Learning to read the mind: A counter-intuitive investigation into mode of stimulus delivery used in data collection, and its influence in the decoding of inner and imagined speech with the use of Machine Learning.

Will Adkins

MSc Data Science
University of Bath
May 2022

Learning to read the mind: A counter-intuitive investigation into mode of stimulus delivery used in data collection, and its influence in the decoding of inner and imagined speech with the use of Machine Learning.

Submitted by: Will Adkins

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: Riccardo Bonzano and Ruthwik Hosur Paramashivaiah. In particular, the data collection, experimental paradigm and main framework of investigation have been developed collaboratively. Otherwise, except where specifically acknowledged, this dissertation is the work of the author. Code for the data collection can be found at "https://github.com/Mithrandir98/BCI_paradigm", and code for the analysis/investigation at "https://github.com/WillSAdkins/MSc_project_EEG"

Abstract

This work challenges the seemingly tacit assumption that mode of stimuli used in stimulus stages of recording structures is unimportant in collecting high quality EEG data. Where there not only appears to be an inherent lack of justification for choices behind the mode of stimuli selected throughout the field of decoding inner and imagined speech, but deeper than that, a lack of research into the topic. There is no/limited information on how mode of stimuli might improve speech imagery, or how it may lead to bias/corrupted signals. This research involves the design of an experimental paradigm for conducting experiments on 9 adults to investigate the effect of mode of stimuli used for prompting participants in the case of performing inner and imagined speech (5 participants performing Inner, 4 participants performing Imagined). It was found that mode of stimuli used in the stimulus delivery/prompting stage did not affect classification accuracy for decoding inner or imagined speech in the case of 4-way classification, binary classification of vowel pairs, or binary classification of mono vs tri syllabic words.

Contents

1 Introduction **1**
- 1.1 Motivation & Background . . . 1
- 1.2 Objectives . . . 2

2 Literature and Technology Survey **3**
- 2.1 Mode of stimulus delivery/prompting . . . 3
- 2.2 Case studies for data collection with use of EEG devices . . . 4
- 2.3 Decoding inner and imagined speech from brain signal data . . . 5
  - 2.3.1 Preprocessing . . . 5
  - 2.3.2 Feature extraction . . . 7
  - 2.3.3 Application of Machine Learning methods . . . 7

3 Methodology **9**
- 3.1 Experimental paradigm . . . 9
  - 3.1.1 Data collection . . . 9
- 3.2 Pre-processing . . . 17
  - 3.2.1 Pipeline structure . . . 18
  - 3.2.2 Epoch rejection . . . 23
  - 3.2.3 Filtering . . . 23
  - 3.2.4 Independent component analysis (ICA) . . . 23
- 3.3 Feature extraction . . . 24
- 3.4 Application of Machine Learning models . . . 24
- 3.5 Performance metrics and hypothesis testing . . . 27

4 Results **29**

5 Discussion **36**
- 5.1 Evoked/average responses . . . 36
  - 5.1.1 Stimulus stage . . . 36
  - 5.1.2 Speaking stage . . . 37
- 5.2 Performance of machine learning models . . . 37

6 Conclusions **39**

Bibliography **40**

List of Figures

2.1 2D/topographic map of the internationally recognized "standard 10-20 montage", which describes the location of scalp electrodes in the context of EEG recordings with each point given an abbreviated label from the terminology commonly used in Neuroscience describing the region for which the electrode is placed upon. 6

3.1 Graphical representation of the experimental paradigm 11
3.2 Visual and text representations of the 4 words plus the rest class 12
3.3 Multi-angle view of the Emotiv EPOCH+ being worn, with corrct positioning on the head. 14
3.4 2D/topographic map of the 14 channel positions. Plot is for visualisation purposes only, and the positions used in later analysis were not based on the values/positions in this graph 14
3.5 Plot of 0.3 seconds unprocessed speaking data across all 14 channels from a single participant, displaying the frozen values within their data/recordings. . 16
3.6 Plot of 2 epochs of unfiltered data (consecutive in the dataset) 20
3.7 Plot of 2 epochs of data band-pass filtered between 1-70Hz (consecutive in the dataset) 21
3.8 Graphical representation of the order in the processing pipeline used in this research from Data Acquisition through the preprocessing steps to feature extraction and classification 22
3.9 Plot of power spectral density throughout frequencies 0 to 128 Hz by channel, displaying at least 16 observable spikes, and the large of amount of contamination within the signals 24
3.10 Example for a component judged to be an eye component, with plot of frequency spectrum, topographical map of brain, and event related potential's thorughout each epoch 25
3.11 Example for a component judged to be an eye component for the right eye, with plot of frequency spectrum, topographical map of brain, and event related potential's thorughout each epoch 25
3.12 Example for a component judged to be an eye component for the left eye, with plot of frequency spectrum, topographical map of brain, and event related potential's throughout each epoch 26
3.13 Graph of the procedure for calculating tangent vectors of Cross channel covariance matrices on training and testing data. Adapted from the paper by Nguyen et al (2017) [18] 26

3.14 Graph of the deep neural network used for 4-way classification, binary classification of vowel pairs, and binary classification of monosyllabic vs trisyllabic words (where i=105, m=50, j=100, c=4 in 4-way classification, and c=2 in binary classification). It must be noted that linear activation functions were used for hidden layers with dropout regularisation applied to the third hidden layer with a rate of 0.2, and a sigmoid activation function was used on the output layer 28

4.1 Topographic map of evoked response throughout the stimulus stage for participant 1, displaying behaviour in line with theory in the literature. 30
4.2 Topographic map of evoked response throughout the stimulus stage for participant 3, displaying behaviour not in line with theory in the literature. 31
4.3 Topographic map of evoked response throughout the performance of imagined speech and during rest class (participant 1). 32
4.4 Topographic map of evoked response throughout the performance of inner speech and during rest class (participant 4). 33

List of Tables

S1 5-fold cross validation classification accuracy to 1 d.p (mean ± standard deviation %) in the case of 4-way classification of Ambulance, Hospital, Lamp, Clock 34

S2 5-fold cross validation classification accuracy to 1 d.p (mean ± standard deviation %) in the case of binary classification of syllable pairs (Ambulance and Hospital versus Lamp and Clock) 34

S3 5-fold cross validation classification accuracy to 1 d.p (mean ± standard deviation %) in the case of binary classification of vowel pairs (Ambulance and Lamp versus Hospital and Clock) 35

S4 p-values corresponding to a One-way ANOVA test between mean classification accuracy between models trained on the audio, Visual and text data. In the case of each type of classification and each algorithm for inner and imagined speech. 35

S5 Table displaying positive correlation between number of samples per class (n) after processing versus the number of models performing significantly above chance level (no. Sig). (Pearsons R = 0.44) 35

Acknowledgements

Special thanks to my collaborators Riccardo Bonzano and Ruthwik Hosur Paramashivaiah, and also my supervisors Scott Wellington and Jim Laird for all the help and support for which i am extremely grateful.

Chapter 1

Introduction

1.1 Motivation & Background

Brain-computer interfaces (BCI) allow for communication or control of external devices using brain signals rather than the brain's normal output pathways of peripheral nerves and muscles [14]. The quantity of research into the application of Brain-computer interfaces has increased over recent decades, whilst companies such as Neuralink, Neurable and Emotiv have experienced a boost in positive sentiment. Applications vary from rehabilition/prevention of serious injuries, mind reading, remote communication, gaming and entertainment [1]. In particular, the use of BCI are potentially revolutionary in providing speech to those unable, through the decoding of inner or imagined speech.

Inner speech can be defined as "the subjective experience of language in the absence of overt and audible articulation" [3]. Or in simpler/layman terms, the use of ones inner voice. This is different to that of imagined speech, which can be defined as "imagining the pronunciation of words as if pronouncing it aloud, but without any articulator movements" [4]. The application of Brain-Computer interfaces extends to the decoding of inner and imagined speech, and this is the area/field with which the investigation carried in this work operates within. Recorded brain activity during the performance of imagined or inner speech can be decoded with the use of machine learning models with both invasive recording methods (for instance ECoG [15]) and non-invasive recording methods (fMRI [10][8] and EEG [6][27]) having been used previously. Data collection methodologies for recording such data follow a very similar process involving (over a number of repeated trials), a participant starting off relaxed/at rest, followed by a prompt of what to speak being presented, followed by a window where the participant must perform such speech imagery. The choice behind mode of stimuli used in prompting a participant with what to speak (for example a text presentation, or an audio presentation, an image) does not appear to be adequately justified within the literature, and there appears to be an unspoken/tacit assumption that this choice is unimportant for successful collection of quality of brain signal data for the purpose of decoding inner and imagined speech using machine learning. This work seeks to explore the influence on decoding inner and imagined speech, of different modes of stimulus being used in prompting participants with what to speak, namely audio, visual and text stimuli.

In their review of decoding covert/imagined speech using EEG, Panachakel et al [19] claim that EEG, ECoG and EMG are the three most common modalities used in the literature for decoding

imagined speech, EEG being the most commonly used. With other modalities such as MEG, fMRI, and fNIRS also observed in the literature. At the cost of lower signal to noise ratios compared to invasive methods such as ECoG and high likelihood of corruption from muscular artifacts. EEG devices are cheap, portable, easy to use, non-invasive and have good temporal resolution [19]. In practical applications for providing a method of communication for those who cannot speak, for an affordable BCI that could be adopted on a large scale, a solution may likely need to be found using low cost, likely non-invasive, easy to use commercial grade devices. Research into the use of EEG devices (in particular commercial grade) is therefore important.

In order for the decoding of speech to be successful, due to low signal to noise ratio, brain signal data often requires vast amounts of pre-processing, feature engineering and decomposition before applying machine learning methods. This leaves researchers with the task of not only exploring the relative merit of various machine learning algorithms for decoding speech, but also recording quality data with EEG devices, and applying appropriate signal processing and feature extraction methods. Active research groups exist in this field such as that at UCSF (Edward Chang), Universität Bremen (Tanja Schultz), and Ciaran Cooney's team at Ulster University.

1.2 Objectives

The focus of this research will be on the influence mode of stimuli used for prompting has on the success of decoding both inner and imagined speech with the use of machine learning. Exploring the influence of using audio, text and visual prompts on the ability to decode inner and imagined speech, and consequently add to the currently available open access EEG datasets for speech processing. Introducing a dataset with a novel component surrounding the method in which participants are prompted, and a detailed experimental paradigm which could be replicated by other research teams. The research questions and hypotheses are as follows:

1. Does mode of stimulus affect the success with which speech can be decoded, in terms of 4-way classification accuracy. With null hypothesis:

   H_0 : Mode of stimulus does not affect classification accuracy for models decoding inner or imagined speech.

2. Does mode of stimulus affect the success with which the speech of monosyllabic vs trisyllabic nouns can be decoded, in terms of binary classification accuracy. With null hypothesis:

   H_0 : Mode of stimulus does not affect the classification accuracy for models decoding inner or imagined speech of monosyllabic vs trisyllabic nouns.

3. Does mode of stimulus affect the success with which speech can be decoded for nouns with vowels /æ/ vs /ɒ/, in terms of binary classification accuracy. With null hypothesis:

   H_0 : Mode of stimulus does not affect the classification accuracy for models decoding inner or imagined speech of nouns with vowels /æ/ vs /ɒ/.

Chapter 2

Literature and Technology Survey

2.1 Mode of stimulus delivery/prompting

Throughout the work relating to speech imagery, auditory, text, visual/image and multi-modal combinations of prompts have been used to not only prompt participants what to speak but also in some cases, when to speak [19]. There however appears to be an inherent lack of justification for the choice of mode of stimuli used to prompt participants, and without supporting evidence it would be naive to assume this choice is unimportant and has no effect on the performance of speech/imagination tasks. On a neurological level Audio and Visual stimuli are processed in different regions. The Auditory cortex is responsible for processing audio information, and the visual cortex, visual information, with the primary auditory cortex located within the temporal lobe, and the primary visual cortex in the occipital lobe [20]. In addition to this, the degree to which stimuli contain speech and language based information will also influence areas of activation. Broca's and Wernickes areas are said to be responsible for speech and language processing along with the angular gyrus, with the motor cortex transmitting information from Broca's area to the articulators to form actualised speech [21]. In the context of EEG signals, Walsh et al (1979) [25] conducted a study into evoked brain responses (from EEG recordings) to auditory and visual stimuli of equal subjective magnitude, with the use of power functions calculated from the evoked brain responses as a metric for physical stimulus intensity. It was found that loudness (audio stimuli) power functions grew faster than those for brightness (visual stimuli) with an increase in subjective magnitude. Interestingly it was also found that Evoked brain response waveforms for bimodal (audio and visual) stimuli appeared to be an "algebraic summation" of unimodal evoked brain response waveforms, with no sensory interaction found between audio and visual Evoked brain response waveforms. In simpler terms, this research suggests a difference exists between the response to audio and visual stimuli. Its therefore likely a difference in the evoked response to stimuli exists within the collection of the EEG data based upon the mode of stimulus used from prompting participants. The audio and visual stimuli used in this research was not speech related however, and Vander Wyk et al (2010) [23] found that when presenting auditory and visual stimuli simultaneously, the cortical regions that got activated depended on the extent to which speech or non-speech percepts were incorporated into stimuli, with more similar regions of activation observed when stimuli contained speech percepts. Although this draws no conclusion on the the link between comprehension of words/phonemes represented by stimuli that get presented, and the subsequent imagination/pronunciation, it introduces the mode of stimuli used (more

specifically the response to different modes of stimuli) as a variable with some form of influence and impact within the data collection stage.

Although research into the evoked response to audio and visual stimuli has been explored, in terms of regions of activation. The influence of mode of stimulus in the data collection stage has not been explored in terms of its effect classification accuracy for models decoding inner and imagined speech, and due to the ambiguity in which regions of the brain are responsible/related to the production of inner and imagined speech, the influence mode of stimulus has in the data collection stage is still somewhat unknown. To fill this gap is the aim of the investigation at hand.

2.2 Case studies for data collection with use of EEG devices

Due to large amounts of noise within brain signal data recorded using non-invasive methods, quality data collection practices are important. Care must taken, with all participants prepared to minimise negative influence from external factors, such as artifacts arising from eye movement [27], audio or visual distractions (within the environment), and mental fatigue etc. In this section, we provide a review of some of the different factors that have importance in the recording process, hence must be considered in decisions about how to best collect and record the data. We provide a review of three datasets similar in format/structure to that which we intend to record in this project (EEG recordings, low number of channels), with the second dataset (FEIS) using the same recording device as in this work (14-Channel Emotiv EPOC+).

Zhao and Rudzicz (2015) present the Kara One dataset recording EEG, facial, and audio data during imagined and vocalized speech of phonemic and single-word prompts, entailing the use of a Microsoft Kinect (v.1.8) camera and a 64-channel Neuroscan Quick-cap for recording [27]. Participants were seated in a chair before a computer monitor, in an office environment, and a format was used where each trial contained 4 successive stages. Firstly a rest stage, followed by a prompt stage, an imagining stage (where a prompt was imagined), and then a speaking stage (where a prompt was spoken). A break was provided every 40 trials, and within the rest stage, participants relaxed/rested their mind for 5 seconds. These strategically placed rest/relaxation periods should in turn slow down the rate at which the participant experiences mental fatigue, and in an abstract sense quieten the mind, ultimately maintaining the quality of recordings. Moreover, during the prompt stage audio and visual stimuli were presented to indicate what should be spoken in the subsequent stages. A 2-second period was then provided between the prompt stage and the imagine stage in which the participant moved their articulators into position to begin pronouncing the prompt, this prevents participants from moving their articulators during the imagined or speaking stage and potentially influencing the brain signals that get recorded during this time, further maintaining the quality of recordings. When designing a data collection methodology its important to factor in mental fatigue, and asking participants to take part in the trials for large amounts of time could dramatically influence the quality of recordings. During the collection of the Kara One dataset, each participant spent 30 to 40 minutes participating in the study, allowing for a reasonable quantity of data to be recorded whilst avoiding a drop off in the quality of data due to the participant becoming mentally fatigued, distracted or disinterested. "7 phonemic/syllabic prompts and 4 words derived from Kent's list of phonetically-similar pairs were used. These prompts were chosen to maintain a relatively even number of nasals, plosives, and vowels, as well as voiced

and unvoiced phonemes"[27]. This exposes subsequent machine learning models to phonemes throughout the phonetic space, which could allow for subsequent research using this data to make scientific claims about how the phonetics of a given prompt may affect machine learning algorithm's ability to decode. The data collection procedure proved not to be perfect however, it is noted that "Data from 4 of the 12 participants were discarded due to unattached ground wires and two participants falling asleep during recording" [27].

