The Efficacy of Machine Learning for Imagined & Inner Speech Decoding from EEG: a Multimodal Analysis

Riccardo Bonzano

MSc Data Science

University of Bath

September 2022

The Efficacy of Machine Learning for Imagined & Inner Speech Decoding from EEG: a Multimodal Analysis

submitted by

Riccardo Bonzano

for the degree of MSc Data Science at the

University of Bath

September 2022

Copyright

Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of Master of Science in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. This project has been conducted with the help of two other University of Bath master's students: Will Adkins and Ruthwik Hosur Paramashivaiah. In particular, the data collection, experimental paradigm and main framework of investigation have been developed collaboratively. Except where specifically acknowledged, this dissertation is the work of the author.

Abstract

Machine learning (ML) is a promising tool for speech decoding and the practical implementation of brain-computer interfaces (BCI). Of the many technologies available, a lot of attention has been directed towards EEG due to its being a non-invasive yet effective method to measure brain activity. In this study we develop a comprehensive experimental paradigm for the study of two of the most important areas of research in EEG-based BCI: imagined and inner speech. We assess which of these paradigms is most suitable for the realization of BCI by analysing the performance of three common ML algorithms: random forests (RF), support-vector machines (SVM) and k-nearest neighbours (KNN). Furthermore, we explore the influence of stimulus modality, word complexity and repeated imagination on speech decoding accuracy.

The output of this project is publicly available for consultation. The code developed and the data collected can be found respectively here and here.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisors, James and Scott, for their help and support throughout the project. A special mention goes to my colleagues Will and Ruthwik, who made this work even more interesting and rewarding.

Chapter 1

Introduction

1.1 Can we read the mind?

In many movies, the ability to 'read the mind' is one of the most powerful and sought-after superpowers. Likewise, mythologies and folklore traditions all over the world feature gods, beings or gifted individuals that can perform telepathy. Clearly, there is something inherently fascinating linked to being able to know what another person is thinking. Science has long held this ambition too (Panachakel and Ramakrishnan, 2021). In fact, the very origins of one of the most popular techniques to measure brain activity, the electroencephalogram (EEG), is closely tied to the quest to read the human mind. The method was invented in 1924 by German neurologist Hans Berger, who was trying to build a machine for synthetic telepathy (Kaplan (2011),Keiper (2013)).Today, almost a hundred years later, we are getting closer and closer to the realization of that machine: brain-computer interfaces (BCI).

1.2 How we all think differently and why it matters

Before delving into the technical details underlying BCI, we will briefly outline what 'reading the mind' entails in practice. BCI aim to translate the electrical activity of the brain into an accurate representation of the corresponding mental action or experience. For example, the experience of seeing an ice-cream gives rise to a unique electrophysiological response in the brain, with different neuronal areas firing simultaneously and giving rise to specific brain

waves. Quite intuitively, hearing a dog barking will cause a completely different response. But what about reading about an ice-cream on a piece of scientific writing? Or hearing the word ice-cream pronounced out loud? To what extent there is a similarity between the brain signals elicited by these ice-cream related events is an open question.

On the other hand, reversing this logic gives rise to interesting considerations as well. When humans think of something, they do it in many different ways, or modalities. Thinking in pure meanings - talking within ourselves with our own 'inner voice' - is defined in the scientific literature as 'inner speech' (Nieto et al., 2022). This is perhaps the most common, instinctive and pervasive way of thinking. Some people tend however to think more in visual clues, like shapes or pictures, while others prefer sounds or written text (not to mention more complicated sensory recollection modes such as taste, smell, tactile or emotional sensations) (Lee et al., 2019). Nonetheless, the most researched modality in the field of speech decoding BCI is a much less intuitive one, termed 'imagined speech'. This consists in imagining to move the mouth muscles to speak without performing any physical movement (i.e., articulatory motor imagery). Summarising the above more formally, thoughts can consist of internally-generated auditory, textual and visual imagery, inner speech and imagined speech, among other media Lee et al. (2019). Going back to our ice-cream example, thinking about an ice-cream might occur in any of the following ways: 1) recalling the image of an ice-cream (visual imagery); 2) visualizing the word 'ice-cream' as if it were written down (textual imagery); 3) hearing the word as if someone spoke it out loud (auditory imagery); 4) thinking about the concept in abstract terms (inner speech); 5) thinking about how one would move their mouth to pronounce the word 'ice-cream' (imagined speech).

The points above have important consequences for the practical realization of BCI, as they highlight the requirement that this technology be flexible and adaptable to different ways of perceiving and thinking about the world (Lee et al., 2019). In other words, even in principle, an effective BCI needs to be a multisensory BCI that can bridge different thought modalities and, ideally, interpret the nuances unique to how each one of us think.

1.3 The applications of BCI

A part from the archetypical curiosity and academic interest sparked by the prospect of 'reading the mind', BCI communication systems would have endless applications. Most importantly, they would accommodate groundbreaking advancements in healthcare, providing an alternative way of communication to those that have lost the ability to speak (Panachakel and Ramakrishnan, 2021). However, the applicability of BCI is not limited to speech, and, more generally, these applications would allow users to translate any thought into action. Coupled with the development of smart homes and the Internet or Things, this technology has the potential to enable people to control their surrounding environment with their minds, from changing the temperature of the room to controlling a robotic arm or even browsing the internet on a laptop (Abdulkader, Atia and Mostafa, 2015).

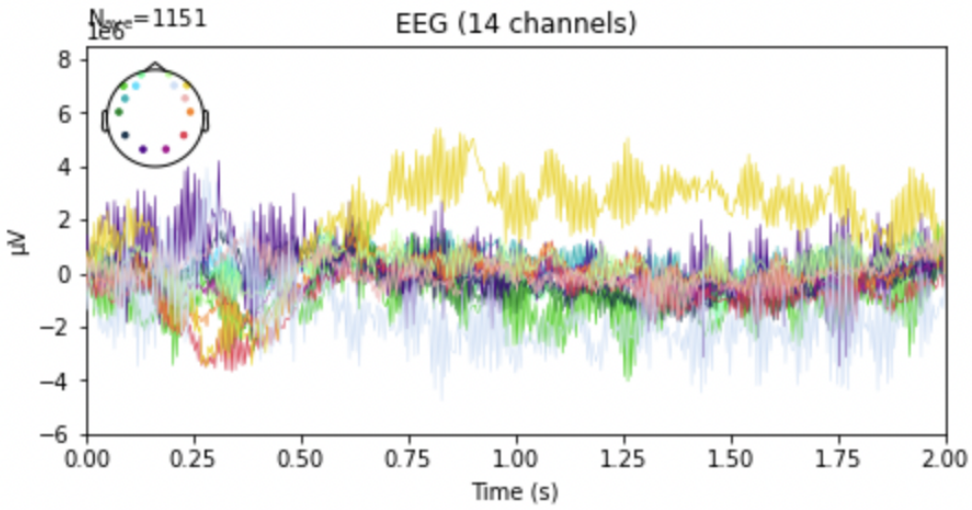
Chapter 2

Background

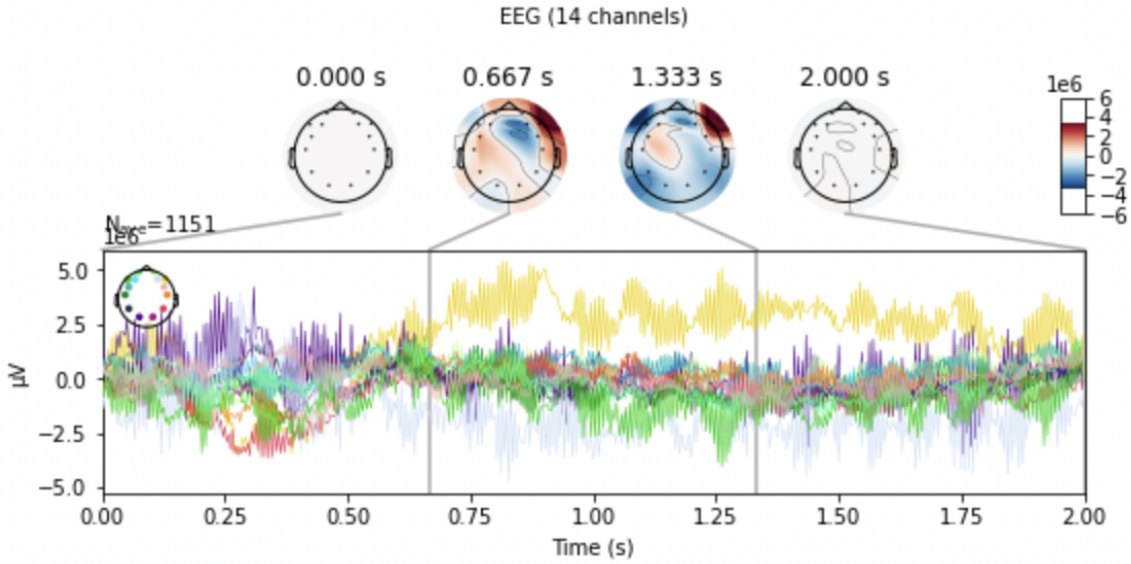
In this chapter, we give an overview of the technology that underlies BCI, and contextualize our line of investigation with respect to the relevant literature. We focus on imagined and inner speech decoding.

2.1 Electroencephalography

Through electrodes placed on the scalp, EEG measures brain activity by recording the voltage fluctuations related to neuronal ionic currents (Kirschstein and Köhling, 2009). The outcome of these measurement takes the form of non-periodic and highly noisy signals, with one signal corresponding to each electrode. An example of EEG data can be observed in Fig. 2.1 below. Notably, EEG signals of different subjects look very different. This is mainly due to people having unique facial and cranial features, which tend to have a much larger impact on EEG readings than the similarity associated to the occurrence of identical brain activity. This inter-subject dissimilarity is so pronounced that a whole landscape of studies into EEG-driven subject identification has developed in recent years (Del Pozo-Banos et al., 2014).



(a)

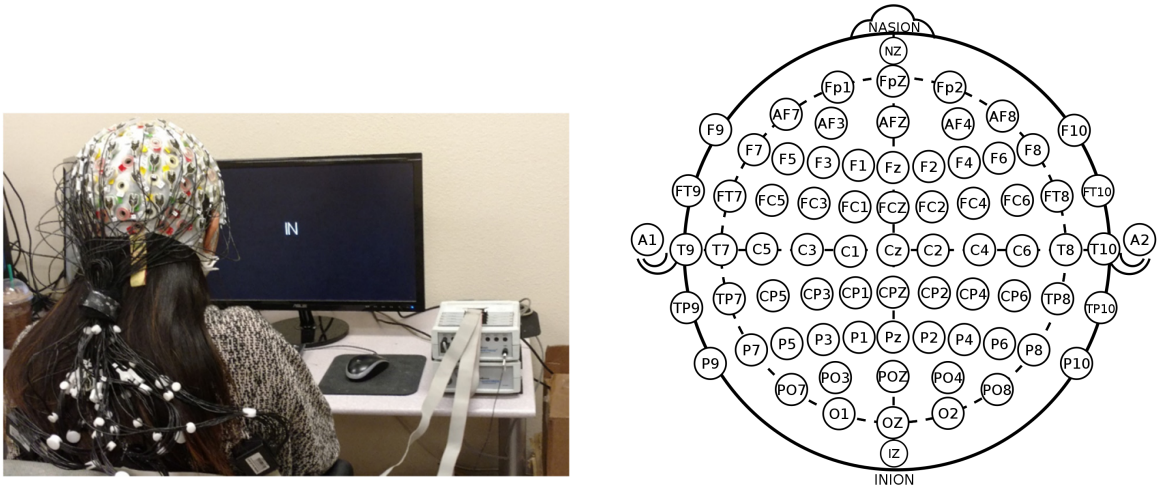


(b)

Figure 2.1: EEG signals of imagined speech from P2. The y axis features voltage in microvolts. Each colored line corresponds to a specific EEG electrode, as can be observed on the top right of graph a). On the top of graph b), we can see scalp topographies, where each black dot is an electrode. These allow to visualize which areas of the brain are active at a particular point in time. At $t = 0$ and $t = 2$ the voltages are set to zero.

While numerous other neuroimaging methods exist, such as electrocorticography (ECoG), functional magnetic resonance imaging (fMRI) and functional near-infrared spectroscopy (fNIRS), most of BCI research focuses on electroencephalography (EEG) (Panachakel and Ramakrishnan, 2021). This is because, although not the best method in terms of sheer

data resolution (especially due to its low signal-to-noise ratio (SNR) and low spectral and spatial resolution), its being non-invasive, easily scalable and relatively cheap makes EEG very promising in terms of scope and applicability (Panachakel and Ramakrishnan, 2021). For a detailed treatment of EEG advantages, disadvantages and limitations compared to other techniques please consult Panachakel and Ramakrishnan (2021). A research-grade EEG device and the positioning of its electrodes can be observed in Fig. 2.2.



(a) **(b)**

Figure 2.2: a) Research-grade 64-channel EEG device. Taken from Nguyen, Karavas and Artemiadis (2017). b) 64 EEG electrodes arrangement according to the 10-20 international system. Taken from here, last accessed on 06/09/2022.

2.2 Imagined speech

As mentioned in the introduction, imagined speech (or covert speech) is the most researched paradigm in the field of EEG-based BCI, with 28 relevant studies identified in a recent review article (Panachakel and Ramakrishnan, 2021). Its popularity is due to its communication potential being virtually unbounded, since any thinkable word can cause a distinct and, at least theoretically, machine-interpretable brain signal. Other methods, albeit successful, such as those based on P300 speller systems (Arvaneh, Robertson and Ward, 2019; Lu et al., 2019; Al-Nuaimi et al., 2020), steady state visually evoked potentials (SSVEP) (Han et al., 2018; Ojha and Mukul, 2021), event-related potentials (ERP) (Rapin et al., 2018; Fouad et al., 2020) and motor imagery (Onose et al., 2012; Kevric and Subasi, 2017), are constrained by

their finite and relatively small number of degrees of freedom (e.g., for motor imagery, the participant could think of moving up, down, left, right, backwards and forwards, allowing for only 6 communicative actions).

