fluent and natural interaction.
An even more fluent and natural concept is the detection of an imagined sentence. However, classifying whole sentences in speech imagery is a rather rare and experimental topic. Beside the regular challenges like finding suitable features, achieving sufficient signal quality and the accurate time window in which the participant produces the silent output, the detection of whole sentences adds parameters like speed and rhythm of pronunciation to the list. This long list of challenges lead most research in the field to avoid this technique and focus on assembling command sentences based on single detected words or assemble words based on syllables etc. Nevertheless, there is existing work as for example Deng et al. [51] who tried to find correlations between heard and imagined sentences within a project for the Defense Advanced Research Projects Agency (DARPA) in 2012.

They recorded the voice of eight English speaking subjects while speaking six sentences from the DARPA TIMIT corpus. The six sentences were chosen to minimize the interclass correlations between sentence power envelopes as:

1. Steve collects rare and novel coins.
2. Herb's birthday occurs frequently on Thanksgiving.
3. Jane may earn more money by working hard.
4. The eastern coast is a place for pure pleasure and excitement.
5. Rock and roll music has a great rhythm.
6. The government sought authorization of his citizenship.

Those recorded audio files of the participants speaking the sentences were later presented to the subjects, again with the instruction to repeat them silently in their head afterwards. Each participant was presented his own voice and had to perform 89 repetitions per sentence leading to a total sequence of 534 repeated sentences collected within 1.5 hours. The EEG was recorded with 128 channels in a dimly-lit room with eyes closed. Deng et al. applied Independent Component Analysis (ICA) on the EEG data and calculated the envelope following response (EFR) to identify sentences. This anaylsis led to a classification which was only significant for two of the eight subjects and could only be increased by artificially concatenating the EEG data of several sentences within the training set. They conclude that "The strength of EFRs found in this EEG study seem unlikely to support the function of a Brain-Computer Interface (BCI) which works in real time to identify heard or imagined speech streams" and suggest to switch to invasive methods like the ECoG to improve spatial resolution.

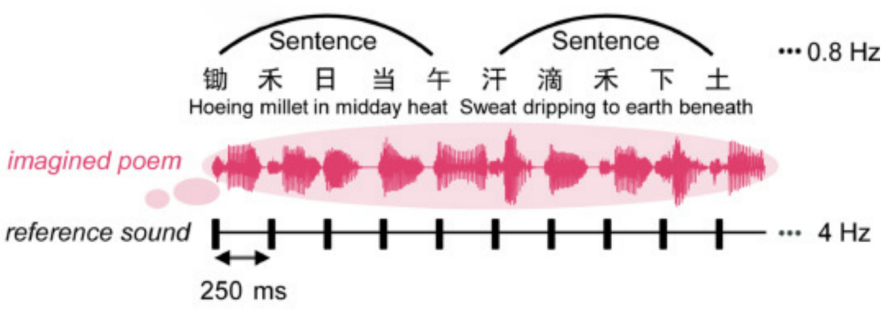

Figure 2.13: Study procedure of [134] for the poem imagery part, illustrating the reference sound for proper rhythmic repetition at the bottom, and the imagined poem in the top row.

One recent approach tried to classify imagined sentences in the form of poems from brain activity [134]. Lu et al. used MEG and a frequency-tagging paradigm which allows for the detection of neural signals that change periodically over time and can therefore be used to detect the rhythm of internally constructed processes. 24 participants listened and silently repeated three traditional Chinese poems each containing 20 syllables and

every five of them forming a sentence. Participants were seated in a dimly lit, magnetically shielded room and the MEG (see section 2.1.1) was recorded with a whole-head MEG system with 306 channels. The study procedure requested the subjects to repeated the poems silently in their head, following a reference sound indicating the rhythm of each syllable of the poem as illustrated in figure 2.13. Lu et al. observed similar neural tracking of imagined and perceived speech induced by the rhythmic constructs of the poems and overlapping neural cohorts. However, their study was conducted via MEG which needs intensive cooling and shielding and is not applicable in real-world BCI applications.
In general, the detection of whole imagined sentences from brain activity might be possible, but appears to this day too challenging with non-invasive portable measures like the EEG. Furthermore, single word detection can be used to construct short command sentence as for example the instruction for a robot to pick up a screw could be assembled by thinking the words "pick" and "screw". Taking those findings into consideration and the more positive results in word-based imagined speech BCIs, as we will see in more detail in the following sections, we decided to proceed our work on the word level and classify single imagined words from EEG activity. Therefore our study setups and implementation presented in this thesis are throughout concerned with word imagination.
All the different concepts described in this section are widely used in the field of SI-BCI research and are applied in approaches using different measurement techniques. A further categorization of related work can be done according to these measurement techniques as provided in the next section.

2.4.3 Speech Imagery BCI Methods

Brain activity can be measured in various ways (see section 2.1.1) and therefore SI-BCIs can be realized with many different measurement methods. According to those measures, we will present in the following related work on Speech Imagery BCIs separated into the different techniques. These works will mostly be considered with word based imagined speech, which we have foreseen as the paradigm of choice for our study setups and implementations due to the reasons explained in the last section.

Electrocorticography (ECoG)

Concerning brain signal measurement, the Electrocorticogramm provides the clearest and most valuable signals. Measuring the activity right at the surface of the cortex, yields to minimum distortions in the signal compared to the EEG where brain activity is measured at the scalp surface. A variety of studies therefor investigated imagined speech approaches measured by ECoG signals.
One of the first works in the field goes back to the year 2004 when Leuthardt et al. enabled users to control a one-dimensional computer cursor by imagined movements and speech, measured with ECoG [129]. Four patients with intractable epilepsy were invited to participate in the study. Each of the patients underwent temporary placement of intracranial electrode arrays to localize seizure foci and had a 48- or 64- electrode grid placed over the left frontal-parietal-temporal region. Brain signals were recorded from a common set of 32 electrode positions between subjects. The participants were given six tasks namely opening and closing right or left hand, protruding the tongue and saying the word "move" which had to be actively performed, and in a second trail imagined to be performed. Over brief training periods of 3 - 24 minutes, the participants then used the imagined signals to control a cursor on a screen in a one-dimensional binary task.

Leuthardt et al. reported a success rate of this control of 74 - 100% depending on the paradigm and could demonstrate the first use of ECoG for online operation of a BCI. With this first work being rather experimental and including motor imagery next to speech imagery, Leuthardt et al. shifted their focus in a follow up work to speech imagery alone [128]. They repeated their study setup with four participants and the same cursor control with imagined phonemes OO, AH, EE and EH. Different combinations of phonemes for binary cursor control (left and right) were investigated with a highest final classification accuracy of 91% for classifying EE versus OO. Notably, those accuracies ranging from 68 to 91% could be achieved within a 15 min training procedure.
Over the years researchers have build up on this approach and managed to decode words [141] but also short sentences [9, 138] from the ECoG signals. One of the most recent and also most impressive works in the field of ECoG-based imagined speech BCIs was provided by Moses et al. [150]. Within their work, they used a high-density multielectrode array over the area of the sensorimotor cortex to record brain signals of a person with paralysis, who was unable to speak. A vocabulary of 50 words enabling the generation of over 1000 sentences was trained over 48 sessions resulting in overall 22 hours of neural recording of imagined speech. Deep-learning algorithms were used to create models and classify imagined words in combination with a natural-language model to determine next-word probabilities. With this setup Moses et al. were able to decode the imagined words in real-time with a median accuracy of 74% at 15 words per minute and a best performance of up to 90% accuracy at 10 words per minute. This real-time sentence decoding pipeline enabled a patient with severe paralysis to communicate via imagined speech, classified from his brain activity alone with a precision and pace never demonstrated before.
The results of these studies show that it is possible to successfully detect and classify imagined speech from brain activity even beyond word-based classification with whole sentences and in a real-time online classification. However, ECoG requires a brain surgery in order to place the electrode grids on the cortex which makes it inconvenient for any user who does not have to undergo such a surgery anyway. Targeting real-world use with healthy patients makes it necessary to transfer these concepts to non-invasive measurement methods.

Magnetoencephalography (MEG)

One of these non-invasive brain activity measures is the Magnetoencephalography (MEG) which measures the magnetic fields emitted by the currents produced by the neurons of the brain. As presented in section 2.4.3 on the different speech imagery BCI methods, most of this work is concerned with the detection of short sentences. Lu et al. [134] tried to decode sentences in poems based on different neural correlations with the rhythm of the sentences, while Dash et al. [49, 48] tried to create speaker independent BCIs and decoded 5 imagined short sentences from MEG. Within their work from 2020 [48] Dash et al. even reported an outstanding average decoding accuracy of up to 93% for the 5 imagined short sentences achieved with a Convolutional Neural Network implementation. Although promising, this measurement method has one major drawback, the required intensive magnetic shielding of the measurement environment. In contrast to fMRI devices this technology might develop into portable solutions in the future, however, until the point of writing this thesis, the existing solutions are in a rather experimental and prototypical state [25, 164] and although being independent of bulky cooling systems, still require the environment of a magnetically shielded room, making it a stationary setup unsuitable in real-world scenarios.

Functional Magnetic Resonance Imaging (fMRI)

The fMRI measures the hemodynamic response of the brain as changes in oxygenated blood flow, and can therefore be used to establish Brain-Computer Interfaces. However, the majority of related work in the field focuses on basic research investigating the underlying principles and neural correlates of imagined speech processes rather than classifying them [196, 72, 194]. Those publications deliver ground braking insights on the activated brain regions during overt and imagined speech production, but none of them attempts to classify words from the recorded brain activity in order to establish a Speech Imagery BCI. This might be due to the poor temporal resolution of the fMRI and the bulky, stationary setups, making them unsuitable for BCI applications. Still, there is one work of Yoo et al. [230] which tries to establish a BCI including imagined speech. Within their study, Yoo et al. enabled their participants to control a cursor and thereby navigate through a 2D maze. The four directions of the cursor could be triggered by performing one of 4 distinct mental tasks, namely, mental calculation, imagined speech, imagined left and right hand movement. Those 4 paradigms are known to activate spatially distinct brain regions which can be classified based on fMRI data. Each of the paradigms had to be performed over a certain period of time, resulting in an overall command time of 2 min 15 seconds, including 1 min 51 seconds fMRI scan time. 3 of the 4 participants were able to navigate without any errors, while the fourth participant managed to trigger commands with a 92.3% accuracy. Those results are impressive, however, the interaction was not solely based on imagined speech but included other mental tasks as well. Furthermore, the comparably long classification time and the stationary setup exclude this technology from the use in real BCI applications.

Functional Near-Infrared Spectroscopy (fNIRS)

A rather recent non-invasive technology, at least in its portable form, is the functional Near-Infrared Spectroscopy (fNIRS). It measures changes of oxygenated blood flow in the brain to conclude on active areas and has gained momentum in brain based speech analysis in the past. First evidence that perceived speech can be identified from the listener's brain signals was shown in [131] and even imagined speech could be decoded from fNIRS already. Sereshkeh et al. [193] developed an online 3-class BCI which enabled participants to answer questions by imagining rehearsing the word yes or no for 15 seconds. The system was evaluated with 12 participant who performed one offline training block and six online blocks over the course of two sessions. A regularized discriminant analysis classifier was trained to detect the changes in oxygenated hemoglobin concentration for the imagination phase of the two words and the resting state. In the final online block 9 of the 12 participants performed significantly above chance level with an average accuracy of 83.8% $\pm$ 9.4%. With their work Sereshkeh et al. could show a working online 3-class imagined speech BCI based on fNIRS signals for the first time. Although impressive this study shows one major drawback of the technology which is the rather poor temporal resolution. fNIRS measures changes in blood flow causing a delayed measure of neural activity due to the hemodynamic lag, as oxygenated blood needs to be delivered to regions with increased activity [91]. This delayed measure makes it inconvenient for real-time speech synthesis [174]. Although Sereshkeh et al. managed to implement an online system it required the users to rehearse the desired word for 15 seconds in order to trigger an interaction which limits the application to fields where reaction time is not the priority.

Electroencephalography (EEG)

The most convenient way for establishing an imagined speech BCI and BCIs in general, is the Electroencephalography. It measures the summed up activity of the neurons at the scalp surface (see section 2.1.1). This method is known to be prone to artifacts by electromagnetic fields and other electrical signals produced by the body (e.g. muscle activity) or just simple movements of the user and therefor the acquisition system, but it can be measured non-invasively at the scalp surface with comparably cheap devices and has an excellent temporal resolution. Recent advances in machine learning and artificial intelligence have improved filtering techniques which makes it possible to consider EEG as a method for BCI applications even in real world scenarios. In the field of SI-BCIs many attempts have been made to decode silent or imagined speech from brain activity as presented in detail in the next section.

2.5 EEG-based Speech Imagery BCIs

The earliest work on EEG-based Speech Imagery BCIs goes back to 1997 when Suppes et al. presented their study on "Brain wave recognition of words" [202]. In their experiment, they recorded electrical and magnetic brain waves of seven subjects, while they were listening to, and imagined speaking, 7 command words, randomly presented via headphones in the auditory, and via a screen in the imagined speech condition. Electrical activity was recorded with 16 EEG sensors arranged according to the 10-20 system at positions F7, T3, T5, FP1, F3,C3, P3, Fz, Cz, FP2, F4, C4, P4, F8, T4, and T6. 100 trails were recorded per word and condition. The 100 trails per word were split in half and the first 50% averaged, in order to create a prototype signal for this specific word. This prototype signal was then compared to a test sample in the later prediction. The test samples were created by taking the remaining 50% of the data, splitting it in 5 equal parts and averaging it, to result in 5 test wave forms per word. The prediction was done by comparing one of the test samples with the created prototypes and selecting the best fit, based on a standard minimum least-square criterion. The comparison was done per channel and the best result later on chosen for analysis. Additionally, Suppes et al. performed filter parameter selection by testing a wide range of upper and lower bounds for their bandpass filters per subject and selected again, the best performing frequency for the analysis. With this setup Suppes et al. could show that for all but one of the predictions they achieved results significantly different from chance but with a wide variation. They conclude their work that "brainwaves carry substantial information about words of which subjects are made consciously aware" and that machine learning methods might in the future overcome the obvious drawbacks of their current method as for example averaging several trails for classification and the intensive filter parameter and electrode selection processes.
Although showing clear evidence for the possibility of classifying imagined speech from EEG data, and beside another paper of Suppes et al. in the following year, in which they repeated their study setup with sentences [201], there was no real follow-up on this approach until 2006, when Wester and Schulz published their work on "Unspoken speech – speech recognition based on electroencephalography" [223]. In their study they used the same number and positions of electrodes as [202] and had participants imagine speaking ten words while the EEG was recorded. The given implementation was able to recognize imagined speech from the recorded EEG signals at an impressive performance, with word error rates on average 4–5 times higher than chance. However,

the stimulus presentation was done in blocks, meaning, that all repetition of one word, were done in a row, resulting in temporal effects in the data detectable by a classifier, as later proven by Probadnigk et al. in 2009 [169]. They repeated the study setup of Wester and Schulz once with the given block-wise stimulus presentation and for three other conditions, namely, randomized, sequential and short block stimulus presentation. In the randomized condition, words were presented randomly over the whole recording. Short blocks were recorded with five repetitions of a word in a row and the sequential condition included concatenating all words in a sequential word order. They invited 21 participants and gave them 5 words to silently repeat while the 16 channel EEG was recorded. Their results showed, that although the block presentation led to classification accuracies of 45.05%, classifying on the data recorded during the other three conditions did in no case exceed chance level, clearly indicating the label leakage occurring in the block-wise presentation, and the importance of the study design in such experiments in general.
The next relevant work in the field, reporting results above chance level, can be considered the study of Torres-García et al. [63] from 2012. EEG data was recorded from 14 electrodes for 21 subjects, while they imagined speaking the five Spanish words “arriba” (up), “abajo” (down), “izquierda” (left), “derecha” (right), and “seleccionar” (select), each 33 times. The selection of words aimed at a later scenario of a cursor control with imagined speech. Discrete Wavelet Transform (DWT) was applied as feature extraction algorithm, to extract information from the frequency bands below 32 Hz. The extracted features were forwarded to 4 different classifiers, namely Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) and a Bagging-RF. Results of the offline analysis showed the best performance for RF and Bagging-RF, significantly above chance level with classification accuracies of up to 55%. However, next to some missing details on the results, as for example the average classification accuracies per classifier, Torres-García et al. are quite careful and reluctant in reporting these impressive results as "evidence to affirm that the EEG signals actually carry useful information to allow the classification of unspoken words". This restraint might be due to two reasons. First, the recorded dataset was comparably small with only 33 samples per word and the applied 10 fold cross validation, shrinking the testset even further. These small testsets result in an increased significance threshold, which will be explained in more detail in section 2.9, and the reported results ending up being only slightly to not significantly above chance level. Second, the hardware used for recording in the study was rather a consumer grade device, the Emotiv EPOC[10]. Back in 2012, the device was freshly released in a first version as an affordable tool for EEG measurements in home use for everybody. Beside the questionable quality of these first cheap consumer EEG-headsets, the electrode positions are fixed to specific regions, not homogeneously distributed over the scalp. With two of the overall 14 electrodes placed on the forehead and another two at the temples, possible accidental facial muscle movements while producing imagined speech, could have had an impact on the measurement, by distinct patterns of electrical muscle activity, recorded from those electrodes. In general, the actual contribution of those electrode positions to imagined speech production remains questionable.
The effect of electrode positions on EEG-based imagined speech classification was also part of the work of Wang et al. in 2013 [219], in which they coined the term Speech Imagery BCI. They used two Chinese characters representing the words "left" and "one", and advised their participants to read those characters silently. This illustrates once more the inconsistency in stimulus paradigms and definitions as mentioned in section 2.4.1. Although referring to speech imagery, the participants were advised to read the

[10] `https://www.emotiv.com/epoc/` Last accessed: 12.07.22

words silently to themselves which is related to speaking but per definition not imagined speech. However, as further explained in 2.4.1 those two are closely correlated and the most important aspect is reliable reproducibility of the selected paradigm. The characters were presented on a screen and the participants advised to repeatedly read it in a period of 4s after stimulus presentation, as often as possible in their normal reading speed. From those 4s, the middle 2 were extracted for classification for each character, as well as a 2s resting period after each imagination, to act as rest condition in the later classification. Words were presented randomly 15 times per run and each participant performed 5 runs, resulting in an overall 75 samples per word and participant. Wang et al. compared two different electrode setups, one covering the whole scalp according to the 10-20 system with 30 electrodes and one with a reduced setup of 15 electrodes focused on the left hemisphere of the brain to target speech related areas. Two of the overall 8 subjects were recorded with the full 30 electrode setup, while the remaining 6 performed the experiment with the 15 electrodes over the left hemisphere. Features were extracted with the Common Spatial Pattern (CSP) algorithm and forwarded to a Support Vector Machine for classification in a 10-fold cross validation approach. Wang et al. report average classification accuracies of 83.97% for left vs rest, 83.22% for one vs rest, and 66.87% for left vs one, taking into account the data of all 8 subjects. The work

Subject	Accuracy ± std (%)		
	左(left) vs Rest	壹(one) vs Rest	左(left) vs壹(one)
A1	82.13 ± 1.56	76.34 ± 3.24	71.06 ± 2.79
A2	80.45 ± 3.34	79.26 ± 2.45	70.23 ± 2.38
B1	73.65 ± 1.98	78.45 ± 1.89	63.76 ± 3.36
B2	88.73 ± 1.67	89.78 ± 1.72	69.27 ± 1.34
B3	79.64 ± 2.35	78.95 ± 1.89	66.15 ± 2.84
B4	93.74 ± 0.97	95.76 ± 0.79	59.96 ± 1.76
B5	85.75 ± 2.78	83.02 ± 1.83	70.13 ± 3.57
B6	87.65 ± 1.38	84.19 ± 1.65	64.37 ± 2.66

Figure 2.14: Individual classification results of Wang et al. as presented in [219]. Participants B1 - B6 were recorded with the reduced setup of 15 electrodes over the left hemisphere while A1 and A2 were measured with a 30 electrode setup over both hemispheres.

of Wang et al. is of interest in the scope of this thesis for several reasons. First of all, it introduces the term Speech Imagery BCI. Second, Wang et al. discuss the possible impact of the different semantic meanings of the two characters and effects of this meaning in brain activity, which should be investigated in the future, supporting our work on semantic category classification in Speech Imagery BCIs in chapter 4. Lastly, Wang et al. investigate two different electrode configurations targeting a homogeneous distribution over the whole cortex and one focused on the left hemisphere. They conclude that, due to the "outstanding" classification results achieved with only the left hemisphere, and

the analysis of the CSP components showing "most of the best channels are covering left brain", that they "can exactly use the left hemisphere as the area to extract features from Chinese characters imagery". However, having a look at the individual classification results in table 2.14, we might agree for the speech versus rest conditions, but for the more interesting character vs character condition in the right column, we can find better classification accuracies in the subjects with the extended electrode configuration. The limitation in this case is of course the small and imbalanced dataset with only two subjects with the extended and six with the left hemisphere setup, but we adopted this configuration and hypotheses in our work on electrode reduction in chapter 5, in order to draw a profound conclusion on the impact of electrode positions on the Speech Imagery BCI classification process.

After Wang et al., many researchers followed the binary classification approach with two words, most notably Sereshkeh et al. in 2017 [191, 192] and Nguyen et al. in 2018 [156]. Sereshkeh et al. tried to classify the imagined words "yes" and "no" once in a multi-session offline analysis [191] and once in a single session online classification experiment [192].

The collected datasets and the recording followed the same procedure in both studies. Simple yes-no-questions were presented to 12 participants on a screen, which they had to answer by silently repeating the correct answer, either yes or no, repeatedly for 10 seconds while a 64 channel EEG was recorded. For the multi-session experiment [191], the same participants were invited on two separate days and performed 90 trails per day, 30 times "yes", "no", and rest condition in which participants were advised to engage in any mental activity except imagined speech. This procedure resulted in overall 180 trails including 60 trails per class. The data was bandpass filtered from 0.5 to 50 Hz, excluding power line noise at 60 Hz, and Independent Component Analysis was applied to remove artifacts as eye movement or blinks. Features were extracted with Discrete Wavelet Transform (DWT) providing information from time and frequency domain, and classified with a Multilayer Perceptron (MLP). This setups was used to perform 20 runs of 10-fold cross validation. In order to evaluate the effect of data from multiple sessions on the classification process, Sereshkeh et al. classified under three different conditions, once for each of the sessions separately and in the third condition with the data of both sessions combined. The results showed, that the second session delivered on average better results as compared to the first one, indicating learning effects of the participants with increasing number of sessions. Regarding the combination of data from both sessions, the overall accuracy reduced compared to the individual sessions in most cases, however, in 10 out of 12 participants the "yes" vs "no" condition achieved results significantly above chance level and for the "yes" vs "no" vs rest even all participants exceeded chance level significantly. The reported avarage classification accuracies for the "yes" vs "no" condition were 67.34% 68.57% and 63.16% for first, second and both sessions combined respectively, and 57.23% 60.46% and 54.07% for the "yes" vs "no" vs rest condition, again for first, second and both sessions combined respectively. This aspect shows, that multi session recordings and use of imagined speech BCIs should in general be possible.

In [192], Sereshkeh et al. applied their approach to an online classification scenario in which they tested two conditions, classification of "no" vs rest and the classification of "yes" vs "no". Again 12 subjects participated, this time in 4 sessions, two training and two online classification sessions, one for each of the conditions. They recorded once more 64 channels of EEG and applied DWT for feature extraction. As classifier an SVM implementation was used, which achieved a 75.90% classification accuracy for the "no" vs rest and a 69.30% for the "yes" vs "no" condition, notably both in the online classification scenario.

These two works of Sereshkeh et al. show the impressive possibilities of imagined speech BCIs even in multi-session and online classification scenarios. However, the limited number of only two words and the paradigm of continuously repeating words for several seconds instead of a single repetition, contradict the idea of imagined speech as an intuitive and natural interaction paradigm.
Nguyen et al. [156] tried to evaluate the possibilities of classifying short versus long words and had 15 subjects repeating short words ("in", "out", "up"), long words ("cooperate", "independent"), and vowels ("a", "i", "u"). EEG was recorded from 64 electrodes while participants were seated in front of a screen for target presentation (see figure 2.15). Subjects were trained to repeat the words 3 times in a certain rhythm, indicated by a beep during a training phase, and without audio guidance during the experiment. Each of those 3 repetitions was regarded as one trail during the later analysis. Nguyen et al. classified under different conditions, the vowels and the short words separately, each in a 3-class classification problem, the long words separately and one long vs one short word in binary classification. Riemannian manifold features were extracted and forwarded to a SVM classifier, resulting in average classification accuracies of 49.00% and 50.10% for the vowel and short word classification, both notably 3-class classification problems with a theoretic chance level of 33.33%. The binary classification of long vs long word and short vs long word yielded 66.20% and 80% respectively. All of the results ended up being above chance level with the highest values achieved for the short vs long word condition. This indicates, that choosing words with different characteristics might have beneficial impact on classification, which supports our work in chapter 4 were we selected words from different semantic categories.

Figure 2.15: Study setup of Nguyen et al. as presented in [156]. The participant wears a stationary 64 channel headset and is seated in front of a screen for target presentation.

The work presents outstanding results, on the other hand it perfectly illustrates the cumbersome training situations of long-lasting recordings with high-resolution stationary EEG setups and word presentation on a screen, as shown in figure 2.15, which was most likely similar used in [191] and [192]. Furthermore, the applied paradigm requires the participants to repeat the word 3 times for classification, equally inconvenient in

natural interaction scenarios as the concept presented in [191]. Lastly, the number of distinguishable words was rather limited with mostly two words at a time in the three presented studies [156, 191, 192].
Researchers also made efforts to exceed such binary classifications as for example González-Castañeda et al. in 2017 [69], where they tried to classify the 5 Spanish words “arriba" (up), “abajo" (down), “izquierda" (left), “derecha" (right), and “seleccionar" (select). However, González-Castañeda et al. had participants perform imagined speech in blocks of 33 repetitions for each word. As mentioned earlier, Probadnigk et al. [169] showed that this type of stimulus presentation most likely leads to temporal artifact in the EEG, positively effecting the later classification. Therefore, we would handle the reported outstanding classification accuracies with care and so all follow-up work using this dataset, as for example Torres-García et al. and their channel selection approach presented in[207], as further explained in chapter 5.
In light of these considerations, the next noteworthy work on multiclass imagined speech classification was presented by Qureshi et al. in 2018 [173]. In their study setup, 8 participants silently repeated the words “go,” “back,” “left,” “right,” and “stop.” one hundred times each. Words were presented in random order, and the EEG was recorded with a 64 channel headset while participants received the target words presented auditory via earphones. Signal processing involved an Independent Component Analysis to filter the signal from artifacts, followed by a covariance based connectivity and a maximum cross-correlation feature extraction. Classification was done by a sigmoid activation function-based linear extreme learning machine (ELM). This classifier is built on a feed-forward neural network-based architecture and has an extremely fast learning speed due to random initialization of input weights. Furthermore, a manual channel selection was applied by selecting four different sets of channels focusing on speech related areas of the brain, namely left inferior frontal lobe (Broca’s Area), left superior temporal lobe (Wernicke’s Area), a combination of the above two areas, and the whole set of electrodes. With this setup Qureshi et al. achieved a best individual classification accuracy of 40.30% and a best average classification accuracy of 32.90% in both cases for the configuration including all electrodes. Those classification accuracies, although not overwhelming, appear plausible and were recorded under appropriate conditions to allow for a conclusion on the performance of actual imagined speech classification.
Cooney et al. tried to extend the number of classes even further by classifying six imagined words from EEG activity [42]. However, the approach based on a Convolutional Neural Network only achieved 24.59% on average. Although being above chance, this can be considered rather low and is most likely due to the limited number of EEG channels. The EEG was recorded with only 6 electrodes at positions F3, F4, C3, C4, P3 and P4, covering speech relevant areas, but far from the usually applied 64 channel setups.
An even more extensive dataset, including EEG data of 12 imagined words, was recorded in 2019 by Lee et al. [122]. Seven healthy subjects performed imagined speech and visual imagery of the twelve Korean words for "ambulance", "clock", "hello", "help me", "light", "pain", "stop", "thank you", "toilet", "TV", "water", and "yes". The words were supposed to provide an essential vocabulary for various use cases, primary targeting patients’ communication. Imagined speech and visual imagery were performed in two different sessions. Both paradigms were recorded in short blocks of 4 repetitions, with 88 trails per class, resulting in 1144 trails overall for the 12 words and a rest condition. Stimuli were randomly presented to the participants via auditory cues for imagined speech and visual cues in the visual imagery condition. The experimental procedure is illustrated in figure 2.16. EEG was recorded with a 64 channel headset and electrodes arranged according to the international 10-20 system. The recorded data was bandpass filtered between 0.5 to 40 Hz with a 5th order Butterworth filter and data epoched in 2 seconds

starting from the beginning of each trail. CSP was applied for feature extraction in a One-versus-rest strategy [226] and a shrinkage regularized linear discriminant analysis (RLDA) as well as a Random Forrest (RF) used for classification in combination with a 10-fold cross validation. For this 13-class classification problem Lee et al. achieved an impressive 20.40% average classification accuracy and best of 34.20%, both with the RF classifier. Notably, all participants exceeded the chance level of 7.7% as well as the confidence bond of 9.2%.

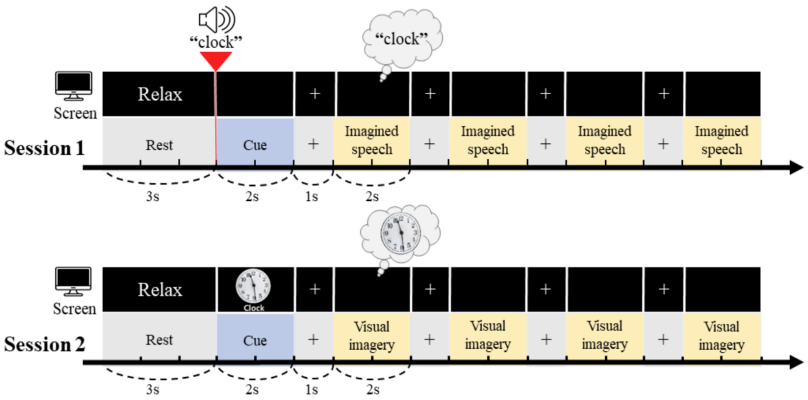

Figure 2.16: Experimental procedure of Lee et al. as presented in [122]. Session 1 in the top row shows the imagined speech condition, while the bottom row in session 2 shows the procedure for the visual imagery.

Although the results are impressive and classification accuracies are significantly above chance level, we can see a drop in classification accuracy in comparison to other works using less words [156, 173]. This issue was further addressed in the follow-up work of Lee et al. in 2020 [125]. This time, 22 subjects were invited to the experiment following the same procedure as in [122]. Lee et al. had a closer look on the relevant cortical regions, the influence of the word properties possibly affecting the decoding performance and the multiclass scalability for both paradigms. Once more, EEG was recorded with a 64 channel EEG headset and each participant performed 88 trails of imagined speech per word in short blocks of 4. This time, data was bandpass filtered from 0.5 to 125 Hz, extending the upper bound in comparison to their previous work. Line noise was removed at 60 Hz as well as the harmonic at 120 Hz and a re-referencing of all electrodes to the common average was applied, followed by artifact filtering with the EEGLAB toolbox. Lee et al. investigated different filter bands and electrode positions in order to conclude on the most valuable components in the EEG setup. All different setups were forwarded to a CSP feature extraction algorithm followed by a SVM classifier. Lee et al. report a superior performance in the high frequency bands especially the gamma regions. These findings are in line with the work of Jahangiri et al. [93] who reported on the contribution of the high-gamma band to imagined speech in EEG data. Furthermore, Lee et al. investigated the effect of meaning of words on decoding performance by discriminating them against each other in binary classification tasks. They conclude that the "meaning" of the words might be the key point to a better decoding of different classes, which supports the idea of our works in chapter 4 on semantic category classification. The evaluation on the effects of expanding the number of distinguishable words showed a clear picture as illustrated in figure 2.17. The graphs show the continuous decrease of the

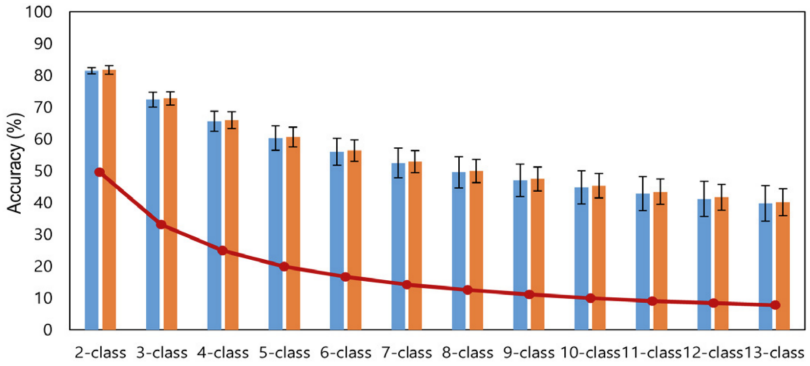

Figure 2.17: Decreasing classification performance with increasing number of classes, as presented by Lee et al. [125]. Blue bars represent imagined speech data, orange bars the visual imagery condition. The red line indicates the chance level.

classification performance with the constantly increasing number of classes. However, Lee et al. compare those rather moderate decrease in performance with the precipitous decrease in emotion classification, object perception and Motor Imagery BCIs, and claim imagined speech to be a strong paradigm for class expansion.
In their most recent work of 2022 [126], Lee et al. provided a pseudo-online implementation of their former study setup with the 12 words. Once more, EEG data was recorded with a 64 channel EEG headset from 9 subjects but this time 100 trails were recorded per word, using 80 trails for training and 20 trails for the pseudo-online test condition. In order to simulate this online condition, classification was performed on 1000 ms time windows which were moved over the recorded EEG signal with a step size of 100 ms, after the experiment. Each of the one second epochs was classified based on a CSP feature extraction followed by an SVM classifier. Within each 2 second time window of imagined speech, the most frequent output of the classifier was selected as final choice of the system. With this pseudo-online classification setup, Lee et al. achieved a remarkable average classification accuracy of 46.54%. However, due to missing details, as for example on the actual study procedure or the CSP implementation, those results must be taken with a grain of salt. Especially the fact, that their previous study was conducted with the same experimental setup, raises the question, what might have lead to the significant performance increase in comparison to their results of 2019 [122]. Although not explicitly mentioned, we suspect the focus on high gamma to be responsible for the performance increase. Lee et al. applied a bandpass from 0.5 to 40 Hz in 2019 [122] while in this work, a bandpass focusing on high-gamma from 30 to 120 Hz was applied, based on the findings in [125]. This highlights once more the importance of this frequency range in imagined speech classification.
We consider the works of Lee et al. [122, 125, 126] as current state-of-the-art in EEG-based Speech Imagery BCIs and the most relevant ones in the scope of this thesis. On the other hand, this state-of-the-art clearly illustrates the problems of the current implementations and challenges of Speech Imagery BCIs. Those challenges together with our targeted contributions and lessons learned from this overview of related work, as well as the impact on our study setups, will be presented in the following.

Summary

Although partially targeting real-world scenarios, all of the previously mentioned studies, and to the best of our knowledge the vast majority of work on EEG-based imagined speech BCIs, is performed in laboratory settings. People are seated in front of a screen and do bluntly repeat words presented to them for several hours (see figure 2.15). These training procedures are not only mentally but also physically exhausting, as participants are advised to avoid any kind of movement during this sessions, to reduce muscle artifacts in the EEG signal. This aspect might not only affect the mood of the participants, but also EEG signals due to frustration or fatigue effects. The currently applied cumbersome training scenarios of Speech Imagery BCIs are one of the biggest roadblocks of the technology being used in a real-world setting, and we will address this issue in chapter 3 by introducing two new methods for Speech Imagery BCI training. One will be based on training during speaking, while the other will involve a natural reading task, making the training procedure not only more engaging but also better controllable and more productive.
Second, although the vast majority of works reports classification accuracies significantly above chance level, the reported numbers are far from a precision expected in real-world use. Especially when increasing the number of distinguishable words, classification performance drops significantly, as reported by [125]. With an imagined speech BCI based on 12 words [122], we slowly start approaching real-world applicability for various fields of application, providing a clear advantage over established concepts as for example Motor Imagery BCIs, which fail to classify more than 4-5 commands simultaneously. However, if those BCIs only manage to provide classification accuracies of roughly 25 percent on average, they will not be applicable in a real-world scenario either. We address this issue in chapter 4 by presenting and evaluating a new concept for imagined speech BCIs including semantic classification of the word, in order to increase overall classification performance.
Beside the mentally and physically exhausting training procedures, the bulky setups as presented in figure 2.15 are a common problem. Most of the works on imagined speech classification record EEG data from as many electrodes as possible, with 64+ electrodes [191, 192, 156, 173, 122, 125, 126], due to the novelty of the concept and the widely spread activity during speech production (see section 2.3). These setups with large numbers of electrodes are accompanied with long setup times and the prices for such devices increase with the number of electrodes. Therefore, the question arises if those bulky setups with 64+ electrodes are required for imagined speech classification or if a smaller subset of electrodes could deliver satisfiable performance in Speech Imagery BCIs. Existing approaches usually only consider certain subsets of electrodes targeting areas responsible for speech production [173, 219, 1, 125] and compare them to the full setup of electrodes. However, this comparison leaves open questions about the performance of setups in between those rather focused and completely distributed configurations. Usually, one would expect that certain channels contribute more to the classification process than others, and some might not even contribute at all, meaning that with the full configuration as comparison we will most likely always include channels which reduce classification performance. Therefore, we target a systematic approach concerning the reduction of electrodes on imagined speech data in chapter 5, in order to conclude on a certain reduced number or subset of electrodes, which can be used in the future to overcome the currently bulky Speech Imagery BCI setups, unsuitable for real-world classification.
Regarding feature extraction and classification algorithms in related work, we could see a clear tendency towards machine learning methods rather than deep learning approaches

and neural networks, most likely due to the comparably small datasets. Although recent work reported an improved performance with neural networks [41] the vast majority is based on machine learning setups as for example Common Spatial Patterns (CSP) or Discrete Wavelet Transform (DWT) for feature extraction and Random Forrest (RF) or Support Vector Machines (SVM) for classification [63, 219, 192, 156, 122, 125, 3]. As our work is concerned with developing and evaluating new concepts and methods for Speech Imagery BCIs, we will mostly rely on those established machine learning algorithms in our signal analysis, and leave adaptions to neural networks as future work, if our concepts prove feasibility with those established methods.
Concerning our study setup procedures, we have seen a variety of different methods in related work, starting with the number of words ranging from binary classification [219, 191] to a dataset including 13 words [122]. We will consider a number of 5 to 9 words as suitable for our experiments. The lower bound of 5 was chosen in order to stay above the maximum of other BCI paradigms, e.g. Motor Imagery. The upper bound was set for similar reasons as the choice for established machine learning methods, namely, that our work was supposed to develop novel concepts for imagined speech training and classification. Maxing out the classifier with a larger number of words would most likely minimize the quested performance differences between standard and newly developed methods, which would contradict the actual aim of this thesis. Regarding stimulus presentation, Sereshkeh et al. [191, 192] and Nguyen et al. [156] used periods of repetitions in which participants were advised to repeat a word continuously in a certain period of time. These continuous repetitions of words do in our opinion contradict the purpose of imagined speech as a natural BCI paradigm. We consider a single repetition of a word to trigger interaction as more natural and will therefore design our study setups according to single word repetitions and short-blocks of repetitions as presented in [122]. The advantage of short-blocks is, that it enables to record more data in a shorter amount of time and was frequently used in imagined speech research [122, 125, 126].

Within this section we have identified some of the most pressing issues in the field of EEG-based Speech-Imagery BCIs, namely cumbersome training procedures, insufficient classification accuracies and bulky high-resolution setups. In the following, we will present related work concerning those three issues in accordance with our plans to solve them, starting with the training procedures.

2.6 Speech Imagery BCI training

Training procedures for Speech Imagery BCIs and BCIs in general are know to be tedious and cumbersome. In order to train a classifier to detect certain patterns in brain activity, participants have to repeat a thought causing a certain brain pattern various times, to collect sufficient training data for the classifier to learn [179, 180]. Figure 2.18 illustrates a typical training scenario for imagined speech recordings as presented in [168] in which a participant is seated in front of a screen and is presented the word to repeat silently.
Those training sessions usually involve several hours in which the participant is advised to remain calmly seated, often fixated with a chin rest, try to avoid blinking and focus on the task, which makes those procedures mentally and also physically exhausting. Mentally exhaustion however can have a negative effect on the recorded EEG data and the later classification process, which leaves the researcher with the difficult trade-off between sufficient number of repetitions and a minimal overall duration during the design of such experiments.

Figure 2.18: Illustration of a standard study setup for EEG-based imagined speech, taken from [168].

Beside being mentally exhausting, the described setup is also not really engaging for the subject and the initial euphoria of being part of a neuroscientific study measuring brain activity usually gives way after an hour to the disillusionment, that this kind of study is very repetitive and less exciting than expected. Loosing the engagement of the participant to fulfill the given tasks properly is of course problematic as well, due to another drawback of imagined speech training scenarios, the lack of verification on the produced output of the participant.
Given that those procedures are long lasting and boring, the participant might loose focus and by accident repeat a different word than shown on the screen, or mix up the starting point of the imagination phase, or even completely stop silently repeating the words, without any chance for the experimenter to notice those mistakes.
All the aforementioned problems lead us to the conclusion that training scenarios for Speech Imagery BCIs need to be reworked from scratch. We therefore propose in this thesis two novel concepts for training Speech Imagery BCIs trying to disguise the training procedure in another task which is less cumbersome and boring, namely, overtly speaking words and silently reading a text. The collected data and trained classifiers shall later be applied to imagined speech data, resulting in a kind of transfer learning concept. Transfer learning in Speech Imagery BCIs is however not entirely new.

2.6.1 Transfer Learning in Speech Imagery BCIs

Transfer learning is motivated by the fact that human beings are capable of intelligently applying previously learned knowledge to facilitate or improve solving new problems. Knowing how to play the electronic organ for example might facilitate to learn the piano [166]. In contrast to standard machine learning approaches we do not necessarily need to stay within the same task, domain or distribution in our test and train data, but try to transfer the knowledge of one domain to another.
Garcia et al. [64] addressed the issue of the tremendous amount of training data needed in SI-BCIs. Usually it requires the user to train the algorithm for each and every word which shall later be used in the interaction. Garcia et al. tried to transfer a feature extraction method from a set of known words to a word unknown to the algorithm in order to investigate the feasibility of saving training time by transferring feature extraction methods between words. They recorded EEG data of 27 subjects while they were silently repeating the five words "Up", "Down", "Left", "Right" and "Select" in Spanish. Each word was repeated 33 times, resulting in a dataset of 165 silent repetitions per participant.

Afterwards they generated a codebook out of this dataset by using K-Means clustering. With this codebook, they generated histograms for each word, similar to a bag-of-words approach. In the first part, they excluded one of the words from the generation of the codebook and just extracted the histograms for this word with the book generated on the remaining 4 words. The resulting data was classified with a Naive Bayes Classifier with a 75/25 train-test-split and a 10-fold cross validation. The classification results were compared to the results on a dataset where the codebook was generated with all five words to investigate differences in classification accuracy and the transferability of the calculated features. The resulting classification accuracy on the dataset including all five words turned out to be 68.93 ± 12.43% and was used as benchmark for several trails of leaving out different words from the feature extraction process. Leaving out the word "Up" resulted in an accuracy of 65.27 ± 12.6% and for the word "Down" in 65.49 ± 12.77%. Garcia et al. concluded, that there is no statistically significant difference between the transfer learning approach and the approach including all 5 words, meaning that the transfer was successful and could be used to reduce training times. Promising results, however, the dataset was comparably small, the algorithms will have to prove feasibility on larger datasets and compared to our targeted approach, the transfer took place on the feature extraction level while we want to transfer between EEG activity of different cognitive processes.
Another transfer learning approach in Speech Imagery BCIs was presented in [41]. Cooney et al. tried to transfer a classifier trained on the data of a group of participants to another group of participants unknown to the algorithm. This between- or cross-subject approach was evaluated on the data collected from 15 subjects while silently repeating the vowels "A", "E", "I", "O" and "U". A Convolutional Neural Network was trained on the power band features form the EEG of 14 subjects and the trained classifier applied on the data of the left out 15th subject. The classification accuracy for the data of all 15 subjects was used as benchmark and turned out to be 35.68%. The transfer learning approach achieved astonishing 32.75% not too far from the benchmark. Considering the chance level of 20% the results are however not overwhelming and once more, the transfer did not take place between different paradigms of speech production but on the same paradigm, imagined speech.

Summing up the work on transfer learning and imagined speech, we can say that the method we target in our work of transferring a classifier from one paradigm to another, has to the best of our knowledge not been done before. Existing work on transfer learning focuses on the transfer of a classifier between data of different subjects [41] and the transfer of feature extraction methods between subjects [64], which is valuable for the reduction of training data needed, however far from our approach of making the training procedure more comfortable and interesting for the user.
Therefore, we developed two new methods for training Speech Imagery BCIs, based on two concepts similar to imagined speech namely, silently reading and overtly speaking.

2.6.2 Neural Correlates of Reading

The process of reading a text can usually be performed in two ways, either reading the words out loud or silently. Reading silently is by definition very similar to producing imagined speech (see also section 2.4.1), not only in terms of the conceptual process but also in terms of the underlying neural correlates.

Various researchers have tried to unravel the neural processes during reading and the effect of different types of stimulus presentation in the past [170, 98], which showed a strong indication of the involvement of the left cerebral hemisphere.
Joubert et al. [99] measured the brain activity via fMRI of ten participants while they were silently reading certain words. Three conditions were investigated, silently reading of very high frequency regular words (lexical task), nonwords (sublexical task) and very low frequency regular words (sublexical task). With this study setup Joubert et al. wanted to compare the brain regions responsible for lexical and sublexical processes in reading. The words were presented in French and selected based on their appearance in the French language, for the high frequency words with a mean frequency of 45.600 out of a corpus of 26.500.000 words, e.g. homme, coeur, and for the low frequency words with a mean frequency of 35.42 out of a corpus of 23.500.000 words, e.g. sonar or tango. The control condition consisted of nonwords which were chosen to be pronounceable and apply to the rules of the French language, i.e., plaud, fosme, but not to resemble known words. Participants were advised to silently read the words inside the fMRI scanner without moving any facial muscles and their tongue. In the offline data analysis Joubert et al. found that the lexical condition primarily activated a border region of the left angular and supramarginal gyri and the sublexical condition the left inferior prefrontal gyrus. Both of those areas lie within the left hemisphere which is also know to be dominantly active during overt and imagined speech (see section 2.3). Furthermore, Joubert et al. conclude, that lexical and sublexical processes in reading activate different regions within a complex network of brain structures, which can be beneficial when switching from fMRI to the EEG with a poor spatial resolution, if corresponding brain patterns are widely distributed over the brain.
Deniz et al. [52] conducted a fMRI study in which they wanted to compare brain activity related to semantic processes during reading of and listening to short stories. Participants had to read and listen to a collection of 10 short stories each 10 - 15 minutes long which were written to cover a wide range of topics and engage the reader/listener. The data analysis was conducted similar to [88]. Maps of brain activity were created and compared between the two conditions. Deniz et al. conclude that the semantic representation in brain activity evoked by listening versus reading are mostly identical and that the representation of language semantics is independent of the sensory modality. Although mainly concerned with semantic processing, this finding strengthens our hypothesis, that silently reading and speaking a word should produce similar patterns in brain activity especially considering our findings on semantics in Speech Imagery BCIs (see chapter 4).
In summary, literature shows, that the two concepts of imagined speech and natural reading mostly activate the same brain regions involving the language center primary dominant in the left hemisphere of the brain. Thus given the concept of production and the neural correlates, one might say that silently reading and imagined speech are basically identical, however, there are slight but significant differences which need to be considered during the implementation of our targeted approach.
First, this work is concerned with Speech Imagery BCIs on a word level, meaning that the repetition only involves single words. In theory also full sentences could be used, nevertheless, as our work addresses the word level solution, our training process will foresee embedding words in a text. This means the text could give the words a different context as expected by the reader, which might give them a different context-dependent meaning, influencing the brain patterns and impeding the later classification process. This aspect needs to be considered in the design of the study setup and analysis of the results.

On the signal processing side, eye artifacts will influence the EEG data during the recording and need to be filtered out in the preprocessing. However, filtering techniques for eye artifacts in EEG data exist and have been applied in several studies involving EEG recording during reading tasks. Even a labeling of EEG data on a word-level based on eye-tracking data has been realized recently in two datasets ZuCo and Zuco 2.0.

2.6.3 EEG and Eye-tracking in Natural Reading

In 2018 Hollenstein et al. recorded the Zurich Cognitive Language Processing Corpus (ZuCo) [85]. In this study 12 healthy adult native English speakers had to read text for 4.6 hours, while EEG and eye-tracking data were recorded simultaneously. The study involved 2 natural reading tasks, where the sentences were once taken from the Stanford Sentiment Treebank and for another session from the Wikipedia relaxation extraction corpus. The data of the Stanford Sentiment Treebank included positive, negative and neutral sentences from movie reviews and participants had to rate the quality of the movie based on the sentence on a scale from 1 (very bad) to 5 (very good). The Wikipedia texts contained specific relations and in this task participants had to answer simple control questions about the presented sentences. In an additional session the participants had to perform task specific reading and were presented different sentences from the Wikipedia corpus, but this time, they had to mark whether a specific relation occurs in the given sentence or not. The main purpose of this work was not to classify imagined speech or words read, but rather to provide a corpus for natural language processing in general with a focus on EEG and eye-tracking during reading, to advance research into the cognitive processes of reading and language understanding. Interestingly for our approach, they used the eye-tracking data for labeling the EEG data on a word level in a natural reading scenario. The participants were shown the complete sentences on a screen and advised to read them at their own speed. Long sentences spanned multiple lines, the lines were triple-spaced, and the words double-spaced resulting in a clearly separated and well readable design as seen in figure 2.19.

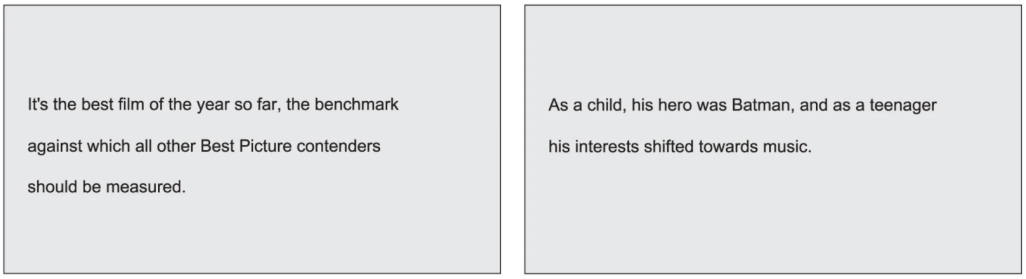

Figure 2.19: Two examples of sentences as presented to the participants in the ZuCo study [85]. Left, a sentence from the Stanford Sentiment Treebank and right a sentence from the Wikipedia relaxation extraction corpus.

The analysis of Hollenstein et al. was concerned with fixation-related potentials in the EEG and not with imagined speech, however, our text design and eye-tracking analysis were inspired by their study as further explained in section 3.1.1 to ensure an appropriate labeling of words during reading for our dataset.
In 2019 Hollenstein et al. presented ZuCo 2.0 [86] a new dataset containing an additional task of annotation which should allow to analyze differences in cognitive processing

between natural reading and annotation. Once more the experiment was not concerned with imagined speech. A closer look on the words inside the sentences of the recorded data did not provide sufficient repetitions of nouns to allow training a classifier on one of them, which could have yielded insights on the brain activity during reading.

In summary, none of the before mentioned studies and, to the best of our knowledge, no other existing study is related to training SI-BCIs on EEG data recorded during reading. However, work on simultaneous recording of eye-tracking and EEG as well as labeling EEG data on a word-level based on eye-tracking data exists [85, 86]. Furthermore, silently reading and speaking seem to involve the same brain areas as shown in [93, 52].
Beside all the remaining challenges and limitations mentioned at the beginning of this section, existing studies give evidence that the transfer from reading to silently speaking can be realized and we foresee it as a promising approach for the training of Speech Imagery BCIs.
Similar promising is the targeted transfer from overt to covert speech in our second Speech Imagery BCI training approach.

2.6.4 Neural Correlates of Overt and Covert Speech

Another paradigm similar to silently speaking words and therefore suited for a transfer approach might actually also be considered the exact opposite of it, overtly spoken speech.
Speech production is considered to be the most complex motor skill covering almost one third of the primary motor cortex with speech related muscles [188]. It is well known that silently speaking triggers partially the same brain regions compared to overtly speaking especially in the motor centers [160, 30] and therefore has similar characteristics in the produced EEG signal, however with a lower amplitude [22]. Studies found a dominantly left lateralized production of overt and covert speech mainly involving the motor, premotor and inferior frontal cortex [165, 108]. A detailed overview and history of speech production models can be found in section 2.3 Speech and the brain. In the following we will focus on related work which has tried to classify overt and covert speech from brain activity and the different methods used with a focus on feature extraction.
Martin et al. [141] investigated similarities in different speech processing procedures by analyzing ECoG recordings of five participants while they were listening, overtly speaking and silently speaking the five words “Spoon”, “Cowboy”, “Battlefield”, “Swimming” and “Telephone”. High gamma features (70-150 Hz) were extracted from the data and a Support Vector Machine trained on those features to do pair-wise classification of the words. The reported average classification accuracies for this binary classification problem with a chance level of 50% were 89.4% for listening, 86.2% for overt speech, and 57.7% for imagined speech. The results might not appear overwhelming with a classification accuracy of 57.7% for the Speech Imagery condition, but they are above chance level and the authors conclude that high gamma frequency should be a valuable feature and involved in all the three different speech processing concepts. In contrast to our studies however, the electrodes used in this work were placed in grids directly on certain parts of the cortex of the left hemisphere of patients who were undergoing neurosurgical procedures due to epilepsy, resulting in a better signal-to-noise-ratio and a higher spatial resolution as compared to the EEG (see section 2.1.1). Nevertheless, as mentioned in section 2.7.2 and applied in our own work in section 4.2 we claim that imagined speech detection based on EEG can benefit from a broader distribution of electrodes covering the whole cortex, which would allow to apply the methods used in

this work also with EEG recordings. A bigger challenge might be the fact, that the EEG is known to be prone to artifacts in this frequency range which will make the analysis complicated but possible.
Lee et al. [124] tried to avoid this problem by investigating spatial and temporal features rather than frequency features for EEG recordings in imagined and overt speech. They collected the data from seven participants, for 12 words “Ambulance”, “Clock”, “Hello”, “Help Me”, “Light”, “Pain”, “Stop”, “Thank You”, “Toilet”, “TV”, “Water”, “Yes” and a resting phase in one session for overt speech and a separate session for imagined speech. Each word was repeated in blocks of four repetitions in a row and 22 blocks for the whole experiment resulting in 88 repetitions per word and participant overall. The EEG data was recorded with a 64-channel EEG headset at a sampling frequency of 1000Hz later downsampled to 250Hz and electrode positions chosen according to the 10-20 system. Surprisingly, the only preprocessing methods applied were a fifth order Butterworth filter in the frequency range of 0.5–40 Hz and a baseline correction by subtracting the average value of 200 ms before the trial onset. No further efforts were made concerning filtering muscle artifacts as for example Independent Component Analysis. Lee et al. used Common Spatial Patterns (CSP) as feature extraction method and a Shrinkage Regularized Linear Discriminant Analysis classifier, resulting in an average classification accuracy of 16.2% for imagined speech and 59.9% for overt speech which is in both cases significantly above the chance level of 7.7%.
Based on the result of the CSP analysis they conclude that equivalent classes of the two speech paradigms share certain parts of the common characteristics (see figure 2.20) and that there is potential in classifying the one based on the other, however they did not perform the transfer. For our work we therefore consider CSP as a promising feature extraction method to achieve our goal of training an imagined speech classifier on overt speech data.
Another feature extraction method targeting the similarities between overt and imagined speech was presented in [146] where wavelet decomposition was used in order to create an Artificial Nerual Network (ANN) Based Multimodal Automatic Speech Recognition System (ASR) with imagined and vocalized speech. Mini et al. used an existing dataset “The KARA ONE Database: Phonological Categories in imagined and articulated speech” from Toronto University [235] consisting of seven phonemic prompts and four words "Pat", "Pot", "Knew", and "Gnaw". Those words and prompts were overtly and silently spoken by 13 participants while a 10-channel EEG was recorded. Wavelet enhanced Independent Component Analysis was applied to remove artifacts in combination with a FIR bandpass filter set to a frequency range of 0.5 - 60 Hz. Wavelet Packet Decomposition (WPD) with a five level decomposition was used for feature extraction along with several statistical features namely average energy, mean, standard deviation and zero-crossing rate. This setup produced promising classification accuracies of 42.03% for the overt speech and 56.29% for the imagined speech condition. These results showed once more that overt and imagined speech can be classified with the same feature extraction and classification methods, supporting our claim that overt speech can be used as input for training an imagined speech classifier.
The similarity in neural activity patterns of overt and imagined speech, measurable even in EEG activity, has been shown in several studies [124, 221, 146], however, a real transfer of a classifier from one to the other has, to the best of our knowledge, not been done before.

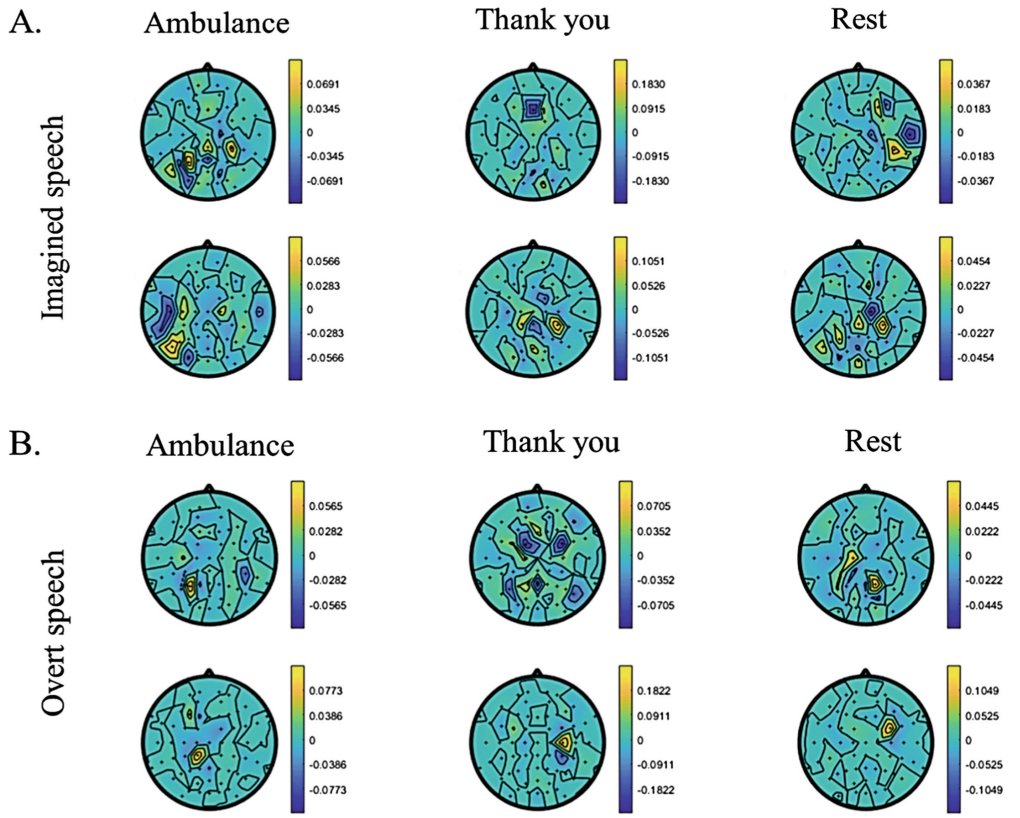

Figure 2.20: First and last CSP patterns for subject 3 discriminating the classes "ambulance", "thank you" and rest as presented in [124]. The picture shows the similar activation patterns for overt (B) and imagined (A) speech.

Summary

Current Speech Imagery BCI training sessions are usually long lasting, mentally and physically exhausting procedures, requiring the participants to bluntly repeat words in front of a screen. In our attempt to make Speech Imagery BCI training procedures more engaging, we foresee high potential in the transfer of classifiers between different paradigms. Work on transfer learning approaches for imagined speech exists, however, only on the transfer of a classifier between data of different subjects [41] and the transfer of feature extraction methods between subjects [64]. To the best of our knowledge, there is no work on transferring a classifier from one paradigm to another, as we plan to do for the transfer from silent reading and overtly speaking to imagined speech. Related work on the neural correlations between silently reading and imagined speech showed, that silently reading and speaking involve similar brain areas as shown in [93, 52]. However, none of the presented studies and, to the best of our knowledge, no other existing study is related to training SI-BCIs on EEG data recorded during reading or speaking. Beside all the remaining challenges and limitations mentioned in section 2.6.2 and 2.6.4, existing studies give evidence that the transfer from silently reading and overtly speaking to imagined speech can be realized and we foresee it as a promising approach for the training of Speech Imagery BCIs.

2.7 Speech Imagery BCIs and semantics

EEG-based Speech Imagery BCIs have proven to provide classification accuracies significantly above chance level, however, the state-of-the art manages to distinguish not more than 4-5 words [69, 1]. A possible solution to this problem is a 2-stage or -level approach which starts with classifying the semantic category of a word before proceeding to classifying the word itself.
In the following, we will give an overview on related work in the field of research on semantic processes in the brain and the possibility to classify those processes in Speech Imagery BCIs.

2.7.1 Semantic Processing in the Brain

Our brain acquires knowledge from experience and associates words with a certain meaning based on this experience. The question how this semantic representation of words is encoded in neural activity is a key challenge in cognitive neuroscience [197] and has been widely addressed over the past decades by researchers analyzing patterns of brain damage in patients with selective impairment for a specific semantic category [220, 79, 54]. Overall, those studies suggest a broadly distributed neural representation, with particular reliance on inferotemporal and posterior inferior parietal regions of the brain [19]. Those investigations have been confirmed also with healthy subjects driven by the advances in brain imaging techniques, most prominently by functional Magnetic Resonance Imaging (fMRI) and Positron Emission Tomography (PET).
Binder et al. [19] provided a review on 187 studies concerned with the detection of semantic processes in brain activity measured by fMRI (126) and PET (61) and projected the resulting 1135 activation foci published in those studies onto a inflated surface of the cortex (see figure 2.21).