Another example of where EEG recordings for decoding brain signals for speech processing has been collected is in the FEIS (Fourteen-channel EEG for Imagined Speech) dataset [6]. In this paper 21 English-speaking participants took part in recordings, with a 14-channel mobile headset with dry electrodes"used. A very similar data collection methodology is employed as in Kara One where each trial has a Rest, Prompt, imagining and speaking stage. Where the Kara One dataset was conducted in an office environment, for the FEIS dataset, participants carried out the experiment alone, sitting in a comfortable chair in front of a laptop screen, inside a hemianechoic chamber [6]. This environment is an improvement upon the one used for the collection of the Kara One dataset in terms of minimising external audio and visual stimulus that could be a potential distraction. The FEIS dataset contains no word prompts, rather sixteen English phonemes that were chosen to "represent a balanced categorical spread of binary phonological features ([±nasal], [±back], [±voice], etc)" also allowing for coverage throughout the phonetic space. Moreover, a combination of audio and text prompts/stimuli were used in the prompt/stimulus stage, however, interestingly for the audio stimuli, participants are played a clip of the prompt looped five times, spoken in their own voice. The use of the participants own voice should further improve the comprehension of a prompt due to familiarity with ones own voice, helping to improve quality of recordings. These small decisions that help control the data, and suggest it towards a particular direction can lead to an increase in the quality of results, this kind of thinking could be pivotal in obtaining a quality dataset within this project. For the reasons aforementioned, the format (Rest, prompt, imagine, speak) in which the EEG recordings from Kara One dataset were recorded and then was adapted for the FEIS dataset would be an excellent choice for a base format in the project at hand, and could be adjusted as required.

A practical BCI for decoding imagined speech would need to be asynchronous (not bound by a particular time period and prompt/cue) and online (processed in real time). This is far more complex than that which is most common in previous work (synchronous and offline)[13], requiring a model that could distinguish whether a word/phone is actually being communicated or whether the user is at rest (not imagining speaking), in real time. It must be noted that both the KARA ONE and the FEIS dataset have speaking/imagination period bound by specific time and hence would not be classed as asynchronous. This is a topic for future work within this area, and one of the next progressions towards a practical BCI for decoding speech.

2.3 Decoding inner and imagined speech from brain signal data

2.3.1 Preprocessing

The aim of pre-processing EEG signals is to increase the signal to noise ratio of the data, and get closer to the signal of interest. Artifacts often exist in the data arising from things such as power-line noise, heartbeat signals, eye movements and motor actions for instance. To combat

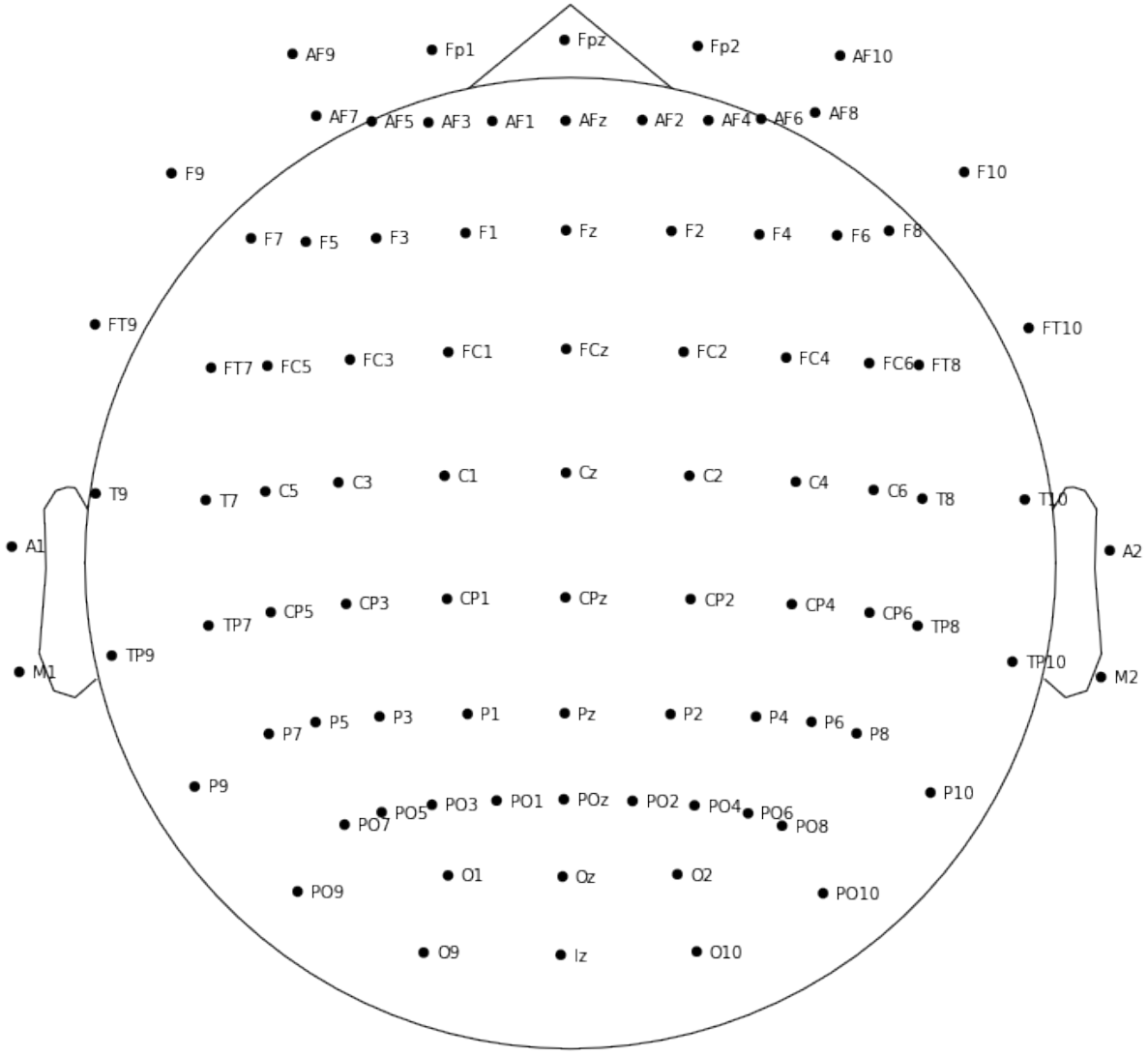

Figure 2.1: 2D/topographic map of the internationally recognized "standard 10-20 montage", which describes the location of scalp electrodes in the context of EEG recordings with each point given an abbreviated label from the terminology commonly used in Neuroscience describing the region for which the electrode is placed upon.

this various techniques can be applied such as resampling, temporal and/or spatial filtering, Independant component analysis, channel selection and epoch rejection to name a few. Choice behind pre-processing techniques are objective and data dependant, however a pattern appears to exist within the literature for decoding inner and imagined speech, and this is elaborated on well in the work by panachakel et al (2021) [19]. Re-sampling is often employed first to reduce the dimensionality of the data, for instance Nguyen et al [18] resample from 1kHz to 256Hz reducing the computational complexity of running feature extraction techniques and training machine learning models, where in this work the feature extraction method ("tangent vectors calculated from Channel-cross covariance matricies") can be time consuming. This is often then followed by removing signals occuring at certain frequencies deemed to be not of interest, and temporal filtering methods can be employed to remove or attenuate such frequencies. There is currently a lack of agreement between researchers on the frequency bands responsible for imagined speech and selected frequency bands are often dependant on the artifacts present in the data [19]. For removal of artifacts such as that from muscle movements and eye blinks, both epoch rejection and Independent component analysis can be employed [13].

2.3.2 Feature extraction

Due to the large amount of data and complexity of the relationship between brain signal and speech, contextualising data is of utmost importance. Cooney et el (2018) [7] suggest that "the difficulty of accurately decoding speech using EEG recordings is due to factors such as weak neural correlates and spatial specificity, and signal noise during the recording process". In their 2018 paper they investigate the use of a "feature-set comprised only of mel frequency cepstral coefficients (MFCC)" for imagineed speech decoding from EEG recordings. By training both a decision tree classifier and a support vector machine using linear, non-linear and MFCC feature sets, it was found that MFCC features lead to higher classification accuracy compared to linear and non-linear feature sets. MFCC features are considered state of the art in speech recognition and results from this study suggest that MFCC features may be a wise choice in the context of this project (BCI context) and will be further considered. Although extracting features from individual channels (as with MFCC's) is easier, extracting features from multiple channels simultaneously helps for the relationship between different areas of the brain/chennels to be better understood [13]. Channel cross-covariance (CCV) matrices are a popular example, for example in Nguyen et al (2017) [18], tangent vectors were calculated from CCV matrices, and using a relevance vector machine algorithm, "shown to outperform other approaches in the field with respect to accuracy and robustness" on an imagined speech task.

2.3.3 Application of Machine Learning methods

Various machine learning algorithms including deep learning based methods such as a 3D convolutional neural network [5] and Recurrent neural network [15] have been previously explored for decoding of speech using brain signal data. As well as some Non-deep learning based methods like Gaussian processes [26], tree based methods (Random Forest [2]) and support vector machines [27]. As per every machine learning task, the best performing algorithm is data and subject dependant, hence its not wise make theoretical claims on which model is most likely to be able to decode inner or imagined speech, however due to the complex patterns in brain activity, one might think that deep learning methods would be best suited given how well they are often able to pick up complex patterns and relationships within data. This is supported by the information from the review by Panachakel et al (2021) [19]

on the frequent best performing algorithms within the field of imagined speech, where deep learning based methods (Deep neural networks, Convolutional neural networks, Recurrent neural networks, etc) were by far the algorithm found to be "best classifier" most often for imagined/covert speech.

Chapter 3

Methodology

3.1 Experimental paradigm

3.1.1 Data collection

Recording structure

With inspiration taken from data collection methodologies observed in the literature, in particular the KARA ONE dataset [27], FEIS dataset [6], and the work of Lee et al (2019) [12], a structure for data collection was designed whereby imagined/inner speech of 4 words (nouns) was recorded. Recordings consisted of 130 trials made up of three successive stages (**Relax**, **Stimulus delivery/prompt**, **Think/Imagine**), as seen in Figure 3.2. Within the **Relax** stage, participants take 5 seconds to clear the mind in some way, with a simple instruction "RELAX" presented on screen. The aim of this "clearing of the mind" is to prevent distracted or tangential thinking within the subsequent stages for which bias could be introduced into the data. In the **Stimulus delivery/prompt** stage, for 2 seconds, a prompt is provided for the word to be spoken/imagined in the following **Think/Imagine** stage. This is provided through the mode of an auditory, a visual or a text prompt (subject to alteration, in line with the research hypotheses). A 1 second period is then provided after this to prepare for the **Think/Imagine** stage, and in the case of performing imagined speech, time for the participant to get their articulators into position to begin imagination of the prompt. Participants are then instructed to perform their inner/imagined speech during the **Think/Imagine** stage. In the design of the experimental paradigm, a trade-off between quality of data, quantity of data, and recording time had to be considered. 100 exemplars/observations per class was deemed the minimum required for the training of accurate machine learning models, and in order to collect this without long/laborious recording times, recording length was constrained to roughly 1 hour. Panachakel et al [19] support the use of repeated imaginations within a single trial, claiming "EEG signatures become more prominent across multiple imaginations in the same trial", and it was calculated that using 10 repeats per trial in the **Think/Imagine** stage would allow 100 exemplars to be recorded within roughly an hour given the pre-determined timings for each stage and trial, all whilst potentially leading to more prominent and better quality signals. To further try and prevent mental fatigue, a 120 second break was included every 20 trials, where participants were allowed to relax their thinking for an extended period whilst remaining seated without touching or moving the headset. In total 130 trials are recorded (28 seconds each), and 6 breaks are provided (120 seconds each), which gives a total recording

time of 1hr 40 seconds (1 hour 12 minutes 40 seconds with breaks).

In order to minimise the possibility of bias between the data pertaining to each mode of stimuli, a decision had to be made on whether it would be best to conduct all recordings for all modes of stimuli within a single session, or whether to conduct three individual sessions, one for each mode of stimuli. Consideration was given on participants experiencing what is known as the "bleed-through" effect, in which during a trial, the memory of having seen the given word for imagination previously in another modality, may influence imagination during the present trial, leading to bias between modalities. Fully avoiding the "bleed-through" effect is somewhat unrealistic, and likely to exist in both setups. In theory conducting three individual sessions would likely minimise the "bleed-through" of information between modalities, with previously observed modalities being less "fresh in the mind" compared to the single session setup. However considering that each of the individual sessions would lead to different electrode placements, and that it may be difficult to get (volunteer) participants to attend three individual sessions. It was decided that conducting all recordings, for all three modes of stimuli, in one session would minimise bias between modalities, albeit not from "bleed-through" of information between modalities. In order to try and minimise the bias both within modalities and within words, and allow for the outlined hypotheses to be tested, the occurrence of each mode of stimuli-word pair was random throughout with the condition that each mode of stimuli and each word were observed an equal number of times.

A practical BCI would need to be online and also asynchronous [13], meaning speech must decoded in real time, with data not bound by a time window, creating the need for BCI to be able to determine whether or not a word is being spoken, and only output a speech prediction when a user is performing speech imagery (rather than at rest). In order to try and make the paradigm closer to the asynchronous goal, a rest class has been included in the dataset. This meant when prompted to rest within the **Stimulus delivery/prompt stage**, participants would intentionally not perform speech imagery of any word during the **Think/Imagine**. Although speech will still be decoded "offline" regardless, there is now a possibility for to classification of whether a word is actually being imagined or not, rather than a word always being predicted, closer to as in a completely asynchronous system. There is no direct need to include the rest class for testing the hypotheses surrounding mode of stimuli used in prompting, however this addition to the paradigm increases the usability of the data, and opens up for other researchers to perform further work using the data. The decision to include a rest class, also contributed to the choice to shorten the **Stimulus delivery/prompt** stage to 2 seconds from 5 seconds as observed in KARA ONE [27] and FEIS datasets [6]. By keeping the **Stimulus delivery/prompt** stage the same length as the **Think/Imagine** stage, the data from the rest classes can be compared to both the imagined/inner speech data and the stimulus data. As in the work by Lee et al (2019) [12], allowing for comparison between when participants are "attending to stimuli verus rest" in the case of the different modes of stimulus, but also "Performing imagery versus rest" in the case of imagined and innner speech". Shortening the **Stimulus delivery/prompt** stage to 2 seconds also saves 3 seconds per trial, which amounts to a long period of time over the course of a whole session, this change allowed for more data to be recorded in less time, which alone provided adequate motivation for this decision.