Given the vastness of the research available on imagined speech, giving a full overview of the state of the art lies beyond the scope of this work. What follows is a brief compendium of the research which this study was the most informed by. Relevant methodology details are omitted and will be discussed in Chapter 4. As laid out in the Introduction, it is opinion of the author that a truly effective BCI needs to be multimodal in order to deal with the intrinsic variability of how different people think. For this reason, we will focus on studies that investigated imagined speech along with other modalities.

The KARA-ONE open-source dataset was specifically developed with the purpose of learning multimodal relationships (Zhao and Rudzicz, 2015). It collected data for 3 modalities: 1) auditory response during speech production, 2) articulatory motor response during speech production, and 3) imagined speech. They investigated 11 prompts in total: 7 phonemic/syllabic prompts (/iy/, /uw/, /piy/, /tiy/, /diy/, /m/, /n/) and 2 pairs of monosyllabic and phonetically similar words ('pat', 'pot', 'knew', and 'gnaw'). To perform binary classification between different phonetic and phonological classes (e.g., vowel-only vs. consonant), they used data from all 3 modalities either separately or in conjuction. With the use of a deep-belief network (DBN), they achieved accuracy of up to 90%, vastly outperforming baseline support-vector machines (SVM) approaches. Notably, they used the fairly lightweight 64-channel Neuroscan Quick-cap, giving good indication commercial-grade EEG devices can obtain excellent results in speech decoding. This is confirmed by the findings of Clayton et al. (2020), who collected the FEIS dataset using the 14-channel Emotiv EPOC+ headset. This is the device we used in this study, which details will be provided in Section 4.1.1.

Another work which strongly informed this study is the paper by Lee et al. (2019), who conducted a multimodal investigation into imagined speech and visual imagery. Keeping real-world applicability as a strong research motivation, they chose 12 prompts from healthcare communication boards that are commonly used in hospitals with paralyzed/aphasia patients ('ambulance', 'clock', 'hello', 'help me', 'light', 'pain', 'stop', 'thank you', 'toilet', 'TV', 'water', and 'yes'). The speaking action of each word was repeated multiple times during each

measurement trial; this techniques is termed 'repeated imagination' and will be covered in 4.1.4. during each Random forests (RF) and shrinkage regularized linear discriminant analysis (RLDA) were their classifiers of choice. The most important ML classification task involved a 13-class classification (12 words plus a rest class) between the prompts. They obtained a mean accuracy of $15.9 \pm 3.8\%$ (shrinkage RLDA) and $20.4 \pm 7.7\%$ (RF) for imagined speech, and $17.0 \pm 3.4\%$ (shrinkage RLDA) and $22.2 \pm 4.3\%$ (RF) for visual imagery. Notably, the differences between both the two speech paradigms and the two classifiers are statistically significant. These are results of great importance for a variety of reasons: 1) they show that it is possible to achieve high classification performance for more than 10 classes for both imagined speech and visual imagery; 2) they show that an array of commonly used words can be decoded successfully for both modalities, despite their dissimilarity across multiple dimensions (i.e., phonetics, semantics and complexity, where complexity is quantified by the number of word syllables); 3) they show that visual imagery is superior to imagined speech in terms of machine-interpretability for RF and RLDA.

The last work we will discuss in this section is by Nguyen, Karavas and Artemiadis (2017). Albeit not strictly a multimodal study by the definition used so far, it does investigate the impact of a variety of word characteristics on ML speech decoding accuracy: semantics, phonetics and complexity. We regard this pursue very worthwhile, as the flexibility requirements of the ideal BCI system will demand exploration into the effects of word variability. They choose three main categories of prompts: vowel phonemes (/a/, /i/ and /u/), short words ('in', 'out' and 'up') and long words ('cooperate' and 'independent'). They use an innovative approach based on Riemannian manifold features for data preprocessing, and employ relevance vector machines (RVM) to perform classification. They achieve impressive results with accuracy up to 70% and 95% for 3-class and binary classification respectively. With regards to complexity, they find that longer words are classified more accurately, probably due to the the strength of EEG signatures being positively correlated to the number of syllables. This is confirmed by Fujiwara, Miyasaka and Sakamoto (2018). With regards to phonetics, performance was very similar between vowels and short words, suggesting that phonetics plays a more substantial role than semantics. Notably, they express interest into expanding their paradigm to include other modalities such as visual and motor imagery in their future works. Once again, this

confirms the relevance of multimodal BCI applications.

2.3 Inner speech

While imagined speech has been thoroughly explored by the scientific community, it seems that inner speech is only now starting to receive the attention it deserves. Arguably, it is a much more natural thinking modality that most humans use all the time by default. Imagined speech, on the contrary, is quite cumbersome and poses stress on the user (M. Rashid, 2020). We believe that this fact alone makes inner speech the most suitable candidate for BCI realization, as it would allow for an intuitive and truly integrated communication system. This opinion is supported by the literature (Panachakel and Ramakrishnan, 2021; Martin et al., 2018).

To our knowledge, only one publicly available dataset for inner speech exists (Nieto et al., 2022), which was published only a year prior to the time of writing. They used a device with 128 active EEG channels and 8 external active electroculography (EOG) / electromiography (EMG) channels, with a sampling frequency of 1024 Hz. The prompts of choice were 4 Spanish words designating spatial directions: "arriba", "abajo", "derecha", "izquierda" (i.e., "up", "down", "right", "left"). For completeness, we mention that a bimodal EEG-fMRI open-source inner speech dataset also exists (Gupta et al., 2022).

We could find only two modern studies which decode inner speech: van den Berg, van Donkelaar and Alimardani (2021) and Jonsson (2022), of which the latter is a Master's thesis. Both use the dataset by Nieto et al. (2022) and implement a 2D Convolutional Neural Network (CNN) based on the EEGNet architecture, achieving very similar results at $\sim 29\%$ accuracy (for reference, chance level is 25%). Further studies that explore inner speech were conducted in the past, however none utilized ML (Suppes, Lu and Han, 1997; Deng et al., 2010; D'Zmura et al., 2009). Additionally, similarly to a large fraction of modern studies on imagined speech, they used auditory cues to trigger the inner speaking action, which is not ideal. The reasons behind this will be explained in Section 4.1.4.

The reasons why inner speech is only now starting to be explored are not clear. It might be because its decoding is expected to be less effective. In fact unlike imagined speech, the phonological characteristics associated to spoken speech brain activity are not retained

(Nieto et al., 2022). Furthermore, thinking in abstract terms seems to be more complex than articulatory motor imagery in terms of neuronal activation, as it involves multiple processes such as phonological and semantic analysis, among others (Pei et al., 2011; Indefrey and Levelt, 2004). Finally, we suspect that the sloppiness with which terminology is used in most of the literature might be a strong factor at play. As we experienced first-hand when explaining to participants how to perform the experiment, imagined speech is not an intuitive action at all. Numerous papers lack descriptive rigour, and it may very well be that many studies that state to be studying imagined speech are in reality studying inner speech. The opposite is certainly true, as we found several articles that use the term inner speech when they are referring to imagined speech (Simistira Liwicki et al., 2022; Fujiwara, Miyasaka and Sakamoto, 2018; Kiroy et al., 2022).

Chapter 3

Motivation

In this section, we state the main objectives and research hypotheses that motivated this work.

As mentioned above, while imagined speech has been extensively investigated, we could find only four modern studies that have focus on inner speech. Since we firmly believe that inner speech is a better candidate for the realization of EEG-based BCI, we aim to fill the above gap in the literature by fulfilling the following objectives.

- **O1.** To verify whether commercial-grade EEG devices are suitable for the decoding of inner speech.
- **O2.** To develop a rigorous and comprehensive EEG experimental paradigm for the study and comparison of inner speech and imagined speech.
- **O3.** To provide a publicly available EEG dataset that will allow other researchers to perform a wide array of studies on inner and imagined speech, including the impact of phonetics, complexity and repeated imagination on decoding performance, with a focus on real-world applicability.
- **O4.** To investigate the efficacy of different ML approaches (ML, RF and KNN) for both imagined and inner speech, and assessing which paradigm is more suitable for the practical realization of EEG-based BCI.

We here expand upon the reasons why these objectives constitute a valuable line of investigation and proceed by formulating our research hypotheses.

Following the approach by Lee et al. (2019), our choice of prompts was guided by real-world applicability in the context of healthcare. The words were also chosen to maximise the number of testable research hypotheses, including analysis of the relationship between ML performance and prompts characteristics such as phonetics, complexity and the effects of repeated imagination. This makes our work innovative in many ways. First, no open-source dataset exists nor any study has ever been conducted on EEG inner speech data that features words with real applications in healthcare. Furthermore, while it is known that phonetics and complexity influence EEG signatures, this has never been investigated in the context of inner speech. Repeated imagination is also associated to a strengthening of EEG signatures (Panachakel and Ramakrishnan, 2021). Whether this is significant for the decoding of either imagined or inner speech, however, has never been experimentally verified. Finally, the only inner speech dataset by Nieto et al. (2022) was collected with a high-end 128 electrodes EEG device. Since we are using a portable and relatively inexpensive headset, this project also explores the suitability of commercial-grade devices to decode inner speech. Combining the above with the notions outlined in our literature review, we state the following research hypotheses:

- **H1.** Commercial-grade EEG devices are suitable for the decoding of inner speech.
- **H2.** The performance of ML classifiers is better for imagined speech than for inner speech.
- **H3.** The performance of ML classifiers is positively correlated to complexity and repeated imagination for both inner and imagined speech.

As we have stressed in the previous sections, we hold the opinion that the ideal BCI needs to be multimodal at its core. Although supported by the literature (Panachakel and Ramakrishnan, 2021), to our knowledge no specific investigation in this area has been conducted. Thus, we decided to develop a suitable research hypothesis to test this belief.

For a BCI to be suitable for practical application, it needs to be able to decode thought in real-world conditions. In these conditions, humans do not think in watertight compartments, and different modalities of thinking might quickly alternate or even overlap one another. As a concrete example, when one thinks in abstract terms (inner speech) and happens to say

the word 'cat' in their mind, the image of a cat might arise to their consciousness (visual imagery), either substituting the inner voice or overlapping with it. Analogously, when one sees a cat (visual stimulus) the abstract thought of a cat (inner speech) might suddenly pop up in their mind. In order to verify whether these mechanisms play a substantial role in speech decoding, we developed an experimental paradigm which aims at mimicking the multimodal characteristics of real-world conditions. Its objective is to allow the quantification of how different modalities of stimulus impact speech decoding. We state our research hypothesis as follows:

- **H4.** The performance of ML classifiers for both imagined and inner speech is impacted by the stimulus modality of prompts.

If the above is discarded, then imagined and inner speech actions are not influenced by the modality of the triggering cue. This would provide evidence against a strong need for multimodal BCIs. While the impact of multimodal stimulation on classification accuracy has been researched for a P300-based BCI (Aloise et al., 2007), to our knowledge no study of this kind has ever been conducted in the context of imagined or inner speech.

Chapter 4

Methodology

In this section we detail the methodology we followed for data collection, design of the experimental paradigm, preprocessing and machine learning implementation.

4.1 Data collection

4.1.1 Device & set-up

The EEG device we used is the mobile, scalable, commercial-grade 14-channels EPOC+ device by Emotiv, which can be observed in Fig. 4.1 below (user manual and full specifications can be found here). It features 16 copper electrodes, of which two are CMS/DRL ground reference electrodes. Their positioning is in accordance to the 10-20 international system and is designed to cover the 'F3', 'FC5', 'AF3', 'F7', 'T7','P7','O1','O2','P8','T8','F8' 'AF4', 'FC6', 'F4', 'P3' and 'P4' areas of the human skull, where the last two are where the reference electrodes are placed. The sampling rate used was 256 Hz.

The data was collected via the personal computers of the researchers. The participants were instructed to sit upright facing the laptop screen, which was set up at eye-level at a distance of $\sim$ 40 cm. Little light was allowed into the room and windows and doors were closed to minimize potential sources of distraction.

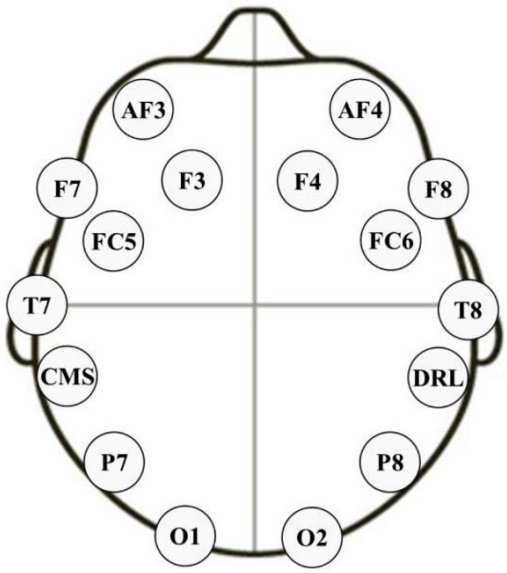
Communication between the Emotiv device and the computer occurred via the OpenVibe software interface, which connected to the Emotiv wireless USB dongle. The experimental

paradigm and data collection were realised respectively with OpenVibe designer and OpenVibe acquisition server. To run the latter, a Lua script was developed. Calibration to ensure suitable connectivity of all the electrodes was performed via the Emotiv interface.

(a)

(b)

(c)



(d)

Figure 4.1: 14-channel EMOTIV EPOC+ EEG Headset. d) shows the positioning of the electrodes, which follows the 10-20 international system. a), b) and c) are own figures. d) is taken from here, last accessed on 06/09/2022.

4.1.2 Subjects

Nine subjects of age between 20 and 30 years old were recruited. All were right-handed and healthy, with no history of neurological diseases. They were informed on the scope of the study

via an information sheet and gave their consent by signing a participant consent form before the start of the experiment. To ensure that they were confident with the actions required, we instructed them on how to perform imagined or inner speech and familiarized them with the relevant auditory, visual and textual prompts. Four participants performed inner speech and five performed imagined speech. The experimental set-up and paradigm were identical for both groups. Only four participants sat the experiment fully, although all but one completed at least 80%. One participant had to interrupt the experiment as they felt nauseous.