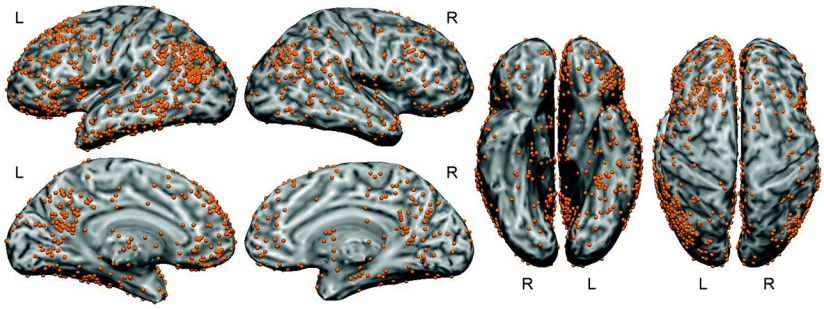

Figure 2.21: Accumulated projection of the foci during semantic processing collected in 187 studies as presented in [19].

The results showed a moderate left-hemisphere lateralization with about 68% of the foci located in the left and 32% located in the right hemisphere. Additionally, the results indicate a rather broad distribution of semantic processes over the cortex, which supports the claim of this work that it should be possible to classify semantic categories of imagined words even with the EEG and its low spatial resolution.
In 2011, Murphy et al. targeted this approach by conducting a study on decoding the semantic category of a word from EEG activity while participants silently named images of mammals and tools [153]. The EEG was recorded with 64 channels and a time-frequency

analysis performed offline on the recorded data. A Support Vector Machine was then used to classify between mammals and tools. Results showed that the category could be detected with an average accuracy of 72% trained on the individual and with an average classification accuracy of 61% for cross-subject training, both well above chance. Beside the classification accuracies Murphy et al. conclude based on their results, that semantic processing and the resulting representations are widely spread across the brain and that is is somewhat shared between the participants. This aspects would both strongly benefit the implementation of an EEG-based Semantic Speech Imagery BCI due to the low spatial resolution of the EEG.
A drawback of Murphys work as well as the vast majority of the works presented in the review of Binder et al. [19] is, that it mainly involves the distinction between only two semantic concepts or categories, living (mammals) versus non-living (tools). These two categories have been the focus of research on semantics for several years and especially in EEG studies this categorization in living and non-living remained untouched [8, 43]. Furthermore, many studies only target certain areas of the brain known to be active while speech processing [206, 197]. One reason might be that research mainly aimed at detecting certain foci in brain activity, as shown in the review of Binder [19], and the reliance on established concepts of speech production being dominantly represented in the left hemisphere.
In 2016 a study of Huth et al. [88] tried to extend those established concepts and managed to reveal semantic maps of the brain spread over the whole cortex beyond certain brain regions or a certain hemisphere. In their study short stories were presented to the participants and the words included later divided into 12 semantic concepts, distinguishable by the individuals brain activity. Furthermore, the study revealed similar maps amongst the different subjects (see figure 2.22), indicating the possibility to apply general mapping approaches in BCI applications for semantic detection, facilitating the implementation and use of these systems, due to reduced training times and increased precision of the classification process.
In a follow-up study Huth et al. showed, that the representation of semantic information across the human cortex is invariant to stimulus modality during listening versus reading [52]. In their study, they used fMRI to record brain activity while participants read or listened to the same narrative stories for several hours. The generated models for listening and reading were highly correlated in most semantically selective regions which suggests, that "the representation of language semantics is independent of the sensory modality through which the semantic information is received"[52]. This fact makes those findings highly interesting for our planned approach of detecting semantic categories during silent repetition of words in Speech Imagery BCIs. However, the studies were conducted with fMRI which has a far better spatial resolution than the EEG, but is less suited for BCI applications due to low time resolution, restrained mobility and high costs of such medical devices. However, there are already isolated approaches in literature in which this concept has been mentioned in the context of SI-BCIs.

2.7.2 Semantic Classification in SI-BCIs

Only few researchers have explored the feasibility of classifying semantic categories of words in in the field of SI-BCIs in the past.
Simanova et al. [197] conducted a study in 2010 in which they presented textual, visual and auditory cues of words from three different categories (animals, tools, clothing/vegetables) to 24 participants. EEG was measured with 64 electrodes in order to capture the

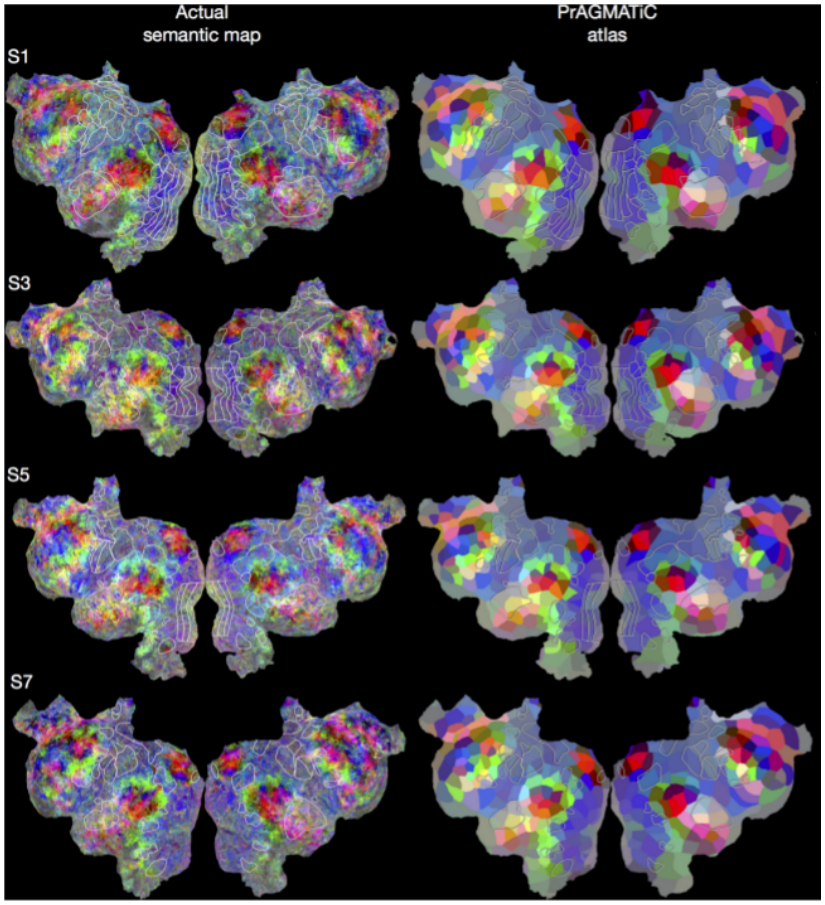

Figure 2.22: Semantic maps as presented in [88] for 4 subjects. The left column shows the actual fMRI data flattened and grouped for the different semantic categories. The right column shows the results of the PrAGMATIC clustering algorithm revealing the similarities between subjects.

reaction of the brain to the different stimuli. The experiment lasted around 80 minutes in which participants were presented the words of the different categories for 300 ms and had to respond upon appearance of the item with a button press of the right index finger according to a previously determined target category (clothing or vegetables). The data analysis was conducted offline and the event related potentials (ERPs) as response to the stimulus from presentation until 600 ms afterwards, were averaged over all trails and compared in between the different categories of the words. Simanova et al. reported a clear differential response to the semantic categories in the case of the visual stimulus. In the spoken and text-based condition categories were less clearly distinguishable showing significant differences only for two out of twenty subjects. Within their study Simanova et al. could once more show, that it is possible to distinguish semantic categories from EEG activity, however, only as a reaction to visually presented stimuli and not during imagined speech. Although their study does not provide a concrete implementation, they still give an outlook on how one could make use of this concept in the context of BCI applications for the first time. Kosmyna et al. described Simanovas idea later as

conceptual imagery and the application as conceptual Brain-Computer Interface [114]. In the field of actual speech imagery, Kim et al. [112] referred to this concept as "meaning based covert speech classification". In their paper from 2013 they conducted a study in which two right-handed healthy subjects performed overt and silent repetitions of 8 Korean monosyllabic words, where 4 of those words could be categorized into a part of the face (cheek nose, eye, mouth) and the other 4 into numbers (three, five, nine, ten). Each participant performed five runs consisting of 80 trails with random word presentation. EEG data was recorded from 32 Ag/AgCl electrodes placed according to the 10-20 system and a sampling frequency of 1000 Hz. The recorded data was preprocessed by applying an ICA in correlation with a simultaneously recorded Electrooculogram (EOG) channel to exclude eye blink artifacts. Additionally the signal was bandpass filtered from 1 to 100 Hz and downsampled to 250 samples per second. The spectogram was calculated for feature extraction in 1 Hz wide bins for 49 frequency bands ranging from 2 to 50 Hz over all 30 channels. The feature vectors were labled according to the category the words belonged to and fed to a linear SVM for classification. After feature selection with SVM-based recursive feature elimination, they achieved an astonishing average accuracy of 92.46%. However, the authors correctly qualify their results as preliminary given that the recording was done for only 2 subjects and 2 categories with a small number of word repetitions.
Surprisingly there was no follow-up work on this promising approach and the next relevant related study was published in 2018 by Kumar et al. [116]. They described their concept as "coarse-to-fine-level" approach and showed their participants pictures from three different categories namely, digits, characters, and objects. Those categories represented the coarse level while on the fine level they included 10 letters for the character category, the digits from 0-9 for the digit category and 10 objects that are commonly used or seen in daily life routine, e.g. a car or a dog. All those fine level items where presented to the participants as pictures as shown in figure 4.1 for 10 seconds followed by the actual task of envisioning the shown item for 10s with eyes closed. After the envisioning phase, participants were given 20 seconds of rest to clear the previous imaginary thought and return to a rest state. Each of the overall 30 items, 10 per category, was repeated 23 times resulting in an overall dataset size of 690 samples. Afterwards, each of the 690 samples was split into multiple parts for fine level in 50ms windows and for the coarse level in 250ms windows. For preprocessing a moving average filter was used for signal smoothing and reducing eye and muscle artifacts. As features they used statistical features, namely, Standard Deviation, Root Mean Square, Sum of Values and Engery of the smoothed signal on all 14 electrodes resulting in a 14x4 feature vector. This feature vector was fed into a Random Forest classifier and trained once to detect the coarse level, consisting of the three categories character, digit and object, and with the same data on the fine level for all the 10 items of each class. The classification on the coarse level achieved an average classification accuracy of 85.20% whereas the fine level classifier achieved 67.03%. In order to show the effectiveness of the proposed framework Kumar et al. computed the performance of the system with and without coarse-level classification. Using the following equation:

$$Total - accuracy = CLA * \frac{(DA + CA + IA)}{3} * \frac{1}{100}$$

with CLA, DA, CA, and IA being the accuracy of coarse level, digits, characters, and images, respectively, they achieved a total accuracy of 57.11 % compared to a 39.34 % without coarse level. These results seem very promising, however, this work has several shortcomings.

First of all, the authors do not clearly state the paradigm they used during the imagination process and are not consistent with the terms they mention. Throughout the whole paper there are various different occurrences of terms describing the paradigm as, e.g. silent, imagined or envisioned speech. Based on this wording the actual process of producing the cognitive task might be different (see section background 2.4.1). In the section on dataset collection they even speak of "envision the shown item for 10s" which would mean that the paradigm in this case was visual imagery of an object, digit or character rather than actual speech imagery. This assumption is strengthened by another study [205] which used the dataset recorded by Kumar et al. and trained a neural network to "generate images from thoughts of a person" which is a clear indication that visual imagery was used, rather than imagined speech.
Secondly, the hardware used for recording was an Emotiv EPOC+ which is rather a consumer device for affordable BCI applications and provides only a a limited set of 14 electrodes, generally not used to evaluate such basic neuroscientific principles as speech imagery or detection of semantic processes. On the other hand, the attempt of classifying imagined speech with such devices seems valid if targeting real-world applications in the long term. The biggest drawback of such consumer headsets usually is the reduced signal quality due to cheaper hardware, resulting in decreased classification accuracies. The reported results however show outstanding accuracies given the astonishing number of overall 30 classes which brings us to the biggest problem with this study, the stimulus presentation.
According to the section on the experimental design Kumar et al. recorded a single block of 10 seconds for each word in this study which was afterwards split into parts for further processing. This kind of block design, where all stimuli of a given class are presented in a row together, is know to lead to a classification of arbitrary brain states based on block-level temporal correlations rather than stimulus-related activity. Li et al. elaborated on that problem in [130] and explicitly mention the dataset recorded in [116] as affected by this problem. Furthermore, Porbadnig et al. [169] emphasized this problem in 2009 already, which questions the quality of the recorded data and therefore the validity of the reported results.

Figure 2.23: Items used for presentation in [116] from the categories a) characters, b) digits, c) objects. The items were presented as pictures and the participants had to imagine seeing them afterwards with eyes closed.

In conclusion, we found that there is only little existing work on semantic classification in SI-BCIs. Kim et al. [112] used a limited set of 2 categories and the results of Kumar et al. [116] are questionable due to their study setup as pointed out in [130]. Furthermore their study was rather concerned with visual imagery than speech imagery and did not really involve repetition of words of specific semantic categories. The concept itself, of classifying semantic categories of an imagined word prior to the classification of the word itself, remains promising however, and has the potential to multiply the number of classifiable words in Speech Imagery and Silent Speech BCIs in general many times over.

Summary

In summary, detecting semantic processes from brain activity has been shown to be feasible on a larger scale in fMRI and EEG studies. Most of the work based on EEG data is however concerned with distinguishing between 2 categories, living, e.g. mammals or non-living, e.g. tools [197, 153, 206]. Related work in SS-BCIs is limited to few studies where [112] used a confined set of 2 categories as well and the results of [116] are questionable due to their study setup as pointed out in [130]. We therefore see a high potential in this method to increase classification accuracies in EEG-based Speech Imagery BCIs and will present concepts and methods on a possible integration into the standard classification process in chapter 4.

2.8 Speech Imagery BCIs and electrode reduction

EEG-based Speech Imagery BCIs usually make use of high resolution EEG-headsets with 64+ channels, as presented in the previous sections. The question in how far all of those electrode contribute valuable information to the classfication process is ongoing research not only in SI-BCIs but BCIs in general. Therefore, we separate our related work presentation by starting with an overview on electrode reduction in BCI applications in general, followed by electrode reduction methods in the BCI paradigm addressed in this thesis, Speech Imagery BCIs. In the following we will give an overview on the identified related works and conclude on the methods to be used in our approach.

2.8.1 Electrode Reduction in BCIs

Electrode reduction is not just a relevant topic in SI-BCIs but BCIs in general. It can be applied in different ways and is usually classified as filter, wrapper or embedded method [207]. The later ones assess channel selection during the process of training and are specific to a given learning function [77]. Filter methods do not use learning functions but rather measure the inherent features from the data and select the relevant channels based on those features [119]. Wrapper methods aim at assessing certain channel subsets based on the accuracy obtained by the algorithm while learning. Those wrapper methods will be the focus of our work, as they have been used successfully in the past in various studies [5, 6, 228, 207], due to better performance and the ease of implementation [140, 117].
Marx et al. [142] attempted to find optimal electrode positions for a Steady-Sate Visually Evoked Potential (SSVEP) BCI. This type of BCI uses stimuli shown on a screen, flickering in a certain frequency and eliciting a brain response in the occipital region of the brain,

which can be measured as a peak in the frequency spectrum at stimulus frequency. For more information on this kind of BCIs see section 2.1.2. Marx et al. used those stimuli in their study to realize a braintyper in which 17 participants had to type the 2 words "PROJECT" and "TRAINS". The EEG data was acquired using 16 electrodes placed according to the 10/5 system [100] and a focus on the occipital region, the primary visual cortex. The number of electrodes was continuously reduced during the experiment based on the minimum energy consumption (MEC) algorithm. The MEC algorithm computes the ratio of signal to noise for each electrode and assigns a weight corresponding to this ratio to the electrode. The higher the noise the less useful information is present in the electrodes signal. Consequently, the electrode with the highest weight and therefore noise, was removed in each step until only 4 electrodes remained. Marx et al. investigated the mean solving time and information transfer rate (ITR) and found, that the solving time increased with decreasing channel number and the ITR decreased accordingly. Although some participants were not able to control the system with a reduced number of electrodes at all, Marx et al. conclude, that they could achieve a high accuracy with a reduced number of electrodes while retaining a fair spelling speed. They furthermore concluded, that the optimal electrode positions differed for each subject and that there was no common setup usable between the participants. A finding that is consistent with our results in [179] and [180], that BCI methods should be tailored to the individual.
Feng et. al [57] investigated an approach to reduce the number of electrodes during motor imagery in BCIs and considered channel selection on multiple frequency bands. They used the dataset from the third BCI competition, where 3 paradigms of motor imagery had to be classified, namely left and right hand, as well as right foot movement. On this dataset they used two special implementations of the Common Spatial Patterns (CSP) algorithm, the CSP-rank and a multifrequency band version of it (CSP-R-MF). This algorithm creates spatial filters for two classes using the dataset's eigenvectors and extracts the longest and the shortest ones. These vectors consist of filter coefficients, assigning weights for each electrode, based upon its influence for the class, implying, that a feature with a larger absolute value is more important. Consequently, electrodes corresponding to the largest remaining coefficient are added to the set of electrodes to be used in classification until the classification accuracy did no longer improve. The CSP-R-MF implementation included a further decomposition of the signal into 7 frequency bands before applying the CSP-rank. The maximum classification was achieved with 30 selected channels, reducing the number of electrodes in this case by more than 50%. The CSP-R-MF scored slightly better than the basic CSP-rank with a difference of about 7%. These results qualify the CSP algorithm and its variants as a promising approach for electrode reduction also for imagined speech BCIs.

2.8.2 Electrode Reduction in SI-BCIs

In the field of imagined speech BCIs, electrode reduction has not been investigated as extensively as in the other domains of BCI research. Most of the existing work in the field is concerned with reducing the number of electrodes to a certain subset based on assumptions on functional areas in the brain related to speech processing [173, 1]. Those studies basically compare the full set of electrode against 3 to 4 subsets of electrodes involving mainly Broca and Wernicke area. Systematic approaches for electrode reduction in SI-BCIs based on algorithms exploring the potential of different electrode subsets and their combinations are rare. Torres-Garcia et al. tried to find the minimal subset of channels required for imagined speech in their work of 2016 [207]. Their task involved the imagination of the five Spanish words "up", "down", "left", "right",

and "select". EEG data was recorded with an Emotiv EPOC headset with 14 electrodes for 27 participants. An ICA was performed on the data to remove artifacts and a common average referencing (CAR) was applied. Features were extracted using discrete wavelet transform (DWT) and four statistical values computed, namely maximum, minimum, average and standard deviation. Furthermore, the relative wavelet energy was added to the feature vector. In the next step, Torres et al. attempted to create a pareto front from all possible channel combinations by using a wrapper function with NSGA II, a genetic algorithm, as search function. As objectives they used minimal error rate and the amount of channels, where the error rate was confirmed using the results of a random forrest classifier. A Mamdami fuzzy inference system with ten rules was used to evalute the pareto optimal subsets. The system selected 6 to 8 channels and a top classification accuracy of 90% was achieved on the dataset of five classes. However, the implementation and methods used in this work appear questionable to us. First of all, the EEG headset used was an Emotiv EPOC, which is a consumer grade headset with semi-dry electrodes using saline to improve conductivity and a limited number of 14 electrodes. Furthermore, those 14 electrodes are mainly placed in frontal areas and at the outer borders of parietal, temporal and occipital regions of the brain, omitting the central region completely. Although partially including speech relevant areas, this setup appears too limited to get a clear picture on relevant brain regions and electrode positions, especially considering that the vast majority of research on SI-BCIs uses 32 channels and above for the recording of brain signals homogeneously spread over the scalp [179, 92, 93, 124, 191, 235, 110]. Thus, beside the obviously worse signal quality of a consumer grade EEG headset to a clinical one the low number of electrodes overall might be a limiting factor when targeting a systematic electrode reduction evaluation. Second, the use of a genetic algorithm as the NSGA II seems outdated as stated in recent works on evolutionary algorithms [55], which show a better performance in comparison to genetic algorithms. One of these evolutionary algorithms is the Grey Wolf Optimization (GWO) algorithm which was recently used by Gosh et al. in SI-BCIs however not for electrode reduction but for feature selection [68]. At least for the feature selection in their work they claim, that GWO outperformed genetic algorithms in optimization. Targeting a systematic approach for electrode reduction in imagined speech BCIs, GWO appears to be a promising method and a better choice as compared to genetic algorithms as NSGA-II.

Summary

Electrode reduction is a well researched topic in the field of BCIs. Several publications addressed the issue in various domains including Motor Imagery [228, 57] and SSVEP [142]. Electrode reduction on imagined speech EEG data is rare however, and mainly focuses on the specific investigation of certain subsets chosen based on functional areas of the cortex [173, 1]. Given the fact that most research targeting imagined speech detection is currently using high resolution setups with headsets including 64 channels and more, we see the necessity of reducing the number of electrodes in order to facilitate the conduction of studies on the one hand, but also paving the way for this technology to be applicable in real-world applications in the future. The lack of work on systematic methods for electrode reduction in SI-BCIs leads us in a first step to provide a comparison of different electrode reduction methods in combination with a variety of feature extraction and classification methods on a single imagined speech dataset in section 5.1. Based on the results of this comparison we want to evaluate in section 5.2, if we can find a single best minimal set of electrodes for SI-BCIs by applying the best performing method from section 5.1.

Within this chapter we have provided a broad overview on Speech Imagery BCIs with a focus on EEG-based Speech Imagery BCIs. The presented studies report impressive classification results with accuracies significantly above chance level. However, how chance levels are calculated and significance is evaluated highly depends on the dataset, the recorded samples, classes etc. and can lead to misinterpretation of results. In the next section, we will shed some light on reporting significance of performance in Speech Imagery BCIs and BCIs in general.

2.9 A Word on Significance of BCI Classification Results

BCI research addresses a topic which is still widely considered to be science fiction by the common population. In the last sections we gave an overview on different work in the field with the purpose to illustrate how this technology has developed over the last decades and that we are not too far from real-world application beyond fiction. Various research institutions from around the world publish studies presenting promising results and report numbers for classification accuracy of the trained classifiers, however, those numbers need to be handled with care as we will elaborate on in the following.
The most common performance evaluation criterion in BCI research is still the classification accuracy of a developed system, given as number of correct predictions over the total number of predictions [17]. Usually this number is compared to the theoretical chance level of the classification problem which is the accuracy under the assumption, that all predictions were made by pure chance, therefore resembling a randomly predicting classifier. It can be calculated as 100 divided by the number of classes in the dataset resulting in a certain percentage value, the percent theoretical chance level of the classification problem. These numbers give an impression on the quality of the classifier output compared to a random one predicting at chance, and are therefor frequently used as performance estimator of the BCI classification. However, accuracy does not take class balance into account which means that if one class is present in the dataset more often than the other, the accuracy might be high, even for a randomly performing classifier. One way of handling this issue is the Cohen's Kappa which is calculated as

$$\kappa = \frac{p - p_0}{1 - p_0}$$

where p is the accuracy of the classifier and p_0 the chance level of the classification problem.

Figure 2.24 (taken from [17]) illustrates the problem of class imbalance and the different effect on classification accuracy and Cohen's Kappa for a binary classification problem. In the confusion matrix on the left we can see that our classifier predicted 50% of the time class 1 and the other 50% class 2, resulting in an accuracy of 50% and a Cohen's Kappa of 0 indicating chance level for both metrics. In the confusion matrix on the right we can see the case in which the classifier solely predicts class 1 resulting in a high accuracy (p) of 90% due to class 1 representing 90% of the classes in the dataset. Cohen's Kappa (κ) on the other hand is not affected by this imbalance and still results in the minimum value of 0, indicating pure chance agreement.
Reporting a high classification accuracy does therefore not always represent a well working BCI classifier. In order to avoid this pitfall, we took care, that all of our studies

		Predicted		
		1	2	
Actual	1	45	45	90
	2	5	5	10
		50	50	100

$p = 0.5 \quad \kappa = 0$

		Predicted		
		1	2	
Actual	1	90	0	90
	2	10	0	10
		100	0	100

$p = 0.9 \quad \kappa = 0$

Figure 2.24: Confusion matrices of a binary classification problem with imbalanced classes and the effect on classification accuracy and Cohen's Kappa, taken from [17]. Left, p and κ manage to detect the chance level classification where on the right only κ can reliably detect the random classification.

resulted in class balanced datasets. Although using Cohen's Kappa would have been a valid alternative, we decided for this option as all of our studies in this work were conducted in controlled laboratory settings without free interaction of the participants, which allowed to predefine parameters like trails and repetitions.
Speaking of trails and repetitions, those parameters obviously affect classification numbers as well. Müller-Putz et al. [152] highlighted this issue in 2008 in their paper "Better than random? A closer look on BCI results.". Within their work they emphasized the importance of the number of trails used in classification problems and their influence on the chance level, because a common problem in BCI studies are the usually small datasets. These rather small datasets with only few repetitions per class result on the one hand from cumbersome study setups, which can last up to one hour, depending on the number of electrodes used. On the other hand, the study procedures and paradigms are usually time-consuming as well when trying to record sufficient number of repetitions. With ongoing study duration the mental condition of participants often changes as it is impossible for most human beings to remain the same mental state as for example level of focus and concentration for several hours. These circumstances lead to changes of cognitive processes and therefore background activity recorded via the EEG affecting the classification process. This aspect leaves BCI researchers always with the trade-off between the quality of the recorded data and a sufficient number of repetitions because few trails, or predictions of the classifier, do not adequately reflect the performance in a long term use.
Within our work we followed the recommendation of existing studies in order to ensure to record a sufficient number of trails which will be presented in more detail in the following chapters during the corresponding methodology sections of our studies.
However, due to the changes in brain activity over the duration of a study, the number of trails will always be limited and in no case come close to the required conditions to meet the theoretical chance level, which would technically require an infinite number of predictions. Therefore, measuring the performance of a BCI classifier based on the comparison of classification accuracy and the theoretic chance level alone can in the majority of cases not be considered valid.
Combrisson and Jerbi addressed this problem in their work from 2015 [39], where they presented a straight-forward method to derive a statistically significant threshold that accounts for sample size given the limited dataset sizes of most brain signal classification experiments. Their assumption is, that the classification errors obey a binomial cumula-

tive distribution taking into account the total number of samples n and classes c. The probability for correct prediction of a class c at least z times by chance is then given by

$$P(z) = \sum_{i=z}^{n} \binom{n}{i} \times \left(\frac{1}{c}\right)^{i} \times \left(\frac{c-1}{c}\right)^{n-1}$$

where z can be calculated as

$$z = \alpha \times n$$

where α is the significance level.

By calculating the binomial inverse cumulative distribution one can compute the statistically significant threshold for a selected significance level. As we will show in the later chapters, a classification with a testset size of 100 samples in a 5 class classification problem leads to a significance threshold of 27% with an α of 0.05.
Compared to the theoretic chance level of 20% calculated as 100 divided by 5 (classes) we can clearly see the difference and the problem that arises with the the theoretic chance level for performance evaluation of BCI applications.
Within this work we will compare our classification results against the significance threshold calculated based on the parameters of the different datasets with reference to this section.
Having those consideration and pitfalls in mind, we took greatest care during the design of our studies and production of our datasets as well as the later analysis of the data to report statistically valid results.

2.10 Summary

In this chapter we provided a broad overview on background including the state-of-the-art in Brain-Computer Interfaces (BCIs), different brain measures to establish BCIs, a summary on silent speech interfaces and a short introduction on neural correlates of speech production processes in the brain.
We concluded, that the combination of silent speech interfaces and BCIs, to establish Speech Imagery BCIs offers in our opinion, the most intuitive way of brain computer interaction and presented different ways and related work on paradigms, concepts and measurement methods for Speech Imagery BCIs, which prove the feasibility of decoding imagined speech from brain activity.
Our conclusion on this related work is, that word based imagined speech, in combination with EEG as measurement method, offers the highest potential to establish Speech Imagery BCIs in real-world applications, due to the flexibility of the word based interaction and the great temporal resolution, non-invasiveness and ease of use of the EEG.
Subsequently we gave an overview on related work in the field of EEG-based Speech Imagery BCIs with a focus on word imagination. The conclusions from this related work support our claims from the motivation presented in chapter 1 that

1. Training procedures for EEG-based Speech Imagery BCIs are tedious long-lasting processes bearing pitfalls in design, which might lead to unintended effects in brain activity.
2. Classification accuracies of EEG-based Speech Imagery BCIs are currently below feasible for real-world application.
3. EEG-based Speech Imagery BCIs are usually established with cumbersome and bulky setups with numerous electrodes.

We will continue with addressing each of those points in the following chapters starting with the training procedures in chapter 3, insufficient classification accuracy in chapter 4, and electrode reduction in chapter 5.

Chapter 3
Improving Speech Imagery BCI Training procedures

In this chapter we introduce two new concepts for training EEG-based imagined speech BCIs, which are supposed to make the usually long-lasting, mentally and physically exhausting training procedures more comfortable for the later user.

The first approach is based on the idea that silently reading is on a conceptual level similar to silently speaking. Hence, we tried to train a classifier on EEG data of participants while they were reading a text including certain keywords. The trained algorithm was later used to predict those keywords silently spoken in an imagined speech task (see section 3.2).
With this setup we wanted to answer the research question:

> **RQ 1.1** Can EEG activity recorded while reading certain words be used to train a classifier to detect those words during imagined speech?

We can foresee several advantages of this concept. First of all, reading a text can be much more entertaining then just repeating single words presented on a screen, depending on the text of course. Second, this method would give the experimenter the freedom to integrate the words into any kind of context in the best case fitting the later application scenario of the SI-BCI. In the case of a scenario in Human-Robot-Interaction, this could for example mean to disguise the words used for later interaction with the robotic system in an instruction manual. Thereby, the human will learn how to interact with the system while simultaneously training a SI-BCI, meaning the robot will learn to interpret the brain activity of the user. We would achieve a bi-directional knowledge transfer from robot to human but also from human to robot making it not only more comfortable for the user to train the system but also more efficient as the required knowledge about how to operate the system could be conveyed simultaneously. Lastly, tracking the positions of the eyes inside the text while reading, gives the experimenter the possibility to verify that the participant has correctly performed the task. Possible skim-reading can however not be eliminated but the verification if the user has moved his eyes over the word in combination with minimum fixation times, can be of help in this case.

The second concept is based on the hypothesis that overtly and silently speaking involves similar brain regions and activity patterns, which should allow to train a classifier on overtly spoken words and let it predict silently spoken words (see section 3.1). The neural similarity of overt and imagined speech measurable even in EEG activity has been shown in several studies [124, 221, 146], however, a real transfer of a classifier from one to the other has, to the best of our knowledge, not been done before. Thus, we consequently applied this approach by letting participants speak certain keywords out loud and silently to train a classifier on the EEG data during spoken interaction and transfer it to the EEG data of the imagined speech interaction. Thereby, we wanted to answer the research question:

> **RQ 1.2** Can EEG activity recorded while speaking certain words be used to train a classifier to detect those words during imagined speech?

We foresee a high potential of this method for training SI-BCIs as it provides several advantages over the standard training procedure.
First, speaking the words might be considered more engaging, but more importantly, a better controllable scenario. If a participant is advised to repeat a word silently for training, the experimenter can never be sure if the person actually followed the advice and repeated the word on the screen, let it be intentional or not. Having the participant saying the word out loud makes it a scenario in which we can control the output and exclude falsely executed repetitions from the training data set. Second, having the later application scenario in mind, the training procedure could actually be performed during interaction with the system. Let us assume we interact with a robot in a calm environment via speech and after a while this environment becomes more noisy, e.g. due to machines running in the background. The system could use the data recorded during spoken interaction and train a classifier simultaneously in order to make it applicable in a later imagined speech interaction. This approach would not only make it more comfortable to collect training data, it would even allow the user to be productively interacting in the given scenario while the classifier is trained in parallel.
Expected complications with muscle artifacts occurring during spoken speech due to the movement of the facial muscles remain challenging but have been addressed in previous work [137, 37, 95] which makes us confident that this promising training approach for SI-BCIs based on overtly spoken words can be realized.

In the following we will present and evaluate our two concepts for automated training starting with the idea of training an SI-BCI based on activity recorded during a natural reading task.

Section 3.1 is based on already published work in [176] while section 3.2 is based on already published work in [178].

3.1 Training SI-BCIs based on EEG Activity Recorded During a Reading Task

In order to make use of the before mentioned benefits in imagined speech BCI training by reading, we designed a study setup in which our participants had to read an instruction manual of a robot including several command words, which were embedded repeatedly in the text. Afterwards silent repetitions of those words were performed in a standard imagined speech training procedure in front of a screen. The EEG-data recorded during reading was then used to train a classifier which was once used to predict on the reading data and additionally on the imagined speech dataset in a transfer approach.
Due to the novelty of the method we further wanted to elaborate on the question if it is possible to classify words during a natural reading task based on EEG activity, as this has, to the best of our knowledge, not been done before as well. Our study setup is described in detail in the following section.

3.1.1 Methodology

The purpose of the reading task study was to measure EEG activity while reading an instruction manual of a robot control and transfer the classifier trained to detect words read on a Speech Imagery dataset. Therefore, the EEG data should be automatically labeled on a word-level based on the position of the eye gaze inside the text, to use this labeled data and create models to detect imagined speech in the later interaction with the robot. In order to record this dataset several preparations had to be made starting with the design of the instruction manual.

Instruction Manual

The instruction manual was written in German and designed in a way to explain the interaction with a robot during an assembly task and how to send commands to the robot by EEG activity alone. 9 command words to be used for the interaction with the robot were repeatedly included in the text namely, "Gehäuse" (case), "Schraube" (screw), "Platine" (circuit board), "Hebe" (lift), "Halte" (hold), "Lege" (put), "Boden" (floor), "Werkbank" (workbench) and "Fließband" (conveyor belt). The command words were chosen to be applicable in an assembly task and to be combinable to formulate certain command expressions e.g., to hold the circuit board over the workbench for assembly. Furthermore, the words were chosen to belong to 3 different semantic categories, namely locational, actions and objects/non-living, in order to be used for the concept of Semantic Silent Speech BCIs as well (see section 4.2). The assignment of each word to a category is illustrated in table 3.1.
In order to collect a sufficient amount of training data for the SI-BCI, the command words were repeated at least 100 times each throughout the text. The instruction manual was designed in a way to appear as natural and close to a normal text as possible, to not disturb the flow of reading. Therefore, it was not possible to limit the number of occurrences of each word to exact 100 repetitions but it was kept above 100 for all of them. This problem was solved during dataset assembly were it was taken care to create a balanced set with the same number of occurrences of each word. Overall the manual included 12 pages with roughly 7000 words and was split into 4 equal parts to ensure sufficient breaks for the participants during the study. Figure 3.1 shows a snippet of the created text with command words highlighted in different colors according to the

Table 3.1: Selected command words and corresponding semantic categories included in the instruction manual, originally in German, translated to English.

Category	**Command word**
Actions	Hold Lift Put
Locational	Conveyer-belt Floor Workbench
Non- living	Case Circuit-board Screw

semantic category they belong to. This highlighting was not shown to the participants during the task and was only added to illustrate the distribution of command words within the sentences.
The text design was kept simple and factual to prevent unintended effects in the EEG due to unexpected formulations or strong emotions to reduce activation of brain regions other than those present in language processing. The template of an instruction manual provided the perfect environment to allow for short, repeating sentence structures with the command words embedded without appearing alienated. However, as the text design was created from scratch we decided to evaluate the quality regarding readability and intelligibility with a questionnaire which all of the participants had to answer during the breaks of the reading sessions. The questionnaire was taken from [62] and included 24 questions divided into 8 categories with 3 questions each. The categories were concerned with word and sentence difficulty, argument and proposition density, effort for inference formation, clearness and language variation as well as intelligibility. The full questionnaire can be found in the appendix 7.4.4.

Der erste Befehl ist der Hebe Befehl. Mit dem Hebe Befehl können Sie dem Roboter befehlen, ein Gehäuse zu greifen. Entsprechend muss nach einem Hebe Befehl eine Angabe kommen, von wo das Gehäuse zu greifen ist. Da der Roboter keine Schraube und auch keine Platine greifen kann, folgt auf einen Hebe Befehl fast immer Gehäuse als Zielobjekt. Als Nächstes kommt der Ort, an dem das Objekt liegt, dass gegriffen werden soll. Ein Gehäuse kann auf dem Fließband, auf der Werkbank oder auf dem Boden liegen. Anschließend wird der Hebe Befehl abgeschlossen, indem dem Roboter mitgeteilt wird, was mit dem per Hebe Befehl gegriffenen Objekt passieren soll. Als erste Möglichkeit kann dem Roboter per Halte Befehl mitgeteilt werden, das Objekt über der Werkbank in eine Halte Position zu bringen, in der Sie das Objekt bearbeiten können. Wollen Sie ein Gehäuse vom Fließband per Halte Befehl in eine Halte Position über der Werkbank bringen, würde der Befehl, den Sie denken müssen, wie folgt lauten:

- "Hebe das Gehäuse von dem Fließband und halte es."

Alternativ können Sie die Reihenfolge des Hebe und Halte Befehls auch kombinieren:

- "Hebe und halte das Gehäuse vom Fließband über der Werkbank."

Figure 3.1: Example excerpt of the instruction manual in German. The target words are highlighted just for demonstration purposes and were not visible to the participants during the task. Actions are marked in blue, locations in green and non-living in red.

Subjects

We conducted the study with 23 healthy subjects (16 male, 7 female) with an average age of 28 years all with normal or corrected-to-normal vision. All subjects were right-handed and native German speakers, as our instruction manual was written in German. Each subject was introduced to the task, and informed consent was obtained from all subjects for scientific use of the recorded data. The study was approved by the ethical review board of the faculty of Mathematics and Computer Science at Saarland University[11].

Recording

The data was acquired in a dim light room with minimized distractions like external sound, mobile devices etc. The voluntary participants were asked to sit in a comfortable chair to prevent unnecessary muscle movements to reduce noise and artifacts in the EEG, which could emerge from mental stress, unrelated sensory input, physiological motor activity and electrical interference. EEG signals were recorded using a wireless 64 channel EEG system namely Brain Products Live Amp 64 [12]. The sampling rate was set to 500 Hz. The 10-20 International System of electrode placement was used to cover the whole scalp resulting in the capturing of spatial information from the brain recordings effectively [144].
For simultaneous eye gaze recording we used the Tobii Pro Fusion eye-tracker [13] attached to the screen to track the participants eye gaze inside the text while reading. Based on this data we were able to label the EEG data on a word level after the experiment. The eye gaze was recorded with a sampling rate of 250 Hz.
We integrated the instruction manual into a Unity application for presentation (see figure 3.3 left) which allowed the adjustment of the text properties like line spacing and margin to the borders of the screen, to improve the reading comfort and facilitate the eye-tracking inside the text. Due to this text adjustments only a small excerpt of the text could be shown on the screen at once and the participants were enabled to switch to the next page via a keyboard button press. This setup and design of the text presentation was adapted from Hollenstein et al. [85]
All the before mentioned components were connected to and their software running on a Windows machine which allowed a synchronization of the recorded data based on timestamps after the experiment.

Study Setup

The objective of this study was to train a classifier for SI-BCIs on EEG-data recorded while reading. In the reading task, participants had to sit down in the chair which was adjusted to have an optimal view on the screen (see figure 3.3 left). The eye-tracker was calibrated before each reading task and the impedances of the EEG were checked to ensure best possible signal quality. The text was split up in 4 parts to allow sufficient breaks in between each session for the participant to rest and prevent to induce too much cognitive load. Each of those parts lasted for approximately 15 - 20 minutes depending on the individual reading speed. There were no time limits set and the participants were allowed to re-read sentences or paragraphs if they were unsure about the content.

[11] https://erb.cs.uni-saarland.de/ Last accessed: 15.09.2022

[12] `https://brainvision.com/products/liveamp-64/` Last accessed: 18.09.2022

[13] `https://www.tobii.com/products/eye-trackers/screen-based/tobii-pro-fusion` Last accessed: 18.09.2022

The navigation between the separate pages of each reading task was realized with a keyboard button press. After each reading part, there were 4 multiple choice questions related to the previously read content in order to exclude the data of participants which might just have skimmed the text without consciously reading it and therefore not able to answer the questions. The question answer session was followed by a 5 - 10 minutes break depending on the individual needs of the participant. During those breaks the participants filled out a questionnaire concerning the quality of the text (see Appendix 7.4.4) and were allowed to stand up, drink some water and clear the mind for the next reading task. Furthermore those breaks were used to check the impedances of the EEG-headset and the eye-tracker was re-calibrated after each break.

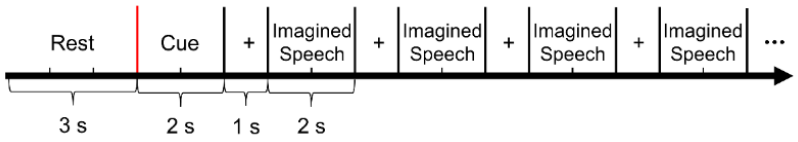

Figure 3.2: Procedure of the silent speech task of our study adapted from [123]. The target word was presented once and had to be repeated silently 5 times, before proceeding to the next word.

In the speech imagery task, each participant had to perform 4 sessions of standard silent repetitions where each of the key words of the text was presented on a screen. In this study we followed the setup of Lee et al. [123], which shows the keyword to be repeated on the screen once and gives the participant several repetition phases in a row afterwards, indicated by a blinking fixation cross, as shown in figure 3.2. This procedure reduces the overall time for the task by increasing the number of repetitions after each key word presentation, saving 2 second each time a word is presented. Over the duration of a whole experiment those 2 seconds sum up to several minutes which allows to collect more samples of each word. In order to prevent the problem of block-wise stimulus presentation as mentioned in section 2.5 the number of repetitions in a row should be kept reasonably low and was chosen to be 5 in our study. Furthermore, we decided to replace the auditory cue in the beginning by another visual one to reduce evoking too many different potentials in the brain activity of the participant by cue onset triggers. The overall 4 session were split up in two parts, two prior and two post reading task. This procedure should allow to investigate effects of reading the words inside a certain context on the brain activity by analyzing the data separately, post and prior to the reading, but also combined. If the reading task had an impact on the process of silently repeating them, as their might be specific associations in a context non existing prior to reading the text, the evaluation on the Speech Imagery dataset recorded post reading task, should deliver better results than the pre reading task dataset. In each of the sessions the target word was presented on the screen (see figure 3.3 right) ten times for each word per session resulting in 40 silent repetitions of each word and 360 silent speech samples overall.

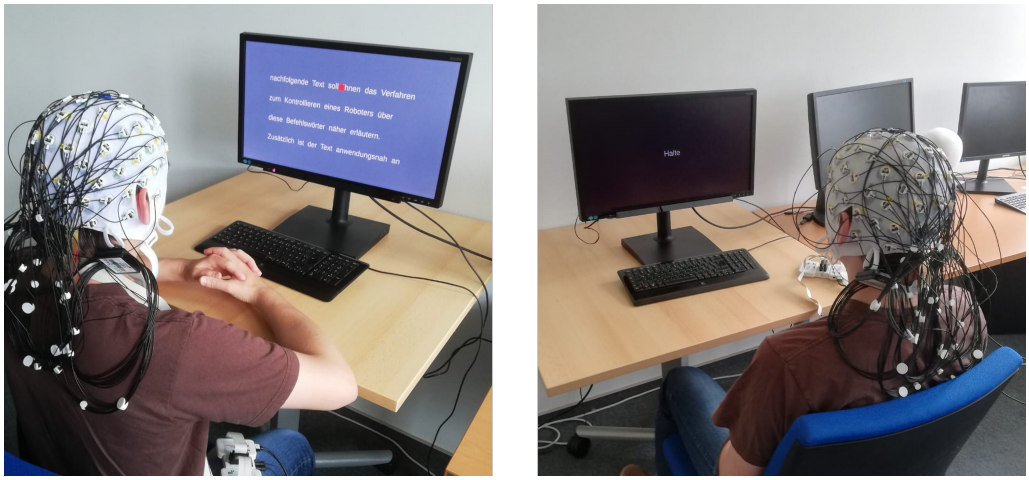

Figure 3.3: Study setup of the reading task (left) and silent repetitions (right). The red dot on the screen (left) shows the eye gaze of the user in the text measured with the eye-tracker attached to the bottom of the screen. The dot is used just for illustration and calibration purposes and not visible to the participant during the experiment. On the right, the word to be repeated silently is presented in the center of the screen.

Data Analysis

We started our analysis with the preprocessing of the different data streams, EEG and eye-tracking, cutting them in epochs and extracting features for the two different paradigms, reading and silently speaking. Afterwards the data was fed to different classifiers to investigate the transfer approach from reading to silently speaking, but also within the same paradigm in order to get a benchmark for comparison against the standard imagined speech classification. Furthermore, we tried to train a classifier on the reading data and tested it on the reading data to explore the potential of classifying words read based on EEG activity. We therefore ended up with three different evaluation scenarios namely:

1. Train on Silent -> Test on Silent (Benchmark 1)
2. Train on Reading - > Test on Reading (Benchmark 2)
3. Train on Reading -> Test on Silent (Transfer)

The different steps of the data analysis will be explained in more detail in the following.

Preprocessing The visual inspection of the EEG data resulted in the rejection of six subjects due to overall noisy data, probably resulting from poor electrode-to-skin contact, leaving a total of 17 subjects (13 male, 4 female).
For filtering we applied a method based on a recent work of Mini et al. [146], which uses a IIR notch filter at 50 Hz and a 20th order FIR band-pass filter with a Hamming window between 0.5 and 60 Hz. This filter performed well on imagined speech data collected in their study. Blind Source Separation was performed on the reading data to remove artifacts induced for example by eye movement during reading via automated Independent Component Analysis (ICA) of the EEGLAB toolbox in Matlab [27]. Rejection thresholds were set to the default values of 0.9 for eye and muscle artifacts while

components without a clear assignment to a group were rejected at 0.95. The eye tracking data was filtered online during reading with an I-VT-Filter based on [161] provided by the manufacturer of the eye-tracker. The purpose of the filter is to separate the saccades from the actual fixation and allow more precise tracking of the gaze of the user. At the time of the start of the fixation of a certain word, as well as at the end, an event is set inside the EEG data for later analysis. Offline, we implemented the methods from Hollenstein et al. [85, 86] to further clean the data of unwanted effects occurring during natural reading e.g., line jumps or repetitions and other events which could cause an unwanted and unstructured fixation and therefore labeling of words not related to the actual reading process. According to their description we excluded fixations of less than 100 ms and over 750 ms, as these are unlikely to reflect fixations relevant for reading. These exclusion criteria led to an average of 100 remaining samples per word for each participant, after leveling the number of samples to the least number of occurrences per participant to ensure a balanced dataset.

Epoching The epoching of EEG data during the reading task was done based on the eye-tracking data after the experiment. Data streams were synchronized and the events collected from the EEG data to investigate the reading durations of the different words and choose the best time window which could be used consistently in the reading and imagined speech data. Choosing epochs of the same length is important for several reasons. First, during the imagined speech task, we do not know when the person started or stopped speaking. The reading data however gives us a good insight on the average duration it took the person to read a word silently, which should be comparable to the time it takes to speak the word silently. Tailoring time windows to individual words however, would give the classifier an advantage as with a changing length of the signal we influence certain feature extraction process and provide unintended additional information. The classifier could potentially make decision based on the epoch length and the differently calculated features based on the word length rather than actual brain activity. Following those assumptions we had a look at the average reading times for the words and all participants calculated based on the eye-tracking data (see table 3.2).

Table 3.2: Average reading times and standard deviation in ms for the different key words averaged for all participants.

Word	**AVG (ms)**	**STDEV (ms)**
Halte (Hold)	276	96.4
Hebe (Lift)	286	79.0
Lege (Put)	276	82.3
Boden (Floor)	272	98.2
Fließband (Convyer Belt)	311	103.2
Werkband (Workbench)	321	91.1
Gehäuse (Case)	290	79.1
Platine (PCB)	284	86.5
Schraube (Screw)	300	69.1
Average	290.6	87.2

We found, that the average reading time of the words did not significantly differ and averaged around 290 ms. Not surprisingly, words with the most characters also scored maximum reading times, namely "Fließband" and "Werkbank" with 311ms and 321ms respectively, as compared to shorter words as "Lege" and "Boden" with 276 ms and 272 ms respectively. Nevertheless, we decided to apply an average window size of 300 ms for all participants and independent of the word. There were two main reason for this decision, first this procedure facilitated the dataset assembly and created equal conditions for the feature extraction methods and classification algorithms due to equal signal length as mentioned earlier. Second, the average reading times of each of the words were almost identical as seen in table 3.2 with a difference of only 50 ms between shortest and longest duration. This time span is very unlikely to include a significant amount of data of other words which might impact the cognitive processes and influence the classification and at the same time include most of the crucial information of the longer words. Based on those results we decided to cut our EEG signal for benchmark condition two, reading on reading, into 300 ms epochs starting from the time point of first fixation.
For benchmark condition one, imagined speech on imagined speech, we used a fixed time window of 2000 ms, the whole time that was given to the participants to perform the repetition.
In order to achieve best possible conditions and resulting accuracy in the transfer task, we experimented using different epoching techniques. We tried to avoid problems induced by parafoveal view, which occurs when the reader picks up visual information of the next word in the text while still reading the currently fixated one [189]. This parafoveal information has a large influence on saccade length reducing the time it takes to fixate and read an upcoming word. We tried to take those circumstances into account by epoching the data with reference to the center point of a word's EEG-signal as shown in figure 3.4, determined based on the simultaneously recorded eye-tracking data. Moreover, we used the approach of starting the epoch at the moment we detected the subjects eyes starting to fixate a word. Both methods were used to create epochs with lengths of 300, 500, 1000 as these window sizes were effective in our previous works [179, 180]. Additionally we attempted classification over the whole span of imagination during the imagined speech task from 0-2000 ms.
The test set, including the imagined speech data, was epoched in different ways as well. A first approach included fixed windows of 300, 500, 1000 and 2000 ms. In an attempt to increase accuracy, the complete interval between the start and end of one imagined speech trial was divided into windows of 300, 500 or 1000 ms length with an overlap of 50 ms. These windows were then used for classification. The individual classification results for each window were then used in a majority vote in order to allow the initial classifier to make more mistakes. In both methods we also tried to either begin epoching at 300 ms from the start, to acknowledge the subject's reaction time, or starting immediately after the beginning of the imagined speech task.

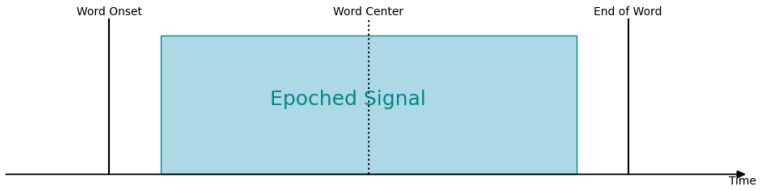

Figure 3.4: A visualization of one example of epoching done on the reading data, acquiring the epoch around the center of the EEG-signal corresponding to a target word.

Feature Extraction In the feature extraction step we decided on two common methods to compare on our dataset, the pyEEG feature vector as mentioned in our works [179, 180] and section 4.1 and 4.2 as well as the Discrete Wavelet Transform (DWT). For the feature vector we followed the exact same implementation as in section 4.1 and 4.2 with different features from time and frequency domain namely, Power Spectral Intensity, Relative Intensity Ratio (for Alpha, Beta, Gamma, Delta and Theta), Hjorth parameter, Spectral Entropy and Skweness.
The DWT implementation followed the approach of Torres-Garcia et al. as presented in [206]. We used the PyWavelets package [121] in Python with the multilevel decomposition function pywt.wavedec() and calculated the four statistic values maximum, minimum, average and standard deviation as well as the relative wavelet energy (RWE). Several kinds of wavelets and different decomposition levels were assessed and compared on a subset of the participants resulting in the best combinations given as ['waveletName',decompositionLevel] = ['rbio3.5', 4], ['rbio3.3', 4], ['rbio3.3', 5], ['coif9', 4]. Those configurations were applied and compared for all participants in the later classification process.

Classification The extracted features were forwarded to an Extreme Gradient Boosting classifier (XGB) which outperformed standard methods as e.g. Random Forrest or Support Vector Machine in a preliminary evaluation of classification performance on the collected data. XGB was implemented according to [34] with a mean error as evaluation metric and instructed to stop if the mean error did not decrease for ten rounds. The objective function was chosen to be softmax for multiple classes. In the benchmark conditions in which we trained and classified on data within the same paradigm, reading on reading and imagined speech on imagined speech, we ended up with a dataset for reading included 100 samples for each participant and word, where for the imagined speech data we collected 40 samples per word. 10-fold cross validation was used on both datasets. In the transfer condition we used the complete set of reading data for training and tested on a randomly chosen 10 percent subset of the imagined speech data ten times to ensure equal conditions as compared to the benchmark scenarios.

Performance metrics and statistical analysis The performance of the targeted transfer and the different classification processes imagined speech on imagined speech, reading on reading, as well as the transfer from reading to imagined speech was assessed based on the classification accuracy in comparison to the respective significance threshold. As explained in section 2.9, BCI experiments and datasets usually include a limited number of samples and the chance level depends on those parameters. Combrisson and Jerbi proposed an adjustment of the chance level for machine learning in neuroscience [39] which we calculated based on their work as explained in section 2.9. With the 9 words and an alpha value of 0.05 we received a significance threshold of 16.66% for the reading

condition with its 90 samples in the testset, for the imagined speech data with 36 samples in the testset a threshold of 19.44% and finally, for the transfer approach with 360 samples in the testset a significant threshold of 13.88%. In comparison, the theoretic chance level for a 9-class classification problem would have resulted in 11.11%.
The classification accuracy of our classifiers was calculated as the number of correct predictions divided by the total number of predictions and compared with the significance threshold to validate statistical significance of the results.

3.1.2 Results and Discussion

In the following we will present the results of the analysis on the recorded dataset, performed as described in the last section. We will start with presenting the results and discuss them afterwards.

Results

The questionnaire on text quality contained 24 questions from 8 different categories as mentioned in the methodology section. The participants had to answer this questionnaire after each of the four reading tasks during the experiment, represented as R1, R2, R3 and R4 in table 3.3, which shows the scores for all the reading tasks averaged for all participants split up in the 8 categories. Each question requested an answer on a Likert scale from 1 (strongly disagree) to 5 (strongly agree). The 8 categories can be grouped again based on the interpretation of the score in difficulty and effort for inference formation, the density values for arguments and propositions, and clearness, variation and intelligibility. In difficulty and effort for inference formation, a lower score is preferable, while for the density values one would hope for results around the median. Best results for clearness, variation and intelligibility correlate with higher values.
As shown in table 3.3 the participants rated the words used in the text as rather easy but gave medium values on the sentence difficulty and the effort for inference formation. There was not significant difference between the values of the four reading tasks aside from the effort for inference formation which was rated slightly higher in the R4 as compared to the other 3 reading tasks.
Concerning the density of arguments and propositions, the score should be preferably around the medium value of 2.5, as a high density could overstrain the reader, while a low density could lead to boredom, which was the initial reason to choose reading, to overcome boring training scenarios. The argument density scored an average value of 2.77 almost perfectly matching the targeted medium value of 2.5. In between the reading tasks only R4 scored slightly higher with a value of 3.21. The proposition density did not differ much in between the different tasks and ended up with an average value of 3.59.
The questions for clearness, language variation and intelligibility were formulated in the way that higher values represented the better results. The values for text clearness showed a consistent picture averaging around 3.77 for all of the tasks. Language variation scored rater low with an average score of 2.81 and a slightly better value only for the last reading part R4 with 3.76. The intelligibility achieved overall good results with an average of 3.87 and no significant difference between the reading tasks.
Our results concerning the classification accuracies of the benchmark and transfer scenarios are listed in table 3.4. Within this table we see the individual accuracies for all 17 participants, each condition and the two feature extraction methods which were applied on the data, the pyEEG feature vector (pyEEG) and the discrete wavelet transform (DWT) after 10-fold cross-validation with our Extreme Gradient Boosting (XGB) classifier. The best classification accuracy for benchmark condition one, train and test on imagined speech data, was achieved by participant three with DWT and 49.44 % exceeding the significance threshold of 19.44%. With the pyEEG feature extraction method an individual best of 21.11% could be achieved for the data of subject twelve, slightly above the significance threshold.

Table 3.3: Results of the text quality questionnaire and the average scores for the 8 categories of the 4 reading tasks R1, R2, R3, R4 as well as the overall average. The number in brackets in the AVG column represents the best possible score.

Category	**R1**	**R2**	**R3**	**R4**	**AVG** (best)
Word Difficulties	1.11	1.08	1.08	1.19	**1.12** (1.00)
Sentence Difficulties	2.59	2.71	2.57	2.44	**2.58** (1.00)
Effort for Inference Formation	2.08	2.11	2.10	3.60	**2.47** (1.00)
Argument Density	2.60	2.65	2.60	3.21	**2.77** (2.50)
Propositions Density	3.86	3.56	3.51	3.41	**3.59** (2.50)
Clearness	3.78	3.73	3.57	3.98	**3.77** (5.00)
Language Variation	2.44	2.43	2.59	3.76	**2.81** (5.00)
Intelligibility	3.83	3.81	3.84	3.98	**3.87** (5.00)

Table 3.4: Individual results for all 15 participants and the 3 conditions, Silent, Reading and Transfer as classification accuracy in percent. Results are presented for both feature extraction methods, pyEEG and DWT separately. The significance threshold is shown in the bottom row, once more in percent.

	Silent		**Reading**		**Transfer**	
Subject Nr.	pyEEG	DWT	pyEEG	DWT	pyEEG	DWT
1	17.22	47.78	18.83	19.32	10.56	**16.39**
2	14.44	47.22	16.10	18.58	12.22	13.89
3	11.94	**49.44**	18.47	22.10	11.67	12.78
4	11.67	41.39	16.91	23.22	11.67	12.78
5	16.11	37.50	19.16	18.93	11.94	11.94
6	13.05	29.17	18.17	21.98	11.67	12.78
7	11.39	27.50	18.52	20.94	12.78	12.22
8	18.89	31.11	18.22	18.78	12.78	14.17
9	13.61	42.50	19.85	20.49	12.22	12.22
10	14.17	24.72	14.97	16.33	11.39	15.83
11	12.22	29.44	20.65	21.33	11.94	12.50
12	**21.11**	45.00	18.74	21.46	**13.06**	13.33
13	18.06	38.61	**23.04**	**25.41**	11.94	12.50
14	16.11	45.28	17.88	18.61	9.44	11.94
15	12.78	41.11	19.31	21.66	10.28	13.89
16	13.33	28.06	20.15	22.70	12.78	12.78
17	14.17	44.44	19.67	22.11	12.22	14.17
Average	14.72	**38.25**	18.74	**20.82**	**11.80**	**13.30**
Sig. Threshold	**19.44**		**16.66**		**13.88**	

Overall, the pyEEG feature extraction method performed throughout worse in comparison to the DWT feature extraction for the imagined speech condition with an average classification accuracy of 14.72% below significance threshold and an astonishing 38.25% average classification accuracy for the DWT exceeding the significance threshold by far. This threshold was exceeded by all participants in the case of DWT analysis but only by one single participant in case of the pyEEG feature extraction.
Similar results were shown for benchmark condition two, training and testing on the EEG data collected during reading the target words. For all participants, except participant five, the DWT outperformed the pyEEG feature extraction method, however, with an individual best classification accuracy of 25.41% for DWT compared to 23.04% for pyEEG, not as clear as in the imagined speech condition. The same holds for the average classification accuracies, with a 18.74% for pyEEG and 20.82% for DWT in both cases above the significance threshold. This time, 16 of the 17 participants exceeded this threshold with DWT and 15 with the pyEEG analysis.
The results for the transfer condition, training on reading and testing on imagined speech, did not yield the expected results. With a single best classification accuracy of 16.39% for the DWT the individual managed to exceed the significance threshold, but the majority of the results ended up being below for both methods. In the case of the pyEEG feature extraction none of the participants managed to exceed this threshold with an individual best classification accuracy of 13.06%. The same holds for the average values with 13.30% for DWT and 11.80% for pyEEG. 5 of the participants managed to exceed the significance threshold of 13.88% in the DWT condition but none for the pyEEG feature extraction, indicating an overall bad performance of the transfer approach.
This impression is supported by the boxplots in figure 3.5 . Due to the throughout better performance of the DWT outperforming the pyEEG feature extraction method in over 98% of the cases, we only included the DWT results in the plots. The red dashed line highlights the theoretic chance level of 11.11% while the green dashed lines show the significance threshold calculated for each of the three conditions separately based on the number of samples in the testset as explained in chapter 2.9. We can see that for benchmark condition one (Silent) all participants exceeded chance level and also the significance threshold by far with a median value of 41.11%. Benchmark condition two (Reading) showed astonishing results as well, with a median value of 21.32% exceeding the chance level for all participant and the significance threshold for all but one. The transfer condition on the other hand did not show results significantly above chance level with a median value of 12.78% and 75% of the participants showing results below or at significance threshold.

Discussion

Summing up the results on the text quality of our created instruction manual we can say, that all 4 text parts were easy to read and clearly understandable. The participants reported that there was not much variation in the language and some of the sentences might have been complex, which was probably due to the nature of an instruction manual being repetitive and the need to include as many repetitions of the target word as possible. However, the high value for intelligibility shows that the text was overall easy to comprehend and fulfilled the purpose of a medium to transport the command words in a way that all participants were able to process the information sufficiently while still being able to focus on the task.

