Accurate Decoding of Speech Information from Neurophysiological Data

Sophie Crowley

A Final Year Project

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

M.Sc. Computer Science (Intelligent Systems)

Supervisor: Giovanni Di Liberto

August 2022

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Sophie Crowley

August 19, 2022

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Sophie Crowley

August 19, 2022

Abstract

Electrophysiological activity in the brain can be recorded as a subject listens to or imagines speech. Decoding speech information from this brain data means attempting to reconstruct elements of the original speech stimulus. The eventual long-term goal of this type of speech reconstruction would allow for subjects to imagine what they want to say and have a brain-computer interface produce it for them. This would help individuals with diseases such as Amyotrophic lateral sclerosis (ALS) to speak when they no longer have the motor ability to do so. Decoding speech information also helps us to discover more about the brain and how it works in relation to speech. Currently, non-invasive methods of recording brain activity have not achieved as much success in decoding as invasive methods due to the noise and inaccuracy of the data. We set a new baseline for decoding speech information using non-invasive brain recording methods. This involves using Multiway Canonical Correlation Analysis (MCCA) to de-noise the data and a Temporal Response Function (TRF) to decode the brain data into vocoder features. The features were then input into a vocoder to attempt to reconstruct the stimulus speech audio. We also analysed how speech that a subject imagines is decoded in comparison to speech that a subject listens to and proved that our methodologies can work for different types of brain data recording.

Acknowledgments

First and foremost, I would like to thank my supervisor, Dr. Giovanni Di Liberto. They have been extremely helpful, knowledgeable, enthusiastic, and supportive throughout this entire process. This project would not be half of what it is today without all of the help and I am extremely grateful. Thank you to everyone else in the Di Liberto lab as well, it has been great to meet you all and help each other throughout the year.

I would also like to thank my family, who have been incredible during this very tough Masters year. I have them and my friends to thank for helping me through it. Finally, I would like to thank Buddy and Remy, the best boy and the evil fox.

Sophie Crowley

University of Dublin, Trinity College
August 2022

Contents

Abstract iii

Acknowledgments iv

Chapter 1 Introduction 1
- 1.1 Motivations . . . 1
  - 1.1.1 Practical Motivation . . . 1
  - 1.1.2 Theoretical Motivation . . . 2
- 1.2 Objectives . . . 2
- 1.3 Overview of Project . . . 3

Chapter 2 Background 5
- 2.1 Recording Brain Activity . . . 5
  - 2.1.1 Electroencephalography (EEG) . . . 5
  - 2.1.2 Magnetoencephalography (MEG) . . . 6
  - 2.1.3 Electrocorticography (ECoG) . . . 6
- 2.2 Decoding Brain Activity . . . 7
  - 2.2.1 Temporal Response Functions . . . 7
  - 2.2.2 Deep Learning . . . 8
  - 2.2.3 MCCA . . . 8
- 2.3 Synthesising Speech . . . 9
  - 2.3.1 Speech Vocoding . . . 10
  - 2.3.2 Imagined Speech . . . 12
- 2.4 Summary . . . 13

Chapter 3 Methodology 14
- 3.1 Overview of the Approach . . . 14
- 3.2 Datasets . . . 15
  - 3.2.1 Dataset 1: EEG . . . 15

3.2.2 Dataset 2: MEG . . . 16
3.3 Speech Vocoder . . . 16
3.3.1 Creating Stimulus Features From Audio . . . 16
3.3.2 Creating Audio From Stimulus Features . . . 17
3.4 Data Preprocessing . . . 18
3.4.1 Stimulus . . . 18
3.4.2 EEG . . . 19
3.4.3 Stimulus-EEG Coordination . . . 19
3.5 Models . . . 19
3.5.1 Individual Subject . . . 19
3.5.2 Average Subject . . . 20
3.5.3 MCCA Subject . . . 21
3.6 Application on MEG Data . . . 22
3.6.1 Extracting MEG Data Epochs . . . 22
3.6.2 Further Adaptation . . . 24

Chapter 4 Results **25**
4.1 Overview . . . 25
4.2 EEG Dataset . . . 25
4.2.1 Stimulus Predictions . . . 25
4.2.2 Audio Produced . . . 33
4.3 MEG Dataset . . . 38
4.3.1 Stimulus Predictions . . . 38
4.3.2 Audio Produced . . . 47

Chapter 5 Conclusions **53**
5.1 Overview . . . 53
5.2 Future Work . . . 54

Bibliography **55**

Appendices **57**
.1 Appendix A . . . 58
.2 Appendix B . . . 58
.3 Appendix C . . . 60

List of Tables

4.1 Objective metrics for 'The Standard' of reconstructed speech. These results comes from getting stimulus features from the vocoder, downsampling them to 32Hz, resampling them to 200Hz, then putting them back through the vocoder again as inputs. 34

4.2 Comparing the reduction in speech intelligibility and quality of compressing the stimulus bands and sampling frequency. No compression means that nothing is done to the stimulus, it is created from the vocoder and immediately put back through it to produce speech. 35

4.3 Comparison of model predictions in regards to speech intelligibility and quality metrics. The difference between Normalised MCCA and MCCA is that the latter did not involve normalising the stimulus training values. . . 35

4.4 Effects of scaling F0 and modulating the audio with the stimulus envelope of speech intelligibility and quality. 37

4.5 Speech intelligibility and quality results when replacing predicted features with real stimulus data. 38

4.6 Objective metrics for 'The Standard' of reconstructed speech with the MEG dataset. These results comes from getting stimulus features from the vocoder, downsampling them to 100Hz, resampling them to 200Hz, then putting them back through the vocoder again as inputs. 48

List of Figures

2.1 Equation to reconstruct a stimulus feature at time t. n represents the current electrode channel and τ represents the current lag. r is the brain data at the current time, lag, and channel, while g is the decoder mapping. Figure adapted from Crosse et al. (2016) 7

2.2 An overview of the MCCA formulation. Each X variable represents a subject, while Y represents the final matrix. Figure adapted from de Cheveigné et al. (2019). 9

2.3 F0 for an example audio. The values range from 0 Hz, which represents silence, to approx. 200 Hz which represents the highest frequency reached during the recorded speech. 10

2.4 Spectrogram for an example word in our dataset. The frequency axis is split into bands of frequency which increase logarithmically. The frequencies at which the most power lies change over time, signalling different phones in the audio. 11

2.5 Aperiodicity for an example audio. It can be seen that there are blocks of near-0 values which represent voiced segments and blocks of near-1 values which represent unvoiced segments. 12

3.1 The overall workflow of this project at a glance. 15

4.1 A comparison of EEG R values based on different models for each stimulus feature. The values are the mean R correlation over 20 trials. The standard error is also shown for each bar. 26

4.2 R value results for models trained on individual MCCA components. . . . 28

4.3 R values for different numbers of components used in an MCCA model. This is used to choose the optimal number of components to include in each feature's final MCCA model. 29

4.4 Comparing stimulus F0 with reconstructed F0. 31

4.5 Comparing stimulus spectrogram with reconstructed spectrogram. 32

4.6 Comparing stimulus aperiodicity with reconstructed aperiodicity. 33

4.7 Comparison of model predictions in regards to normalised speech intelligibility and quality metric results. 36
4.8 Spectrograms reconstructed using imagined MEG data with 32-component MCCA models. The first used a model with the highest lambda value possible. The frequencies with the most power are constantly within the approx. 6th frequency band. This rarely changes over time, showing an overgeneralised model. The second used a model the second highest lambda value. While the predictions may not be as accurate in terms of R values, the frequencies with the most power at least change as time passes. 39
4.9 A comparison of listened MEG R values based on different models for each stimulus feature. The values are the mean R correlation over 20 trials. The standard error is also shown for each bar. 40
4.10 A comparison of imagined MEG R values based on different models for each stimulus feature. The values are the mean R correlation over 20 trials. The standard error is also shown for each bar. 41
4.11 Spectrograms for a sample audio from the stimulus and MCCA models using listened MEG data and imagined MEG data. 43
4.12 Average correlation of word pairs that are the top 100 most highly correlated and lowly correlated in the stimulus spectrogram. 44
4.13 Average correlation of word pairs that are the top 100 most highly correlated and lowly correlated in the stimulus spectrogram when a smaller lambda is used in the MCCA models. 45
4.14 Average correlation of word pairs that are the top 100 most highly correlated and lowly correlated in the stimulus spectrogram when a large number of components are used in the MCCA model. 46
4.15 Spectrograms for two words that are highly correlated. The stimulus spectrogram is compared to the MCCA model predictions using listened and imagined MEG data. 47
4.16 Comparison of model predictions in regards to normalised speech intelligibility and quality metric results with listened MEG data. 49
4.17 Comparison of model predictions in regards to normalised speech intelligibility and quality metric results with imagined MEG data. 50
4.18 Comparison of audio metric results for reconstructions with an array of improvements on the initial MCCA model. 52