4.1.3 Word prompts

The choice of words was guided by **O3** and **O4**, which both demanded a large enough set of datapoints for the ML to work optimally. As we were constricted by a time limit of ~ 90 minutes per experimental session and by the impossibility of conducting multiple sessions per participant, we opted for 4 word prompts: 'ambulance', 'clock', 'hospital', 'lamp' (from now on, we will refer to these as the "words"), plus a rest class. The choice of words was also informed by **O2**. More specifically, the selection was carried out in light of 5 characteristics or dimensions: applicability, syntax, semantics, phonetics and complexity. We shall consider each in turn.

- Applicability: two words ('ambulance', 'clock') were taken from the set used by Lee et al. (2019), to which we added 'hospital' and 'lamp'. These words have clear practical applications and could constitute part of an essential vocabulary for the patients' communication (Patak et al., 2006).

- Syntax: all words are nouns. This eliminates any potential variability that might originate from how different syntactic categories are processed by the brain (Gansonre et al., 2018). This is an improvement over many imagined speech paradigms in the literature which do not fix the syntactic dimension (e.g., Lee et al. (2019); Nguyen, Karavas and Artemiadis (2017) use a mixture of nouns, verbs and adjectives).

- Complexity: 'lamp' and 'clock' are monosyllabic (we will refer to these as "short words") while 'hospital' and 'ambulance' are trisyllabic (we will refer to these as "long words"). Ideally, tetrasyllabic words (as in Nguyen, Karavas and Artemiadis (2017)) would have been a better

choice, but we could not find any that also satisfied the other selection criteria.

- Phonetics: 'ambulance' and 'lamp' have the /æ/ vowel, while 'hospital' and 'clock' both have the /ɒ/ vowel. Both are monophthongs. Note that only the vowel nucleus of the monosyllabic words and the first syllable of the trisyllabic words informed our selection. This is similar to the method adopted by Nguyen, Karavas and Artemiadis (2017).

We chose the auditory, visual and textual stimulus modalities. This resulted in a total of 13 different classes of prompts (4 words per 3 modalities + 1 rest class). For example, the prompts included an image of a clock ('image-clock' class), the word 'lamp' written in text ('text-lamp' class), an audio recording of a person saying 'hospital' ('audio-hospital' class).

4.1.4 Experimental paradigm

The experimental session lasted 72 minutes and 40 seconds in total. It was divided into 6 sub-sessions (500 s each), separated by a break of 110 s, plus one last sub-session (250 s). During the breaks, the participants could take a rest and move a bit (albeit without removing the device or standing up). This allowed several moments where, if the participants wished, they could withdraw from the experiment without disrupting the data collection. At the end of each break, a message was displayed on the screen for 10 s to allow the participants to get ready to restart the experiment.

We define one trial as the experimental procedure whereby the action of interest is recorded. Trials consisted in an initial relaxation phase, a stimulus phase and an action phase. More specifically, at the beginning of each trial the participants cleared their mind for 3 s, then they experienced the stimulus for 2 s, and finally they performed the speaking action for 20 s. The total trial duration was 25 s. In the action stage, participants were instructed to say in their mind (either by performing imagined or inner speech) what they just heard/saw/read. For example, if they just saw the image of a clock, they would say 'clock' in their mind. For the rest class, the word 'rest' appeared on the screen in the stimulus phase. Then, in the action phase, participants were instructed to watch the screen without performing any action. This structure was mainly inspired by Zhao and Rudzicz (2015) and Lee et al. (2019). 20 trials constituted a sub-session (a part from the last sub-session that contained only 10).

Each class was recorded for 10 trials, so that 130 trials were recorded in total. The order of trials was hardcoded to be random and was the same for all participants. This minimized bias in the action stage as there is no pattern in the succession of trial classes. Each speaking repetition corresponded to one epoch in the EEG raw signal (OpenVibe was ideal in this regard as it automatically executed the subdivision of data into epochs).

To fulfil **O3** and **O4** and maximise the data available for learning, we implemented repeated imagination. Participants performed the speaking action 10 times per trial. Thus, across the whole experiment 100 datapoints were recorded for each class (10 trials per 10 repetitions). This is done extensively in the literature (Brigham and Kumar, 2010; Deng et al., 2010; Koizumi, Ueda and Nakao, 2018). In particular Nguyen, Karavas and Artemiadis (2017), who performed classification tasks similar to ours, also recorded exactly 100 trials per class. Additionally, on top of being a more time effective way to collect data, repeated imagination might improve decoding accuracy. Quoting Panachakel and Ramakrishnan (2021), *'EEG signatures become more prominent across multiple imaginations in the same trial but deteriorate across multiple trials in the same recording session'*. This effectively motivates our investigation into repeated imagination set by **H1**.

To ensure that always the same number of speaking actions is performed in the action stage, we set a rythm by using a visual cue. This avoids having an unknown number of speaking repetitions in each measurement window, which would constitute unlabelled data with an unknown number of instances. While auditory cues such as clicks are more common in the literature, they elicit responses in both Broca's and Wernicke's areas, which are involved in general speech processing and consequently in both imagined and inner speech (Panachakel and Ramakrishnan, 2021; Hesslow, 2002; Poeppel and Idsardi, n.d.). This makes the removal of the auditory trigger from the data complicated. Using a visual cue circumvents this problem. Since visual stimuli are processed in the occipital lobe, which is not involved in neither production nor comprehension of speech, the data can be cleaned by simply discarding the occipital lobe EEG channel (Nguyen, Karavas and Artemiadis, 2017; Panachakel and Ramakrishnan, 2021).

4.2 Data visualization

4.2.1 Raw data

Raw data was automatically saved into csv files, one collective file for each stage of trials (e.g., all stimuli data are collected in a file, all data for the first speaking repetition of the action stage in another file, etc.). The data had 21 columns (16 EEG channels, epoch, time, word prompt, speaking stage and modality of stimulus), and a number of rows equal to the number of epochs recorded multiplied by the number of rows per epoch. Given a sampling rate of 256 Hz and an epoch duration of 2 seconds, each epoch was comprised of 512 rows. Hence, the total amount of epochs was 1300 (130 trials x 10 epochs per trial). Less epochs were recorded for those who did not complete the experimental session.

4.2.2 Signal corruption

A large portion of the data collected was unfortunately corrupted. Specifically, up to about a third of the values recorded are frozen, as can be seen in Fig. 4.2 in the next page. The exact reasons why this happened are not clear. A possible explanation is that the processing speed of the laptop used for data collection exceeded the data transfer speed of either the Emotiv EPOC+ device or the OpenVibe acquisition server. This would have caused the data collection tool to freeze and keep transmitting the last measured value until it caught up with the laptop. Unfortunately, the aperiodic nature of EEG signals makes it impossible to recover the frozen data.

The fact that a large portion of our data is corrupted and not recoverable is far from ideal, and completely prevents the attainment of **O3**. However, the ML might be able to circumvent this issue and still deliver good results.

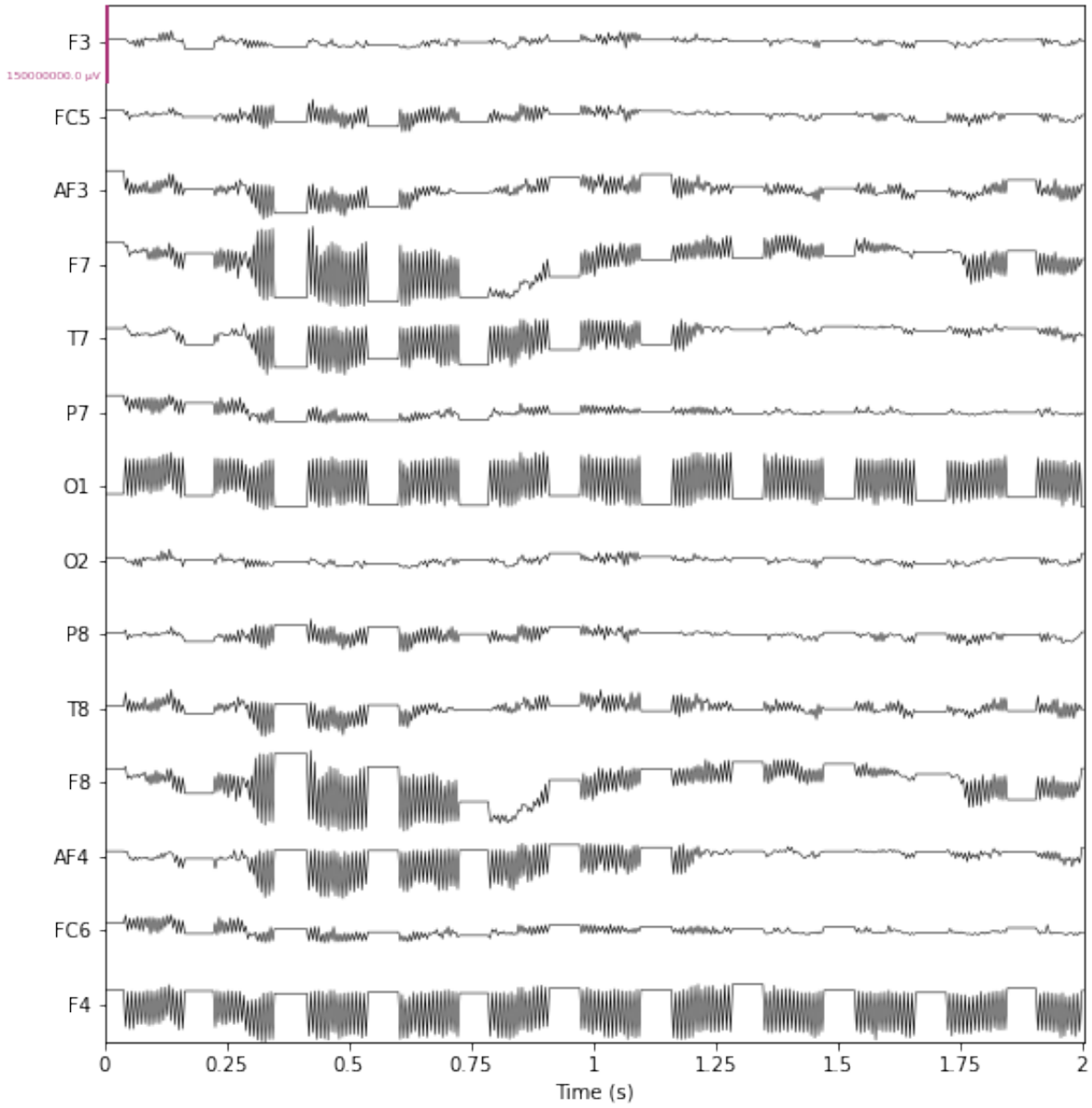


Figure 4.2: The graph shows the signal corresponding to one epoch of P2 speaking data (i.e., one single word repetition) for all the 14 EEG channels. The x-axis is time [s] while the y-axis is voltage [μV]. The data is frozen for random subsets of the signals across all channels simultaneously. Own figure.

4.3 Data preprocessing

4.3.1 Class equalisation

Either due to the removal of epochs with extreme outliers, or due to participants not completing the experimental session, datasets can be imbalanced with respect to data class (e.g., before

preprocessing, P5 contained 80 instances of 'audio-hospital' and only 40 instances of 'text-clock').

Clearly, this can substantially skew the classification accuracy towards the most represented classes in the dataset. To solve this problem, we randomly remove datapoints from each class until all have equal amounts. Albeit necessary, this is a drastic measure which can considerably decrease the data available (e.g., half of the 'audio-hospital' data in P5 is lost).

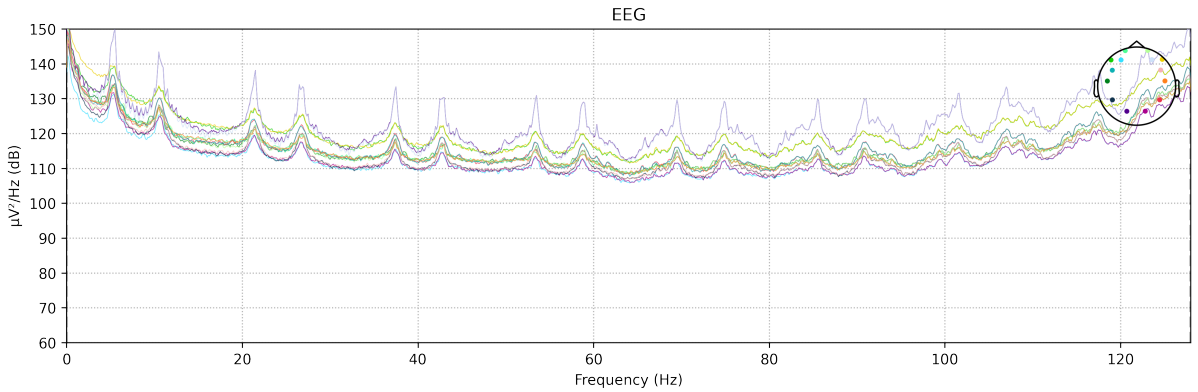
4.3.2 Resampling & filtering

The first step in EEG data preprocessing procedures is usually resampling. This consists in decreasing the sampling rate and hence the computational complexity associated to handling the data. For example, Brigham and Kumar (2010); Nguyen, Karavas and Artemiadis (2017) downsample their sampling rate from 1 KHz to 256 Hz. Since our sampling rate is 256 Hz already, we can skip this step.