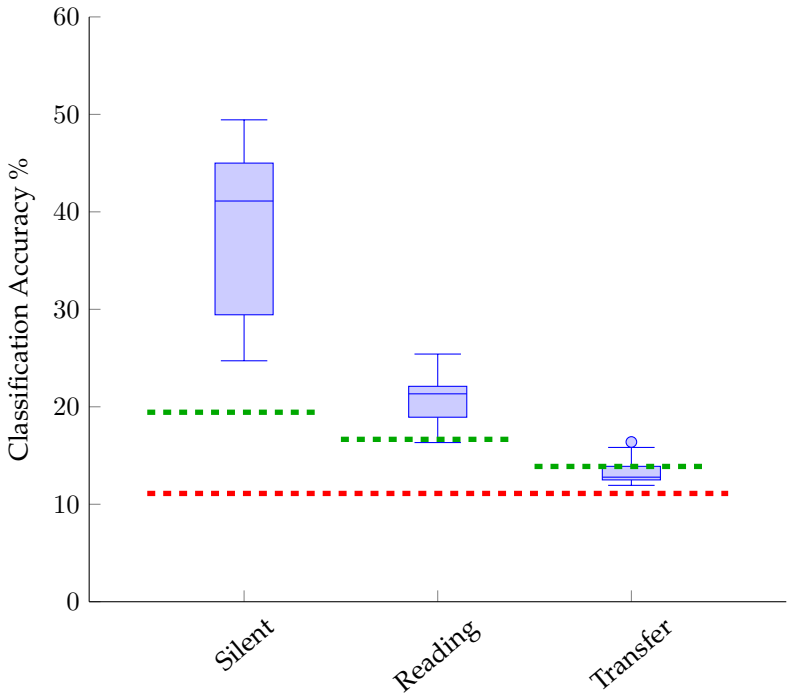

Figure 3.5: Boxplots of the classification accuracies for the 9 different words of the two benchmark and the transfer approach of the individual configurations. The green dashed lines represents the significance thresholds for each conditions, while the red dashed line shows the theoretic chance level.

The results of the classification on the benchmark scenarios, namely train and test on the same condition for EEG activity recorded during reading and silently speaking, are significantly above chance level. With this setup we could show that it is possible to detect silently spoken words but also words silently read in a natural reading task from EEG activity for the first time. The classification accuracies for the reading classification achieved a median value of 21.32% for 9 words which is farily above the significance threshold of 16.66% and clearly above the theoretic chance level of 11.11%. Concerning our research question, if it is possible to classify words during a natural reading task based on EEG data, we would answer with yes. In our case, for the 9 different words embedded into an instruction manual presented as natural reading task, we were able to classify the command words embedded in the text in our offline analysis.
However, the initial idea of the study, to train a classifier on the EEG activity during reading certain words, and classify those command words in an imagined speech task later on, did not yield the expected results. The classification accuracies for this transfer approach did not significantly exceed chance level and RQ 2.1, if the EEG activity recorded while reading certain words can be used to train a classifier to detect those words during imagined speech, has to be answered with no. With our setup it was not possible to exceed the significance threshold and we can imagine several reasons for that. First, the eye movements during natural reading might have corrupted the signal in a way, that even advanced methods for artifact removal, as for example the applied ICA, might not have worked without using an additional channel to record eye movement during the recording. By measuring the Electrooculogram (EOG) with an additional

electrode near the eye simultaneously, eye artifacts can be filtered much better than with automated standard procedures based on comparison with artifact databases which is the case for most automated ICA preprocessing methods. Including EOG in the analysis could improve the filtering process and provide a clearer EEG signal less corrupted by muscle activity. One could also imagine to make use of the eye tracking data and try to create a model, which predicts electrical field distribution based on eye movements.
However, signal quality and corruption with eye artifacts might play a minor role in our opinion, although there is room for improvement of course. A more promising adjustment could be the adaption of the epoching of the signals. As mentioned in the methodology section we have chosen the same time window for all the words in the reading as well as in the imagined speech recording. The reasons for this decision were on the one hand, that it provides a straight forward implementation but is also more robust during feature extraction as some features we used in this study depend on the signal length and might therefore provide additional information to the classifier, risking a classification based on signal length rather than patterns in brain activity. It is also not a common practice in imagined speech classification to use epochs of different length but rather because the start and endpoint of the output of the user can not clearly be defined when silently speaking. During reading with eye tracking we can do so and would suggest for future work to emphasize this approach by cutting the EEG data during reading based on the eye tracking input and focus on features that are not affected by signal length. This would then allow to also cut the imagined speech data more precisely based on the average reading time of the word for example and tailor the time interval extracted from the EEG activity while the participant was silently speaking to this average during reading. As the reading on reading classification exceeded chance level however, we would still consider the approach presented in this study as valid.
Another improvement of the given implementation that we foresee most promising is the adaption of the stimulus presentation. Currently, the train and test task differ in terms of the procedure during interaction. In the natural reading task, the participants were free to choose reading speed and move the eyes in a way it was comfortable for them. During the speech imagination phase, we advised the participant to sit still, pay attention to the screen and repeat a single word presented on the screen whenever the fixation cross appears. This procedure might have on the one hand induced another base state of the users, being focused on an unknown input, influencing their brain activity in general and on the other hand the different context in which the words were presented could have had an impact on the users brain activity as well. Reading words inside a text in comparison to silently repeating them without any context most likely creates overlapping patterns where the patterns for single words can not clearly be isolated. This might not only be due to overlapping time windows that were cut from the continuous signal during epoching, but also due to different brain activity produced by combined words in sentences in general. At this point this is just speculation which needs to be verified in a study more tailored to finding those differences, but for the given approach we see possible adjustments which could improve performance of the transfer classifier.
No matter in which direction, for the reading or the imagined speech part, an adaption to make the train and test task more similar might help to improve the performance of the system. We could imagine for an adaption of the training scenario during reading, to use rapid serial visual presentation. In this method each single word of the text would be shown in the middle of the screen for a small time window and the user reads word by word. Target words could be shown longer and highlighted by different font size or style, to increase the brain response to those special words. On the side of the imagined speech we could foresee to adjust the task in a way, that the person has to imagine a whole sentence which was trained in beforehand and maybe even to change

the paradigm to instruct the participant imagining reading the sentence on a blank screen. This might however have the unintended effect of classifying eye movement rather than brain activity and should be handle with care and adjusted filtering methods.
Concerning our created instruction manual that we used for the reading task, we would conclude, that it did not have a negative impact on the classification results. Our questionnaire regarding the quality of the text and its readability, was throughout rated as positive concerning difficulty of words, intelligibility, density of arguments and propositions from the study participants. Given by this feedback, the text was clearly understandable and only the sentence difficulty might have been slightly to high, probably due to the nature of an instruction manual including the description of complex tasks of interaction. So there is always room for improvement, however, a negative impact on the classification results can most likely be excluded.

3.1.3 Conclusion

In this study we tried to train an imagined speech classifier on EEG activity recorded during a natural reading task. The reading task was designed as an instruction manual, describing the interaction with a robot based on a BCI. Within this instruction manual we have embedded 9 keywords which can be assembled to form commands for a robot in the given scenario. We recorded eye-tracking and EEG data simultaneously during the reading task and labeled the EEG data on a word level based on the eye-tracking data. The 9 command words were then separately recorded during a standard imagined speech training scenario to create a test set for the targeted transfer from reading to imagined speech classification. The study resulted in a dataset of 17 healthy subjects which performed the reading task in 4 parts in addition to 4 imagined speech tasks including the 9 command words. The text quality of the created instruction manual was assessed immediately after each reading part with a questionnaire and resulted in overall good results concerning clearness, intelligibility and difficulty of the text. EEG data was extracted on a word level based on the recorded eye-tracking data and a classifier was trained for three different classification scenarios. Two benchmark scenarios consisted of a training within the same paradigm, reading on reading, and imagined on imagined speech, while in the third scenario the classifier was trained on the reading data and applied on the imagined speech data. The two benchmark scenarios yielded results significantly above chance level. For the imagined speech, an average classification of 38.25% was achieved with a significance threshold of 19.44% and for the reading on reading an average of 20.88% with a significance threshold of 16.66%. Our transfer approach did unfortunately not exceed chance level. As possible reasons we could foresee the strong difference in between the test and the training scenario as silently repeating single words is too different from reading them in a text. For future work we would recommend to include an additional channel to measure eye movement if proceeding with the given natural reading task or adapt the tasks to be more similar. For the reading we foresee rapid serial visual presentation as promising stimulus presentation method, as it is close to the imagined speech scenario. The silent repetitions could be adjusted to include full sentences rather than single words, which would be closer to natural reading.
Overall, our study delivered valuable results with insights on the pitfalls of a transfer training approach from silently reading to imagined speech. Our successful classification of 9 words from EEG activity during natural reading with a classification accuracy significantly above chance level opens up new areas of research in the field of BCIs and has to the best of our knowledge, not been done before.

3.2 Training SI-BCIs based on EEG Activity Recorded During Speaking

Another paradigm similar to silently speaking words, and therefore a promising candidate for a transfer approach, might actually also be considered the exact opposite of it, overtly spoken speech. As presented in section 2.6.4, neural activity patterns of overt and imagined speech have been shown to be similar and activate related brain regions, measurable even in EEG activity [124, 221, 146]. However, a real transfer of a classifier from one paradigm to the other has, to the best of our knowledge, not been done before and shall therefore be addressed within this section. Additionally, we wanted to investigate, if our new training method delivers comparable results to a standard training approach for imagined speech. Therefore, we created a study setup in which the participants had to move a virtual robot presented on a screen through a maze like factory environment with 5 command words. This robot control was performed in two sessions with overt and for another two sessions with imagined speech in order to transfer the classifier between the two different interaction paradigms. The study setup and data analysis are explained in detail in the following.

3.2.1 Methodology

The objective of this study was to train a classifier on EEG activity recorded during overtly speaking words and let it try to classify on EEG data recorded while speaking the same words silently. This setup had the purpose of making the usually long-lasting training sessions more interesting for the participants. Furthermore, we decided to present the task of silently and overtly speaking certain words in a game-like scenario as an additional step towards more comfortable Speech Imagery BCI training scenarios. Recently, related works have proven that a gamified study setup can increase engagement [182] and performance [198] of the participants in BCI training which would benefit the performance of the system and the quality of the collected data as well. We therefore designed our study to record data from overtly and silently spoken speech in a game-like interaction scenario with a virtual robot based on the guidelines provided in [133].

Game Design

The game was developed with GDevelop[14] an open-source game development software and was designed to resemble a real-world task of guiding a robot through a factory which might be applicable in a teleoperation scenario. This setting was chosen in accordance to Lotte et al. [133] to provide a relevant scenario with an understandable purpose to motivate participants over the long duration of the study and prevent brain states non related to imagined speech, for example due to boredom or stress. The user was presented a birds-view of the surrounding with the robot in the middle (see figure 3.6) and had to decide about its next step. The interaction was triggered for one part of the study via overt and in the second part via imagined speech. The view was limited to a small excerpt covering just the near surrounding of the robot to allow the users a decision about the obvious next step and at the same time preventing them from thinking too far ahead, which might lead to distraction. The interaction consisted of moving the robot in 3 different directions resulting in the command words "left", "right" and "up".

[14] https://gdevelop-app.com/

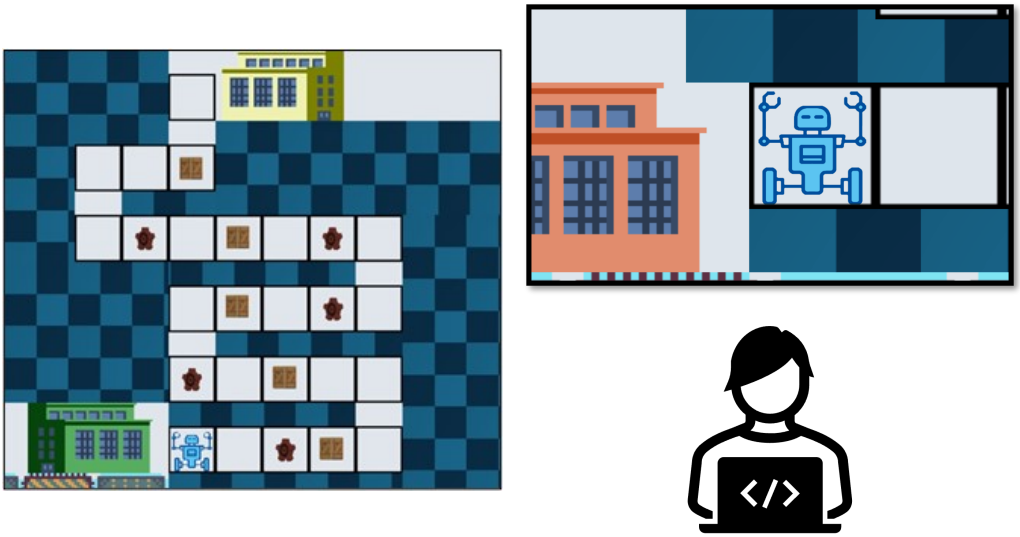

Figure 3.6: Birds view of one level of the robot game, left, and the excerpt shown to the participant during the study on the right.

In order to include another category of words apart from directions into the task, we furthermore used the words "pick" and "push" as robot actions. Those actions were embedded into the scenario with icons of gears and boxes. So whenever the next field on the robots way contained either a gear or a box the interaction required the word "pick" or "push", respectively. If the next field was blank, a movement in the corresponding direction was required via the directional commands left, right and up. Those words were chosen to be easy understandable even for non-native speakers and because they could be integrated into a Human-Robot-Interaction scenario in the future as previously mentioned. Furthermore, the interaction should be kept as simple and clear as possible, meaning that the next command to trigger should be obvious. Those requirements were made in order to induce as less stress and confusion as possible to prevent negative impacts on the EEG recording. Furthermore we wanted to break with the controlled training procedures in standard SI-BCI applications in which the user bluntly repeats the word presented on a screen and rather pretend the existence of a free will and control over the scenario, although the way of the robot was of course determined in beforehand. The interaction process during the game is illustrated in figure 3.7. The user was shown an excerpt of the whole factory on the screen with the robot at its current position in the middle. In the case of figure 3.7 we are at the very start of the robots way through the factory with our starting point to the left and an empty spot to our right. Obviously, the next step should be moving the robot to the right. Whenever the user has made a decision about the next command a spacebar press indicates the desire for interaction. This setting was supposed to provide a self-paced interaction by letting participants decide themselves when to interact with the system. According to Lotte et al. [133] this procedure helps to improve the overall experience during BCI training. A blank screen will be shown for two seconds in order to allow the user to prepare for the task and clear the mind. After the two seconds a fixation cross will appear in the middle of the screen, indicating the user to start thinking or speaking the command, depending on the current part of the experiment.

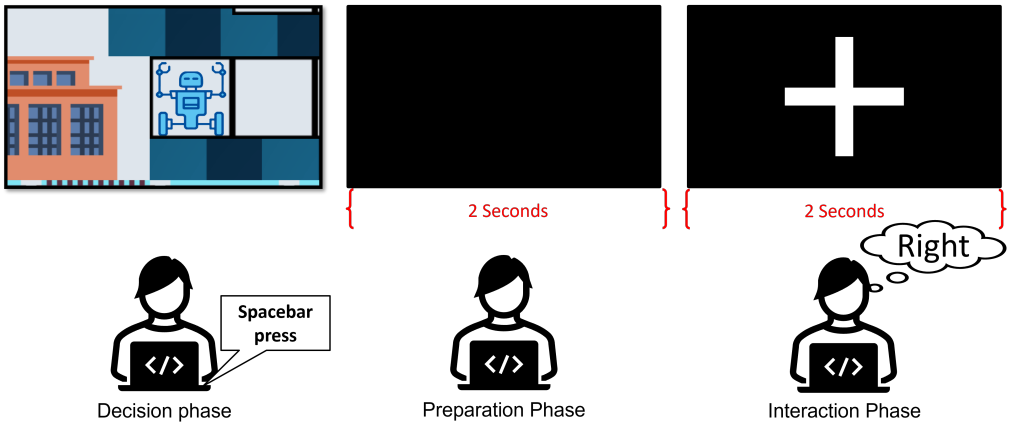

Figure 3.7: Illustration of the interaction process during the game. The user presses the spacebar to trigger a command (left). A blank screen is shown to the user for 2 seconds in order to prepare for the interaction (middle). If the screen shows a white fixation cross in the middle, the user can interact by either saying or thinking the command (right). After 2 seconds the fixation cross disappears and the updated game view is shown.

The cross will disappear after the 2 seconds period automatically and the view will switch back to the robot, now at the updated location, in our example one spot further to the right, revealing the next required action. In order to prevent inducing stress resulting from the pressure of producing correct imagined speech commands, the interaction was triggered correctly independent of the input of the user. The same held for the spoken interaction. In this case however, the experimenter kept track of the output of the participant and marked the wrongly spoken commands to exclude those parts of the recording from the analysis. The participants were informed about this fact prior to the experiment again in order to avoid recording activity resulting from emotional states as frustration or euphoria depending on the performance of the system.

Concerning the game procedure, the BCI approach produced some more requirements. First of all, the words needed to be repeated numerous times to provide sufficient data for training a classifier. Based on the related work in the field which recorded between 50 [221] and 88 [123] repetitions, we decided to settle in the upper half and record 80 repetitions per word and paradigm, meaning for the 5 words 400 imagined and 400 spoken repetitions per participant, resulting in 800 interactions overall. Secondly, the commands should not appear blockwise in a row but rather randomly to avoid the accidental classification of arbitrary brain states as mentioned in section 2.5.

Furthermore, we would need to integrate breaks into the experiment, as doing all repetitions in one session would in the best case take around 70 minutes (5-6 sec per task times 800 tasks), far too long to remain focused. Therefore we decided to split up the interaction in levels of 25 interactions including 5 repetitions of each word in a random order but without repeating a word directly.

For our 800 repetitions, this meant that we had to create 32 unique levels with a random order of the 5 different commands with 5 repetitions of each word in each level. Those 32 unique levels were then split into 4 parts, two for silent and two for overt repetitions. The order of the parts during the experiment was overt, silent, overt, silent, again to prevent recording the data per paradigm blockwise, accidentally resulting in classifying arbitrary brain states rather then cognitive processes (see section 2.7.2).

Additionally a tutorial level was created to let the participant practice and make them familiar with the task to feel comfortable during interaction as suggested by Lotte et al. [133].

Subjects

We conducted the study with 15 healthy subjects, 11 male and 4 female, with an average age of 26.8 years all with normal or corrected-to-normal vision and right-handed. All subjects were non-native English speakers but fluent and experienced with the language as our command words were selected to be English. There was no additional survey of participants' English skills because the 5 words were chosen to be as simple as possible. Each subject was introduced to the task, and informed consent was obtained from all subjects for scientific use of the recorded data. The study was approved by the ethical review board of the faculty of Mathematics and Computer Science at Saarland University[15].

Recording

The data was acquired in a dim light room with minimized distractions like external sound, mobile devices and others. The voluntary participants were asked to sit in a comfortable chair to prevent unnecessary muscle movements to reduce noise and artifacts in the EEG, which could emerge from mental stress, unrelated sensory input, physiological motor activity and electrical interference. EEG signals were recorded using a wireless 64 channel EEG system namely Brain Products LiveAmp[16]. The sampling rate was set to 500 Hz. The 10-20 International System of electrode placement was used to cover the whole scalp resulting in the capturing of spatial information from the brain effectively [144].
The robot game was compiled and executed on the same Windows PC as the recording software of the EEG-Headset, to allow synchronization of the data and events recorded in the game, for example keyboard press or fixation cross. The data was stored on the PC for offline analysis.

Study Procedure

Participants were seated in a chair which was adjusted to have an optimal view on the screen. The EEG-Headset was setup and impedances were checked to ensure best possible signal quality. When the 64-channel EEG headset was fixed on the participants head, they were introduced to the study procedure by playing the tutorial level until feeling accustomed and comfortable with the words and the game mechanic. The participants were informed, that the game would trigger the correct interaction independent from their input, however, that they should still try to produce commands as precise as possible. For the spoken speech the input of the participant was verified by the experimenter for a later exclusion of wrongly expressed commands. After the tutorial the actual experiment and game started. The game was split up in 4 parts to allow for sufficient breaks in between each session for the participant to rest and prevent to induce too much cognitive load. Furthermore those breaks were used to check the impedances of the EEG-headset. Additionally, the participants were asked after each of the 8 levels

[15] `https://erb.cs.uni-saarland.de/` Last accessed: 23.09.2022
[16] `https://brainvision.com/products/liveamp-64/` Last accessed: 23.09.2022

inside each task if they needed a break. Each participant started with a block of overt speech, followed by an imagined speech part, continued with overt speech and did a final block of imagined speech. This shift was chosen mainly to keep the participants attention and provide some sort of variety over the duration of the experiment but also to prevent blockwise recording of the two paradigms.
On average the participants needed around 2.5 minutes per level resulting in 80 minutes of pure interaction for the 32 levels. Including breaks, which were provided based on the individual needs, the preparation of the headset and briefing of the participant, the whole study took roughly two hours on average. After finishing the study, an informal spoken feedback was taken from the participants in order to evaluate how they perceived the study, the game and the training procedure in general.

Data Analysis

Based on the recording in four parts and the two paradigms used, we ended up with four blocks of data, two overt speech blocks and two imagined speech blocks for each of the 15 participants, each including 20 repetitions per word. The data underwent the common steps of signal processing in BCIs, namely Preprocessing, Feature Extraction and Classification.

Preprocessing In a first step, we discarded the trails of overt speech in which the participant did not express the correct command. The reason for exclusion instead of re-labeling was, that the words were mistakenly expressed due to negligence occurring over the duration of the experiment and immediately noticed by the participant. This immediate notice of the self-produced error can lead to the formation of error-related potentials within milliseconds which might corrupt the brain signals [111] and a re-labeling with the actually expressed word would not be beneficial. Therefore those trails were noted by the experimenter during the study and later on excluded from the analysis. However, those mistakes only occurred occasionally and on average only 3 words had to be excluded per participant.
In the second step, the EEG signal was cut into time frames of interest (epochs) in our case the speech production phase with a duration of two seconds. We applied an overlap of 100ms prior to the the actual speech production phase and 100ms post speech production phase. The 100 ms prior to stimulus onset were used for baseline correction of each of the extracted epochs [135, 126]. This involved calculating the average of the EEG signal in the period 100 ms before the epoch which was then subtracted from the actual speech production phase, to eliminate any external events prior to it interfering with the signal during speech production. The 100 ms post stimulus onset were added based on observations of the experimenter during the study, that some participants did slightly exceed the given 2 seconds window to produce the words. This was actually not due to the words being too long, resulting in insufficient time for expressing them, but rather resulting from an individual pace of the participants during execution of the task. Although advised to start imagining or speaking the word as soon as the fixation cross appeared, some participants eventually took longer to react to the onset or maybe lost focus over the duration of the experiment. As already mentioned this was not the case for all participants and all epochs and furthermore not significantly beyond stimulus offset, nevertheless we decided to include a short overlap of 100 ms to cover those possible outliers.
In the next step bad channels were dropped due to poor conductivity and the signal was filtered in two different ways. Especially for the overt speech data this was a crucial step, as the EEG signal is corrupted by the activity produced by the facial muscles during

speech production. As mentioned previously, filtering speech related artifacts from EEG data has been addressed in research and can be summarized to certain frequency ranges being affected by different groups of muscles. The band between 30–40 Hz is affected by the frontalis muscles in the forehead, 50–60 Hz from the masseter muscles in the jaw area, and 40–80 Hz from the temporal muscles [221]. Cutting the signal with a lowpass filter at 30 Hz however would remove too much information most likely also relevant for speech production and therefore the classification process. Varying the filter method between the different paradigms of overt and imagined speech might also induce problems for the classifier because features extracted from the signals originating from different frequency ranges might differ too much to allow the targeted transfer between the 2 paradigms. We therefore decided to apply two different filter methods, one which applies the same filter parameters to both paradigms and a second method in which we differentiate between overt and imagined speech. For the first method we adapted a recent filter setup of Mini et al. [146] used to classify overt and imagined speech separately. The method involved an IIR notch filter at 50Hz to remove powerline noise and a 20th order FIR band-pass filter with a Hamming Window between 0.5 and 60 Hz. The second method was based on one of our recent works [180] for imagined speech detection. We used a 2nd order Butterworth notch filter with cut-off frequencies of 48 and 52 Hz to remove powerline noise. Aftwards a 4th order Butterworth high-pass filter at 1 Hz and a 17th order type two Chebyshev low-pass filter at 200 Hz was used for the imagined speech data, to include the high-gamma bands as we did not expect any speech muscle artifacts in this case. For the overt speech data we applied an additional 4th order Butterworth low-pass filter at 50 Hz. This bandpass filter appeared to be a good compromise by excluding the jaw and temporal muscles partially but including frontalis muscles completely, therefore however reducing the risk of loosing too much information inside the signal.
These two filter methods were applied to the data of all subjects for imagined and overt speech. The dataset assembly however, required several steps as described in the following.

Dataset assembly The study recordings resulted in 4 datatsets, two for overt and two for imagined speech. Special care had to be taken during the dataset assembly as the calculation of the later features with Common Spatial Patterns (CSP) takes into account all of the given samples from one class. Therefore, training and testing data had to be split beforehand to not share information in their features. For the planed evaluation concerning the transfer of a classifier, a test train split was not an issue as the datasets were recorded separately for overt and imagined speech and the train and test set were processed separately anyway. The overt speech data was therefore pre-processed and forwarded to the feature extraction without any further processing. The imagined speech data however, was intended to be used for a standard training procedure within the same paradigm as well, to compare the transfer to a standard method. In order to do so, the data was randomly shuffled and a test-train-split was applied resulting in 75% training and 25% test data. This split was performed 4 times resulting in a 4-fold cross validation and 4 datasets for the imagined speech data for training and 4 different datasets for testing. In conclusion the feature extraction method received 3 different types of datasets, imagined speech training data, overt speech training data and the imagined speech test data, each of them four times due to the 4-fold cross validation. The process of assembling the different datasets is illustrated in figure 3.8. The created datasets were forwarded to the feature extraction algorithm in the next step.

Feature Extraction In the feature extraction step we decided to focus only on one method, Common Spatial Patterns (CSP) and emphasize on the parameter tuning of this one method. One reason for this decision was, that the combination with two different filter

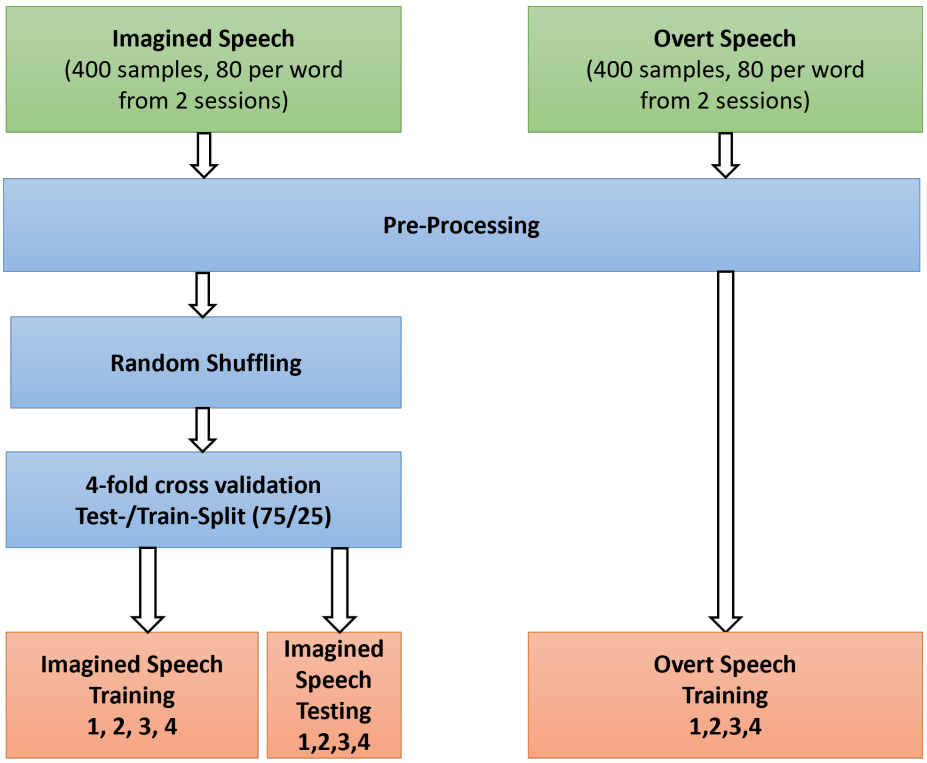

Figure 3.8: Dataset assembly of the recorded data with a random shuffling and 75/25 train test split for the imagined speech data. Due to the separate recording of the overt speech data and no overt on overt classification, the overt data did not need to undergo shuffling and test-/train-split.

methods as well as the two different paradigms overt and imagined speech provided a broad field of evaluation already. A second reason were the remarkable results of Lee et al. [124] with this technique clearly highlighting the similar brain regions of overt and imagined speech.
The CSP algorithm tries to distinguish between two classes by creating a covariance matrix which maximizes the variance for one of the classes and minimizes the variance for the other class, thereby maximizing the difference between the two classes [113]. The results from the CSP computations are different spatial filters equal in number to the input vector, in our case, the 64 channels of the EEG. Those different spatial filters and the appropriate selection for classification is a crucial step in every BCI signal processing pipeline. Clearly not all of those filters contain valuable information for classification and including all of them might end up in increasing training times and computational effort. As those filters or components are ordered inside the resulting covariance matrix according to variance in the signal in ascending order for class A, and descending order for class B, the question is usually just about how many components, starting from the first with the highest variance, to include, and at which point the information is not valuable for the classifier anymore. If the number of CSP components is too small, the classifier will not be able to discriminate the two classes correctly, if it is too high it will lead to overfitting. The number of components can be chosen based on previous

studies, or a cross-validation can be performed to determine the best number for the exact use case [21]. In our work we decided for the later implementation and adjusted the number of components for the data of each individual participant. The CSP features were computed for nine different components between 4 and 12. The 4 components are frequently appearing in literature and are set as the default parameter in most of the signal processing libraries. 12 components were chosen as upper bound as this equals about 20% of our number of channels of the used 64-channel EEG headset. The results of each number of components was calculated for the individual and the best performing setup included into the final evaluation of results, see section 3.2.2.
For calculating the CSP features we used the python library MNE [73] in a one versus rest implementation. In its basic implementation the CSP tries to find spatial filters separating two classes. In the given case however we have a multiclass approach to distinguish between five different words. In order to create features for all the five words and enable a classifier to distinguish between them, we created five different pairs of data. In all of those pairs one class was combined with a set including a combination of the other classes, randomly selected in an equal amount for each class from the remaining data, resulting in two sets of equal sizes, one for the target class and one mixed set of the other classes.
After creating the different training pairs, we applied the feature extraction itself and used each of the training data pairs to create a CSP mask. This mask was then fitted to the training data for each equivalent word and used on its testing counterpart as well to extract the features for classification. This feature extraction process is illustrated in figure 3.9.

Classification Similar to the filtration methods we decided to test different classification methods. In this case we implemented 4 commonly used machine learning algorithms, namely Support Vector Machine (SVM) and Random Forrest (RF) as presented in our recent work [179, 180], a K-Nearest-Neighbor (KNN) and a Gradient Boosting (GB) algorithm. We used the standard parameters as given by the Scikit-Learn library [17] and adjusted the classifiers only slightly. The SVM was implemented with three different kernel types "RBF", "Poly" and "Sigmoid". The RF was tested with different numbers of trees, the standard 100 as well as 50, 300 and 450. For the KNN, the parameter K was chosen to be 7 and the GB was used with its standard parameters except that the maximum depth of the individual regression estimators was set to 1.
Additionally to the two different filter methods and the 9 different CSP components during feature extraction, we trained and tested the data of all subjects for all the 4 classifiers for all of the before-mentioned configurations.

Performance metrics and statistical analysis As a measure of performance of the classifiers we used the classification accuracy given as the number of correct predictions divided by the total number of predictions. Due to the fact that BCI experiments in general include a limited number of samples caused by long and cumbersome study preparations and paradigms, Combrisson and Jerbi proposed an adjustment of the chance level for machine learning in neuroscience [39]. We calculated the significance threshold based on their work as explained in section 2.9. With the total of 800 repetitions, five different classes and an alpha value of 0.025, given because the set included the same words for imagined as for spoken speech, we received significance threshold of 28.00%. In order to answer our two research questions, we performed a Wilcoxon rank sum test as explained in the following section presenting the results.

[17] https://scikit-learn.org/

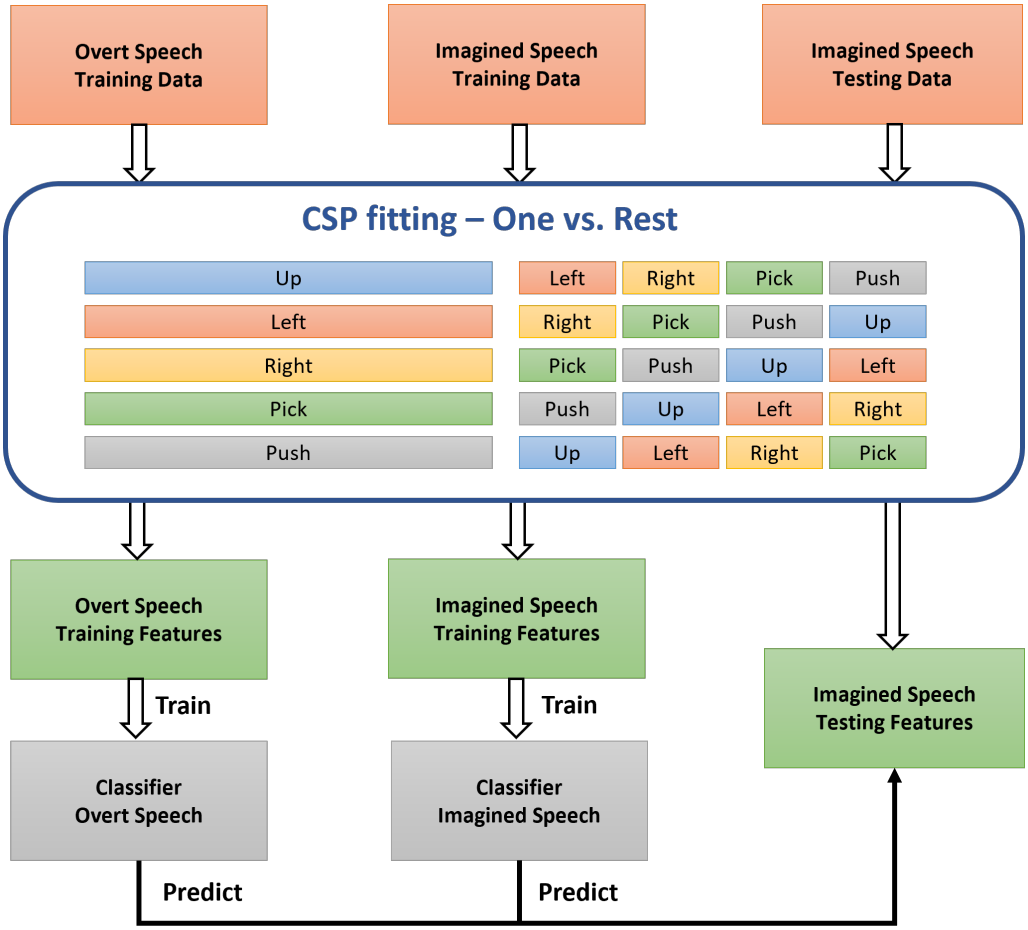

Figure 3.9: Illustration of the CSP processing. Each of the data sets was fitted according to a one versus rest scheme. The resulting features were used for training and testing the classifier for overt on imagined and imagined on imagined speech.

3.2.2 Results and Discussion

In the following we will present the results of the analysis on the recorded dataset performed as described in the last section. We will start with presenting the results and discuss them afterwards.

Results

The analysis included a variety of filter, classification and feature extraction methods as presented in section 3.2.1 which were evaluated on the data of the individual for each possible combination of methods. Table 3.5 shows the best average classification results for the 4 fold cross validation and the standard training approach on the imagined speech data alone. Table 3.6 presents the best classification results averaged for the 4 fold cross validation for the transfer from spoken to imagined speech. Within the table we have listed only the best average classification accuracy for the individual along with the corresponding parameters of the algorithms, namely the number of components used in the CSP, the classifier and the filtering method. Out of those parameters we have furthermore selected the overall most occurring setup of parameters and present the average for this common configuration of the parameters for all subjects as "Common Avg.".

Subject	Accuracy (%)	CSP components	Classifier	Filter method
1	55.75	11	RF	2nd
2	50.25	8	RF	2nd
3	71.00	10	RF	2nd
4	66.00	11	RF	2nd
5	58.50	9	RF	1st
6	90.75	12	RF	2nd
7	96.00	12	RF	2nd
8	69.00	12	RF	2nd
9	63.25	11	RF	2nd
10	95.00	11	RF	2nd
11	94.00	9	RF	2nd
12	76.00	4	GB	1st
13	62.25	6	GB	1st
14	64.50	5	RF	2nd
15	60.00	10	RF	2nd
Unique Avg.	**71.48**	**Individual**	**Individual**	**Individual**
Common Avg.	**69.10**	**11**	**RF**	**2nd**

Table 3.5: Classification results for the standard approach, imagined speech on imagined speech, including the setup parameters, namely number of CSP components, filtering method and the classifier used. The bottom rows show the average classification accuracy for unique parameters and an average accuracy calculated for common best setup.

Both methods, standard and transfer approach clearly exceeded the adjusted significance threshold of 28.00% for all the participants. The standard training approach achieved an astonishing best single subject average classification accuracy for participant 5 with 96.00%. This accuracy resulted from a setup with the second filtration method, 12

CSP components and a RF classifier. This setup was also clearly rated as the best performing overall for the standard training approach, the second filter method was chosen 12 out of 15 times and the RF classifier even 13 out of 15 times. The number of components appeared to perform better in the upper half of the chosen window of evaluation ranging from 4 to 12 components. On average, 9.4 components were present in the best performing sets and the most occurring number was 11 with 4 out of 15 participants. The average accuracy for the individual configurations reached 71.48% as indicated at the bottom of the table with "Unique Avg.". The "Common Avg." as calculated with the most occurring configuration with 11 components, a RF classifier and the second filtration method scored closely behind it with 69.10%.
The transfer approach delivered slightly worse results as shown in table 3.6 with an average classification accuracy for the unique configuration of 61.78% however, still far above the significance threshold of 28.00%. The average classification accuracy calculated with a common configuration on the overall most occurring parameters performed a little worse as well. It ended up 8% lower than the unique configuration with 53.37% for 6 CSP components, a RF classifier and the second filtration method. This might result from the overall more varying individual configurations. The classifier showed once more a preference to the RF with 11 out of 15 participants. The filter method on the other hand did not show the clear preference as for the standard approach although still 9 out of 15 times the second method was chosen. The CSP components ranged a little lower this time with 6 being the most chosen option from the individual setup and an close average number of 6,4 components over all. Surprisingly the highest classification accuracy overall was achieved by subject 7 with 11 components far above this on average lower number of components, the RF classifier and the second filtration method.

Subject	Accuracy (%)	CSP components	Classifier	Filter method
1	62.00	6	RF	2nd
2	35.00	7	RF	1st
3	65.00	5	RF	2nd
4	72.75	6	RF	2nd
5	62.00	5	GB	1st
6	58.25	4	GB	1st
7	83.75	11	RF	2nd
8	69.00	7	RF	2nd
9	62.00	4	RF	2nd
10	79.75	6	RF	2nd
11	76.25	9	RF	2nd
12	35.50	11	RF	1st
13	58.75	6	GB	1st
14	63.75	4	RF	2nd
15	43.00	4	GB	1st
Unique Avg.	**61.78**	**Individual**	**Individual**	**Individual**
Common Avg.	**53.37**	**6**	**RF**	**2nd**

Table 3.6: Classification results for the transfer approach including the setup parameters, namely number of CSP components, filtering method and the classifier used. The bottom rows show the average classification accuracy for unique parameters and an average accuracy calculated for common best setup.

The confusion matrices for this subject are shown in figure 3.10 for the standard and in figure 3.11 for the transfer approach.

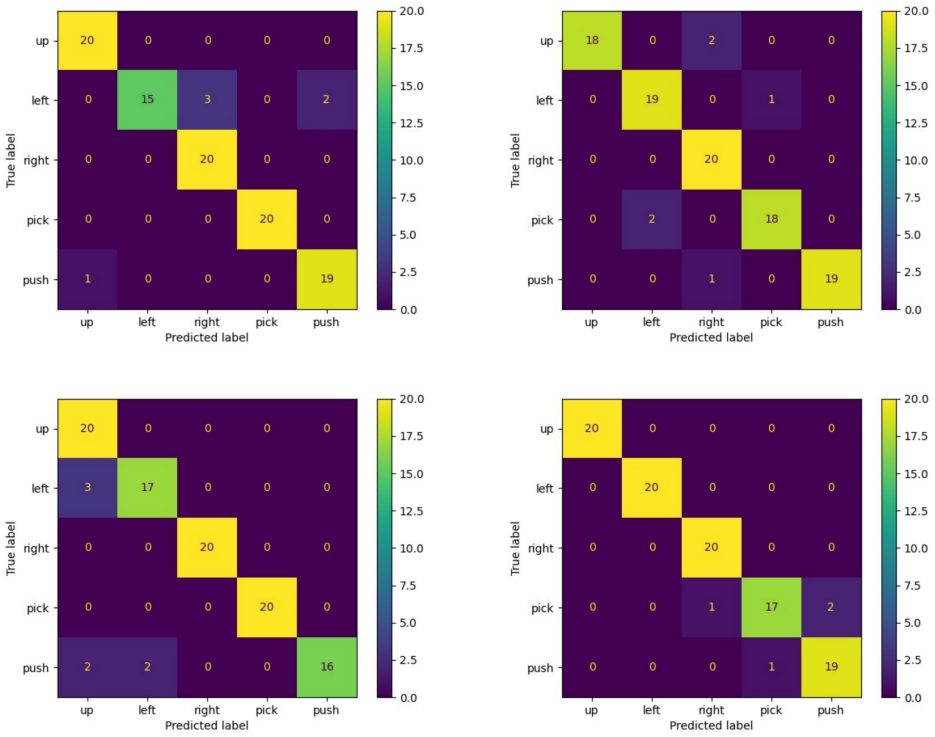

Figure 3.10: Confusion matrices for the classification results of the best performing subject (7) for each of the 4 folds in the standard approach, silent training on silent testing.

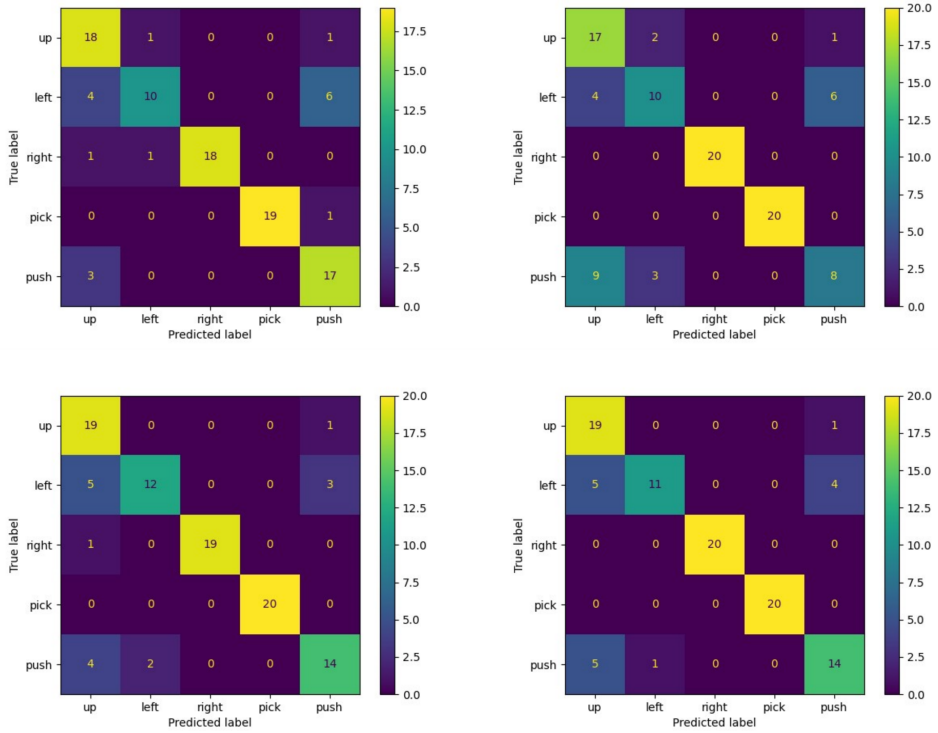

Figure 3.11: Confusion matrices for the classification results of the best performing subject (7) for each of the 4 folds in the transfer approach, spoken training on silent testing.

Figure 3.12 illustrates the classification accuracies of the individual for the standard training approach using silent repetitions and the transfer approach in which the classifier was trained on overt speech and predicted on imagined speech data. The dashed red line represents the significance threshold of 28.00% calculated taking into account the dimensions of the dataset in order to highlight significance as described in section 2.9. This figure illustrates that all participants achieved classification accuracies far above the significance threshold for both of the approaches. A clear tendency towards which method performed better can not clearly be determined from this diagram, although one could claim that the standard approach performed better on average. Figure 3.13 supports this impression by showing the boxplots for the two approaches with a slightly lower median value for the Transfer approach as compared to the standard approach.

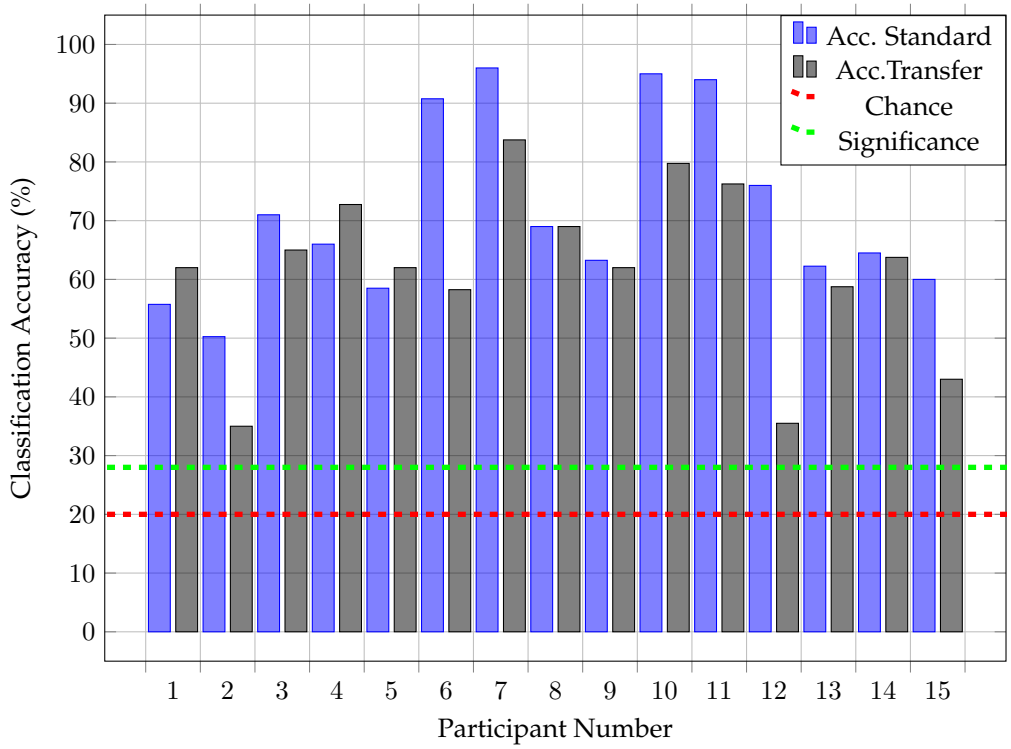

Figure 3.12: Histogram of the individual classification accuracies for each participant in blue for the standard training and in grey for the transfer learning approach. The green dashed line represents the significance threshold and the red dashed line the theoretic chance level.

In order to find the appropriate statistical test for verification of the results, we first performed a Shapiro-Wilk test which resulted in a p-value of 0.05309 for the standard and a p-value of 0.2137 for the transfer approach. These values did not show evidence of non-normality for the transfer but due to the relatively low value for the standard accuracies and after visual inspection of the histograms of the data we, decided to use a non-parametric test. We chose the Wilcoxon test implemented as signed-rank test because the two distributions contained accuracies from the same participants therefore being dependent.
The alpha value was chosen as 0.025 after applying a Bonferroni correction because we used the same datasets in two different tests, in order to answer our two research questions (1) if we can train an imagined speech classifier for a set of words on the EEG signals recorded while speaking those words out loud and (2) if this new training method can deliver comparable results to a standard training approach for imagined speech.

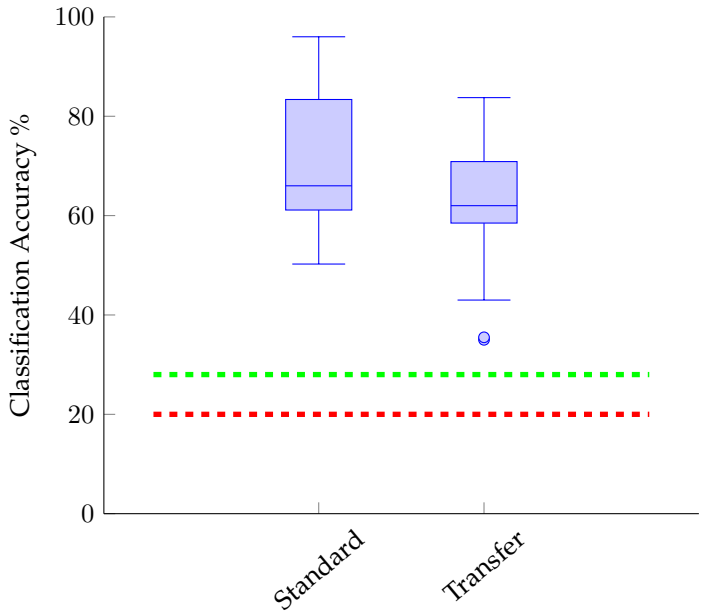

Figure 3.13: Boxplots of the classification accuracies of the standard and transfer approach of the individual configurations. The red dashed line represents the chance level and the green dashed line the significance threshold.

Research Question 1
For research question one we performed a Wilcoxon test to evaluate the accuracies of the transfer learning approach in comparison to the accuracies produced by a classifier performing on pure chance. This chance level accuracy was achieved by forwarding the classifier a sine wave signal for training which did not contain any valuable information and have it classifying on the test set. This approach resulted in a classification accuracy of 20.02% equaling almost exactly the theoretical chance level of 20%. Those two distributions, the actual transfer learning classification accuracies and the chance level accuracies, were used to perform the Wilcoxon test. The null hypothesis H0 was chosen as:

H0: There is no difference between the classification accuracies delivered by the transfer learning approach and a random classifier.

and the alternative hypothesis H1 to be:

H1: The transfer learning approach delivers better classification accuracies than a random classifier.

The test resulted in a p-value of 3.052e-05 meaning that the null hypothesis could be rejected and that our transfer approach was performing significantly better than a random classifier. Research question one could therefore be answered with yes, with our setup it was possible to train a classifier on EEG data recorded while overtly speaking and apply it on data recorded during silently speaking the same words.

Research Question 2
The second research question aimed at comparing the performance of the developed transfer training approach to a standard approach for imagined speech training. Therefore, four-fold cross validation was applied to split the imagined speech EEG data into train-test-pairs, and the test data was once used for a classifier trained on imagined speech and once for a classifier trained on the overt speech data. The resulting average classification results as shown in the tables 3.5 and 3.6 where used to perform a one-sided Wilcoxon signed-rank test to see whether the standard approach delivered significantly better results compared to our newly developed transfer approach. The test was conducted under the same assumptions as for research question one with an alpha value of 0.025 after Bonferroni correction.The null hypothesis was formulated as follows:

H0: There is no difference between the classification accuracies delivered by the transfer learning approach and the standard learning approach.

The alternative hypothesis H1 was formulated to be:

H1: The standard learning approach delivers better classification accuracies than the transfer learning approach.

The test resulted in a p-value of $0.01288 < 0.025$ meaning that the null hypothesis could be rejected and that the standard approach delivered significantly higher classification results as compared to the transfer approach. Thus, although showing classification accuracies throughout above the significance threshold, close median values and similar boxplots to the standard approach as illustrated in figure 3.13, the second research question had to be answered with no, the transfer training approach did not give us results comparable to the standard training approach.

Discussion

The presented results clearly indicate the successful transfer of the classifier trained on spoken to imagined speech. The classification accuracies for all subjects for this approach are significantly above chance level as confirmed by the results of the Wilcoxon test. Furthermore, the classifier exceeded the significance threshold for all of the participants as shown in figure 3.12. This means that we successfully trained an imagined speech classifier on EEG data recorded on spoken speech for the first time with a best individual classification accuracy of 83.75% and an average classification accuracies of 61.78% and 53.37% for individual and common configuration respectively. These results are astonishing given the fact that they were achieved in a 5-class classification problem. The same holds for the standard approach in which we trained and tested on imagined speech data and achieved an accuracy of 71.48% and 69.10% for unique and common configuration respectively. The single best accuracy even ended up being 96% for subject 7.
Comparing the unique versus the common configuration we can say for the standard approach, that there was not much of a difference between the two setups. The two calculated accuracies differed by only 2%. For the transfer approach however, we see

8% worse results for the common configuration. A possible explanation for this larger number in the transfer case might be the higher variance in the individual configurations. In the standard approach the second filtering method provided better results for most of the subjects, where for the transfer approach, more than a third of the best individual results were achieved with filter method 1. As for the common analysis setup the most occurring method was chosen, this might have had an impact on the overall classification accuracy, as for one third of the participants, this was not the preferred method of filtration. Concerning the CSP components for both approaches there was no clear evidence for one best performing number of components. While for the standard approach one could interpret a tendency towards higher number of components with 11 as the most occurring, the transfer approach seemed to prefer a lower number of components with 6 chosen for the common setup. If we calculate the average number of components over all participants we also receive a higher value for the standard approach with 9.4 as compared to the transfer approach with 6.34. Nevertheless, the best result for the transfer condition was still achieved with 11 components which might lead to the conclusion that the number of components has less influence and can be chosen more flexible as compared to the filtering method.
A limitation of the CSP that needs to be mentioned at this point are problems concerning the applicability of this method in a real-world scenario. The method presented in this work splits the data into separate train and test set and evaluates the classifiers on unseen data, however the features of the CSP are extracted in one step for both datasets. Still split up in train and test set but for the group of test data together in one process, which requires the data to be present and stored at this part of the processing pipeline. In a real world scenario one would rather have single incoming pieces of EEG data than a set of data on which the CSP could be performed. One possible solution to this problem could be, to extract the spatial filters created by the CSP on the training data, and use them to extract features on incoming single data snippets, thereby evaluating the systems performance under simulated real-world online classification conditions.
The classification method itself shows a clear picture with a preference to the RF classifier in both approaches. Another interesting fact is, that all the selections using GB performed better with the first filtration method. Given the small number of times GB was chosen overall, however, this can not clearly be concluded as the best combination for the GB classifier. A clear preference of a subject between the two approaches could also not be verified. There is no clear tendency from one subject to a certain classifier or filtering method between standard and transfer condition. This lets us conclude once more, that the signal analysis setup needs to be tailored to the individual and the use case.
What can be observed however is, that subject 7 scored the highest values for both, standard and transfer approach as illustrated in figures 3.10 and 3.11 showing the confusion matrices. More interesting, there appears to be a correlation between good performance in the standard and a good performance in the transfer approach. Subjects 6,7, 10 and 11 performed best in the standard condition with classification accuracies above 90%. Except from subject 6, the other 3 participants also performed best in the transfer condition. This observation was verified by calculating the Pearsons correlation coefficient resulting in a value of 0.58, showing a moderate positive correlation between the two conditions. This would again strengthen the hypothesis, that our transfer worked, and that well distinguishable imagined speech data can be classified by training on overt speech data.
The comparison of the two approaches concerning classification performance did not provide a clear picture. Although showing similar median values and boxplots (see figure 3.13) for classification accuracy with values throughout significantly better than chance level, the Wilcoxon test revealed a significantly better performance of the standard approach as compared to the newly developed transfer approach. Given the fact that

training by speaking provides various advantages in comparison to standard approaches, as for example being better controllable, more engaging and productive as the classifier can be trained during interaction, we would claim that the slightly worse classification accuracy can be considered a trade-off, which is compensated by the benefits of the transfer approach. Although the second research question has to be answered with no, the transfer approach did not deliver comparable results to the standard approach, we still want to highlight the average classification accuracy of 61.78% achieved with this newly developed method being significantly larger than the theoretic chance level of 20% for a 5-class classification problem as well as the significance threshold of 28%.
However, our results are preliminary and require further investigation in several directions. First of all, our preprocessing included only basic filtering methods, in order to make our approach applicable in online classification scenarios. Although we achieved good values for the transfer approach, we can imagine, that more sophisticated artifact removal methods, especially for the overt speech data, could further improve the performance of the system. The CSP feature extraction method in the current implementation does on the other hand not allow for an online classification and would need adjustment for that purpose. Our developed solution therefore acts as a proof of concept which needs to show feasibility in an online setting with a larger group of participants. In addition to being tested with a larger group of participants, increasing the number of words and the effect on classification accuracy should be researched in the future in order to investigate to which extent the approach is still feasible with a larger vocabulary. Lastly, with larger datasets collected from an increased number of participants and words, the whole setup could be tested using neural networks for classification, which might improve the performance of the system.
The informal interviews at the end of each session with the participants showed, that they liked the training scenario and the way of interacting with the system, however, after a few trails they felt bored and unfocused. This is probably due to fact that the study took on average 2 hours and the repetition involved only 5 words in a rather monotonous scenario. Based on our own experience and compared to the current state-of-the-art in BCI training we would still consider it a step towards more entertaining training scenarios. Maybe a change to 3D or virtual environments could help keeping the participants more engaged. In any case we will keep in mind to design the tasks in upcoming studies less monotonous with more variety in interaction with the system.

3.2.3 Conclusion

In this study we presented a new method for training imagined speech classifiers based on EEG data recorded during overtly speaking words. We designed a game-like scenario in which the participants had to control a robot on a computer screen through a factory by overtly and silently speaking 5 command words. The game was designed to leave the participants the freedom to interact in their own pace and they were informed, that independent of their input, the robot would perform the correct action in any case. During this interaction we recorded the electrical activity of the brain with an EEG headset and labeled the EEG data according to the overt and imagined output of the participant based on the state of the game. The recorded data was used to train and evaluate an imagined speech classifier under two conditions, first, trained on imagined speech data in a standard approach and second trained on the EEG data recorded during the spoken speech trails in a transfer approach. The evaluation was done on a separated imagined speech test set. Classification accuracies were calculated for both approaches with individual configurations including several different methods concerning signal

analysis, namely 2 filter methods, 2 classifiers and 9 different numbers of components for the CSP feature extraction. The single best configuration among all subjects was calculated as well. For the unique configuration the transfer training approach achieved an average classification accuracy of 61.78% compared to 71.48% for the standard approach. Remarkable considering the theoretic chance level of 20% and the significance threshold of 28%. This difference in average accuracies of 10% was also represented in the results of our Wilcoxon signed-rank test which showed, that the standard approach performed significantly better than the transfer approach. We can therefore not conclude, that the transfer approach delivered comparable results, however, results for all participants significantly above the theoretic chance level of 20% and even above the significance threshold of 28%, which lets us conclude, that the transfer worked. This is, to the best of our knowledge, the first time, that a classifier has been trained on overtly spoken speech and applied on an imagined speech dataset successfully. These promising results have the potential to tremendously improve training procedure for Speech Imagery BCIs in the future by providing a more engaging, controllable and productive training of the system during actual interaction by spoken speech.

As these results are preliminary, there is future work that needs to be addressed. First, we would like to explore the possibility of cross subject evaluation, meaning to train a classifier on the spoken data of all participants. If successful, this approach would allow to use general models as a starting point during training or in the best case even result in a baseline classifier which is precise enough to be applied on a variety of participants.

Second, we would like to extend our training scenarios into a more interactive environment. In the current study the interaction took place in front of a screen and a rather controlled scenario. In the future, such applications need to prove their applicability under real world conditions.

Third, not only a different environment but also different days on which the participants perform the training might be taken into consideration, as brain activity is know to vary from day to day, depending on the mental condition. These multi-day experiments are another domain in which the transfer learning approach on spoken speech needs to prove its applicability in comparison to a standard silent speech training scenario.

Lastly, the developed solution needs to be evaluated in an online classification scenario with adapted feature extraction methods and a larger group of participants which might even allow for Deep Learning methods to be applied on the data.

3.3 Summary

In this chapter we presented two novel methods for training Speech Imagery BCIs using transfer training approaches. One based on recording EEG data during reading and the other based on recording EEG data during speaking. In the following we will summarize the overall results of the conducted studies and discuss them once more on a higher level in the context of the topic of this thesis. We will continue with explicitly highlighting the contributions of those studies to the field and close the chapter by mentioning limitations of the here presented approaches.