1 Correlation of MCCA Components to EEG Electrodes. The absolute values of the correlations were averaged over the 19 EEG subjects and the 20 audio trials. 59
2 R Value of Individual Spectrogram Bands for MCCA Model. 60
3 R values for different numbers of components used in an MCCA model with listened MEG data. This is used to choose the optimal number of components to include in each feature's final MCCA model. 61
4 R values for different numbers of components used in an MCCA model with imagined MEG data. This is used to choose the optimal number of components to include in each feature's final MCCA model. 62

Chapter 1

Introduction

1.1 Motivations

1.1.1 Practical Motivation

Deciphering speech information from brain data can be useful in many ways. It can have both theoretical and practical applications. For example, many people are born without or lose the ability to speak due to diseases such as Amyotrophic Lateral Sclerosis (ALS). Without control over their own muscles, they cannot make the articulatory gestures needed to produce sound. However, the brain structures that are responsible for the cognition and production of speech may be unaffected. Devices that help people in these situations to communicate already exist but are generally limited in their expression and speed of communication, e.g. using eye-tracking to spell out a word one letter at a time.

With the advances in neural recording, machine learning, and the processing power of devices in recent years, the use of these types of brain-computer interfaces (BCI) is more promising and viable than ever. Furthermore, new insights into how our brains work mean that researchers are better equipped to develop more complex BCIs. One particular area of interest is continuous speech. The ideal long-term application of a BCI would allow a user to think of what they want to say and have it immediately be produced by the device. This would allow for a much wider breadth of expression and would be significantly quicker.

While this is not yet a possibility, there has been significant progress in the field. In particular, electrocorticography (ECoG) has already been used to decode speech by recording a subject's brain data as they listen to speech (Akbari et al., 2019). This is a large improvement on previous research. The issue, however, is that ECoG is a very invasive form of recording that requires direct access to the brain. Non-invasive electroencephalography (EEG), on the other hand, is a form of recording that is significantly less

invasive as the sensors are placed on the outside of the scalp. EEG is already the most popular form of recording for BCI devices due to how non-invasive and cheap it is in comparison to other methods.

The question is whether there is a way to decode speech information using EEG as this would be extremely useful. The success with ECoG leads us to believe there is a chance. Can EEG record enough relevant, accurate information as someone is listening to speech in order for it to be decoded? While EEG can definitely record enough useful brain data to conduct simple tasks (Värbu et al., 2022), is it too noisy and inaccurate for complex and continuous data such as speech?

1.1.2 Theoretical Motivation

A more short-term goal is to discover more about how the brain works. The more knowledge we have about the brain and how it functions, the more possible these BCI applications will also be in the future. Analysing brain data has already shown us a lot about how humans process speech, such as where the brain processes speech and when. Different parts of the brain process different levels of speech, i.e. low level acoustics versus high level meanings (Price, 2012).

In particular, if a successful method for decoding listened speech information from EEG is found, can we use this with better, cleaner data to decode imagined speech? Imagined speech is different to listened speech as there is no actual audio or motor input that the subject listens to, everything is within the brain. The term often used in the literature for speech that a subject imagines is 'speech imagery', but the less ambiguous term of 'imagined speech' will be used here for ease and clarity.

Magnetoencephalography (MEG) is another form of recording which is non-invasive like EEG but more expensive. If MEG data is less noisy than EEG and EEG is producing significant results for decoding listened-to speech, it may be possible to prove that MEG can pick up significant information from imagined speech. If MEG can solve a harder problem such as decoding imagined speech, it would justify its use and its cost. The question is whether there is enough information in the MEG data for this to be possible. It would also help us learn about how imagined speech in general is processed by the brain.

1.2 Objectives

The main aim of this project is to determine the upper bound of decoding EEG data, i.e. what and how much speech information is actually encoded in EEG and what is the

best that linear models of decoding can achieve. A more long-term goal is to decode EEG brain data recorded as a subject listens to speech back into said speech. This is a difficult goal that may not ever be possible, but we hope to contribute to and further the research effort. Either way, we aim to set a baseline for future EEG research in this area based on our methods. This can be broken into the following sub-goals:

- Transform the EEG data into a form that is de-noised and summarises multiple subjects. This will be done using a variety of pre-processing methods along with Multiway Canonical Correlation Analysis (MCCA).
- Transform the stimulus speech data into a form that can be used as an output/goal for our model.
- Use a linear model to decode the input EEG data into a reconstruction of the stimulus data.
- Assess how effective this process and model is in order to understand what information the EEG data contains about speech.

Following this, the second aim is to apply the developed methods on another form of brain data, MEG for imagined speech. The primary goal will be for these models to allow for the decoding of imagined speech data into useful information. This allows us to attempt to answer questions such as whether the brain processes imagined speech like listened speech and if the content of imagined speech can even be decoded. This will entail the following:

- Adapt the current methods to work for a new MEG dataset.
- Run and analyse these methods on both listened-to and imagined speech MEG data.
- Understand what is and is not being included and decoded well.

1.3 Overview of Project

Chapter 2 provides essential background information in order to give context to the project, as well as a review of state of the art related work in the field.

In Chapter 3, the approach taken for this project is detailed, as well as a description of the key elements and decisions involved in producing the final implementation. This includes the preprocessing and transformation of data, the creation of linear models, the decoding and production of the speech itself, and the adaptation of the final implementation to work with MEG data.

Chapter 4 contains an evaluation of the developed process. This includes both the EEG data results as well as the imagined MEG data results.

Finally, Chapter 5 provides final conclusions in relation to the project and the progress made, as well as suggestions for possible future work.

Chapter 2

Background

In this section, an overview of the ways in which brain data for speech can be recorded is provided, as well as an analysis of how this data is decoded back into speech. This section also aims to present an outline of the current state of the art approaches in producing reconstructed speech.

2.1 Recording Brain Activity

A variety of methods can be used to record brain activity. They can largely be split into two categories. The first concerns those based on measuring metabolic processes in the brain, i.e. how much oxygen-carrying blood is in each part of the brain (Herff and Schultz, 2016). Methods in this category, such as functional magnetic resonance imaging (fMRI), can provide detailed information but are inherently slow. As such, they are not very suitable for speech data as it is continuous and changes quickly.

The other category is more focused on electrophysiological signals. This involves using electrodes to measure electrical currents in the brain that create electrical and magnetic fields (da Silva, 2013). These methods are more suited to recording speech data as the responses are quicker and can pick up changes in things such as phones. As such, we will focus on the main three electrophysiological methods: EEG, MEG, and ECoG (Herff and Schultz, 2016). Each has its own benefits and drawbacks, mostly relating to how invasive the recording is and how detailed or noisy the data itself is.

2.1.1 Electroencephalography (EEG)

EEG is the least invasive and least expensive form of electrophysiological recording. As such, it has been widely used to record brain data for a variety of uses. One notable example would be the P300 Speller BCI which allows the user to input text into a computer

based on solely thinking about it (Guan et al., 2004). However, the non-intrusiveness of EEG is also why the data recorded using it is generally very noisy. Electrodes are placed around the scalp which leads to any sort of motion by the subject affecting the data, as well as the distance of the electrodes from the brain itself limiting the resolution of the data (Panachakel and Ramakrishnan, 2021).

While EEG has achieved much success with challenges in other research areas, speech is continuous and complicated. It is currently unclear whether EEG recordings of brain data alone are accurate enough to be used to decode speech. This project attempts to contribute towards answering that question.

2.1.2 Magnetoencephalography (MEG)

Rather than electrical fields, MEG records magnetic fields in the brain using magnetometers around the scalp. It has been shown to give better results than EEG when it comes to tracking speech brain data (Destoky et al., 2019). MEG is less prone to noise than EEG and has an overall higher spatial and temporal resolution (Herff and Schultz, 2016). The downside, however, is that MEG is harder and more expensive to set-up and use. This makes it less convenient to use for BCI applications as the apparatus used by subjects should be accessible.

MEG has already shown promise in regards to decoding speech (Dash et al., 2020b). What has not yet been proven, however, is whether MEG can accurately decode imagined speech, rather than speech that the subject is actively listening to. This project aims to contribute towards the argument that MEG can record imagined speech well and be used to decode it into useful speech information. If this is true, it would also contribute towards the argument that MEG is worth the resources needed to use it and is a viable method for speech BCI applications in the future.