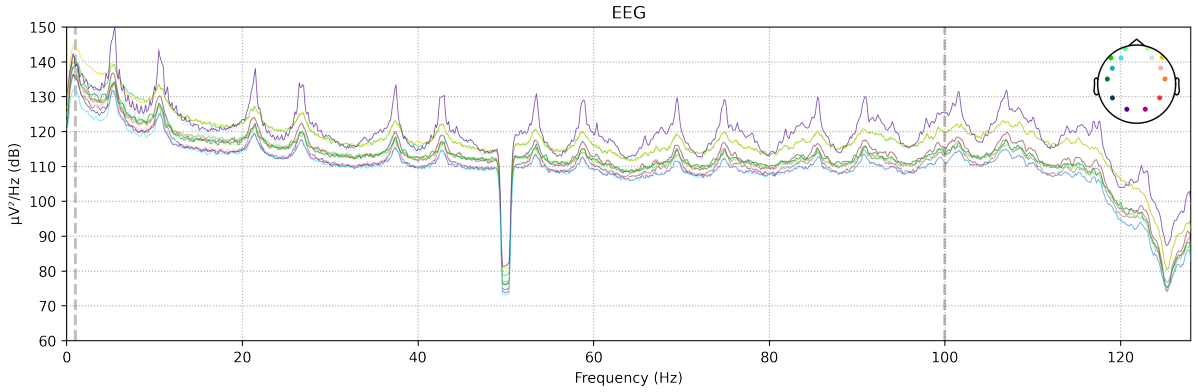
Temporal filtering involves removing data that falls outside a particular frequency band in order to maximize the SNR and remove artifacts. There is no consensus in the literature regarding which frequency band is optimal for EEG imagined speech decoding, and different researchers adopt many different approaches. Most commonly the 8-20 Hz band is used (Panachakel and Ramakrishnan, 2021). Among the works that we are expanding upon, Lee et al. (2019) and Nguyen, Karavas and Artemiadis (2017) selected the 0.5-40 Hz and 8-70 Hz bands respectively. Nieto et al. (2022), which is the only existing open-source dataset on inner speech, utilizes the 0.5-100 Hz band. This is done to maintain the data as raw as possible and allow other researchers to select a frequency band of their choice. Removing artifacts with the Python package MNE, which will be covered in Section A.1, requires a high pass filter at 1 Hz, which is hence our lower limit. Since many of our hypotheses are the first of their kind, we need to consider the eventuality that utilizing a specific frequency band may affect imagined and inner speech data differently, thus introducing undesired variability. Furthermore, despite data corruption, we still aim to develop a methodology that is suitable for the creation of an open-source dataset. As a consequence, we follow the approach by Nieto et al. (2022) and conservatively select the 1-100 Hz frequency band. Additionally, we apply a notch filter of width 1 Hz around 50 Hz (i.e., the band 49-51 Hz is filtered out) to remove the UK power line

frequency. Please find below power spectra graphs for the unfiltered and filtered data in Fig. 4.3 a) and b) respectively.

Finally, we did not consider spatial filtering since most works in the literature do not use it (Panachakel and Ramakrishnan, 2021).



(a)



(b)

Figure 4.3: These graphs show the Raw temporally unfiltered data. The x-axis is frequency (Hz) and the y-axis is amplitude (dB). Each coloured electrode on the top right corresponds to a coloured signal in the graph. Own figure.

4.3.3 Channel selection

To improve the SNR, some studies only keep data from a selected subset of EEG channels. Two families of approaches to channel selection exist: 1) selecting the channels that measure the brain activity of the regions involved in the task at study; 2) selecting the subset of channels that delivers best ML performance. For example, García, García and Pineda (2012) manually choose the channels that cover the cortical region, which is particularly significant for speech

processing (Marslen-Wilson and Tyler, 2007; Alderson-Day and Fernyhough, 2015). On the other hand, Nguyen, Karavas and Artemiadis (2017) and Lee et al. (2019) use common spatial patterns (CSP) to select those channels which optimize classification results. Of the channels selected by either one of the approaches above, our device only has 4 ('F7', 'F3', 'FC5' and 'T7'). We attempted to use these channels yet found that classification accuracy was severely impacted, most likely due to the data being too scarce for the ML to train properly. Notably, this is coherent with the findings of Panachakel, Ramakrishnan and Ananthapadmanabha (2020), who reported that less than 9 channels was detrimental for decoding performance. We operate no channel selection.

4.3.4 Artifact removal: Epoch rejection & ICA

While EEG allows us to measure the brain activity associated to inner and imagined speech, it also picks up unrelated electrophysiological signals. In other words, noise. Typical sources include EOG, EMG and ECG (electrocardiography) activity, the most common being eye blinks. Myologic electrical activity (i.e., electrical activity that involves muscular tissue) is much stronger in magnitude ($\sim$ mV) then the EEG signal of interest ($\sim \mu$V), and hence needs to be removed. Integrating different approaches from the literature, we combine heuristics with an algorithmic method and we proceed as follows.

First, we discard epochs which maximum value exceeds an upper threshold M. We estimate M by extracting the maximum value for all epochs and computing the mean μ_{max} and standard deviation σ_{max} of the collection of maxima. Through some trial and error, we identify a suitable upper bound as

$$M = \mu_{max} \pm 7\sigma_{max}, \tag{4.1}$$

which cuts $3-6\%$ data for most participants. This is to be expected since the data is incredibly noisy. We recognise that (4.1) is a very rough estimation, even more so given that the maxima of the epochs are not normally distributed. However, it works well to automatically generate an appropriate upper bound and remove epochs that are corrupted by artifactual outliers. This operation can cause class imbalances, which we address by applying the method described

above in Section 4.3.1.

Independent component analysis (ICA) is the most common algorithm employed for EEG artifact removal. It works by decomposing the signals into statistically independent sources. Of all the sources, some will originate from processes that are of no interest (e.g., eye blinks), which are then removed. To perform ICA, we utilize MNE built-in `mne.preprocessing.ICA` object. The decomposition delivers 14 distinct components. Following the instructions on the ICLabel website, we identify and exclude exclude artifacts by visually inspecting the plots and power spectra of the sources. You can observe a plot of ICA components before and after artifact exclusion in Fig. 4.4 below. We find eye blinks and/or cardiac activity in most participants, as displayed in Fig. A.1. For some participants the algorithm is not able to discern any artifacts, see Appendix A. Please also find a list of the removed ICA components per participant, a plot of all the signal sources and a diagram of artifacts with relative power spectra, respectively in Appendix A.

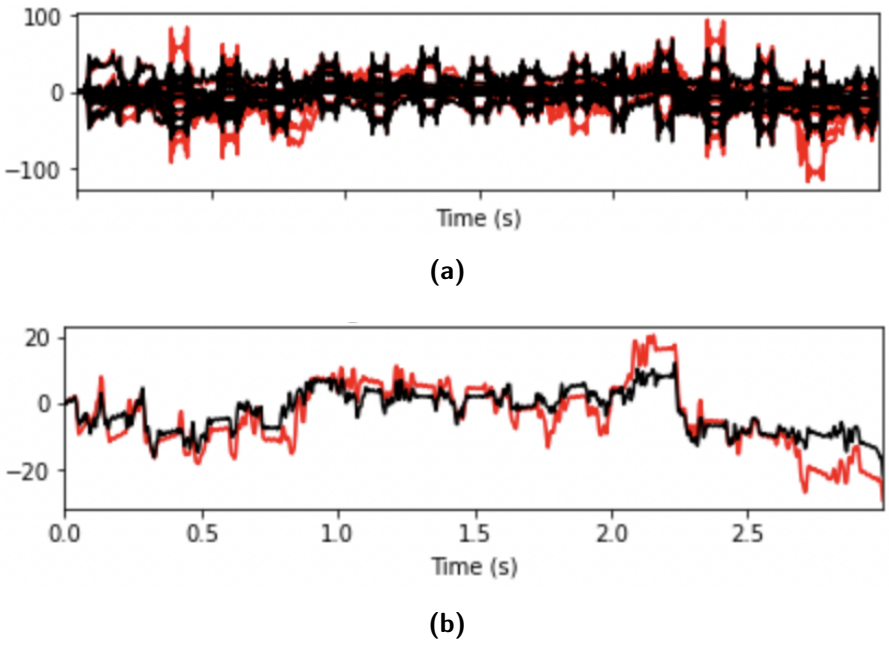
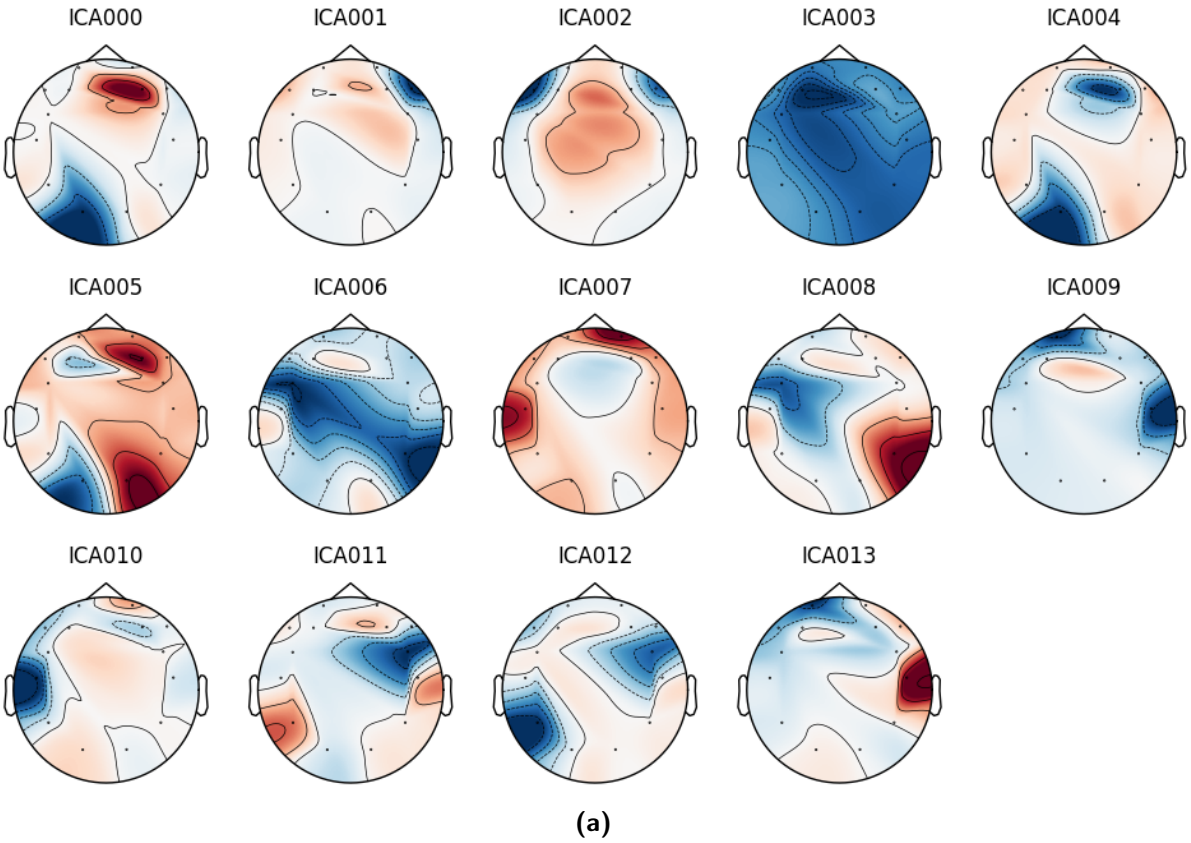


Figure 4.4: In both graphs the y-axis is voltage. The y-axis in a) is in μV, while in b) it is in 10^{-5} V. a) displays all the ICA sources including (red) and excluding (black) artifacts. b) displays the average across all ICA sources before (red) and after (black) having removed artifacts. P2 data. Own figure.

We ensure full reproducibility of results by fixing the `mne.preprocessing.ICA` `random_state` parameter. If this is not set, the algorithm can perform different decompositions for different

runs. Finally, for the sake of rigor, we highlight that all of the preprocessing steps before ICA do not involve acting upon the data in such a way that might cause data leakage.



(a)

Figure 4.5: The graph shows the ICA decomposition for P1. We identified sources ICA001 and ICA002 as eye blinks, and ICA003 as ECG activity. Own figure.

ICA, on the other hand, does incur this issue, as the decomposition into signal sources involves analyzing and operating on the dataset as a whole. This formally constitutes data leakage in relation to 1) train-test ML split for classification; 2) **H2**, **H3** and **H4** since they all involve investigating the effects of distinct classes of data. In practice, however, this is done quite ubiquitously in the literature (Panachakel and Ramakrishnan, 2021) and does not compromise the validity of results.

4.4 Event-related potentials: a sanity check

At this point, we perform a premilinary test of our hypotheses by plotting the event-related potentials (ERP). In simple terms, an ERP is the electrophysiological response of the brain to a particular stimulus or event (Nidal and Malik, 2014). By examining the ERP graphs of the speaking data for the different classes that the ML will work with, we can do a rough sanity check on the suitability of the hypotheses. For example, let us assume that the ERP averaged across all EEG channels for short words and for long words look the same. This would likely mean that word complexity does not play a significant role, and the ML would most likely not perform well. You can observe the relevant ERP plots in Fig. 4.6 below.

In all cases, ERP look distinct. In particular, the ERP of long words and short words are very different. The ERP for long words peaks later than for short words, which is the behaviour one would expect. From this visual inspection, it seems that the data for all the categories of interest is dissimilar enough for the ML to work well.