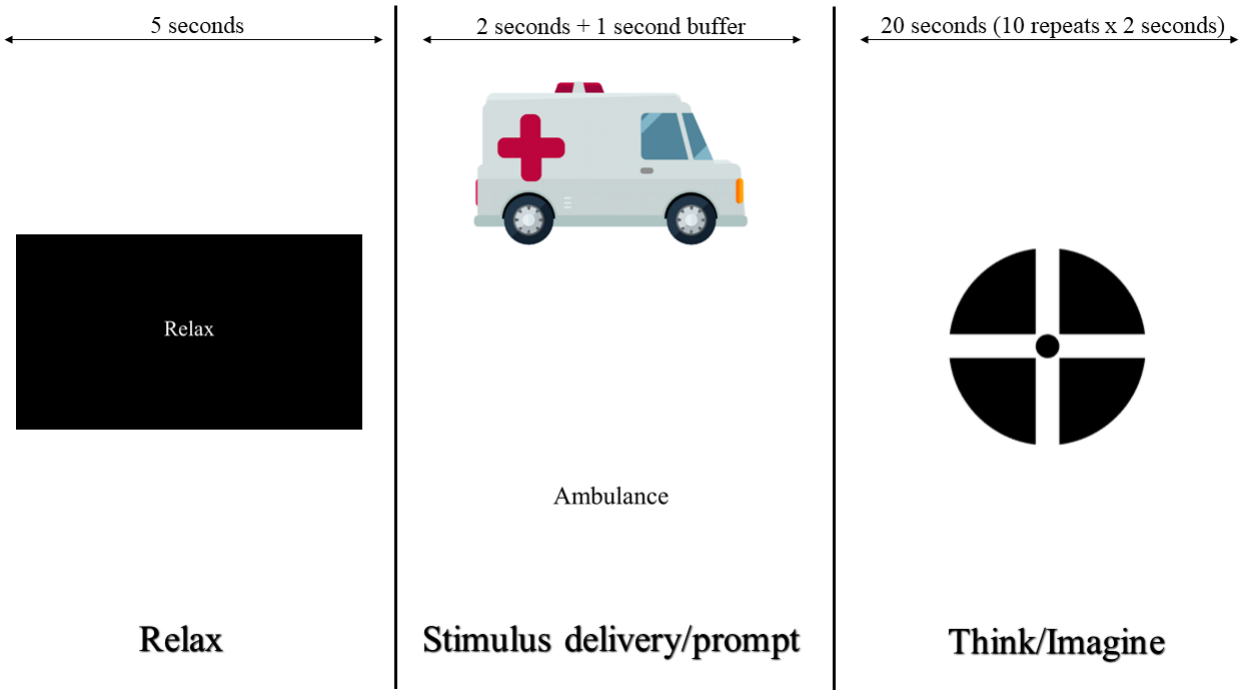

Figure 3.1: Graphical representation of the experimental paradigm

Choice of words

Both the practicality of words for a working BCI and the number of empirical claims that could be made during the investigation were considered in the choice of words to include in the paradigm, with word phonetics, complexity (number of syllables), and syntax all considered. Similar to the work by Lee et al (2019) [12], healthcare related words were used to fuel the "possibility of expanding the BCI communication system into the real world for various purposes". To allow for more detailed analysis within the investigation and to increase the usability of the data as a consequence, the choice of words was variable on the phonetic level, variable in the number of syllables, with fixed syntax. This allows for comparison across two additional dimensions, allowing for the investigation into research questions 2 and 3. The two sets of phonetically similar pairs were chosen with "Ambulance" and "Lamp" both having the /æ/ vowel, and "Hospital" and "Clock" both having the /ɒ/ vowel. Within these two pairs, the number of syllables was also varied with one monosyllabic and one trisyllabic word included. "Ambulance" and "Hospital" being trisyllabic, "Clock" and "Lamp" being monosyllabic. With all four words being nouns and hence fixed in terms of syntax.

Choice of prompts

In order to test the hypotheses surrounding the influence of mode of stimuli used to prompt participants in decoding speech, mode of stimuli provided in the **Stimulus delivery/prompt** stage were altered between trials, with one of a Visual, audio, or text representation used for a given word. Each of the 4 words are observed an equal number of times, throughout the 3 modes of stimulus, leading to 13 different classes observed within the data, when including the rest class. For the visual representations, cartoon images were chosen in order to keep the most simple representations possible. Real life visualisations introduce other phenomena not related to the subject of the image. For instance, A real life picture of an ambulance may

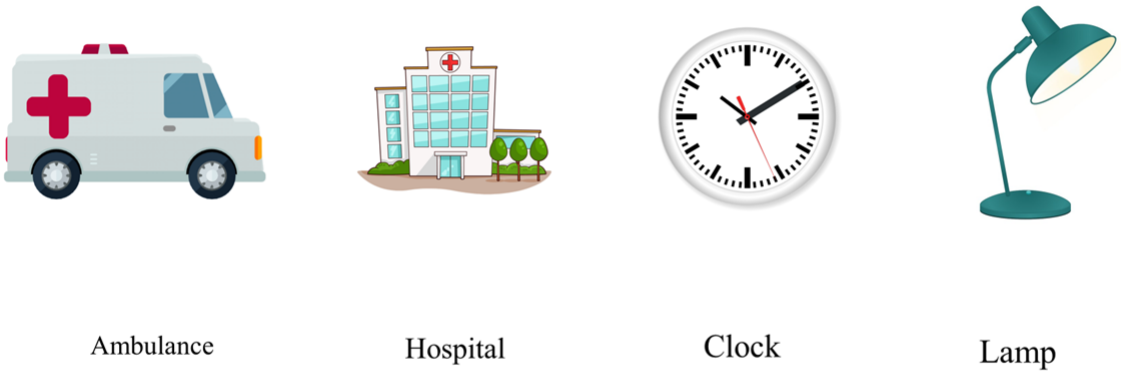

Figure 3.2: Visual and text representations of the 4 words plus the rest class

include a road, trees, or signposts, which could lead to bias in the response to stimuli. In the case of text representations, the word was simply presented in the center of the screen, with each word being of the same font size. Audio representations were recorded by the researcher (Will Adkins) who speaks English as a first language with a south-western English accent, and in the 2 second **Stimulus delivery/prompt** stage, words were spoken/played a single time.

In order to prompt participants when to perform speech imagery in the **Think/Imagine** stage, a marker was used to indicate the rhythm and timing (identical to that used in the FEIS dataset [6], visible in the **Think/Imagine** portion of Figure 3.2). Starting from the far left of the screen, the marker moves across in 10 equally sized steps to the far right of the screen, with one step taking place every 2 seconds. Participants are instructed to focus on the marker throughout the **Think/Imagine** stage, and perform each imagination upon the movement of the marker, leading to 10 imaginations being recorded per trial. The idea of having the marker move across the screen instead of just flash on and off in the center, is to try and control as much as possible for ocular artifacts, and prevent uncontrolled/unpredictable artifacts from arising. Having the marker flash on and off the screen leaves freedom for participants to move thier eyes around during the imaginations. Having the moving marker creates an ocular artifact that is consistent throughout, and hence can be removed during preprocessing. The alternative option for setting the rhythm and timing is to use an auditory cue for instance a "click" or a "beep" sound, however Panachakel et al [19] claim it is difficult to remove the signature of the auditory cue from the EEG signal recorded during speech imagery, and that it has been shown by Nguyen et al [18] that visual cues elicit responses in the occipital lobe, and since the occipital lobe is involved neither in production nor comprehension of speech, there is the option of discarding the EEG channels over the occipital lobe, eliminating the interference of the visual cue on the EEG recorded during imagined/inner speech.

Conducting speech imagery

Both imagined and inner speech data was collected in this research. Originally 10 participants without any known neurological or speech related disorders (6 males, 4 females, average age 24.9) had agreed to take part in the recordings, and 5 participants were randomly assigned to do inner speech, and the other 5 assigned imagined speech, however the data from one of the participants doing imagined speech was dropped due to the identification of some major flaws, and a change in the recording device used (more detail provided in section 3.1.1), with that participant unable to rerecord their data. Hence in total 5 participants performed inner speech, and 4 imagined. Participants who were allocated to do inner speech, were given the instruction to "speak using your inner voice, do not move, or imagine moving your articulators". And participants who were allocated to do imagined speech, the instruction to "imagine that you are speaking the given word, with imagination of the articulator movements included". What was meant by "articulators" was explained to all participants as "any movable organ that is involved in speech production (tongue and lips etc)". Participants were also requested to try not to blink, and remain as still as possible during speech, in order to try and minimise the number of artifacts in the data, and hence allow for better quality data to be recorded.

Use of 14-channel 256Hz EEG recording device (Emotiv Epoch +)

Originally it had been planned to use the 128Hz EPOCH recording device to conduct recordings, however after analysing recordings from a participant using the device it was observed that excessive noise existed in the data. Ordinarily values exist around $4000\mu V$ as observed in the FEIS dataset [6], however values ranging from 50 μV and 9000 μV were observed, this was thought to be down to hardware damage in multiple areas of the device arm that uncovered wires connected to electrodes, leading to corruption of the signals and large amounts of additional noise in the signal. After experimentation with other available devices, it was found that the 14-channel 256Hz EPOCH + device available at the University of Bath (seen in figure 3.3) did not encounter the same problems surrounding signal noise, and was opted for instead.

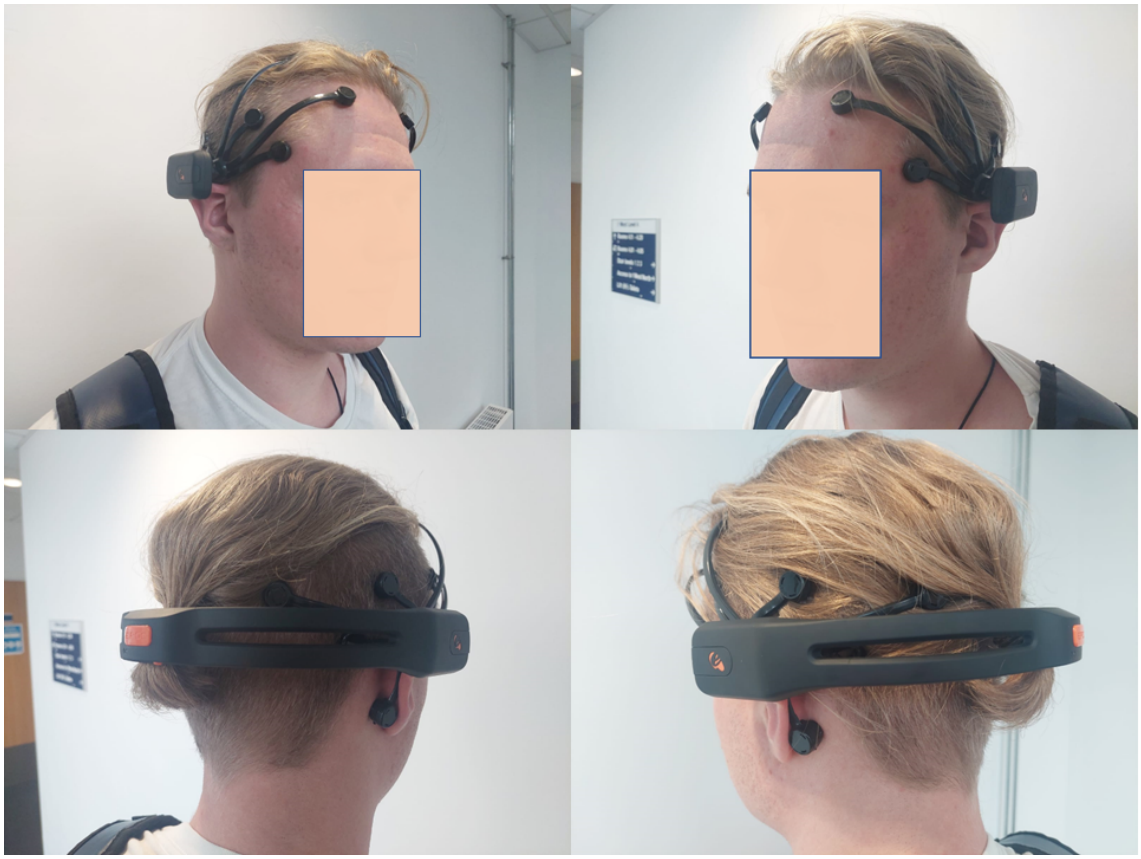

Figure 3.3: Multi-angle view of the Emotiv EPOCH+ being worn, with corrct positioning on the head.

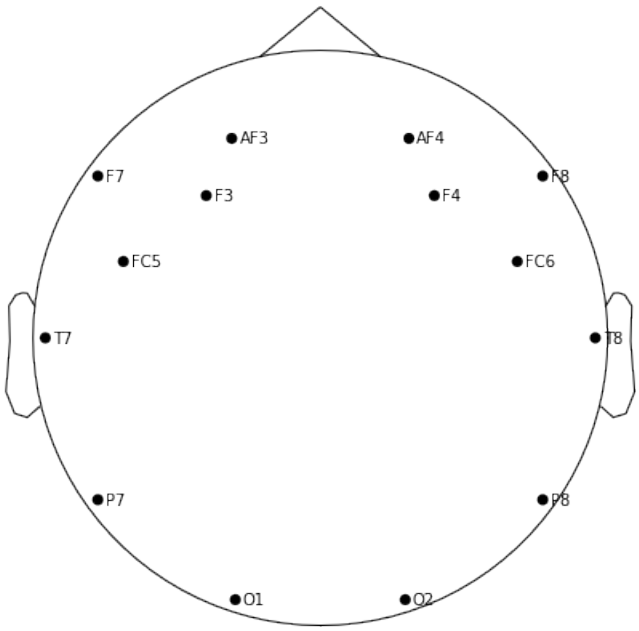

Figure 3.4: 2D/topographic map of the 14 channel positions. Plot is for visualisation purposes only, and the positions used in later analysis were not based on the values/positions in this graph

The most difficult part of using the device was setting up the device on a participant such that each electrode had good impedance levels (<5kΩ). The step by step procedure outlined by Ramele et al (2022)[22] was implemented whereby:

1. Saline solution (0.75% sodium chloride) was applied to each electrode **before** connecting electrodes to the device arm.
2. Pressing electrodes firmly to the scalp, each electrode was adjusted such that the device arm could exert a force on the scalp.
3. The electrode configuration was reset by "pressing at the same time with two fingers the CMS - T7 and DRL - T8 electrodes (they act like a reset switch when pressed)".[22]
4. Continue this process until all electrodes have impedance <5kΩ, with electrodes placed in the correct position (depicted in figures 3.3 and 3.4)

In some cases due to specific electrodes having poor impedance, saline solution was carefully applied to such electrodes whilst the device was on a participants head/scalp. Extra care was taken to not apply too much solution for the risk of creating salt bridges and connecting two electrodes together and leading to bias in the data.