3.3.1 Overall Results and Discussion

Overall, we can say that based on our study results, the transfer between different paradigms of EEG activity for training and classifying imagined speech appears to be possible.
The transfer from EEG activity recorded during speaking certain words to an imagined speech dataset including silent repetitions of the same words in section 3.2, did not only show classification accuracies significantly above chance level but also above the calculated significance threshold for all participants of our study. The results confirmed the similarity in brain patterns of the two paradigms as reported in [124, 221, 146] and lay the foundation for more efficient and engaging Speech Imagery training scenarios in the future. Being able to train an imagined speech classifier while interacting with the system via overt speech provides more productive training scenarios as interaction is already possible during the training phase. The different interaction paradigms can be dynamically switched to best suit the given situation, after a first interaction phase with overt speech. In an industrial environment for example the user could interact with imagined speech if there is too much environmental noise affecting audio based speech recognition. As soon as the noise reduces, the classifier could switch back to detect overtly spoken speech based on audio data.
Our work showed that this transfer between overt and imagined speech is possible for the first time, paving the way for more flexible, productive and entertaining Speech Imagery training in the future.
The reading task study on the other hand did not show the same promising results. We were not able to achieve a successful transfer from the EEG activity recorded while reading to silently speaking the same words. The classification results did not exceed the significance threshold for the transfer, however, we could show accuracies significantly above chance level for classifying on the reading data. These results indicate, that the brain activity during reading certain words in a text appears to produce distinct patterns which can be classified with machine learning algorithms. Transferring this classifier to imagined speech data was not possible and we can imagine several reasons for those results.
First of all, reading and especially the movement of the eyes induces artifacts in the EEG data, which requires more precise filtering techniques. We see the chance to improve on our current methods by explicitly measuring the eye movement with an additional electrode placed next to the eyes or include the eye tracking data into the filtering process. Sophisticated models of eye movement combined with electrical field modeling might be a possible solution to receive less corrupted signals.
Second, the epoching process plays an important role and could be improved in a way to tailor the extracted time intervals from the continuous data to the individual fixation time, rather than choosing a common best time window for all words. Our reasons to use

a common time window are explained in detail in section 3.1.1 and included issues with balanced dataset assembly and possible label leakage during feature extraction based on different epoch length. It is not common practice in imagined speech classification to use epochs of different length for different words because the output during imagined speech production is not detectable in terms of start and end of speech production, as it happens in silence. During reading with eye tracking we can do so and would suggest for future work to emphasize this approach by cutting the EEG data during reading based on the eye tracking input, and focus on features that are not affected by signal length. This would then allow to also cut the imagined speech data more precisely, based on the average reading time of the word for example.
Most importantly however, we see improvement in the design of the study setups of the two paradigms, natural reading and imagined speech.
In this study we designed a natural reading task in which the participants had to read a text in their own speed presented in several rows on a screen. The silent repetition however was presented in a standard imagined speech training approach, similar to the speaking scenario, with a fixation cross in the middle and no other context than the target word presented prior to the cross.
The patterns in brain activity during reading words in a text and a specific context are most likely too different from a single silent word repetition. In the case of the transfer from overt to imagined speech, we used the exact same task for the two types of speech. The participants of the study controlled a robot through a maze with 5 different words and indicated the intention to interact via a spacebar press. Imagined and overt repetition therefore always happened under the same conditions, with a blank screen and a fixation cross and the exact same purpose, to control the robot through the maze. This might have improved results due to the data being better transferable between the two paradigms, as compared to the reading task. In the future, the text design of the reading task should be adjusted to isolate single words in parts of the text or rather switch to a rapid serial visual presentation paradigm which shows each of the words sequentially in the center of the screen for a short time window. The word presentation duration of the target words could be extended to emphasize those words and make them better transferable to the Speech Imagery scenario afterwards. This procedure would contradict of course the goal of producing engaging and comfortable ways of training Speech Imagery BCIs, where a natural reading task is most likely considered more comfortable as the rapid visual serial presentation of single words. However, we see a clear chance of improving transfer classification accuracies with this method, as it would make the two paradigms more similar.
Concerning adjustments on the silent repetitions, we can imagine to switch to repetitions of short command sentences, which would be easier to embed into a text and are more similar to the reading task. This adjustment might have a positive impact on classification accuracies of the transfer approach as well.
Concluding on our study about the transfer from reading to silently speaking we could not show the feasibility with our current setup but still foresee it as a promising method to train Speech Imagery BCIs in the future and a good chance to improve on the results given the previously mentioned adjustments on the setup. The potential of this method, beside providing more engaging and comfortable training tasks, lies in the possibility to provide knowledge to the user inside the text about the later interaction scenario. While the system tries to understand the users brain patterns during silently reading, the user can learn how to control the system in the later interaction simultaneously, hence establishing a bidirectional information flow.

Furthermore this text can include any possible topic as long as it follows certain design parameters as mentioned in 3.1.1 which provides endless possibilities for the later field of application, making it a promising approach for flexible and engaging Speech Imagery training scenarios in the future.

Summing up our results we wanted to answer the following research questions within this chapter:

> **RQ 1.1** Can EEG activity recorded while reading certain words be used to train a classifier to detect those words during imagined speech?
>
> **RQ 1.2** Can EEG activity recorded while speaking certain words be used to train a classifier to detect those words during imagined speech?

Concerning **RQ 1.1** we were not able to successfully transfer the classifier trained on the data of the natural reading task to the imagined speech data with our study setup. We can see possible improvements in our setup as for example to design the reading task more similar to the imagined speech task, with rapid serial visual presentation instead of natural reading, or to adjust filtering and epoching methods. Nevertheless we managed to show within our study that it was possible to classify words read during a natural reading task based on EEG activity alone for the first time which leads us to the conclusion that the cognitive patterns during natural reading can be classified from EEG data. In how far those patterns have the potential to provide a basis for the classification of silently spoken words has to be evaluated in further research with improved signal processing methods and adjusted study setups.
Concerning **RQ 1.2** we could clearly show that it was possible to train a classifier on spoken speech and let it successfully predict the words in an imagined speech scenario. The classification accuracies for this transfer clearly exceeded chance level in the given study setup and significance threshold for all participants which will provide the foundation for more engaging and productive Speech Imagery training based on spoken speech in the future.

Further conclusions and recommendations are summarized in the following as contributions and limitations of the findings in this chapter on improving training procedures in EEG-based Speech Imagery BCIs.

3.3.2 Contributions

Within this chapter we have introduced two novel concepts for training EEG-based Speech Imagery BCIs, which will not only make training procedures more engaging and comfortable for the user but also more productive. The transfer of EEG-activity during spoken speech allows a direct interaction with the system during training procedure which will provide productive, engaging and interactive training of SI-BCIs.
In the reading task we enable the user to train the SI-BCI classifier during naturally reading a text and at the same time provide knowledge on the system to interact with later on by reading the instruction manual. Although our first attempt to transfer the classifier from EEG data recorded during reading to imagined speech data was not successful, our study provided valuable insights on pitfalls of this technique. In addition with our detailed recommendations for adjustments of the study setup, we have paved the way for further research on this technique in the future.

Both of the developed concepts will enable a more comfortable way of training Speech Imagery BCIs for possible future real-world application, but will furthermore facilitate research study setups and make them less mentally and physically exhausting compared to the state-of-the-art approaches.
Finally, we have investigated brain activity during different paradigms related to speech production and could confirm the findings of other existing research in this field with our EEG-based classification approaches. This will allow to build up on our research and further expose relations between the concepts to extend our understanding of neural processes during speech production in the brain.

3.3.3 Limitations

The results presented in this chapter about new training procedures for Speech Imagery BCIs are promising, there are however several limitations to these preliminary findings which need to be addressed in the future.

One limitation which needs to be mentioned is the applicability in a real world scenario and online classification. The presented CSP feature extraction method in the case of the transfer from spoken to imagined speech, is applied on separate train and test data, evaluating the classifiers on unseen data. The CSP features however are extracted in one step together, also for the test set. In a real world scenario data would rather be provided in single incoming pieces than a continuous stream on which the CSP would then be performed. A possible adaption of our method to comply with those requirements would be, to extract the spatial filters created by the CSP on the training data and use them for calculating features on the single incoming data snippets. This would allow to test the performance of the classifier under online conditions.
Additionally, our experiments were conducted in rather controlled scenarios inside the lab with environmental influences reduced to a minimum. The developed methods still have to prove their feasibility in environments of their targeted application, which will definitely rise further challenges in terms of filtering data from artifacts, either produced from sources in the surrounding, or for example movement of the user in the environment.
Furthermore, the data was recorded for each subject in only one session on one day. In how far the recorded training data could be used to classify on data recorded on different days is questionable and one of the biggest remaining challenges in BCI research in general which should be addressed in the future.
Extending the dataset with recordings from several days and maybe also more participants would finally lead to a bigger data corpus which would allow for an analysis with deep learning methods. The current implementations are based on established machine learning techniques, however, a deep learning approach might provide new insights, given a sufficient amount of data, and improve classification accuracies even further.

The problem of usually low classification accuracies is common in imagined speech BCIs and known to decrease even further with the increase of words to distinguished simultaneously. Within the next chapter we will try to overcome this issue by including semantic category detection into the process of imagined speech classification.

Chapter 4
Semantic Category Detection in Speech Imagery BCIs

EEG-based Speech Imagery BCIs have proven to provide classification accuracies significantly above chance level in a variety of application scenarios (see section 2.5). However, the state-of-the art manages to distinguish not more than 4-5 words [69, 1] with satisfying classification accuracy, which is far from a sufficient number necessary for communication (see section 1.3). A possible solution to this problem is a 2-stage or -level approach which starts with classifying the semantic category of a word before proceeding to classifying the word itself.
As mentioned in 2.7.2, detecting semantic processes from brain activity has been shown to be feasible on a larger scale in fMRI and EEG studies. Most of the work based on EEG data is however concerned with distinguishing between 2 categories, living, e.g. mammals or non-living, e.g. tools [197, 153, 206]. Related work in SI-BCIs is limited to few studies where [112] used a confined set of 2 categories as well and the results of [116] are questionable due to their study setup as pointed out in [130]. The lack of existing work on the topic lead us to a three-fold study setup for our research on semantic category detection in Speech Imagery BCIs:

1. In a first step, we wanted to extend the number of distinguishable semantic categories in EEG studies and explore the possibilities of classifying multiple semantic categories from EEG activity in a standardized procedure, namely decision tasks, to show feasibility of this approach for EEG-based SI-BCIs. (Section 4.1)

2. In a second step we transferred the approach of classifying semantic categories from EEG activity to the actual application scenario, silent repetitions of words. (section 4.2)

3. In the last step we consequently applied the concept of classifying semantic categories prior to word classification on two Speech Imagery datasets containing words from different semantic categories. (Section 4.3)

With this study setup we aim at answering the following research questions:

RQ 2.1 Can semantic categories be classified from EEG activity during imagined speech production?

RQ 2.2 Can semantic classification increase classification accuracies in EEG-based imagined speech BCIs?

Section 4.1 is based on already published work in [179] while section 4.2 is based on already published work in [180].

4.1 Semantic Category Detection in Object-based Decision Tasks

In order to leverage semantic category classification based on EEG data beyond the state-of-the-art and apply the developed methods later on in imagined speech tasks, we started with a standard procedure in neuroscientific research to evoke conscious semantic processing, a decision tasks. In a decision task a sequence of items belonging to one of several categories is shown to the participant, prompted by a question, if a certain attribute can be assigned to the presented item [53]. This process of evaluating one individual item from the current sequence is in the following referred to as a object-based decision task, as in our study several objects in the form of words were presented to the participants.
The aim of this study therefore was to classify semantic categories from EEG activity in those tasks to show the feasibility for a future integration of semantic category classification into speech imagery BCIs. Based on the study of Huth et al. [88], we selected 5 semantic categories for classification during object-based decision tasks, which required the participant to answer simple yes/no-questions regarding a word to ensure conscious semantic processing [53]. We investigated different feature extraction and machine learning techniques and methods as well as different epoching intervals during data pre-processing in order to evaluate the best possible setup for semantic category detection from EEG activity.

In detail we targeted the following three research goals:

- To decode semantic processes in EEG-activity of a single subject for the first time in a 5-class classification task.
- To find similar patterns amongst different subjects during semantic processing, as shown in [88] with fMRI, based on EEG-activity.
- To provide recommendations on the best possible setup for semantic category detection in future EEG-based Semantic Silent Speech BCIs.

4.1.1 Methodology

In order to prove feasibility of our targeted approach to classify 5 different semantic categories from EEG activity for the first time, greatest care had to be taken concerning the methods of our study, starting with the study setup.

Study Setup

The tasks for the subject were designed similar to a study by [53], where transcranial magnetic stimulation of the left inferior prefrontal cortex was used to investigate involvement in semantic and phonological processing. A sequence of items belonging to one of five semantic categories was shown, prompted by a question, if a certain attribute can be assigned to the presented item, and the task was to choose between an affirmative ("Yes") and a negative ("No") answer for each individual item after it was shown on the screen (see figure 4.1) via a button press. This process of evaluating one individual item or object from the current sequence is referred to as a object-based decision task.
5 semantic categories were selected based on the findings of [88] with the prerequisite to be best distinguishable concerning their position on the cortex and to fit into a Smarthome context of a possible BCI application. Living, non-living, numbers, locations and action verbs were chosen as categories, each category containing 10 words to be processed in the decision task, Figure 4.3 gives a detailed overview.
In order to keep the experiment as short as possible and therefore more comfortable for the participants, we decided on short-block stimulus presentation [169]. As the task involved answering a question to a certain word, we wanted to avoid on the one hand to present the question prior to each task, which would make the experiment time consuming, and on the other hand avoid having the subjects to remember the questions over the whole procedure. Therefor the words were presented in short blocks of 10, according to their category with the question presented in the beginning. The words inside the blocks were randomly shuffled, each block was presented randomly once per trail and the experiment consisted of 6 trails, resulting in a total number of 300 samples per participant. Figure 4.2 provides an illustration of the short-block stimulus presentation. With this type of presentation we wanted to facilitate study setup without risking unwanted side effects during classification, as reported in [130] for plain block-wise stimulus presentation.

Subjects

The study was conducted with 20 healthy subjects (age 21–29). All subjects were native German speakers who were right-handed and had a normal or corrected-to-normal vision. Subjects where chosen to have the same mother tongue, to prevent confounding neurolinguistic effects on the EEG due to foreign language use [71], and prevent multilingual requirements to the setup and subject population [84]. The subjects were asked not to consume caffeinated substances at least three hours before the starting of data collection as they have a proven potential to affect the EEG recordings [185]. Each subject was introduced to the task, and informed consent was obtained from all subjects for scientific use of the recorded data. The data was acquired in a dim light room with minimized distractions like external sound, mobile devices and others, where the voluntary participants were asked to sit in a comfortable chair to prevent unnecessary muscle movements to reduce noise and artifacts in the EEG, which could emerge from mental stress, unrelated sensory input, physiological motor activity and electrical interference.

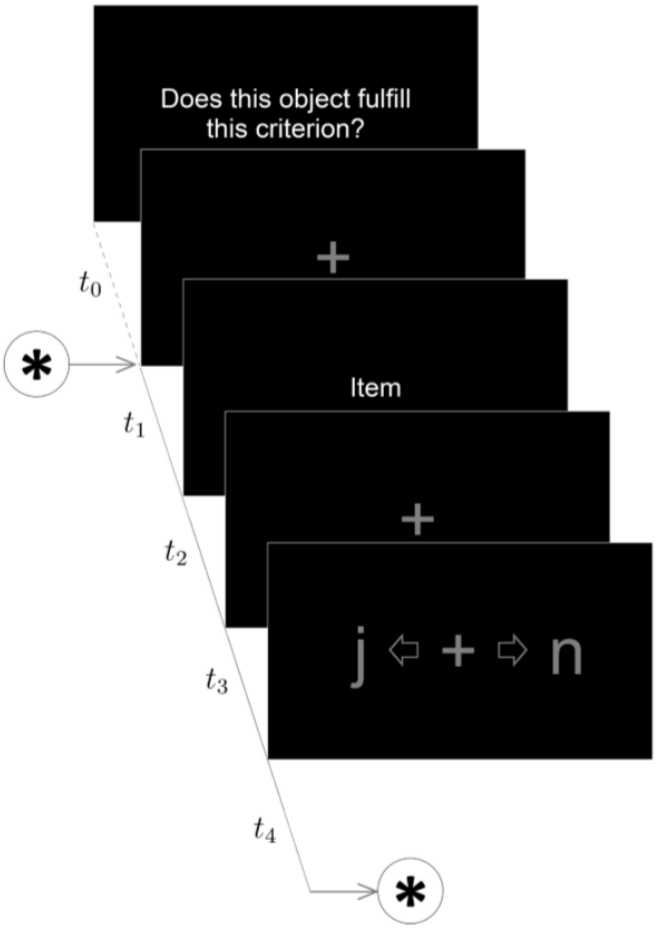

Figure 4.1: Procedure of visual output on the screen during the decision tasks with exemplary question and item, as done for 5 semantic categories in each trial($t_0 = \infty$, $t_1 = 1.0s$, $t_2 = 0.1s$, $t_3 = (2. \pm 0.2)s$ and $t_4 \leq 5.0s$).

Two subjects were excluded because of poor electrode-to-skin contact later, leaving a total of 18 subjects (9 male, 9 female).

Recording

EEG signals were recorded using a wireless 32 channel electroencephalograph system namely g.Nautilus with g.Scarabeo electrodes (g.tec medical engineering GmbH, Austria). The sampling rate was set to 500 Hz. The 10-20 International System of electrode placement was used to locate the electrodes. This configuration is believed to cover the whole scalp resulting in the capturing of spatial information from the brain recordings effectively which provided the optimal setup for our study based on the findings of [88]. Figure 4.3 b) illustrates the positions of electrodes chosen for the study.

Table 4.1: Selected semantic categories, items and questions for the decision task, originally in German, translated to English.

Category	Question	Items
Locational	Do you usually find a tap at this location?	Kitchen, Bathroom, Cellar, Garden, Court, Bedroom, Living room, Staircase, Corridor, Attic
Actions	Do you associate this action with a movement in positive direction or positive change of state?	Throw, Open, Lift, Lower, Switch on, Switch off, Put, Place, Push, Pull
Living	Do you associate this living being with social properties?	Dog, Cat, Peacock, Lamb, Pigeon, Mother, Father, Grandmother, Grandfather, Doctor
Non-living	Do you associate this object with the living room?	Light, Shutters, Heater, Television, Telephone, Computer, Stove, Refrigerator, Washing machine, dryer
Numbers	Is this number larger than six?	One, Two, Three, Four, Five, Six, Seven, Eight, Nine, Ten, Eleven

Data analysis

Preprocessing After the EEG data acquisition, a basic preprocessing including filtering and referencing, was applied to remove artifacts. While recording the EEG activity, scenarios like poor electrode-to-skin contact, broken recording device, low signal quality, and others can hinder the quality of signals. Therefore, channels containing such issues were labeled bad and thus, excluded from further analysis. Moreover, data segments identified to contain strong artifacts throughout all channels are also considered as bad and removed from further analysis. The EEG was filtered with a second-order Butterworth notch filter with a lower cutoff frequency of 48Hz and an upper cutoff frequency of 52Hz, to remove power line noise. A fourth-order Butterworth was applied for high pass

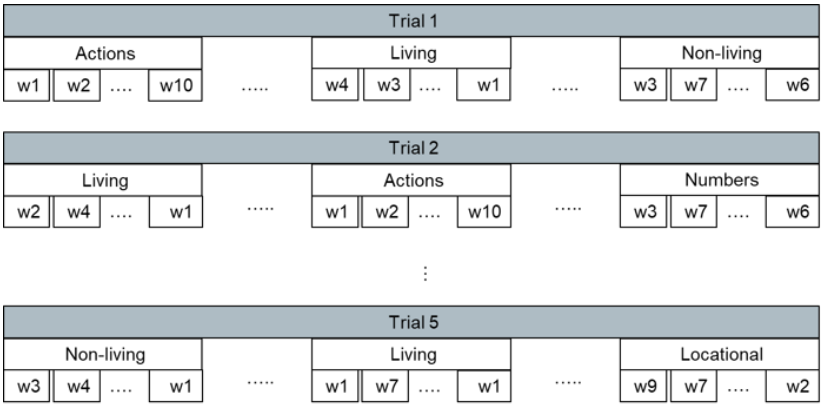

Figure 4.2: Illustration of the short-block presentation. Each trail included the 5 categories, consisting of 10 words each, where the category, as well as the words inside, were presented randomly shuffled.

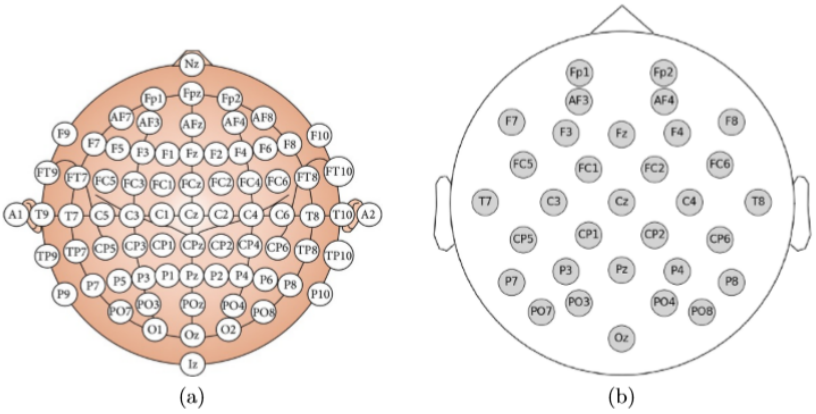

Figure 4.3: (a) Conventional electrode placement according to 10-20-system as presented in [96] (b) Subset of electrodes used in this study with the 32 available electrodes.

filtering at a cutoff frequency of 1Hz as recommended by [224], to remove low-frequency noise, but also to remove low-frequency shifts before applying the Independent Component Analysis (ICA). Lowpass filtering was done using a 17th order Chebyshev type 2 filter with a cutoff frequency of 200Hz preserving the frequency range up to high-gamma [46], and a minimum of 60dB attenuation in the stop band, to obtain adequate roll-off. All the electrodes were referenced to common average over all electrodes to achieve low signal-to-noise ratio [136], and the features being scattered across the whole scalp thus local changes of a feature are vulnerable to close references [127]. Further denoising of the EEG was done using a joined method of ICA and wavelet denoising, referred to as wavelet-enhanced independent component analysis (wICA) [7] to remove the artifacts.

Epoching Epoching in EEG refers to the procedure of extracting specific time-windows from the continuous EEG signal [60] for further analysis. Epochs were extracted in reference to the stimulus onset from the pre-processed data. To further analyze the impact of different epoching duration, three intervals were chosen, namely, $T_1 = 0.3s - 0.8s$,

$T_2 = 0.3s - 1.5s$ and $T_3 = 0s - 2s$. T_1 and T_2 were supposed to cover the point around 400 ms after stimulus onset, as studies on semantics assume this to be the earliest point information is consciously processed after passing the visual pathway and being directed to higher cortical structures for semantic processing [112]. The third interval was chosen as the entire period between stimulus onset and end of the task in order to cover as much information as possible. This setup was supposed to provide evidence on the best interval length of EEG signals for the recognition of semantics in envisioned speech.

Feature Extraction We investigated two feature extraction methods, first we assembled a feature vector based on the different state-of-the-art literature in imagined speech detection [183, 115, 235, 145]. The Thirteen features were chosen in time and frequency domain and included, Power Spectral Intensity and Relative Intensity Ratio (Alpha, Beta, Gamma, Delta, Theta), Petrosian and Higuchi Fractal Dimension, Hjorth parameters, Spectral Entropy, Skewness, Fisher Information, Approximate Entropy, Detrended Fluctuation Analysis and Hurst Exponent. The features were extracted with the open-source python module PyEEG [14].
As a second feature extraction method we chose Common Spatial Pattern (CSP) which is a frequently used technique in BCI applications [109, 156]. The basic principle behind CSP is to apply a linear transformation to project the multi-channel EEG signal data to a lower-dimensional spatial subspace. The transformation results in the maximization of the variance of one class while minimizing the variance of other class at the same time [4]. Originally CSP algorithms are used to solve binary classification problems. In order to perform multi-class classification with CSP, we trained 5 different classifiers in a one-against-all approach for the 5 semantic categories. Therefor the input data was split in test and train set, both uniformly distributed, containing equal instances of all the five semantic categories. After training, we received 5 models M1, M2, M3, M4 and M5 which were used to make predictions on the previously created test set. Suppose, the predicted values are P1, P2, P3, P4 and P5 for each model then the maximum value is chosen from the five predicted values and the result evaluated. In this way, the theoretic chance level is 20% as the prediction is computed in a combined way by taking all the five models into consideration.

Classification Classification was done based on two strategies, cross-subject and within-subject. In the cross-subject analysis, the data from all subjects is taken together as a single input in order to find similar patterns amongst subjects while in within-subject analysis, the performance of the classifier is computed on the individual subject's data set. We used Support Vector Machine (SVM) and Random Forrest Classifiers (RF), because of their frequent use in EEG-based Speech Imagery BCIs [219, 206, 147], both evaluated using grid search and ten-fold cross-validation with a test train split of 0.1. For the cross-subject data set we further implemented a deep artificial neural network, the Multilayer Perceptron using the ADAM and SGD optimizer with mean-squared error as loss function. In this case grid search was performed to find the optimal number of hidden layers and chose the best learning rate in addition to the activation function. The within-subject condition did not provide a sufficient database for training neural networks and could therefor not be evaluated.

Performance metrics and statistical analysis The performance of the implemented classifiers for semantic category detection was assessed based on the classification accuracy in comparison to the respective significance threshold. All the experiments were multi-class classification problems with five labels resulting in a theoretic chance level of 20%. However, as explained in section 2.9, this threshold theoretically requires an infinite number of classifications. Combrisson and Jerbi proposed an adjustment of the chance level for machine learning in neuroscience [39] which we calculated based on

their work as explained in section 2.9. With the 5 categories and an alpha value of 0.05 we received a significance threshold of 25.33% for the testset including 30 samples after cross-validation.
The classification accuracy of our classifiers was calculated as the number of correct predictions divided by the total number of predictions and compared with the significance threshold to validate statistical significance of the results.

4.1.2 Results and Discussion

In the following we will present the results of the analysis on the recorded dataset performed as described in the last section. We will start with presenting the results and discuss them afterwards.

Results

Table 4.4 shows the results for the cross-subject dataset evaluated with the different classifiers (SVM, RF, MLP), feature extraction methods (CSP, FV) and concerning the different epoching intervals (T_1, T_2, T_3). For the CSP feature extraction condition, the classifiers did not manage to exceed chance level. In all other cases, the classifiers evened out around 40% which is significantly above chance level and RF classifier achieved the highest accuracy for all epoching intervals.

Table 4.2: Mean Accuracy for the cross-subject classification depending on the different epoching intervals (T_1, T_2, T_3), feature extraction methods (CSP RV) and classifiers (SVM, RF, MLP) used.

Classifier	T_1 **= 0.3s - 0.8s**	T_2 **= 0.3s - 1.5s**	T_3 **= 0s - 2s**
SVM-CSP	$20.09 \pm 0.36\%$	$20.52 \pm 0.80\%$	$18.17 \pm 0.76\%$
RF-CSP	$20.04 \pm 0.72\%$	$23.03 \pm 0.46\%$	$20.04 \pm 0.72\%$
SVM-FV	$40.00 \pm 2.46\%$	$40.29 \pm 0.67\%$	$38.55 \pm 0.79\%$
RF-FV	$41.97 \pm 1.95\%$	$39.98 \pm 2.58\%$	$40.97 \pm 2.70\%$
MLP-ADAM	$40.03 \pm 0.46\%$	$40.47 \pm 1.48\%$	$41.31 \pm 1.29\%$
MLP-SGD	$38.59 \pm 3.02\%$	$39.79 \pm 0.98\%$	$39.30 \pm 1.46\%$

Figures 4.4 – 4.7 show the results for the within-subject condition and the different classifiers used on the assembled feature vector and the CSP feature extraction methods. For within-subject analysis using the Support Vector Machine and the feature vector, figure 4.4 summarizes the results for all three intervals. Overall, the highest accuracy achieved is 51.72% by subject 13 in the epoching duration of T_3. In the interval, T_2, the highest accuracy of 43.34% is achieved by subjects 2, 5, and 12. In the interval, T_1, the highest accuracy of 46.67% is achieved by subject 17. However, subject 17 is below chance level (blue horizontal line) in the interval of T_3 and subject 2 below significance threshold (green dashed line) for interval T_3. The mean accuracies were reported to be 35.78%, 35.41% and 34.42% in the duration of T_1, T_2 and T_3 respectively.

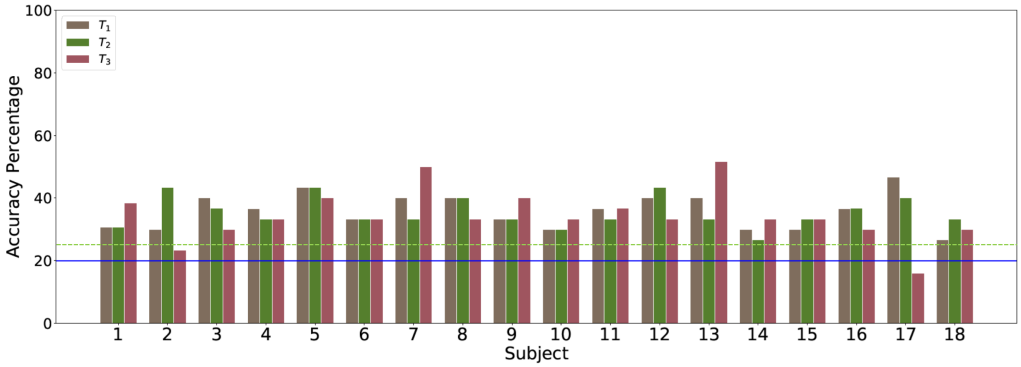

Figure 4.4: SVM within-subject accuracy using the assembled feature vector for three different epoching intervals, $T_1 = 0.3s - 0.8s$, $T_2 = 0.3s - 1.5s$, $T_3 = 0s - 2s$. The solid blue line indicates chance level, the green dashed line the significance threshold.

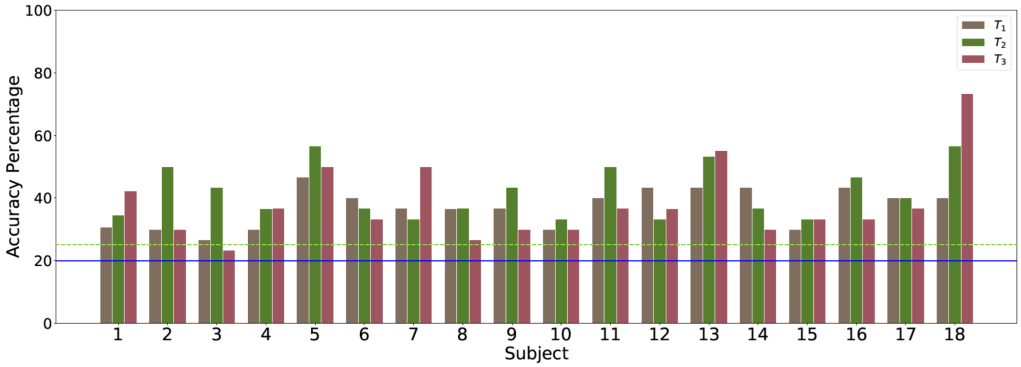

Figure 4.5: RF within-subject accuracy using the assembled feature vector for three different epoching intervals, $T_1 = 0.3s - 0.8s$, $T_2 = 0.3s - 1.5s$, $T_3 = 0s - 2s$. The solid blue line indicates chance level, the green dashed line the significance threshold.

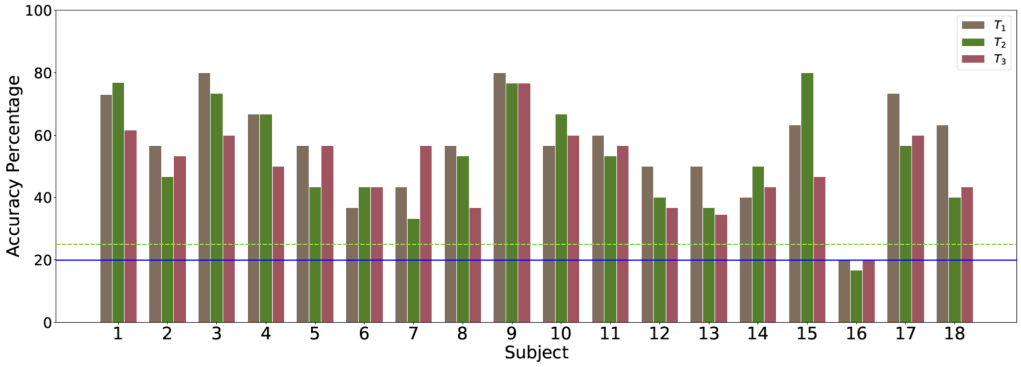

Figure 4.6: SVM within-subject accuracy using CSP for three different epoching intervals, $T_1 = 0.3s - 0.8s$, $T_2 = 0.3s - 1.5s$, $T_3 = 0s - 2s$. The solid blue line indicates chance level, the green dashed line the significance threshold.

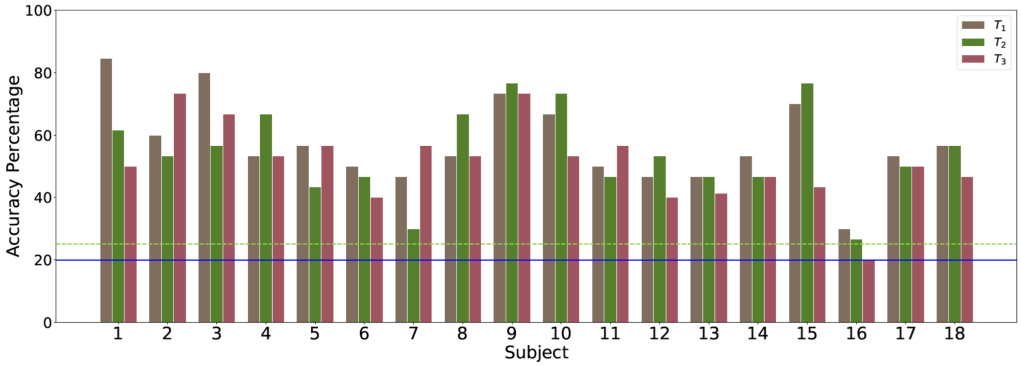

Figure 4.7: RF within-subject accuracy using CSP for three different epoching intervals, $T_1 = 0.3s - 0.8s$, $T_2 = 0.3s - 1.5s$, $T_3 = 0s - 2s$. The solid blue line indicates chance level, the green dashed line the significance threshold.

Figure 4.5 shows the accuracy versus subject plot for all three epoching intervals using the Random Forest classifier and the feature vector. The highest accuracy of 73.34% is achieved by subject number 18 in the epoching interval of T_3. In the interval, T_2, subject 5 and 18 achieved the highest accuracy of 56.67% while in T_1 subject 5 achieved the highest accuracy of 46.67%. The mean accuracies are 37.08, 41.92% and 38.19% for interval T_1, T_2 and T_3 respectively. All the subjects managed to score above chance level, i.e., 20% represented by the horizontal blue line and only subject 3 stayed below the significance threshold as represented by the green dashed line in interval T_3.
Figure 4.6 presents the results obtained with the implemented CSP algorithm using a Support Vector Machine. The overall highest accuracy achieved is 80% by subjects 3,9 and 15 for intervals T_1 and T_2. In the interval, T_3, the highest accuracy of 76.67% is achieved by subject 9. The mean accuracies were reported to be 52.02%, 57.98% and 49.78% in the duration of T_1, T_2 and T_3 respectively. The worst performing participant is subject 16 for all the epoching intervals with results at chance level.
Figure 4.7 illustrates within-subject accuracy for all three epoching durations using random forest. Overall, the highest accuracy achieved is 84.61% by subject number 1 in the interval of T_1. In the interval T_2, best performing subjects are subject number 9 and 15 with an accuracy of 76.67%. In the interval T_3, best performing subjects are subject number 2 and 9 with an accuracy of 73.34%. The worst performing participant is again subject 16 for all the epoching intervals with T_1 and T_2 slightly above significance threshold and T_3 at chance level. The mean accuracies reported are 57.29%, 54.34% and 51.19% in interval T_1, T_2 and T_3 respectively.

Discussion

One goal of this study was to find similar patterns amongst different subjects during semantic processing as shown in [88], with fMRI, for the first time with EEG. For the CSP feature extraction our implementation did not manage to find patterns of similar activity concerning the data of all subjects in neither of the intervals. For our feature vector however SVM, RF as well as the neural network implementation managed to exceed chance level and evened out around 40% for all given conditions. Further research is needed to verify those results but it seems possible to train a classifier with certain features on data of several subjects and use it to classify for the individual. As the

Multilayer Perceptron delivered promising results, we plan to extend our studies and gather further data, which is needed for the training of artificial neural networks.
Another goal of the study was to decode semantic processes in EEG-activity of a single subject for the first time in a 5-class classification task. This goal was clearly achieved, as for all conditions regarding different classifiers and feature extraction methods, all subjects managed to exceed chance level and the significance threshold, with the exception of subject 16 for CSP feature extraction. The highest accuracy was achieved by subject 1 with CSP and RF classifier in the interval of T_1 with 84.61%.
On average, the CSP feature extraction achieved better results than the assembled feature vector, for both classifiers. Concerning the epoching intervals, there was no significant best, but the CSP feature extraction perform better on the shorter intervals T_1 and T_2 starting after 300ms. This might be the case due to event related potentials which occur within the first 300ms after the presentation of a stimulus as part of unconscious processing of visual or auditory input. Those potentials have characteristic forms and locations in which they occur and might have complicated classification due to similar patterns in the beginning of the signal for this interval and certain subjects.
This brings us to goal number three of the study, to make an assumption on the best possible setup for semantic category detection for a possible future use in EEG-based Speech Imagery BCIs. Concerning our results, we claim, that there is no single best setup which could be applied generally amongst different individuals. Taking into account the pure numbers for average classification accuracy, CSP feautre extraction performed better than our feature vector and SVM and RF delivered equal good results in this case. Having a look at single subject performance however, each subject has its own preferences and conditions under which classification performed best. Subject 9 for example achieved pretty stable and good results for all epoching intervals with CSP around 80%, however feature vector classification did not exceed 50% and varied amongst different epoching intervals. For this subject the clear recommendation for a possible Semantic Silent Speech Interface would be the feature extraction with CSP and no specific epoching interval, whereas one would probably choose the smallest interval regarding performance issues during data processing. Subject 18 on the other hand, achieved the highest accuracy with the feature vector, RF and interval T_3 outperforming the rest of the participants within this condition. Subject 16 did not even manage to exceed chance level for the CSP feature extraction but reached nearly 50% accuracy for our assembled feature vector.

Summing up the results we can conclude that on our dataset:

- Distinguishing 5 semantic categories from EEG data is possible for the individual.
- CSP performed on average better than our assembled feature vector.
- In CSP, epoching intervals starting after 300ms should be chosen.
- For CSP feature extraction, SVM and RF perform equally well.
- If using our feature vector, RF performs better than SVM.
- Methods should be tailored to the individual.

4.1.3 Conclusion

In this study we explored the potential of decoding semantic categories from EEG activity for the use in Speech Imagery BCIs. Object-based decision tasks were used to evoke conscious semantic processing for 5 different semantic categories in the participants cerebral cortical structures. We implemented different feature extraction and classification methods to provide an estimation on the best possible setup for Semantic Silent Speech BCIs. Although we could not find one single best setup, all of our different classification methods exceeded chance level and significance threshold for training and testing on the data of the individual and even for a cross-subject condition. The best individual accuracy achieved was 84.61% for a CSP feature extraction method and RF classification illustrating the potential of this approach for a possible future use in Speech Imagery BCIs.
In the next section we will transfer the developed methods and findings into a speech imagery setup to proof the feasibility of the here presented passive approach in object-based decision tasks during the actual imagination of words from the 5 different categories.

4.2 Semantic Category Detection in Speech Imagery Tasks

The goal of this study was to classify semantic categories of imagined words from EEG activity and to compare the results to our previous study for classification of semantic categories in object based decision tasks presented in the previous section. This study therefore followed a similar setup and used the same data analysis methods in order to make the results of both studies comparable. The results will give insights in how far the concept of semantic category detection of imagined words can be included into a SI-BCI. The following section gives an overview over the material and methods used in this study.

4.2.1 Methodology

Although our motivation behind the study and the analysis of the data were based on similar approaches, our methodology differed in several points, starting with the study setup.

Study Setup

The objective of this study was to decode semantic categories of imagined words using EEG signals. Therefor the task included the imagination of words from different semantic categories. The overall experimental paradigm for recording the speech imagery part is depicted in Figure 4.8. Each trial starts with a fixation cross shown in the time interval of t_1. Subsequently, the word was presented on the computer screen, which was later followed by the blinking of the fixation cross once. The participant was instructed to imagine speaking the word after this blinking once by inner voice, i.e., saying the word in your mind without actually producing a sound or moving the facial muscles. A short break of 2s followed until the next trial.
The same 5 semantic categories were selected as during our last study [179], originally based on the findings of [88], namely living, non-living, numbers, locations and action verbs. Each category contained 10 words to be processed, Table 4.3 gives a detailed overview.
In order to prevent classifying arbitrary brain states based on block-level temporal correlations rather than stimulus-related activity[130], we chose the short-block presentation paradigm as described in [179] and illustrated in figure 4.2. In this procedure the words were presented in blocks of 10 according to their category. The words inside the blocks were randomly shuffled, each block was presented randomly once per trail and the experiment consisted of 4 trails, resulting in a total number of 200 samples per participant.

Subjects

The study was conducted with 20 healthy subjects (age 21–29). All subjects were native German speakers who were right-handed and had a normal or corrected-to-normal vision. Subjects where chosen to have the same mother tongue, to prevent confounding neurolinguistic effects on the EEG due to foreign language use [71], and prevent multilingual requirements to the setup and subject population [84]. The subjects were asked not to consume caffeinated substances at least three hours before the starting of data collection as they have a proven potential to affect the EEG recordings [185]. Each

subject was introduced to the task, and informed consent was obtained from all subjects for scientific use of the recorded data. The data was acquired in a dim light room with minimized distractions like external sound, mobile devices and others, where the voluntary participants were asked to sit in a comfortable chair to prevent unnecessary muscle movements to reduce noise and artifacts in the EEG, which could emerge from mental stress, unrelated sensory input, physiological motor activity and electrical interference. Three subjects (1,2 and 19) were excluded due to poor electrode-to-skin contact later, leaving a total of 17 subjects (8 male, 9 female).

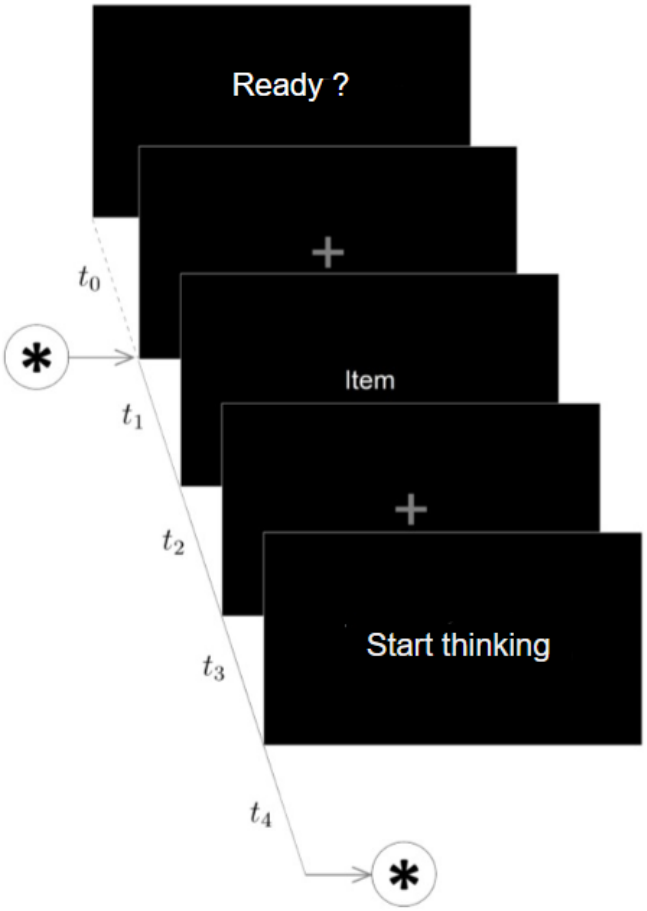

Figure 4.8: Procedure of visual output on the screen during the speech imagery task, as done for 5 semantic categories in each trial ($t_0 = \infty$, $t_1 = 1.0s$, $t_2 = 0.1s$, $t_3 = 0.3s$ and $t_4 = 2.5s$).

Table 4.3: Selected semantic categories and items presented in the imagined speech task, originally in German, translated to English.

Category	Items
Locational	Kitchen, Bathroom, Cellar, Garden, Court, Bedroom, Living room, Staircase, Corridor, Attic
Actions	Throw, Open, Lift, Lower, Switch on, Switch off, Put, Place, Push, Pull
Living	Dog, Cat, Peacock, Lamb, Pigeon, Mother, Father, Grandmother, Grandfather, Doctor
Non-living	Light, Shutters, Heater, Television, Telephone, Computer, Stove, Refrigerator, Washing machine, dryer
Numbers	One, Two, Three, Four, Five, Six, Seven, Eight, Nine, Ten, Eleven

Recording

EEG signals were recorded using a wireless 32 channel electroencephalograph system namely g.Nautilus with g.Scarabeo electrodes (g.tec medical engineering GmbH, Austria). The sampling rate was set to 500 Hz. The 10-20 International System of electrode placement was used to locate the electrodes. This configuration is believed to cover the whole scalp resulting in the capturing of spatial information from the brain recordings effectively which provided the optimal setup for our study based on the findings of [88].

Data analysis

Preprocessing A basic preprocessing, including filtering and referencing, was applied to remove unwanted artifacts. Channels containing obvious signal quality issues as well as data segments identified to contain strong artifacts were labeled bad and thus, excluded from further analysis. The EEG was filtered with a second-order Butterworth notch filter with a lower cutoff frequency of 48Hz and an upper cutoff frequency of 52Hz, to remove power line noise. A fourth-order Butterworth was applied for high pass filtering at a cutoff frequency of 1Hz as recommended by [224]. Lowpass filtering was done using a 17th order Chebyshev type 2 filter with a cutoff frequency of 200Hz preserving the frequency range up to high-gamma [46], and a minimum of 60dB attenuation in the stop band, to obtain adequate roll-off. All the electrodes were referenced to common average over all electrodes to achieve low signal-to-noise ratio [136]. Further denoising of the EEG was done using a joined method of Independent Component Analysis (ICA) and wavelet denoising, referred to as wavelet-enhanced independent component analysis (wICA) [7] to remove artifacts.

Epoching The continuous EEG signal was cut into smaller epochs of different length for further analysis. Epochs were extracted in reference to the stimulus onset from the pre-processed data. To further analyze the impact of different epoching duration, three

intervals were chosen, namely, $T_1 = 0.3s - 0.8s$, $T_2 = 0.3s - 1.5s$ and $T_3 = 0s - 2s$. T_1 and T_2 were supposed to cover the point around 400 ms after stimulus onset, as studies on semantics assume this to be the earliest point information is consciously processed [112]. The third interval was chosen as the entire period between stimulus onset and end of the task in order to cover as much information as possible.

Feature Extraction We investigated two feature extraction methods, first we assembled a feature vector containing features from the time and frequency domain and based on the different state-of-the-art literature in Speech Imagery detection [183, 115, 235, 145]. The features were extracted with the open-source python module PyEEG [14]. As a second feature extraction method we chose Common Spatial Patterns (CSP) which is a frequently used technique in BCI applications [109, 156]. The basic principle behind CSP is to apply a linear transformation to project the multi-channel EEG signal data to a lower-dimensional spatial subspace. The transformation results in the maximization of the variance of one class while minimizing the variance of other class at the same time [4]. Originally the CSP algorithm is developed for binary classification problems, but there are some studies showing the multiclass classification with One-vs-all scheme [175, 32, 75]. In this study, we implemented the multiclass classification setup as mentioned in [179] with five labels and a chance level at 20%.

Classification Classification was done based on two strategies, cross-subject and within-subject. In the cross-subject analysis, the data from all subjects is taken together as a single input in order to find similar patterns among subjects while in within-subject analysis, the performance of the classifier is computed on the individual subjects data set. We used Support Vector Machine (SVM) and Random Forest (RF), because of their frequent use in EEG-based Silent Speech BCIs [219, 206, 147], both evaluated using grid search and ten-fold cross-validation with a test train split of 0.1. For the cross-subject data set, we further implemented a deep artificial neural network, the Multilayer Perceptron using the ADAM and SGD optimizer with mean-squared error as loss function. In this case grid search was performed to find the optimal number of hidden layers and chose the best learning rate in addition to the activation function. Unfortunately, within-subject analysis with neural networks was not possible due to the limited availability of data of an individual subject. All the experiments were multi-class classification problems with five labels resulting in a chance level of 20%.

Performance metrics To evaluate the performance of the classifiers, we have calculated the classification accuracy. Accuracy is the most commonly used performance metric and defined as the ratio of the total number of correct predictions to the total number of predictions overall. All the experiments were multi-class classification problems with five labels resulting in a theoretic chance level of 20%. However, as explained in section 2.9, this threshold theoretically requires an infinite number of classifications. Combrisson and Jerbi proposed an adjustment of the chance level for machine learning in neuroscience [39] which we calculated based on their work as explained in section 2.9. With the 5 categories and an alpha value of 0.05 we received a significance threshold of 27.00% for the testset including 20 samples after cross-validation.
The classification accuracy of our classifiers was calculated as the number of correct predictions divided by the total number of predictions and compared with the significance threshold to validate statistical significance of the results. We have furthermore created the confusion matrix and classification reports including F1-score, Precision and Recall and present them in this work for the best performing subjects.

4.2.2 Results and Discussion

In the following we will present the results of the analysis on the recorded dataset performed as described in the last section. We will start with presenting the results and discuss them afterwards.

Results

Table 4.4 shows the results for the cross-subject data set evaluated with the different classifiers (SVM, RF, MLP), feature extraction methods (CSP, FV) and concerning the different epoching intervals (T_1, T_2, T_3). For the CSP feature extraction condition, the classifiers did not manage to exceed chance level. In all other cases, the classifiers evened out around 40% which is significantly above chance level and RF classifier achieved the highest accuracy for all epoching intervals.

Table 4.4: Mean Accuracy for the cross-subject classification depending on the different epoching intervals (T_1, T_2, T_3), feature extraction methods (CSP and FV) and classifiers (SVM, RF, MLP) used.

Classifier	T_1 **= 0.3s - 0.8s**	T_2 **= 0.3s - 1.5s**	T_3 **= 0s - 2s**
SVM-CSP	$19.93 \pm 1.90\%$	$20.05 \pm 2.11\%$	$20.94 \pm 1.35\%$
RF-CSP	$22.54 \pm 2.64\%$	$20.64 \pm 2.07\%$	$21.65 \pm 1.24\%$
SVM-FV	$38.45 \pm 2.27\%$	$37.21 \pm 2.06\%$	$41.59 \pm 1.79\%$
RF-FV	$42.18 \pm 1.61\%$	$40.05 \pm 2.25\%$	$43.54 \pm 2.13\%$
MLP-ADAM	$38.92 \pm 2.75\%$	$38.69 \pm 2.77\%$	$39.64 \pm 1.10\%$
MLP-SGD	$40.17 \pm 1.08\%$	$40.94 \pm 3.74\%$	$40.88 \pm 2.50\%$

Figures 4.9 – 4.13 show the results for the within-subject condition and the different classifiers used on the assembled feature vector and the CSP feature extraction methods. For within-subject analysis using the Support Vector Machine and the feature vector, figure 4.9 summarizes the results for all three intervals. Overall, the highest accuracy of 68.42% is achieved by subject 4 in the duration T_1. The mean accuracies were reported to be 46.08%, 45.81%, and 47.81% in the duration of T_1, T_2 and T_3 respectively. Except for subjects 5 and 6, all the other subjects managed to achieve classification results above chance level and significance threshold, as represented by the blue and dashed green horizontal lines across all the data arrangements and epoching intervals.

Figure 4.10 shows the accuracy versus subject plot for all three epoching intervals using the random forest classifier and the feature vector. The overall highest accuracy of 70% is achieved by subjects 16 and 11 in the epoching interval T_3 and T_2, respectively. The mean accuracies are 52.81%, 52.52%, and 55.18% in the duration of T_1, T_2 and T_3 respectively. In this case, all the subjects managed to score above chance level and the significance threshold. As compared to SVM, RF reported better accuracies for all the subjects.

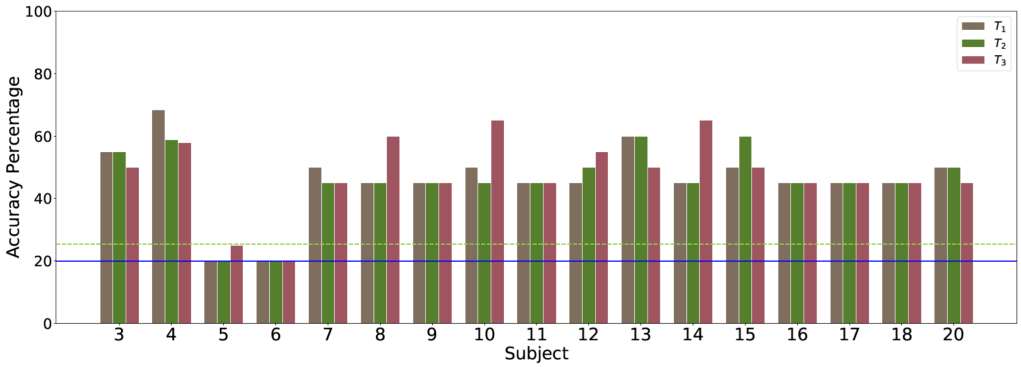

Figure 4.9: SVM within-subject accuracy using the assembled feature vector for three different epoching intervals, $T_1 = 0.3s - 0.8s$, $T_2 = 0.3s - 1.5s$, $T_3 = 0s - 2s$. The solid blue line indicates chance level, the green dashed line the significance threshold.

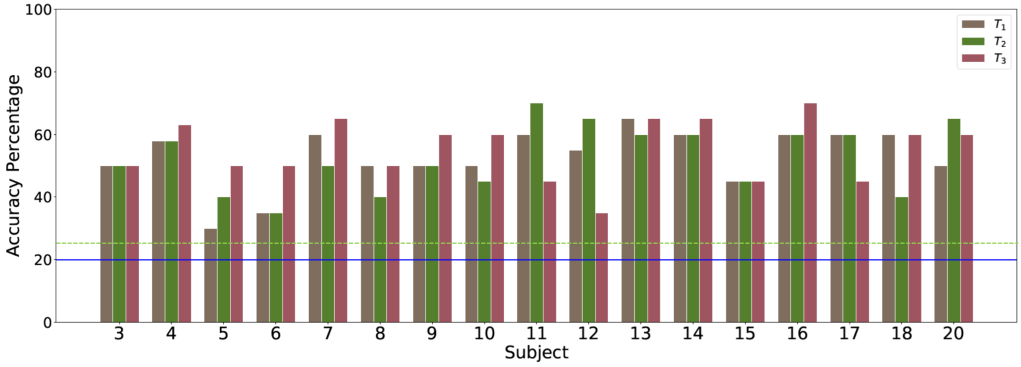

Figure 4.10: RF Within-subject Accuracy using the assembled feature vector for three different epoching intervals, $T_1 = 0.3s - 0.8s$, $T_2 = 0.3s - 1.5s$, $T_3 = 0s - 2s$. The solid blue line indicates chance level, the green dashed line the significance threshold.

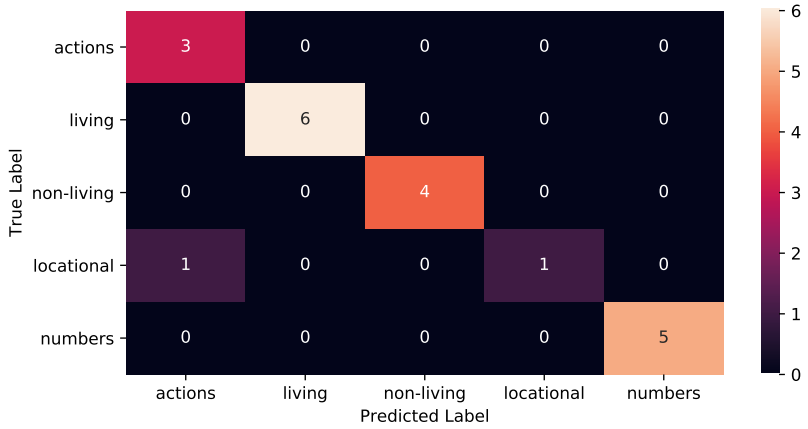

Figure 4.11: Confusion matrix for subject 10 in the interval T_3 using SVM-CSP.

Table 4.5: Classification report for subject 10 using SVM-CSP and T_3

Labels	F1-Score	Precision	Recall	Support
actions	0.85	0.75	1	3
living	1	1	1	6
Non-living	1	1	1	4
Locational	0.67	1	0.5	2
Numbers	1	1	1	5

Figure 4.12 presents the results obtained with the implemented CSP algorithm using a Support Vector Machine. The overall highest accuracy of 95% is achieved by subject 10 in the interval of T_3. The confusion matrix for this condition can be found in figure 4.11 and the classification report in table 4.5. The highest accuracies of 90%, 85%, and 95% are achieved in the interval T_1, T_2, and T_3, respectively. The mean accuracies are 56.91%, 51.08%, and 55.77% in interval of T_1, T_2 and T_3 respectively. The worst performing subject is subject 13, in the interval T_1. However, in the other two intervals T_2 and T_3, subject 13 manages to achieve accuracies above chance and for T_3 also above significance threshold.
Figure 4.13 illustrates within-subject accuracy for all three epoching durations using Random Forest. Overall, the highest accuracy achieved is 90% by subject 20 in the epoching interval of $T_1 = 0.3s - 0.8s$. The confusion matrix for this condition can be found in figure 4.14 and the classification report in table 4.6. The mean accuracies are 60.44%, 51.43%, and 51.93%, in T_1, T_2, and T_3, respectively. The worst performing subject is 13, below chance level in the interval T_2. However, in the other two intervals T_1 and T_3, subject 13 manages to achieve classification accuracies even above significance threshold. Subject 15 is at chance level in the interval T_2, while in the interval T_1, subject 15 manages to reach an accuracy of 60%. Thus, we can observe a lot of fluctuation in accuracy in different epoching durations.

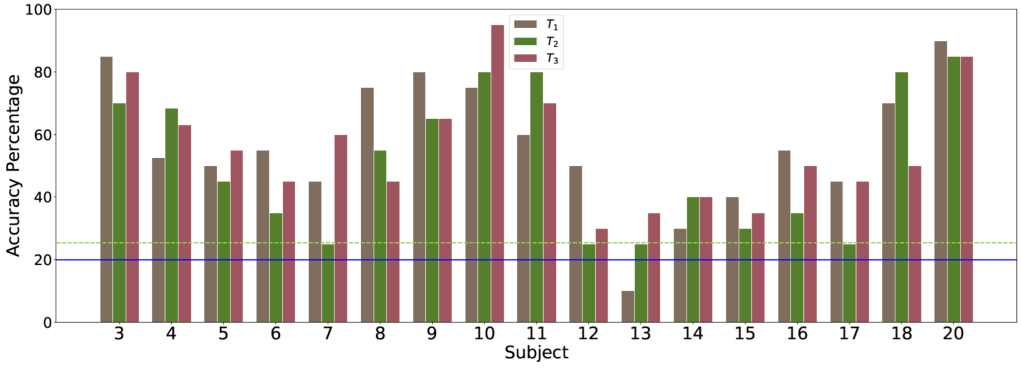

Figure 4.12: SVM within-subject accuracy using CSP for three different epoching intervals, $T_1 = 0.3s - 0.8s$, $T_2 = 0.3s - 1.5s$, $T_3 = 0s - 2s$. The solid blue line indicates chance level, the green dashed line the significance threshold.

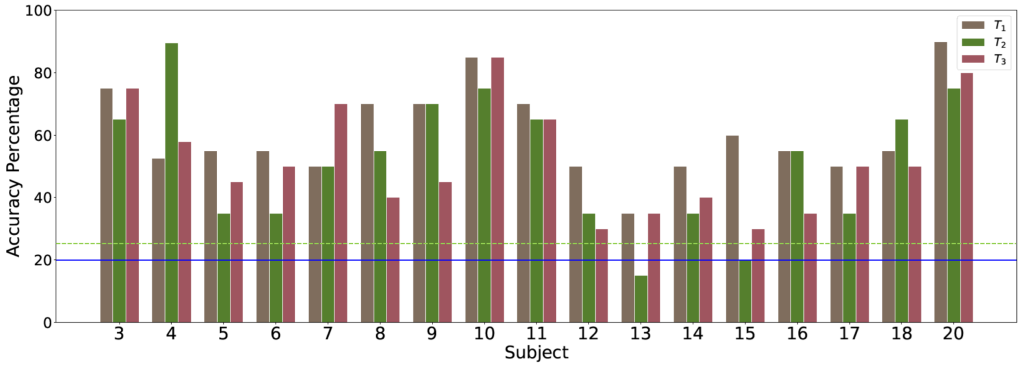

Figure 4.13: RF within-subject accuracy using CSP for three different epoching intervals, $T_1 = 0.3s - 0.8s$, $T_2 = 0.3s - 1.5s$, $T_3 = 0s - 2s$. The solid blue line indicates chance level, the green dashed line the significance threshold.

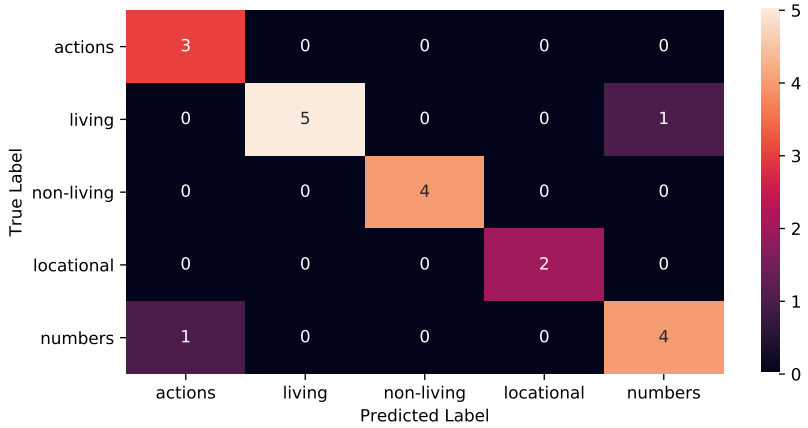

Figure 4.14: Confusion matrix for subject 20 in the interval T_1 using RF-CSP.

Table 4.6: Classification Report for Subject 20 using RF-CSP and T_1.