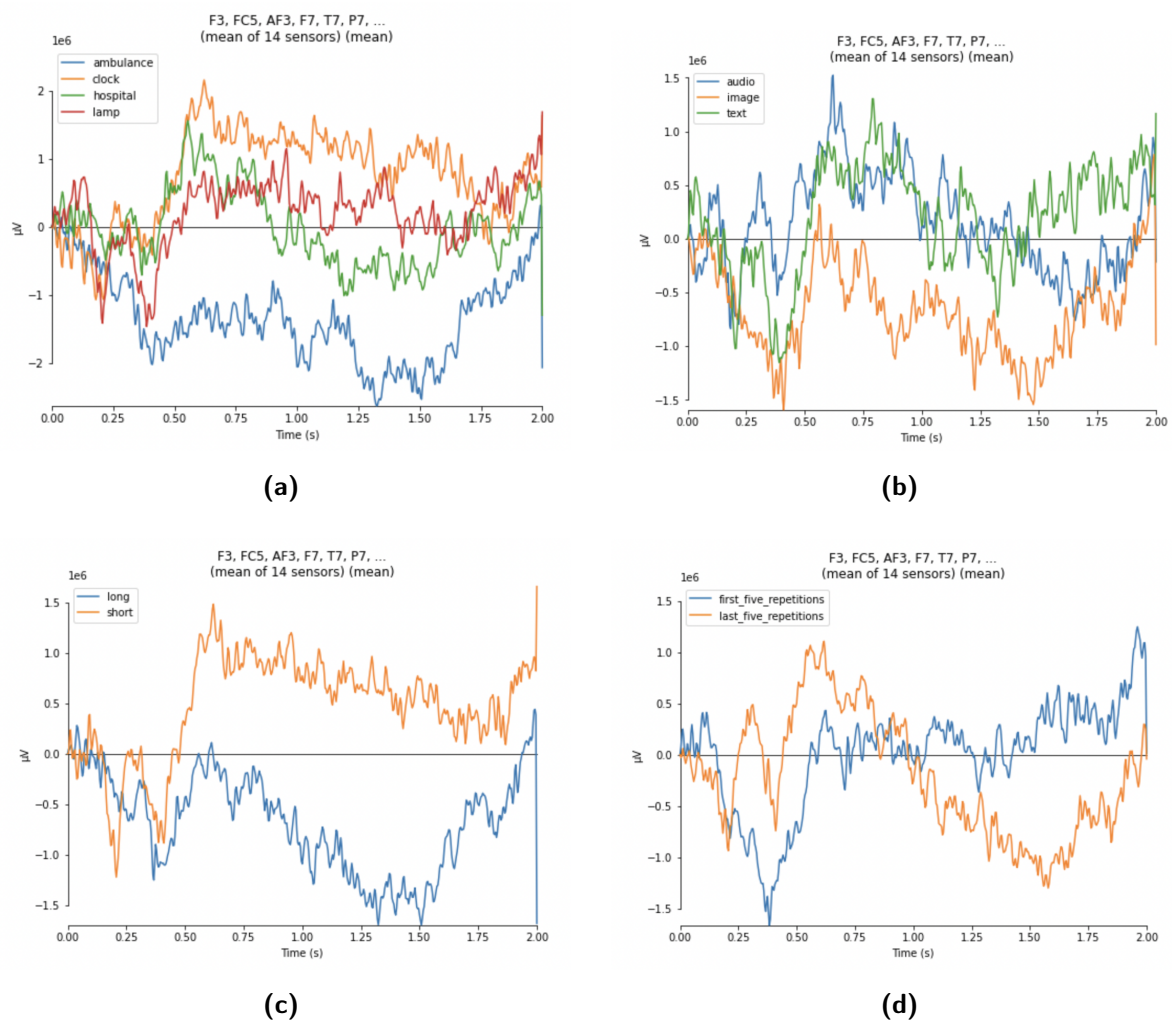


Figure 4.6: ERP plots for P2. The y-axis is voltage (μV). The signals are the average of all 14 EEG channels. Clockwise starting from a), they show the ERP of imagined speech for the 4 words (a), the different modalities of stimulus (b), complexity (c) and repeated imagination (d). Own figure.

4.5 Machine learning

We ensure reproducibility of results by fixing the random state where relevant. In order to estimate performance in the most accurate way possible, we use the Scikit-learn class `RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=32)`. We have the same folds and the same train-test split for all 3 classifiers.

4.5.1 Feature selection & hyperparameter tuning

Initially, we were planning to perform feature selection and hyperparameter optimization for each one of the classification tasks. We programmed this to happen automatically for each classifier for a selected set of encoders, scalers, feature selectors and hyperparameters. To do so, we combined Scikit-learn classes `Pipeline` and `GridSearchCV` to automatically sweep through all the parameters and select the optimal ones. However, we decided in the end to not perform neither feature selection nor hyperparameter tuning. If we did, a different optimal pipeline would have been selected for each ML classifier for each classification task (for each participant). This would have introduced great variability between classifiers, between imagined and inner speech, between different modalities of stimulus, and between participants. As a result, performance comparisons across the categories would have lost validity.

4.5.2 Classifiers

We chose to use RF, SVM and KNN, as they are all quite ubiquitous in literature. For example, excluding the papers that we have talked about extensively in Sections 2.2 and 2.3, they are used in García, García and Pineda (2012); Tøttrup et al. (2019); Moctezuma and Molinas (2018); Qureshi et al. (2017); Hashim, Ali and Mohd-Isa (2017). For a comprehensive review of the methods used for imagined speech decoding please consult Panachakel and Ramakrishnan (2021). On the other hand, to our knowledge no study exists that implements RF, SVM or KNN for inner speech.

Albeit neural networks are one of the most popular and effective approaches to EEG speech decoding, we decided to not implement them as two works already exist on 4-class classification in the context of inner speech (Jonsson, 2022; van den Berg, van Donkelaar and Alimardani, 2021). Furthermore, while we have not carried out these tests on more than two or three participants, in the initial stages of our analysis we have tried a variety of other algorithms commonly used in the literature, including linear discriminant analysis (LDA), naive Bayes classifiers (NB) and shrinkage RLDA. We have decided to discard them as their performance was lower than our three algorithms of choice.

For each participant, the mean accuracy μ_p and standard deviation μ_p of a classification test

are computed by Scikit-learn class `cross_val_score`. For details on how the mean and standard deviation are computed when combining data of different participants please consult Appendix B on uncertainty propagation.

4.6 Hypothesis evaluation

To test our hypotheses, we carry out the following classification tasks. Note that 4-class classification (4CC) always refers to the classification of our 4 words of choice. All of the tests below test **H1**.

- **T1.** 4CC irrespective of modality of stimulus, for both inner and imagined speech. This tests H2.
- **T2.** 4CC by modality of stimulus, for both inner and imagined speech. This involves splitting the dataset into 3 subsets, one for each stimulus modality, and performing 4CC on each separately. This tests H4.
- **T3.** Binary classification by complexity, irrespective of modality of stimulus, for both inner and imagined speech. This involves splitting the dataset into 2 subsets, long words and short words, and performing binary classification on each separately. This tests the part of H3 pertaining to complexity.
- **T3.** 4CC by repeated imagination, for both inner and imagined speech. This involves splitting the dataset into 3 subsets, one for the first, one for the middle and one for the last 5 repetitions in each trial. This tests the part of H3 pertaining to repeated imagination.

Chapter 5

Results

We report here our results, which will be discussed in the next section. Please find below a table for each investigation.

5.1 4-class classification

Table 5.1: 4CC, irrespective of the modality of stimulus. The results are reported for RF, SVM and KNN, for both imagined and inner speech. The accuracy is expressed as Mean ± Std.

(a) Imagined Speech					
	P1	P2	P6	P7	Mean
RF	24.2 ± 4.8	30.6 ± 4.7	24.7 ± 5.2	27.4 ± 4.5	**26.7** $\pm$ 4.8
SVM	25.2 ± 4.6	29.5 ± 2.9	24.8 ± 4.8	25.9 ± 5.0	**26.4** $\pm$ 4.4
KNN	24.3 ± 4.9	27.9 ± 3.6	24.2 ± 4.6	24.1 ± 3.5	**25.1** $\pm$ 4.2

(b) Inner Speech						
	P3	P4	P5	P8	P9	Average
RF	25.3 ± 7.0	26.6 ± 4.1	27.2 ± 5.1	27.1 ± 4.5	26.5 ± 4.4	**26.5** $\pm$ 5.1
SVM	26.6 ± 3.3	27.4 ± 4.8	29.3 ± 6.2	29.7 ± 5.5	27.0 ± 4.3	**28** $\pm$ 4.9
KNN	24.5 ± 3.1	25.8 ± 3.2	24.5 ± 4.5	26.7 ± 5.2	26.5 ± 4.9	**25.6** $\pm$ 4.3

5.2 4-class classification by stimulus modality

Table 5.2: 4CC by stimulus modality. The results are reported for RF, SVM and KNN, for both imagined and inner speech. The accuracy is expressed as Mean ± Std.

(a) Imagined Speech					
	P1	P2	P6	P7	Mean
(i) Audio					
RF	27.9 ± 6.7	41.2 ± 6.9	25.5 ± 6.5	31.6 ± 9.3	**31.6** ± 7.4
SVM	30.1 ± 8.1	31.8 ± 5.6	30.1 ± 7.7	33.4 ± 7.8	**31.4** ± 7.3
KNN	28.2 ± 5.3	33.8 ± 7.0	27.8 ± 6.0	28.3 ± 6.5	**29.5** ± 6.2
(ii) Image					
RF	24.1 ± 9.1	40.8 ± 7.9	25.2 ± 8.6	24.4 ± 6.7	**28.6** ± 8.1
SVM	28.7 ± 7.2	39.3 ± 7.3	28.9 ± 7.2	26.7 ± 8.3	**30.9** ± 7.5
KNN	27.9 ± 6.7	33.3 ± 6.2	26.8 ± 8.8	25.3 ± 7.0	**28.3** ± 7.2
(iii) Text					
RF	23.4 ± 6.8	33.0 ± 7.2	25.9 ± 8.2	24.4 ± 6.9	**26.7** ± 7.3
SVM	29.3 ± 7.4	27.9 ± 7.7	29.0 ± 7.4	26.7 ± 9.0	**28.2** ± 7.9
KNN	24.3 ± 8.1	26.2 ± 8.1	25.5 ± 7.3	26.2 ± 6.0	**25.6** ± 7.4

(b) Inner Speech						
	P3	P4	P5	P8	P9	Mean
(i) Audio						
RF	26.1 ± 7.2	32.4 ± 8.4	31.7 ± 9.0	29.7 ± 10.8	24.7 ± 8.4	**28.9** ± 8.8
SVM	28.8 ± 7.2	30.4 ± 6.6	32.1 ± 9.5	28.3 ± 7.7	31.0 ± 7.2	**30.1** ± 7.7
KNN	25.6 ± 4.5	27.4 ± 5.6	30.4 ± 8.5	27.5 ± 11.4	25.6 ± 5.6	**27.3** ± 7.5
(ii) Image						
RF	27.9 ± 7.0	24.9 ± 6.5	26.3 ± 9.0	24.8 ± 8.9	29.1 ± 7.3	**26.6** ± 7.8
SVM	28.7 ± 7.6	25.6 ± 5.1	19.2 ± 10.2	22.8 ± 6.8	30.4 ± 5.8	**25.3** ± 7.3
KNN	23.9 ± 6.4	27.1 ± 6.1	27.3 ± 10.5	28.3 ± 8.7	29.6 ± 9.3	**27.2** ± 8.4
(iii) Text						
RF	26.3 ± 6.5	23.1 ± 7.7	23.3 ± 9.9	19.7 ± 8.4	27.5 ± 9.8	**24** ± 8.6
SVM	23.6 ± 6.1	25.6 ± 6.1	21.9 ± 9.4	32.1 ± 8.6	30.9 ± 7.6	**26.8** ± 7.7
KNN	24.2 ± 8.2	24.6 ± 7.9	27.5 ± 10.2	30.0 ± 10.5	22.5 ± 8.5	**25.8** ± 9.1

5.3 Binary classification by word complexity

Table 5.3: Accuracy for the binary classification for long and short words (i.e., 'ambulance' vs 'hospital' and 'clock' vs 'lamp'). The results are reported for RF, SVM and KNN for both imagined and inner speech. The accuracy is expressed as Mean ± Std.

(a) Imagined Speech					
	P1	P2	P6	P7	Mean
RF	54.4 ± 8.1	61.2 ± 8.9	44.1 ± 7.7	56.2 ± 8.3	**54** $\pm$ 8.3
SVM	54.0 ± 6.8	59.8 ± 8.8	43.6 ± 7.3	49.1 ± 8.2	**51.6** $\pm$ 7.8
KNN	51.3 ± 9.1	55.1 ± 7.8	52.3 ± 8.2	51.1 ± 6.4	**52.4** $\pm$ 7.9
(ii) Short					
	P1	P2	P6	P7	Mean
RF	50.4 ± 6.2	58.7 ± 5.5	51.8 ± 7.4	56.1 ± 7.1	**54.2** $\pm$ 6.6
SVM	50.1 ± 7.4	60.2 ± 7.6	49.8 ± 7.7	48.6 ± 7.5	**52.2** $\pm$ 7.6
KNN	53.1 ± 8.1	51.1 ± 5.7	51.3 ± 8.3	50.1 ± 6.8	**51.4** $\pm$ 7.3
(b) Inner Speech					
	P3	P4	P5	P8	P9	Mean
RF	52.4 ± 5.8	49.6 ± 6.3	54.6 ± 10.3	57.1 ± 7.8	52.9 ± 7.7	**53.3** $\pm$ 7.7
SVM	49.1 ± 5.9	49.4 ± 6.3	48.7 ± 9.1	55.7 ± 6.2	52.1 ± 6.1	**51** $\pm$ 6.8
KNN	50.2 ± 5.1	47.3 ± 7.8	51.1 ± 11.1	55.7 ± 10.8	50.8 ± 5.6	**51** $\pm$ 8.5
(ii) Short					
	P3	P4	P5	P8	P9	Mean
RF	49.7 ± 6.8	55.1 ± 6.1	51.1 ± 12.2	47.8 ± 9.7	53.5 ± 8.7	**51.4** $\pm$ 9
SVM	50.4 ± 5.9	51.5 ± 6.2	51.2 ± 8.5	53.5 ± 7.6	51.4 ± 7.4	**51.6** $\pm$ 7.2
KNN	50.6 ± 6.4	53.1 ± 5.1	46.7 ± 10.7	51.5 ± 8.2	51.2 ± 6.8	**50.6** $\pm$ 7.7

5.4 4-class classification by repeated imagination

Table 5.4: 4-class classification of the words: Ambulance, Clock, Hospital and Lamp, performed on data from the 1-5, 3-7 or 6-10 trial repetition intervals. The results are reported for RF, SVM and KNN, for both imagined and inner speech. The accuracy is expressed as Mean ± Std.