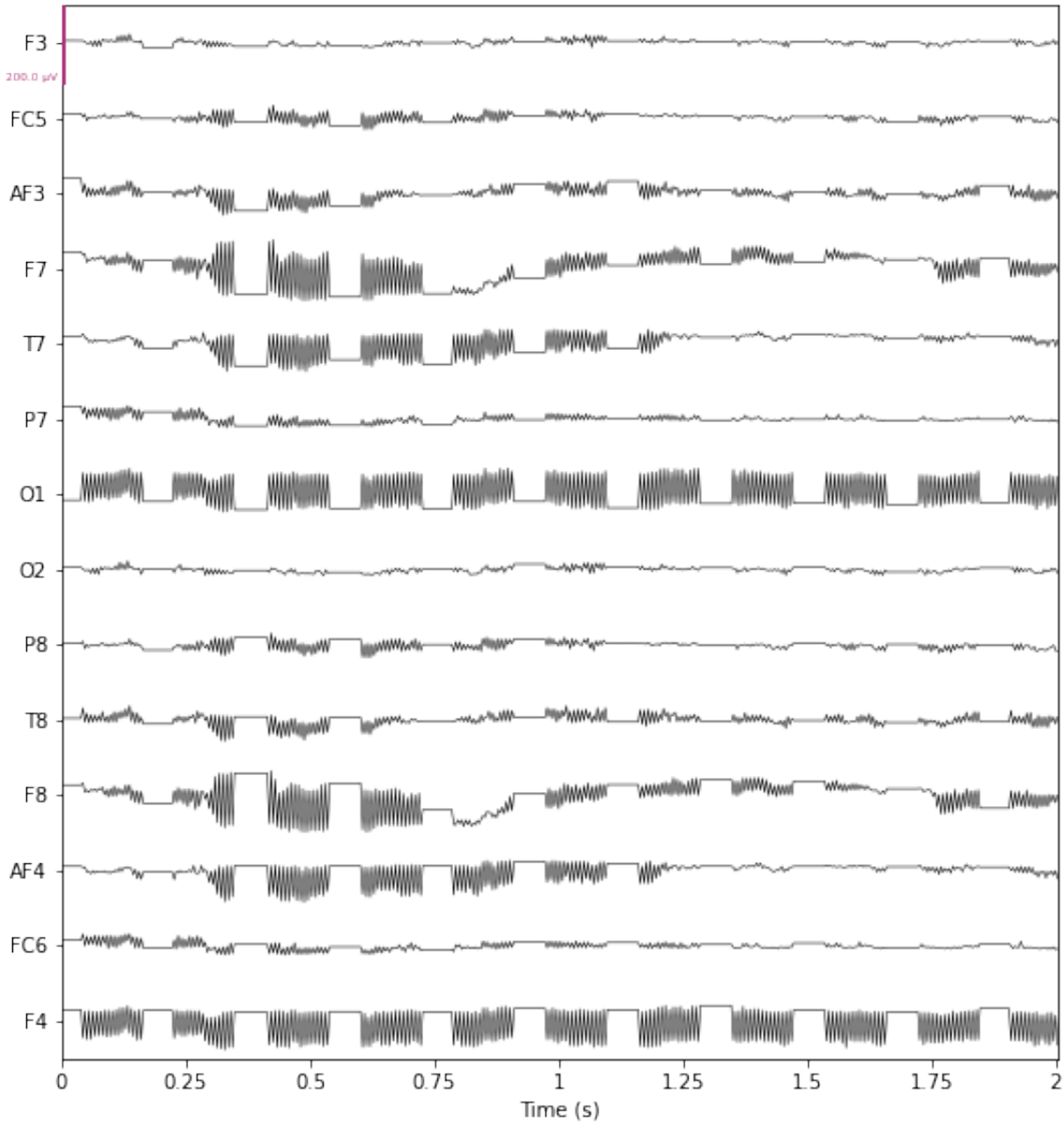

Figure 3.5: Plot of 0.3 seconds unprocessed speaking data across all 14 channels from a single participant, displaying the frozen values within their data/recordings.

Design flaws

After the recording of the data, some clear design flaws became evident:

1. **All modalities recorded in a single session** - Recording all data within a single session, means that all three modalities will have to either be processed together as one signal potentially leading to data leakage, or the data will need to be split into 3 datasets, one for each modality, leading to risk of bias from variation in processing (filtering and ICA)
2. **Frozen data** - Due to actions having to be taken in the data collection stage in order to combat device drift/lag, a problem coined as "frozen" or "dead" data was created where system and device clocks becoming out of synchronisation caused small and irregular periods in the data to be repeated, where the software used to program the paradigm (OpenVibe) performed corrections to re-synchronise, with some datasets having as much as a third of the dataset consisting of frozen data. This has a negative effect on the investigation not only due to the information lost, but it may also negatively impact filtering and ICA procedures due to the unorthodox/non-brain signal shaped data it creates. This problem can be observed in figure 3.5 as the flat-lining across all 14 channels which across a full recording appears in unequal time periods and repeats/occurs with what appears to be an irregular frequency/period. Throughout the 9 participants, 8 had roughly 30-34% frozen data, and one participant had roughly 15%
3. **Incomplete recordings** - Though all but one of the participants were able to complete at least 100/130 epochs total, only 4/9 participants were able to finish recordings fully, with all those having to stop being for reasons related to mental fatigue and feelings of nauseousness. Measures were put in place to try and manage fatigue, however it appears that these were not as successful as planned, although it is thought that the poor completion rate could be due to lack of incentive to complete (with participants not receiving any reward for completion).
4. **Discontinuous data.** Because of a mistake in the coding stage, the signal that was recorded for each person was not whole/complete, and not all parts of the signal were saved. This meant that the speaking data had to be concatenated for pre-processing to be peformed.

3.2 Pre-processing

The aim of the pre-processing carried out was to remove as much of the noise and unwanted signal as possible, to get closer to the signal of interest (increase the signal to noise ratio). In order to try and do this various different methods were deployed including epoch rejection, temporal filtering, and individual component analysis (ICA). Decisions taken within the processing steps, were heavily influenced by the design flaws outlined in subsection 3.1.1, and in particular design flaws 1 and 2. In regards to design flaw 2, current research into processing EEG data is not geared towards data with frozen values, therefore some creative and inventive solutions had to be thought up and discussed between collaborators, with the following solutions considered as potentially actionable:

1. Interpolating over frozen areas using live data either side of frozen areas.
2. Imputing values (for the data recorded in the **Stimulus delivery/Prompting** and

Think/Imagine stages) in place of those that were frozen by using average values for the time point within the 2 second window over all examplars and adding a noise coefficient to create variation in the data and a more orthodox shape.

3. Imputing values (for the data recorded in the **Stimulus delivery/Prompting** and **Think/Imagine** stages) in place of those that were frozen by using a Gaussian Process model trained to predict values based on time point within the 2 second window to the same effect as in 3.

In the case of all three proposed solutions, none introduced any new information into the data, and only sought to fix the "unorthodox" shape. Imputing values as in solution 2 and 3 will lead to information being shared between samples, causing data leakage between train and test sets, hence were not deemed viable. Upon research and attempting to apply solution 1, it was found to be far too computationally complex to perform given available resources (RAM storage), and was deemed unviable also. It was decided that it would be neither time efficient nor effective to attempt to directly alter the frozen areas. Hence the investigation continued without alteration of the frozen/dead data.

3.2.1 Pipeline structure

The process of temporal filtering involves attenuating frequencies that aren't of interest, attempting to improve the signal to noise ratio. This relies on the appliaction of fourier transforms, and removing sinusoidal waves of specific frequencies which are deemed to be not related to the signal of interest. The presence of the frozen values within the data is highly disruptive, it breaks up the signal, disrupting its sinusoidal nature and prevents waves within the live data from being fully expressed, and limits the success of temporal filtering. In theory, the flat patches created by the frozen data could be described as sinusoidal waves with frequency 0Hz, it is not mathematically possible to remove waves of this frequency however, and the data lost over these periods is not retrievable. This disruption appears in fact to cause temporal filtering to have a negative impact on the data creating artifacts potentially lowering the signal to noise ratio. The impact of applying band pass filtering on two consecutive epochs can be seen through figures 3.6 and 3.7 (For reference an epoch describes a single 2 second speaking observation/examplar in this work). In comparison to the variation in the live data areas, large fluctuations exist in the filtered data at the start and end of where each frozen area was, these dramatic changes will likely hinder the decoding of speech. As a consequence, temporal filtering was kept to a minimum, although for reasons elaborated on in 3.2.3, some temporal filtering was still required. It can also be seen in figures 3.6 and 3.7 clear evidence of information sharing between epochs where in the unfiltered data the discontinuity between epochs (design flaw 4) is obvious, and in the filtered data, epochs appear to have been levelled out and look continuous, this is likely due to the rapid change in value being removed as theoretically it could be interpreted as part of a sinusoidal wave with frequency above 70Hz. This information sharing between epochs may be solved by filtering each epoch independently (iteratively), or by setting filter lengths such that only values within the same epoch are used in filtering, the method for this though could not be found within the literature and was not applied.

As each mode of stimuli was randomly assigned throughout the recording phase (for reasons mentioned in 3.1.1), a decision had to be made on whether it would be better to perform the filtering and ICA steps on the whole dataset together, or whether audio, visual and text data should be processed separately. If the data was to be processed whole with audio, text and

visual classes together, there is a risk of leakage of information between audio, visual and text data, leading to each of the three becoming more similar and reducing the likelihood of picking up on inter modality performance differences, and hindering the power of statistical tests for hypotheses 1, 2, and 3. On the other hand the risk of processing each modality individually is the introduction of bias in the results based on the success of preprocessing steps, in particular the independant component analysis (ICA) reducing the validity of the results, potentially increasing the type II error rate. The data leakage incurred by performing processing on the data as a whole, without splitting by train and test sets or by modality would ordinarily not be seen as problematic for data without frozen values and without discontinuity. However given the observed impact across separate windows of applying temporal filtering, the chance that leakage between modalities would be problematic appears high, hence it was decided to perform processing separately on audio, visual and text data.

Unlike recommended by Panachakel et al (2021) [19] when using visual cues to prompt timing and rhythm for performance of speech imagery, channels over the occipital lobe were not discarded. Instead ICA was used as a means to identify potential artifacts arising from such visual cues. Very few of the datasets contained individual components with source around the occipital lobe (O1, O2), and within datasets containing such components, these did not occur wihtin all three datasets, hence were not removed. On the other hand, during the ICA, eye components thought to be the ocular artifacts from following the marker on screen were present within audio, visual and text datasets, giving confidence that the intentionally created artifacts had successfully been removed (to an adequate degree).

Figure 3.9 shows a graphical representation of the processing pipeline deployed. Epoch rejection and class equalisation gets applied first, followed by temporal filtering and ICA on audio, visual and text data (separetely), finshing with feature extraction and classification/machine learning. 5-fold cross validation was performed with the same folds opted for in 4-way classification, binary classification of the vowel pairs, and binary classification of the syllable pairs. This was done to make results across the three forms of classification comparable within results relating to each mode of stimuli. This allows for interesting/unexpected results for a particular model to be compared to other models trained on the same fold, ruling out a "good fold" as a reason for good performance for example.

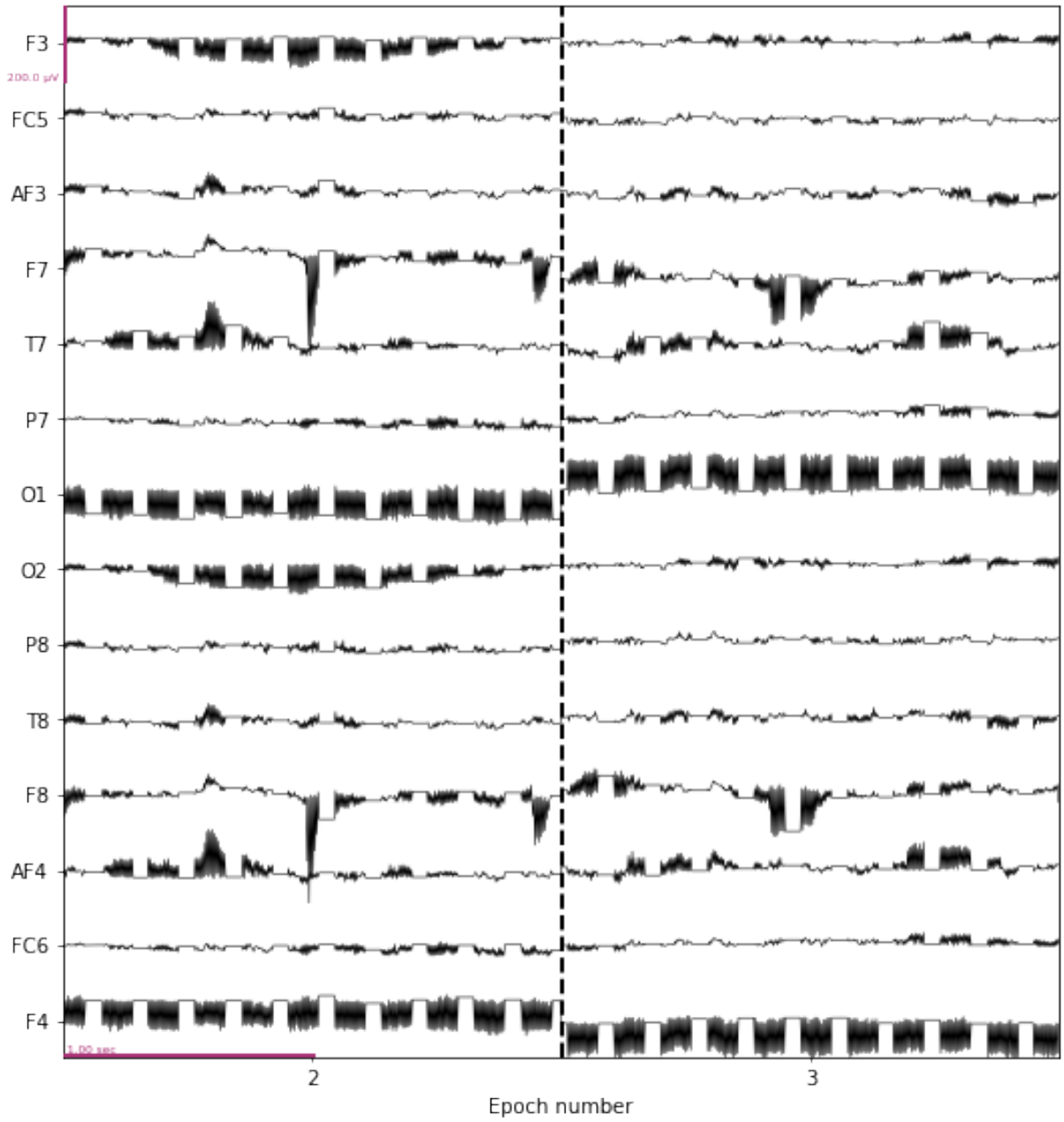

Figure 3.6: Plot of 2 epochs of unfiltered data (consecutive in the dataset)