2.1.3 Electrocorticography (ECoG)

ECoG is an invasive electrophysiological method that has shown the most success in decoding speech from brain data. ECoG is not as prone to noise as the other methods and has the highest spatial and temporal resolution (Martin et al., 2018). Overall, it is the ideal method for recording continuous speech. There have already been multiple studies that show a successful decoding of ECoG data back into speech, such as that of Herff et al. (2015) and Akbari et al. (2019).

However, the reason ECoG data is so accurate is because it is measured by placing grids directly on the brain itself, such as during brain surgery. This means it is currently not very viable in the BCI field as an on-the-go device. Furthermore, the opportunities

to record ECoG data are naturally a lot more limited than that of EEG or MEG. This is why this project aims to use some of the methods others have applied to ECoG data on EEG and MEG data to see if we can produce similar results.

2.2 Decoding Brain Activity

Brain data, once recorded, then needs to be analysed and transformed so that we can try to predict, or 'reconstruct', the original stimulus. The stimulus in this case means the speech that a subject listened to or imagined. There are a variety of approaches and methods that a researcher can use to do this. This section outlines some of the most prominent and relevant methods in the field for decoding speech from brain data.

2.2.1 Temporal Response Functions

The Temporal Response Function (TRF) is the most common linear way of relating continuous stimuli such as speech to recorded brain data. A TRF is a form of regularised linear regression, ridge regression, mapping the stimulus to the response (Crosse et al., 2016). What we use is a decoding, or 'backwards', model, which predicts the stimulus from the brain data. A TRF is also capable of a 'forwards' encoding model which does the opposite, predicting the brain's response to a stimulus. A point of note on TRFs is that they use the concept of 'lag' in their formula. Humans do not process speech instantly, so there is a certain amount of time between hearing something and the brain processing it. As such, the TRF uses a time lag range to see when this happens (Crosse et al., 2016). The formula below shows us how EEG data is used to reconstruct the stimulus audio.

$$\widehat{s}(t) = \sum_{n} \sum_{\tau} r(t+\tau, n) g(\tau, n),$$

Figure 2.1: Equation to reconstruct a stimulus feature at time t. n represents the current electrode channel and τ represents the current lag. r is the brain data at the current time, lag, and channel, while g is the decoder mapping. Figure adapted from Crosse et al. (2016)

While there are other methods in the field that map discrete or short stimuli, such as those related to measuring event-related potentials (Luck, 2005), TRFs have the advantage of working with continuous data. The relatively recent development of TRFs means we can analyse responses to continuous natural speech. For example, multiple studies such

as that of Vanthornhout et al. (2018) have successfully used TRFs to predict the speech envelope of an audio. A TRF is a linear method, which generally makes it easier and quicker to implement and run than more complex deep learning methods. Linear methods also have the advantage of being easier to analyse and understand.

2.2.2 Deep Learning

Many researchers have turned to deep learning methods such as CNNs (Angrick et al., 2019) and DNNs (Akbari et al., 2019) to try and decode data as complex as speech. Results can be very successful but it is harder to analyse the inner workings of the model. Furthermore, deep learning methods often require a large amount of data to work well (Alzubaidi et al., 2021). In cases where there is limited data or resources, deep learning may not be ideal. For these reasons, this project does not use any deep learning methods and instead uses the linear TRF method. If the TRF method works well, it is very possible in future to replace this with a deep learning method in the hopes of improving the results.

2.2.3 MCCA

Brain data, especially non-invasive data, is noisy. In order to produce acceptable results with our models, we first need to find the most important and useful parts of the raw brain data. The main way of achieving this with at most two subjects is canonical correlation analysis (CCA). CCA involves finding linear combinations between two datasets in order to maximise their correlation (Zhuang et al., 2020). In the case of EEG data, CCA would be looking at data across electrodes for two subjects to find the brain activity common to the both of them. It can then provide us with components to use instead of the raw data.

However, many datasets include more than two subjects which led to the creation of multiway canonical correlation analysis (MCCA). This method works similarly to CCA except across more than two subjects. First, each subject's data is transformed using principal component analysis into a common representation so that they can be compared to each other. They are then concatenated into one matrix which is put through PCA again. This final matrix is used to retrieve MCCA components.

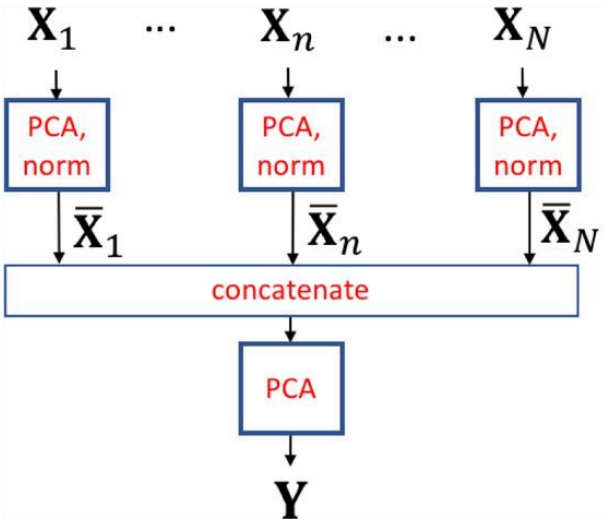

Figure 2.2: An overview of the MCCA formulation. Each X variable represents a subject, while Y represents the final matrix. Figure adapted from de Cheveigné et al. (2019).

MCCA is useful for denoising data, summarising it, reducing its dimensionality, and has already been shown to be more effective than just averaging the data (de Cheveigné et al., 2019). The datasets used throughout this project all have multiple subjects, leading to MCCA being a good choice to implement all of the above without using more involved deep learning methods.

2.3 Synthesising Speech

Input data and linear methods can only work if there is also appropriate output data to be trained on. In other words, we need the models to predict something that can be turned into speech. This can come in the form of many things such as phones, text, spectrograms, articulatory features, and so on.

Throughout the history of synthetic speech production, different forms of vocoders have been used to produce speech. A vocoder attempts to synthesise a human voice when given certain parameters. Recently, there have been numerous successful research projects that use ECoG data combined with a vocoder to reproduce various forms of speech. Most prominently, a study by Akbari et al. (2019) used ECoG data, a DNN, and the WORLD Vocoder (Morise et al., 2016) to produce intelligible continuous speech. Our study aims to adapt some of these ideas to work with more noisy EEG data and linear methods. The use of a vocoder allows for direct reproduction of continuous speech, rather than segments or patterns that need to be put together. Furthermore, using a vocoder has been shown by Akbari et al. (2019) to provide more accurate results than that of a spectrogram, which is the other continuous option.

2.3.1 Speech Vocoding

The vocoder developed by Morise et al. (2016) requires three inputs to produce audio: The fundamental frequency (F0), the spectrogram, and the aperiodicity.

F0 is known as the frequency that vocal chords vibrate at and is expressed in Hertz (Hz). More definitively, it is the lowest frequency of a periodic waveform. When there is no sound, F0 would be 0. It is closely related to pitch so high pitched sounds such as /s/ would have a higher F0 and a male speaker would generally have a lower F0 during speech on average. The method used by WORLD to estimate the F0 of an audio is called DIO (Morise et al., 2009) and it has been shown to be faster than its competitors without sacrificing accuracy (Morise et al., 2016).

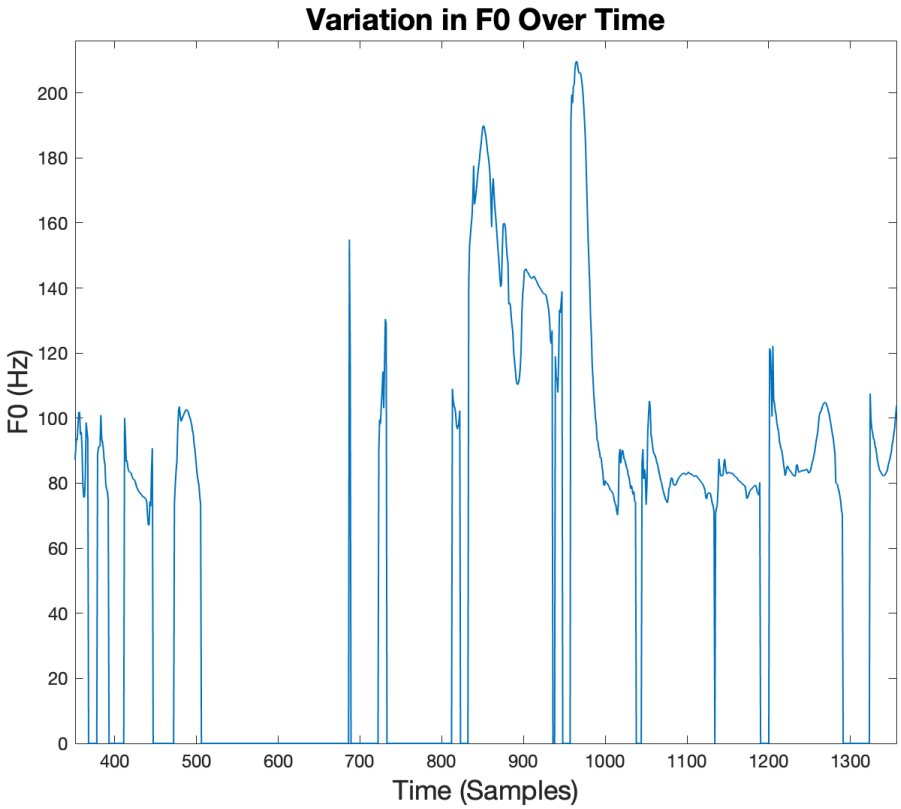

Figure 2.3: F0 for an example audio. The values range from 0 Hz, which represents silence, to approx. 200 Hz which represents the highest frequency reached during the recorded speech.