Labels	**F1-Score**	**Precision**	**Recall**	**Support**
actions	0.85	0.75	1	3
living	0.90	1	0.83	6
Non-living	1	1	1	4
Locational	1	1	1	2
Numbers	0.8	0.8	0.8	5

In order to evaluate how well the different semantic categories could be discriminated we selected the two most promising conditions regarding classification accuracy, namely CSP-SMV for T_1 and CSP-RF for T_3, and calculated the error percentage for the different categories.
Figure 4.15, illustrates the evaluation of how well the semantic categories could be discriminated for CSP-RF in $T_1 0.3s - 0.8s$. The plot shows the error percentage for each category and each subject for the epoching interval of T_1. For subject 20 the classifier managed to classify all the instances of categories, actions, numbers, and non-living correctly, while categories living and locational were misclassified by 16.67% and 20%. On average for all subjects, categories actions, living, non-living, numbers, and locational are correctly classified with 71.17%, 48.16%, 64.70%, 72.06%, and 60.59%. Hence, the best discriminated category across all the subjects in this condition is numbers.

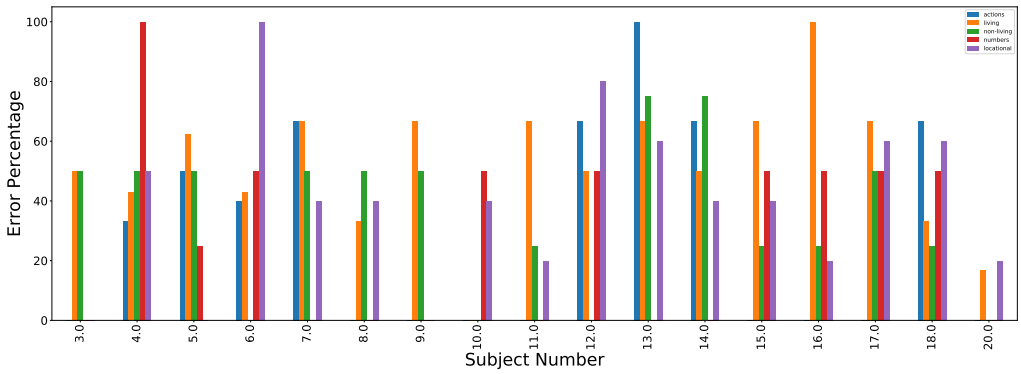

Figure 4.15: Error percentage per subject for discriminating the five semantic categories (actions: blue, living: orange, non-living: green, numbers: red, locational: purple) using CSP and RF in the interval T_1.

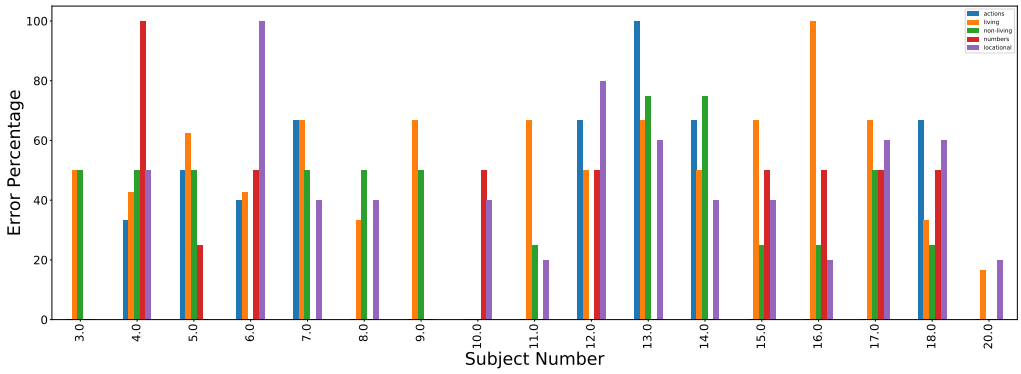

Figure 4.16: Error percentage per subject for discriminating the five semantic categories (actions: blue, living: orange, non-living: green, numbers: red, locational: purple) using CSP and SVM in the interval T_3.

Figure 4.16 depicts the error percentage for each category in the interval $T_3 = 0s - 2s$ using CSP and SVM. In this Figure, subject 10 manages to classify all the instances of categories actions, living, non-living, and locational correctly, leaving zero error. On average, categories actions, living, non-living, numbers, and locational are correctly classified with 56.48%, 56.91%, 66.18%, 54.91%, and 47.65% across all subjects. Hence, the best discriminated category across all the subjects in this condition is non-living.

Discussion

The main goal of this study was to classify EEG activity related to semantic processing during word imagination. In our attempt to train different classifier and feature extraction methods on the recorded EEG data of all participants in the cross-subject condition, we can say, that the assembled feature vector as well as the Multi-Layer-Perceptron approach managed to exceed chance level witch classification accuracies of up to 43.54% for the Random Forrest Classifier and the assembled Feature Vector. These results indicate that the classification of semantic categories of imagined words based on cross-subject data might be possible. This hypothesis should be further investigated however with a larger dataset, especially to make use of the potential of the neural networks. The hypothesis, that the imagination of words from the same semantic category produces similar spatial patterns of EEG activity across all subjects, as shown for fMRI measurements in [88], could not be proven by the Common Spatial Pattern method, with classification accuracy around the chance level of 20%. These results are consistent with the findings of our previous study for the classification of semantic categories during object based decision tasks [179].
The within-subject condition yielded similar results. Our analysis clearly indicates that it is possible to classify the semantic category of a word a person was subvocalizing in their head with a remarkable single best classification accuracy of 95% for subject 10 with CSP feature extraction and SVM classifier. However, there does not seem to be one best setup or combination of epoching, feature extraction or classification methods. Comparing the two feature extraction methods used we can say that CSP deliverd better results on average than the assembled feature vector. While the feature vector managed to produce a more uniform distribution amongst the subjects with no significant outliers, the results for CSP vary strongly among subjects. The two classifiers do not produce significantly different results within the two feature extraction methods but again on average the Random Forrest classifier performed slightly better than the Support Vector Machine with one exception for SVM and CSP at T_3 where the single best subject was reported.
Although our results show that on average the best method for semantic category classification in word imagination tasks (among those presented in this study) is a Common Spatial Pattern feature extraction combined with a Random Forrest classifier and an epoching interval of 300 - 800 ms after word imagination onset, we clearly recommend to tailor those methods to the individual. As shown in our previous study for object based decision tasks, the word imagination task appears to be highly subject specific. While subject 14 managed to achieve classification accuracies of above 60% for the feature vector it hardly manages to exceed chance level for the Common Spatial Pattern method. The opposite holds for subjects 5 and 6 although improvement can be shown when switching from a Support Vector Machine to a Random Forrest Classifier in the Feature Vector condition. While subject 10 achieved the best results in all feature extraction and classification methods within the epoching interval T_3, subject 3 performed best within the interval T_1, and the results of subject 4 regarding the epoching interval vary completely under the different conditions.

4.2.3 Conclusion

The goal of this study was to explore the potential of classifying the semantic categories of imagined words from EEG data for the use in Speech Imagery BCIs. We implemented various feature extraction and classification methods as well as different epoching intervals with the aim to provide a recommendation for a best possible setup concerning semantic category classification during word imagination. We furthermore analysed the data in a cross-subject approach with the intention to find similar patterns in brain activity among subjects but also investigated a within-subject condition where the classification was done on the data of each subject individually.
The cross-subject results did not show the expected common spatial patterns among all participants but promising results could be achieved on temporal and spectral features. With classification accuracies of up to 43.54% the proposed method with a Random Forrest classifier clearly exceeded chance level but is still far from an accuracy needed for real world applications. This approach should be further evaluated in the future on a larger dataset to fully exploit the potential of Neural Networks for classification which usually work with a tremendous amount of input data, larger than what could be provided in this study.
The results for the individual subjects were promising as well but highly distributed among the different epoching intervals, feature extraction and classification methods. The results clearly indicate that it is possible to classify the semantic category of a word during word imagination, with a best average classification accuracy of 60.44% for CSP feature extraction and SVM classifier in a time interval of 0.3s - 0.8s after imagination onset, and a best single subject classification accuracy of even 95% for CSP and RF classifier for the full length time interval of 0s - 2s. As shown in our last study on semantic category classification in object based decision tasks, the epoching intervals, feature extraction and classification methods are highly subject specific. There was no clear best setup to recommend for semantic category classification but rather the conclusion to select the different methods tailored to the individual based on predefined training sessions. Nevertheless, with our study setup we could show that it is possible to classify semantic categories from EEG activity during word imagination, for the first time for 5 different categories simultaneously, with individual classification accuracies throughout exceeding significance thresholds. These results illustrate the potential of the method for the use in Semantic Silent Speech BCIs.

In the following section we will go the last step on the way to a real Semantic Silent Speech or Speech Imagery BCI by actually implementing the 2-layer approach of classifying the semantic category of an imagined word followed by the classification of the word itself.

4.3 Semantic Silent Speech BCI

A Semantic Silent Speech BCI makes use of the fact that semantic processing triggers a broad network of different regions on the cortex of the human brain (see section 2.7.2). This widely spread activation makes it a perfect use case for EEG-based BCIs which usually suffer from a low spatial resolution and can benefit from this distribution. The concept of the Semantic Silent Speech BCI involves a second layer of semantic category classification prior to the actual word classification layer. This concept can in theory be applied to any kind of Silent Speech interface based on brain activity (see section 2.4). Thus, we will not specifically tailor this definition to Speech Imagery BCIs in the following. The idea is based on a study of Huth et al [88] which showed, that it is possible to cluster patterns of similar brain activity for words with the same semantic category, while the participants were listening to short stories with those words embedded. Those similar patterns could not only be shown for the individual but also among the brain activity of all subjects, which indicates that it might be possible to train a classifier on certain patterns of brain activity for semantic categories and achieve cross-subject classification. The main advantage and purpose in the scope of our work however, lies on the possibility to increase classification accuracy and thereby the number of simultaneously classifiable words by adding this additional semantic layer to Silent Speech BCIs, given that it provides sufficiently precise classification. We have define a Semantic Silent Speech BCI in [180] as follows:
A Semantic Silent Speech BCI tries to classify imagined words of different semantic categories and consists of a word and a semantic category classifier. The system classifies the semantic category in a first step, followed by the word itself with a separate word classifier trained on the subset of words inside this category. Let n be the number of different semantic categories and m be the number of different words equally distributed among those categories in a Silent Speech task. With the additional semantic layer, we can divide the original m-class classification task (resulting from the overall m numbers of words) into a n-class classification task followed by an $\frac{m}{n}$-class classification task. Given the fact that semantic processing in the brain was shown to be present spread over the whole cortex [88] and that imagined speech production mainly evokes the left hemisphere around Wernicke's area [210], the feature space and sources of those two related cognitive processes, speech production and semantic processing, should be separable and provide distinguishable features. By decomposing the original n-class classification task into smaller classification tasks with a separated feature space, we expect the number of classifiable words to increase by n (= number of semantic categories) in the best case.

Within our works in section 4.1 and 4.2 we could show that it is possible to classify semantic categories of imagined words from EEG data, which were inspired by the studies of Huth et al. [88] who achieved similar results, however with fMRI instead of EEG data and not in a BCI context. Based on those findings and assumptions we will in the following use the two datasets recorded in chapter 3 to prove the feasibility of the concept of a Semantic Silent Speech BCI in an offline analysis on those two datasets.

4.3.1 Methodology

Due to the novelty of the concept and the lack of related work in this field of research we decided for an offline analysis of recorded data instead of an online BCI application. This allowed for a broad investigation of a variety of different methods without frustrating study participants with unreliable online classification accuracies which might have an impact on the recorded brain activity conversely. We used the recorded imagined speech data from chapter 3 as described in more detail in sections 3.1 and 3.2. In the following we will give an overview on those datasets, the methods used for preprocessing, feature extraction, classification and the performance measures applied to evaluate the newly developed approach.

Datasets

The datasets used for this evaluation were recorded during the work on improving SI-BCI training scenarios in chapter 3. The detailed description of the recording procedure and its purpose can be found in section 3.1 and 3.2. In this section we will only give a high-level overview on the dataset with the most important properties for the data analysis within this chapter.
Dataset one included 9 German words from 3 different semantic categories. The categories were chosen as locational, including the German words for "workbench", "floor" and "conveyer belt", non-living, including the German words for "screw", "case", "PCB", and action verbs, including the German words for "hold", "lift" and "put". The words were recorded in a standard imagined speech training procedure as described in section 3.1. The dataset consists of 40 silent repetitions of each word, leading to 120 silent repetitions per semantic category and 360 imagined speech samples overall.
Dataset two was recorded in a setting where participants had to navigate a robot through a factory with three directional commands, up, left, right, and the two action verbs, pick and push. During this recording, the words had to be silently produced in English and could be freely selected by the participant, however, the study setup ensured that they had to be repeated equally often to create a balanced dataset. Each of the words was repeated 80 times resulting in overall 400 imagined speech samples. For more details on the procedure see section 3.1. The focus of this dataset was not on distinguishing semantic categories of imagined words but rather on the improvement of the training procedures. Therefore the semantic categories had to be assigned subsequently after the study and were chosen to be action verbs, pick and push, and directional words, up, left and right. To ensure a balanced dataset we excluded one of the action verbs to end up with 80 silent repetitions of the directions left and right, and 80 repetitions of the action verbs pick and push.
These two datasets provided silent repetitions of words with different semantic categories and therefore the perfect foundation for the implementation of a Semantic Silent Speech BCI evaluation.

Preprocessing

The EEG data was filtered with a second-order Butterworth notch filter with a lower cutoff frequency of 48Hz and an upper cutoff frequency of 52Hz, to remove power line noise. A fourth-order Butterworth was applied for high pass filtering at a cutoff frequency of 1Hz as recommended by [224]. Lowpass filtering was done using a 17th order Chebyshev type 2 filter with a cutoff frequency of 200Hz preserving the frequency

range up to high-gamma [46], and a minimum of 60dB attenuation in the stop band, to obtain adequate roll-off. Further denoising of the EEG was done using an Independent Component Analysis (ICA) based on the automated Matlab EEGLAB toolbox [27] to remove artifacts.

Epoching

The continuous EEG signal was cut into smaller epochs of different length for further analysis. Our novel Semantic Silent Speech approach requires to train a classifier to detect the semantic category of a word as well as a classifier for the subsequent word detection itself. In both cases the epochs were extracted in reference to the stimulus onset from the pre-processed data for each word. For the semantic classifier the three intervals from our previous studies as presented in sections 4.1 and 4.2 were chosen, namely, $T_1 = 0.3s - 0.8s$, $T_2 = 0.3s - 1.5s$ and $T_3 = 0s - 2s$. T_1 and T_2 were supposed to cover the point around 400 ms after stimulus onset, as studies on semantics assume this to be the earliest point information is consciously processed [112]. The third interval was chosen as the entire period between stimulus onset and end of the task in order to cover as much information as possible. As our previous studies showed that those intervals performed differently between subjects, we decided once more to tailor the analysis to the individual and investigate all 3 time windows for each subject. For the word classifier we only chose T_3 in order to cover as much information of the whole process of the silent repetition as possible. In both cases epoching was performed with the MNE library [73], a software packaged designed for processing and visualization of EEG and MEG data.

Feature Extraction

We investigated several feature extraction methods for the semantic and word level. On the semantic level we chose the two methods evaluated in our previous studies as presented in sections 4.1 and 4.2. First we assembled a feature vector containing features from the time and frequency domain based on the different state-of-the-art literature in imagined speech detection [183, 115, 235, 145]. The Thirteen features included, Power Spectral Intensity and Relative Intensity Ratio (Alpha, Beta, Gamma, Delta, Theta), Petrosian Fractal Dimension, Hjorth Parameters (Mobility and Complexity), Spectral Entropy, Hurst Exponent, Detrended Fluctuation Analysis, Skewness, and Kurtosis. The features were extracted with the open-source python module PyEEG [14]. As a second feature extraction method we chose the Common Spatial Pattern (CSP) algorithm which is a frequently used technique in BCI applications [109, 156]. The CSP was realized with the multiclass implementation of the MNE library [73] with the default parameters.
On the word level we decided on the Common Spatial Pattern (CSP) algorithm and the Discrete Wavelet Transform (DWT). The CSP was realized once more using the multiclass implementation of the MNE library [73] with the default parameters. The DWT was implemented based on the PyWavelets library [121]. As mother wavelet we applied biorthorgonal 2.2 (bior2.2) as suggested in [57]. The data was decomposed until fourth level. Afterwards a wavelet feature vector was created out of the data as presented in [207].

Classification

The purpose of our work was to compare the performance of our newly developed concept of a Semantic Silent Speech BCI with a standard word-based SI-BCI. This comparison and in particular our Semantic Silent Speech approach, required the implementation of two different classifiers, one for word classification alone and another one for semantic category classification in combination with a word classifier for each category containing the corresponding words. Both approaches used the algorithms implemented in our previous studies, namely Random Forrest (RF) and Support Vector Machine (SVM) as described in sections 4.1 and 4.2. We furthermore implemented an Extreme Gradient Boosting classifier (XGB) based on [34] with a mean error as evaluation metric and instructed to stop if the mean error did not decrease for ten rounds. The objective function was chosen to be softmax for multiple classes. Classification was done based on two strategies, a standard word-based approach in which all words were classified at once and our two step Semantic Silent Speech BCI approach in which a semantic classifier tries to classify the category prior to the word itself. Word and semantic classifier were implemented using the same algorithms and datasets but trained with different labels according to their purpose with the labels of the semantic category and the corresponding words in those categories or with the labels of all words. For example, in dataset one this resulted in a classifier trained to detect the three semantic categories and another three classifiers trained on the 3 words inside those three categories, which were applied subsequently after category classification and according to the detected category. The control condition with a standard word-based SI-BCI approach, used a classifier which was trained on all 9 words resulting in a single step 9-class classification problem. All of the classifiers were trained for all combinations of the different epoching windows and feature extraction methods using ten-fold cross-validation with a test train split of 0.1.

Performance metrics

To evaluate the performance of the classifiers, we have calculated the classification accuracy which is the most commonly used performance metric in BCI research and defined as the ratio of the total number of correct predictions to the total number of predictions overall. All the experiments were multi-class classification problems with five labels for dataset two, resulting in a theoretic chance level of 20% and 9 labels for dataset one, resulting in a theoretic chance level of 11.11%. However, those values assume an infinite number of predictions which is a problem especially in the usually limited smaller BCI datasets. In order to allow for a fair and profound decision on the performance of the developed approach and its significance, we calculated the significance threshold for an alpha of 0.05 to investigate if our algorithms performed significantly above chance level. The significance threshold takes into account the number of samples and classes of the dataset and was calculated as described in section 2.9. For dataset one this calculation resulted in threshold of 19.44 % compared to the theoretic chance level of 11.11% and threshold of 37.50 % for dataset to in comparison to the initial 20%.

4.3.2 Results and Discussion

In the following we will present the results of the performance of our newly developed Semantic Silent Speech BCI in comparison to a standard classification approach evaluated on the two different datasets and discuss them subsequently.

Results

Table 4.7 and 4.8 show the classification results on the two datasets as comparison between Standard and Semantic Speech Imagery approach including the feature extraction and classification methods which lead to this result, as well as the time interval in the case for the semantic classification for the individual.
Concerning classification accuracy we compared our results against the significance threshold which was calculated as stated in chapter 2.9 and takes into account the dataset size and classes to be distinguished. For dataset one this calculation resulted in 19,44% as compared to the theoretic chance level of roughly 11% for the 9 classes and in 37,50% instead of the theoretic chance level of 20% for the 5 classes of dataset two. Classification accuracies above those thresholds were considered as significantly above chance level and therefore a successful imagined speech classification.
Taking into account the significant thresholds we achieved classification results significantly above chance level for all participants in dataset one for standard as well as for the semantic approach. In 6 out of 15 times, the standard approach delivered better classification accuracies as compared to the newly developed Semantic SI-BCI. On average the implementation managed to classify 35.11% and 37.01% of the labels correctly for standard and semantic condition respectively. Single best results were achieved in the standard condition by subject 1 with 45.55% and in the semantic condition by subject 14 with 48.05%.
For dataset two, we achieved classification results significantly above chance level as well, however only for the newly developed Semantic SI-BCI. In this semantic condition, all participants exceeded significance threshold, while for the standard approach, the achieved values settled sightly below, with no participant exceeding it. This time, the newly developed semantic approach outperformed the standard approach and delivered better classification results for all participants. On average the implementation managed to classify 34.06% and 39.19% of the labels correctly for standard and semantic condition respectively. Single best results were achieved in the standard condition by subject 11 with 37.18% and in the semantic condition by subject 5 with 40.94%.
Concerning the feature extraction and classification methods we can see a clear dominance of the Extreme Gradient Boosting classifier (XGB), which was responsible for the vast majority of the top classifications for the standard as well as the semantic approach in both datasets. In only one case of the overall 64 classification processes, resulting from 15 subjects in dataset one and 17 subjects in dataset two for standard and semantic approach, the setup with Support Vector Machine (SVM) for subject 12 in dataset two delivered better results with the standard classifier. It needs to be emphasized that dataset one and two were recorded in separate sessions with different subjects which means that we cannot draw any conclusion on the selected methods in between the subjects of the two datasets. On the other hand this fact clearly supports the conclusion that XGB is a promising method for imagined speech classification with the potential to outperform established classifiers as for example SVM or Random Forrest.

The feature extraction methods are presented for the standard approach and separated into word and semantic classifiers for the semantic approach. The methods in the table represent the feature extraction method that lead to the corresponding result and show a similar picture as for the classifiers. For dataset one we observe a clear dominance of the Discrete Wavelet Decomposition (wav) for both, semantic and word based feature extraction. Only once we identified the feature vector (vec) as best performing method for subject 5 in the Semantic SI-BCI condition. In dataset two, we can observe differences in the two approaches. For the Standard SI-BCI we experience a clear preference to the CSP feature extraction method, while the Semantic SI-BCI appears to prefer the wav method for category detection and again CSP for the word classification. While the standard condition delivers a quite clear picture with 11 out of 15 times CSP as the best method and the remaining 4 times for wav, the semantic condition is more distributed among all methods. We experience in the category classifier 9 times the wav, 4 times CSP and twice the vec method. The word classifier delivered its best results with 6 times the CSP, 5 times wav and 3 times vec.

	Standard SI-BCI			**Semantic SI-BCI**				
Subject	**Acc (%)**	**Feat**	**Clf**	**Acc (%)**	**Feat**		**Clf**	**Int**
					Cat	**Word**		
1	45.55	wav	XGB	44.45	wav	wav	XGB	1
2	44.44	wav	XGB	42.50	wav	wav	XGB	3
3	38.33	wav	XGB	37.22	wav	wav	XGB	2
4	32.78	wav	XGB	33.89	wav	wav	XGB	3
5	34.72	wav	XGB	34.44	wav	vec	XGB	1
6	27.78	wav	XGB	28.61	wav	wav	XGB	1
7	26.94	wav	XGB	29.44	wav	wav	XGB	3
8	24.44	wav	XGB	29.44	wav	wav	XGB	1
9	37.50	wav	XGB	37.50	wav	wav	XGB	1
10	31.11	wav	XGB	33.89	wav	wav	XGB	3
11	32.22	wav	XGB	30.56	wav	wav	XGB	1
12	40.00	wav	XGB	38.89	wav	wav	XGB	3
13	35.28	wav	XGB	41.94	wav	wav	XGB	3
14	43.61	wav	XGB	48.05	wav	wav	XGB	3
15	36.94	wav	XGB	41.67	wav	wav	XGB	3
16	27.22	wav	XGB	33.61	wav	wav	XGB	2
17	38.06	wav	XGB	43.06	wav	wav	XGB	3
Average	**35.11**	**wav**	**XGB**	**37.01**	**wav**	**wav**	**XGB**	**3**

Table 4.7: Classification results on dataset one for the standard approach and the Semantic Speech Imagery BCI including the setup parameters, namely feature extraction method, classifier and time interval for semantic classification.

	Standard SI-BCI			Semantic SI-BCI				
Subject	**Acc (%)**	**Feat**	**Clf**	**Acc (%)**	**Feat**		**Clf**	**Int**
					Cat	**Word**		
1	36.87	CSP	XGB	40.62	vec	CSP	XGB	3
2	32.50	wav	XGB	39.68	CSP	wav	XGB	1
3	34.37	CSP	XGB	37.50	CSP	wav	XGB	3
4	33.43	CSP	XGB	40.31	wav	wav	XGB	3
5	32.18	CSP	XGB	40.94	wav	vec	XGB	3
6	35.31	CSP	XGB	39.38	wav	vec	XGB	2
7	32.81	wav	XGB	38.75	vec	CSP	XGB	3
8	31.87	wav	XGB	38.43	CSP	wav	XGB	1
9	35.00	CSP	XGB	40.00	wav	wav	XGB	3
10	34.06	CSP	XGB	39.37	CSP	CSP	XGB	1
11	37.18	CSP	XGB	38.75	wav	vec	XGB	2
12	32.50	wav	SVM	39.06	wav	CSP	XGB	2
13	31.87	CSP	XGB	38.44	wav	CSP	XGB	1
14	36.87	CSP	XGB	38.75	wav	vec	XGB	3
15	34.06	CSP	XGB	37.81	wav	CSP	XGB	1
Average	**34.06**	**CSP**	**XGB**	**39.19**	**wav**	**CSP**	**XGB**	**3**

Table 4.8: Classification results on dataset two for the standard approach and the Semantic Speech Imagery BCI including the setup parameters, namely feature extraction method, classifier and time interval for semantic classification.

Discussion

Having a look at possible correlations and differences in feature extraction methods, we need to keep in mind that the datasets were collected with different subjects and that we therefore cannot investigate correlations or a consistency of the methods within the data of a certain subject in between the datasets. However, within each dataset and the different classification approaches on the word level we can do so. In dataset one only one of the 15 subjects showed a difference in the feature extraction method in between the word classifiers for the standard and semantic SI-BCI approach while in dataset 2 the methods did not match for 9 subjects. With 21 out of 30 overall classifications using the same features for the same participant we can not clearly argue for a within-subject consistency when it comes to feature extraction, at least for dataset two.
The feature extraction method for the semantic classifier resulted in a clear dominance of the wav feature extraction in dataset one, where all best classification results were achieved with wav, and a more distributed picture for dataset two with 9 times the wav, 4 times CSP and twice the vec method. In this case we see a confirmation of our results from chapter 4.1 and 4.2 on semantic category detection, that feature extraction is highly depending on the brain activity of the individual and should therefore be chosen tailored to the individual.
Regarding the conclusion on best performing methods, our results on the feature extraction and classification algorithms need to be put into perspective in the sense, that we only considered the top classification accuracy. In order to determine a single best common method for all subject one could use a fuzzy inference system for example, to determine the best configuration with the least loss of classification accuracy based

on a certain set of rules for all subjects. However, as Speech Imagery BCIs cannot be used without prior training, we would always recommend to tailor the analysis to the individual as the slightest improvement in classification accuracy of only a few percent, can have an impact on usability and the user experience of this technology in the end.
In order to investigate if the average classification accuracy of our newly developed Semantic SI-BCI performed significantly better as compared to a standard implementation, we conducted statistical tests on the results of both datasets. For dataset one we started with a Shapiro-Wilk test on both sets of classification accuracies, standard and semantic, to check for normal distribution. As both datasets showed normal distribution, with a p-value of 0.7104 for the standard and 0.4288 for the semantic approach, we performed a paired one-tailed t-test with the null hypothesis H_0 being:

H_0: There is no difference in the average classification accuracies of standard and semantic approach.

and the alternative hypothesis H_1 being:

H_1: The average classification accuracy achieved with the semantic approach is significantly greater than the one achieved with the standard approach.

The test resulted in a p-value of 0.009306, rejecting H_0 after Bonferroni correction and an alpha of 0.25, leading to the conclusion, that our newly developed Semantic SI-BCI delivered significantly better classification results as compared to the standard approach on dataset one.

For dataset two we followed the same procedure and started once more by applying a Shapiro-Wilk test on both sets of classification accuracies to check for normal distribution. With a p-value of 0.1085 for the standard and 0.5857 for semantic condition, we could not clearly claim for normal distribution. Instead of a t-test we therefore applied a Wilcoxon signed-rank test as it is more robust against outliers and heavy tail distributions. The null hypothesis H_0 was chosen to be:

H_0: There is no difference in the average classification accuracies of standard and semantic approach.

and the alternative hypothesis H_1 being:

H_1: The average classification accuracy achieved with the semantic approach is significantly greater than the one achieved with the standard approach.

The test resulted in a p-value of 0.0003606, clearly rejecting H_0 after Bonferroni correction and an alpha of 0.25, leading to the conclusion, that our newly developed Semantic SI-BCI delivered significantly greater classification results as compared to the standard approach on dataset two.
The classification results are visualized in the boxplots in figure 4.17 including the theoretic chance level in red and the significance threshold in green. The plots clearly illustrate that for dataset one all of the participants for both conditions significantly exceeded chance level with median values of 35.28% for standard and 37.22% for semantic classification.

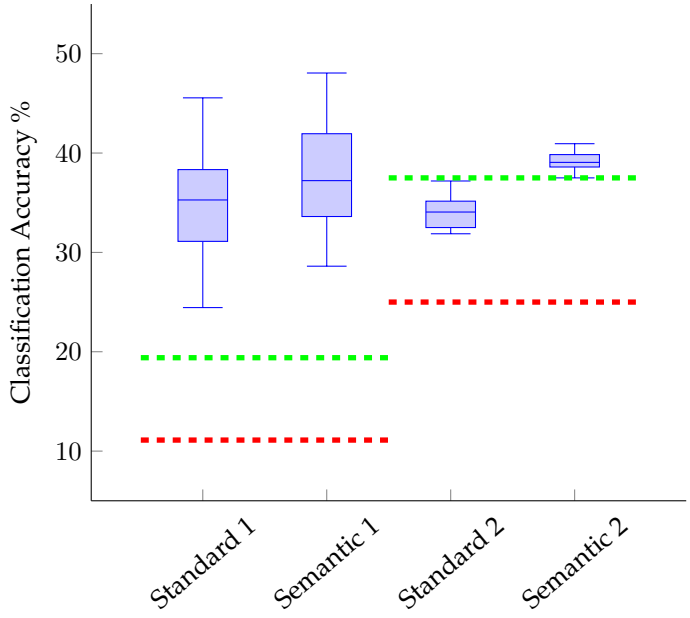

Figure 4.17: Boxplots of the classification accuracies of the standard and semantic approach and the two datasets. The red dashed line represents the adjusted chance level.

Furthermore, we can observe that for dataset two the standard approach remains below significance threshold with a median of 34.06%, while for the semantic approach all participants achieve values above significance threshold, with a median value of 38.13%. The average improvement that could be achieved with the semantic approach in comparison to the standard classification was 1.9% for dataset one with a highest improvement achieved for subject 13 with 6.66%. For dataset two the average improvement resulted in 5.13% with a highest difference achieved for subject 5 with 8.31%. Those differences, although significant, are below the desired improvement which would raise the classification accuracies into the required range for real-world applicability and we can see several reasons for this issue.
Concerning semantic classification, the semantic classifier for dataset one achieved an average accuracy of 58.33% for the three categories, while the classifier trained on the two categories of dataset two scored an average accuracy of 60.56%. Compared to the average accuracies in our previous studies in section 4.1 and 4.2 of 60.44% achieved on a dataset of 5 categories, those values might appear low, as one could expect an increase of classification accuracy with decreasing classes. However, the studies reported in 4.1 and 4.2 were designed to specifically target semantic processes and best distinguishable categories were chosen. The datasets used for evaluation in this section were collected during our attempts to improve imagined speech training procedures in chapter 3, primarily designed to suit a certain training scenario with a robot, and the semantic categories arose as byproduct of the study design. Given these circumstances, the semantic classifiers delivered acceptable classification results, with room for improvement in the future, either by selecting better distinguishable semantic categories or by improving on the classification process.
Surprisingly, the average accuracy of the semantic classifier achieved on dataset one was almost equal to the accuracy achieved on dataset two, although including one additional semantic category. Similar surprising results could be observed for the almost equal

average classification accuracies for the two datasets in standard and semantic condition. Those were rather unexpected, due to the different number of words and therefore classes for the classifier to predict. Dataset one included 9 words assigned to 3 categories resulting in a 3 class classification problem on the word level for our newly developed Semantic SI-BCI approach and a 9 class classification problem for the standard approach. In dataset two there were only 4 words belonging to 2 semantic categories resulting in a binary classification problem on the word and semantic level and a 4 class classification problem in the standard condition. In both cases, we would have expected the average classification accuracies to differ, with a significantly higher accuracy for dataset two due to the lesser number of words.
Furthermore, the classification accuracies varied a lot between subjects for dataset one, while the analysis on dataset two yielded more dense results, as expected with a 10-fold cross validation. Finally, the fact that for dataset one, the Discrete Wavelet Decomposition delivered best results for all participants for semantic and word classification, let's us conclude, that dataset one might be affected by temporal effects leading to label leakage in the frequency domain, resulting from the short-block stimulus presentation paradigm. As mentioned in chapter 2.5, Probadnig et al. [169] revealed the temporal effects in the blockwise stimulus presentation of a former study [223] and compared different types of presentations, including random, blockwise and short-block presentation. They could show, that the blockwise presentation of words lead to significant improvements of classification accuracy due to the classifier identifying frequency shifts over the time of the experiment resulting from the words being presented in blocks as compared to the random and short-block presentation. Although Probadnig did not explicitly evaluate short-block versus random presentation, the classification accuracies for the short-block appeared to be slightly better than the completely randomized stimulus presentation in their dataset as well. Having a detailed look on the stimulus presentation of our two datasets, we used random stimulus presentation in dataset two and the short-block presentation as presented in [123] for dataset one. In our setup, words were presented in short-blocks of 5. Taking into account the overall only 40 repetitions of each word this means, that 1/8 of the data of each class was recorded in one block, which could have produced similar temporal signatures in the EEG of those blocks, positively effecting the classification process, especially when using frequency based features. We will further elaborate on this topic in chapter 5, when comparing the performance on another dataset, recorded with short-block presentation as well.
In the future, a data collection more focused on the semantic categories, with consequent random stimulus presentation, including more repetitions of words and participants could shed some light on this issue and further help to improve the performance of the developed system.
However, concerning the validity of our evaluation, these circumstances rather support the success of the newly developed concept by achieving those impressive results on two separate datasets recorded for different purposes and the better results with higher difference in between the two approaches for the dataset recorded with random stimulus presentation.
Furthermore, it has to be emphasized, that the classification accuracy of the Semantic SI-BCI approach did not only show better results on average but for all classifications of all subjects on dataset two. This clearly demonstrates the success of our newly developed approach and the potential of the Semantic Speech Imagery BCI to significantly improve classification accuracy in comparison to standard word-based Speech Imagery BCIs in the future. Our research question RQ 1.2, if semantic classification can increase classification accuracies in EEG-based imagined speech BCIs, can therefore be answered with yes.

4.3.3 Conclusion

Within this chapter we went the last step towards a Semantic Speech Imagery BCI and implemented the concept of a two-level approach with the classification of the semantic category of a word prior to the word itself. We evaluated our approach on two datasets, one consisting of 4 words and 2 semantic categories and the other one consisting of 9 words and 3 semantic categories, where in both cases the words were equally distributed among the number of categories. We compared the classification accuracy of our Semantic Speech Imagery BCI against the significance threshold calculated for both datasets and achieved accuracies above this threshold in both cases showing the significance of the achieved results. The classification accuracy was furthermore compared with the accuracy of a standard Speech Imagery approach in which a classifier was only trained on word level. Our Semantic Speech Imagery BCI achieved an average improvement of 1.9% for dataset one with a single best improvement of 6.66%, and an average improvement of 5.13% for dataset two with a single best improvement of 8.31%. Furthermore, our approach showed higher classification accuracies for the majority of participants in both datasets, and even all participants in dataset two, highlighting the success of the Semantic Speech Imagery BCI and its potential to improve classification accuracy even further in the future. The effect of short-block stimulus presentation on the data needs to be evaluated in the future as well as the selection of specific categories best distinguishable by the semantic classifier.

4.4 Summary

EEG-based Speech Imagery BCIs currently suffer from the problem, that the number of simultaneously distinguishable words is insufficient for communication. Setups with 3-4 words deliver satisfying classification accuracies of up to 80% [93] which would allow for a real-world application of this technology, but the higher the number of classes, the higher the decrease in classification accuracy [126].
In order to overcome this problem we presented the approach of a Semantic Speech Imagery BCI in this chapter, which tries to classify the semantic category of a word prior to classifying the word itself. Studies based on fMRI showed promising results in clustering semantic processing in the brain into certain regions depending on the categories of the words, and revealed a functional network of these processes widely distributed over the whole cortex [88]. We wanted to make use of those new insights on a widely spread semantic network in the brain, as this would possibly allow also methods with poor spatial resolution, e.g. the EEG, to detect those processes reliably. In the following we summarize the overall results of the conducted studies and discuss them once more on a higher level in the context of the topic of this thesis. We will continue with explicitly highlighting the contributions of those studies to the field and close the chapter by mentioning limitations of the here presented approaches.

4.4.1 Overall Results and Discussion

Due to the fact that research on detecting semantic processing in the brain based on EEG activity was limited to mainly 2 categories, living and non-living [197, 153], we started with an attempt to prove the feasibility of the detection of more complex setups of semantic categories from EEG activity. In a first step we wanted to extend the existing work on semantic category detection based on EEG data and improve on the commonly used binary classification by decoding 5 different semantic categories of words from EEG activity for the first time. We selected 5 categories from the fMRI study of Huth et al. [88] and assigned 10 words per category. Due to the novelty of the approach we did not start with silent repetitions of those words but a standardized study setup commonly used in neuroscience to trigger conscious processing of a certain word or object, namely decision tasks. This task requires the participant to answer a simple yes/no question concerning the word presented subsequently and therefore enforce conscious processing of the word.
Our approach to decode semantic categories of words from EEG activity in object-based decision tasks was successful. We managed to distinguish 5 semantic categories from EEG data of the individual and achieved classification accuracies significantly above chance level for all 18 participants of our study with a single best classification accuracy of 84.61%.
The attempt to find similar patterns in brain activity among subjects during the processing of the words, as presented in the fMRI study of Huth et al. [88], was not successful. Our Common Spatial Pattern based classification trained on the data of all participants did not achieve accuracies above chance level. An implementation using a feature vector assembled from features of time and frequency domain however, achieved classification accuracies of around 40% . Considering the chance level of 20% these results indicate that a cross-subject classification might be possible in the future with larger datasets.
We furthermore investigated different feature extraction and classification methods to provide a recommendation on a best setup for semantic category classification based on EEG data. The inspection of the configurations for the individual best classification

results of the participants revealed however, that there was no single best setup that could have been selected for all participants. The selected methods and combinations of different feature extraction and classification methods varied strongly between subjects and lead us to the conclusion, that those methods should be tailored to the individual based on prior evaluation on the training set.
The task in the follow-up study required the participant to produce imagined speech, once more including the 5 categories and ten words per category, to prove the feasibility of semantic classification of imagined words from EEG activity.
The study showed a similar outcome as the object-based decision task. The results of the classification on the data of the individual were promising and exceeded chance level for all 17 participants of our study with an average best classification accuracy of 60.44% and a single best classification accuracy of 95% for a 5 class classification problem and a chance level of 20%.
We were once more not able to find common spatial patterns in a cross-subject classification approach, but the feature vector implementation showed again an average accuracy of around 40%. This supports or findings from the object-based decision task study and lets us conclude that having a shared semantic model for a broad field of participants might be possible in the future with a larger amount of data. Those models could be used as a starting point for a self-learning online Speech Imagery BCI to be continuously improved and tailored to the specific brain signal characteristics of the individual.
The highly subject specific configuration concerning feature extraction and classification methods could be confirmed for the imagined speech task study, as we were also not able to find one common best setup for all participants. Just like for the results of the object-based decision task we conclude once more, that semantic category detection from EEG data should make use of implementations tailored to the individual, based on prior evaluation of different methods on a training dataset.
Finally, we integrated the concept of semantic category detection into a Speech Imagery BCI setup to improve the classification accuracy, by implementing a Semantic Speech Imagery BCI which tries to classify the semantic category of an imagined word prior to the word itself. We evaluated our developed concept in an offline classification approach on two datasets containing the words belonging to different categories and compared it to a standard word-based Speech Imagery classifier.
Our Semantic Speech Imagery BCI achieved an average improvement of 1.9% for dataset one with a single best improvement of 6.66%, and an average improvement of 5.13% for dataset two with a single best improvement of 8.31%. Furthermore, our approach showed higher classification accuracies for the majority of participants in both datasets, and even all participants in dataset two, highlighting the success of the Semantic Speech Imagery BCI and its potential to improve classification accuracy even further in the future.
An overall worse average classification accuracy for the semantic categories, as compared to the results in the previous chapters, might result from the recording procedure of the two datasets, which was not focusing on the semantic aspect but rather included additional tasks in between the imagined speech parts which might have had an impact on the brain activity of the participants an therefore the overall classification accuracy. Furthermore, the possible effect of short-block presentation, together with general questions on variability of brain activity by various factors in between different Speech Imagery BCI sessions and their impact on classification accuracy, need to be evaluated in further studies in the future.

Summing up our results, we wanted to answer the following research questions within this chapter:

> **RQ 2.1** Can semantic categories be classified from EEG activity during imagined speech production?
>
> **RQ 2.2** Can semantic classification increase classification accuracies in EEG-based imagined speech BCIs?

Concerning **RQ 2.1**, the results of the two studies on the classification of semantic processing from EEG activity and the concrete results from section 4.3 lead us to the conclusion that we can answer the question with a clear yes. Both conditions, object-based decision tasks and the imagined speech paradigm, lead to classification accuracies significantly above chance level, supporting our claims. With this setup we could show for the first time, that it is possible to classify semantic categories from imagined words based on EEG data.
Concerning **RQ 2.2**, we would once more answers the question with yes. In our analysis on two imagined speech datasets, we achieved an average improvement of 1.9% for dataset one with a single best improvement of 6.66%, and an average improvement of 5.13% for dataset two with a single best improvement of 8.31%. This clearly indicates the enhanced performance achieved with our newly developed Semantic Speech Imagery BCI approach, and illustrates the potential to refine classification accuracy even further in the future.

Further conclusions and recommendations are summarized in the following as contributions and limitations of the findings in this chapter on semantic category detection in EEG-based Speech Imagery BCIs.

4.4.2 Contributions

Although our work has a focus on Human-Computer-Interaction and follows an engineering approach, we were able to show that it is possible to classify more than just the common two semantic categories, living and non-living, from EEG activity. These results illustrates the potential of EEG experiments in neurolinguistic research on semantic processing in the brain, beyond the existing applications and might encourage further research to consider EEG as a potential measure, hence replacing cumbersome and cost intensive fMRI setups.

Within our study setups and analysis in this chapter we evaluated a variety of feature extraction and classification methods for imagined speech detection based on EEG activity. Our results on feature extraction suggest that it should be tailored to the individual and there is no clear single best feature extraction method to consider for Speech Imagery classification. Concerning results on different classifiers we can say that our Extreme Gradient Boosting (XGB) implementation outperformed all other methods in a direct comparison on two imagined speech datasets, which makes it our clear recommendation for further Speech Imagery research and application.

Our main contribution of this chapter consists of the conceptualization and actual implementation of a Semantic Speech Imagery BCI which aims at classifying the semantic category of a silently spoken word prior to the word itself. Our implementation of this Semantic Speech Imagery BCI showed throughout better classification results in direct comparison to a standard word-based classifier on two different Speech Imagery datasets.

This shows the potential of the method to cope with the usually low classification accuracies, especially with increasing number of words to be classified simultaneously. With the Semantic Speech Imagery BCI we provide a valuable tool for further Speech Imagery BCI research and application.

4.4.3 Limitations

Our results on semantic category detection are based on machine learning methods and follow an engineering approach. The cross-subject evaluation is based on the Common Spatial Pattern (CSP) algorithm and the assumption, that a successful classification based on this feature extraction method should speak for common patterns of EEG activity between subjects. The fact that our classification with CSP did not exceed chance level does therefor not allow for a clear conclusion on the active regions or similarity in the patterns and should be further investigated including a holistic neuroscientific analysis of the recorded EEG data.

One of the reasons for this missing neuroscientific analysis of the data is the limited number of participants in our experiments with on average 15 per study. This number allows for an evaluation of the performance of a developed system and implemented algorithms and methods but does only hardly allow for generalization of detected patterns in brain activity. Those effects but also the performance of the developed system need to be investigated further on a larger dataset to strengthen the achieved results and conclusions.

The Semantic Speech Imagery approach was able to improve the classification accuracy for all the subjects of the different datasets, however only by around 1.9% and 5.13% on average. Furthermore, a clear conclusion in how far this improvement could cope with an increase of simultaneously detectable words is questionable. The method is still in its infancy and needs to be investigated further on larger datasets and with more fine tuning on the classification and feature extraction algorithms.

Additionally, our setups were still quite controlled in a lab environment and the data analysis was conducted offline. In order to prove real-world applicability, the results need to be validated during a real online classification task and show feasibility also in multi-day application.

The limited number of participants and therefore limited samples for classification is a common problem in EEG-based BCI research. One of the main reasons for this limitation is the cumbersome preparation of the experiment with setup times for EEG headsets of 30 to 60 minutes, depending on the number of electrodes. In order to facilitate data collection for larger corpora of imagined speech data in the future we consider electrode reduction a necessary step. Given that for both of our recorded datasets in chapter 3 we used a cumbersome 64-channel setup, and the vast majority of imagined speech research uses such high-resolution headsets [122, 156, 125, 219], the question arises how much each of those electrodes actually contributes to the imagined speech classification process, and if a subset of those electrodes would also be sufficient for successful imagined speech classification. This question will be addressed within our next chapter on electrode reduction in Speech Imagery BCIs.

Chapter 5
Electrode Reduction in Speech Imagery BCIs

EEG-based Speech Imagery research usually focuses on high resolution EEG recordings involving 64 channels and more [124, 191, 235, 110]. Covering the whole brain during recording appears to be a reasonable approach, as shown in our studies in chapter 4, especially due to the low spatial resolution of the EEG. Researchers therefore often aim at recording as many channels as possible. However, those setups including 64+ channels are tedious and time consuming in preparation, especially with gel-based electrodes, but even dry electrode headsets require a longer preparation time until the quality of the signal and the correct placement of all channels is verified. Furthermore, wearing those headsets is accompanied with various deficits in user comfort, as the electrodes need to have a good and stable contact to the scalp, meaning that they are strapped to the head with a certain pressure. The more electrodes, the more pressure on the head, the more gel to be applied, the more wires and connectivity issues, etc.
Beside usability aspects of SI-BCIs one has to take into consideration the computation time and complexity of the feature space that rises with each additional channel. The higher the number of channels, the higher the feature space the higher the complexity and in the end also the computation time and resources needed.
Given the fact that speech production and understanding have been shown to be related to certain areas of the brain (see section 2.3) the question arises if one could also only focus on those areas, and if those setups with 64+ electrodes for basic research can be reduced to a certain subset, in order to make Speech Imagery BCIs applicable in real-world scenarios with less electrodes and therefore less effort in preparation.
In this chapter we elaborate on the possibilities of electrode reduction in Speech Imagery BCIs and aim at answering the following research questions:

RQ 3.1 Is there a single best minimal set of electrodes for imagined speech BCIs?

RQ 3.2 Can we determine certain electrode positions related to good imagined speech classification accuracies?

We will start by evaluating several electrode reduction methods in combination with various feature extraction and classification algorithms on an imagined speech dataset, in order to conclude on a best setup and method for electrode reduction for imagined speech data in section 5.1. In section 5.2, we will use the determined methods in a systematic approach for electrode reduction on three imagined speech datasets, with the goal of finding a certain subset of electrodes best suited for imagined speech classification.

5.1 Comparison of Electrode Reduction Methods on Imagined Speech Data

Our literature research on electrode reduction methods in section 2.8 led us to the conclusion to use wrapper methods due to their better performance as compared to filter and embedded methods [140, 117], the ease of implementation and the numerous existing implementations in previous research [5, 6, 228, 207]. These methods aim at assessing certain channel subsets based on the accuracy obtained by the algorithm while learning. In our implementation we reduced the number of electrodes systematically from the maximum to the minimum, as presented in [142], with three different algorithms for electrode reduction. Given that feature extraction plays an important role in BCIs and highly varies depending on the individual [180], we further wanted to elaborate in this section on the effect of electrode reduction on feature extraction methods and the resulting classification accuracy. Therefore, we included a variety of feature extraction and classification methods and evaluated them in combination with the electrode reduction algorithms, in order to find a suitable setup for our systematic electrode reduction in SI-BCIs in section 5.2.
Accordingly, this pre-study aimed at finding a suitable method for electrode reduction in SI-BCIs including various feature extraction and classification methods as explained in more detail in the following.

5.1.1 Methodology

For our comparison of electrode reduction methods in Speech Imagery BCIs we identified 3 promising methods in related works:

- Grey Wolf Optimization (GWO)
- Common Spatial Pattern - Rank (CSP-Rank)
- Independent Component Analysis (ICA)

These methods have already been used in a BCI context and have proven to work reasonably well with EEG data [68, 57, 33], however not for electrode reduction in imagined speech BCIs. Details concerning the algorithms and the implementation in our work are further explained in the course of this section. We will continue with the presentation of the overall concept of the implemented electrode reduction approach.

Concept

As previously mentioned our implementation reduced the number of electrodes systematically from the maximum to the minimum, as presented in [142], based on the classification accuracy achieved with each subset. Hence, the data needs to be processed with a standard pipeline for EEG-based imagined speech BCIs as illustrated in figure 5.1. Our pipeline starts with pre-processing the data in order to remove artifacts and limit the time signal to a certain frequency spectrum relevant for analysis. In the next step the features are extracted and forwarded to a classifier which trains to predict those features and is tested on a separate test set of data. After this classification the electrode reduction algorithm selects an electrode to be excluded from the classification process, based on the three different electrode reduction methods. The data, reduced by the channel selected by the algorithm, is then forwarded to the feature extraction again and this circle continues until only one channel is left.

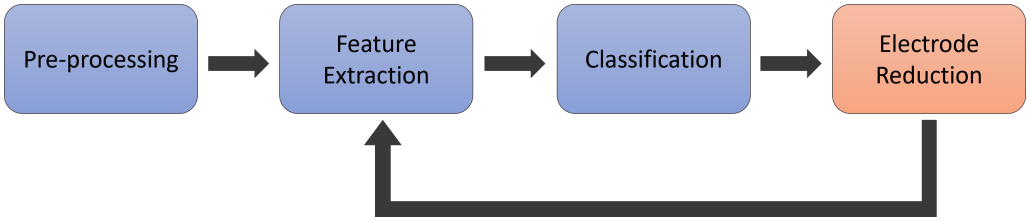

Figure 5.1: Conceptual illustration of the processing pipeline. The data is pre-processed in a first step, afterwards features are extracted and those features forwarded to a classifier. Finally, the electrode reduction method selects one electrode to be excluded and the remaining data is forwarded again to the feature extraction.

This procedure was implemented for all combinations of the different feature extraction, classification and electrode reduction methods and the intermediate results of the classification processes were stored for comparison, as presented in the pseudo code shown in figure 5.2.

The number of electrodes in this algorithm is given by the dataset which will be explained in the following.

Dataset

We applied our algorithms on the existing imagined speech data from the 2020 international BCI competition which is publicly available [18]. It consists of silent repetitions from 15 subjects aged between 20 and 30 years which were asked to perform imagined speech of the five Korean words "Hello", "Help me", "Stop", "Thank you" and "Yes". Subjects were seated inside of a comfortable chair in front of a 24 inch LCD-screen and instructed to imagine silent pronunciation of a given word. Distractions in the surrounding were minimized and the participants were instructed to imagine the silent pronunciation of the given word as if they would speak it, but without moving any articulatory muscle and without producing any sound. Furthermore, they were advised not to perform any other brain activity but the given task, not to move and avoid eye blinks during the task. The stimulus presentation started with three seconds of initial rest time followed by an audio cue of the given word. The cue presentation was followed by 0.8 to 1.2 seconds

[18] `https://osf.io/pq7vb/` Last accessed: 24.10.2022

Algorithm 1: Pipeline pseudocode

Result: Ordered List of Electrodes and corresponding accuracy

```
num_electrodes = 64;
# prepare outputs ele_list = [];
acc_list = [];
while num_electrodes > 1 do
    #select electrode to be removed
      ele = electrodeReduction(data, num_electrodes);
    #remove electrode from data
      data = remove_ele(data);
    #extract features
      fe_data = feature_extraction(data);
    #append results
      ele_list.append(ele);
    acc_list.append(classify(fe_data));
    num_electrodes -= 1;
end
```

Figure 5.2: Pseudocode of the algorithm used for electrode reduction.

of resting time, randomly selected, during which a fixation cross was shown. Subjects were then instructed to perform imagined speech of the given word four times, before moving on to the next one. The recording scheme was illustrated in the previous chapter in figure 3.2. Stimulus presentation was randomized and 70 repetitions per word were recorded resulting in a dataset size of 350 trails overall per participant. The data was split into 300 trials training and 50 trials test set. The trails for test and train set were predefined for the BCI competition and used accordingly for our evaluation.
EEG data was recorded from 64 channels using a Brain Products Live Amp signal amplifier (BrainProducts GmbH). The electrodes were placed according to the international 10-20 system [100] with the ground at position FPz and the reference at position FCz. The impedance of all electrodes between the sensors and the skin of the scalp were maintained below 15kΩ.

Preprocessing

The data was bandpass- filtered between 2 and 50 Hz and notch filtered again at 50 Hz to remove any overlying powerline noise. The parameters for the filtering methods were chosen according to our previous work [180] as explained in detail in section 4.2. After filtering the data was epoched into pieces of two seconds in order to reduce the signal to the relevant sections containing the imagined speech.

Feature Extraction

Feature extraction is one of the most important steps in BCI applications and studies and the performance of a classifier highly depends on the chosen feature extraction method. Due to the novelty of our approach to compare different electrode reduction methods on imagined speech data, we decided to include some of the most common feature extraction methods in BCI research to the comparison as well.
We implemented Common Spatial Patterns (CSP), Discrecte Wavelet Transform (DWT), and a feature vector used in one of our previous works [180].

Discrete Wavelet Transformation Discrete Wavelet Transformation (DWT) tries to model variations in EEG-data within the scale-time domain using the powers of two restrictions. DWT is computed by highpass and lowpass filtering the input time-series data and downsampling it by two [206]. This operation can be repeated iteratively which creates components in a time-frequency representation of the frequency bands with a higher resolution than e.g. short time Fourier transform. This is helpful in EEG analysis to find important patterns in the frequency domain of a signal and the points in time when they occur, which is vital for BCI classification [157].
In our work we used the PyWavelets library [121] for wavelet decomposition of the signal. As mother wavelet we applied biorthorgonal 2.2 (bior2.2) as suggested in [57]. The data was decomposed until fourth level. Afterwards a wavelet feature vector was created out of the data as presented in [207]. This method uses the maximum and minimum of a given time series T as well as its average and the standard deviation combined with the relative wavelet energy of the signal.

Feature Vector The feature vector has been mentioned several times before in this thesis and is explained in detail in section 4.1.

CSP The Common Spatial Pattern algorithm is frequently used in BCI applications [109, 156] and applies a linear transformation to project the multi-channel EEG signal data to a lower-dimensional spatial subspace. The transformation results in the maximization of the variance of one class while minimizing the variance of other class at the same time [4]. In our implementation we used the multiclass CSP algorithm as provided by the mne python library [74]. As this algorithm provides a spatial filtering of the signal, we decided to use it beside its base implementation, additionally filtered by our feature vector and the wavelet decomposition, as described in the two other feature extraction methods above. Both of the methods were applied after performing the CSP on the signal and are referred to in the following as CSPfv for the feature vector and CSPwav for the wavelet decomposition combination. The standard implementation is referred to as CSP.

Classification

Similar to the feature extraction step we used several classification methods in combination with the former mentioned feature extraction and electrode reduction methods, which are commonly used in imagined speech research. We implemented a Random Forest (RF) and a Support Vector Machine (SVM) algorithm as presented in our previous works [179, 180]. Furthermore we integrated an Extreme Gradient Boosting (XGB) based on [34]. It used mean error as an evaluation metric and was instructed to stop if the mean error did not decrease for ten rounds. The objective function was chosen to be softmax for multiple classes. We furthermore implemented a Neural Network (NN) provided in the paper of Panachakel et al. [167]. This neural network features five hidden layers using ReLU and hyperbolic tangent activation functions, except for the final layer which uses the sigmoid function. Dropout and batch normalization were applied between each hidden layer. The network was instructed to stop training once validation loss did no longer decreased for 5 epochs. Categorical crossentropy was used as loss function and Adam as optimizer.

Electrode Reduction

In the following we will give a detailed overview on the 3 electrode reduction methods used in our work.

Grey Wolf Optimization (GWO)
As described in [68] the Grey Wolf Optimization Algorithm is part of the family of evolutionary algorithms which are based on the idea of the "survival of the fittest". This principle was implemented in the algorithm by evaluating a fitness function, in this case represented by the classification accuracy using a dataset in which a randomly selected electrode was excluded. This process was repeated for all electrodes consecutively and the electrode, without which the classification achieved the highest accuracy, was finally rejected. This implementation is inspired by the hunting behavior of the grey wolf, where the alpha animal guides the hunt to the prey, encircling it to come to a closest point to capture it [55]. Although promising, this algorithm has so far only been used for feature selection in imagined speech but not for electrode reduction.

Common Spatial Pattern - Rank (CSP-rank)
The CSP-Rank electrode reduction uses the Common Spatial Pattern algorithm as its basis. This method is usually used for spatial filtering of EEG data in BCI applications, it can however also be modified to reduce the number of electrodes. In this case, the CSP creates the spatial filters for the different classes, using the datasets eigenvectors. From these eigenvectors the longest and the shortest are extracted. The vectors consist of filter coefficients, assigning weights to each electrode, based upon its influence for the class, which also implies that a feature with a larger absolute value is more important. Thus, electrodes corresponding to the largest remaining coefficient are added to the set of electrodes to be used in classification until a stopping condition is reached.
We included the CSP algorithm in two different implementations, the multiclass CSP as mentioned in the feature extraction section and the One-Vs-All implementation.
The multiclass variant was implemented by using the mne python library [74]. CSP was applied to the data belonging to all classes after filtering and epoching. This resulted in the mixing matrix W from which the two vectors with the most common information are selected, as they correspond to EEG signals [75]. For all vectors v in W, each of them corresponding to one channel of the signal, the smallest absolute values are selected, as they contain the least information related to classification. Thus, this electrode is then rejected.
Another way of transforming the originally binary algorithm of CSP into a multiclass approach, is the One-Vs-All method. In this implementation the data is split up in pairs corresponding to their label. The dual class CSP algorithm is then computed, resulting in sets of eigenvectors. As presented in [57], from each set of eigenvectors, the largest and smallest are extracted, as they contain the most information on which electrodes are most relevant for their respective classes. Each vector out of this set is a spatial filter for EEG-data, whose values correspond to a respective electrode. Thus, from all of the chosen vectors the largest absolute value, v_{max}, is selected, as it indicates that the corresponding electrode was significant in distinguishing the classes. The electrode corresponding to the position of v_{max} in the eigenvector is then added to the set of chosen electrodes. This process is repeated until a stopping condition is reached. The unselected electrodes are discarded. In this case the stopping condition was a certain threshold of electrodes being reached. As by the nature of the processing pipeline, as explained in the concept at the beginning of this section, the electrodes are always reduced by one. This

made the threshold out to be always one less than the current amount of electrodes. The standard multiclass CSP for electrode reduction is in the following referred to as CSP and the One-Vs-All approach is referred to as oCSP.

Independent Component Analysis (ICA)
The third electrode reduction method we included into our evaluation is usually applied to EEG data in order to remove artifacts, the Independent Component Analysis (ICA). The ICA algorithm transforms a set of time-series data vectors X into their separate independent components S using a weight matrix W. Thus, the ICA problem can be formalized as $S = W \cdot X$. This algorithm assumes that each vector x in X can be created by a linear mixture of n independent components. The weight matrix W can be derived from these vectors by searching the matrix that minimizes the mutual information of all vectors and thus is able to find the linear mixture S [89]. Commonly ICA is used to remove artifacts, like eyeblinks or heartbeats in EEG-data [207, 179]. This is done by identifying components belonging to artifacts in S. Afterwards the respective vector in the weight matrix is set to be the 0-vector, thus removing the artifact when doing the inverse transform. However, these components can also be used to find information on the neural components of the signal as proposed by [33]. Similar to the CSP-Rank algorithm, one can try to rank the vectors within the ICAs mixing matrix W and reduce the electrodes based on this ranking.

With those 4 electrode reduction methods we started our evaluation based on the criteria described in the following.

Evaluation Criteria

Given that our implementation consisted of various different combinations of feature extraction, classification and electrode reduction methods which we wanted to compare, we needed to define criteria concerning the performance of the different methods. Our main evaluation criterion was the classification accuracy of the developed methods, given as the sum of correct classifications divided by the amount of total classifications. Higher classification accuracy is then linked to a better performance of the combined method. By comparing the classification accuracies of the different methods we selected the top sets, which performed best on the data, consisting of the methods used for electrode reduction, feature extraction and classification as well as the number of electrodes used in this set. This was supposed to give us insights on the best suited number of electrodes for imagined speech classification but also on the feature extraction and classification methods in those best performing subsets.

5.1.2 Results and Discussion

In the following we will present the results of the evaluation on the performance of the different electrode reduction methods and discuss them subsequently.

Results

As mentioned in section 5.1.1 our implementation included the combination of various electrode reduction, feature extraction and classification methods. The results were calculated on the data of the individual and the classification accuracies for all possible combinations and each step of reduction of the electrodes saved for evaluation. Our algorithm reduced one electrode per iteration based on the four different electrode reduction algorithms, which resulted in a large table of classification accuracies of all subjects, for each number of electrodes and each combination of feature extraction and classification methods.
In a first step we wanted to get an overview of the performance of the 4 different electrode reduction algorithms implemented in this work. Therefore, we calculated the mean classification accuracy on the test set over all participants and all possible combinations of feature extraction and classification methods for each step of the different electrode reduction methods.