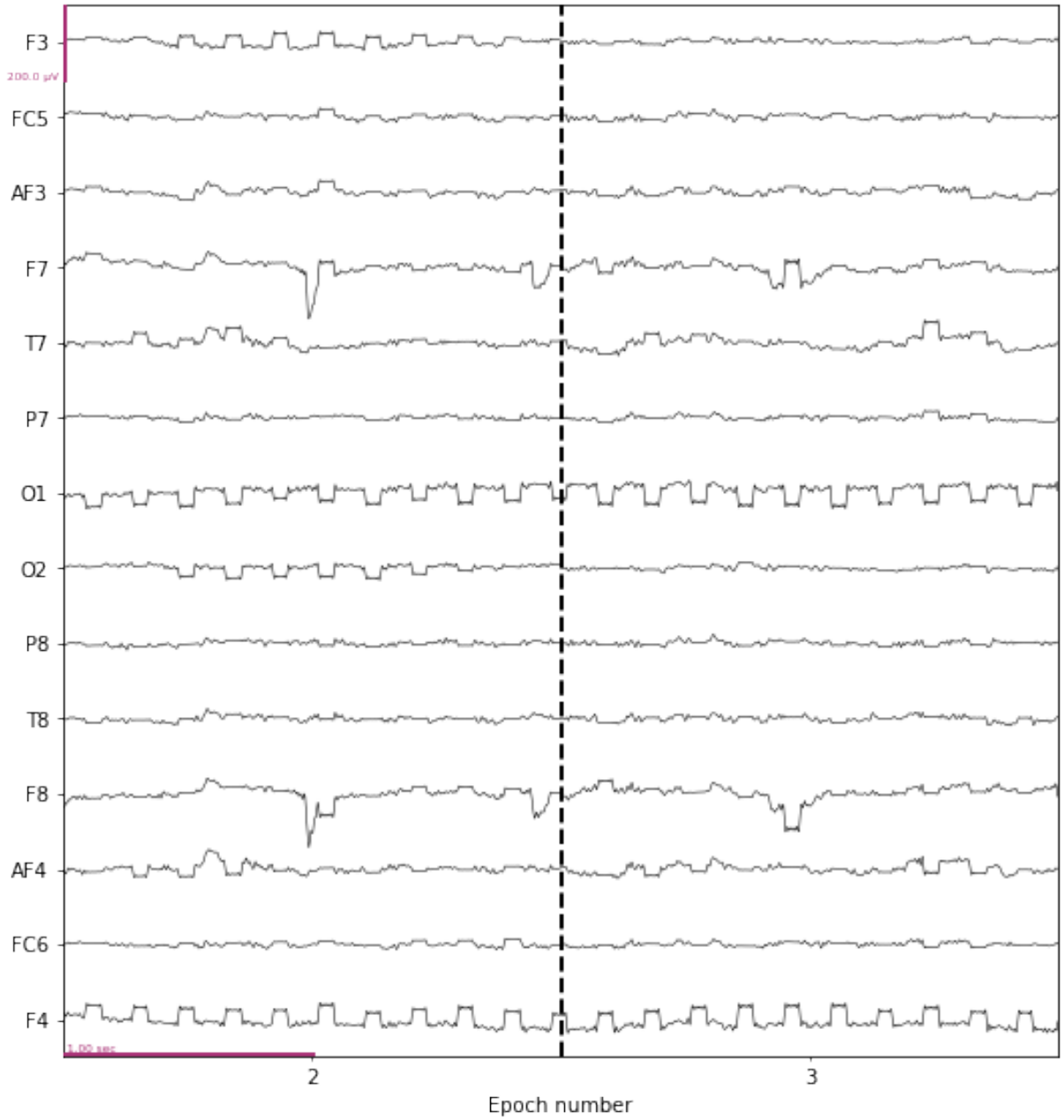

Figure 3.7: Plot of 2 epochs of data band-pass filtered between 1-70Hz (consecutive in the dataset)

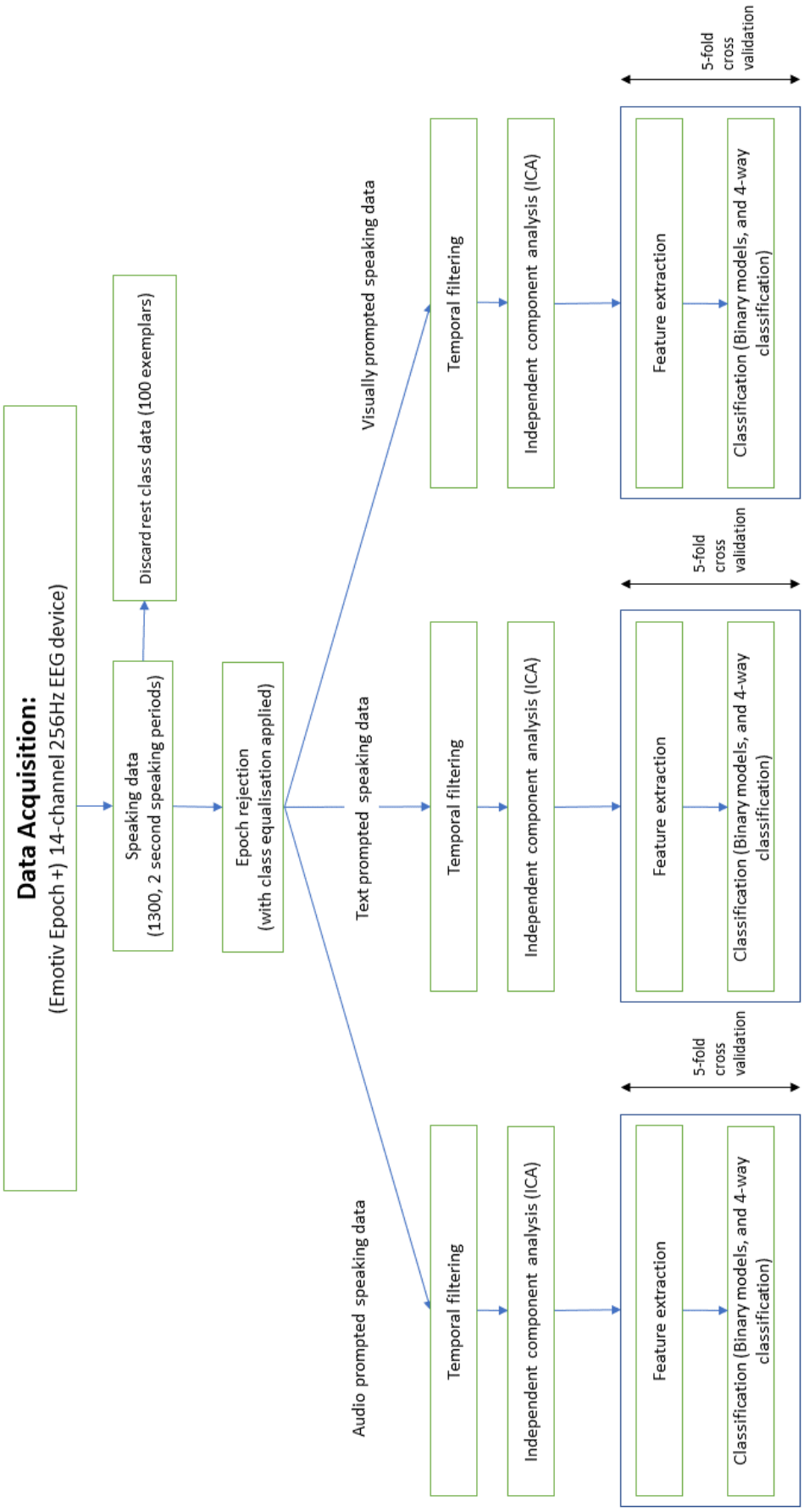

Figure 3.8: Graphical representation of the order in the processing pipeline used in this research from Data Acquisition through the preprocessing steps to feature extraction and classification

3.2.2 Epoch rejection

An Epoch rejection procedure was employed based on peak-to-peak signal amplitude. The optimal threshold in terms of peak-to-peak amplitude is dependent on factors relevant to the research (Choice of recording device, recording structure and hypotheses etc), hence there is no universally agreed optimal value. A process of trial and error was used to find a balance between data quality and sample size, and throughout all of the participants data, rejecting epochs with at least one channel having peak-to-peak amplitude greater than $2000\mu V$ was deemed optimal, with an average of 13.1 epochs dropped per participant. Along with design flaw 3, epoch rejection led to imbalances in the number of classes in the data. To prevent imbalanced classification from taking place, classes in the data were dropped such that all classes were observed an equal number of times in the data. This was done using a time based method that ensures remaining epochs occur as close in time as possible, working on the idea that "if there happened to be some time-varying (like on the scale of minutes) noise characteristics during a recording, they could be compensated for (to some extent) in the equalisation process" [11].

3.2.3 Filtering

Disruption to the filtering process is well depicted by figure 3.9, where certain frequencies of wave (that would be present without the data loss) are now not represented within the data, this can be seen as the multiple dips in the amplitude throughout the frequency spectrum. Therefore not only do the frozen values hinder the ability to repair artifacts but they also make it more difficult to identify them. The removal of slow drifts from the data, with high pass filtering around 1-2 Hz has been found to improve the effectiveness of independent component analysis (ICA) [9]. Hence for all participants, a high pass filter with cutoff frequency 1Hz was applied. Because of the negative impacts observed by applying filtering to the data, the high pass filtered data was only used during the ICA algorithm, and regardless of whether any components were removed from the data, it was the unprocessed data (not high pass filtered) that was used in the subsequent feature extraction and machine learning/classification stages. This comes from the recommendation of of Viola et al (2010) [24] for the problem of wanting to perform artifact removal without having to use high pass filtered data within the subsequent feature extraction and machine learning stages. Transformation of the data given the ICA weights is calculated on the high pass filtered data, and the transformation then applied on the unfiltered data.

3.2.4 Independent component analysis (ICA)

Because of the number of problems with the data recorded and how sensitive the data is to bias from applying processing, a manual inspection of the event related potentials, frequency spectrum, and brain regions/channels for each individual component was conducted with the use of a fastICA algorithm [11], rather than an automatic procedure (MARA [17], etc). Within the ICA procedure, the only type of component that got removed were eye components. Figure 3.10 shows a good example of an eye component found where the source can be seen in the scalp topography to be around both left and right eyes, with power concentrated around the lower frequencies (<5Hz), the likely frequency in which this would occur given not many people could move their eyes faster than that. The plot of event related potentials shows this component occuring somewhat periodically throughout the data, and it was thought that this and similar components removed were mainly from the intentionally created ocular artifacts

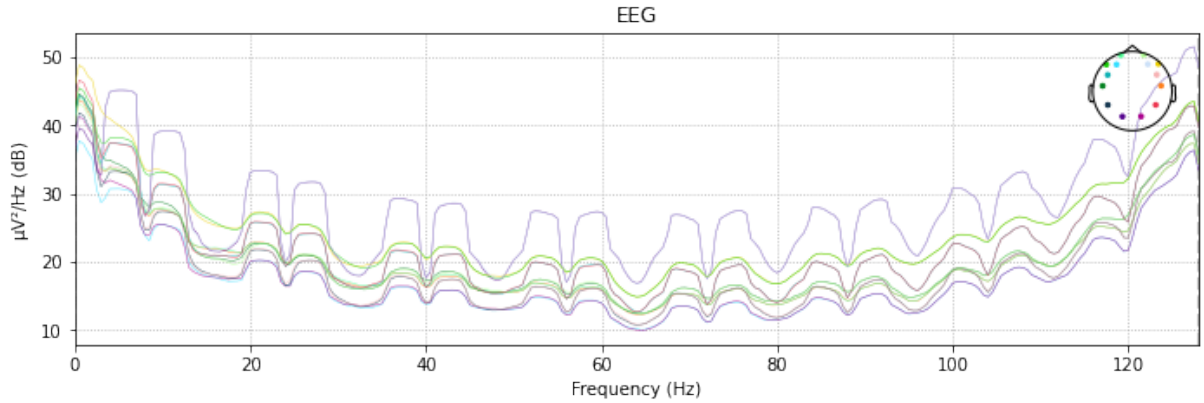

Figure 3.9: Plot of power spectral density throughout frequencies 0 to 128 Hz by channel, displaying at least 16 observable spikes, and the large of amount of contamination within the signals

of following the moving on screen marker. Not all eye components removed looked exactly like that in figure 3.10 though, and several components demonstrated sources around a single eye as in figures 3.12 and 3.11 with the same characteristic of concentration around lower frequencies and periodic occurrence throughout the data. It was considered whether it would only be right to remove a component containing a source over a single eye if there was an accompanying component with source from the other eye within the same data, however due to no knowledge of the ocular health of participants, all eye components were removed.

3.3 Feature extraction

With the regions of activation during inner and imagined speech said to ambiguous, its likely that a highly complex process takes place in the brain between different regions, hence it was decided that due to simultaneous feature extraction methods said to better at capturing the reationship between channels [13], this would be opted for over single channel feature extraction methods. The method proposed by Nguyen et al (2017) [18] was used for feature extraction in this research, whereby tangent vectors are calculated from channel cross-covariance (CCV) matrices. The procedure for calculation of features for training and testing data is presented in figure 3.13, where mean values are calculated from covariance matrices in the training data, and used for normalisation on both the training and testing data, where tangent vectors are then calculated and used as features. Full explanation with mathematical notation is provided via Nguyen et al (2017) [18].

3.4 Application of Machine Learning models

For investigation into the hypotheses 1, 2 and 3, 4-way classification between Ambulance, Hospital, Clock and Lamp was performed, as well as binary classification between the vowel pairs (Ambulance and Lamp, Clock and Hospital), and also binary classification between syllable pairs (Ambulance and Hospital, Clock and Lamp). It would not be time efficient nor wise to train a large number of different types of algorithm for each case. Not only would this lead to extensive/unmanageable runtimes, but also may lead to what is known as "p-hacking" or "data dredging", where as the number of outcome measures increases the

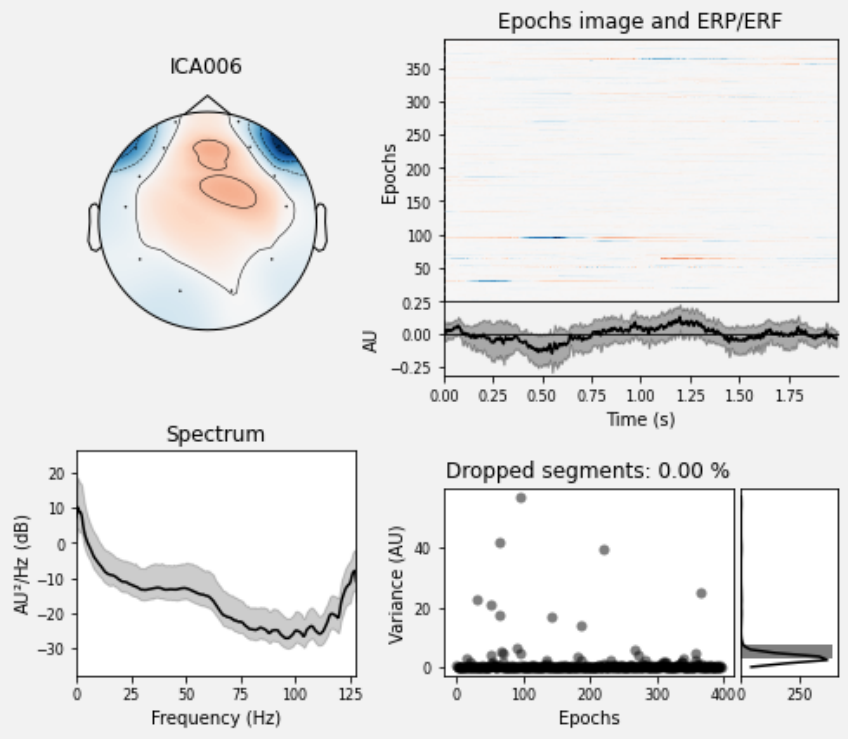

Figure 3.10: Example for a component judged to be an eye component, with plot of frequency spectrum, topographical map of brain, and event related potential's thorughout each epoch

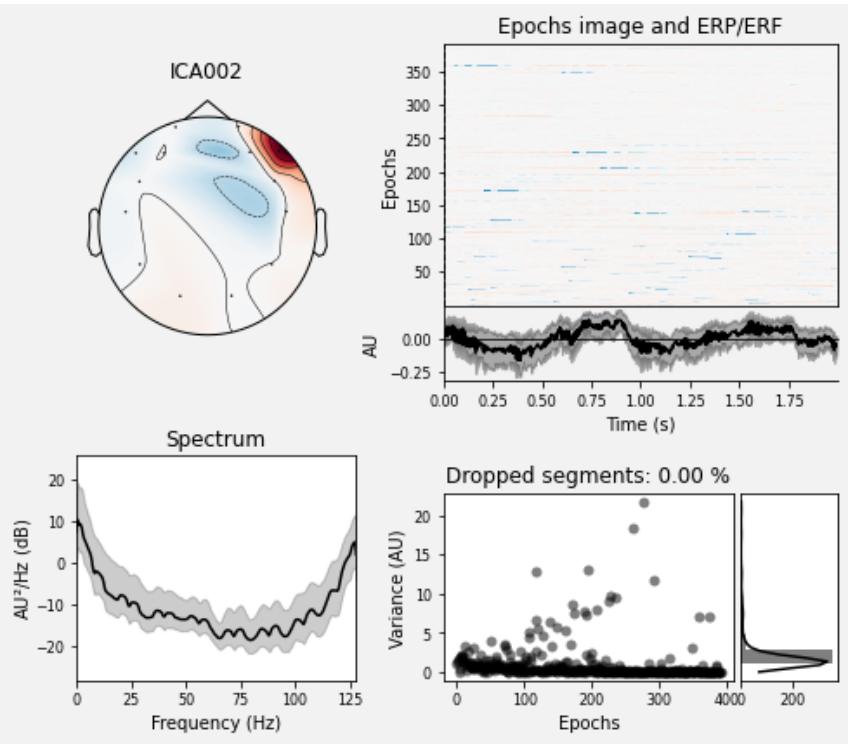

Figure 3.11: Example for a component judged to be an eye component for the right eye, with plot of frequency spectrum, topographical map of brain, and event related potential's thorughout each epoch