(a) Imagined Speech					
	P1	P2	P6	P7	Mean
(i) 1-5					
RF	24.7 ± 6.5	28.4 ± 4.8	27.4 ± 5.7	25.1 ± 6.5	**26.4** ± 7.4
SVM	28.9 ± 6.5	28.7 ± 6.1	28.7 ± 6.2	26.2 ± 4.8	**28.1** ± 7.3
KNN	24.6 ± 5.5	27.7 ± 4.9	24.6 ± 5.7	26.5 ± 5.5	**25.8** ± 6.2
(ii) 3-7					
RF	23.3 ± 8.1	28.9 ± 5.2	25.2 ± 5.4	29.0 ± 6.3	**26.6** ± 6.4
SVM	23.4 ± 5.8	27.8 ± 3.9	23.0 ± 6.1	21.3 ± 5.3	**23.9** ± 5.3
KNN	24.9 ± 6.5	29.5 ± 4.8	25.4 ± 6.3	24.6 ± 5.7	**26.1** ± 5.9
(iii) 6-10					
RF	26.3 ± 7.2	29.9 ± 5.7	29.7 ± 5.2	28.9 ± 5.6	**28.7** ± 6
SVM	25.6 ± 6.3	30.7 ± 5.2	25.1 ± 6.5	25.8 ± 6.5	**26.8** ± 6.1
KNN	27.3 ± 7.5	27.1 ± 4.6	25.4 ± 7.4	25.5 ± 5.7	**26.3** ± 6.4

(b) Inner Speech						
	P3	P4	P5	P8	P9	Mean
(i) 1-5						
RF	28.5 ± 6.6	25.2 ± 5.2	29.4 ± 9.6	29.0 ± 6.7	26.6 ± 6.6	**27.7** ± 7.1
SVM	28.5 ± 5.7	25.0 ± 7.1	28.1 ± 8.9	31.7 ± 6.1	27.1 ± 5.3	**28.1** ± 6.7
KNN	25.6 ± 5.5	25.4 ± 6.5	26.4 ± 9.8	24.7 ± 7.8	28.3 ± 6.6	**26.1** ± 7.4
(ii) 3-7						
RF	26.2 ± 5.9	26.4 ± 5.4	22.9 ± 9.1	25.9 ± 7.2	25.2 ± 6.7	**25.3** ± 7
SVM	30.5 ± 6.9	28.5 ± 5.1	21.1 ± 6.5	26.6 ± 7.6	27.5 ± 6.2	**26.8** ± 6.5
KNN	25.9 ± 4.7	25.2 ± 5.1	22.6 ± 6.1	23.6 ± 7.3	28.4 ± 5.8	**25.1** ± 5.9
(iii) 6-10						
RF	25.2 ± 5.5	24.5 ± 5.3	24.7 ± 8.2	25.5 ± 7.6	22.5 ± 6.1	**24.5** ± 6.6
SVM	24.8 ± 5.0	27.7 ± 4.9	25.8 ± 7.6	22.8 ± 6.4	22.4 ± 5.7	**24.7** ± 6
KNN	25.2 ± 4.2	23.9 ± 6.1	28.5 ± 9.4	24.7 ± 7.7	24.5 ± 5.2	**25.4** ± 6.8

Chapter 6

Discussion

In this section we discuss our findings. Given that a third of our data is unrecoverably corrupted, we expected our results to be chance-level. We find that, while this is true for a substantial portion of the results, a non-negligible fraction lies above the randomness threshold. Despite the frozen values, it seems that in a few instances the ML was able to spot and capitalize on the residual patterns preserved in the signals. This is surprising and motivating, and one cannot help but wonder what the results would have looked like if the data collection went as expected. We will analyse each investigation in turn.

6.1 4-class classification

We plot the results of this classification task in Fig. 6.1 below. The great performance of P2 is immediately evident, a finding that recurs across all tests (see tables in the results section). Comparing the ML models for both speech paradigms, SVM performs best overall, obtaining up to 28% accuracy for inner speech. Notably, KNN is the lowest performer for all participants. For reference, Nguyen, Karavas and Artemiadis (2017) achieve 50% for 4CC of short words with SVM, while the best performance that we could find in the literature was 73.4% with artificial neural networks (ANN) (Balaji et al., 2017). Although this is not outstanding performance by any means, considering the data corruption and the absence of any form of feature selection and hyperparameter tuning, the fact that most findings are above chance-level is impressive.

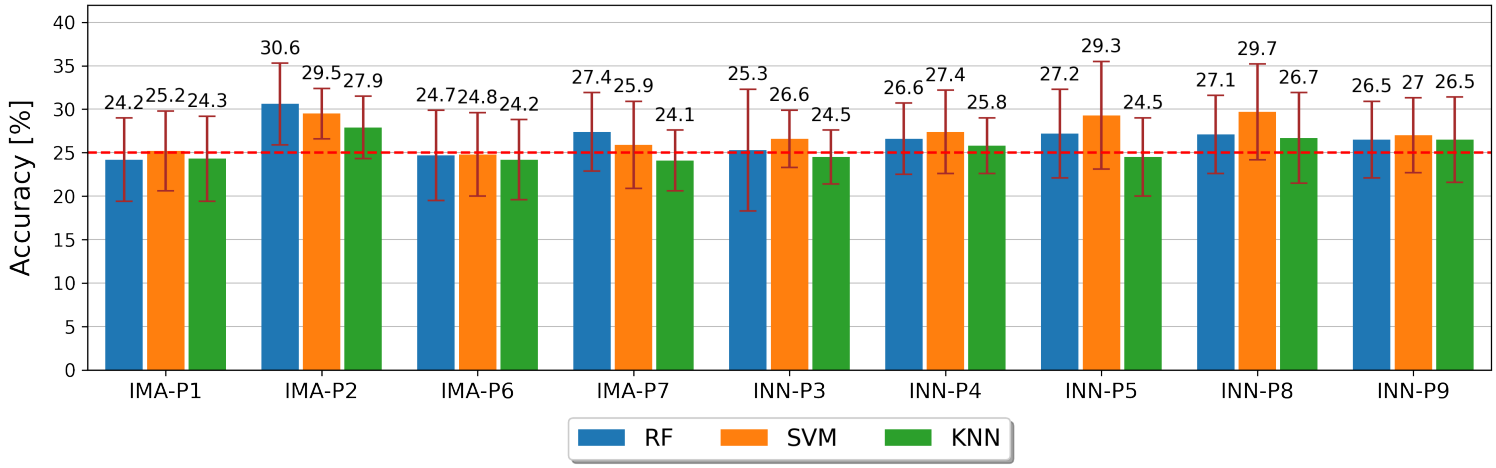


Figure 6.1: Mean 4CC accuracy irrespective of the modality of stimulus. The labels 'IMA' and 'INN' indicate that the participant performed imagined and inner speech, respectively. The results are reported for RF, SVM and KNN. The error bars are standard deviation. Own figure.

Comparing the speech paradigms, RF and KNN deliver similar results in both, while SVM performs substantially better (+1.6% accuracy with comparable Std) for inner speech. Averaging across all ML models, the accuracy is 26% for imagined speech and 26.7 for inner speech. Given that the outperformance of inner speech is slight and it occurs for one ML model only, it could very well be statistical noise. This means that we cannot draw any definite conclusion as to which paradigm is best, and **H2** can not be validated nor rejected.

6.2 Stimulus modality

You can observe the results of this classification task in Fig. 6.2 below. Similarly to the previous section, SVM performs best overall (up to 31.4%), followed by RF and KNN. Remarkably, in one instance RF beats SVM achieving 31.6%. KNN is the lowest performer for all instances but visually-elicited inner speech. Albeit no similar investigation has ever been conducted in the literature, this is still 4-class classification, and we can hold the same comparison metrics as in the previous section for imagined speech(50% by Nguyen, Karavas and Artemiadis (2017), 73% (Balaji et al., 2017)). For inner speech, however, the picture is brighter, as the only available studies we could find achieve $\sim$ 29% with CNN (van den Berg, van Donkelaar and Alimardani, 2021; Jonsson, 2022).

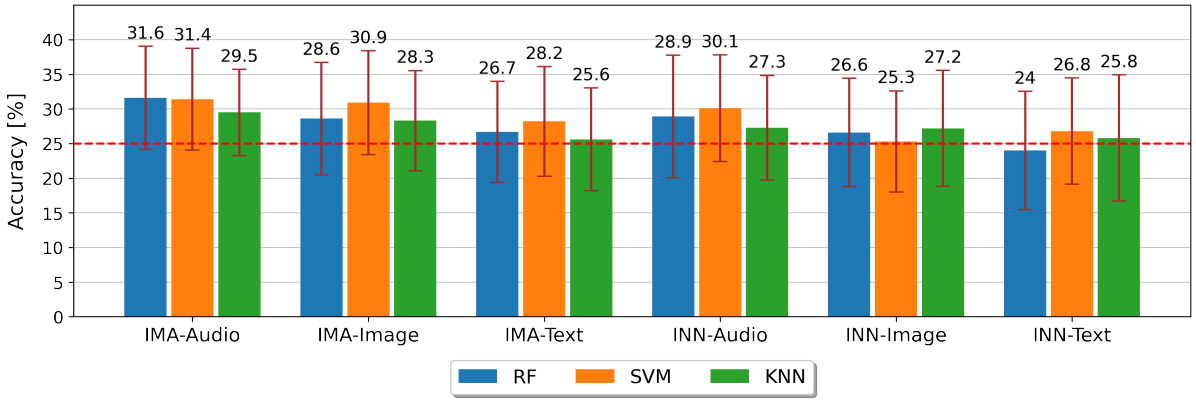


Figure 6.2: Mean 4CC accuracy for different modalities of stimulus. The values displayed are the average across all participants for each speech paradigm. The labels 'IMA' and 'INN' indicate imagined and inner speech, respectively. The results are reported for RF, SVM and KNN. The error bars are standard deviation. Own figure.

Imagined speech outperforms inner speech in all modalities, achieving an inter-modality accuracy of 29% vs 26.8%. This is interesting because the present classification task is effectively a zoom-in into the general 4CC of the previous section. The zoom-in aims to assess the influence of different stimulus modalities. Previously, we found that inner speech just slightly outperformed imagined speech. Here, we find a more significant opposite result (given its magnitude and pervasiveness across modalities). In terms of modalities, audio performs best, followed by images and text, for both imagined and inner speech. This matches our expectations and the findings of the ERP plot in Fig. 4.6, where aurally and visually elicited speaking had ERP that were larger in magnitude. This indicates that paradigms focusing on auditory imagery, which are currently not prevalent in literature, are worthy of exploration.

Although data corruption casts a shadow over the reliability of our conclusions, we believe that the results in this section are worth of consideration. They clearly show that modality of stimulus has a significant impact on the accuracy with which both imagined and inner speech can be decoded. This highlights the importance of focusing on multimodality in EEG-based BCI research.

6.3 Complexity

We plot results in Fig. 6.3 below.

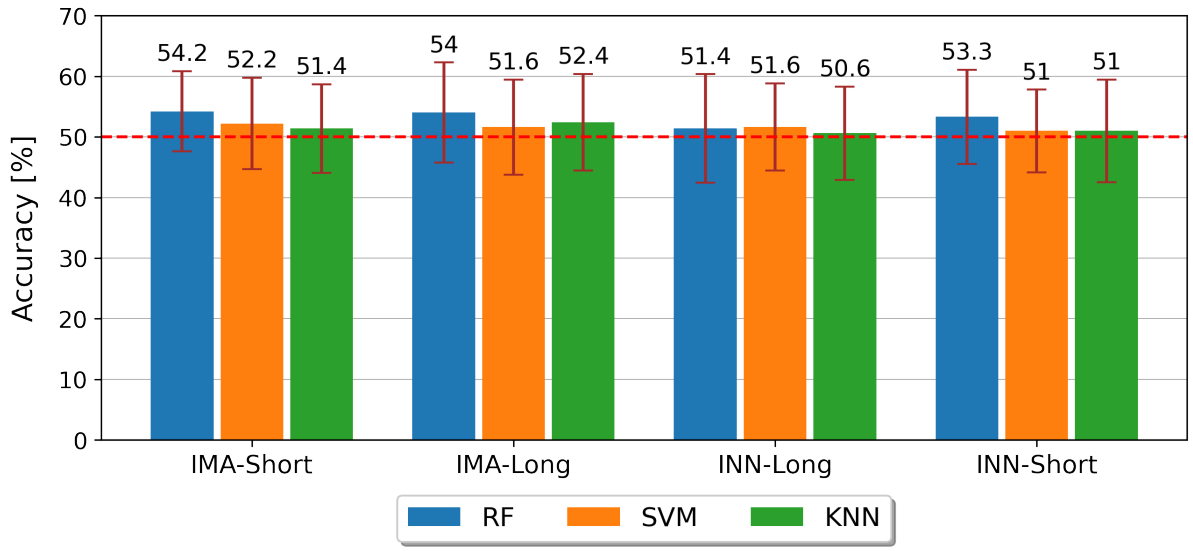


Figure 6.3: Mean accuracy for the binary classification of long and short words (i.e., 'ambulance' vs 'hospital' and 'clock' vs 'lamp'). The values displayed are the average across all participants for each speech paradigm. The labels 'IMA' and 'INN' indicate imagined and inner speech, respectively. The results are reported for RF, SVM and KNN. The error bars are standard deviation. Own figure.

Most results here are clearly chance-level, and no category of data seems to display statistically significant better performance. The only exception is RF, which in three cases out of four outperforms by a margin of $\sim 2\%$. For reference, Nguyen, Karavas and Artemiadis (2017) achieve 66% for the binary classification of tetrasyllabic long words. While the impact of complexity on inner speech has never been investigated, we know from the literature that this is the case for imagined speech. The absence of this effect in our results might stem from the multimodality of the paradigm. If stimulus modality has a strong enough correlation to complexity, whether it being positive or negative, patterns in the data might be washed out. Alternatively, it might be that the impact of complexity is more particularly sensitive to data corruption and was destroyed.

6.4 Repeated imagination

We plot results in Fig. 6.4 below.

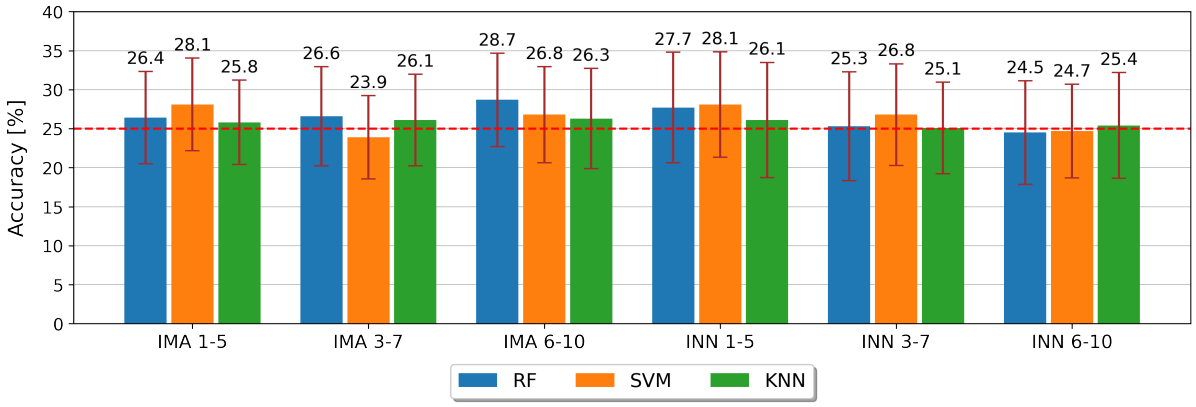


Figure 6.4: 4CC mean accuracy for different repetition intervals during trials. '$1-5$', '$3-7$ and '$6-10$' indicate the first five, middle five and last five repetitions. The values displayed are the average across all participants for each speech paradigm. The labels 'IMA' and 'INN' indicate imagined and inner speech, respectively. The results are reported for RF, SVM and KNN. The error bars are standard deviation. Own figure.