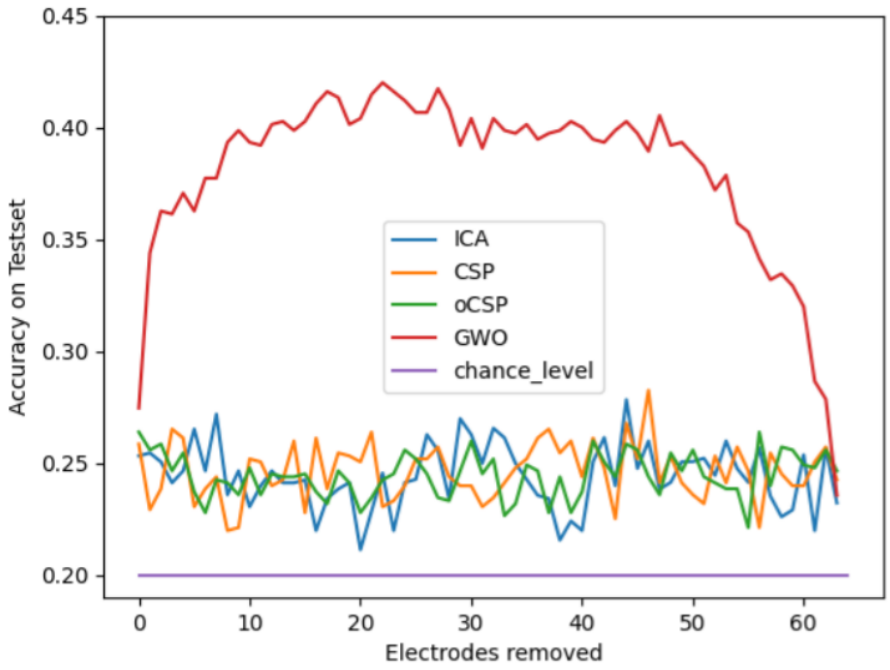

Figure 5.3: Mean classification accuracy on the test set over all subjects, for all possible combinations of feature extraction and classification methods and each step of the different electrode reduction methods, presented as classification accuracy over number of electrodes removed.

	CSP	CSPfv	CSPwav	wav	featvec	Acc Avg	Elec Avg
XGB	(0.45, 20)	(0.46, 24)	(0.56, 25)	(0.50, 53)	(0.34, 59)	0.464	36
NN	(0.38, 48)	(0.34, 43)	(0.36, 59)	(0.36, 49)	(0.36, 24)	0.36	45
RF	(0.44, 39)	(0.34, 58)	(0.56, 14)	(0.48, 36)	(0.28, 35)	0.42	36
SVM	(0.38, 15)	(0.40, 39)	(0.36, 62)	(0.42, 48)	(0.32, 63)	0.376	45
Acc Avg	0.415	0.385	0.46	0.44	0.325	0.405	-
Elec Avg	31	41	40	47	45	-	41

Table 5.1: Example of individual top sets for all combinations of feature extraction and classification methods for GWO on the data of subject 11. The left number in brackets represents the classification accuracy, the right number the corresponding number of electrodes for which this accuracy was achieved.

The results of this calculation are shown in figure 5.3 as the accuracy on the testset over the number of electrodes removed for the 4 different electrode reduction methods ICA, CSP, oCSP and GWO as well as the chance level given at 20 %. Notably, all methods scored consistently above chance level on average. Furthermore, the graph indicates that Grey Wolf Optimization massively outperformed the other methods by a large margin. This impression is further strengthened after the analysis of the individual data where all best performing setups, referred to as top sets, resulted from the GWO algorithm. An example of these results of the top sets is shown in Table 5.1 for all combinations of different feature extraction and classification methods of the best performing subject 11. The left number in the brackets represents the classification accuracy achieved with the given configuration and the number on the right the corresponding number of electrodes with which this accuracy was achieved. From the average classification accuracy concerning the given classifier on the right side of the table, it can be seen, that the XGB delivered the best classification accuracy.
Having a look at the results of all subjects we found, that the XGB clearly outperformed the other classification methods, with a 100% best performance on the GWO electrode reduction methods and still above 50% at least for the remaining methods as shown in figure 5.4. A detailed view on the results shows that the XGB was chosen 45 times out of the 60 configurations resulting of the 4 electrode reduction methods for 15 participants. This equals 75 % of all the possible configurations. The NN classifier was chosen 9 times, the RF 5 and the SVM only once.
The feature extraction methods did not show such a clear picture as presented in figure 5.5. In this diagram, the feature extraction methods are listed according to their presence in the top sets of the subjects for all of the 4 electrode reduction methods. Looking at GWO the DWT can be considered as the best performing method with 7 out of 15 occurrences in the top set, followed by the standard CSP with 4 occurrences and 3 for the CSPwav. The feature vector was only chosen once and the CSPfv method not at all. For the other electrode reduction algorithms we can see a rather balanced distribution of the methods including the CSPfv.

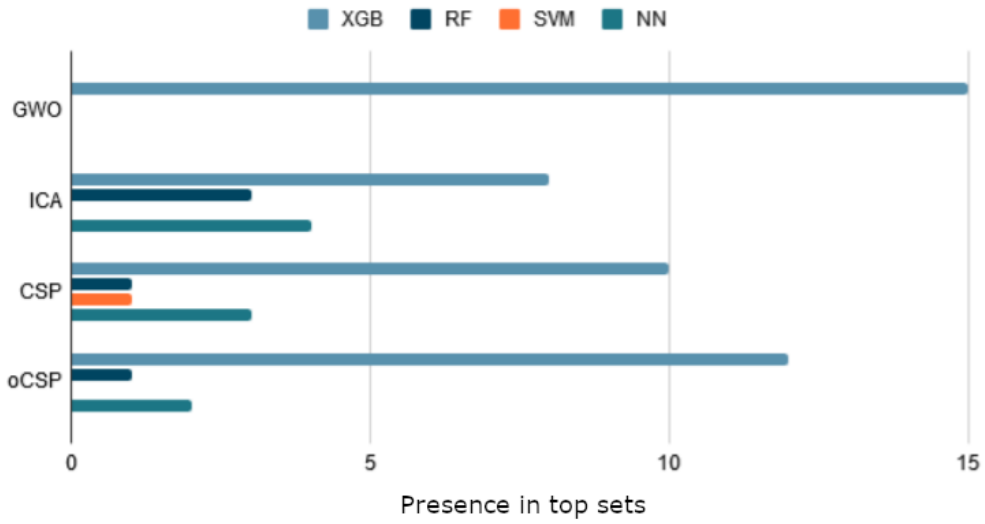

Figure 5.4: Presence of the 4 different classifiers in the top sets listed for the 4 electrode reduction methods.

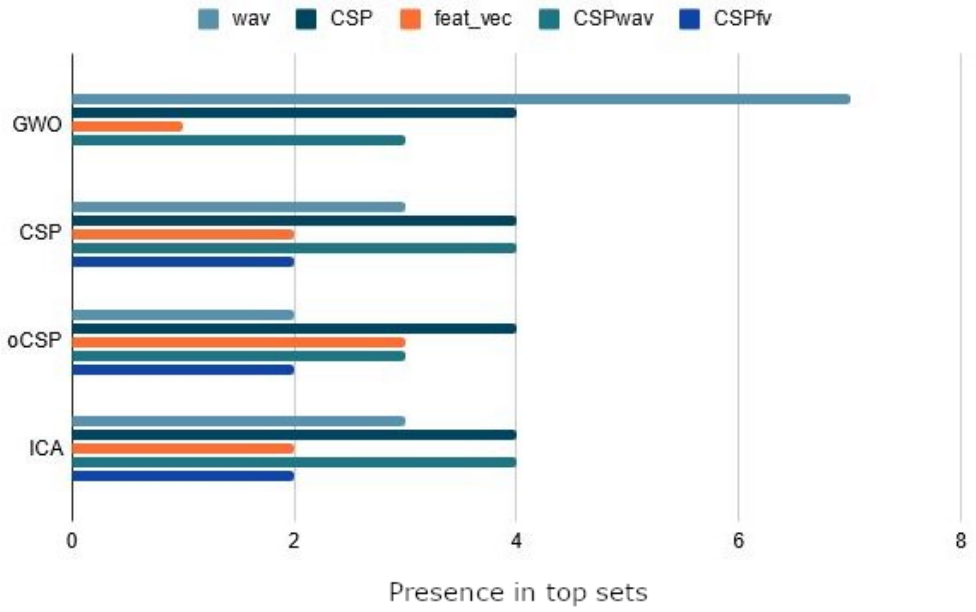

Figure 5.5: Presence of the 5 different feature extraction methods in the top sets listed for the 4 electrode reduction methods.

Discussion

In our attempt to find a best performing setup for electrode reduction in Speech Imagery BCIs, we can clearly say, that the GWO algorithm outperformed the other implemented methods for electrode reduction in our pre-study. A certain dominance of this algorithm was expected, as it evaluates the contribution of each electrode to the overall classification by calculating the accuracy for each and every subset after removing one electrode. Thus, the method is based on brute-force and examines the data set down to the smallest detail, which makes it computationally intensive but on the other hand very precise. Furthermore it showed the expected shape of the curve in figure 5.3 for classification accuracy over the number of electrodes removed. One would expect a rise in the accuracy in the beginning, as clearly not all electrodes contribute to the classification process equally well and some might be corrupted by noise or contain information not relevant for the imagined speech production at all. After a certain number of electrodes removed however, the data does not contain sufficient information anymore and the classification accuracy drops again as can be seen in figure 5.3 at around 45 electrodes removed.
Compared to the GWO the other algorithms did not show this behavior and remained in the lower classification accuracy range oscillating around a value of approximately 25% classification accuracy without a clear indication of peak or best setup. One could conclude on a peak for CSP and ICA at around 27% and 45 electrodes removed, in comparison to the clear difference to the GWO this can however be neglected. Due to the fact, that this expected behavior was not present in the ICA, CSP and oCSP electrode reduction method and the significantly worse classification accuracy compared to the GWO algorithm, we decided to focus on the GWO in our further studies on electrode reduction in imagined speech BCIs.
Concerning the classifier, the XGB clearly outperformed the other methods, especially on the GWO electrode reduction method. In this case all of the best performing configurations used this classifier as shown in figure 5.4. As mentioned previously this method calculates the accuracies for each and every subset and therefore gives us the perfect basis for the evaluation of the classifier. But also for the other electrode reduction method we can see a clear preference towards the XGB with its overall 75% appearance in the top sets without any real competitor among the other 3 classifiers. This circumstance and the fact, that we decided on GWO as electrode reduction method for further evaluation, in which XGB scored all of the top sets, we decided to proceed with this classifier in our following study on systematic electrode reduction in Speech Imagery BCIs.
The results concerning the feature extraction methods seem to support our findings from previous work [179, 180] as they are highly individual. For the GWO one could claim a preference towards the wavelet decomposition, closely followed by the CSP algorithm and the CSPwav. Taking into account the other electrode reduction methods however, there is no clear evidence for the preference of a certain feature extraction method. The occurrence in the top sets is rather balanced and it seems to depend highly on the individual. In order to preserve this individual preferences in our follow-up study, we decided to select the 3 best performing feature extraction methods of the GWO configuration namely CSP, DWT (wav) and the combination of the two, CSPwav. As the feature vector was only responsible for one top set in the GWO condition and the combination of feature vector and CSP (CSPfv) not for a single one, we decided to exclude them from further investigation.

For our follow-up study on electrode reduction we therefore conclude on the following setup:

- **Electrode reduction:** Grey Wolf Optimization (GWO)
- **Classification:** Extreme Gradient Boost (XGB)
- **Feature Extraction:**
    1. Common Spatial Patterns (CSP)
    2. Discrete Wavelet Transform (DWT)
    3. DWT + CSP (CSPwav)

5.1.3 Conclusion

This pre-study aimed at finding a best suited setup for systematic electrode reduction in imagined speech BCIs by evaluating different combinations of electrode reduction, feature extraction and classification methods. In our attempt to determine this setup we compared 4 different electrode reduction methods namely Grey Wolf Optimization (GWO), Independent Component Analysis (ICA) and the Common Spatial Pattern - Rank (CSP-rank) algorithm in two variations together with a variety of feature extraction and classification methods on an existing dataset of 15 participants performing imagined speech. The dataset was recorded with a 64-channel EEG-headset and our algorithm consecutively removed channels in steps of one based on the decision of each electrode reduction method. The intermediate classification accuracy for all implemented feature extraction and classification methods was stored and used for evaluation concerning the quality of each subset of electrodes based on the given signal processing setup. Our results show that the GWO algorithm outperformed all other electrode reduction methods and was the only one to deliver the expected behavior concerning classification accuracy with proceeding electrode reduction. Within this method, the Extreme Gradient Boosting (XGB) classifier was responsible for all the top classification accuracies of all subjects. Consequently, we suggest this classifier for the further evaluation. As the feature extraction method did not provide such a clear picture and rather supports previous findings, that this step of the signal processing pipeline is highly subject specific and needs to be tailored to the individual, we decided to include the 3 best performing methods in the follow-up study, namely Discrete Wavelet Transformation (DWT), Common Spatial Patterns (CSP) and the combination of the two wavelet transformed CSP (CSPwav). With this implementation we expect to provide a sound setup for the systematic electrode reduction in imagined speech BCIs in the best possible way and want to proceed to evaluate this method on a broader basis on different datasets in the next section.

5.2 Systematic Electrode Reduction in SI-BCIs

Our pre-study on the comparison of electrode reduction methods on imagined speech data in the previous section resulted in the Grey Wolf Optimization (GWO) method outperforming all other implemented algorithms. Despite being the computationally most expensive one, we decided to continue our evaluation on electrode reduction in imagined speech BCIs with this method. We proceeded with a systematic reduction of electrodes by applying this algorithm on 3 different imagined speech datasets with the methods as described in section 5.1.1 and the setup we concluded on in section 5.1.2. This procedure will be explained in the following.

5.2.1 Methodology

Our previous findings confirm, that a maximum number of electrodes does not always correlate with the maximum classification accuracy of a SI-BCI and therefore its performance (see figure 5.3). In order to evaluate the methods we have concluded on in the previous section (5.1.2), we will try to verify the findings in a systematic approach of reducing the electrodes of three different imagined speech datasets in an attempt to find a minimal best performing subset of electrodes for SI-BCIs. Furthermore, we will compare our results of the best performing setups with two predefined subsets including electrodes chosen based on certain functional areas of the cortex, known to be related to speech production, namely all electrodes over the left hemisphere and a subset specifically targeting Broca and Wernicke area of the brain (see chapter 2.3).

Concept

The implementation concept was taken from our pre-study as described in section 5.1.1. The only differences in this case was, that 3 different datasets were fed to the pipeline illustrated in figure 5.1 one after another. The details concerning the recording of these datasets will be explained in the following.

Datasets

All three dataset consisted of EEG-data recorded from at least 15 participants while performing imagined speech. In all studies, the EEG was recorded with 64 electrodes placed according to the 10-20 system and with the same hardware, the Brain Products Live-Amp[19].
Although the datasets were recorded by different research groups, they differed only slightly in the procedure of the recording and the number of words used as presented in more detail in the following.
Dataset 1 was taken from the 2020 international BCI competition as presented in section 5.1.1 and already used in our pre-study. It was included in this part of the study for benchmark reasons and to provide a more detailed analysis of the previously reported results. Participants repeated the 5 Korean words "Hello", "Help me", "Stop", "Thank you" and "Yes". The dataset consisted of 70 repetitions per word and was recorded as illustrated in figure 3.2. Information about the handedness of the participants was not provided in the dataset description.

[19] `https://brainvision.com/products/liveamp-64/` Last accessed: 24.10.2022

Dataset 2 was acquired in the scope of the study presented in section 3.1. It consists of the EEG-data of 17 participants while producing imagined speech of the 9 German words "screw", "case", "circuit-board", "floor", "conveyer belt", "workbench", "push", "hold" and "lift". Overall we recorded 40 repetitions per word using the study setup similar as shown in figure 3.2. More details about the recording can be found in section 3.1. For better comparability with the other two datasets, we randomly excluded two participants from the dataset and ended up with 15 participants in the final analysis. All participants were right-handed.
Dataset 3 was taken from our study as presented in section 3.2. It consists of the data of 15 right-handed participants silently repeating the five English words "up", "left", "right", "push" and "pick" to navigate a robot through a maze-like game on a computer screen. Using this setup we collected 80 imagined repetitions per word. More details on the data acquisition and the study itself can be found in section 3.2.

Although partially recorded under different circumstances like varying number of repetitions per word, the language and the words themselves, the overall important characteristics of the datasets are the same, namely, the number and position of electrodes used, the number of participants, the paradigm of imagined speech and even the recording hardware. This makes it the perfect setup for systematic evaluation of electrode reduction and important electrode positions in SI-BCIs. Important to mention at this point is, that although all of the datasets included 15 participants, those participants were not the same subjects. The data was recorded from different participants for each study.

Preprocessing

The preprocessing followed the steps of our pre-study. The data was bandpass- filtered between 0.5 and 60Hz and notch filtered again at 50 Hz to remove any overlying powerline noise. The parameters for the filtering methods were chosen according to our previous work [180] as explained in detail in section 4.2. After filtering the data was cut into epochs of two seconds starting from the onset of the fixation cross prior to silent repetition in order to reduce the signal to the relevant sections containing the imagined speech.

Feature Extraction

Based on the results of our pre-study we decided to implement the 3 best performing feature extraction methods, as they have an impact on the performance of the classifier on the data of the individual. Those were the Common Spatial Pattern (CSP) algorithm, the Discrete Wavelet Transform (DWT) and a combination of the two which we refer to as CSPWav. The CSP was realized once more using the multiclass implementation of the mne library [73] with the default parameters. The DWT was again implemented based on the PyWavelets library [121]. As mother wavelet we applied biorthorgonal 2.2 (bior2.2) as suggested in [57]. The data was decomposed until fourth level. Afterwards, a wavelet feature vector was created out of the data as presented in [207]. The CSPwav feature extraction included the combination of the previously explained methods by first applying CSP and creating the DWT of the resulting time signal.

Classification

For classification we used the Extreme Gradient Boosting implementation of our pre-study as it managed to outperform all other implemented classifiers especially in the Grey Wolf Optimization electrode reduction method, which will be used in this study setup. The XGB was again implemented based on [34] with a mean error as evaluation metric and instructed to stop if the mean error did not decrease for ten rounds. The objective function was chosen to be softmax for multiple classes.

Electrode Reduction

As the results of the pre-study showed significantly better results for the Grey Wolf Optimization (GWO) in comparison to all other algorithms, we decided to solely use this method for electrode reduction in our final evaluation. Although being computational expensive, the great difference in performance in comparison to the other evaluated electrode reduction methods were decisive for the choice of the GWO. It was implemented as presented in section 5.1.1.

Evaluation Criteria

Our evaluation criteria followed the methods of the pre-study and included the classification accuracy as performance measure given as the sum of correct classifications divided by the amount of total classifications in comparison to the chance level. As described in 2.9, the theoretic chance level of dividing 100% by the number of classes to predict, assumes an infinite number of predictions. BCI datasets however, are usually rather small due to the efforts in recording them and the chance level should therefore be adjusted according to the size of the dataset [39]. We calculated the significance threshold based on the procedure given in 2.9 and achieved values of 24.29%, 14.44% and 24.00% for dataset 1, 2 and 3 respectively.
Top performing sets were calculated using a Fuzzy Inference System implemented according to Torres et al. [206]. The purpose of this system was to prevent excluding sets with only slightly lower classification accuracy but significantly lower number of electrodes. This system makes a decision based on a set of rules to provide a good balance between classification accuracy and number of electrodes. This decision is based on the FIS membership function, which is computed from accuracy and amount of channels used, and illustrated in figure 5.6. In our implementation we used the subject's maximum classification accuracy as upper bound instead of the error rate used by Torres et al., since visual inspection of both methods yielded better results for the maximum accuracy, which led the FIS to not favoring lower channel amounts for higher drops in accuracy.
Based on the top sets defined by our fuzzy inference system we concluded on a possible best minimal number and subset of electrodes for SI-BCIs and furthermore tried to determine the most valuable electrode positions in the classification process by looking at the number of occurrence of each electrode in the top sets per participant. We furthermore applied k-means clustering on the top sets for all the data and the data split up into the three different feature extraction methods (wav, CSP, CSPwav) to investigate valuable electrode positions in relation to a certain cluster of lower or higher amount of electrodes.
Finally, we compared our results to subsets of electrodes chosen based on functional areas of the cortex related to speech production, namely, all electrodes of the left hemisphere, the right hemisphere, and electrodes targeting Broca and Wernicke. The electrodes for Broca and Wernicke were selected according to [210] as "F7" and "P3".

Additionally we included the electrodes directly adjoining those two positions in order to cover a wider variety of spatial information, resulting in "F7", "AF7", "AF3", "F5", "F9", "F3", "FT9", "FT7", "FC5", "CP5", "CP3", "CPz", "P5", "P1", "P3", "PO3", "PO8", "POz".

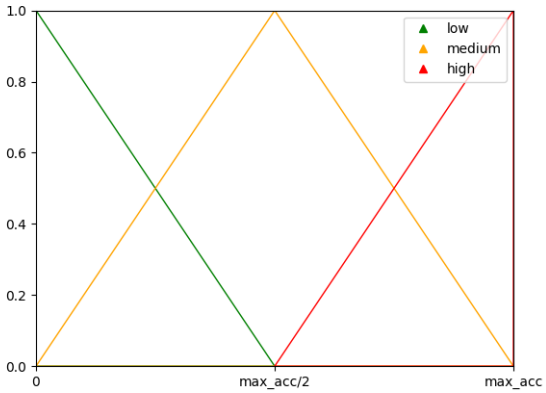

Figure 5.6: Membership functions used for the fuzzy inference system with the maximum classification accuracy as upper bound.

5.2.2 Results and Discussion

In the following we will present the results of our evaluation on the three different datasets and discuss them subsequently. Whenever we speak of the number of electrodes in this section in text and figure captions, it depicts the number of electrodes removed from the set, not the remaining ones. We will repeat this fact several times throughout the text in order to avoid misunderstandings or misinterpretation.

Results

Figure 5.7 shows the average results for all participants, the three different datasets and the three feature extraction methods, as classification accuracy over the amount of electrodes removed. All the diagrams clearly show the expected shape of an early increase of classification accuracy after reducing the first electrodes and decreasing again if too many electrodes are removed towards the end. This shape was most distinct for the discrete wavelet transform (wav) feature extraction method represented in the right column. It confirmed the success of our electrode reduction method and the results from the pre-study for the two additional imagined speech datasets.
Having a closer look on the individual results, tables 5.2, 5.3 and 5.4 show the individual top sets of each participant for the three datasets and the three feature extraction methods as determined by our fuzzy inference system. The number in brackets on the left represents the classification accuracy and the number on the right the number of electrodes removed from the original set of 64, which led to this classification accuracy. As a reminder, our fuzzy inference system selected those combinations out of all single results of the individual which does not necessarily mean, that it was the highest accuracy achieved, but rather the highest accuracy with the least amount of electrodes according to the parameters mentioned in the methodology section.

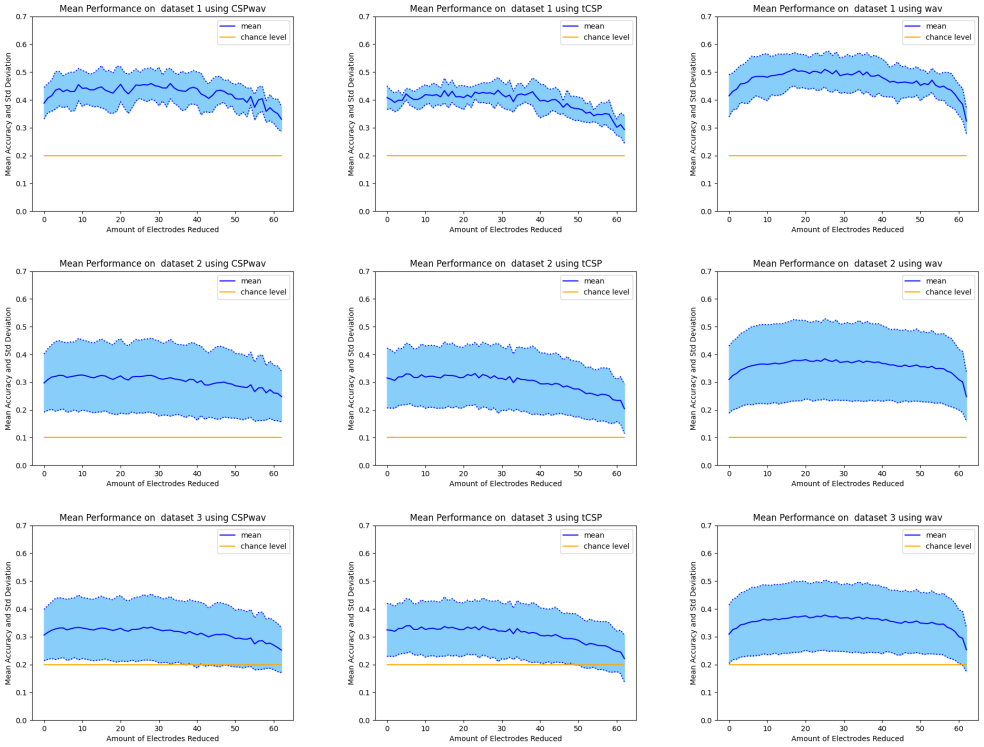

Figure 5.7: Mean classification accuracy of all participants over the number of electrodes removed. In the first row for the dataset one, second row dataset two and in the third row dataset three. The first column shows the results for CSPwav, second CSP and third DWT feature extraction method. The solid blue line represents the mean value and the yellow one the chance level.

For dataset one, 38 electrodes were removed on average and resulted in a classification accuracy of 51% being significantly above chance level with the significance threshold of 24.29%. The average classification accuracies did not differ much between the different feature extraction methods, however, the boxplots, shown in figure 5.8, confirm the superior performance of the wavelet transform (wav) method, as indicated by the 54% average classification accuracy and the fact, that in 12 out of 15 times, the highest classification accuracy was achieved with wav. The 3 remaining top sets resulted from the CSPwav feature extraction method. Having a look at the electrodes which were removed in the top sets of this dataset, we can see, that the average numbers do not differ too much as well, but the values come with a rather high standard deviation, indicating that the numbers strongly differ between the participants. Figure 5.9 illustrates this fact in the boxplots, showing a wide range of removed electrodes from 30 to 56 in the case of the CSPwav feature extraction method. Given this strong distribution and wide range of different numbers of electrodes removed, it is not really possible to conclude on a single best subset or number of electrodes which would be suitable for all participants, although the medians center around 36 electrodes removed.

Subject	CSP	CSPwav	wav	Acc Avg	Elec Avg
1	(0.44, 47)	(0.54, 37)	(0.60, 37)	0.53±0.08	40 ± 6
2	(0.42, 35)	(0.48, 48)	(0.50, 53)	0.47±0.04	45 ± 9
3	(0.46, 32)	(0.48, 49)	(0.60, 33)	0.51±0.08	38 ± 10
4	(0.56, 41)	(0.52, 33)	(0.58, 40)	0.55±0.03	38 ± 4
5	(0.52, 40)	(0.52, 34)	(0.62, 36)	0.55±0.05	37 ± 3
6	(0.48, 29)	(0.48, 33)	(0.54, 34)	0.50±0.03	32 ± 3
7	(0.50, 31)	(0.52, 36)	(0.60, 34)	0.54±0.05	34 ± 3
8	(0.42, 37)	(0.48, 30)	(0.44, 37)	0.45±0.03	35 ± 4
9	(0.56, 38)	(0.62, 32)	(0.50, 61)	0.56±0.06	47 ± 15
10	(0.44, 34)	(0.48, 38)	(0.58, 35)	0.50±0.07	36 ± 2
11	(0.50, 35)	(0.52, 56)	(0.58, 44)	0.53±0.04	45 ± 11
12	(0.48, 32)	(0.52, 46)	(0.44, 42)	0.48±0.04	40 ± 7
13	(0.42, 35)	(0.48, 48)	(0.50, 53)	0.47±0.04	45 ± 9
14	(0.50, 30)	(0.54, 30)	(0.54, 35)	0.53±0.02	3 ± 3
15	(0.48, 38)	(0.50, 30)	(0.54, 32)	0.51±0.03	33 ± 4
Acc Avg	**0.48 ± 0.05**	**0.51 ± 0.04**	**0.54 ± 0.06**	**0.51 ± 0.05**	-
Elec Avg	**36 ± 5**	**39 ± 8**	**40 ± 9**	-	**38 ± 7**

Table 5.2: Top sets as determined by our fuzzy inference system for each participant and each feature extraction method in dataset one. The number in brackets on the left represents the classification accuracy and the number on the right the number of electrodes removed from the original set of 64 which led to this accuracy.

Dataset two shows a similar behavior as dataset one. Again we see an average number of 38 electrodes removed from the initial set of 64 electrodes but with a even higher standard deviation as compared to dataset one. Classification accuracies are stable and do not differ too much with an average classification accuracy of 36%. Having a look at the boxplots in figure 5.8, we can see that although the average values let us conclude on a rather dense distribution the wavelet feature extraction (wav) outperformed the other two methods. However, given the significance threshold of 14.44% those results are all significantly above chance level for the 9 words to be distinguished. The highest classification accuracies in this case were produced 10 times by the wav feature extraction method and in the remaining 5 times by the CSP method. The boxplots for the number of removed electrodes in figure 5.9 show an even clearer picture of the broad distribution of different numbers than in dataset one. There is no clear evidence on a certain number to remove in order to receive best classification results, in this case even the median differs strongly between the different feature extraction methods and we can not determine a single best subset or number of electrodes. Distinguishing in between feature extraction methods we can see a preference of the wav method to remove more electrodes and therefore a smaller number of electrodes for classification, while the CSP-based setups seem to prefer larger numbers of electrodes.

In dataset three we can again observe similar numbers for the average values in table 5.4. Average classification accuracies lie closely together at around 39% and a low standard deviation. In the boxplot in figure 5.8 we can this time confirm the average results from the table, as the classification accuracies do not differ significantly and beside a few outliers for the wavelet feature extraction method (wav) the accuracies seem to center around the average value of 39%. Best classification accuracies were achieved 6 times by the CSP, 5 times by the CSPwav and 4 times by the wav feature extraction method.

Subject	CSP	CSPwav	wav	Acc Avg	Elec Avg
1	(0.37, 32)	(0.30, 43)	(0.40, 55)	0.36±0.05	43 ± 12
2	(0.38, 21)	(0.27, 47)	(0.36, 50)	0.34±0.06	39 ± 16
3	(0.36, 32)	(0.36, 29)	(0.41, 30)	0.38±0.03	30 ± 2
4	(0.43, 25)	(0.33, 33)	(0.33, 33)	0.37±0.06	30 ± 5
5	(0.33, 23)	(0.29, 35)	(0.37, 32)	0.33±0.04	30 ± 6
6	(0.38, 26)	(0.36, 62)	(0.45, 55)	0.40±0.05	48 ± 19
7	(0.36, 25)	(0.31, 29)	(0.38, 58)	0.36±0.03	37 ± 18
8	(0.33, 31)	(0.30, 27)	(0.40, 34)	0.35±0.05	31 ± 4
9	(0.31, 45)	(0.29, 42)	(0.47, 56)	0.36±0.10	48 ± 7
10	(0.36, 29)	(0.31, 28)	(0.34, 55)	0.34±0.02	38 ± 15
11	(0.37, 23)	(0.31, 59)	(0.40, 41)	0.37±0.04	41 ± 18
12	(0.34, 32)	(0.29, 36)	(0.33, 58)	0.32±0.03	42 ± 14
13	(0.37, 28)	(0.29, 32)	(0.45, 44)	0.38±0.08	35 ± 8
14	(0.33, 28)	(0.30, 26)	(0.40, 43)	0.35±0.05	32 ± 9
15	(0.37, 32)	(0.27, 27)	(0.34, 57)	0.33±0.05	39 ± 16
Acc Avg	**0.36±0.04**	**0.31±0.03**	**0.39±0.04**	**0.36±0.04**	-
Elec Avg	**29 ± 6**	**37 ± 11**	**47 ± 11**	-	**38 ± 12**

Table 5.3: Top sets as determined by our fuzzy inference system for each participant and each feature extraction method in dataset two. The number in brackets on the left represents the classification accuracy and the number on the right the number of electrodes removed from the original set of 64 which led to this accuracy.

Subject	CSP	CSPwav	wav	Acc Avg	Elec Avg
1	(0.41, 34)	(0.42, 41)	(0.38, 33)	0.41±0.02	36 ± 4
2	(0.36, 33)	(0.40, 31)	(0.38, 35)	0.38±0.02	33 ± 2
3	(0.40, 57)	(0.42, 31)	(0.38, 48)	0.40±0.02	45 ± 13
4	(0.37, 32)	(0.36, 35)	(0.36, 41)	0.37±0.01	36 ± 5
5	(0.41, 31)	(0.37, 28)	(0.40, 33)	0.40±0.02	31 ± 3
6	(0.38, 33)	(0.38, 42)	(0.40, 43)	0.39±0.01	39 ± 6
7	(0.36, 34)	(0.40, 42)	(0.38, 41)	0.38±0.02	39 ± 4
8	(0.38, 34)	(0.37, 29)	(0.33, 30)	0.37±0.03	31 ± 3
9	(0.37, 43)	(0.37, 53)	(0.40, 30)	0.38±0.01	42 ± 12
10	(0.40, 31)	(0.36, 35)	(0.40, 44)	0.39±0.02	37 ± 7
11	(0.40, 44)	(0.38, 48)	(0.37, 53)	0.39±0.01	48 ± 5
12	(0.43, 34)	(0.40, 42)	(0.41, 52)	0.42±0.02	43 ± 9
13	(0.41, 46)	(0.36, 30)	(0.42, 36)	0.40±0.03	37 ± 8
14	(0.40, 30)	(0.41, 33)	(0.40, 52)	0.40±0.01	38 ± 12
15	(0.37, 30)	(0.37, 33)	(0.40, 35)	0.38±0.01	33 ± 3
Acc Avg	**0.39±0.02**	**0.38±0.02**	**0.39±0.02**	**0.39±0.02**	-
Elec Avg	**36 ± 8**	**37 ± 7**	**40 ± 8**	-	**38 ± 8**

Table 5.4: Top sets as determined by our fuzzy inference system for each participant and each feature extraction method in dataset three. The number in brackets on the left represents the classification accuracy and the number on the right the number of electrodes removed from the original set of 64 which led to this accuracy.

Having a look at the number of electrodes removed we can see once more the average number of 38 but the boxplot also reveals once more a wide distribution of the numbers among the different subjects and no clear evidence on a concrete number which might be suitable for all the participants. Similar to dataset two, we can see a preference of the wav feature extraction to a smaller number of electrodes as compared to the CSP-based methods.
This impression is confirmed by the black boxplots on the right in figure 5.9, which give an overview of the the distribution of the number of reduced electrodes per feature extraction method, for the data of all three datasets. We can see a tendency of the wavelet transform (wav) to reduce more electrodes than the other two feature extraction methods. Although there appears to be a wide range and again some outliers, the interquartile range of the different methods indicates that the CSP and CSPwav feature extraction tend to use more electrodes than the wavelet transform. Furthermore, we can see that the top sets of the wavelet transform for all the participants in all datasets lie above 30 removed electrodes, which means that for this feature extraction method we could have achieved the top set configurations with only 34 of the initial 64 electrodes. This represents a significant saving of electrodes and therefore efforts in terms of setup times for imagined speech experiments. Although the CSP based feature extraction methods seem to prefer more electrodes to achieve their top sets, we can observe the first quartile for both implementations lying at 30 removed electrodes. This means that for 75% of the participants we would have achieved the top results with 30 electrodes less even for those feature extraction methods. Taking the numbers for all three methods together we can include 83% of the top sets overall, and can therefore conclude, that in our study, with those three different datasets, we could have achieved the same results with roughly half the number of electrodes.

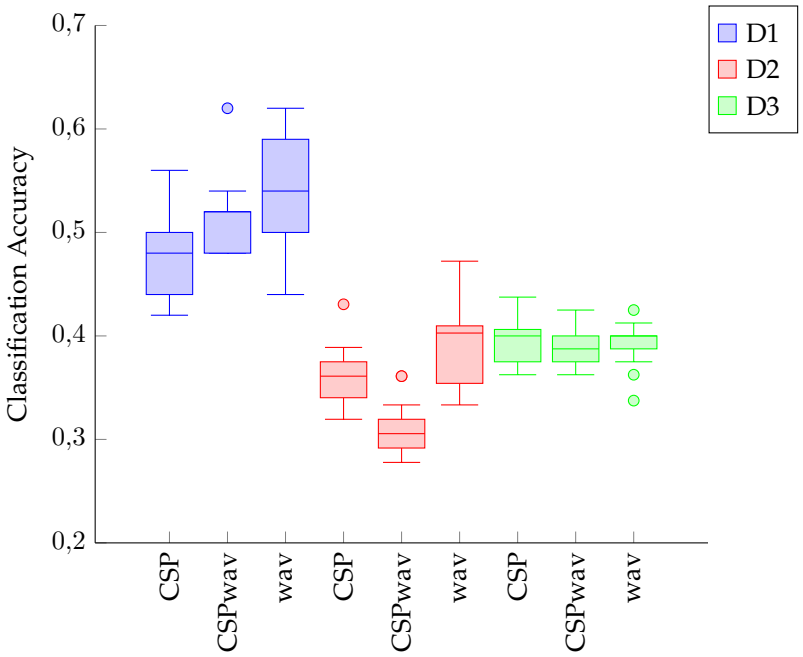

Figure 5.8: Boxplots of the classification accuracies calculated for the 3 different datasets (D1, D2, D3) and feature extraction methods.

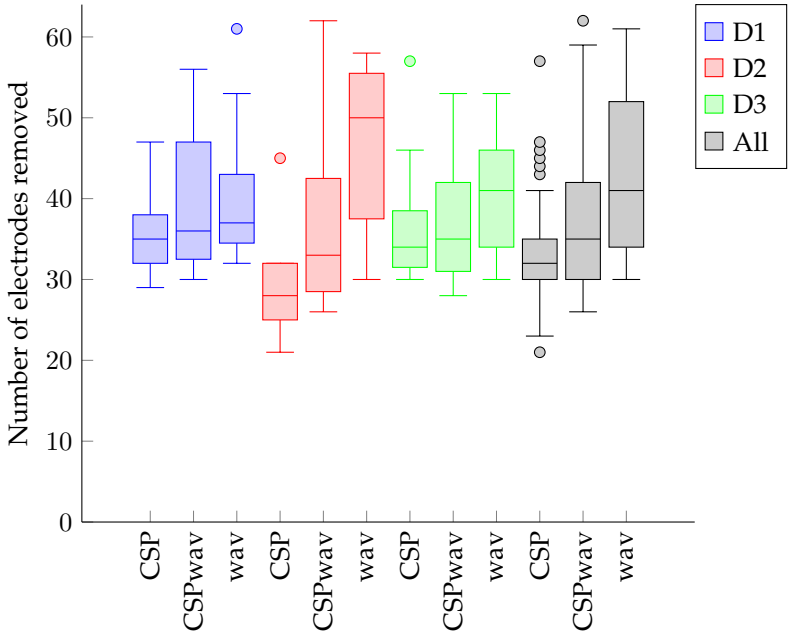

Figure 5.9: Boxplots of the number of electrodes removed for each of the three datasets (D1, D2, D3) and feature extraction methods. The black boxes on the right show the results for all datasets combined (All) separated by feature extraction methods.

In order to determine relevant electrode positions in the defined top sets, we counted the occurrences of each individual electrode in those sets. The resulting values were calculated according to the number of times the algorithm chose the specific electrode in the classification process within the top set for all participants. We decided to not only have a look at all the top sets combined but rather also cluster the sets into certain numbers of electrodes to get an impression on possible electrode hotspots in sets with higher or lower number of electrodes overall. We therefore applied k-means clustering on the top sets of all participants and within the top sets of the three different feature extraction methods. The number of clusters for each of the sets was determined by plotting the sum of the squared error (SSE) over the first 10 clusters and finding the elbow point to be at three resulting in three clusters for each of the sets of electrodes.
The results of the clustering are shown in figures 5.10, 5.11, 5.12, 5.13. For each of the figures we can clearly see a cluster for low, medium and high number of electrodes removed. Methods including the wavelet transform (wav, CSPwav) seem to create clusters around 30, 40 and 50 electrodes, while the pure CSP appears to be separable into clusters at 25, 35 and 45 once more indicating the demand for more electrodes in the classification process with those feature extraction methods.

A clear conclusion on the relevant positions of electrodes could however not be drawn. As an example we have included the results for the electrode positions in the top sets of all dataset separated by the feature extraction methods in figure 5.14. Each electrode position is visualized at its position of the head, colored according to the percentage of occurrence in the top sets as given in the legend on the right. As we can see there is a rather homogeneous distribution for all three feature extraction methods which does not allow a clear conclusion on a certain brain region being dominantly involved in the classification process.

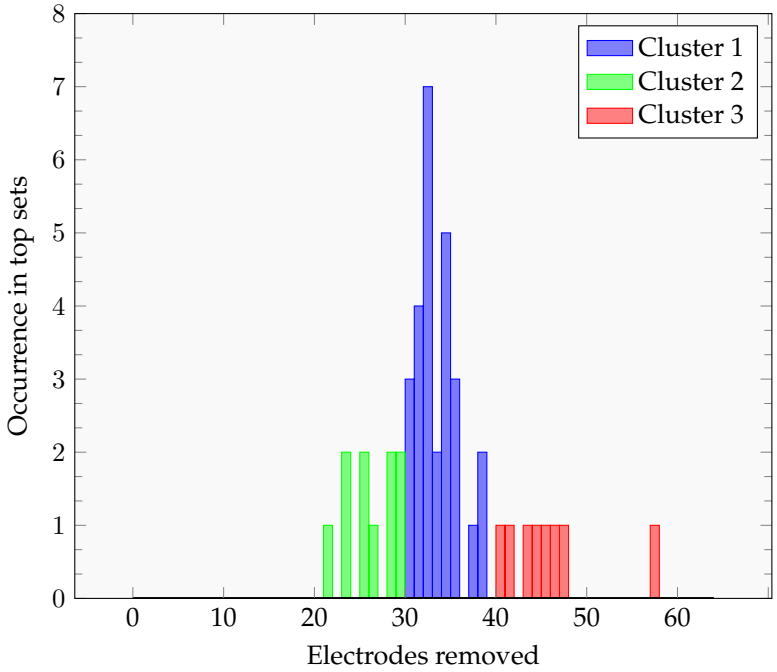

Figure 5.10: Histogram of the removed electrodes for all datasets and the CSP feature extraction method including the three clusters determined for this distribution.

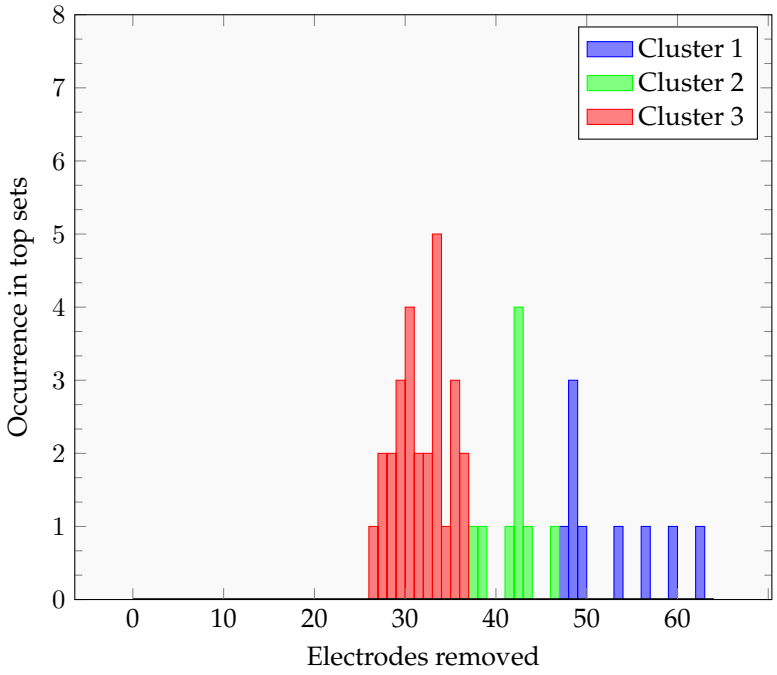

Figure 5.11: Histogram of the removed electrodes for all datasets and the CSPwav feature extraction method including the three clusters determined for this distribution.

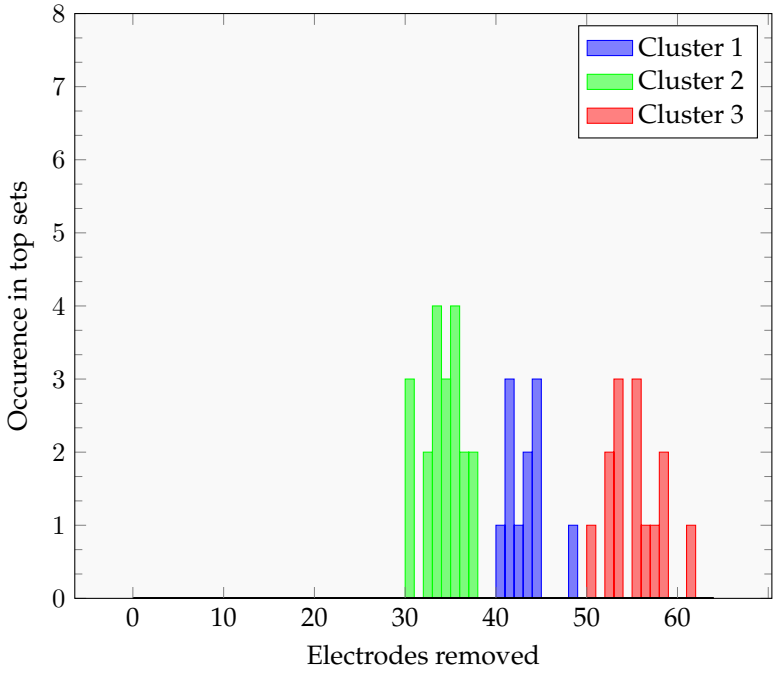

Figure 5.12: Histogram of the removed electrodes for all datasets and the wav feature extraction method including the three clusters determined for this distribution.

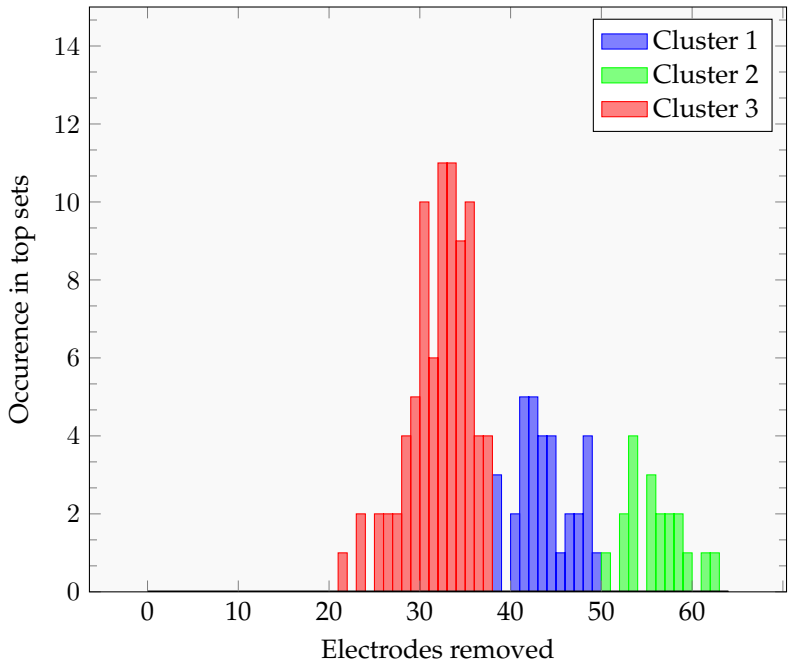

Figure 5.13: Histogram of the removed electrodes for all datasets and feature extraction methods including the three clusters determined for this distribution.

One could conclude on a more dominant role of the occipital region in the case of the discrete wavelet transform, however none of them passed the 80% threshold and the remaining electrodes are all spread rather homogeneously. This homogeneous distribution is overall represented also in the results for the other configurations which are included in the Appendix 7.4.4.

Furthermore, these findings are in line with our classification results based on certain subsets representing functional areas of the cortex related to speech production. Table 5.5 shows the average results for these subsets namely, all electrodes of the left hemisphere (Left) and electrodes targeting Broca and Wernicke area (B & W), in comparison to the average results of our Grey Wolf Optimization algorithm. For the sake of completeness we furthermore included all electrodes of the right hemisphere as subset for comparison. The numbers for the specific subsets represent the overall best value from the three different feature extraction methods used, while the GWO is presented in detail for the different feature extraction methods, in order to compare the values to the highest and lowest performance of the GWO implementation. Within this table the results clearly show, that the GWO implementation outperformed the predefined subsets of electrodes from left and right hemisphere as well as Broca Wernicke area with a difference in average classification accuracies of 20% for dataset one, 18% for dataset two and 11% for dataset three. This was not only the case for the average values but for all participants and all individual accuracies as shown in table 1 in the Appendix, clearly supporting the results that those electrode positions in the subsets are widely distributed across the cortex. This hypothesis is further supported by the fact, that classification on the electrodes of the left and right hemisphere alone, yielded equal average classification results for all datasets, highlighting that both hemispheres on average equally contributed to the classification process. Thus, we would conclude once more, that the position of the electrodes is highly subject specific and can not be generalized to apply in a cross-subject condition.

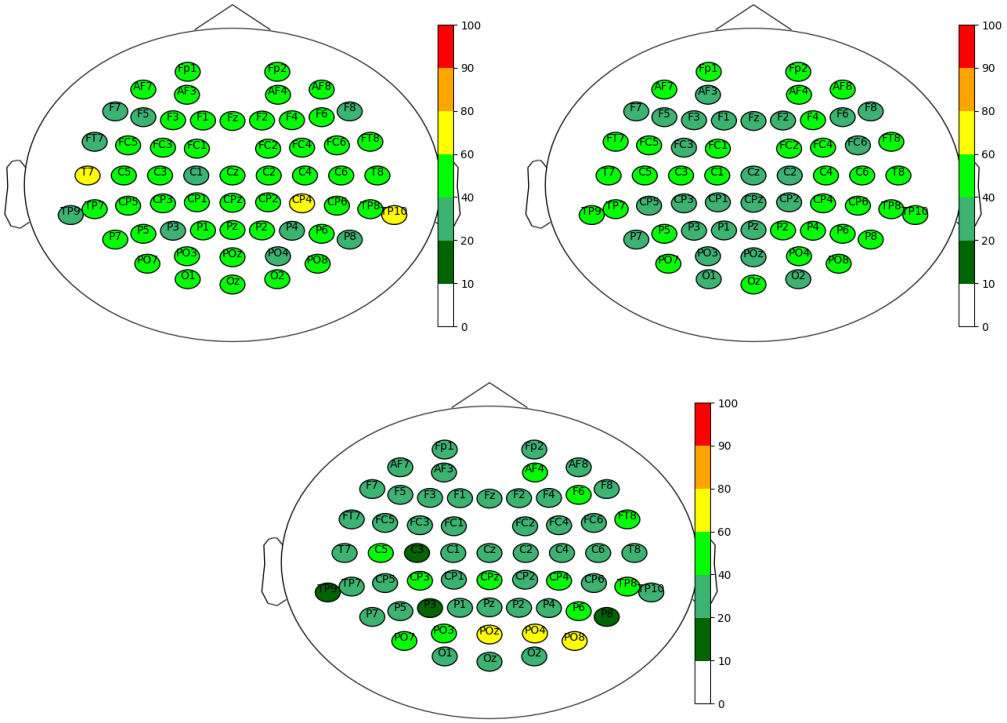

Figure 5.14: Electrode positions for the top sets of all three datasets and the three different feature extraction methods in percent. Top left: CSP, top right: CSPwav and bottom: wav.

D1	CSP	CSPwav	wav	B & W	Left	Right
Acc Avg	0.48	0.51	0.54	0.31	0.34	0.34
Elec Avg	28	25	24	18	35	35

D2	CSP	CSPwav	wav	B & W	Left	Right
Acc Avg	0.36	0.31	0.39	0.19	0.21	0.21
Elec Avg	35	27	17	18	35	35

D3	CSP	CSPwav	wav	B & W	Left	Right
Acc Avg	0.39	0.38	0.39	0.27	0.28	0.28
Elec Avg	28	27	24	18	35	35

Table 5.5: Average classification accuracies and number of electrodes, in this case remaining ones, for the top sets of the GWO, separated according to feature extraction method, CSP, CSPwav and wav. The right half of the table shows the average values for Broca and Wernicke (B & A), the electrodes of the left and the right hemisphere.

Discussion

The results as presented in the previous section showed, that it was not possible to determine one specific subset or number of electrodes which can be applied as a standard setup for EEG-based imagined speech BCIs. Although on average 38 electrodes were removed in all the three datasets, we see a strong variation among subjects concerning the number of electrodes. Having a look at the three feature extraction methods, one could conclude that if feature extraction is performed with the Discrete Wavelet Transform, less electrodes are sufficient as compared to the CSP-based methods which seemed on average to include more electrodes in the top sets for classification. A concrete number could however again not be determined as the results vary and are highly subject specific. What can be concluded concerning the number of electrodes is, that the 64 channel setup as used in all 3 studies appears to be over-sized. Figure 5.8 illustrates, that we could clearly reduce the number of electrodes for each participant by 20 without excluding any of the top sets determined by our fuzzy inference system. Furthremore, this was the case in all three different imagined speech datasets evaluated in our study. Increasing this number to 30 would still include all top sets of the wavelet feature extraction method and 75% of the two CSP based methods which shows, that EEG-based imagined speech BCIs appear to work perfectly fine with smaller setups of electrodes. Our recommendation in this case is to limit further studies in imagined speech to setups with 32 electrodes as they seem to have worked reasonably well for the 3 investigated datasets in our study. Still starting with a higher number of electrodes and systematically reducing it to the best performing subset of the individual seems necessary, as we could see a high variation in the remaining results above 30 electrodes removed and in the individual case, the number could be reduced even further.
The results on specific electrode positions did unfortunately not yield a clear picture. Although the different numbers of electrodes for the individual top sets could be clearly divided into separate clusters for low, medium and high number of electrodes, these clusters did not reveal certain positions on the cortex which could be seen as significantly contributing to the classification process. We expected to see a dominant role of the speech related brain areas as for example Broca or Wernicke, however not even in those regions we could see an increase in the number of electrodes in the top sets. Furthermore, speech related processes are usually handled by the left hemisphere of the brain in 90% of right-handed and also in about 60 - 70% of left-handed persons [31]. However, our classification results on subsets of exactly those left-lateralized brain regions were outperformed by our customly selected electrode positions by the Grey Wolf Optimization algorithm with differences of up to 20% on average for dataset one.
Investigating the different hemispheres and their occurrence in the top set, we could rather see a minimal dominance of the right hemisphere, however small enough to be neglected. The handedness of participants of dataset 1 was not available in the notes of the BCI competition, for the two other datasets all of the participants were right-handed, which should have led to a dominant presence of electrodes of the left hemisphere.
One possible explanation for those results could be, that the process of producing imagined speech and the resulting brain activity is widely spread over the cortex rather than focused on a specific area as we concluded in chapter 4. The average classification accuracies achieved with the subset of electrodes of the right hemisphere supports or hypothesis, that the most relevant electrodes can not be assigned to a certain hemisphere but are rather spread over the whole cortex, as they were equal to the results achieved with the subset of electrodes over the left cortex. This would also speak for the individual results concerning preferences on feature extraction methods and number of electrodes, which turned out to be highly subject-specific.

Finally, the main focus of this work was on reducing the number of electrodes with performance metric being the classification accuracy and therefore the methods were tailored to electrode reduction not source localization. Our analysis followed a systematic step-wise reduction according to the classification accuracy independent from the overall contribution of this electrode or specific areas it belongs to. A more detailed look on source localization independent from channel reduction might therefore be a next step for future work towards the reduction of electrodes in EEG-based imagined speech BCIs. Concerning the dataset selection for our evaluation we stated the differences in between them in the methodology section and showed, that they are as similar as possible, considering that they were recorded from different research groups. All three studies used the same headset, the same electrode layout, the same recording configuration and the same paradigm, imagined speech, in their studies. The studies differed in the language of the words, being Korean for dataset 1, German for dataset 2 and English for dataset 3, notably English by non-native speakers. While dataset 1 and 3 used 5 words for silent repetitions, dataset 2 used 9 which has an impact on the chance level, however for the reduced number of electrodes we could not see a difference in our results. Dataset 1 and 2 used short-block repetitions of the word in a row in a standard training scenario with words to repeat shown on the screen, while for the recording of dataset 3 the participants repeated the words only once and had a certain task given to move a robot around in a maze with their input by imagined speech. Taking into account the average number of electrodes removed from the three datasets, which was the same for all of them, and the same wide variation and distribution of electrodes as shown in the boxplots in figure 5.9 we do not see any evidence for differences between the sets which would make them not comparable. However, having a closer look at the average classification accuracies it remains suspicious, that the 5-class dataset recorded with random stimulus presentation (dataset 3) preformed worse on average than the 5-class dataset using short-block presentation (dataset 1). The fact that dataset 2, including 9 classes also recorded with short-block, produced equal results as the dataset 3 with only 5 classes, strengthens the doubts we have in the short-block stimulus presentation, that we have already addresses in section 4.3. We can once more see a clear preference of the short-block datasets towards Discrete Wavelet Transformation, while dataset 2 with random stimulus presentation showed a balanced distribution between feature extraction methods with no clear preference. Our conclusion on this kind of stimulus presentation is, that the blocks can induce detectable patterns in the EEG data leading to label leakage benefiting the classification process and increasing classification results. Those effects are less strong than in a complete blockwise presentation and do probably not appear in all participants, however, we do conclude on a moderate improvement, also supported by the wide distribution of results and larger deviations for those datasets, especially for the wav feature extraction method. Those findings are preliminary and are claims that we can not prove with the current study setups. In order to evaluate the effects of this stimulus presentation paradigm appropriately, we suggest a similar setup to Porbadnigk et al. [169] to shed light on this issue in the future.

5.2.3 Conclusion

With our work on electrode reduction on imagined speech data in this section, we tried to find a certain subset of electrodes best suited for EEG-based imagined speech BCIs. In our attempt to find this set we systematically reduced the number of electrodes during classification of imagined speech for three different datasets including the data of 15 participants each. We had a look at three different feature extraction methods and used Grey Wolf Optimization for electrode reduction. A fuzzy inference system determined the top sets out of the single individual results for all electrode configurations based on classification accuracy in relation to the electrode number. Those top sets were compared for all the subjects of the three studies and the three feature extraction methods used. A single best set of electrodes could not be defined, as the number of electrodes turned out to be highly subject specific for each of the datasets. However, we found that for 83% of the determined top sets, 30 electrodes could have been removed indicating that the currently common high resolution setups with 64+ electrodes are most likely oversized and a comparable performance of the imagined speech classifier could have been achieved with half of the electrodes as well. As our attempts to find specific electrode positions which contribute most to the imagined speech classification process did not show a clear picture and the positions were rather homogeneously spread in between the different subjects, it remains unclear if a standard 32 channel EEG with electrodes placed according to the 10-20 system would deliver the same results. Our recommendation resulting from this work is therefore to start with a higher number of electrodes and a wide and homogenous distribution over the cortex and determine the best suited electrodes based on a measurement with the individual. Future work regarding this topic should address this issue, as the given work had a look at 3 rather extensive datasets which required the participants to record several hours of imagined speech recordings. A more focused view on the contribution of each electrode during an online experiment could be used to reduce the electrodes consecutively and evaluate the contribution of each electrode on the fly in a more rapid manner. Furthermore, the here presented results should be verified by comparing the recording of a single subject with a high resolution EEG-headset and the reduced number of electrodes determined for the participant in a separate recording.

5.3 Summary

As we have seen in our studies in chapter 3, and the majority of work on imagined speech in the literature [92, 93, 124, 191, 235, 110], EEG-based Speech Imagery BCI research commonly uses high resolution devices with 64+ electrodes. In this chapter we have elaborated on the possibility to reduce the number of electrodes, as such setups usually come with extensive preparation times, higher costs and reduced user comfort. A possible future real-world application of this technology requires a reduced number of electrodes and also research would benefit from a conclusion on a reduced set of electrodes relevant for imagined speech studies, as it would allow more comfortable and less time-consuming study setups.
In the following we will summarize and discuss our results on electrode reduction for Speech Imagery BCIs and highlight the contributions as well as the limitations of our work.

5.3.1 Overall Results and Discussion

Within our first pre-study we compared different methods for electrode reduction on imagined speech EEG data, as existing research on this topic is rare and mainly focuses on specific regions of the brain, as for example the whole left hemisphere or Broca and Wenicke area, known to be related to speech processing. A systematic approach on reducing electrodes over the whole cortex however, was missing and we therefore started with comparing different methods for electrode reduction in combination with a variety of feature extraction and classification methods on imagined speech EEG data, in order to find the best suited configuration for further investigation of the topic. The pre-study showed that the Grey Wolf Optimization (GWO) outperformed the other algorithms and was chosen to be best suited for electrode reduction in combination with the Extreme Gradient Boosting (XGB) classifier. The results on the feature extraction methods appeared to be highly subject specific, supporting our findings in chapter 4, but led us to include three feature extraction methods in the follow-up study which performed best on the GWO and XGB configuration, namely, Commen Spatial Patterns (CSP), Discrete Wavelet Transform (DWT) and a combination of CSP and DWT (CSPwav).
In our follow-up study on the systematic electrode reduction in Speech Imagery BCIs, we applied this configuration on three different imagined speech datasets. The datasets were all recorded with the same 64-channel EEG headset by different research groups however, and each set included the data of 15 different subjects while producing imagined speech. Our setup continuously removed electrodes from the initial set of 64 and recorded the classification accuracy for each step together with the electrodes removed. A fuzzy inference system was used to determine the top sets, representing the set with the least number of electrodes possible while remaining a high classification accuracy. Those top sets were then further evaluated in order to find a best possible subset of electrodes and compared to the performance of three standard subsets covering the electrodes of the left and the right hemisphere as well as the electrodes surrounding Broca and Wernicke area, both known to be related to speech processing. On average 38 electrodes were removed from the initial 64 and the results showed that the CSP-based methods seemed to prefer higher number of electrodes while the wavelet transform could also achieve the top results with far less electrodes. Having a look at the results in detail we found that for the wavelet transform we could have removed 30 electrodes without excluding any of the top sets while for the two CSP-based approaches still 75% of top sets would have been included.