The spectrogram, or the spectral envelope, is a representation of the power spectrum of frequencies over time. It is a three-dimensional representation made up of time, frequency, and amplitude. A spectrogram is very useful and can be used to differentiate phones and identify words based on what frequencies contain the most power over time. It is a

very important part of speech construction for a vocoder. The WORLD Vocoder uses CheapTrick (Morise, 2015) to estimate the spectrogram.

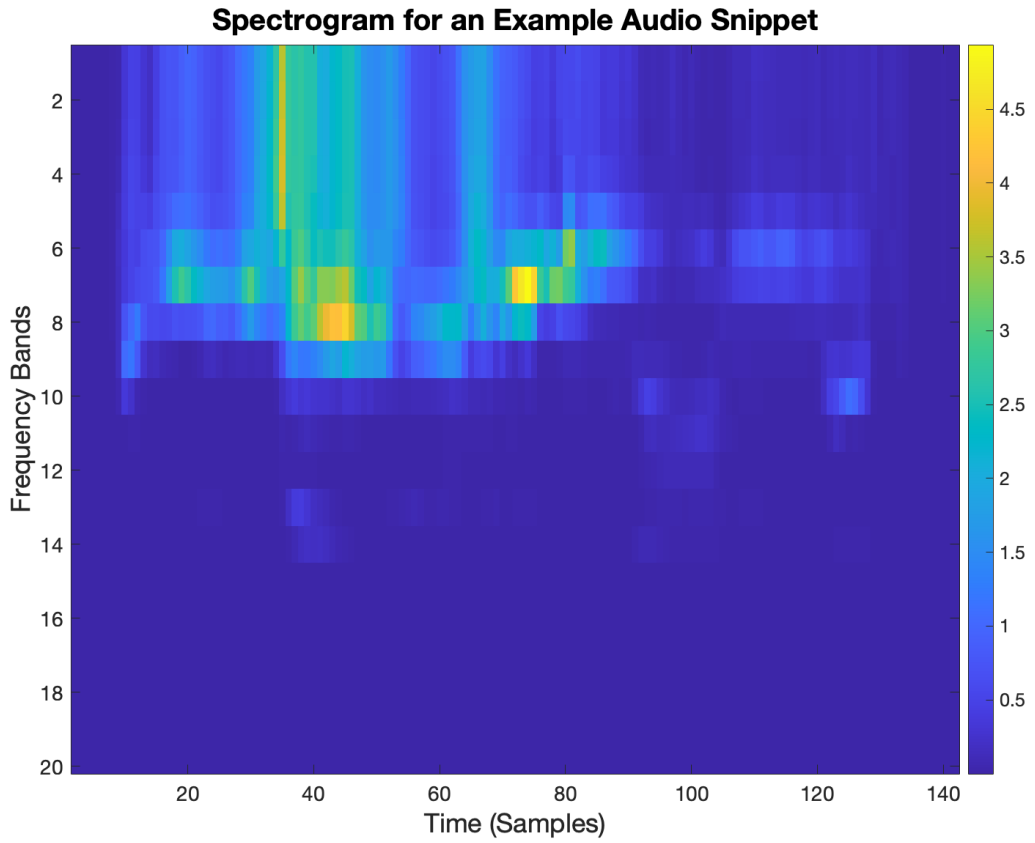

Figure 2.4: Spectrogram for an example word in our dataset. The frequency axis is split into bands of frequency which increase logarithmically. The frequencies at which the most power lies change over time, signalling different phones in the audio.

The final feature, the aperiodicity, is also known as the excitation signal. This is calculated directly from the audio's waveform, F0, and spectrogram (Morise et al., 2016). It is related to whether the speech at any given time is voiced or unvoiced.

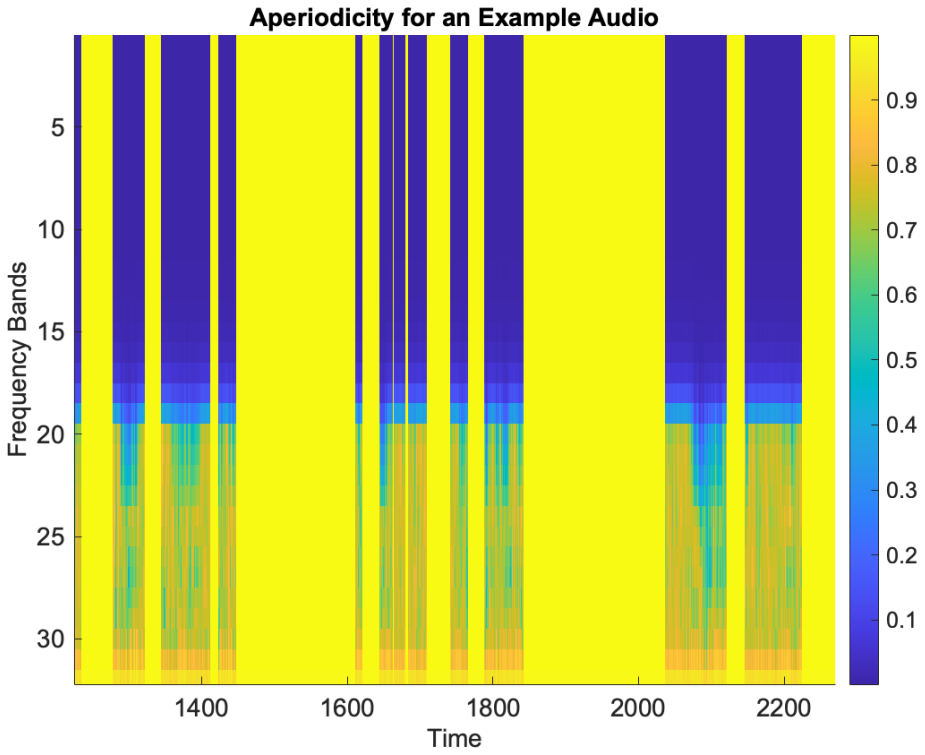

Figure 2.5: Aperiodicity for an example audio. It can be seen that there are blocks of near-0 values which represent voiced segments and blocks of near-1 values which represent unvoiced segments.

2.3.2 Imagined Speech

It has already been shown that it is possible to reconstruct speech from ECoG data of the subject listening to said speech. The next step in order to create BCI devices to help those who cannot speak would be reconstructing imagined speech. Imagined speech data entails the subject imagining the speech being spoken rather than listening to an audio clip. There have been some signs of progress on this front with ECoG data (Proix et al., 2022), but less so when using non-invasive forms of recording like EEG and MEG.

In order to reconstruct speech using this type of data, it must first be proven whether significant and useful data can even be retrieved from a subject imagining the audio. MEG, being the less noisy of the two non-invasive forms of recording, is the more promising candidate. Some work has begun to investigate this, such as that by Dash et al. (2020a), but there is still no definitive consensus. Our research hopes to contribute to a positive consensus with significant results for the reconstructed vocoder features.

2.4 Summary

An overview of the state of the art methods in the fields of speech recording, brain data decoding, and speech reproduction has been given. From this overview, it is clear that EEG, MEG, and ECoG all come with different benefits and drawbacks. EEG, as a non-invasive and cheap form of recording, is a prime candidate for BCI devices if accurate and useful information can be derived from the data. Understanding the extent to which EEG can be used would be very beneficial, i.e. can EEG record enough speech information for it ever to be decoded accurately. MEG is also promising in that it has the potential to contain information on imagined speech.

There has also been a lot of success in the field with the use of various deep learning methods. The TRF, on the other hand, is the new standard in terms of linear methods. If it could produce the same success as more complex methods, it would be both easier to understand and use. There is some hope that it can be used to decode speech with the help of MCCA and a speech vocoder.