A number of number of instances lies above chance level and most standard deviations are positively skewed. However, no distinct pattern seems to be emerging. The only exception is the lowest performance of KNN, which is in line with the rest of our findings. Although it seems that repeated imagination does not influence speech decoding accuracy, we can not draw any conclusions from the data for the same reasons as above. We cannot validate nor invalidate both parts of **H3**t.

Chapter 7

Conclusions & Outlook

In this project, we have built on the state of the art to develop an innovative, rigorous and comprehensive experimental paradigm for the study of imagined and inner speech. It has a focus on real-life applicability and is designed to allow the investigation of phonetics, complexity and repeated imagination on speech decoding performance. Furthermore, eliciting the speaking action with different modalities of stimulus, it is the first of its kind and it opens up the way for new research into this area.

Unfortunately, our data was corrupted. This prevents us from being able to supply an open-source dataset and hinders our ability to make definite claims about our findings, including whether commercial-grade devices are suitable for the study of inner speech. However, remarkably, the performance of our SVM on 4CC for inner speech (28%) aligns with the only other existing studies in this area (29%) (van den Berg, van Donkelaar and Alimardani, 2021; Jonsson, 2022). In relation to modality of stimulus, we find that: 1) imagined speech outperforms inner speech when the data is analysed with respect to modality of stimulus; 2) modality of stimulus impacts speech decoding accuracy. In particular, auditory stimuli result in substantially better performance for both imagined and inner speech. This reinforces the importance of developing multimodal BCI, and gives a clear indication as to which direction future research in this area could follow.

The most natural development of this work would involve analysing the data acquisition software to ensure that the data is not corrupted. Alternatively, the current data could be

interpolated, although we doubt that quality would improve drastically even if state-of-the-art interpolation software were to be used. After having produced valid baseline results with default classifiers, the next stage would involve optimizing the ML models by implementing feature selection (e.g., channel cross-covariance (CCV) matrices) and hyperparameter tuning. This would be incredibly worthwhile, as it would allow for a fairer comparison with the literature, where sophisticated models are the norm. Another line of development could look at 1) investigating phonetics and 2) collecting more data per participant to allow a more granular and rigorous analysis into the interdependence of phonetics, complexity and stimulus modality. It would also be interesting to implement other ML approaches, focusing on how to best deal with sparsely-distributed time series data with low SNR. Finally, the rigor of the analysis could be vastly improved by implementing statistical methods to evaluate the significance of results (e.g., hypothesis testing).

Bibliography

Abdulkader, S., Atia, A. and Mostafa, M., 2015. Brain computer interfacing: applications and challenges. egypt. inform. j. 16 (2), 213–230 (2015).

Al-Nuaimi, F.A., Al-Nuaimi, R.J., Al-Dhaheri, S.S., Ouhbi, S. and Belkacem, A.N., 2020. Mind drone chasing using eeg-based brain computer interface. *2020 16th international conference on intelligent environments (ie)*. IEEE, pp.74–79.

Alderson-Day, B. and Fernyhough, C., 2015. Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychological bulletin*, 141(5), p.931.

Aloise, F., Lasorsa, I., Schettini, F., Brouwer, A., Mattila, D., Babiloni, F., Salinari, S., Marciani, M. and Cincotti, F., 2007. Multimodal stimulation for a p300-based bci. *Int. j. bioelectromagn*, 9(3), pp.128–130.

Arvaneh, M., Robertson, I.H. and Ward, T.E., 2019. A p300-based brain-computer interface for improving attention. *Frontiers in human neuroscience*, 12, p.524.

Balaji, A., Haldar, A., Patil, K., Ruthvik, T.S., Valliappan, C., Jartarkar, M. and Baths, V., 2017. Eeg-based classification of bilingual unspoken speech using ann. *2017 39th annual international conference of the ieee engineering in medicine and biology society (embc)*. IEEE, pp.1022–1025.

Berg, B. van den, Donkelaar, S. van and Alimardani, M., 2021. Inner speech classification using eeg signals: A deep learning approach. *2021 ieee 2nd international conference on human-machine systems (ichms)*. IEEE, pp.1–4.

Brigham, K. and Kumar, B.V., 2010. Imagined speech classification with eeg signals for silent

communication: a preliminary investigation into synthetic telepathy. *2010 4th international conference on bioinformatics and biomedical engineering*. IEEE, pp.1–4.

Clayton, J., Wellington, S., Valentini-Botinhao, C. and Watts, O., 2020. Decoding imagined, heard, and spoken speech: Classification and regression of eeg using a 14-channel dry-contact mobile headset. *Interspeech*. pp.4886–4890.

Del Pozo-Banos, M., Alonso, J.B., Ticay-Rivas, J.R. and Travieso, C.M., 2014. Electroencephalogram subject identification: A review. *Expert systems with applications*, 41(15), pp.6537–6554.

Deng, S., Srinivasan, R., Lappas, T. and D'Zmura, M., 2010. Eeg classification of imagined syllable rhythm using hilbert spectrum methods. *Journal of neural engineering*, 7(4), p.046006.

D'Zmura, M., Deng, S., Lappas, T., Thorpe, S. and Srinivasan, R., 2009. Toward eeg sensing of imagined speech. *International conference on human-computer interaction*. Springer, pp.40–48.

Fouad, I.A., Labib, F.E.Z.M., Mabrouk, M.S., Sharawy, A.A. and Sayed, A.Y., 2020. Improving the performance of p300 bci system using different methods. *Network modeling analysis in health informatics and bioinformatics*, 9(1), pp.1–13.

Fujiwara, M., Miyasaka, K. and Sakamoto, Y., 2018. Study on eeg-based bci using difference of length and number of syllable in inner speech. *2018 joint 10th international conference on soft computing and intelligent systems (scis) and 19th international symposium on advanced intelligent systems (isis)*. IEEE, pp.1190–1193.

Gansonre, C., Højlund, A., Leminen, A., Bailey, C. and Shtyrov, Y., 2018. Task-free auditory eeg paradigm for probing multiple levels of speech processing in the brain. *Psychophysiology*, 55(11), p.e13216.

García, A.A.T., García, C.A.R. and Pineda, L.V., 2012. Toward a silent speech interface based on unspoken speech. *Biosignals*. pp.370–373.

Gupta, V., Saini, R., De, K., Abid, N., Rakesh, S., Wellington, S., Wilson, H., Liwicki, M.,

Eriksson, J. et al., 2022. Bimodal pilot study on inner speech decoding reveals the potential of combining eeg and fmri.

Han, C., Xu, G., Xie, J., Chen, C. and Zhang, S., 2018. Highly interactive brain–computer interface based on flicker-free steady-state motion visual evoked potential. *Scientific reports*, 8(1), pp.1–13.

Hashim, N., Ali, A. and Mohd-Isa, W.N., 2017. Word-based classification of imagined speech using eeg. *International conference on computational science and technology*. Springer, pp.195–204.

Hesslow, G., 2002. Conscious thought as simulation of behaviour and perception. *Trends in cognitive sciences*, 6(6), pp.242–247.

Indefrey, P. and Levelt, W.J., 2004. The spatial and temporal signatures of word production components. *Cognition*, 92(1-2), pp.101–144.

Jonsson, L., 2022. Using machine learning to analyse eeg brain signals for inner speech detection.

Kaplan, R.M., 2011. The mind reader: the forgotten life of hans berger, discoverer of the eeg. *Australasian psychiatry*, 19(2), p.168.

Keiper, A., 2013. The age of neuroelectronics. *Nanotechnology, the brain, and the future*. Springer, pp.115–146.

Kevric, J. and Subasi, A., 2017. Comparison of signal decomposition methods in classification of eeg signals for motor-imagery bci system. *Biomedical signal processing and control*, 31, pp.398–406.

Kiroy, V.N., Bakhtin, O., Krivko, E., Lazurenko, D.M., Aslanyan, E., Shaposhnikov, D. and Shcherban, I.V., 2022. Spoken and inner speech-related eeg connectivity in different spatial direction. *Biomedical signal processing and control*, 71, p.103224.

Kirschstein, T. and Köhling, R., 2009. What is the source of the eeg? *Clinical eeg and neuroscience*, 40(3), pp.146–149.

Koizumi, K., Ueda, K. and Nakao, M., 2018. Development of a cognitive brain-machine

interface based on a visual imagery method. *2018 40th annual international conference of the ieee engineering in medicine and biology society (embc)*. IEEE, pp.1062–1065.

Lee, S.H., Lee, M., Jeong, J.H. and Lee, S.W., 2019. Towards an eeg-based intuitive bci communication system using imagined speech and visual imagery. *2019 ieee international conference on systems, man and cybernetics (smc)*. IEEE, pp.4409–4414.

Lu, Z., Li, Q., Gao, N., Yang, J. and Bai, O., 2019. A novel audiovisual p300-speller paradigm based on cross-modal spatial and semantic congruence. *Frontiers in neuroscience*, 13, p.1040.

M. Rashid, N. Sulaiman, A.P.A.M.R.M.M.A.F.A.N.B.S.B.e.a., 2020. Current status challenges and possible solutions of eeg-based brain-computer interface: A comprehensive review. *Frontiers in neurorobotics*, 14, p.15.

Marslen-Wilson, W.D. and Tyler, L.K., 2007. Morphology, language and the brain: the decompositional substrate for language comprehension. *Philosophical transactions of the royal society b: Biological sciences*, 362(1481), pp.823–836.

Martin, S., Iturrate, I., Millán, J.d.R., Knight, R.T. and Pasley, B.N., 2018. Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis. *Frontiers in neuroscience*, 12, p.422.

Moctezuma, L.A. and Molinas, M., 2018. Eeg-based subjects identification based on biometrics of imagined speech using emd. *International conference on brain informatics*. Springer, pp.458–467.

Nguyen, C.H., Karavas, G.K. and Artemiadis, P., 2017. Inferring imagined speech using eeg signals: a new approach using riemannian manifold features. *Journal of neural engineering*, 15(1), p.016002.

Nidal, K. and Malik, A.S., 2014. *Eeg/erp analysis: methods and applications*. Crc Press.

Nieto, N., Peterson, V., Rufiner, H.L., Kamienkowski, J.E. and Spies, R., 2022. Thinking out loud, an open-access eeg-based bci dataset for inner speech recognition. *Scientific data*, 9(1), pp.1–17.

Ojha, M.K. and Mukul, M.K., 2021. Detection of target frequency from ssvep signal using

empirical mode decomposition for ssvep based bci inference system. *Wireless personal communications*, 116(1), pp.777–789.

Onose, G., Grozea, C., Anghelescu, A., Daia, C., Sinescu, C.J., Ciurea, A.V., Spircu, T., Mirea, A., Andone, I., Spânu, A. et al., 2012. On the feasibility of using motor imagery eeg-based brain–computer interface in chronic tetraplegics for assistive robotic arm control: a clinical test and long-term post-trial follow-up. *Spinal cord*, 50(8), pp.599–608.

Panachakel, J.T., Ramakrishnan, A. and Ananthapadmanabha, T., 2020. A novel deep learning architecture for decoding imagined speech from eeg. *arxiv preprint arxiv:2003.09374*.

Panachakel, J.T. and Ramakrishnan, A.G., 2021. Decoding covert speech from eeg-a comprehensive review. *Frontiers in neuroscience*, p.392.

Patak, L., Gawlinski, A., Fung, N.I., Doering, L., Berg, J. and Henneman, E.A., 2006. Communication boards in critical care: patients' views. *Applied nursing research*, 19(4), pp.182–190.

Pei, X., Barbour, D.L., Leuthardt, E.C. and Schalk, G., 2011. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal of neural engineering*, 8(4), p.046028.

Poeppel, D. and Idsardi, W., n.d. van wv (2008). *Speech perception at the interface of neurobiology and linguistics. philos. trans. r. soc. lond. b, biol. sci*, 363, pp.1071–108610.

Qureshi, M.N.I., Min, B., Park, H.j., Cho, D., Choi, W. and Lee, B., 2017. Multiclass classification of word imagination speech with hybrid connectivity features. *Ieee transactions on biomedical engineering*, 65(10), pp.2168–2177.

Rapin, M., Braun, F., Adler, A., Wacker, J., Frerichs, I., Vogt, B. and Chetelat, O., 2018. Wearable sensors for frequency-multiplexed eit and multilead ecg data acquisition. *Ieee transactions on biomedical engineering*, 66(3), pp.810–820.

Simistira Liwicki, F., Gupta, V., Saini, R., De, K. and Liwicki, M., 2022. Rethinking the methods and algorithms for inner speech decoding and making them reproducible. *Neurosci*, 3(2), pp.226–244.