This lets us conclude, that EEG-based imagined speech setups do not necessarily require this large number of electrodes to be efficient and that instead of the 64+ channels, a 32 channel setup would most likely be sufficient.
Our results on the position of those electrodes did however not show a clear picture. We counted the occurrences of the different electrode positions inside the top sets and tried to find patterns in the visualized data to conclude on positions with highest contributions. We compared those plots for the different feature extraction methods and applied k-means clustering to group the data into sets of low, medium and high numbers of electrodes, but there was no clear tendency towards a certain set of electrodes or even a certain area. The distributions looked rather homogeneous which was supported by the classification results on the subsets of left and right hemisphere as well as Broca & Wernicke area, which were outperformed by our individually selected subsets with the GWO algorithm.
Thus, we conclude once more, that those positions are highly subject specific and should be tailored to the individual.

Summing up our results we wanted to answer the following research questions within this chapter:

> **RQ 3.1** Is there a single best minimal set of electrodes for imagined speech BCIs?
>
> **RQ 3.2** Can we determine certain electrode positions related to good imagined speech classification accuracies?

Concerning **RQ 3.1** we were not able to find one specific minimal subset of electrodes which can be used for EEG-based imagined speech classification. The results vary strongly between subjects but we could see a tendency towards a maximum number of roughly 34 electrodes below which 83% of all the top performing configurations could be achieved. Within the wavelet transform feature extraction, even all the top sets were achieved with less than 34 electrodes and with an average value of 40 removed electrodes. Thus, only 24 electrodes remained on average in the top sets, which clearly speaks for a preference of the method towards lower numbers of electrodes. Overall we conclude that the currently used high-resolution setups with 64+ channels are over-sized for imagined speech detection and that also smaller setups with half the number of electrodes can achieve comparable results.
Concerning **RQ 3.2** we could once more see a highly subject specific behavior in our results and we could not find certain brain regions or electrode positions which contributed most to the classification process. Figure 5.14 illustrates the homogeneous distribution of the electrodes in the top sets among the single subjects. We therefore once more conclude that, just like for feature extraction, the electrode positions should be tailored to the individual.

Further conclusions and recommendations are summarized in the following as contributions and limitations of the findings in this chapter on electrode reduction in EEG-based imagined speech BCIs.

5.3.2 Contributions

Within this chapter we provide several contributions to the field of EEG-based imagined speech BCIs arising from our presented results.
First we could show in our pre-study, that the Grey Wolf Optimization appears to be best suited for electrode reduction on imagined speech EEG data. Although computationally expensive, due to its nature of an evolutionary algorithm based on the idea of the "survival of the fittest", the algorithm clearly outperformed the other methods used (ICA, CSP, oCSP). We could show the expected behavior of a rise in classification accuracy after removing a first set of electrodes due to exclusion of channels with no contribution to the imagined speech classification, and the drop after removing too many electrodes, resulting in removing too much valuable information from the classification process. In further research on this topic we would therefore clearly recommend this algorithm to reduce electrodes.
Secondly, the Extreme Gradient Boosting (XGB) classifier showed its potential in the classification of imagined speech on EEG data, especially in combination with the Discrete Wavelet Transform (DWT) feature extraction algorithm. The XGB classifier was responsible for all the top set classifications in our pre-study and the combination of XGB and DWT for 54% of the top sets in our follow-up study on the systematic electrode reduction on three imagined speech datasets. These results make it a clear recommendation for further use in Speech Imagery BCI research and application.
Finally, we were able to show, that the currently used setups of 64+ electrodes in imagined speech classification are most likely over-sized and that a comparable classification accuracy can be achieved with around 34 to 32 electrodes as well. For a real-world application but also for further experiments in this field these findings will allow to reduce the number of electrodes drastically by almost a half and therefore save setup times and costs significantly.

5.3.3 Limitations

Our strongest contribution, that imagined speech EEG recordings can be conducted successfully with half the amount of electrodes, comes with a limitation however. As we could not see clear patterns in those reduced sets of electrodes, we can not conclude on specific positions or groups how to reduce those electrodes. All of our datasets were recorded with 64 electrodes placed according to the standard 10-20 system which foresees a widely spread homogeneous distribution of the electrodes over the cortex.
By systematically reducing the number of electrodes based on classification accuracy and occurrence in top sets, independent of their specific location, we loose the original homogeneous distribution of the 10-20 system and proceed with an individual distribution of the electrodes. Our results therefore still depend on the layout of a 64-channel setup based on the 10-20 system but with a reduced number of electrodes on this specific layout. The results can therefore also not be directly transferred into a 32 channel standard 10-20 layout but remain to be determined based on a initial measurement with the 64 channel headset as mentioned previously.
The overall homogeneous distributions as seen in the electrode plots on the other hand, could lead to the conclusion, that a standard 10-20 setup with 32 electrodes might have led to the same results, but remains to be verified in a direct comparison of a 64 and 32 channel measurement of imagined speech for the same subject.

As one of our conclusions is once more, that the analysis should be tailored to the individual, we also have to add once more the limitation, that the data evaluated in our experiments has been recorded from single participants on single days. A closer look on the data of the individual over a certain period of time, could help to further investigate the rather widely spread patterns of electrode positions and if they remain constant also within the data of the same subject over time.

Chapter 6
Discussion

The research field of imagined speech Brain-Computer Interfaces has seen a rapid development over the last decade, showing the potential of classifying imagined words from brain activity, even with non-invasive measures as for example the EEG. However, state-of-the-art research and implementations concerning Speech Imagery BCIs faces major challenges in three different fields:

Nr 1: **Training:** Exhausting and inefficient training scenarios.

Nr 2: **Classification:** Insufficient accuracy and number of words distinguishable.

Nr 3: **Usability:** Cumbersome setups inappropriate for real-world application.

Within this thesis we aimed at providing solutions to those problems by developing new methods and concepts for the implementation and use of EEG-based imagined speech BCIs. Those works and our goals can be summarized in accordance with the three aforementioned problems as follows:

Nr 1: Train an imagined speech classifier during interaction, while silently reading a text and overtly speaking.

Nr 2: Include semantic categories detection into the classification process to improve the overall classification accuracy of imagined speech BCIs.

Nr 3: Reduce the number of electrodes needed for imagined speech detection and find a minimum subset of electrodes required for Speech Imagery BCIs.

In the following, we will present the discussion on our results from those three domains and put them into perspective in the context of the overall topic of the thesis, Speech Imagery BCIs. We will further more provide a discussion on Speech Imagery BCIs in general and the insights we gained throughout our studies. We will close the chapter by discussing ethical issues that might arise from the presented work.

6.1 Speech Imagery BCI Training procedures

One major roadblock of Speech Imagery BCIs being used in real-world applications, is the tedious training sessions in which participants have to collect training data for the imagined speech classifiers. Due to the repetitive nature of these sessions and the tremendous amount of required training data, procedures are mentally and physically exhausting. Within chapter 3 we have introduced two novel concepts of collecting data for imagined speech classification which shall make training procedures more engaging and less tedious.
The first concept is based on the idea, that imagined speech is in principle similar to the process of silently reading. Thus we designed a natural reading task with 9 target words embedded, which should later be used for imagined speech classification. The EEG activity of participants was recorded and synchronized with eye-tracking data during reading the text, to label EEG data based on the word the participant was reading. This activity was then used in an offline analysis to classify the 9 target words based on the EEG activity recorded during the natural reading task. In a second step this classifier trained on reading was transferred to EEG activity recorded while the participants silently repeated the same words in a standard imagined speech task (RQ 1.1).
The classification of words read achieved accuracies significantly above chance level and we could show for the first time, that it is possible to classify words read based on EEG activity during a natural reading task. With an average classification accuracy of 20.82%, we significantly exceeded the chance level of 11.11% with the significance threshold being at 16.66%. Taking into account that the classifier was only trained on a certain subset of words of the text, and the overall rather low classification accuracy, a unimodal approach of classifying words read based on EEG activity seems outside the realm of possibility with the current implementation. A multimodal approach however, in which an EEG-based classifier supports an eye-tracking solution in detecting the words a user was reading in a text, could be suitable in a future application.
The targeted transfer approach in which the classifier trained on reading data was applied on imagined speech data, did not yield the expected results and showed values around chance level. We can foresee several reasons for those resuts.
First of all, natural reading involves heavy eye movement, which is usually avoided during imagined speech recordings. This movement leads to artifacts in the EEG data overlaying the actual brain activity data. Those artifacts can be filtered to a certain extend, however, without additional channels measuring eye movement, possibilities are limited. A stronger focus on the filtering of eye movement artifacts, possibly based on the eye-tracking data and calculation of electric field models induced by eye movement, might improve results in the future.
Second, selecting suitable time intervals for cutting the EEG data remains a challenge. In our current implementation we have investigated several methods to epoch the EEG data in several length and positions around the fixation point to avoid problems arising from parafoveal view [189]. Still, our implementation required epochs of equal length for all words due to preconditions with the selected feature extraction methods as mentioned in section 3.1. Another improvement on the method might be to cut the EEG data precisely according to the reading times of the single words. Choosing different feature extraction methods independent of temporal components might make this possible and improve on the transfer condition.
Finally, and probably most relevant, the chosen setup of a natural reading task appears to be too different from the imagined speech setup to evoke similar brain patterns. Inside a text, words are presented in a certain context surrounded by other words and not

isolated as in the word-based Speech Imagery BCIs. Beside the previously mentioned parafoveal view, this aspect might make brain activity while reading a text vanish between words, to a rather connected pattern related to the sentences and their meaning. Making the paradigms more similar could improve the transferability of the classifier and the achieved results. On the one hand, using short sentences as imagined speech paradigm might be more similar to natural reading, while presenting the words one after another in the reading condition, in some sort of Rapid Serial Visual Presentation (RSVP) paradigm, would make it more similar to our word-based imagined speech BCI condition. In any case we see foresee the transfer from reading to imagined speech as promising, as it would allow for more engaging training scenarios with words embedded in basically every text suitable for the given scenario. Furthermore, knowledge can be transferred to the user by embedding the target words into a description about the later use of the BCI, therefore enabling a bidirectional knowledge transfer from man to machine and vice versa.
Our attempt of training an imagined speech classifier based on EEG activity recorded during overtly speaking words was successful (RQ 1.2). The transfer of a classifier trained on EEG data recorded while speaking 5 words to a dataset recorded while participants imagined speaking the same 5 words delivered classification accuracies significantly above chance level and, although not equal, but a performance close to the standard training approach. As the results are preliminary there is however still room for improvement on the method. In our current setup the data analysis is done offline and the implementation of the Common Spatial Pattern algorithm for feature extraction applied on the whole test and train set separately. Although valid in terms of evaluation of the developed classifier, this method can not be applied in an online classification in which data is provided step-wise in pieces for each interaction. The CSP algorithm needs to be adapted to handle those smaller pieces of data by applying the spatial filters, created on the train set, on each incoming piece of data. We do not foresee drastic changes in classification accuracy, nevertheless, this approach needs to prove feasibility in the given scenario.
With the presented approach of training while speaking, we provide a better controllable, more engaging way of training imagined speech classifiers in which the output of the participant can actually be verified. This paradigm will allow to train a SI-BCI during interaction with a system by recording EEG activity while speaking commands and train an imagined speech classifier simultaneously in the background. Thus not only making the training procedure more engaging but also more productive as the user can interact and use the system while training.

Summing up our results in chapter 3 we can answer our research question

> **RQ 1.1** Can EEG activity recorded while reading certain words be used to train a classifier to detect those words during imagined speech?

with no. With our current implementation as presented in 3.1 we were not able to transfer a classifier from reading to imagined speech.
Our second research question

> **RQ 1.2** Can EEG activity recorded while speaking certain words be used to train a classifier to detect those words during imagined speech?

could be answered with yes.

Our newly developed method to train an imagined speech classifier based on data recorded during overt speech was successful and achieved classification accuracies significantly above chance level for all our participants.

Addressing our overall **RQ1**, if we can automatically train SI-BCIs based on other modalities, we mangaged to provided preliminary results clearly indicating that it is possible.

6.2 Speech Imagery and Semantic Classification

In our attempt to improve classification accuracies and to increase the number of words distinguishable in Speech Imagery BCIs, we had a closer look on semantic classification of imagined words. We introduced the concept of a Semantic Silent Speech BCI, which tries to classify the category of a silently spoken word prior to the word itself. The goal of this approach is to decompose the original classification problem on a word level into a smaller number of classification problems based on the semantic category of the word (see section 4.3). Although in our case we applied the concept on imagined speech, it is not limited to this paradigm and could theoretically be applied to other forms of silent speech as well, therefore we chose the name Semantic Silent Speech BCI. The idea is based on the work of Huth et al. [88], who showed, that semantic processing is spread over the whole cortex, creating complex patterns of brain activity, which should be detectable even with brain measures with reduced spatial resolution, as for example the EEG.
In our first step we could verify the results of Huth et al., originally achieved with fMRI, in a EEG study. We showed that 5 different semantic categories can be classified from EEG activity during imagined speech (RQ 2.1), with classification accuracies significantly above chance level. Those results extend the previously existing work in the field of semantic category detection with EEG, which usually only focuses on two categories, living and non-living [197, 153, 206]. The common patterns of brain activity between subjects, as shown in the fMRI study, could however not be reproduced. Our attempt to classify cross-subject with the Common Spatial Pattern (CSP) feature extraction method did not yield results above chance level. Our implementation based on a time-frequency feature vector on the other hand, did show results above chance level, which let's us conclude, that there might still be common patterns of brain activity shared between subjects during semantic processing, which could just not be found by the CSP algorithm. Given that our analysis was based on classification results, having a closer look on the actual visualized patterns of brain activity and improved methods for source localization could shed light on this issue in an advanced neuroscientific analysis of the underlying processes in the future.
In an applied scenario of a Semantic Silent Speech BCI, we would still recommend to follow a within-subject analysis and tailor the classification process to the individual. The same holds for feature extraction and classification methods in general, as our results showed highly individual preferences between subjects. We compared different combinations of feature extraction and classification methods in order to provide a conclusion on a single best setup for imagined speech classification. The results did however show a rather homogeneous distribution of the different methods among the participants, which let's us conclude, that those parameters should be determined for the individual prior to usage of such a system.

Our second research question was concerned with the actual implementation of the Semantic Silent Speech BCI and the performance in comparison to a standard word based Speech Imagery BCI (RQ 2.2). We let the two classifiers compete on two datasets recorded during our work on improved training procedures for SI-BCIs as presented in chapter 3.
Our Semantic Silent Speech BCI achieved an average improvement of 1.9% for dataset 1 with a single best improvement of 6.66%, and an average improvement of 5.13% for dataset 2 with a single best improvement of 8.31%. Furthermore, our approach showed higher classification accuracies for the majority of participants in both datasets, and even all participants in dataset 2, highlighting the success of the Semantic Silent Speech BCI and its potential to improve classification accuracy even further in the future.
An overall worse average classification accuracy for the semantic categories, as compared to the results in sections 4.1 and 4.2, might result from the recording procedure of the two datasets, which was not focusing on the Semantic Silent Speech aspects but rather included additional tasks in between the imagined speech parts as presented in chapter 3, which might have had an impact on the brain activity of the participants an therefore the overall classification accuracy.
Furthermore, the achieved accuracies are still far from feasible for real world applications and the concrete effect on consequently increasing number of words needs to be investigated in the future. Finally, the possible effect of short-block presentation, and the impact on classification accuracy, need to be evaluated in further studies, as we could observe a comparable performance on the 9-class dataset recorded with the short-block presentation and the 5-class dataset recorded with random presentation. This issue will be further discussed in section 6.4, covering our general findings in the field of Speech Imagery BCIs.

Summing up our results in chapter 4 we can answer our research question

> **RQ 2.1** Can semantic categories be classified from EEG activity during imagined speech production?

with a clear yes. Our results in section 4.1 and 4.2 showed, that this is possible for 5 different semantic categories with classification accuracies significantly above chance level.
Our second research question

> **RQ 2.2** Can semantic classification increase classification accuracies in EEG-based imagined speech BCIs?

could also be answered with yes. Our newly developed Semantic Silent Speech BCI approach delivered throughout better classification accuracies on two datasets as compared to a standard word based classifier.

Addressing our overall **RQ2**, if semantic category detection of words can be used to improve the decision making process in SI-BCI, we could clearly show improved performance of our Semantic Silent Speech BCI in comparison to a standard approach and therefore the potential of the method.

6.3 Speech Imagery and Electrode Reduction

In our attempt to find a minimal subset of electrodes for imagined speech classification, we compared several electrode reduction algorithms in a first step. In our evaluation on an imagined speech dataset the Grey Wolf Optimization (GWO) algorithm outperformed all other tested algorithms namely, Indepentend Component Analysis (ICA), Common Spatial patterns (CSP) and a CSP-rank implementation. GWO resulted in highest classification accuracies and was the only algorithm to show the expected pattern of a rise in classification accuracy after removing the first few electrodes, unrelated to imagined speech production, while dropping after removing too many, thereby discarding too much valuable information. Despite being computationally expensive the algorithm can be highly recommended for electrode reduction in imagined speech BCI applications.
The subsequent analysis on systematic electrode reduction for Speech Imagery BCIs involved three imagined speech datasets, each recorded with 64 channels, on which we consequently applied our GWO algorithm implementation. We used a fuzzy inference system to determine the best possible subset for each participant with the lowest number of electrodes while retaining a high classification accuracy. On average 38 electrodes were removed from the initial 64 electrodes in all three datasets, showing a clear sign, that the commonly used 64 channel setups are over-sized. Given a high standard deviation and the wide distribution of the number of electrodes removed in between subjects, we could not conclude on one single best minimal set of electrodes for Speech Imagery BCIs (RQ 3.1). The number of electrodes required for reliable imagined speech classification appears to be as individual as the preference for feature extraction methods found in chapter 4. Differences were also found in between the two feature extraction methods. We observed that the Discrete Wavelet Transform (DWT) managed to produce high classification accuracies also with a smaller number of electrodes while the Common Spatial Pattern (CSP) algorithm preferred larger numbers of electrodes. With the DWT implementation the top sets with highest accuracies were throughout achieved by removing at least 30 electrodes for all participants in all the three datasets. Given that DWT was resonsible for 75% of the top classification accuracies in our analysis of all datasets we conclude that a setup with 32 channels is in most cases sufficient for imagined speech BCIs especially when using DWT as feature extraction algorithm (RQ 3.1). Concerning the question if there are certain areas or electrode positions which contribute most to the classification process of an imagined speech BCI, we were not able to find specific locations with our methods (RQ 3.2). We could once more see a highly subject specific behavior and a homogeneous distribution of electrodes with highest impact on the classification between participants and no highlight in specific brain regions. This was even more surprising, as two of the datasets contained only data from right-handed participants, where speech related processes should mainly trigger activity in the left hemisphere of the brain. But also our attempts to classify on subsets of electrodes solely involving electrodes from the two hemisphere separately and speech related regions, resulted in throughout worse classification accuracies as compared to the subsets determined by the GWO algorithm. This supports the impression, that the activity most relevant for imagined speech classification can not solely be found in the left hemisphere of the brain, but is rather a widely spread network distributed over the whole cortex. Given the poor spatial resolution of the EEG we can not make clear claims on the concrete underlying neural processes and break with established concepts, as for example the dominance of the left hemisphere of the brain in speech production. However, given our results on subsets of electrodes, delivered by our GWO implementation and based on classification accuracy, we could not determine clear patterns, which leads us to the

conclusion to once more tailor the methods for imagined speech classification to the individual.
Nevertheless, our approach delivered useful insights on reasonable numbers of electrodes highlighting the currently over-sized setups with 64+ channels. Starting with a calibration measurement with 64 electrodes in a first training session, would allow to determine the best possible individual subset of electrodes with our GWO implementation. This reduced number of electrodes could then be applied for further continuous use of the SI-BCI and the specific user. Such a procedure can however only work, if the subset of electrodes remains its performance over several sessions, a general aspect which applies to all of the previously discussed methods on semantic category detection and novel training scenarios. This multi-session issue will be discussed in detail in section 6.4 on our findings concerning SI-BCIs in general.

Summarizing our results in chapter 5 we can answer our first research question

> **RQ 3.1** Is there a single best minimal set of electrodes for imagined speech BCIs?

with no. In our evaluation on three different imagined speech datasets the number of electrodes responsible for top performance varied between subjects. However, we could achieve individual top performances in 83% of the cases with 30 or more electrodes removed from the initial configuration, which indicated, that Speech Imagery BCIs can be realized with a significant reduced number of electrodes as it is currently common practice.
Our second research question

> **RQ 3.2** Can we determine certain electrode positions related to good imagined speech classification accuracies?

had to be answered with no as well. Electrode positions in the top sets of the participants varied greatly and were homogeneously spread over the cortex rather than focused on certain areas. This impression was confirmed by our classification results on subsets of electrodes targeting the left hemisphere which throughout performed significantly worse than our GWO sets.

Addressing our overall **RQ3**, if we can make SI-BCIs more user-friendly by reducing the required number of electrodes, we could clearly show that common 64+ channel setups are most likely oversized and would recommend to switch to 32 channel setups for imagined speech classification, which should provide similar results, but with a more user-friendly setup.

6.4 EEG-based Speech Imagery BCIs

Our work on new methods targeting the improvement of EEG-based Speech Imagery BCIs, delivered a variety of valuable insights and new findings with respect to the three core fields of this thesis, improvement of classification accuracy, training scenarios and usability of such systems. Discussing EEG-based Speech Imagery BCIs and the current state of the technology more generally, we can start with confirming, that it is possible to classify imagined words from EEG data. Our results on existing and self-recorded datasets showed classification accuracies significantly above chance level for up to 9 words distinguished at a time. We foresee this way of establishing a BCI, by just speaking from the mind, as the most intuitive and natural way of brain computer interaction. With our work we were able to alleviate some of the major roadblocks on the way of this technology to real-world application. Our works on semantic classification, improved training scenarios and electrode reduction pave the way to more applied implementations of SI-BCIs and provide valuable insights and novel methods, which will enable real-world application in the future. Speaking about real-world application, there are however several remaining issues to be addressed.
All of our studies, as well as the vast majority of existing related studies, were conducted in laboratory environments. Real-world environments set different requirements involving artifacts induced by noise of different types, e.g acoustic, electromagnetic and many more. Especially user movement will be an issue due to distortions in the EEG signals caused by the usually stronger electric signals of muscle movement detected by the EEG. However, filter and feature extraction methods are evolving and will be able to cope with those artifacts in the future, probably in combination with additional sensor data, as for example gyroscopes for movement measurement. Furthermore, context information could be included in the classification process and provide some situation-based auto-adaption or -correction to relax the strict conditions and support decision making, maybe also in combination with other modalities in a Hybrid-BCI approach.
Another important next step is the transfer of the developed methods from an offline, to an online analysis. All of our analysis, and again the vast majority of related work, was done offline. Most of our methods can be applied to an online scenario without further adaption and should work with comparable performance in such a setting. However, algorithms as for example the Independent Component Analysis (ICA) for filtering of the data can only hardly be applied in real-time. Thus, standard filtering methods will have to prove feasibility in an online scenario. The same holds for some of our feature extraction methods. In case of the Common Spatial Pattern (CSP) algorithm, our implementation as presented in chapter 3.2, can not be used in real-time as it processes test and train set in complete blocks. An adaption to use the produced spatial filter masks of the CSP on single incoming data snippets, should theoretically perform equally well, but needs further evaluation. Furthermore, extending the currently established machine learning methods for feature extraction and classification with neural network approaches, might leverage the technology even further. There are existing approaches on Deep learning methods in Speech Imagery BCIs [186, 146], however, the state-of-the-art presents better performance achieved with standard machine learning methods as for example CSP and SVM [122, 125, 126]. One reason for this issue might be the usually small datasets, which do no provide the sufficient basis for deep learning. This aspect could be solved in the future with the help of our improved training methods, as presented in chapter 3.

Speaking of established methods, we do see problems in stimulus presentation during imagined speech experiments. A common practice as presented mainly by Lee et al. [123, 125, 126], includes the presentation of words in short blocks. In this procedure, the target word is presented once and the subject advised to repeat it subsequently a certain number of times. This type of stimulus presentation saves time by reducing the overall target word presentation, which usually takes several seconds per repetition, summing up to several minutes over the duration of a whole experiment. The problem occurring with pure block-wise presentation and the classification of arbitrary brain states as discovered in [169], is avoided by using only few repetitions in short-blocks and presenting them randomly throughout the procedure of the experiment. However, we claim, that this type of presentation still induces temporal effects, benefiting the later classification process by label leakage, due to several reasons. First, in all of our analysis on datasets recorded with short-block presentation, we could experience a significantly better performance of the Discrete Wavelet Transform (DWT) for feature extraction. This feature extraction makes use of a frequency analysis of the signal, in which the effects mentioned in [169] have the strongest impact. In our work on electrode reduction in chapter 5 including three datasets of which two were recorded with short-blocks and one with random presentation, the random presentation showed comparable results between the CSP and DWT feature extraction. For the datasets with short-block presentation we could observe clearly better results with DWT as compared to CSP. In chapter 4, we compared two of the datasets with a focus on the classification of semantic categories of the words. One of those datasets was recorded with short-blocks, the other with random stimulus presentation. For the short-block presentation, the DWT outperformed the CSP method with better classification accuracies for all participants in 98% of the investigated configurations. The dataset recorded with randomized stimulus presentation on the other hand, showed a more balanced occurrence of the feature extraction method with an overall better performance with the CSP.
Second, the reported results of Porbadnigks study [169] showed slightly better results for the short-block presentation as compared to the random presentation. As both of the methods scored below significance threshold and due to the purpose of the paper to highlight the negative effects of block-wise stimulus presentation, this observation was however not further evaluated in the paper.
Lastly, the classification accuracies achieved in chapter 4 and 5, during our comparison of the different datasets, showed similar classification accuracies for the 9-class and the 5-class dataset. This is surprising, as one would expect the classifier to perform worse on the 9-class dataset, when applying the same signal analysis methods. Still, they ended up with a comparable clasification performance, and once more, the 9-class classification problem used short-blocks, while the 5-class problem used random presentation.
All the aforementioned issues and observations let us conclude, that the short-block presentation method might lead to similar effects as block-wise stimulus presentation, however, with a more moderate effect due to some randomness remaining in the paradigm, but still benefiting the later classification process. A profound conclusion can however not be made, as the observed effects might arise from differences in the recording sessions or other factors as well. Therefore, we suggest to investigate this topic in the future with a study setup similar to [169] and systematically compare effects of the different stimulus presentation methods on the performance of the classification process.
Concerning a recommendation on a best possible setup for the implementation of EEG-based Speech Imagery BCIs, we would therefore avoid short-blocks in study setups and use completely randomized stimulus presentation instead. Although the DWT showed promising results for feature extraction, we can not clearly exclude the possibility of the influence of the stimulus presentation paradigm and would suggest, to tailor the

methods to the individual. Within our studies, the CSP and DWT showed the most promising results and should be considered for comparison on the individual data. Regarding the classifier, we observed throughout better classification accuracies with the Extreme Gradient Boosting (XGB) algorithm, outperforming established methods as for example Support Vector Machines (SVM) or Random Forrest (RF), making it our clear recommendation for future use in EEG-based imagined speech detection.
Nevertheless, these recommendations need to be taken with a grain of salt, as our analysis was done on data recorded on a single day and mostly within-subject. Another big challenge to be addressed in the future, not only in imagined speech but BCIs in general, is the inter- and intra-subject variability. Inter-subject variability is a common issue in BCI applications and describes the problem of variations in brain activity between subjects [56]. These differences lead to classifiers being trained on data of the individual, resulting in applications which need to be retrained for every new user. Collecting larger datasets of multiple users might help to overcome those challenges by applying deep learning algorithms on those datasets and find similar patterns which could be used for classification processes between users. But also the data of the individual varies, e.g. in between sessions. This phenomenon is referred to as intra-subject or inter-session variability [184] and leads to problems with training a classifier on the one day and applying it on the other, even on the same subject. Depending on the condition of the users, the brain activity might differ, for example if they slept well on day one, when training the system, and had a hard night before using it on day two. Those changes lead to significant drops in classification accuracies in between sessions.
Although this problem is known in the BCI community, it is only poorly understood. What are the sources causing variability and how these sources affect EEG signals and BCI control, is still mostly unknown, and yet to be solved. In order to do so, research needs to systematically collect data in multi-day experiments to unravel responsible factors and measures to cope with those changes in brain activity in future research.

In summary, Speech Imagery BCIs offer a natural and intuitive way of brain computer interaction, and have shown to be feasible even based on EEG measurements. However, most research is still done in controlled laboratory environments and single session recordings. Addressing this issues in the future, by moving study setups to more applied online multi-session experiments, will provide crucial insights on challenges ahead, to finally bridge the gap to a successful implementation and use in a real-world scenario of EEG-based Speech Imagery BCIs. Successfully decoding imagined speech from brain activity provides a tremendous potential in Brain Computer Interaction, on the other hand, ethical questions arise with those possibilities, which will be discussed in the next section.

6.5 Ethical Issues

Imagined speech is often considered as an approach to read a user's mind. With such a technology ethical concerns arise and some might see the freedom of individual thoughts of each human endangered, creating dystopic scenarios in which e.g. a government would be able to control the thoughts of the people by this technology. Critical thinking and recapitulating ones own work from a variety of perspectives, is an important part of research and ethical issues are a major concern in all kind of BCI applications.
We would like to start with a reflection of the studies conducted in this work, the processing of the collected data and the measures taken to protect privacy of the data as well as to ensure ethical standards. We will furthermore reflect on the possibilities arising with the collected data, given the current state-of-the-art in BCI technology, in order to put the potential of this technology and possible threats, as well as ethical questions resulting from it, into perspective. Finally, we will discuss a broader view on ethical issues within research and possible future applications of Speech Imagery and BCIs in general, and measures taken to handle this critical topic appropriately in the future.

6.5.1 Ethical Issues and Measures taken in this Thesis

Within the scope of this work we have conducted several studies in which we recorded brain activity via EEG of healthy participants. Concerning the privacy of the data, we have taken the greatest care, already in the stage of planing and designing the studies, to value the later participants privacy, by avoiding to measure data which can only hardly be anonymized, as for example speech or video recordings, and to record the least amount of personal data required. During the experiment we have collected demographic information of the participants as for example, gender, age and handedness, which were used for the later analysis, but allow for a complete anonymization of the data without any direct reference to a specific person. Contact tracing forms were used due to the ongoing Covid-19 pandemic, however, they were not linked in any way to the recorded data and destroyed 2 weeks after the participation. The resulting datasets of the studies are therefore completely anonymized, without any personal data left, beside the measured brain activity. Furthermore, all studies took place inside the facilities of the German Research Center for Artificial Intelligence (DFKI) and the recorded data has neither left DFKIs infrastructure nor the network.
All participants were informed prior to the study about the procedure of the experiment, the hardware used, and the non-hazardous nature of the EEG and other recording techniques. The participants all signed written consent, that they have been informed about the study, are participating by free will and agree to the use and publication of the anonymized data in scientific context. Additionally, our studies were approved by the ethical review board of the Universtiy of Saarland [20].
With all the aforementioned measures, we claim that we have taken the greatest care valuing the participants privacy and applying to all ethical standards. Nevertheless, the recorded brain data is personal data, one might consider it the most personal and private data to be measurable from a human being. It is therefore important to elaborate on the possibilities that arise with this measured data in the scope of the current stat-of-the-art in BCIs, but also with an outlook on possible future applications.

[20] `https://erb.cs.uni-saarland.de/` Last accessed: 04.12.2022

6.5.2 Ethics in Present and Future BCI Research

Regarding the current state-of-the-art in BCI research we can clearly say, that the data, as it is available within our databases, is completely anonymized and does not allow any assignment to a specific person. Brain patterns or features which allow for the classification of an individual based on recorded EEG data is to the best of our knowledge not possible. Although brain data varies between human beings, it mainly consists of certain oscillating patterns without specific individual traces, which would allow for a precise tracking of the individual.
Addressing the concerns on "mind reading" and the possibility to read the thoughts of a person against their own will, we hope to have provided a holistic overview, as well as detailed insights, into the research on Speech Imagery BCIs within this work. This overview should illustrate, that such dystopic scenarios are, with the current state-of-the-art, outside the realm of possibility. All of the presented methods require the active and willing participation of the user to train a system before using it, and will not allow to "read the mind" of a person by just putting an EEG on the head and measure the brain activity. Furthermore, current applications manage to distinguish several words recorded during controlled sessions looking for the optimal measurement conditions, far from a generalizable approach for continuous decoding of imagined or silently spoken words of a broader vocabulary.
Thinking about future applications of imagined speech BCIs and BCIs in general it is important to consider the streaming of brain data to the cloud as a likely development. In order to provide sufficient computational power and mobile solutions for brain recording at the same time, the signal processing and classification will have to be outsourced to the cloud, and the devices for recording will just forward the data for processing. If this data is stored in the cloud it will be available to train deep learning algorithms, increasing the chances of being able to generalize brain activity models, and use BCIs without prior training. A bright and promising future which will allow to overcome the constraints of current tedious training procedures in BCI applications, however, it brings back the fear of misuse of the technology. With the increased availability of brain data in the cloud and further advances in machine and deep learning algorithms, a common agreement on appropriate ways of storing brain data in databases becomes essential. There are ongoing discussions in the community about establishing standards for storing data, and a foundation was established, with the goal to protect the human rights of all people from the potential misuse or abuse of neurotechnology. The Neurorights foundation is working to incorporate five specific Neuro-Rights into international human rights law, national legal and regulatory frameworks, and ethical guidelines stated on their website[21] as follows:

1. Mental Privacy - Any NeuroData obtained from measuring neural activity should be kept private. If stored, there should be a right to have it deleted at the subject's request. The sale, commercial transfer, and use of neural data should be strictly regulated.
2. Personal Identity - Boundaries must be developed to prohibit technology from disrupting the sense of self. When neurotechnology connects individuals with digital networks, it could blur the line between a person's consciousness and external technological inputs.
3. Free Will - Individuals should have ultimate control over their own decision making, without unknown manipulation from external neurotechnologies.

[21] `https://neurorightsfoundation.org/` Last accessed: 04.12.22

4. Fair Access to Mental Augmentation - There should be established guidelines at both international and national levels regulating the use of mental enhancement neurotechnologies. These guidelines should be based on the principle of justice and guarantee equality of access.

5. Protection from Bias - Countermeasures to combat bias should be the norm for algorithms in neurotechnology. Algorithm design should include input from user groups to foundationally address bias.

It has to be emphasized, that these rights target future BCI applications, which will need to rely on personal data for continuous classification in a real-world use. As the law includes the right to have the data deleted and explicitly states that the use should be strictly regulated, we can see those kind of adaptions as reasonable for future real-world applications. Such applications should take care to convey to those Neuro-Rights in order to ensure the safe and responsible handling of neurological data, and brain activity in particular, as it is one of the most private and personal data we can record from a human being. Therefore, any research measuring this data should be aware of the risks arising with future development in this field, and carefully think through the design of study procedures and the handling of the recorded data.

Chapter 7
Conclusion

In this last chapter we give a conclusion on the work done in the scope of this thesis starting with a summary of the overall purpose and scientific research questions that were answered, followed by the concrete contributions to the field of EEG-based Speech Imagery BCIs. As all the provide insights and results come with certain limitations, we will elaborate on those with regard to the contributions and conclude the thesis with an outlook on future work.

7.1 Summary

Within this thesis we wanted to contribute to the field of EEG-based Speech Imagery BCIs by providing new concepts and methods to solve some of the biggest roadblocks of this technology, which prevent it from being used in real-world application scenarios. The existing problems to be solved were categorized in three blocks concerning mentally and physically exhausting training procedures, insufficient classification accuracies and cumbersome setups with inconvenient high-density electrode EEG-headsets.
Driven by the experiences during former experiments on imagined speech data collection, we wanted to overcome the mentally and physically exhausting training procedures of Speech Imagery BCIs and developed two novel training procedures. The first approach presented in chapter 3.1 is based on the similarity of the concept of imagined speech to silent reading and we wanted to answer the research question RQ 1.1, if EEG activity recorded during reading certain words can be used to train a classifier and detect the same words during imagined speech. In our effort to answer this question we designed a text with 9 repeatedly embedded words and recorded EEG and eye-tracking data simultaneously from participants reading this text silently. The classifier that was trained on the recorded data during reading did not significantly exceed chance level when applied on imagined speech data of the same 9 words silently spoken by the same subject, which lead us to the conclusion that natural reading and imagined speech are not similar enough for a direct transfer of the EEG data between the two paradigms. Within the chapter we provided several possible adjustment of the method which might lead to a successful transfer of this promising training technique in the future.

The second approach presented in chapter 3.1 is based on the similarity of the neural correlates of spoken and imagined speech and we wanted to answer the research question RQ 2.2, if EEG activity recorded during speaking certain words out loud can be used to train a classifier and detect the same words during imagined speech. We developed a study procedure in which participants had to control a virtual robot in a maze-like game through a factory, once by overtly and once by silently speaking 5 command words. With classification accuracies significantly above chance level and even close to the standard imagined speech classification process, we could answer RQ 2.2 with yes. Training during speaking offers a promising way of improving Speech Imagery BCI training procedures to be more engaging and productive in the future.
Problems concerning classification accuracy come along with the oftentimes insufficient number of distinguishable command words and usually decreasing classification accuracy with increasing number of words. While binary classification tasks manage to provide satisfying results for imagined speech detection from EEG data, they are usually insufficient for communication, which is actually the main purpose of those types of BCIs. Our suggestion to overcome this problem was presented as Semantic Silent Speech BCIs, which try to classify the semantic category of a word before proceeding to classify the word itself. The development of this concept was presented in chapter 4 starting from the classification of semantic categories of words during object based decision tasks (chapter 4.1), followed by the actual classification of the categories during silent repetitions of those words. Within these two studies we could show for the first time, that it is possible to classify 5 different semantic categories of imagined words from EEG activity, which was until that point only researched within fMRI studies. Our research question RQ 2.1, if semantic categories can be classified from EEG activity during imagined speech production could therefore be answered with yes.
In the second step we wanted to make use of this semantic classification by implementing a Semantic Silent Speech BCI and classify the semantic category of a word prior to the word itself. Within our evaluation on two datasets containing EEG data recorded during silent repetitions of words with different semantic categories, our newly developed semantic classification approach outperformed the standard word-based approach for all subjects which enabled us to answer our research question RQ 2.2, if semantic classification can increase the classification accuracy of Speech Imagery BCIs, with yes. With this novel paradigm we provide the foundation for more precise and robust Speech Imagery BCIs in the future by including the semantic categories of words in the classification process.
Based on the datasets collected during the extensive studies in our previously mentioned work, recorded with 32 and 64 channel EEG-headsets, we wanted to elaborate on the possibility of reducing the number of electrodes needed for the implementation of an EEG-based Speech Imagery BCI in chapter 5. RQ 3.1 was therefore concerned with finding a minimal subset of electrodes, which can be used for imagined speech classification. Our evaluation included two of our datasets and one publicly available dataset, all recorded with 64 electrodes. We systematically removed electrodes from the classification process with Grey Wolf Optimization and determined the subset with highest classification accuracy and lowest number of electrodes, based on a fuzzy inference system. Although the results varied greatly between participants and we were not able to determine one specific subset, the numbers showed, that on average 38 electrodes were removed from the initial 64 and that the Common Spatial Patterns (CSP)-based feature extraction methods seemed to prefer higher number of electrodes, while the Discrete Wavelet Transform (DWT) could also achieve the top results with far less electrodes. Having a look at the results in detail we found that for the DWT we could have removed 30 electrodes without any decrease in performance, while for the two CSP-based ap-

proaches still 75% of top performing sets would have been included. Hence we conclude, that EEG-based imagined speech setups with usually 64 electrodes are oversized, and that reliable imagined speech classification can be achieved with half the amount of electrodes
Furthermore, we could show, that the positions of electrodes varied between participants and that we were not able to detect certain brain regions, which contribute most to the classification of imagined speech, as stated in RQ 3.2. The positions of the electrodes in the best subsets for each participant showed a homogeneous distribution of the electrodes over the whole cortex, without a clear indication on certain high impact areas with major influence on the classification process. With the given methods we can and will not question existing common agreement and understanding on neural correlates of speech production as presented in chapter 2.3. However, we could confirm the often mentioned widely distributed brain activity, emphasized by newer models on speech production in the brain.

In the following we will summarize the concrete contributions of our work to the research field of EEG-based Speech Imagery BCIs.

7.2 Contributions

As presented in section 3.1, we developed a new concept for Speech Imagery BCI training based on EEG and eye-tracking data recorded during natural reading. The approach is supposed to facilitate the usually mentally and physically exhausting training procedures of Speech Imagery BCIs and train a classifier simultaneously while reading a text with the later SI-BCI target words repetitively embedded. Our first attempt using a natural reading task could not achieve satisfying classification accuracies, however, with an adjustment of either the reading task to be more similar to the imagined speech part, e.g. by Rapid Serial Visual Presentation (RSVP) for reading, or an adjustment of the imagined speech part to use short command sentences and be more similar to natural reading, we can see a large potential in this method for engaging and productive Speech Imagery BCI training in the future.
In section 3.2, we presented our newly developed concept for EEG-based Speech Imagery BCI training during spoken speech, which has the potential to facilitate Speech Imagery BCI training procedures and make them more user friendly as well as productive, as the BCI can be trained automatically on the fly while interacting with a system. Our developed method showed classification accuracies significantly above chance level and even delivered values close to a standard training approach. This evaluation shows the potential of the method to facilitate Speech Imagery BCI training in the future and automatically train imagined speech during overtly spoken interaction with a system, making it a more engaging and productive process.
Concerning the improvement of classification accuracies, we provided a completely new concept for imagined speech classification with the Semantic Silent Speech BCI, which has the potential to significantly improved classification accuracies of Speech Imagery BCIs in the future. The integration of semantic categories into the classification process will change the way, command words are selected and grouped for interaction and provides completely new possibilities to design imagined speech BCIs. The fact that our evaluation against a standard word-based imagined speech classification resulted in throughout better performance of our newly developed approach, highlights the potential of our method, to increase classification accuracies and therefore potentially the number of simultaneously detectable words of such BCIs in the future.

With our research on electrode reduction in EEG-based Speech Imagery BCIs, we determined the Grey Wolf Optimization as a promising method. Beside being computationally expensive, the algorithm outperformed the other methods in our study and is a clear recommendation from the results of this work, when it comes to reducing electrodes in Speech Imagery BCIs.
Furthermore, our investigation of three imagined speech datasets recorded with 64 electrodes showed, that those setups as they are commonly used in Speech Imagery BCI research, are most likely oversized. For all 3 datasets our fuzzy inference system determined comparable classification accuracies with half of the electrodes around 32 to 34, which is a clear indication that imagined speech detection does not require time consuming and costly high-resolution setups with 64+ electrodes. These insights will help researchers in the future to avoid cumbersome high resolution setups and reduce the number of electrodes, therefore saving time and effort during their experiments.
Overall this work provides a comprehensive overview on different methods for EEG-based Speech Imagery BCI training and classification procedures providing valuable insights for future applications. Within our work, we have conducted 4 studies and collected 4 extensive datasets with 15 - 20 participants and EEG recordings of 32 - 64 channels. Based on those studies we could draw conclusions and recommendations on best practice for establishing Speech Imagery BCIs and summarized them in chapter 6.4. Furthermore, we have provided a holistic overview on the research field of EEG-based Speech Imagery BCIs in 2.5 and provided recommendations based on the state-of-the-art and our own experiences for a best possible setup of EEG-based imagined speech BCIs in chapter 6.4. We highlighted the importance of random stimulus presentation and the individual feature method selection for potential users and identify the Extreme Gradient Boosting (XGB) algorithm as promising classifier for imagined speech detection, outperforming established methods as for example Support Vector Machines (SVM) or Random Forrest (RF). These insights can act as foundation for researchers in the future during design and analysis of their experiments on imagined speech.
All the aforementioned contributions will help to improve the performance and usability of EEG-based Speech Imagery BCIs, and thus leverage the successful transfer of this technology into real-world applications in the near future.

Although providing a long list of valuable contributions, research usually does not come without any limitations, and so does our work.

7.3 Limitations

Starting with our results on the newly developed training methods for Speech Imagery BCIs in chapter 3, we have to emphasize that those results are preliminary, based on one study setup with a limited number of participants. The methods need to prove their feasibility on a larger number of participant in the future to show applicability in the targeted area of Speech Imagery BCIs. Extending the dataset with more recordings from several days and participants would finally lead to a bigger data corpus and also allow for an analysis with deep learning methods. The current implementations are based on established machine learning techniques trained on the data of the individual. Hence, a deep learning approach might provide new insights on the possibilities of cross-subject classification, given a sufficient amount of data, and improve classification accuracies in general.

Concerning our work in chapter 4 on semantic category detection, the results are based on machine learning methods and follow an engineering approach. The fact that our cross-subject classification with Common Spatial Patterns did not exceed chance level does therefor not allow for a clear conclusion on the active regions or similarity in the brain patterns between subjects, and should be further investigated including a holistic neuroscientific analysis of the recorded EEG data in the future.
The Semantic Speech Imagery approach was able to improve the classification accuracy for all the subjects of the different datasets in comparison to the standard approach. However, with an average classification accuracy of 37% and 39% the results of the newly developed method still remain below the required performance for real-world application. Furthermore, a clear conclusion in how far this improvement could cope with an increase of simultaneously detectable words is questionable. The method is still in its infancy and needs to be investigated further on larger datasets and with more fine tuning on the classification and feature extraction algorithms.
Regarding our work on electrode reduction in chapter 5, our strongest contribution, that imagined speech EEG recordings can be conducted successfully with half the amount of electrodes, comes with a limitation. By systematically reducing the number of electrodes based on classification accuracy and occurrence in top sets, independent of their specific location, we loose the original homogeneous distribution of the 10-20 system and proceed with an individual distribution of the electrodes. Our results therefore still depend on the layout of a 64-channel setup based on the 10-20 system but with a reduced number of electrodes on this specific layout. The results can therefore not be directly transferred into a 32 channel standard 10-20 layout but remain to be determined based on an initial measurement with the 64 channel headset.
The overall homogeneous distributions as seen in the electrode plots on the other hand, could lead to the conclusion, that a standard 10-20 setup with 32 electrodes might have led to the same results, but remains to be verified in a direct comparison of a 64 and 32 channel measurement of imagined speech for the same subject.
Additionally, all of our experiments were conducted in rather controlled scenarios inside the lab with environmental influences reduced to a minimum. The developed methods still have to prove their feasibility in environments of their targeted application, which will definitely rise further challenges, as real-world application of this technology comes with several factors influencing the performance of the system. Starting with environmental factors possibly having a negative effect on the quality of the recorded data and demand for different artifact filtering methods, followed by human factors including certain mental states which might affect the classification process, for example if too many wrong predictions lead to frustration of the user and change brain activity patterns unconsciously. Changes in mental states are a general problem, which leads us to next limitation, single day experiments.
All data recordings of a single experiment or study in our work were conducted on one day. Beside the immediate effect of frustration or excitement of the user as a reaction to outputs of the SI-BCI system, brain activity and its recording depends on a variety of parameters that might change between different recording sessions on different days. Starting again with environmental factors and the placement of the electrodes, ensuring perfectly reproduced electrode positions in between sessions is almost impossible, in spite of the standardizes positioning procedures as for example the 10-20 system (see section 2.1.1). Data processing pipelines often include spatial filtering and the following classification process might experience complications, if the data is recorded on different positions in between session or if one of the electrodes has slightly different conductivity.

Additionally, the cognitive state of the user might change and therefore certain background activity, which depends on the mental and physical condition. A lack of sleep or stress can significantly affect cognitive capabilities and therefore the resulting brain patterns during BCI tasks.
Future research needs to take into account this variability in brain data and our plans for future work are aligned with the previously mentioned limitations.

7.4 Future Work

Based on the limitations as stated in the previous section we see the potential for future work in various directions. This future work is split up according to the different topics of this thesis into work on training procedures, the semantic classification, electrode reduction and general topics concerning SI-BCIs.

7.4.1 Training Procedures

Concerning our newly developed training scenarios we still see room for improvement in the case of the reading condition. We plan to make the actual application scenario more similar to the natural reading during training and will try to classify short command sentences instead of single words. Those sentences will be embedded in the text and can be formulated as commands towards the system the user wants to interact with, thus making the training and application scenario more similar. Classifying whole sentences is a rather new approach and especially in the field of EEG-based Speech Imagery BCIs the focus is still on single word repetitions. The challenges ahead might be strong variation in the formulation of the command sentences as for example in reading speed. In how far such variations have an impact on the recorded data needs to be evaluated. In any case there is still a lot of potential in this method enabling bidirectional knowledge transfer during training and the increased user comfort as compared to standard training methods.
The developed training by speaking scenario delivered promising results. However, in the current condition we solely rely on the EEG data recorded during speaking and imagined speech independently. We plan to continue our research with a hybrid approach in which the classifier will be trained with data from both conditions and thereby include brain patterns from both paradigms, spoken and imagined speech, to further increase the transferability between the two concepts and therefore performance. Furthermore, we will investigate the effect of noise on the developed concept. The main benefit of the technology will be situated in noisy environments, where audio-based speech classification is not applicable and communication can be established with imagined speech. In how far imagined speech is affected by environmental noise however, remains a subject of research as well. We will therefore consequently transfer and evaluate the developed method in a real-world context with realistic environmental influences as for example acoustic but also electromagnetic noise.
Concerning training techniques in general, gamification is a promising method to increase user-comfort but also performance during training sessions [133]. We did a first step into this direction by implementing a game-like scenario for interaction with a robot by spoken and imagined speech, in how far this approach does have benefits over standard training procedures has however not been evaluated. Our future plans include a more detailed evaluation of the effect of gamification on Speech Imagery BCIs training concerning user comfort and performance.

7.4.2 Semantic classification

Regarding the Semantic Silent Speech BCI, we want to investigate the effect of semantic classification in Speech Imagery BCIs in more detail. Especially the desired aim of receiving better classification results when increasing the number of words, in comparison to a standard training approach, remains to be evaluated. With the current study setups we were able to show an improved performance in comparison to the standard approach, however, in our study setups we chose a fixed and rather small set of words. Increasing this number consequently beyond the scope of the current state-of-the-art will reveal the full potential of this method. In a future setup we will therefore choose a larger pool of words from each semantic category and analyze the performance of standard and semantic classification with increasing number of words. Due to the larger number of samples required for this evaluation resulting from the larger number of words to classify, we can foresee a synergy with our newly developed training methods and plan to use one of those methods into the study setup, in order to make the intensive training session more engaging for the participants.
We furthermore plan to have a closer look on source localization in the recorded EEG data to spot brain regions which are dominantly involved in semantic processing. Investigating in how far a certain region contributes more to the classification process will help to better tailor the number and position of electrodes used for classification and shed light on the question which categories are best discriminable by EEG activity. Our current selection of categories was based on fMRI studies which enable brain recordings with a much higher spatial resolution in comparison to EEG. Although we carefully selected categories from those studies to trigger clearly separatable regions of the cortex, we cannot ensure the same distribution and evoked patterns on the scalp surface for our EEG recordings. With the recorded data we want to evaluate which categories are best separatable across subjects and define certain meta categories on a higher level which can be better distinguished by EEG data.

7.4.3 Electrode reduction

Similar to the semantic concepts, we will further investigate our results on electrode reduction from a more advanced neuroscientific point of view. Our achieved result suggest that 32 electrodes are sufficient to reliably classify imagined speech from EEG activity. However, the positions of those electrodes initially followed a 64 channel 10-20 layout, which was reduced until the best classification performance was achieved. The performance of a standard 32 channel 10-20 layout could therefore not be measured and compared. Similar regions could unfortunately also not be determined among the participants and the reduced sets turned out to be individually distributed. A closer look on source localization techniques of the EEG and the recorded datasets might reveal a more precise view on important regions during imagined speech classification in general and allow for a more individual selection of electrodes. Apart from this individual selection we plan to extend our database with imagined speech EEG data recorded with further standard configurations with less channels as for example a 32 and 20 channel configuration arranged according to the standard 10-20 system. Such standard configurations will foster the transfer and ensure reproducibility of the developed methods among research groups around the world.
Beside electrode reduction for the sake of performance improvement, we will have a closer look on reduced sets of electrodes for improved user comfort. In the last decade, EEG devices have made rapid advancements and are produced in various shapes and

form factors. A quite recent development is the concept of in-ear and around the ear electrodes. Those electrodes are placed in the ear channel or around the pina, similar to headphones. Although providing only a limited view on the brain activity from a small region around the ear, researchers have started to investigate the possibility of classifying imagined speech with those electrodes [102, 103]. We foresee a tremendous potential for such electrodes especially concerning real-world applications of SI-BCIs, as the acceptance of a later user towards wearing an EEG device just around the ear will be much higher in comparison to wearing a standard EEG cap. Research in this field is however still in its infancy and we will contribute to it by testing our newly developed concepts and methods as for example the training methods with those types of electrodes in the future.

7.4.4 SI-BCI

In general, our newly developed methods need to show their performance under online BCI task conditions. Although similar on a technical level, the human factor during an online classification including feedback about the systems decision should not be underestimated. Excitement and frustration depending on the performance of the system can have a tremendous effect on the brain activity of the user and change or overlay the brain patterns as recorded during the offline training without any feedback of the system. Thus, the performance can change to the better or worse. A generally more focused state of the user during an actual task could furthermore improve the quality of the data and therefore the performance of the system. However, wrong classifications might lead to frustration and long term use to fatigue effects, which could have a negative impact on performance. We will therefore consequently transfer our methods evaluated in offline scenarios, to online applications and systematically evaluate the impact of the previously mentioned human factors on the performance.
Beside the previously mentioned immediate short-term effects in users brain activity, there are also long-term effects as for example variability of brain data which needs to be considered. Brain activity and therefore EEG data varies between users, but variation can also be found in the data of one individual in between separate days. Depending on the current state of the user, mental and physical, a classifier trained on the data of one day might not be applicable to data recorded on another day, creating the need for cumbersome recalibration of the system on a daily basis. Within our future study setups we aim at collecting larger datasets including multi-day experiments to investigate on the effect of brain signal variability on BCI performance and leverage more robust user- and variability-independent models for SI-BCIs.
Another extension of the developed methods in this thesis could be the integration of contextual information as additional support in the classification process. Speech interfaces usually operate in a certain situation with environmental factors and use additional information to support the classification process, e.g. autocomplete in text editors or spoken speech interfaces in general. The current experiments make use of setups in which isolated words are repeated without a clear task and context for the participant. A more defined use case scenario would allow for an integration of contextual information and a hybrid classification based on the knowledge about the surrounding. If the classifier provides an output for example to control a smart environment and switch on the light in the kitchen but this light is already switched on, the system could rely on that contextual knowledge and instead proceed with the next likely option which might be to switch the light off.

Concerning stimulus presentation in SI-BCI training scenarios, we want to have a closer look on the effects of different stimulus presentation methods on the later performance of the system. Within our studies we found first evidence, that the studies which used short-block presentation of stimuli achieved throughout higher classification accuracies in comparison to those using a random presentation. This effect might embellish the results achieved in offline evaluation and create unjustified high expectations for the transfer of this technology into online classification without such a short-block repetitions. Our procedure for future studies on SI-BCIs will in any case solely be based on randomized stimulus presentation, however, we want to investigate the concrete effect and differences in between those stimulus presentation methods to put research results of the community into perspective and correct the expectations towards a real-world online application of this technology.

In terms of interaction concepts and the paradigms used for imagined speech production, we can imagen the use of short sentences instead of single command words. Given that the repetition of single words requires the combination of several single interactions, an interaction based on short sentences would allow for a more natural and flexible interaction with a system. While a user would have to express the three words "put", "screwdriver" and "box" subsequently in order to advise a robot to put a screwdriver in a box, the short sentence "Put the screwdriver in the box." would be able to handle that command in one interaction. With a clear structure of such sentences as for example action word followed by an object followed by a destination, one could imagine numerous and flexible combinations of different commands with a certain set of words combined to short sentences. Classification of those sentences could realized by detecting single words inside a command string or trying to classify the whole structure of the sentence. Within our future study setups we will investigate the possibilities and potential of classifying imagined short sentences from EEG activity.

Given the classification results in general and the current implementation, we plan to have a closer look on different epochs of imagined speech signals. Research shows, that even with EEG it is possible to detect different parts of the speech production process as for example linguistic processing, motor planning and articulation [93]. Investigating these different stages in detail and applying more precisely tailored epochs to the classification process might significantly improve classification accuracies.

Finally, we will have a closer look on more advanced feature extraction and classification methods including Deep Learning approaches, as presented for example in [186] or [146]. Although the vast majority of work on imagined speech classification based on EEG data focuses on standard machine learning methods and hand-crafted features, using a neural network to analyze raw EEG data without extensive signal processing has great potential especially regarding real-world applications. We will extend our signal analysis pipelines to include neural networks and advanced deep learning techniques, in order to improve the performance of imagined speech classification based on EEG data and pave the way for real-world application of Speech Imagery BCIs in the future.

Appendix

Questionnaire on text eligibility

Fragebogen zur Textverständlichkeit

Im Folgenden finden Sie 24 Aussagen dazu, wie verständlich Sie persönlich den Text fanden. Bitte geben Sie zu jeder Aussage an, wie sehr diese Ihres Erachtens zutrifft oder nicht zutrifft. Falls Ihnen die Entscheidung einmal schwer fallen sollte, geben Sie einfach die Antwort, die am ehesten passt. Achten Sie darauf, dass Sie keine Aussage auslassen und überlegen Sie bitte bei einzelnen Aussagen nicht zu lange.

ID: _____ **Teil:**____

		stimmt nicht	stimmt eher nicht	stimmt teilweise	stimmt eher	stimmt genau
01	Bei manchen Wörtern war ich mir nicht sicher, was sie bedeuten.	☐	☐	☐	☐	☐
02	Es fiel mir leicht, mir den Inhalt bildlich vorzustellen.	☐	☐	☐	☐	☐
03	Die Sätze waren kompliziert geschrieben.	☐	☐	☐	☐	☐
04	Eine Liste aller Gegenstände oder Themen die im Text vorkamen, wäre sehr lang.	☐	☐	☐	☐	☐
05	Der Text widersprach an mehreren Stellen dem, was ich erwartet hatte.	☐	☐	☐	☐	☐
06	Ich fand die Sprache lebhaft.	☐	☐	☐	☐	☐
07	Im Text wurden viele Zusammenhänge dargestellt.	☐	☐	☐	☐	☐
08	Ich fand den Text verständlich.	☐	☐	☐	☐	☐
09	Ich kannte viele Wörter nicht.	☐	☐	☐	☐	☐
10	Beim Lesen hatte ich immer gleich ein Bild vom Gesagten vor Augen.	☐	☐	☐	☐	☐
11	Ich fand den Satzbau oft zu kompliziert.	☐	☐	☐	☐	☐
12	Der Text enthielt Aussagen über viele verschiedene Gegenstände oder Themen.	☐	☐	☐	☐	☐

Figure 1: Questionnaire on text eligibility part one, taken from [62]

		stimmt nicht	stimmt eher nicht	stimmt teilweise	stimmt eher	stimmt genau
13	Der Text ging an mehreren Stellen anders weiter, als ich es erwartet hatte.	☐	☐	☐	☐	☐
14	Der Text war monoton.	☐	☐	☐	☐	☐
15	Im Text wurden nur wenige Zusammenhänge dargestellt.	☐	☐	☐	☐	☐
16	Der Text könnte deutlich verständlicher sein.	☐	☐	☐	☐	☐
17	Ich wusste bei allen Wörtern sofort, was sie bedeuten.	☐	☐	☐	☐	☐
18	Ich fände es sehr leicht, eine Zusammenfassung zu geben.	☐	☐	☐	☐	☐
19	Viele Sätze waren sehr lang.	☐	☐	☐	☐	☐
20	Im Text kamen viele verschiedene Gegenstände oder Themen vor.	☐	☐	☐	☐	☐
21	Ich war manchmal überrascht, wie der Text weiter ging.	☐	☐	☐	☐	☐
22	Ich fand die Sprache abwechslungsreich.	☐	☐	☐	☐	☐
23	Der Text enthielt sehr viele Informationen.	☐	☐	☐	☐	☐
24	Alles in allem war der Text leicht zu verstehen.	☐	☐	☐	☐	☐

Figure 2: Questionnaire on text eligibility part two, taken from [62]

Extended results on electrode reduction

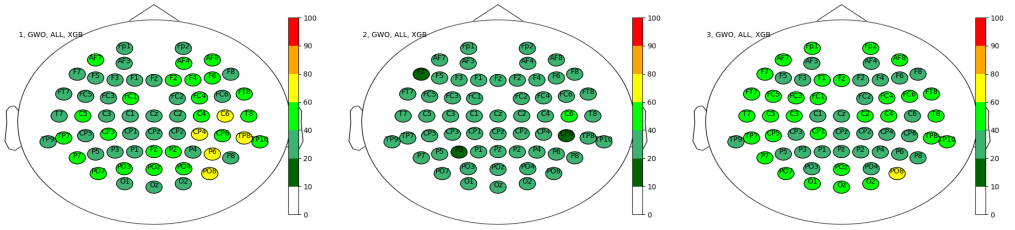

Figure 3: Occurrences of electrodes by position for the top sets, separated by the three datasets including all three feature extraction methods. Left: dataset 1, middle: dataset 2, right: dataset 3.

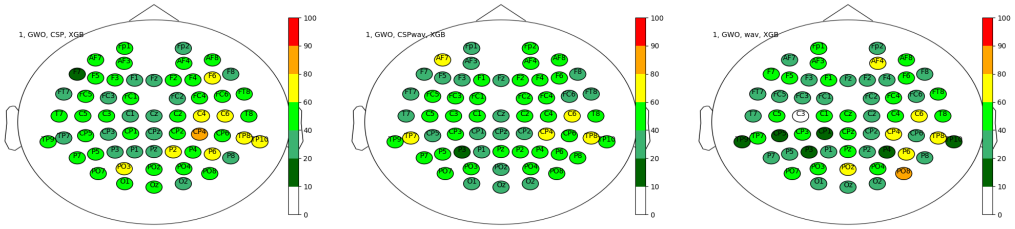

Figure 4: Occurrences of electrodes by position for the top sets, separated by feature extraction method for dataset 1. Left: CSP, middle: CSPwav, right: wav.