Chapter 3

Methodology

3.1 Overview of the Approach

This chapter will provide a comprehensive look at the design and implementation aspects of the project, i.e. why certain things were done and how exactly they were executed. First, the chosen datasets are outlined. Following this, each section represents a step in the process.

The first step involves preparing the stimulus audio files by using the vocoder to transform them into vocoder features. The next step concerns preprocessing the EEG data using methods such as filtering and resampling. Once the EEG data is ready, MCCA transforms the data into a representational single subject with the 'best' information. The modelling step involves running a multivariate TRF (mTRF) model for each stimulus vocoder feature respectively. The final step begins once the optimal parameters are chosen for each model. The final models are used to get predictions for the three stimulus features, which are then input into the vocoder to produce speech. The final section describes how these steps are used on the MEG dataset.

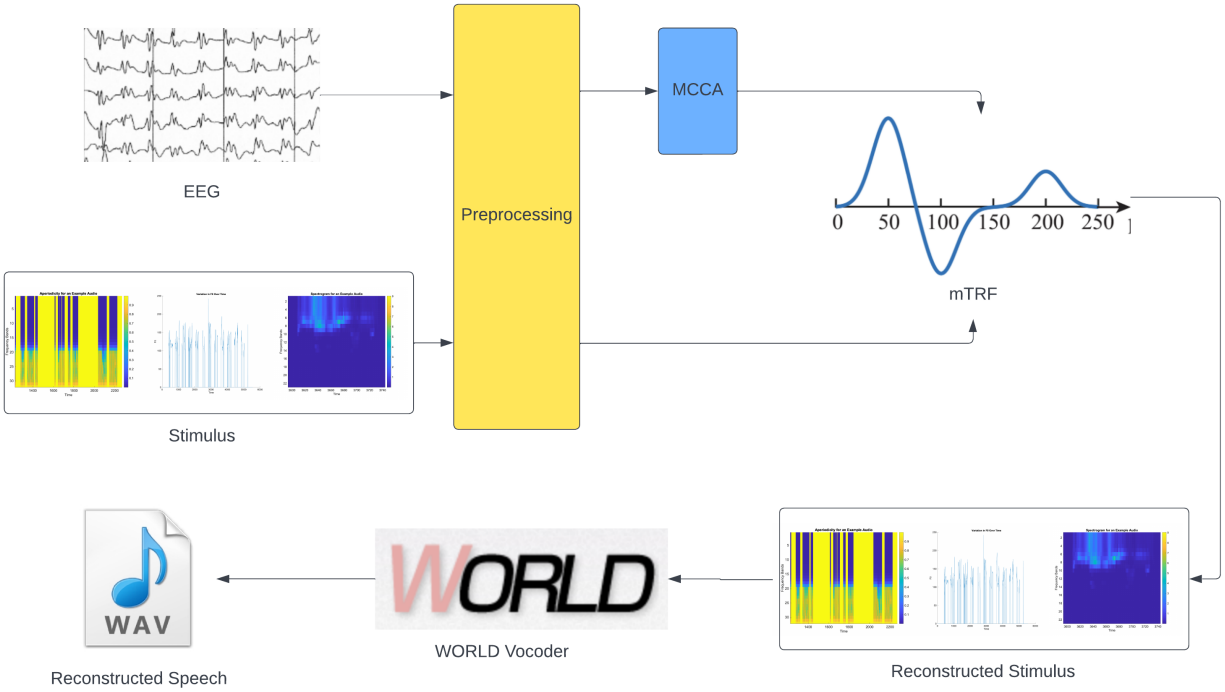

Figure 3.1: The overall workflow of this project at a glance.

3.2 Datasets

3.2.1 Dataset 1: EEG

The EEG dataset used is that of Broderick et al. (2017) and contains brain data for 19 different subjects. 20 trials were conducted for each subject, with each trial playing a different piece of natural speech for the subject to listen to. For each trial, 128 electrodes called 'channels' were placed around the subject's scalp to record data. As such, the data for each subject consists of 20 trials, each of which has the dimensions of time by channels. The sampling frequency of the trials is 128Hz so there are approx. 23040 samples for each trial, i.e. the size of each trial is roughly 23040×128.

Each audio was an approx. three minute snippet from the audio-book of The Old Man and the Sea by Ernest Hemingway. The WAV files themselves have a sampling frequency of 44100Hz. The dataset also contains extra information on each audio, namely its speech envelope vector and speech onset vector sampled at 128Hz.

3.2.2 Dataset 2: MEG

The MEG dataset used in this project is preliminary unpublished data, fully anonymised, that was freely shared as part of the CogHear workshop. The full version of the dataset will be published in the coming months (Razaeizadeh et al., TBD). The data we are using consists of five subjects who listened to one poem 10 times and another poem 10 times. There is also corresponding data for each subject where instead of listening to the audio, they imagined it. This means there is a total of 40 trials per subject, 20 per mode (listening or imagined). The sampling frequency is 1000Hz and there are 157 MEG channels along with three reference channels. There is a file for each subject which indicates which poem and mode each of their trials correspond to.

Similar to the EEG dataset, continuous natural speech was used. The poems were approx. 27 seconds long each. The poems were read out along with a metronome, and follow a regular rhythmic pattern. Considering this, it should be easier to predict F0 for this dataset as the silences will be regular and predictable. Like with the EEG dataset, the speech onset vector for each poem is provided as part of the dataset. The WAV files for the stimulus audios also have a sampling frequency of 44100Hz.

3.3 Speech Vocoder

The WORLD Vocoder (Morise et al., 2016) can both estimate the F0, spectrogram, and aperiodicity of an audio as well as generate speech with only these three estimated parameters as inputs. These three features are what connect the models and brain data to the actual speech. This particular vocoder was shown by Morise and Watanabe (2018) to be superior in terms of sound quality of the synthesized speech to other vocoders in the field.

3.3.1 Creating Stimulus Features From Audio

In order to predict speech from EEG, it needs to have training and test output data to have as a goal. The 20 *The Old Man and the Sea* audios are put through the vocoder to produce F0, spectrogram, and aperiodicity values for each. The process to achieve this will be outlined in this section.

First, the audio files need to be ordered alphabetically so that they match the order of the EEG data trials. Otherwise, there will be mismatches in the training data and the models will not learn well. Each audio is then processed by the WORLD vocoder to produce the features. The results are an array of F0 values, a matrix of spectrogram values, and a matrix of aperiodicity values. Each of these is automatically sampled at 200Hz by

the vocoder. Following this, the features are formatted so that they will work properly in Matlab, which is where the models are created and trained. This involves creating a data structure that matches the Continuous-event Neural Data structure (CND) format (Di Liberto and Nidiffer, 2021), which is also how the EEG data is structured.

Due to resource constraints, the large spectrogram and aperiodicity matrices have to be reduced in size. The initial dimensions are *time* × 1025, with 1025 representing the range of frequency values split into 1025 bands. This is reduced to 32 bands using the Greenwood equation (Souza and Rosen, 2009) and the code below. The Greenwood equation provides us with the *lower* and *upper* cutoff frequency values for each of the 32 new bands.

For each new band, the old bands in the original spectrogram that contains these lower and upper frequency values must be found. They are found by dividing the *lower* and *upper* frequency values by the step size. The step size is how many frequencies each band in the original spectrogram contains, i.e. max_frequency / 1025. The values between the old bands are then averaged and the result is assigned to the new band value, which encompasses the frequencies between *lower* and *upper*.

```
num_samples, old_filters = data.shape
new_mat = zeros((num_samples, new_filters))
for i in range(0, num_samples):
    sample_data = data[i, :]
    for j in range(0, new_filters):
        bottom_filt = floor(lower[j] / step_size)
        top_filt = floor(upper[j] / step_size)
        if top_filt > old_filters:
            top_filt = old_filters
        vals = sample_data[bottom_filt:top_filt+1]
        avg_freq = np.sum(vals) / len(vals)
        new_mat[i, j] = avg_freq
```

Listing 3.1: Reducing the number of bands from *old_filters* (1025) to *new_filters* (32) for a single trial, *data*.

Once the filters have been reduced, the F0, spectrogram, and aperiodicity values for all of the trials are saved into a Matlab structure to be used in the models.

3.3.2 Creating Audio From Stimulus Features

Once the mTRF models are finalised, predictions can be made for each audio's reconstruction in the form of F0, the spectrogram, and the aperiodicity. Before plugging them

back into the vocoder to recreate speech, they first need to be resampled back to 200Hz. Following this, the number of bands needs to be expanded back from 32 to 1025. This is done in a very similar manner to when reducing the filters, except the other way around. For example, instead of averaging the values across multiple bands, the value of one band is spread out and used for multiple bands.