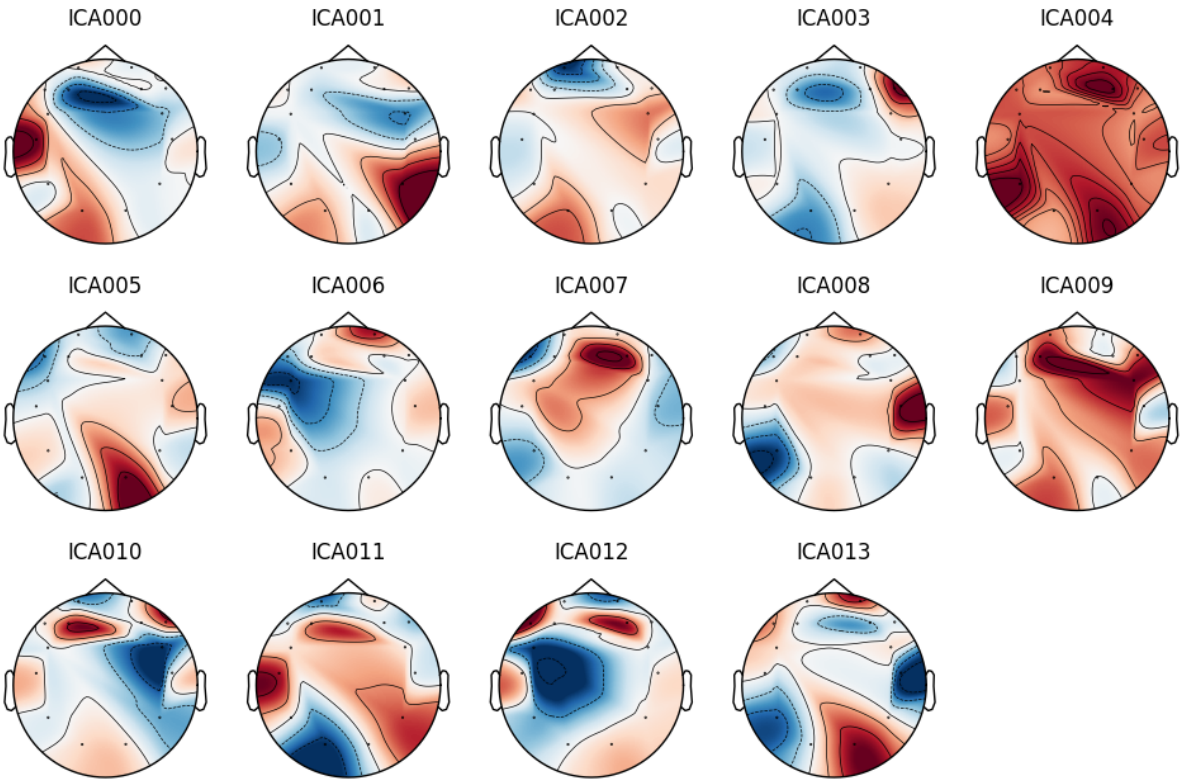
Suppes, P., Lu, Z.L. and Han, B., 1997. Brain wave recognition of words. *Proceedings of the national academy of sciences*, 94(26), pp.14965–14969.

Tøttrup, L., Leerskov, K., Hadsund, J.T., Kamavuako, E.N., Kæseler, R.L. and Jochumsen, M., 2019. Decoding covert speech for intuitive control of brain-computer interfaces based on single-trial eeg: a feasibility study. *2019 ieee 16th international conference on rehabilitation robotics (icorr)*. IEEE, pp.689–693.

Zhao, S. and Rudzicz, F., 2015. Classifying phonological categories in imagined and articulated speech. *2015 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, pp.992–996.

Appendix A.

ICA figures and tables



(a)

Figure A.1: ICA decomposition for P1. We could not identify any artifact.

Table A.1: ICA signal sources that have been excluded for each participant (P1-P9). Each column features the mne.preprocessing.ICA source number (ICA0 + number in Fig. A.1) and the artifact that has been identified in that source.

Imagined Speech								Inner Speech									
P1		P2		P6		P7		P3		P4		P5		P8		P9	
/	/	1	Eye	8	Eye	3	Eye	0	Eye	0	Heart	4	Eye	0	Eye	0	Heart
/	/	2	Eye	/	/	5	Eye	5	Heart	6	Eye	5	Eye	1	Eye	1	Eye
/	/	3	Heart	/	/	/	/	10	Eye	/	/	9	Heart	/	/	2	Eye

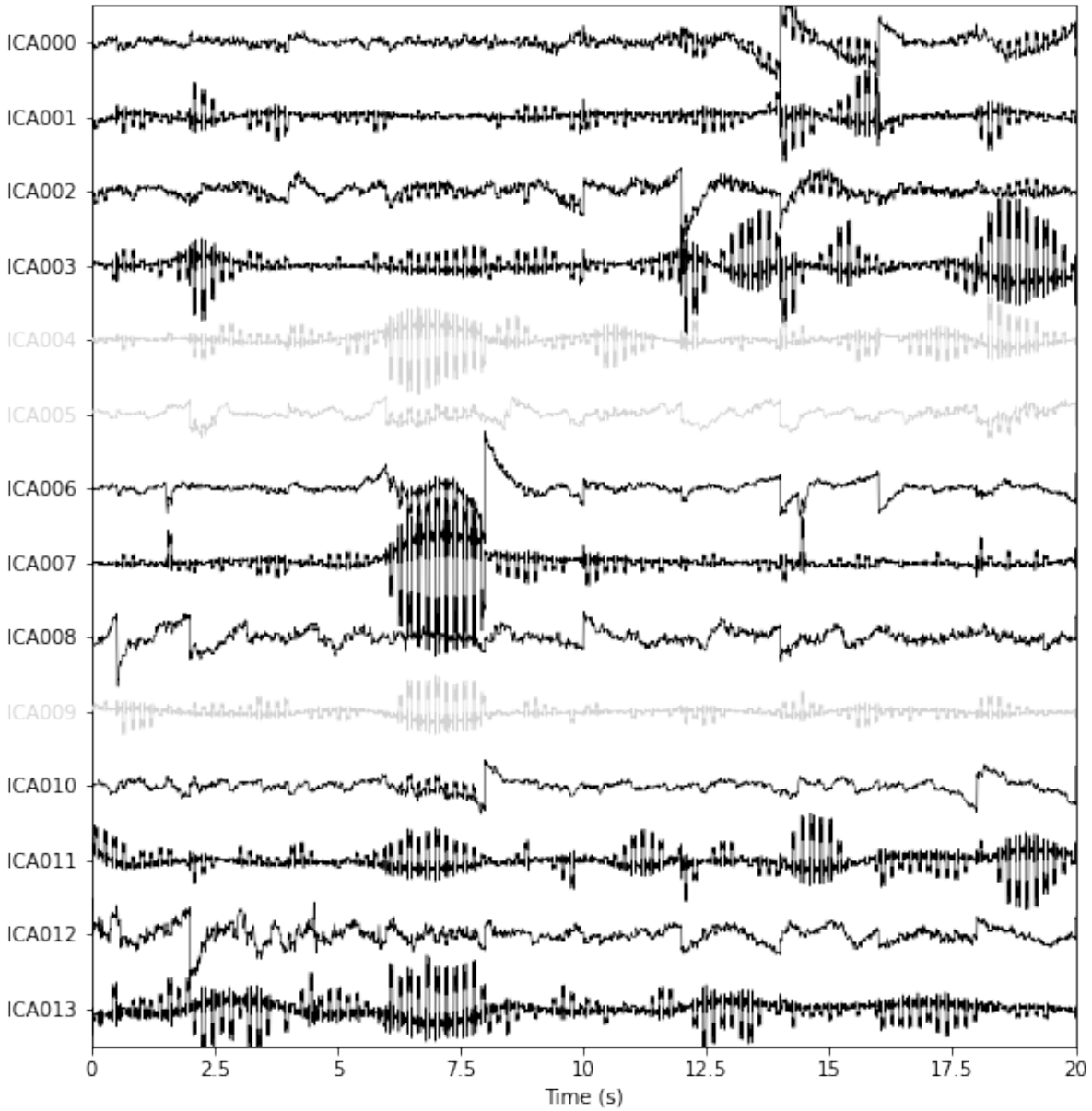
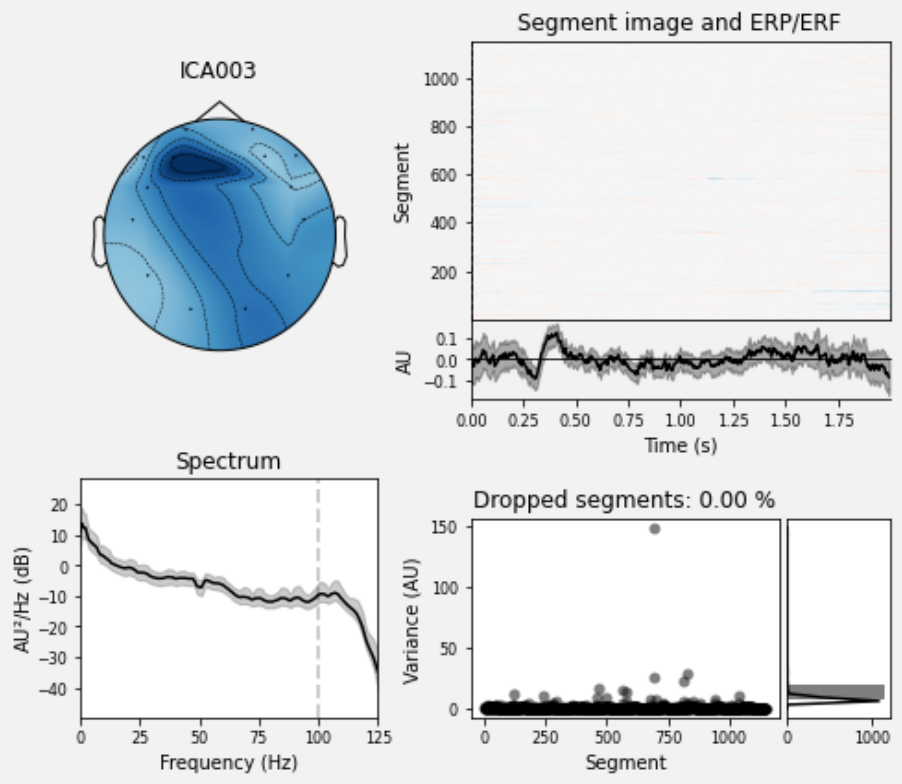
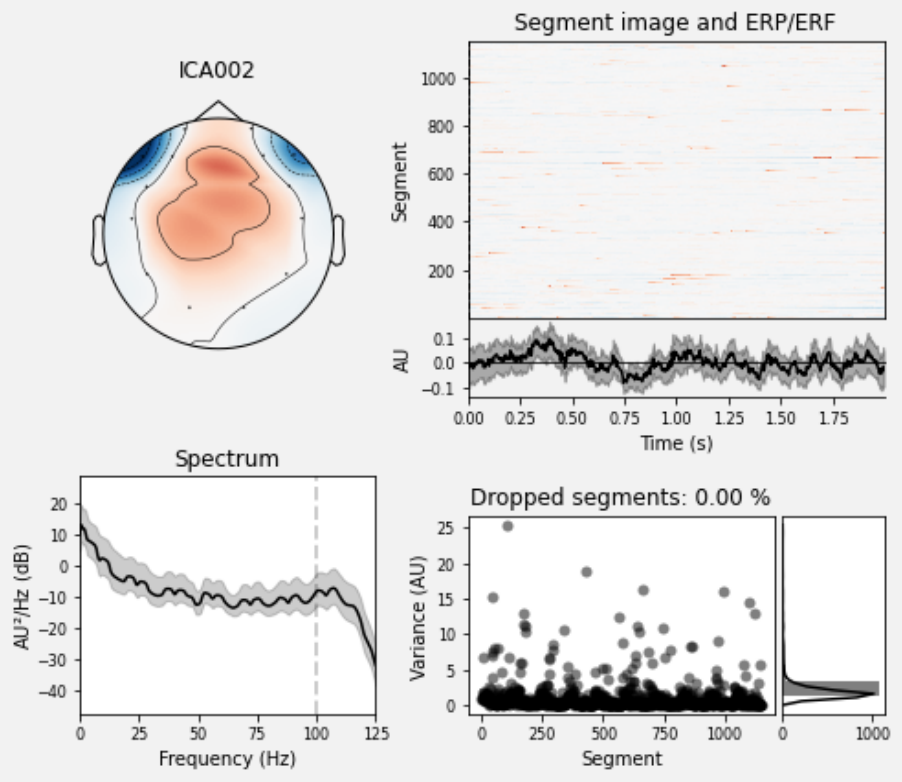


Figure A.2: ICA sources for P5. The faded signals are the excluded artifacts. Own figure.



(a)



(b)

Figure A.3: Clockwise from the left upper corner, the graphs show the artifacts scalp topography, ERP, dropped segments, and power spectra. a) is the cardiac and b) the the eye-blinks source. Own figure.

Appendix B.

Uncertainty propagation

When combining the data of N participants, we calculate the combined mean accuracy μ as

$$\mu = \frac{1}{N}\sum_{p=1}^{N}\mu_n. \tag{B.1}$$

However, to estimate the standard deviation σ_μ of μ, applying the usual formula for a discrete random variable

$$\sigma_\mu = \sqrt{\frac{1}{N}\sum_{p=1}^{N}(\mu_p - \mu)^2} \tag{B.2}$$

to $\mu_1, \mu_2, ..., \mu_N$ would lead to severe underestimation. In our results we find μ_p to be quite close to each other, with large associated σ_p. In this scenario, (B.2) would give $\sigma_\mu < \sigma_p$, which is not sensible.

Let N datasets contain respectively $n_1, n_2, ..., n_N$ observations, with associated means $\mu_1, \mu_2, ..., \mu_N$ and variances $\sigma_1^2, \sigma_2^2, ..., \sigma_N^2$. We can calculate the combined mean with

$$\mu = \frac{\sum_{k=1}^{N} n_k\mu_k}{\sum_{k=1}^{N} n_k}, \tag{B.3}$$

while the combined variance is given by

$$\sigma_\mu^2 = \frac{\sum_{k=1}^{N} n_k \sigma_k^2 + n_k \left(\mu_k - \mu\right)^2}{\sum_{k=1}^{N} n_k}. \tag{B.4}$$

In our case, we want the mean accuracy μ_p of each participant to carry the same weight. This means that we can set $n_1 = n_2 = ... = n_N = 1$, which beautifully reduces (B.3) to (B.1), while (B.4) becomes

$$\sigma_\mu = \sqrt{\frac{1}{N} \sum_{p=1}^{N} \sigma_p^2 + \left(\mu_p - \mu\right)^2}. \tag{B.5}$$

We use (B.1) and (B.5) to estimate combined mean and variance when combining data from different participants.