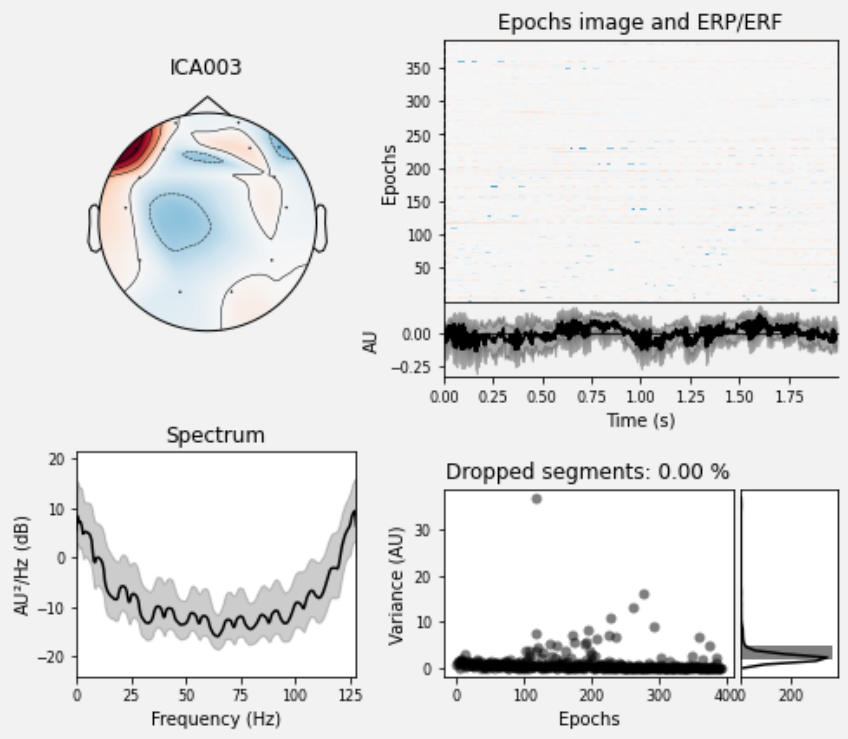

Figure 3.12: Example for a component judged to be an eye component for the left eye, with plot of frequency spectrum, topographical map of brain, and event related potential's throughout each epoch

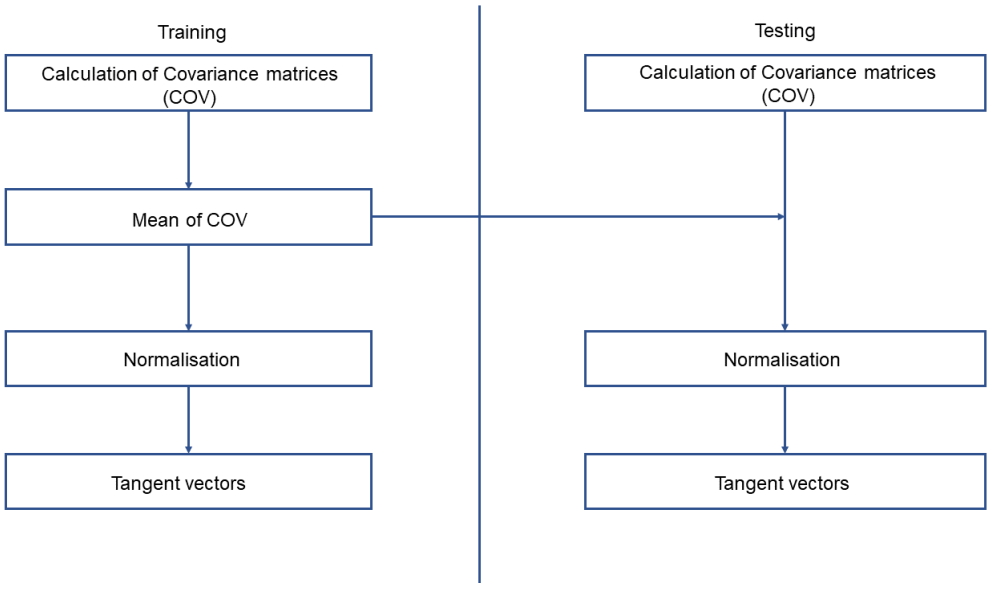

Figure 3.13: Graph of the procedure for calculating tangent vectors of Cross channel covariance matrices on training and testing data. Adapted from the paper by Nguyen et al (2017) [18]

probability of false positives increases. In other words testing a large number of models will increase the chance of seeing a significant difference in classification accuracy between models trained on data for each of the modes of stimuli. Hence the number of different machine learning models used was kept low (2). The machine learning algorithms used were selected using information compiled in the work by Panachakel et al (2021) [19] on the frequent best performing algorithms within the field of inagined speech. Deep learning based methods (Deep neural networks, Convolutional neural networks, Recurrent neural networks, etc) were by far the algorithm found to be "best classifier" most often, with Random Forests, Support vector machines and nearest neighbour also ranking highly. For variation, one deep learning, and one non-deep learning based method were chosen from this group, and factoring in the need for algorithms that do not require extensive training and optimisation time due to the large number of models per participant (18), the random forest algorithm was deemed more desirable than both support vector machines (due to the need for choosing/testing different kernels) and nearest neighbours (due to large training times).

The same structures and training/tuning procedures were applied in the case of all three types of classification, allowing for both comparison in results and also simplicity in application. Once again due to time constraints arising from the large number of total models that will need to be trained throughout all participants (162), an automated procedure for training and tuning all models had to be created for both the RF and DNN models. Within the RF selected, trees used gini impurity for calculating optimal splits with number of features set as the square root of the number of features in the data (10 rounding down). For hyper-parameter tuning, a random search was performed throughout the number of estimators/trees in the ensemble, maximum tree depth, minimum samples per tree, and minimum samples per split with cross validation using train set data used to gage a best set of hyperparameters. The structure of the deep neural network can be seen in figure 3.14, and throughout models, number of epochs and batch size were set to 500 and 4 respectively, with Adam optimiser (learning rate of 0.001) employed for optimisation with a categorical cross entropy loss function.

3.5 Performance metrics and hypothesis testing

In order to perform statistical tests to test against the null hypotheses set out, a One-way ANOVA (analysis of variance) test statistic was used. One-way ANOVA tests for unequal variance between 3 or more independant samples. In this context it tests for equal variance in the mean 5-fold classification accuracy for models across audio, text and visual modes of stimuli. This provides a measure of whether mode of stimulus has an **effect** on classification accuracy, hence the ability to decode inner and/or imagined speech. In the case of significant differences in variances between groups being found, post-hoc analyses could be performed to determine a potential superior mode of stimulus, however this was not required. In order to test whether a individual model performed significantly above chance level, a test of whether the probability of the observed 5-fold classification accuracies being drawn from a normal distribution with mean of the random chance level (50% in 4-way, 25% in binary classification) was less than 0.05 was conducted by calculating a z-statistic and comparing to the critical value for 5% significance on a one way test (1.64). Models with z-statistic above the critical value were deemed to be performing significantly above the chance level.

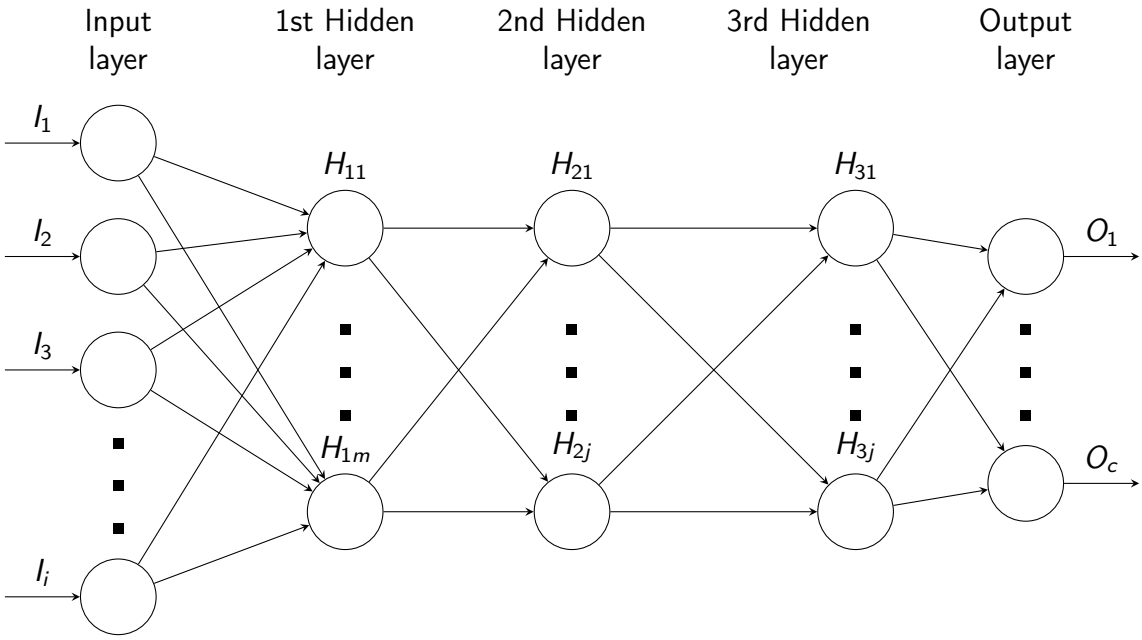

Figure 3.14: Graph of the deep neural network used for 4-way classification, binary classification of vowel pairs, and binary classification of monosyllabic vs trisyllabic words (where i=105, m=50, j=100, c=4 in 4-way classification, and c=2 in binary classification). It must be noted that linear activation functions were used for hidden layers with dropout regularisation applied to the third hidden layer with a rate of 0.2, and a sigmoid activation function was used on the output layer

Chapter 4

Results

This section contains the results in relation to both the evoked response in stimulus and speaking stages, as well as classification accuracy obtained by machine learning models. For information purposes, both the evoked response to the presentation of Audio, visual and text stimuli and the average response from the start of each speaking epoch have been calculated by taking the average across epochs. It must be noted that audio stimuli gets presented starting from 0.5s onwards within the stimulus stage of recording, whereas text and visual stimuli are presented from the beginning. Evoked responses are calculated from 0s onwards in all three cases (with no baseline correction applied).

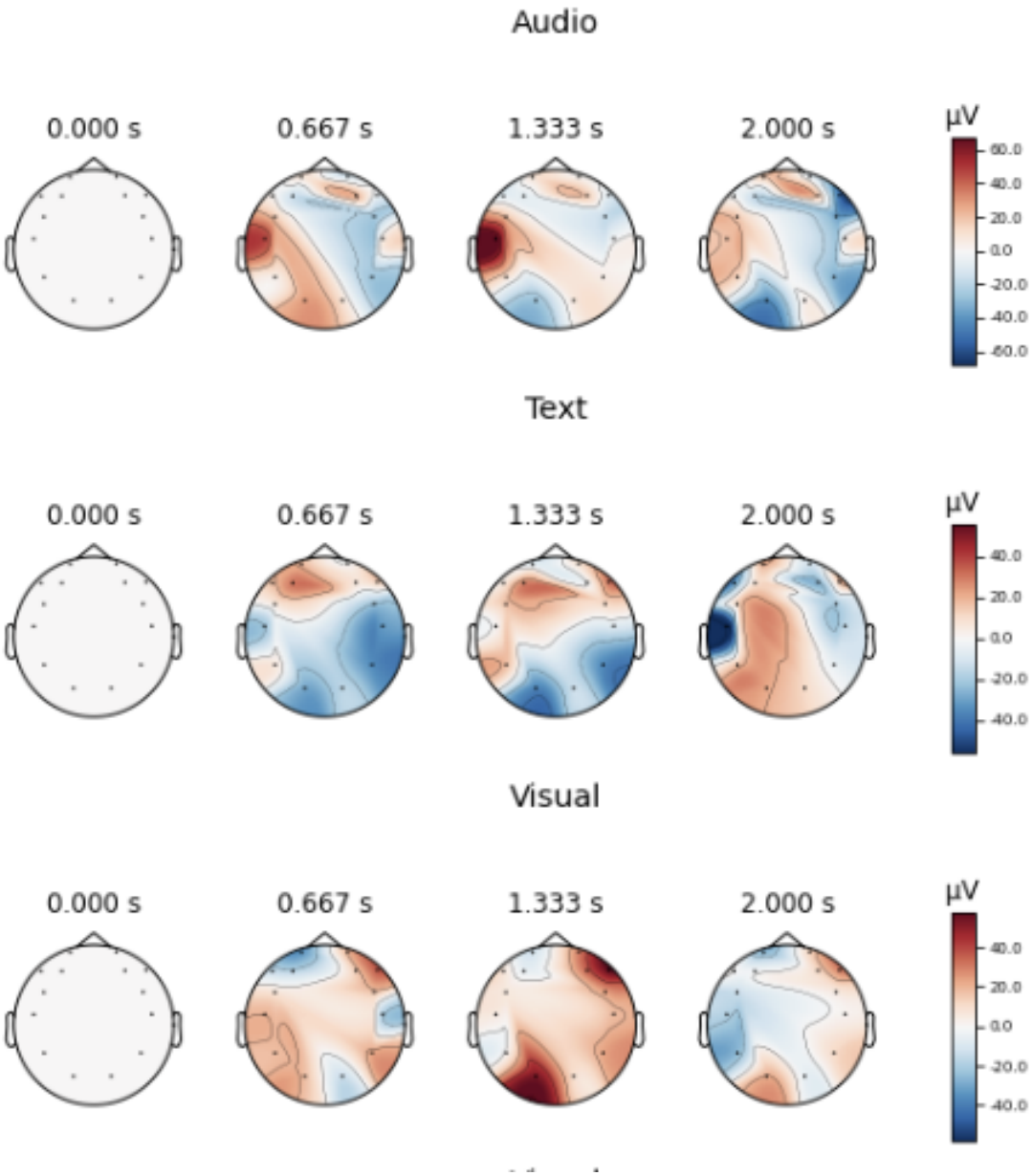

Figure 4.1: Topographic map of evoked response throughout the stimulus stage for participant 1, displaying behaviour in line with theory in the literature.