Naturally, this means that in the overall process of reducing and then re-expanding the number of bands, inaccuracy is introduced. However, it does not significantly affect the results, e.g. the real audio put through the vocoder twice sounds very similar to what it did initially, just with a minimal amount of noise added. Once the number of bands has been expanded back to 1025, the features are put into the vocoder and a WAV file of speech is produced for each audio.

The speech reconstructions have trouble predicting silence so one option is to modulate the output sound by its amplitude envelope which cleans the noise somewhat. There is also the option to cheat and use the original audio's envelope in order to see what the best possible audio can be achieved envelope-wise, i.e. an idea as to what information we do have and what is missing. Another optional step is to scale F0 before putting it into the vocoder. The predicted values may have a smaller range than a human voice typically has, or go down into the negatives. Scaling it with, for example, the subject's F0 range can make the final audio sound more realistic.

3.4 Data Preprocessing

This section will outline the data preprocessing completed regardless of which model is being created. Extra steps may be needed depending on whether an individual, average, or MCCA model is being trained.

3.4.1 Stimulus

The mTRF has been shown to be robust to noise so we can minimise the amount of preprocessing done, particularly to our output data. Less is better as the more steps that are added, the more room for error and damage done. The only preprocessing that is done to the stimulus at this point is to resample it from 200Hz to 32Hz. The sampling rate is reduced in order to facilitate a more efficient workflow and to stay within resource limits. Nearest-neighbour interpolation is used for F0's resampling so as to keep as many 'silent' 0 values as possible. Without this, the trained models are worse at predicting moments of silence.

Initially, the sampling rate was reduced to 128Hz as that was the sampling rate of

the EEG. However, there were multiple issues with this. First, the file size was too large for Matlab to load so the stimulus file had to be split into four, which makes the process more complicated. Secondly, running the model took approximately half a day per feature, which is unreasonable for making iterative improvements. A sampling rate of 32Hz allows all of the data to be stored in one file and for the models to be trained much more quickly, allowing for fast iterations to improve results.

3.4.2 EEG

The EEG data preprocessing requires multiple steps. First, the data is run through a low-pass filter which attenuates any EEG values above 8Hz. The data is then also downsampled to 32Hz. After this, a high-pass filter is applied to the data which attenuates values lower than 1Hz. Any bad channels are then removed from the data and replaced with a spline interpolation of the other channels. Finally, the data is re-referenced based on its average. These are industry standard preprocessing steps, all of which done using the mTRF toolbox (Crosse et al., 2016).

3.4.3 Stimulus-EEG Coordination

Before using the stimulus and EEG data in our models, we make sure that they have the same number of trials and that their sampling rate is the same. The length of each trial also needs to be the same for the stimulus and EEG so one often needs to be cut short slightly. The EEG is also normalised using its standard deviation, while nothing is done to the stimulus in this regard. The stimulus values are kept as-is as our output is put into a vocoder so the actual values themselves matter and should be as realistic as possible. If the values are normalised, it affects F0 in particular as the predicted values are quieter and do not vary as much.

3.5 Models

3.5.1 Individual Subject

The generic case for training a model is that of using a single subject's data. Each of the three stimulus features for this subject have their own models that are trained separately from the other two. For each feature, the subject's EEG data is used as input and the stimulus feature is used as output.

The mTRF has two modes, a forward encoding mode and a backward decoding mode. Encoding entails predicting the EEG brain data from the stimulus, while decoding in-

volved predicting the stimulus from the EEG brain data. As such, we use the backward direction. Another parameter that needs to be chosen is the time lag range. Time lags are used in order to correctly match the EEG samples to the stimulus samples as there may be a delay between a subject hearing something and their brain processing it. The model will find the optimal lag values between the range given and use that. For this model, the minimal lag was set to -200 and the maximum lag was set to 600. This means the model accounts for the relationship between the stimulus and the neural response up to 200 ms earlier and 600 ms later. This wide range was used so as to make sure we did not miss information.

The final parameter is lambda (λ), the scalar value that controls how much regularisation impacts the model. An array of different λ values are provided to choose from, ranging from $1e-6$ to $1e4$ and the optimal value is chosen using cross validation. Cross validation is run for each trial/audio. For each trial, the data is comprised of all the training data minus that of the current trial itself, i.e. leave-one-out cross validation. The data is used in an mTRF cross validation function that also implements a leave-one-out process of its own to find the best lambda and its corresponding r value. The r value is a correlation coefficient based on Pearson's linear correlation coefficient. In this way, for each of the 20 models created we have a train and test set within the cross validation as well as a validation set of the current trial.

After cross validation is done for each model, its best λ is chosen and the actual model is trained using the chosen parameters and saved. We then test each of the models on its validation set and get a prediction and corresponding r value for each trial. These predictions are what are eventually put back into the vocoder, while the r values are used for evaluation.

3.5.2 Average Subject

In order to have a middle-ground comparison between the MCCA model and an individual subject's model, we also decided to create a model based on the average of the subjects. This should remove some of the noise that is in the EEG data as some subject's outliers will cancel each other out and so on. Technically, another option would have been to just average the predictions of the 19 subjects after their individual models had been trained. However, that is not a model but more of a postprocessing step. Furthermore, averaging the subjects beforehand allows for a more direct comparison with MCCA, as both create a single subject from the data and create a model from it. The process is straightforward, the data for each trial for each feature is averaged across subjects. This data is then used in the same way as for the individual subject.

3.5.3 MCCA Subject

As previously mentioned, EEG data is noisy and speech is complex. Due to this, we need to find a way to de-noise the data and find its most useful parts. We can use the fact that there is data for multiple subjects to try and achieve this. MCCA allows for this in a manner that is more intelligent than just averaging the subjects. It creates an ideal single subject by extracting the components that are most common to all of the initial subjects.

The Subject Matrix

First, all of the subjects' data must be combined into a single matrix. For technical reasons, the sample length of each trial in this matrix must be the same in order for MCCA to work. The length of the smallest trial is found and all of the trials are shortened to be this length. Once the trials are the same length, they are put into a Matlab data structure which has the dimensions of 19×20, i.e. subjects by trials. Each 'value' in this structure is actually a $min\_trial\_length \times 128$ matrix. The samples for each trial are then concatenated so that the overall matrix now has the dimensions of $19 \times (20 * min\_trial\_length) \times 128$.

MCCA Calculations

The MCCA process itself can now begin. We go through each of the 19 subjects, one by one, and get their data from the matrix. This will have the dimensions $(20 * min\_trial\_length) \times 128$. PCA is run on this data, with the second input being the number of principal components you want to use, $num\_components$. This retrieves the top $num\_components$ components across the 128 channels and does matrix multiplication on the input data using them. The results are stored for each subject in a second matrix M with dimensions of $19 \times (20 * min\_trial\_length) \times num\_components$.

Once the results for all of the subjects have been stored, the matrix M is transformed so that it now has the dimensions $(20 * min\_trial\_length) \times num\_components \times 19$. Using M as input, calculate its time shift covariance matrix C. In this case, the $(20 * min\_trial\_length)$ is considered the main data, the $num\_components$ components are considered the time shifts, and the 19 subjects are the weights.

Finally, run MCCA with C and $num\_components$ as the inputs to get a transformation matrix A. Compute the matrix multiplication of M and A, which produces a matrix with the dimensions $(min\_trial\_length * 20) \times A\_dimension$, where $A\_dimension$ is the dimension of the square matrix A. This transformation represent using the components with the highest correlation across subjects. The final step is to take only the values for the first $num\_components$ MCCA components from this matrix, leaving you with

a $(min\_trial\_length * 20) \times num\_components$ matrix. Seeing as the number of initial channels was 128, we chose to take 128 MCCA components so that comparisons can be easily made.

A Single Subject

All that is left to do before this data can be used in an mTRF is to separate it back out into its 20 trials and put it into an appropriate CND data structure. This is straightforward as all of the trials were shortened to be the same length. The EEG data structure for one of the initial subjects can be used as a base, with the new MCCA data being inserted instead. An optional extra step is to find the best components of the 128 and only use them in the model. The process as to how exactly the best components were found will be outlined in the Results section. For this project, 16, 7, and 64 components were used for the F0, spectrogram, and aperiodicity models respectively.

3.6 Application on MEG Data

The majority of the process remains similar when using the MEG dataset. However, some extra steps need to be taken in order to make it compatible with our workflow. The stimulus audios are the same vocoder-wise, except that there are just two of them rather than 20. There are, however, 10 trials for each audio, both for the listening data and the imagined data. In order to make the dimensions match, we recreate the stimulus feature data structures and simply duplicate each stimulus 10 times, so that there are now 20 'audios' in total.