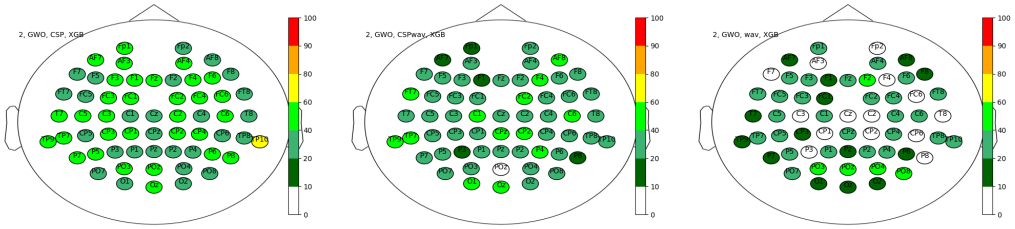

Figure 5: Occurrences of electrodes by position for the top sets, separated by feature extraction method for dataset 2. Left: CSP, middle: CSPwav, right: wav.

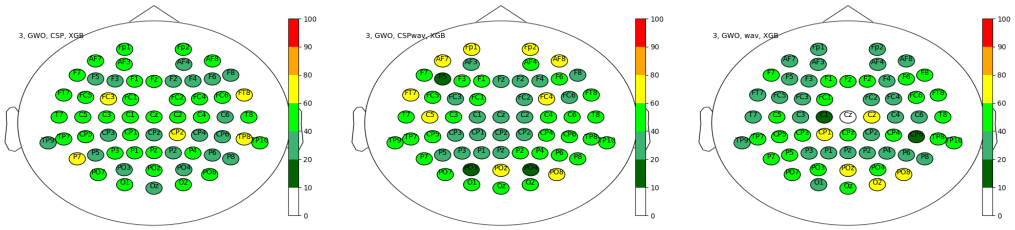

Figure 6: Occurrences of electrodes by position for the top sets, separated by feature extraction method for dataset 3. Left: CSP, middle: CSPwav, right: wav.

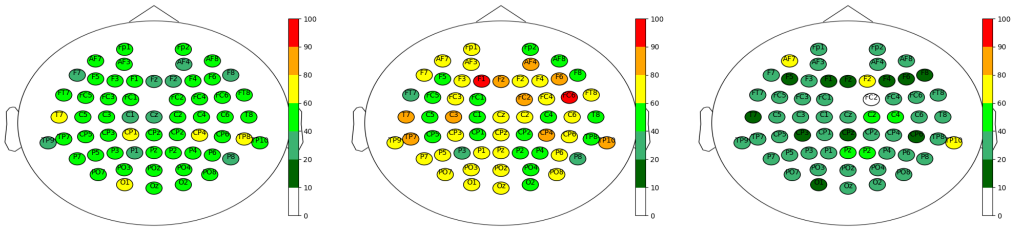

Figure 7: Occurrences of electrodes for the top sets, separated by cluster for all datasets and CSP feature extraction method. Left: Cluster 1, middle: Cluster 2, right: Cluster 3.

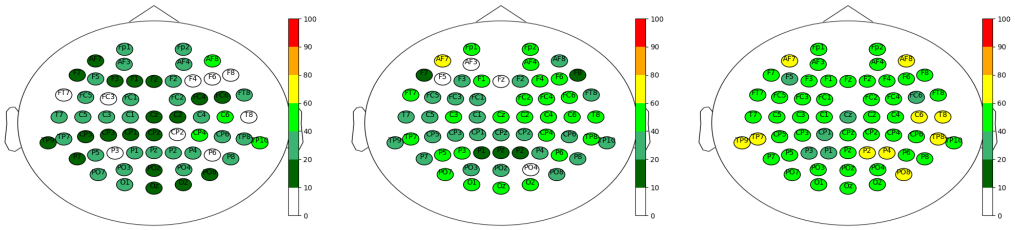

Figure 8: Occurrences of electrodes for the top sets, separated by cluster for all datasets and CSPwav feature extraction method. Left: Cluster 1, middle: Cluster 2, right: Cluster 3.

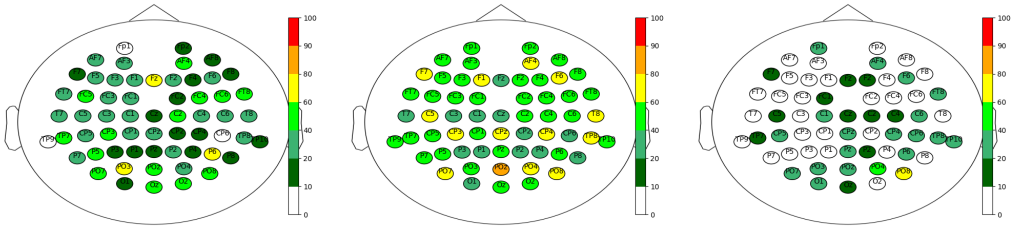

Figure 9: Occurrences of electrodes for the top sets, separated by cluster for all datasets and wav feature extraction method. Left: Cluster 1, middle: Cluster 2, right: Cluster 3.

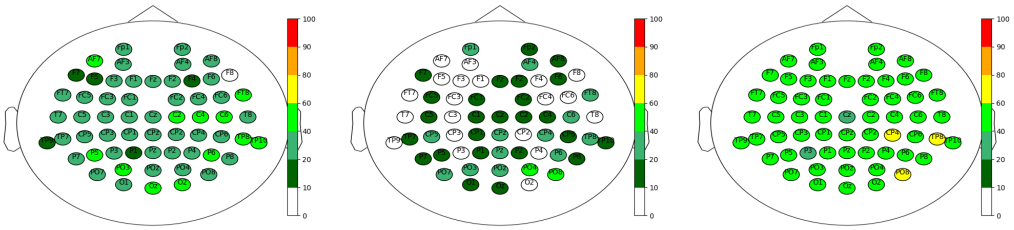

Figure 10: Occurrences of electrodes for the top sets, separated by cluster for all datasets and all feature extraction methods. Left: Cluster 1, middle: Cluster 2, right: Cluster 3.

		Dataset 1				Dataset 2				Dataset 3			
Nr	**Param**	**GWO**	**B & W**	**Left**	**Right**	**GWO**	**B & W**	**Left**	**Right**	**GWO**	**B & W**	**Right**	**Left**
	Acc	0.60	0.36	0.44	0.38	0.40	0.24	0.22	0.17	0.42	0.26	0.28	0.25
1	Elec	37	18	35	35	55	18	35	35	41	18	35	35
	Feat	wav	CSPw	wav	wav	wav	CSPw	wav	CSP	CSPw	wav	wav	CSPw
	Acc	0.50	0.26	0.32	0.28	0.38	0.22	0.18	0.22	0.40	0.29	0.23	0.36
2	Elec	53	18	35	35	21	18	35	35	31	18	35	35
	Feat	wav	wav	wav	wav	CSP	wav	CSP	wav	CSPw	CSP	wav	wav
	Acc	0.60	0.3	0.32	0.40	0.41	0.18	0.21	0.19	0.42	0.24	0.29	0.30
3	Elec	33	18	35	35	30	18	35	35	31	18	35	35
	Feat	wav	wav	wav	wav	wav	CSPw	wav	CSP	CSPw	wav	CSP	CSPw
	Acc	0.58	0.32	0.38	0.42	0.43	0.24	0.22	0.17	0.37	0.31	0.25	0.30
4	Elec	40	18	35	35	25	18	35	35	32	18	35	35
	Feat	wav	CSP	wav	wav	CSP	CSP	wav	CSP	CSP	CSP	CSP	CSPw
	Acc	0.62	0.38	0.28	0.30	0.37	0.15	0.22	0.19	0.41	0.25	0.29	0.26
5	Elec	36	18	35	35	32	18	35	35	31	18	35	35
	Feat	wav	CSPw	wav	CSPw	wav	CSPw	CSPw	CSP	CSP	CSPw	CSPw	CSPw
	Acc	0.54	0.3	0.32	0.34	0.45	0.18	0.24	0.24	0.40	0.2	0.29	0.24
6	Elec	34	18	35	35	55	18	35	35	43	18	35	35
	Feat	wav	wav	wav	wav	wav	CSP	CSP	wav	wav	CSP	CSP	wav
	Acc	0.60	0.34	0.34	0.40	0.38	0.18	0.17	0.29	0.40	0.31	0.28	0.26
7	Elec	34	18	35	35	58	18	35	35	42	18	35	35
	Feat	wav	CSPw	CSPw	wav	wav	wav	wav	wav	CSPw	CSPw	CSP	wav
	Acc	0.48	0.36	0.34	0.24	0.40	0.19	0.24	0.21	0.38	0.31	0.24	0.29
8	Elec	30	18	35	35	34	18	35	35	34	18	35	35
	Feat	CSPw	CSPw	wav	CSP	wav	wav	wav	CSP	CSP	CSP	CSPw	CSP
	Acc	0.62	0.36	0.4	0.34	0.47	0.21	0.17	0.18	0.40	0.24	0.28	0.25
9	Elec	32	18	35	35	56	18	35	35	30	18	35	35
	Feat	CSPw	wav	CSPw	CSPw	wav	CSP	wav	wav	wav	CSP	wav	CSP
	Acc	0.58	0.28	0.42	0.42	0.36	0.18	0.19	0.25	0.40	0.25	0.25	0.26
10	Elec	35	18	35	35	29	18	35	35	31	18	35	35
	Feat	wav	CSP	wav	wav	CSP	CSP	wav	wav	CSP	wav	wav	CSPw
	Acc	0.58	0.28	0.3	0.38	0.40	0.19	0.19	0.17	0.40	0.28	0.29	0.23
11	Elec	44	18	35	35	41	18	35	35	44	18	35	35
	Feat	wav	CSPw	wav	wav	wav	CSP	CSP	CSP	CSP	CSP	CSP	CSPw
	Acc	0.52	0.24	0.36	0.30	0.34	0.21	0.22	0.21	0.43	0.32	0.29	0.26
12	Elec	46	18	35	35	32	18	35	35	34	18	35	35
	Feat	CSPw	wav	CSP	CSPw	CSP	CSP	wav	wav	CSP	CSPw	CSPw	CSPw
	Acc	0.50	0.26	0.32	0.28	0.45	0.18	0.19	0.26	0.42	0.31	0.29	0.31
13	Elec	53	18	35	35	44	18	35	35	36	18	35	35
	Feat	wav	wav	wav	wav	wav	wav	wav	wav	wav	wav	wav	CSP
	Acc	0.54	0.3	0.34	0.36	0.40	0.17	0.21	0.22	0.41	0.23	0.3	0.34
14	Elec	30	18	35	35	43	18	35	35	33	18	35	35
	Feat	CSPw	CSP	CSP	CSPw	wav	CSP	CSPw	CSP	CSPw	CSP	wav	CSP
	Acc	0.54	0.36	0.28	0.30	0.37	0.19	0.25	0.18	0.40	0.26	0.28	0.25
15	Elec	32	18	35	35	32	18	35	35	35	18	35	35
	Feat	wav	wav	wav	wav	CSP	wav	wav	wav	wav	wav	wav	wav
	Acc	0.56	0.31	0.34	0.34	0.40	0.19	0.21	0.21	0.40	0.27	0.28	0.28
Avg	Elec	38	18	35	35	wav	18	35	35	35	18	35	35
	Feat	wav	wav	wav	wav	wav	CSP	CSP	CSP	wav	CSP	wav	CSPw

Table 1: Best top sets for each individual compared to the Broca and wernicke as well as left and hemisphere configuration of electrodes.

List of Figures

List of Tables

Bibliography

[1] Nassib Abdallah, Pierre Chauvet, Abd El Salam Hajjar, and Bassam Daya. 2018. Optimized brain computer interface system for unspoken speech recognition: Role of wernicke area. *International Journal of Biomedical and Biological Engineering* 12, 10 (2018), 456–462.

[2] Jayant N Acharya, Abeer J Hani, Janna Cheek, Parthasarathy Thirumala, and Tammy N Tsuchida. 2016. American clinical neurophysiology society guideline 2: guidelines for standard electrode position nomenclature. *The Neurodiagnostic Journal* 56, 4 (2016), 245–252.

[3] Prabhakar Agarwal, Rajiv Kumar Kale, Manish Kumar, and Sandeep Kumar. 2020. Silent speech classification based upon various feature extraction methods. In *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 16–20.

[4] Amjed S Al-Fahoum and Ausilah A Al-Fraihat. 2014. Methods of EEG Signal Features Extraction Using Linear Analysis in Frequency and Time-Frequency Domains. *ISRN Neuroscience* (2014).

[5] Noura Al Moubayed, Bashar Awwad Shiekh Hasan, John Q Gan, Andrei Petrovski, and John McCall. 2010. Binary-SDMOPSO and its application in channel selection for brain-computer interfaces. In *2010 UK Workshop on Computational Intelligence (UKCI)*. IEEE, 1–6.

[6] Noura Al Moubayed, Bashar Awwad Shiekh Hasan, John Q Gan, Andrei Petrovski, and John McCall. 2012. Continuous presentation for multi-objective channel selection in brain-computer interfaces. In *2012 IEEE Congress on Evolutionary Computation*. IEEE, 1–7.

[7] Laurent Albera, Amar Kachenoura, Pierre Comon, Ahmad Karfoul, Fabrice Wendling, Lotfi Senhadji, and Isabelle Merlet. 2012. ICA-based EEG denoising: a comparative analysis of fifteen methods. *Bulletin of the Polish Academy of Sciences: Technical Sciences* 60, 3 (2012), 407–418.

[8] Sarah Alizadeh, Hamidreza Jamalabadi, Monika Schönauer, Christian Leibold, and Steffen Gais. 2017. Decoding cognitive concepts from neuroimaging data using multivariate pattern analysis. *Neuroimage* 159 (2017), 449–458.

[9] Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. 2019. Speech synthesis from neural decoding of spoken sentences. *Nature* 568, 7753 (2019), 493–498.

[10] Alfredo Ardila, Byron Bernal, and Monica Rosselli. 2016. How localized are language brain areas? A review of Brodmann areas involvement in oral language. *Archives of Clinical Neuropsychology* 31, 1 (2016), 112–122.

[11] Pietro Aricò, Gianluca Borghini, Gianluca Di Flumeri, Alfredo Colosimo, Stefano Bonelli, Alessia Golfetti, Simone Pozzi, Jean-Paul Imbert, Géraud Granger, Raïlane Benhacene, and others. 2016. Adaptive automation triggered by EEG-based mental workload index: a passive brain-computer interface application in realistic air traffic control environment. *Frontiers in human neuroscience* 10 (2016), 539.

[12] Aamir Arsalan, Muhammad Majid, Amna Rauf Butt, and Syed Muhammad Anwar. 2019. Classification of perceived mental stress using a commercially available EEG headband. *IEEE journal of biomedical and health informatics* 23, 6 (2019), 2257–2264.

[13] Alan Baddeley, Marge Eldridge, and Vivien Lewis. 1981. The role of subvocalisation in reading. *The Quarterly Journal of Experimental Psychology* 33, 4 (1981), 439–454.

[14] Forrest Sheng Bao, Xin Liu, and Christina Zhang. 2011. PyEEG: an open source python module for EEG/MEG feature extraction. *Computational intelligence and neuroscience* 2011 (2011).

[15] Mark Bear, Barry Connors, and Michael A Paradiso. 2016. *Neuroscience: Exploring the Brain, Enhanced Edition: Exploring the Brain*. Wolters Kluwer.

[16] Hans Berger. 1929. Über das elektroenkephalogramm des menschen. *Archiv für psychiatrie und nervenkrankheiten* 87, 1 (1929), 527–570.

[17] Martin Billinger, Ian Daly, Vera Kaiser, Jing Jin, Brendan Z Allison, Gernot R Müller-Putz, and Clemens Brunner. 2012. Is it significant? Guidelines for reporting BCI performance. In *Towards Practical Brain-Computer Interfaces*. Springer, 333–354.

[18] Jeffrey R Binder. 2017. Current controversies on Wernicke's area and its role in language. *Current neurology and neuroscience reports* 17, 8 (2017), 1–10.

[19] Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex* 19, 12 (2009), 2767–2796.

[20] Benjamin Blankertz, Claudia Sannelli, Sebastian Halder, Eva M Hammer, Andrea Kübler, Klaus-Robert Müller, Gabriel Curio, and Thorsten Dickhaus. 2010. Neurophysiological predictor of SMR-based BCI performance. *Neuroimage* 51, 4 (2010), 1303–1309.

[21] Benjamin Blankertz, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and Klaus-Robert Muller. 2007. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal processing magazine* 25, 1 (2007), 41–56.

[22] Florent Bocquelet, Thomas Hueber, Laurent Girin, Stéphan Chabardès, and Blaise Yvert. 2016. Key considerations in designing a speech brain-computer interface. *Journal of Physiology-Paris* 110, 4 (2016), 392–401.

[23] Erik Bojorges-Valdez, Juan C Echeverría, and Oscar Yanez-Suarez. 2015. Evaluation of the continuous detection of mental calculation episodes as a BCI control input. *Computers in biology and medicine* 64 (2015), 155–162.

[24] Laurent Bonnet, Fabien Lotte, and Anatole Lécuyer. 2013. Two brains, one game: design and evaluation of a multiuser BCI video game based on motor imagery. *IEEE Transactions on Computational Intelligence and AI in games* 5, 2 (2013), 185–198.

[25] Elena Boto, Niall Holmes, James Leggett, Gillian Roberts, Vishal Shah, Sofie S Meyer, Leonardo Duque Muñoz, Karen J Mullinger, Tim M Tierney, Sven Bestmann, and others. 2018. Moving magnetoencephalography towards real-world applications with a wearable system. *Nature* 555, 7698 (2018), 657–661.

[26] Michael Breakspear. 2017. Dynamic models of large-scale brain activity. *Nature neuroscience* 20, 3 (2017), 340–352.

[27] Clemens Brunner, Arnaud Delorme, and Scott Makeig. 2013. Eeglab–an open source matlab toolbox for electrophysiological research. *Biomedical Engineering/Biomedizinische Technik* 58, SI-1-Track-G (2013), 000010151520134182.

[28] Richard B Buxton. 2009. *Introduction to functional magnetic resonance imaging: principles and techniques*. Cambridge university press.

[29] Richard Caton. 1875. The electric currents of the brain. *The British Medical Journal* (1875), 278.

[30] Shreya Chakrabarti, Hilary M Sandberg, Jonathan S Brumberg, and Dean J Krusienski. 2015. Progress in speech decoding from the electrocorticogram. *Biomedical Engineering Letters* 5, 1 (2015), 10–21.

[31] V Srinivasa Chakravarthy. 2018. *Demystifying the Brain: A Computational Approach*. Springer.

[32] Soumyadip Chatterjee, Saugat Bhattacharyya, Amit Konar, DN Tibarewala, Anwesha Khasnobish, and Ramadoss Janarthanan. 2013. Performance analysis of multiclass common spatial patterns in brain-computer interface. In *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 115–120.

[33] Maximilien Chaumon, Dorothy VM Bishop, and Niko A Busch. 2015. A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *Journal of neuroscience methods* 250 (2015), 47–63.

[34] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[35] Xiaogang Chen, Yijun Wang, Masaki Nakanishi, Xiaorong Gao, Tzyy-Ping Jung, and Shangkai Gao. 2015. High-speed spelling with a noninvasive brain–computer interface. *Proceedings of the national academy of sciences* 112, 44 (2015), E6058–E6067.

[36] Xiaogang Chen, Yijun Wang, Masaki Nakanishi, Tzyy-Ping Jung, and Xiaorong Gao. 2014. Hybrid frequency and phase coding for a high-speed SSVEP-based BCI speller. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 3993–3996.

[37] Xun Chen, Xueyuan Xu, Aiping Liu, Soojin Lee, Xiang Chen, Xu Zhang, Martin J. McKeown, and Z. Jane Wang. 2019. Removal of Muscle Artifacts from the EEG: A Review and Recommendations. *IEEE Sensors Journal* 19 (2019), 5353–5368. Issue 14. DOI:http://dx.doi.org/10.1109/JSEN.2019.2906572

[38] Michael X Cohen. 2017. Where does EEG come from and what does it mean? *Trends in neurosciences* 40, 4 (2017), 208–218.

[39] Etienne Combrisson and Karim Jerbi. 2015. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of neuroscience methods* 250 (2015), 126–136.

[40] Stephen Coombes, Peter beim Graben, Roland Potthast, and James Wright. 2014. *Neural fields: theory and applications*. Springer.

[41] Ciaran Cooney, Raffaella Folli, and Damien Coyle. 2019. Optimizing layers improves CNN generalization and transfer learning for imagined speech decoding from EEG. In *2019 IEEE international conference on systems, man and cybernetics (SMC)*. IEEE, 1311–1316.

[42] Ciaran Cooney, Attila Korik, Raffaella Folli, and Damien Coyle. 2020. Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG. *Sensors* 20, 16 (2020), 4629.

[43] Michelle E Costanzo, Joseph J McArdle, Bruce Swett, Vladimir Nechaev, Stefan Kemeny, Jiang Xu, and Allen R Braun. 2013. Spatial and temporal features of superordinate semantic processing studied with fMRI and EEG. *Frontiers in human neuroscience* 7 (2013), 293.

[44] Damien Coyle, Jhonatan Garcia, Abdul R Satti, and T Martin McGinnity. 2011. EEG-based continuous control of a game using a 3 channel motor imagery BCI: BCI game. In *2011 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)*. IEEE, 1–7.

[45] Shirley M Coyle, Tomás E Ward, and Charles M Markham. 2007. Brain–computer interface using a simplified functional near-infrared spectroscopy system. *Journal of neural engineering* 4, 3 (2007), 219.

[46] Nathan E Crone, Alon Sinai, and Anna Korzeniewska. 2006. High-frequency gamma oscillations and human brain mapping with electrocorticography. *Progress in brain research* 159 (2006), 275–295.

[47] Charles S DaSalla, Hiroyuki Kambara, Yasuharu Koike, and Makoto Sato. 2009. Spatial filtering and single-trial classification of EEG during vowel speech imagery. In *Proceedings of the 3rd International Convention on Rehabilitation Engineering & Assistive Technology*. 1–4.

[48] Debadatta Dash, Paul Ferrari, and Jun Wang. 2020. Decoding imagined and spoken phrases from non-invasive neural (MEG) signals. *Frontiers in neuroscience* 14 (2020), 290.

[49] Debadatta Dash, Alan Wisler, Paul Ferrari, and Jun Wang. 2019. Towards a Speaker Independent Speech-BCI Using Speaker Adaptation.. In *INTERSPEECH*. 864–868.

[50] Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, Jim M Gilbert, and Jonathan S Brumberg. 2010. Silent speech interfaces. *Speech Communication* 52, 4 (2010), 270–287.

[51] Siyi Deng, Ramesh Srinivasan, and Michael D'Zmura. 2013. *Cortical signatures of heard and imagined speech envelopes*. Technical Report. CALIFORNIA UNIV IRVINE DEPT OF COGNITIVE SCIENCES.

[52] Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. 2019. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience* 39, 39 (2019), 7722–7736.

[53] Joseph T Devlin, Paul M Matthews, and Matthew FS Rushworth. 2003. Semantic processing in the left inferior prefrontal cortex: a combined functional magnetic resonance imaging and transcranial magnetic stimulation study. *Journal of cognitive neuroscience* 15, 1 (2003), 71–84.

[54] Nina F Dronkers, David P Wilkins, Robert D Van Valin Jr, Brenda B Redfern, and Jeri J Jaeger. 2004. Lesion analysis of the brain areas involved in language comprehension. *Cognition* 92, 1-2 (2004), 145–177.

[55] Eid Emary, Hossam M Zawbaa, and Aboul Ella Hassanien. 2016. Binary grey wolf optimization approaches for feature selection. *Neurocomputing* 172 (2016), 371–381.

[56] Stephen H Fairclough and Fabien Lotte. 2020. Grand challenges in neurotechnology and system neuroergonomics. *Frontiers in Neuroergonomics* (2020), 2.

[57] Jian Kui Feng, Jing Jin, Ian Daly, Jiale Zhou, Yugang Niu, Xingyu Wang, and Andrzej Cichocki. 2019. An optimized channel selection method based on multifrequency CSP-rank for motor imagery-based BCI system. *Computational intelligence and neuroscience* 2019 (2019).

[58] Thalía Fernández, Thalía Harmony, Mario Rodríguez, Jorge Bernal, Juan Silva, Alfonso Reyes, and Erzsébet Marosi. 1995. EEG activation patterns during the performance of tasks involving different components of mental calculation. *Electroencephalography and clinical neurophysiology* 94, 3 (1995), 175–182.

[59] Marco Ferrari and Valentina Quaresima. 2012. A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *Neuroimage* 63, 2 (2012), 921–935.

[60] Matteo Fraschini, Matteo Demuru, Alessandra Crobe, Francesco Marrosu, Cornelis J Stam, and Arjan Hillebrand. 2016. The effect of epoch length on estimated EEG functional connectivity and brain network organisation. *Journal of neural engineering* 13, 3 (2016), 036015.

[61] Jo Freitas, Ant Teixeira, Miguel Sales Dias, Samuel Silva, and others. 2017. *An Introduction to Silent Speech Interfaces*. Springer.

[62] Marcus Friedrich. 2017. *Textverständlichkeit und ihre Messung: Entwicklung und Erprobung eines Fragebogens zur Textverständlichkeit*. Waxmann Verlag.

[63] Alejandro Antonio Torres García, Carlos A Reyes García, and Luis Villaseñor Pineda. 2012. Toward a Silent Speech Interface based on Unspoken Speech.. In *Biosignals*. 370–373.

[64] Jesús S García-Salinas, Luis Villaseñor-Pineda, Carlos A Reyes-García, and Alejandro A Torres-García. 2019. Transfer learning in imagined speech EEG-based BCIs. *Biomedical Signal Processing and Control* 50 (2019), 151–157.

[65] Antoine Gaume, François Vialatte, and Gérard Dreyfus. 2014. Transient brain activity explains the spectral content of steady-state visual evoked potentials. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 688–692.

[66] Norman Geschwind. 1970. The Organization of Language and the Brain: Language disorders after brain damage help in elucidating the neural basis of verbal behavior. *Science* 170, 3961 (1970), 940–944.

[67] P Ghane, G Hossain, and A Tovar. 2015. Robust understanding of EEG patterns in silent speech. In *2015 National Aerospace and Electronics Conference (NAECON)*. IEEE, 282–289.

[68] Rajdeep Ghosh, Nidul Sinha, Saroj Kumar Biswas, and Souvik Phadikar. 2019. A modified grey wolf optimization based feature selection method from EEG for silent speech classification. *Journal of Information and Optimization Sciences* 40, 8 (2019), 1639–1652.

[69] Erick F González-Castañeda, Alejandro A Torres-García, Carlos A Reyes-García, and Luis Villaseñor-Pineda. 2017. Sonification and textification: Proposing methods for classifying unspoken words from EEG signals. *Biomedical Signal Processing and Control* 37 (2017), 82–91.

[70] Melvyn A Goodale and A David Milner. 1992. Separate visual pathways for perception and action. *Trends in neurosciences* 15, 1 (1992), 20–25.

[71] Roland H Grabner, Clemens Brunner, Robert Leeb, Christa Neuper, and Gert Pfurtscheller. 2007. Event-related EEG theta and alpha band oscillatory responses during language translation. *Brain research bulletin* 72, 1 (2007), 57–65.

[72] Vincent L Gracco, Pascale Tremblay, and Bruce Pike. 2005. Imaging speech production using fMRI. *Neuroimage* 26, 1 (2005), 294–301.

[73] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and others. 2013b. MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience* 7 (2013), 267.

[74] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. 2013a. MEG and EEG Data Analysis with MNE-Python. *Frontiers in Neuroscience* 7, 267 (2013), 1–13. DOI:http://dx.doi.org/10.3389/fnins.2013.00267

[75] Moritz Grosse-Wentrup and Martin Buss. 2008. Multiclass common spatial patterns and information theoretic feature extraction. *IEEE transactions on Biomedical Engineering* 55, 8 (2008), 1991–2000.

[76] Christoph Guger, Gunter Edlinger, W Harkam, I Niedermayer, and Gert Pfurtscheller. 2003. How many people are able to operate an EEG-based brain-computer interface (BCI)? *IEEE transactions on neural systems and rehabilitation engineering* 11, 2 (2003), 145–147.

[77] Isabelle Guyon and André Elisseeff. 2006. An introduction to feature extraction. In *Feature extraction*. Springer, 1–25.

[78] Matti Hämäläinen, Riitta Hari, Risto J Ilmoniemi, Jukka Knuutila, and Olli V Lounasmaa. 1993. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of modern Physics* 65, 2 (1993), 413.

[79] John Hart Jr and Barry Gordon. 1990. Delineation of single-word semantic comprehension deficits in aphasia, with anatomical correlation. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 27, 3 (1990), 226–231.

[80] Christian Henle, Martin Schuettler, Jörn Rickert, and Thomas Stieglitz. 2012. Towards electrocorticographic electrodes for chronic use in BCI applications. In *Towards Practical Brain-Computer Interfaces*. Springer, 85–103.

[81] Gregory Hickok. 2012. The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *Journal of Communication Disorders* 45, 6 (2012), 393–402.

[82] Gregory Hickok. 2022. The dual stream model of speech and language processing. In *Handbook of Clinical Neurology*. Vol. 185. Elsevier, 57–69.

[83] Gregory Hickok and David Poeppel. 2007. The cortical organization of speech processing. *Nature reviews neuroscience* 8, 5 (2007), 393–402.

[84] Barbara Höhle. 2011. *Psycholinguistik*. Walter de Gruyter.

[85] Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific data* 5, 1 (2018), 1–13.

[86] Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 138–146.

[87] Scott A Huettel, Allen W Song, Gregory McCarthy, and others. 2004. *Functional magnetic resonance imaging*. Vol. 1. Sinauer Associates Sunderland, MA.

[88] Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532, 7600 (2016), 453.

[89] Aapo Hyvärinen and Erkki Oja. 2000. Independent component analysis: algorithms and applications. *Neural networks* 13, 4-5 (2000), 411–430.

[90] Khalid Ibrahim and Richard Appleton. 2004. Seizures as the presenting symptom of brain tumours in children. *Seizure* 13, 2 (2004), 108–112.

[91] Farzin Irani, Steven M Platek, Scott Bunce, Anthony C Ruocco, and Douglas Chute. 2007. Functional near infrared spectroscopy (fNIRS): an emerging neuroimaging technology with important applications for the study of brain disorders. *The Clinical Neuropsychologist* 21, 1 (2007), 9–37.

[92] Amir Jahangiri, Juan M Chau, David R Achanccaray, and Francisco Sepulveda. 2018. Covert speech vs. motor imagery: a comparative study of class separability in identical environments. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2020–2023.

[93] Amir Jahangiri and Francisco Sepulveda. 2019. The relative contribution of high-gamma linguistic processing stages of word production, and motor imagery of articulation in class separability of covert speech tasks in EEG data. *Journal of medical systems* 43, 2 (2019), 1–9.

[94] Kyle M Jasmin, Carolyn McGettigan, Zarinah K Agnew, Nadine Lavan, Oliver Josephs, Fred Cummins, and Sophie K Scott. 2016. Cohesion and joint speech: right hemisphere contributions to synchronized vocal production. *Journal of Neuroscience* 36, 17 (2016), 4669–4680.

[95] Xiao Jiang, Gui-Bin Bian, and Zean Tian. 2019. Removal of artifacts from EEG signals: a review. *Sensors* 19, 5 (2019), 987.

[96] Suwicha Jirayucharoensak, Setha Pan-Ngum, and Pasin Israsena. 2014. EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *The Scientific World Journal* 2014 (2014).

[97] Anna Jo and Brian Yongwook Chae. 2020. Introduction to real time user interaction in virtual reality powered by brain computer interface technology. In *ACM SIGGRAPH 2020 Real-Time Live!* 1–1.

[98] Gael Jobard, Mathieu Vigneau, Bernard Mazoyer, and Nathalie Tzourio-Mazoyer. 2007. Impact of modality and linguistic complexity during reading and listening tasks. *Neuroimage* 34, 2 (2007), 784–800.

[99] Sven Joubert, Mario Beauregard, Nathalie Walter, Pierre Bourgouin, Gilles Beaudoin, Jean-Maxime Leroux, Sherif Karama, and André Roch Lecours. 2004. Neural correlates of lexical and sublexical processes in reading. *Brain and language* 89, 1 (2004), 9–20.

[100] Valer Jurcak, Daisuke Tsuzuki, and Ippeita Dan. 2007. 10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems. *Neuroimage* 34, 4 (2007), 1600–1611.

[101] Julia WY Kam, Sandon Griffin, Alan Shen, Shawn Patel, Hermann Hinrichs, Hans-Jochen Heinze, Leon Y Deouell, and Robert T Knight. 2019. Systematic comparison between a wireless EEG system with dry electrodes and a wired EEG system with wet electrodes. *NeuroImage* 184 (2019), 119–129.

[102] Netiwit Kaongoen, Jaehoon Choi, and Sungho Jo. 2021. Speech-imagery-based brain–computer interface system using ear-EEG. *Journal of neural engineering* 18, 1 (2021), 016023.

[103] Netiwit Kaongoen, Jaehoon Choi, and Sungho Jo. 2022. A novel online BCI system using speech imagery and ear-EEG for home appliances control. *Computer Methods and Programs in Biomedicine* 224 (2022), 107022.

[104] Robert M Kaplan. 2011. The mind reader: the forgotten life of Hans Berger, discoverer of the EEG. *Australasian Psychiatry* 19, 2 (2011), 168.

[105] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. Alterego: A personalized wearable silent speech interface. In *23rd International conference on intelligent user interfaces*. 43–53.

[106] Arnav Kapur, Utkarsh Sarawgi, Eric Wadkins, Matthew Wu, Nora Hollenstein, and Pattie Maes. 2020. Non-invasive silent speech recognition in multiple sclerosis with dysphonia. In *Machine Learning for Health Workshop*. PMLR, 25–38.

[107] Aneta Kartali, Milica M Janković, Ivan Gligorijević, Pavle Mijović, Bogdan Mijović, and Maria Chiara Leva. 2019. Real-time mental workload estimation using eeg. In *International Symposium on Human Mental Workload: Models and Applications*. Springer, 20–34.

[108] Christian Keller and Christian A Kell. 2016. Asymmetric intra-and interhemispheric interactions during covert and overt sentence reading. *Neuropsychologia* 93 (2016), 448–465.

[109] Javeria Khan, Muhammad Hamza Bhatti, Usman Ghani Khan, and Razi Iqbal. 2019. Multiclass EEG motor-imagery classification with sub-band common spatial patterns. *EURASIP Journal on Wireless Communications and Networking* 2019, 1 (2019), 1–9.

[110] Jongin Kim, Suh-Kyung Lee, and Boreom Lee. 2014. EEG classification in a single-trial basis for vowel speech perception using multivariate empirical mode decomposition. *Journal of neural engineering* 11, 3 (2014), 036010.

[111] Su Kyoung Kim, Elsa Andrea Kirchner, Arne Stefes, and Frank Kirchner. 2017. Intrinsic interactive reinforcement learning–Using error-related potentials for real world human-robot interaction. *Scientific reports* 7, 1 (2017), 1–16.

[112] Taekyung Kim, Jeyeon Lee, Hoseok Choi, Hojong Lee, In-Young Kim, and Dong Pyo Jang. 2013. Meaning based covert speech classification for brain-computer interface based on electroencephalography. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 53–56.

[113] Zoltan J Koles, Michael S Lazar, and Steven Z Zhou. 1990. Spatial patterns underlying population differences in the background EEG. *Brain topography* 2, 4 (1990), 275–284.

[114] Nataliya Kosmyna, Franck Tarpin-Bernard, and Bertrand Rivet. 2015. Towards brain computer interfaces for recreational activities: Piloting a drone. In *IFIP Conference on Human-Computer Interaction*. Springer, 506–522.

[115] Gautam Krishna, Yan Han, Co Tran, Mason Carnahan, and Ahmed H Tewfik. 2019. State-of-the-art speech recognition using eeg and towards decoding of speech spectrum from eeg. *arXiv preprint arXiv:1908.05743* (2019).

[116] Pradeep Kumar, Rajkumar Saini, Partha Pratim Roy, Pawan Kumar Sahu, and Debi Prosad Dogra. 2018. Envisioned speech recognition using EEG sensors. *Personal and Ubiquitous Computing* 22, 1 (2018), 185–199.

[117] Binita Kumari and Tripti Swarnkar. 2011. Filter versus wrapper feature subset selection in large dimensionality micro array: A review. *International Journal of Computer Science and Information Technologies* 2, 3 (2011), 1048–1053.

[118] Abraham Kuruvilla and Roland Flink. 2003. Intraoperative electrocorticography in epilepsy surgery: useful or not? *Seizure* 12, 8 (2003), 577–584.

[119] Tian Lan, Deniz Erdogmus, Andre Adami, Misha Pavel, and Santosh Mathan. 2006. Salient EEG channel selection in brain computer interfaces by mutual information maximization. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, 7064–7067.

[120] Dae-Hyeok Lee, Ji-Hoon Jeong, Hyung-Ju Ahn, and Seong-Whan Lee. 2021. Design of an EEG-based drone swarm control system using endogenous BCI paradigms. In *2021 9th International Winter Conference on Brain-Computer Interface (BCI)*. IEEE, 1–5.

[121] Gregory Lee, Ralf Gommers, Filip Waselewski, Kai Wohlfahrt, and Aaron O'Leary. 2019. PyWavelets: A Python package for wavelet analysis. *Journal of Open Source Software* 4, 36 (2019), 1237.

[122] Seo-Hyun Lee, Minji Lee, Ji-Hoon Jeong, and Seong-Whan Lee. 2019. Towards an EEG-based intuitive BCI communication system using imagined speech and visual imagery. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 4409–4414.

[123] Seo-Hyun Lee, Minji Lee, and Seong-Whan Lee. 2019. EEG representations of spatial and temporal features in imagined speech and overt speech. In *Asian Conference on Pattern Recognition*. Springer, 387–400.

[124] Seo Hyun Lee, Minji Lee, and Seong Whan Lee. 2020a. EEG representations of spatial and temporal features in imagined speech and overt speech. In *5th Asian Conference on Pattern Recognition, ACPR 2019*. Springer, 387–400.

[125] Seo-Hyun Lee, Minji Lee, and Seong-Whan Lee. 2020b. Neural decoding of imagined speech and visual imagery as intuitive paradigms for BCI communication. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28, 12 (2020), 2647–2659.

[126] Seo-Hyun Lee, Young-Eun Lee, and Seong-Whan Lee. 2022. Toward Imagined Speech based Smart Communication System: Potential Applications on Metaverse Conditions. In *2022 10th International Winter Conference on Brain-Computer Interface (BCI)*. IEEE, 1–4.

[127] M Sande Lemos and BJ Fisch. 1991. The weighted average reference montage. *Electroencephalography and clinical Neurophysiology* 79, 5 (1991), 361–370.

[128] Eric C Leuthardt, Charles Gaona, Mohit Sharma, Nicholas Szrama, Jarod Roland, Zac Freudenberg, Jamie Solis, Jonathan Breshears, and Gerwin Schalk. 2011. Using the electrocorticographic speech network to control a brain–computer interface in humans. *Journal of neural engineering* 8, 3 (2011), 036004.

[129] Eric C Leuthardt, Gerwin Schalk, Jonathan R Wolpaw, Jeffrey G Ojemann, and Daniel W Moran. 2004. A brain–computer interface using electrocorticographic signals in humans. *Journal of neural engineering* 1, 2 (2004), 63.

[130] Ren Li, Jared S Johansen, Hamad Ahmed, Thomas V Ilyevsky, Ronnie B Wilbur, Hari M Bharadwaj, and Jeffrey Mark Siskind. 2018. Training on the test set? an analysis of spampinato et al.[31]. *arXiv preprint arXiv:1812.07697* (2018).

[131] Yichuan Liu and Hasan Ayaz. 2018. Speech recognition via fNIRS based brain signals. *Frontiers in neuroscience* 12 (2018), 695.

[132] Nikos K Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. 2001. Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412, 6843 (2001), 150–157.

[133] Fabien Lotte, Florian Larrue, and Christian Mühl. 2013. Flaws in current human training protocols for spontaneous brain-computer interfaces: lessons learned from instructional design. *Frontiers in human neuroscience* 7 (2013), 568.

[134] Lingxi Lu, Jingwei Sheng, Zhaowei Liu, and Jia-Hong Gao. 2021. Neural representations of imagined speech revealed by frequency-tagged magnetoencephalography responses. *NeuroImage* 229 (2021), 117724.

[135] Steven J Luck. 2014. *An introduction to the event-related potential technique*. MIT press.

[136] Kip A Ludwig, Rachel M Miriani, Nicholas B Langhals, Michael D Joseph, David J Anderson, and Daryl R Kipke. 2009. Using a common average reference to improve cortical neuron recordings from microelectrode arrays. *Journal of neurophysiology* 101, 3 (2009), 1679–1689.

[137] Junshui Ma, Peining Tao, Sevinç Bayram, and Vladimir Svetnik. 2012. Muscle artifacts in multichannel EEG: Characteristics and reduction. *Clinical Neurophysiology* 123 (2012), 1676–1686. Issue 8. DOI:http://dx.doi.org/10.1016/j.clinph.2011.11.083

[138] Joseph G Makin, David A Moses, and Edward F Chang. 2020. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience* 23, 4 (2020), 575–582.

[139] Dimitra Makri, Christina Farmaki, and Vangelis Sakkalis. 2015. Visual fatigue effects on steady state visual evoked potential-based brain computer interfaces. In *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 70–73.

[140] Sebastián Maldonado and Richard Weber. 2009. A wrapper method for feature selection using support vector machines. *Information Sciences* 179, 13 (2009), 2208–2217.

[141] Stephanie Martin, Peter Brunner, Iñaki Iturrate, José del R Millán, Gerwin Schalk, Robert T Knight, and Brian N Pasley. 2016. Word pair classification during imagined speech using direct brain recordings. *Scientific reports* 6, 1 (2016), 1–12.

[142] Elena Marx, Mihaly Benda, and Ivan Volosyak. 2019. Optimal Electrode Positions for an SSVEP-based BCI. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2731–2736.

[143] Mariko Matsumoto and Junichi Hori. 2014. Classification of silent speech using support vector machine and relevance vector machine. *Applied Soft Computing* 20 (2014), 95–102.

[144] Volker Milnik. 2006. Instruction of electrode placement to the international 10-20-system. (2006).

[145] Beomjun Min, Jongin Kim, Hyeong-jun Park, and Boreom Lee. 2016. Vowel imagery decoding toward silent speech BCI using extreme learning machine with electroencephalogram. *BioMed research international* 2016 (2016).

[146] PP Mini, Tessamma Thomas, and R Gopikakumari. 2021. Wavelet feature selection of audio and imagined/vocalized EEG signals for ANN based multimodal ASR system. *Biomedical Signal Processing and Control* 63 (2021), 102218.

[147] Kusuma Mohanchandra and Snehanshu Saha. 2016. A communication paradigm using subvocalized speech: translating brain signals into speech. *Augmented Human Research* 1, 1 (2016), 1–14.

[148] Takahiro Morooka, Kazumi Ishizuka, and Nobuaki Kobayashi. 2018. Electroencephalographic analysis of auditory imagination to realize silent speech BCI. In *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*. IEEE, 683–686.

[149] David A Moses, Matthew K Leonard, Joseph G Makin, and Edward F Chang. 2019. Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nature communications* 10, 1 (2019), 1–14.

[150] David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, and others. 2021. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine* 385, 3 (2021), 217–227.

[151] Sandra Mara Torres Müller, Teodiano Freire Bastos-Filho, and Mário Sarcinelli-Filho. 2011. Using a SSVEP-BCI to command a robotic wheelchair. In *2011 IEEE International Symposium on Industrial Electronics*. IEEE, 957–962.

[152] Gernot Müller-Putz, Reinhold Scherer, Clemens Brunner, Robert Leeb, and Gert Pfurtscheller. 2008. Better than random: a closer look on BCI results. *International journal of bioelectromagnetism* 10 (2008), 52–55.

[153] Brian Murphy, Massimo Poesio, Francesca Bovolo, Lorenzo Bruzzone, Michele Dalponte, and Heba Lakany. 2011. EEG decoding of semantic category reveals distributed representations for single concepts. *Brain and language* 117, 1 (2011), 12–22.

[154] Masaki Nakanishi, Yijun Wang, Xiaogang Chen, Yu-Te Wang, Xiaorong Gao, and Tzyy-Ping Jung. 2017. Enhancing detection of SSVEPs for a high-speed brain speller using task-related component analysis. *IEEE Transactions on Biomedical Engineering* 65, 1 (2017), 104–112.

[155] Noman Naseer and Keum-Shik Hong. 2013. Classification of functional near-infrared spectroscopy signals corresponding to the right-and left-wrist motor imagery for development of a brain–computer interface. *Neuroscience letters* 553 (2013), 84–89.

[156] Thanh Nguyen, Imali Hettiarachchi, Amin Khatami, Lee Gordon-Brown, Chee Peng Lim, and Saeid Nahavandi. 2018. Classification of multi-class bci data by common spatial pattern and fuzzy system. *IEEE Access* 6 (2018), 27873–27884.

[157] Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. 2012. Brain computer interfaces, a review. *Sensors* 12, 2 (2012), 1211–1279.

[158] Soheyl Noachtar and Jan Rémi. 2009. The role of EEG in epilepsy: a critical review. *Epilepsy & Behavior* 15, 1 (2009), 22–33.

[159] Standard Electrode Position Nomenclature. 1991. American electroencephalographic society guidelines for. *Journal of clinical Neurophysiology* 8, 2 (1991), 200–2.

[160] Jussi Numminen and Gabriel Curio. 1999. Differential effects of overt, covert and replayed speech on vowel-evoked responses of the human auditory cortex. *Neuroscience letters* 272, 1 (1999), 29–32.

[161] Anneli Olsen. 2012. The Tobii IVT Fixation Filter Algorithm description.

[162] Rupert Ortner, Brendan Z Allison, Gerd Korisek, Herbert Gaggl, and Gert Pfurtscheller. 2010. An SSVEP BCI to control a hand orthosis for persons with tetraplegia. *IEEE transactions on neural systems and rehabilitation engineering* 19, 1 (2010), 1–5.

[163] Thomas J Oxley, Peter E Yoo, Gil S Rind, Stephen M Ronayne, CM Sarah Lee, Christin Bird, Victoria Hampshire, Rahul P Sharma, Andrew Morokoff, Daryl L Williams, and others. 2021. Motor neuroprosthesis implanted with neurointerventional surgery improves capacity for activities of daily living tasks in severe paralysis: first in-human experience. *Journal of neurointerventional surgery* 13, 2 (2021), 102–108.

[164] Andrew Y Paek, Atilla Kilicarslan, Branislav Korenko, Vladislav Gerginov, Svenja Knappe, and Jose L Contreras-Vidal. 2020. Towards a portable magnetoencephalography based brain computer interface with optically-pumped magnetometers. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 3420–3423.

[165] Erica D Palmer, Howard J Rosen, Jeffrey G Ojemann, Randy L Buckner, William M Kelley, and Steven E Petersen. 2001. An event-related fMRI study of overt and covert word stem completion. *Neuroimage* 14, 1 (2001), 182–193.

[166] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.

[167] Jerrin Thomas Panachakel, AG Ramakrishnan, and TV Ananthapadmanabha. 2020. A novel deep learning architecture for decoding imagined speech from EEG. *arXiv preprint arXiv:2003.09374* (2020).

[168] Jerrin Thomas Panachakel and Angarai Ganesan Ramakrishnan. 2021. Decoding covert speech from EEG-a comprehensive review. *Frontiers in Neuroscience* (2021), 392.

[169] Anne Porbadnigk, Marek Wester, Jan-P Calliess, and Tanja Schultz. 2009. EEG-based speech recognition - Impact of Temporal Effects. (2009).

[170] Cathy J Price, RJS Wise, John DG Watson, Karalyn Patterson, David Howard, and RSJ Frackowiak. 1994. Brain activity during reading The effects of exposure duration and task. *Brain* 117, 6 (1994), 1255–1269.

[171] David Prutchi and Michael Norris. 2005. *Design and development of medical electronic instrumentation: a practical perspective of the design, construction, and test of medical devices*. John Wiley & Sons.

[172] Yunyong Punsawad and Yodchanan Wongsawat. 2012. Motion visual stimulus for SSVEP-based BCI system. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 3837–3840.

[173] Muhammad Naveed Iqbal Qureshi, Beomjun Min, Hyeong-jun Park, Dongrae Cho, Woosu Choi, and Boreom Lee. 2017. Multiclass classification of word imagination speech with hybrid connectivity features. *IEEE Transactions on Biomedical Engineering* 65, 10 (2017), 2168–2177.

[174] Qinwan Rabbani, Griffin Milsap, and Nathan E Crone. 2019. The potential for a speech brain–computer interface using chronic electrocorticography. *Neurotherapeutics* 16, 1 (2019), 144–165.

[175] Herbert Ramoser, Johannes Muller-Gerking, and Gert Pfurtscheller. 2000. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE transactions on rehabilitation engineering* 8, 4 (2000), 441–446.

[176] Maurice Rekrut, Andreas Fey, Matthias Nadig, Johannes Ihl, and Antonio Krüger. 2022. Classifying Words in Natural Reading Tasks Based on EEG Activity to Improve Silent Speech BCI Training in a Transfer Approach. In *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE) (2022 IEEE MetroXRAINE)*. Rome, Italy.

[177] Maurice Rekrut, Tobias Jungbluth, Jan Alexandersson, and Antonio Krüger. 2021. Spinning Icons: Introducing a Novel SSVEP-BCI Paradigm Based on Rotation. In *26th International Conference on Intelligent User Interfaces*. 234–243.

[178] Maurice Rekrut, Abdulrahman Mohamed Selim, and Antonio Krüger. 2022. Improving Silent Speech BCI Training Procedures Through Transfer from Overt to Silent Speech. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2650–2656. DOI:http://dx.doi.org/10.1109/SMC53654.2022.9945447

[179] Maurice Rekrut, Mansi Sharma, Matthias Schmitt, Jan Alexandersson, and Antonio Krüger. 2020. Decoding semantic categories from eeg activity in object-based decision tasks. In *2020 8th International Winter Conference on Brain-Computer Interface (BCI)*. IEEE, 1–7.

[180] Maurice Rekrut, Mansi Sharma, Matthias Schmitt, Jan Alexandersson, and Antonio Krüger. 2021. Decoding Semantic Categories from EEG Activity in Silent Speech Imagination Tasks. In *2021 9th International Winter Conference on Brain-Computer Interface (BCI)*. IEEE, 1–7.

[181] G Reshmi and A Amal. 2013. Design of a BCI system for piloting a wheelchair using five class MI Based EEG. In *2013 Third International Conference on Advances in Computing and Communications*. IEEE, 25–28.

[182] Joe T Rexwinkle, Gregory Lieberman, Matthew Jaswa, and Brent J Lance. 2019. Development of a Game with a Purpose for Acquisition of Brain-Computer Interface Data. *arXiv preprint arXiv:1910.00106* (2019).

[183] Anaum Riaz, Sana Akhtar, Shanza Iftikhar, Amir Ali Khan, and Ahmad Salman. 2014. Inter comparison of classification techniques for vowel speech imagery using EEG sensors. In *The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014)*. IEEE, 712–717.

[184] Simanto Saha and Mathias Baumert. 2020. Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: a review. *Frontiers in computational neuroscience* 13 (2020), 87.

[185] M Saifudinova, M Bachmann, J Lass, and H Hinrikus. 2015. Effect of coffee on EEG spectral assymmetry. In *World Congress on Medical Physics and Biomedical Engineering, June 7-12, 2015, Toronto, Canada*. Springer, 1030–1033.

[186] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. 2017. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping* 38, 11 (2017), 5391–5420.

[187] Donald L Schomer and Fernando Lopes Da Silva. 2012. *Niedermeyer's electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins.

[188] Geoffrey D Schott. 1993. Penfield's homunculus: a note on cerebral cartography. *Journal of neurology, neurosurgery, and psychiatry* 56, 4 (1993), 329.

[189] Elizabeth R Schotter, Bernhard Angele, and Keith Rayner. 2012. Parafoveal processing in reading. *Attention, Perception, & Psychophysics* 74, 1 (2012), 5–35.

[190] J Anthony Seikel, David G Drumright, and Douglas W King. 2015. *Anatomy & physiology for speech, language, and hearing*. Cengage Learning.

[191] Alborz Rezazadeh Sereshkeh, Robert Trott, Aurélien Bricout, and Tom Chau. 2017a. EEG classification of covert speech using regularized neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 12 (2017), 2292–2300.

[192] Alborz Rezazadeh Sereshkeh, Robert Trott, Aurélien Bricout, and Tom Chau. 2017b. Online EEG classification of covert speech for brain–computer interfacing. *International journal of neural systems* 27, 08 (2017), 1750033.

[193] Alborz Rezazadeh Sereshkeh, Rozhin Yousefi, Andrew T Wong, and Tom Chau. 2018. Online classification of imagined speech using functional near-infrared spectroscopy signals. *Journal of neural engineering* 16, 1 (2018), 016005.

[194] SS Shergill, DK Tracy, M Seal, K Rubia, and P McGuire. 2006. Timing of covert articulation: An fMRI study. *Neuropsychologia* 44, 12 (2006), 2573–2577.

[195] Revati Shriram, M Sundhararajan, and Nivedita Daimiwal. 2013. EEG based cognitive workload assessment for maximum efficiency. *Int. Organ. Sci. Res. IOSR* 7 (2013), 34–38.

[196] Linda I Shuster and Susan K Lemieux. 2005. An fMRI investigation of covertly and overtly produced mono-and multisyllabic words. *Brain and language* 93, 1 (2005), 20–31.

[197] Irina Simanova, Marcel Van Gerven, Robert Oostenveld, and Peter Hagoort. 2010. Identifying object categories from event-related EEG: toward decoding of conceptual representations. *PloS one* 5, 12 (2010), e14465.

[198] Filip Škola, Simona Tinková, and Fotis Liarokapis. 2019. Progressive training for motor imagery brain-computer interfaces using gamification and virtual reality embodiment. *Frontiers in human neuroscience* 13 (2019), 329.

[199] J David Smith, Margaret Wilson, and Daniel Reisberg. 1995. The role of subvocalization in auditory imagery. *Neuropsychologia* 33, 11 (1995), 1433–1454.

[200] Mojtaba Soltanlou, Maria A Sitnikova, Hans-Christoph Nuerk, and Thomas Dresler. 2018. Applications of functional near-infrared spectroscopy (fNIRS) in studying cognitive development: The case of mathematics and language. *Frontiers in psychology* 9 (2018), 277.

[201] Patrick Suppes, Bing Han, and Zhong-Lin Lu. 1998. Brain-wave recognition of sentences. *Proceedings of the National Academy of Sciences* 95, 26 (1998), 15861–15866.

[202] Patrick Suppes, Zhong-Lin Lu, and Bing Han. 1997. Brain wave recognition of words. *Proceedings of the National Academy of Sciences* 94, 26 (1997), 14965–14969.

[203] Guillaume Thierry. 2008. Neurophysiological examination methods: electrophysiology and neuroimaging. *Cognitive Neurology: A clinical textbook* (2008).

[204] Marieke E Thurlings, Jan BF van Erp, Anne-Marie Brouwer, and Peter J Werkhoven. 2010. EEG-based navigation from a human factors perspective. In *Brain-computer interfaces*. Springer, 71–86.

[205] Praveen Tirupattur, Concetto Spampinato, Yogesh Singh Rawat, and Mubarak Shah. 2018. ThoughtViz: Visualizing human thoughts using generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*. 950–958. DOI:http://dx.doi.org/10.1145/3240508.3240641

[206] Juan Manuel Mayor Torres, Evgeny A Stepanov, and Giuseppe Riccardi. 2016. Eeg semantic decoding using deep neural networks. In *Rovereto Workshop on Concepts, Actions, and Objects (CAOS)*.

[207] Alejandro A Torres-García, Carlos A Reyes-García, Luis Villaseñor-Pineda, and Gregorio García-Aguilar. 2016. Implementing a fuzzy inference system in a multi-objective EEG channel selection model for imagined speech classification. *Expert Systems with Applications* 59 (2016), 1–12.

[208] Pascale Tremblay and Anthony Steven Dick. 2016. Broca and Wernicke are dead, or moving past the classic model of language neurobiology. *Brain and language* 162 (2016), 60–71.

[209] Lau Troy M, Gwin Joseph T, and Ferris Daniel P. 2012. How many electrodes are really needed for EEG-based mobile brain imaging? *Journal of Behavioral and Brain Science* 2012 (2012).

[210] Akihiko Tsukahara, Masayuki Yamada, Keita Tanaka, and Yoshinori Uchikawa. 2019. Analysis of EEG Frequency Components and an Examination of Electrodes Localization during Speech Imagery. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 4698–4702.

[211] István Ulbert, Eric Halgren, Gary Heit, and George Karmos. 2001. Multiple microelectrode-recording system for human intracortical applications. *Journal of neuroscience methods* 106, 1 (2001), 69–79.

[212] Anirudh Vallabhaneni, Tao Wang, and Bin He. 2005. Brain—computer interface. In *Neural engineering*. Springer, 85–121.

[213] Jacques J Vidal. 1973. Toward direct brain-computer communication. *Annual review of Biophysics and Bioengineering* 2, 1 (1973), 157–180.

[214] Eric J Wadkins. 2019. *A continuous silent speech recognition system for AlterEgo, a silent speech interface*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[215] Christoph Wagner, Petr Schaffer, Pouriya Amini Digehsara, Michael Bärhold, Dirk Plettemeier, and Peter Birkholz. 2022. Silent speech command word recognition using stepped frequency continuous wave radar. *Scientific Reports* 12, 1 (2022), 1–12.

[216] Michael Wand, Jan Koutník, and Jürgen Schmidhuber. 2016. Lipreading with long short-term memory. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6115–6119.

[217] Michael Wand and Tanja Schultz. 2011. Session-independent EMG-based Speech Recognition.. In *Biosignals*. 295–300.

[218] Cong Wang, Bin Xia, Jie Li, Wenlu Yang, Alejandro Cardona Velez, Hong Yang, and others. 2011. Motor imagery BCI-based robot arm system. In *2011 Seventh International Conference on Natural Computation*, Vol. 1. IEEE, 181–184.

[219] Li Wang, Xiong Zhang, Xuefei Zhong, and Yu Zhang. 2013. Analysis and classification of speech imagery EEG for BCI. *Biomedical signal processing and control* 8, 6 (2013), 901–908.

[220] Elizabeth K Warrington and Tim Shallice. 1984. Category specific semantic impairments. *Brain* 107, 3 (1984), 829–853.

[221] Hiroki Watanabe, Hiroki Tanaka, Sakriani Sakti, and Satoshi Nakamura. 2020. Synchronization between overt speech envelope and EEG oscillations during imagined speech. *Neuroscience research* 153 (2020), 48–55.

[222] C Wernicke. 1874. The symptom complex of aphasia. A psychological study on an anatomical basis (Translated from German; GH Eggert, Trans.). *Wernicke's work on aphasia: A sourcebook and review* (1874).

[223] Marek Wester. 2006. Unspoken speech-speech recognition based on electroencephalography. *Master's Thesis, Universitat Karlsruhe (TH)* (2006).

[224] Irene Winkler, Stefan Debener, Klaus-Robert Müller, and Michael Tangermann. 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, 4101–4105.

[225] Marc E Wolf, Anne D Ebert, and Anastasios Chatzikonstantinou. 2017. The use of routine EEG in acute ischemic stroke patients without seizures: generalized but not focal EEG pathology is associated with clinical deterioration. *International Journal of Neuroscience* 127, 5 (2017), 421–426.

[226] Wei Wu, Xiaorong Gao, and Shangkai Gao. 2006. One-versus-the-rest (OVR) algorithm: An extension of common spatial patterns (CSP) algorithm to multi-class case. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, 2387–2390.

[227] Madoka Yamazaki, Don Tucker, Marie Terrill, Ayataka Fujimoto, and Takamichi Yamamoto. 2013. Dense array EEG source estimation in neocortical epilepsy. *Frontiers in neurology* 4 (2013), 42.

[228] Jianhua Yang, Harsimrat Singh, Evor L Hines, Friederike Schlaghecken, Daciana D Iliescu, Mark S Leeson, and Nigel G Stocks. 2012. Channel selection and classification of electroencephalogram signals: an artificial neural network and genetic algorithm-based approach. *Artificial intelligence in medicine* 55, 2 (2012), 117–126.

[229] Erwei Yin, Zongtan Zhou, Jun Jiang, Yang Yu, and Dewen Hu. 2014. A dynamically optimized SSVEP brain–computer interface (BCI) speller. *IEEE Transactions on Biomedical Engineering* 62, 6 (2014), 1447–1456.

[230] Seung-Schik Yoo, Ty Fairneny, Nan-Kuei Chen, Seh-Eun Choo, Lawrence P Panych, HyunWook Park, Soo-Young Lee, and Ferenc A Jolesz. 2004. Brain–computer interface using fMRI: spatial navigation by thoughts. *Neuroreport* 15, 10 (2004), 1591–1595.

[231] Yang Yu, Yadong Liu, Jun Jiang, Erwei Yin, Zongtan Zhou, and Dewen Hu. 2018. An asynchronous control paradigm based on sequential motor imagery and its application in wheelchair navigation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26, 12 (2018), 2367–2375.

[232] Thorsten O Zander and Sabine Jatzev. 2009. Detecting affective covert user states with passive brain-computer interfaces. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 1–9.

[233] Thorsten O Zander, Christian Kothe, Sabine Jatzev, and Matti Gaertner. 2010. Enhancing human-computer interaction with input from active and passive brain-computer interfaces. In *Brain-computer interfaces*. Springer, 181–199.

[234] Yu Zhang, Guoxu Zhou, Jing Jin, Xingyu Wang, and Andrzej Cichocki. 2015. SSVEP recognition using common feature analysis in brain–computer interface. *Journal of neuroscience methods* 244 (2015), 8–15.

[235] Shunan Zhao and Frank Rudzicz. 2015. Classifying phonological categories in imagined and articulated speech. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 992–996.