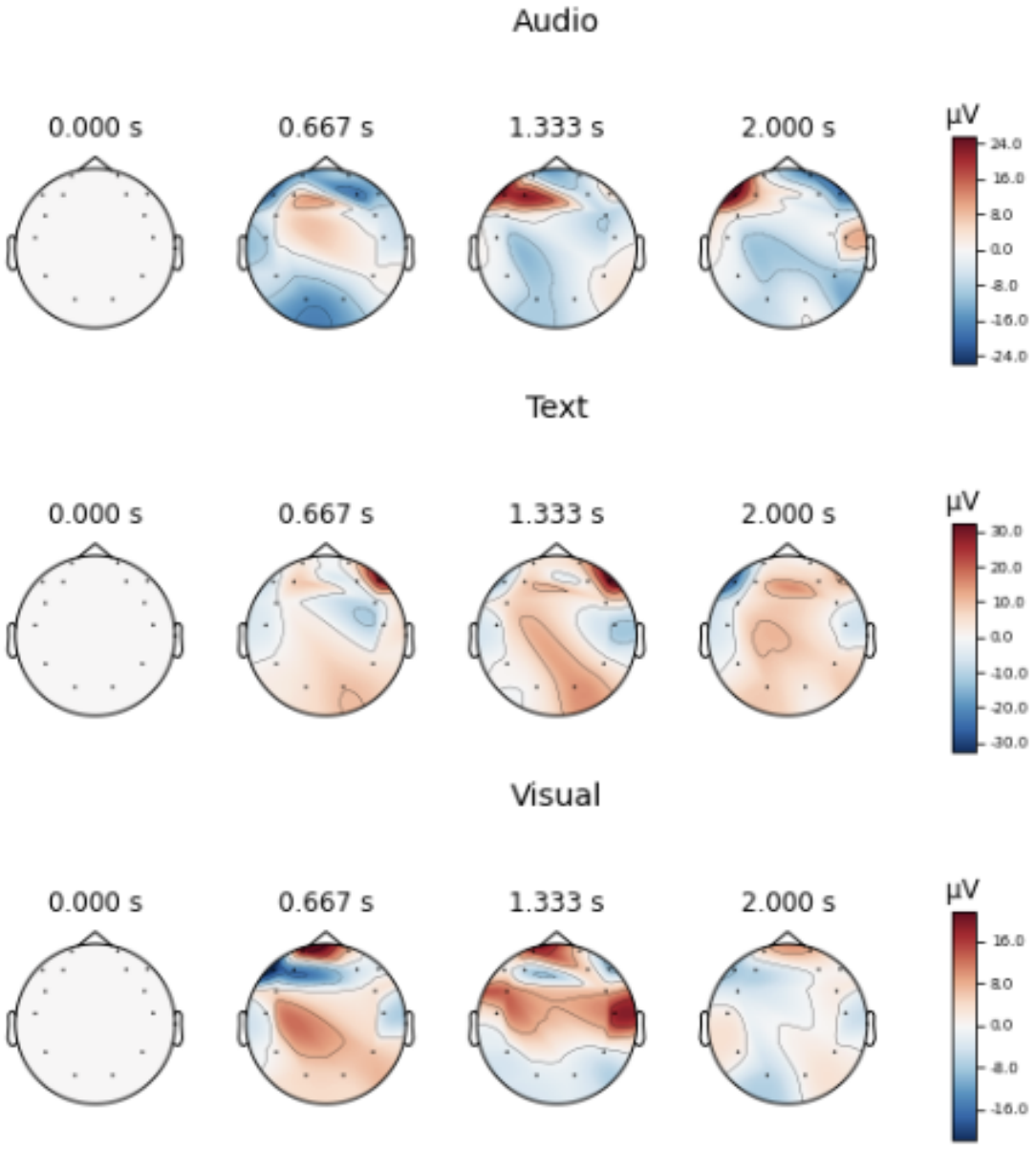

Figure 4.2: Topographic map of evoked response throughout the stimulus stage for participant 3, displaying behaviour not in line with theory in the literature.

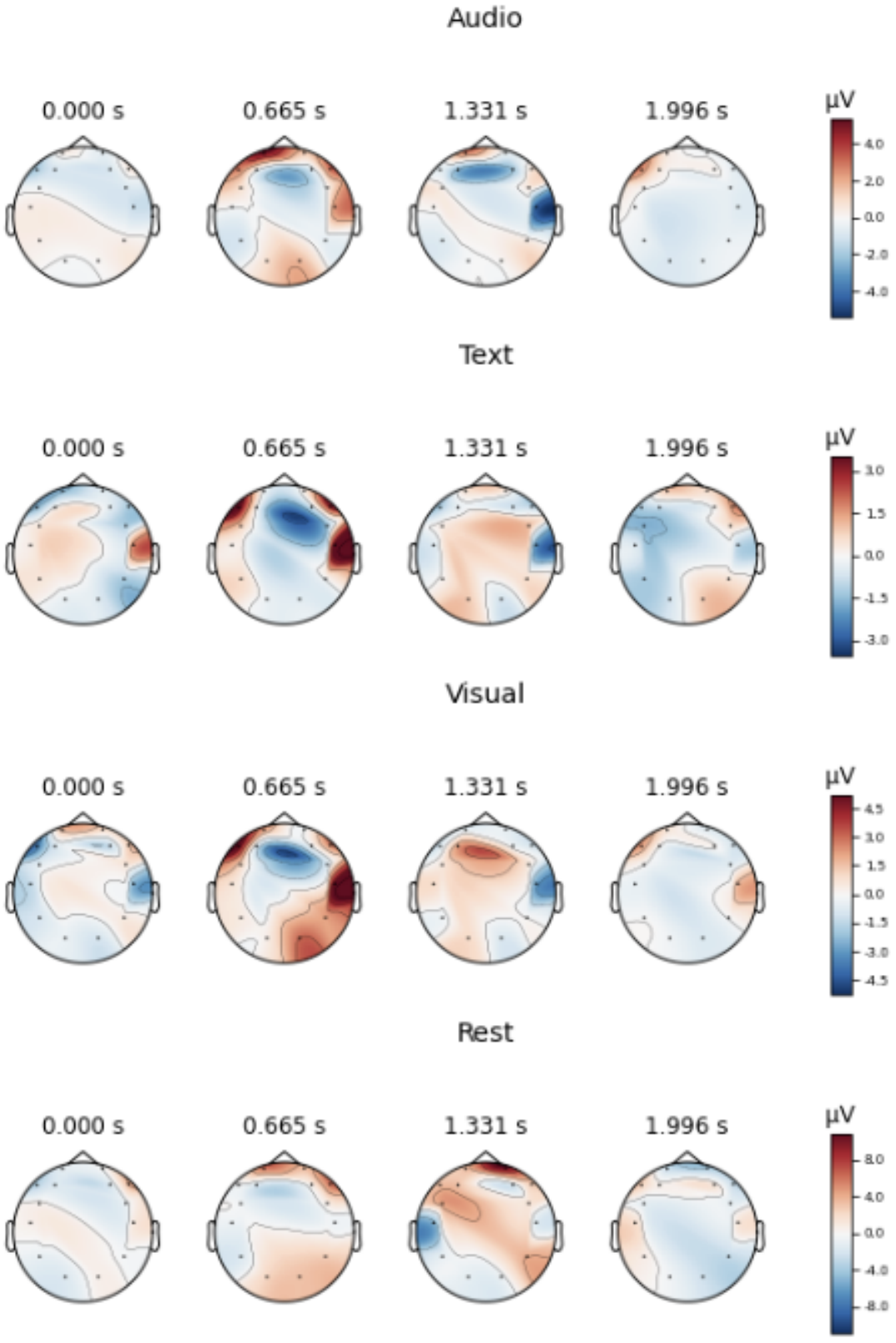

Figure 4.3: Topographic map of evoked response throughout the performance of imagined speech and during rest class (participant 1).

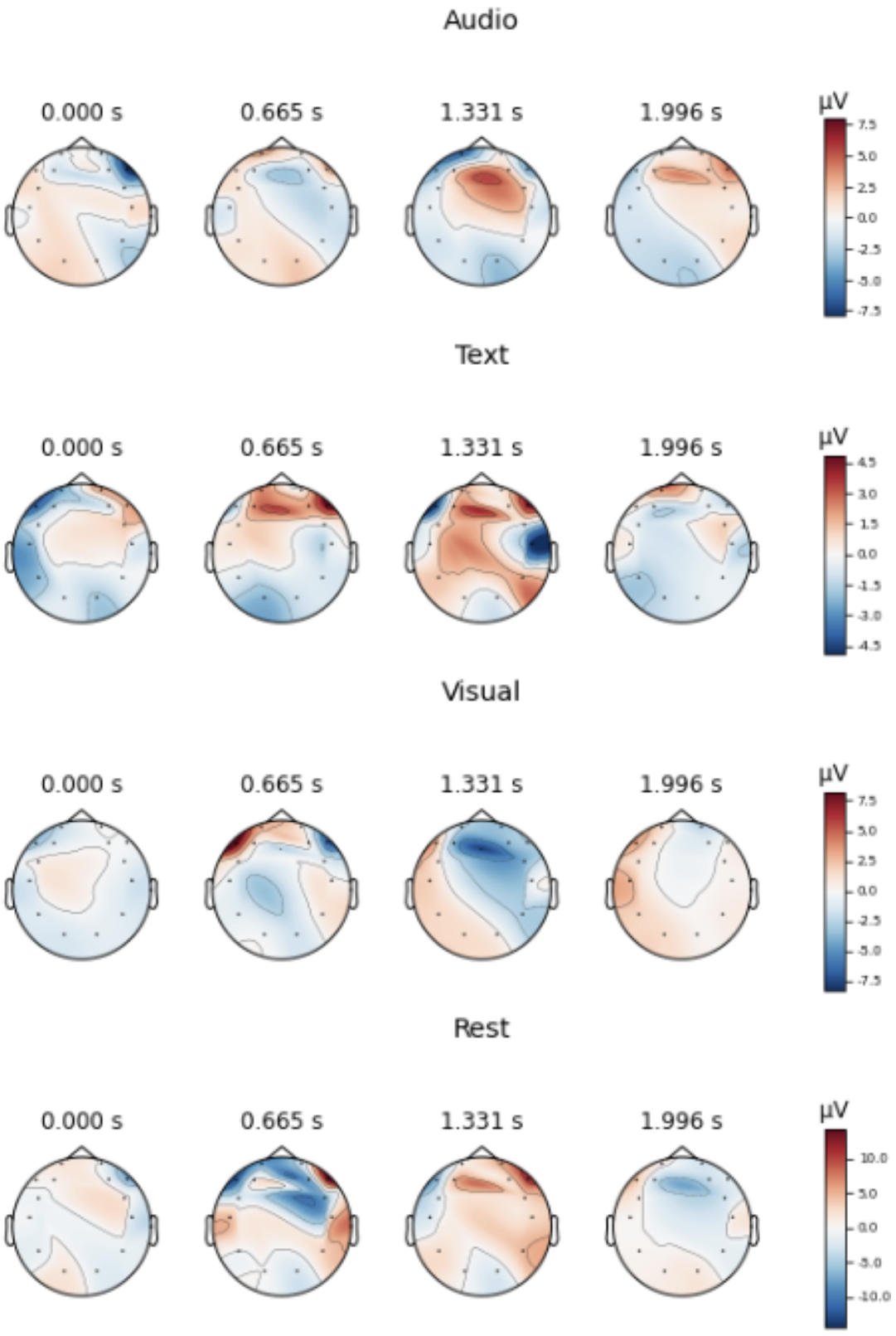

Figure 4.4: Topographic map of evoked response throughout the performance of inner speech and during rest class (participant 4).

Table S1: 5-fold cross validation classification accuracy to 1 d.p (mean ± standard deviation %) in the case of 4-way classification of Ambulance, Hospital, Lamp, Clock

Participant	DNN accuracy (Mean ± Std %)			RF accuracy (Mean ± Std %)		
	Audio	Text	Visual	Audio	Text	Visual
Inner						
3	**33.0 ± 3.8**	26.0 ± 3.2	23.2 ± 3.3	**34.9 ± 4.2**	**34.1 ± 2.8**	28.5 ± 3.6
4	25.5 ± 3.0	25.5 ± 4.7	22.2 ± 2.9	26.3 ± 5.1	**30.3 ± 2.6**	20.7 ± 3.8
5	25.0 ± 8.4	20.4 ± 5.0	20.4 ± 3.7	23.1 ± 10.0	14.5 ± 5.5	24.3 ± 6.0
8	19.8 ± 6.6	22.9 ± 3.8	28.1 ± 8.8	23.4 ± 5.9	26.0 ± 3.4	**31.2 ± 1.7**
9	22.9 ± 3.5	22.1 ± 3.7	32.9 ± 5.5	18.6 ± 4.9	18.2 ± 3.8	**38.9 ± 5.6**
Mean	24.1 ± 4.5	24.7 ± 2.6	25.2 ± 4.0	25.2 ± 6.0	24.6 ± 8.1	28.8 ± 7.0
Imagined						
1	23.1 ± 3.8	29.0 ± 3.6	24.2 ± 2.5	25.2 ± 1.6	22.8 ± 2.6	22.9 ± 3.5
2	26.6 ± 4.0	25.5 ± 3.5	25.8 ± 4.6	28.3 ± 2.7	27.2 ± 4.2	25.2 ± 6.8
6	24.3 ± 3.2	23.9 ± 7.2	21.3 ± 1.9	20.6 ± 5.5	19.5 ± 4.3	27.9 ± 4.8
7	16.9 ± 2.4	26.6 ± 4.7	28.2 ± 3.8	20.7 ± 2.7	26.2 ± 2.0	31.1 ± 6.1
Mean	22.7 ± 4.1	26.2 ± 2.2	24.9 ± 2.9	23.7 ± 3.7	24.0 ± 3.5	26.8 ± 3.6

Bold implies result is significantly above the chance level at the 5% significance level.

Table S2: 5-fold cross validation classification accuracy to 1 d.p (mean ± standard deviation %) in the case of binary classification of syllable pairs (Ambulance and Hospital versus Lamp and Clock)

Participant	DNN accuracy (Mean ± Std %)			RF accuracy (Mean ± Std %)		
	Audio	Text	Visual	Audio	Text	Visual
Inner						
3	56.6 ± 5.7	49.0 ± 6.2	48.5 ± 5.1	**62.3 ± 2.4**	**56.6 ± 4.0**	45.2 ± 3.0
4	49.5 ± 1.4	**54.0 ± 2.0**	44.4 ± 5.9	49.5 ± 6.5	54.8 ± 6.9	43.4 ± 2.8
5	40.2 ± 6.5	51.0 ± 3.3	37.5 ± 5.1	40.2 ± 5.8	50.6 ± 6.5	47.4 ± 8.0
8	44.7 ± 8.8	58.3 ± 7.9	55.2 ± 7.2	53.1 ± 11.1	**61.4 ± 6.9**	60.9 ± 7.9
9	51.8 ± 7.7	42.9 ± 10.0	**61.1 ± 4.3**	44.6 ± 7.1	39.4 ± 7.5	**63.2 ± 2.7**
Mean	48.6 ± 6.3	51.0 ± 5.8	49.3 ± 9.2	50.0 ± 8.6	52.6 ± 3.1	52.0 ± 9.3
Imagined						
1	47.9 ± 2.5	54.5 ± 3.3	49.7 ± 4.5	48.7 ± 5.1	53.7 ± 2.0	49.5 ± 2.1
2	51.3 ± 3.6	52.3 ± 3.5	44.4 ± 5.1	53.3 ± 2.1	51.3 ± 3.7	49.2 ± 6.2
6	53.6 ± 10.0	46.3 ± 6.2	44.8 ± 5.8	47.4 ± 6.7	46.4 ± 8.4	52.9 ± 2.0
7	43.2 ± 2.7	44.2 ± 2.7	56.8 ± 4.9	44.8 ± 3.8	49.4 ± 4.3	62.7 ± 8.0
Mean	49.0 ± 4.5	49.3 ± 4.9	48.9 ± 5.8	48.5 ± 3.6	50.2 ± 3.1	53.6 ± 6.3

Bold implies result is significantly above the chance level at the 5% significance level.