3.6.1 Extracting MEG Data Epochs

The process for the MEG data itself is somewhat more involved. Each mode, 'listening' or 'imagined', is processed separately. For a given mode such as 'listening', we iterate through the five subjects. The MEG data for that subject is read in as well as the CSV file that indicates which trial concerns which mode and poem audio. Using this CSV file, we find the trial numbers for the 10 'poem 1' trials and the 10 'poem 2' trials. The trial numbers are put in an order array such that the 'poem 1' trial numbers are at the beginning and the 'poem 2' trial numbers after them. This is done so that when the MEG data is eventually ordered using these trial numbers, it matches the ordering of the stimulus in terms of 'poem 1' and 'poem 2'.

It is important to note that the raw MEG data we are using is not yet separated into trials. All of the trials are currently concatenated into one long piece of data. There is a

channel within the MEG data that is not an actual MEG channel, but instead a timing channel that tells us when each trial starts i.e. at which sample. We found this channel manually for each subject in advance. It is the 183rd channel for all subjects except for subject R2383, where it is the 173rd channel. The timing channel is made up of all 0s except for a number of 1s which represent when a trial begins. There are actually 80 1s in total because there are 80 trials in total. This is because, along with the 40 trials association with the two poems in listening and imagined modes, there are two other music-related audios. However, these extra 40 trials are not relevant to this project.

Once the timings are retrieved for the current subject and mode, we can remove all of the channels past the first 157 channels, as they are not MEG channels. The timings and order calculated can now be used to split the data up into trials. The following code encompasses how this is done.

```
function [trials] = split_data(data, times, order)
    trials = cell(20,1);
    for i = 1:size(order,1)
        start = times(order(i));
        finish = size(data,1);
        if order(i) ~= size(times,1)
            finish = times(order(i)+1)-1;
        end
        curr_data = data(start:finish,:);
        trials{i} = curr_data;
    end
end
```

Listing 3.2: Epoching MEG data given the order and timings of relevant trials.

For each trial number value in the order array, we find the sample at which it begins. The timing array has all of the trials in order so the trial number is simply used as an index. For example, the first value in the order could be 12, indicating that the first trial concerning 'poem 1' for this mode is the 12th trial of 80. In the timing array, the 12th value is 5000. This means that the 12th trial begins at sample 5000. In order to find out when the trial ends, we check the timing array to see when the 13th trial begins. If, for example, the 13th trial begins at sample 6000, this means that the 12th trial ends at sample 5999. This is how we find the beginning and end samples for each trial. The sample numbers can then be used to cut out the trial data and store it. There is one edge case where the trial number in the order array is the 80th and final trial, in which case

we automatically know that it ends at the final sample rather than at a hypothetical next trial's beginning.

3.6.2 Further Adaptation

Once the trial data is found and stored in the correct order, any bad channels are removed. Channels 86 and 65 are removed from all of the MEG data. As the data needs to have the same dimensions, if one subject removes a channel then all of the subjects will also need to remove the channel. This could be improved in later iterations by simply re-referencing a bad channel rather than removing it altogether. Once this is done, the data is restructured to match the CND format of the EEG dataset. The EEG stimulus data structure can be used as a basis.

The sampling frequency of the MEG data is initially at 1000Hz. This is downsampled to 100Hz for resource and timing reasons. The stimulus is also resampled to 100Hz. Past this point, the process is generally the same as with the EEG dataset. Some parameter values are changed, such as to facilitate the number of subjects being 5, not 19. The lower and upper bandpass ranges were also increased from 1 and 8 to 0.1 and 35 as this wider range for MEG significantly improved results.

One alternative method, however, is the 'averaging' method. This means that, instead of repeating each stimulus feature 10 times, repeat each poem just twice, totalling four 'audios'. Then, take the average of the first five MEG trials for 'poem 1' to be the 'first' trial, the next five MEG trials to be the 'second' trial, and so on. This means there are four MEG trials, each somewhat denoised. This takes advantage of the fact that multiple trials use the same stimulus. The reason we cannot simply average all 10 trials for each audio is that we need enough trials for mTRF cross-validation to work well.

Chapter 4

Results

4.1 Overview

This chapter will provide a detailed account of the evaluation conducted for this project. The results of this project, the model predictions and the reconstructed audios made using them, will be analysed. The first section discusses the results when using the EEG dataset. The focus here is on the MCCA model and how to improve results in general. The second section involves looking at the MEG dataset, particularly the data where a subject imagines the audio, and how applying our developed methods works on it. The analysis of these results and the iterative improvement of methods and results bases on analyses was a large part of this project. We hope that the outputs can be used as a baseline for future research.

4.2 EEG Dataset

4.2.1 Stimulus Predictions

The main metric used to measure the reconstruction success is an R correlation value based on Pearson's linear correlation coefficient. This was used both for the cross-validation to choose optimal parameters as well as general evaluation of the final results. This metric measures how correlated the reconstructed feature is to the actual stimulus feature, with values ranging from -1 to 1. A value of 0 indicates no correlation, 1 indicates that there is a purely positive correlation, and -1 means there is a purely negative correlation. A good result would be a high value such as 0.8, while small values close to 0 mean our reconstructions are not performing well.

For each feature and each audio, the R correlation was averaged for our models. The models consist of an individual model, an average model, an MCCA model, and a dummy

'random' model. This dummy model was trained on trials put in a random order so should not perform well and is used as a point of comparison. It can be seen in Figure 4.1 below that all of our properly trained models are significant improvements on the random one. This is a good sanity check as it shows that the models are picking up at least some valuable information and there are some successful patterns to follow. It is also clear from this figure that MCCA is by far the best model and gives significantly better results in comparison to the other models. Results are generally double that of the individual model for all features. However, MCCA results of 0.19627, 0.1937, and 0.12734 for the aperiodicity, F0, and spectrogram respectively are still far from perfect. This is particularly true for the spectrogram, which makes sense as it is a complex feature.

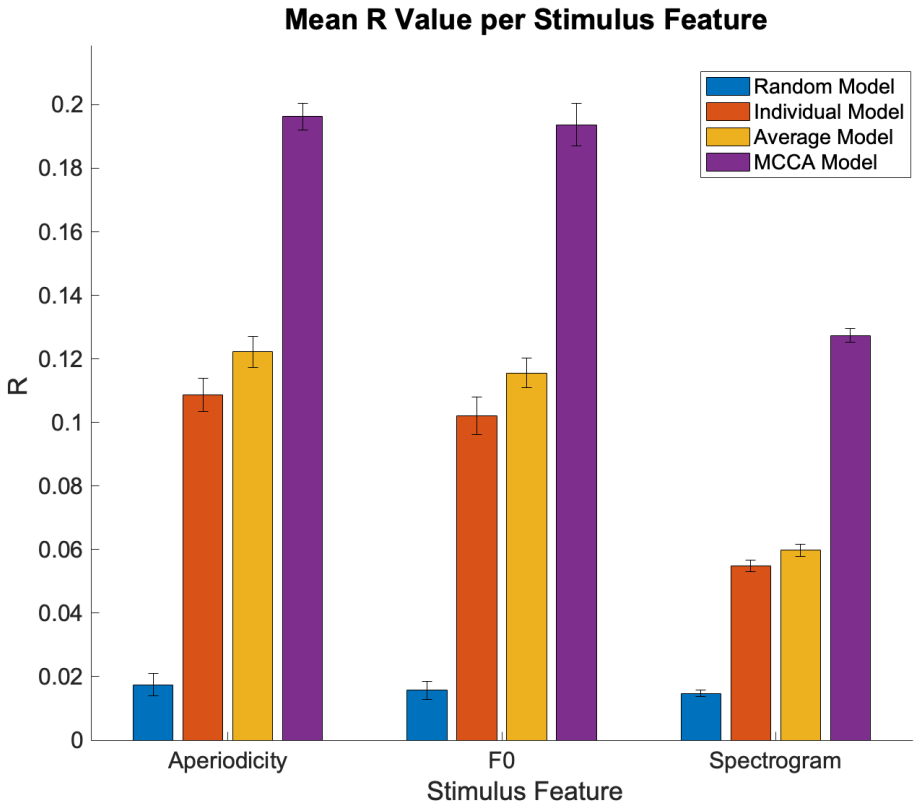

Figure 4.1: A comparison of EEG R values based on different models for each stimulus feature. The values are the mean R correlation over 20 trials. The standard error is also shown for each bar.

A paired t-test was conducted to confirm these results. The mean R value across frequency bands for each trial was compared for the average model and the MCCA model. This means that, for each feature, 20 corresponding R values were used to compare the two models. The t-test's null hypothesis is that the data in X, the average model, and Y, the MCCA model, comes from a normal distribution with a mean equal to 0. The null

hypothesis was rejected for all features, meaning that the p value was less than 5% and our results are significant. MCCA does improve on the baseline of using the average data over subjects.