Table S3: 5-fold cross validation classification accuracy to 1 d.p (mean ± standard deviation %) in the case of binary classification of vowel pairs (Ambulance and Lamp versus Hospital and Clock)

Participant	DNN accuracy (Mean ± Std %)			RF accuracy (Mean ± Std %)		
	Audio	Text	Visual	Audio	Text	Visual
Inner						
3	**59.6 ± 2.7**	55.8 ± 9.0	47.7 ± 3.8	**57.6 ± 1.5**	57.5 ± 7.1	52.1 ± 7.4
4	47.3 ± 6.4	48.2 ± 3.9	46.7 ± 5.0	48.2 ± 5.5	**61.3 ± 4.4**	50.0 ± 4.2
5	56.6 ± 4.3	36.8 ± 4.7	50.1 ± 7.8	57.2 ± 4.5	38.2 ± 3.1	51.3 ± 5.1
8	52.0 ± 8.2	48.9 ± 7.4	53.6 ± 9.0	55.2 ± 5.7	50.0 ± 6.1	45.8 ± 6.2
9	51.4 ± 3.8	47.8 ± 9.5	53.2 ± 4.1	41.7 ± 3.5	38.9 ± 5.7	**57.5 ± 3.1**
Mean	53.4 ± 4.8	47.5 ± 6.8	50.3 ± 3.1	52.0 ± 6.8	49.2 ± 10.5	51.3 ± 4.2
Imagined						
1	48.4 ± 6.8	55.0 ± 3.5	47.6 ± 1.1	52.4 ± 3.1	48.6 ± 3.3	50.3 ± 4.4
2	51.8 ± 4.8	**58.1 ± 4.2**	54.9 ± 3.4	52.3 ± 2.8	**61.7 ± 1.9**	52.1 ± 5.4
6	43.0 ± 6.9	46.7 ± 3.6	47.4 ± 2.8	47.1 ± 5.8	47.3 ± 5.7	43.7 ± 3.0
7	49.7 ± 2.6	52.3 ± 2.5	50.3 ± 5.4	50.0 ± 3.3	55.2 ± 5.4	55.9 ± 6.5
Mean	53.1 ± 4.9	48.2 ± 3.7	50.1 ± 3.5	52.0 ± 6.8	53.2 ± 6.6	50.5 ± 5.1

Bold implies result is significantly above the chance level at the 5% significance level.

Table S4: p-values corresponding to a One-way ANOVA test between mean classification accuracy between models trained on the audio, Visual and text data. In the case of each type of classification and each algorithm for inner and imagined speech.

Case	DNN	RF
4-way - Inner	0.72	0.63
4-way - Imagined	0.33	0.43
Binary (vowel pairs) - Inner	0.86	0.88
Binary (vowel pairs) - Imagined	0.99	0.32
Binary (syllable pairs) - Inner	0.23	0.82
Binary (syllable pairs) - Imagined	0.29	0.68

Table S5: Table displaying positive correlation between number of samples per class (n) after processing versus the number of models performing significantly above chance level (no. Sig). (Pearsons R = 0.44)

Participant	n	no. Sig
1	94	0
2	98	2
3	99	7
4	99	3
5	38	0
6	68	0
7	77	0
8	48	2
9	70	3

Chapter 5

Discussion

5.1 Evoked/average responses

Both the evoked response to the presentation of Audio, visual and text stimuli and the average response from the start of each speaking epoch has been calculated by averaging over all epochs, in the case of each participant. In this section some of the interesting and informative results have been picked out and discussed, with particular interest paid to areas of activation within the brain. It must be noted that in data collection the audio stimuli gets presented starting from 0.5s onwards within the stimulus stage, whereas text and visual stimuli are presented from the beginning (0s), however the evoked response is calculated from 0s onwards in all three cases (with no baseline correction applied), and this must be considered in the interpretation of the stimulus stage evoked response to audio stimuli. Special attention is placed on how the percentage of frozen values influences the observed responses, via a comparison of one participant with roughly 15% frozen data and a selection of the other participants who all have around 30-34%.

5.1.1 Stimulus stage

Topographic mappings (figures 4.1 and 4.2) of the evoked response to each of the audio, visual and text stimuli have been included for two participants who fully completed the recordings (Participants 1 and 3) displaying responses most and least alike to that which is expected given past research. As can be seen in Figure 4.1, participant 1 demonstrated evoked responses very much in line with that which is expected in terms of regions of activation. For the audio stimuli, there is clear activation of the temporal lobe (T7) where the primary auditory cortex is located [20]. For the visual stimuli, activation of the occipital lobe (O7), the location of the primary visual cortex [20]. And for text stimuli, activation of areas in the left frontal lobe (F7, F3, AF3) where brocas area is located, and in the left temporal lobe, the location of wernickes area [21]. On the other hand, participant 3 did not demonstrate the behaviour expected with activation existing predominantly in the frontal lobe throughout audio, visual and text stimuli, as can be seen in figure 4.2. It is notable that the dataset corresponding to participant 3 had the joint highest percentage of frozen data out of all participants (roughly 34% of values), whereas the dataset corresponding to participant 1 had the least frozen data out of all participants (roughly 15% of values). Both participants 1 and 3 completed recordings fully, and this difference in correspondence is a good representation of the negative impact the

frozen values have on the quality of data collected. This pattern of non-correspondance is true of all the datasets with 30-34% frozen values, and although there is a possibility of these participants just having abnormal thinking patterns, this seems unlikely seeing how many other problems the frozen values have caused. Given how well the observed regions of activation correspond to theory for participant 1, it suggests that without the frozen values, the data recorded in this work could have been reasonable quality.

5.1.2 Speaking stage

For participant 1 seen in figure 4.3, the presence of a difference in activation regions between the rest state and during speech imagery following the audio, visual and text stimuli (when actually performing speech imagery), that imagery is taking place, in particular around 0.665s. Once again though, as in the stimulus stage, for data with a higher frozen percentage, the pattern that theory would explain is not nearly as well expressed. This can be seen in figure 4.4 where there is not such a clear distinction between behaviour between when a participant is at rest versus performing speech imagery.

Activation does not appear vastly different between modes of stimuli however for participant 1, with frontal lobe (F3, F4, F7), right temporal lobe (T8), and right occipital lobe (O2) showing activation in all 3 modalities. This is interesting with the right temporal lobe said to involved in processing non verbal information. It is also interesting that the occipital lobe (responsible for visual processing) appears to see larger activation prior to receiving a visual stimuli during prompting. This could imply that the visual cue is still being processed in this participants mind, and biasing speech imagery as a consequence. Participant 1 performed Imagined speech and the activation in the frontal lobe could be down to the imagined movement of speech articulators which has required the use of the motor cortex. Figure 4.4 shows the topographical map of participant 4's response throughout the performance of inner speech. Understanding the the areas of activation within the human cortex when performing inner speech remains an elusive challenge within the field [16]. Although unreliable given the low quality of data, results from participant 4 would suggest the frontal lobe is involved with strong activation observed throughout all three modalities compared to the rest state.

5.2 Performance of machine learning models

In total 162 models were trained in this research and results of which can be seen in tables S1, S2, and S3. As previously mentioned, as well as potential enhanced speaking and comprehension after receiving a given mode of stimuli, there is also a possibility of bias introduced into speech imagery which machine learning algorithms could become fit to. Its ambiguous and unknown what exactly is going through a participants mind during the speaking stage, and after audio stimuli a "playback" may go through someones mind, whilst after visual stimuli a visualisation may go through a participants mind for instance. In simpler terms, its not possible to infer that a particular mode of stimuli caused a difference in the ability to decode speech from EEG for the right reasons, hence its not wise to make claims that a given mode of stimulus is "superior" even if observing superior performance.

Prior to conducting analyses and performing what was outlined in the processing pipeline (3.9) discussed in 3.2.1, it was realised that the decision to perform filtering and ICA on the dataset before the cross validation, and not perform it for each fold, risks data leakage between

prospective train and test samples. Having the consequence of some of the information from train data contained within the test data, biasing machine learning algorithm performance. Given the perceived superiority of data from participant 1 compared to the other participants mentioned previously, at face value its surprising that no model from this participant performed significantly above chance level compared with some of the other participants seeing multiple models above the chance level, and this was the only fully complete dataset to not see a model perform significantly. However, when looking a bit further, the reason/causation behind good model performance in this research may not be "successful learning" of patterns in speech imagery and the ability to decode the signal of interest, rather a result of such data leakage with the algorithms learning the noise within the train set. More frozen values within the data means more disruption to the signal, hence it makes sense why models for participant 1 see such inferior performance. Furthermore, it was found that the participants with the highest number of models significantly above chance level were those with the highest number of samples per class used in training/test after processing, with a moderate positive correlation found between the two (R=0.44). This appears to be contrary to intuition, as lower number of samples should mean even more similar train and test sets. Out of the 162 models trained, 18 were found to perform significantly above the chance level, 16 being on inner speech, 12 of those being random forests. Given the data leakage, superior performance from the random forest is likely down to a greater propensity for over-fitting. However, there doesn't seem to be any reason relating to data leakage as to why models perform better on inner speech than imagined, and this seems contrary to findings in the literature.

Given the issues within the data and subsequent processing pipeline, it is difficult to determine the influence mode of stimulus has on the ability to decode inner and imagined speech, it can be seen in some of the evoked responses that potential bias may arise from particular modes of stimuli (namely visual stimuli), however problems relating to frozen values and data leakage prevent this from being properly tested with any validity. From the p-values in table S4 relating to each individual one-way ANOVA test. Its seen that in no case throughout 4-way classification, binary classification of vowel pairs or binary classification of syllable pairs was there a significant difference in accuracy between models trained on data with the different modes of stimulus. Therefore it must be stated that in the case of the hypotheses 1, 2, and 3 there is insufficient evidence to reject the null hypothesis.

Chapter 6

Conclusions

Technical issues occurring in data collection made results of the investigation unreliable with very little external validity. On a positive note, this work can serve as a lesson to those looking to research the decoding of inner and imagined speech, and on a broader sense, to those looking to record EEG data. Within this work a data collection methodology was created for which adequate quality and quantity of data could be recorded given some minor adjustments in programming, and sets up the possibility for another research team (potentially with more resources) to apply the methodology with more success. Firstly the introduction of frozen values into the data was near fatal for data quality, and lowers the signal to noise ratio in data that already has low signal to noise ratio. Efforts taken to try and subvert the problem proved unsuccessful and ended up further hindering results, and introducing data leakage between train and test samples. The occurence of the frozen values was only a programming error however and was not caused by poor theoretical design directly. The same can be said of design flaw 4, where a tweak in the code used for conducting recordings could allow for the whole signal throughout recordings to be saved. On the other hand, the decision to process the data as outlined in figure 3.9 to try and account for design flaws 2 and 4 was poorly thought out and lead to obvious data leakage. Design flaw 3 is also likely down to poor experimental setup, and not properly managing fatigue levels, making recordings less intensive and/or providing greater incentive to complete recordings may be required to prevent such drop out rates.

Results suggested that mode of stimuli used in the stimulus stage of recordings does not influence the ability to decode subsequent inner or imagined speech. Moreover, the random forest algorithm appeared to largely outperform the deep neural network, either suggesting that the danger of overfitting is worse for the random forest or the ability to decode speech imagery is better, with the former being the most likely given the problems with the data. However it was found that models performed better on the task of decoding inner speech, with an explanation based off methods applied in this work not found.

Bibliography

[1] Abdulkader, S.N., Atia, A. and Mostafa, M.S.M., 2015. Brain computer interfacing: Applications and challenges. *Egyptian informatics journal*, 16(2), pp.213–230.

[2] Agarwal, P., Kale, R.K., Kumar, M. and Kumar, S., 2020. Silent speech classification based upon various feature extraction methods. *2020 7th international conference on signal processing and integrated networks (spin)*. IEEE, pp.16–20.

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

[4] AlSaleh, M., Moore, R., Christensen, H. and Arvaneh, M., 2018. Discriminating between imagined speech and non-speech tasks using eeg. *2018 40th annual international conference of the ieee engineering in medicine and biology society (embc)*. IEEE, pp.1952–1955.

[5] Angrick, M., Herff, C., Mugler, E., Tate, M.C., Slutzky, M.W., Krusienski, D.J. and Schultz, T., 2019. Speech synthesis from ecog using densely connected 3d convolutional neural networks. *Journal of neural engineering*, 16(3), p.036019.

[6] 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.

[7] Cooney, C., Folli, R. and Coyle, D., 2018. Mel frequency cepstral coefficients enhance imagined speech decoding accuracy from eeg. *2018 29th irish signals and systems conference (issc)*. IEEE, pp.1–7.

[8] Correia, J.M., Jansma, B.M. and Bonte, M., 2015. Decoding articulatory features from fmri responses in dorsal speech regions. *Journal of neuroscience*, 35(45), pp.15015–15025.

[9] Dowding, I., Debener, S., Müller, K. and Tangermann, M., 2015. On the influence of high-pass filtering on ica-based artifact reduction in eeg-erp. *Paper presented at the 37th annual international conference of the ieee engineering in medicine and biology society (embc)*. Milan Italy.

[10] Du, B., Cheng, X., Duan, Y. and Ning, H., 2022. fmri brain decoding and its applications in brain–computer interface: A survey. *Brain sciences*, 12(2), p.228.

[11] Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L. and Hämäläinen, M.S., 2013. MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience* [Online], 7(267), pp.1–13. Available from: `https://doi.org/10.3389/fnins.2013.00267`.

[12] 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.

[13] Lopez-Bernal, D., Balderas, D., Ponce, P. and Molina, A., 2022. A state-of-the-art review of eeg-based imagined speech decoding. *Frontiers in human neuroscience*, 16.

[14] Mak, J.N. and Wolpaw, J.R., 2009. Clinical applications of brain-computer interfaces: current state and future prospects. *Ieee reviews in biomedical engineering*, 2, pp.187–199.

[15] Makin, J.G., Moses, D.A. and Chang, E.F., 2020. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature neuroscience*, 23(4), pp.575–582.

[16] 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* [Online], 12. Available from: `https://doi.org/10.3389/fnins.2018.00422`.

[17] Mur, A., Dormido, R. and Duro, N., 2019. An unsupervised method for artefact removal in eeg signals. *Sensors*, 19(10), p.2302.

[18] 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.

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

[20] Purves, D., Augustine, G., Fitzpatrick, D., Katz, L., LaMantia, A., McNamara, J. and Williams, S., 2001. Neuroscience 2nd edition. sunderland (ma) sinauer associates. *Types of eye movements and their functions*.

[21] Purves, D., Augustine, G., Fitzpatrick, D., Katz, L., LaMantia, A., McNamara, J. and Williams, S., 2001. Neuroscience 2nd edition. sunderland (ma) sinauer associates. *Types of eye movements and their functions*.

[22] Ramele, R., Villar, A.J. and Santos, J.M., 2022. Report: Epoc emotiv eeg basics. *arxiv preprint arxiv:2206.09051*.

[23] Vander Wyk, B.C., Ramsay, G.J., Hudac, C.M., Jones, W., Lin, D., Klin, A., Lee, S.M. and Pelphrey, K.A., 2010. Cortical integration of audio-visual information. *Brain and cognition*, 74(2).

[24] Viola, F.C., Debener, S., Thorne, J. and Schneider, T.R., 2010. Using ica for the analysis of multi-channel eeg data. *Simultaneous eeg and fmri: Recording, analysis, and application: Recording, analysis, and application*, pp.121–133.

[25] Walsh, J.K., 1979. Evoked brain responses to auditory and visual stimuli of equal subjective magnitude. *Perception & psychophysics*, 26(5), pp.396–402.

[26] Wu, H., Pan, C., Li, M. and Chen, F., 2021. A method to map eeg signals to spoken speech using gaussian process modeling. *2021 29th european signal processing conference (eusipco)*. IEEE, pp.1140–1144.

[27] 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.