One important lesson learned during this evaluation stage was that the order of the preprocessing steps is very important. At first, MCCA was run on the raw data and this resulting MCCA subject was put through the general preprocessing pipeline, i.e. resampling, removing bad channels, etc. However, this resulted in R values that were the same as or lower than the individual model's R values and worse than the average model for all features. The R values were 0.1191 for the aperiodicity, 0.1101 for F0, and 0.0598 for the spectrogram. Clearly, preprocessing beforehand is important for cleaning the data and making it more uniformly comparable. This makes it easier for MCCA to find correlated components across subjects and channels. The final pipeline involves preprocessing the data before applying MCCA, which resulted in the MCCA R values shown in Figure 4.1. Interestingly, normalising the stimulus during preprocessing does not affect the R values despite it significantly affecting how the eventual vocoded audio reconstruction sounds.

Another important step in achieving these final MCCA results was picking the number of components used in each feature's model, as well as which specific components. It is likely the case that the top MCCA components are the best components to include in the model. When MCCA returns its results, the components are in order from highest to lowest in terms of correlation across subjects. However, we wanted to confirm this as well as see which features were most related to which components. Models were run for each feature with each of the first 16 components, i.e. 16 models with each based on one component. The R value results for these models can be seen in Figure 4.2.

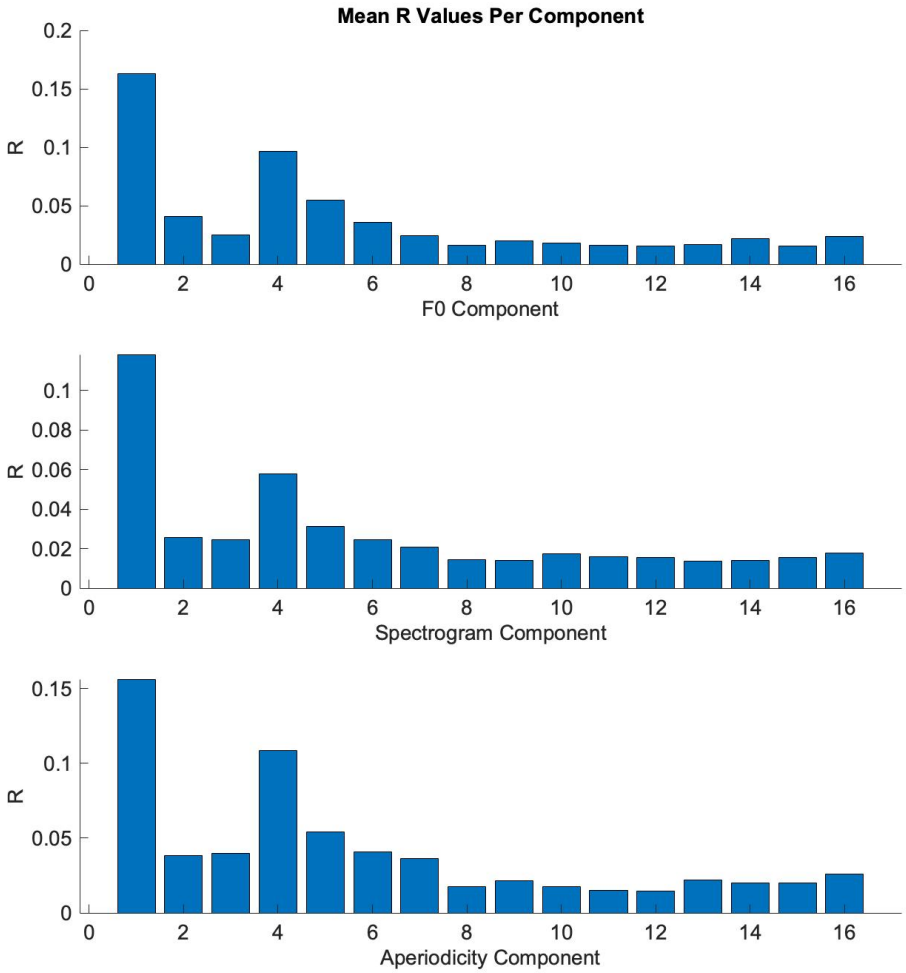

Figure 4.2: R value results for models trained on individual MCCA components.

The results show that the first block of components seem to give the best results, with the R values petering out after around the seventh component. This confirms that, generally, the 'best' components translate to better models. It is also interesting to note that the first and fourth components produced the best models for all three features. Furthermore, the first component alone had a very high R value, especially considering it was just a single component. There is not a huge difference between the final model and this model with just the first component, showing that this component is very impactful.

Considering these results, a number of different combinations of components were tried out for each feature. For models with 16 or fewer components, the components were chosen based on which single-component models produced the highest R values. For models with more than 16 components, the first n components were simply chosen as it is

unlikely to have made a significant difference manually choosing them. A comparison of the R values for the resulting models can be seen in Figure 4.3. The optimal number of components for F0 was 16, for the spectrogram it was 7, and for the aperiodicity it was 64. If a smaller model is needed for the aperiodicity, 16 components can be used instead of 64 as there is just an approx. 0.004 difference in value between the two.

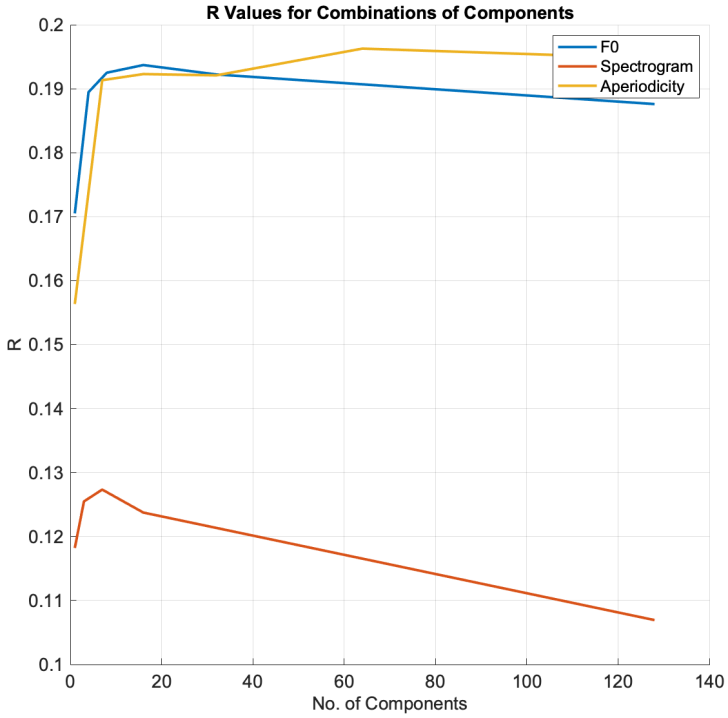

Figure 4.3: R values for different numbers of components used in an MCCA model. This is used to choose the optimal number of components to include in each feature's final MCCA model.

Further evaluation using a sampling frequency of 64Hz instead of 32Hz did not significantly improve the results. This, combined with how much more resources and time it takes to run this larger model, led to the continued use of 32Hz instead. Another approach that evaluation showed to not significantly improve things was to train separate models for each band in the spectrogram. This would mean that a different lambda can be chosen for each band, possibly allowing for a more accurate prediction. This did not turn out to be the case, R values were the same as or worse than the current MCCA model.

We also tried to find the audio that was best reconstructed based on R value and then also find the best segment of that audio. This was in the hopes of finding a particularly well reconstructed piece of audio. However, while a section with high R values was found, e.g. average spectrogram R value of approx. 0.22 in comparison to 0.13, it did not result

in intelligible speech. The speech metrics also did not show any improvements.

It is also important to note that models have a hard time predicting zero values. While the actual stimulus F0 values drop to 0 often, i.e. whenever there is silence, the reconstructed F0 values do not. Furthermore, it is hard to say whether the model and predictions are trying to accurately represent the changes in F0 values or are just differentiating between low or zero F0 values and higher values. In order to investigate this, we also calculated the R value solely based on samples which did not have zero values in the original stimulus. While this did lead to a 15% decrease in R value, this still leaves an R value of approx. 0.16 which is not an extreme drop. According to a paired t-test, this is still a significant improvement on the average model. These values show that the model is actually attempting to predict the change in F0 values rather than just differentiating between silence and sound.

A comparison of a sample of the stimulus F0 along with its corresponding MCCA reconstruction can be seen in Figure 4.4. As mentioned previously, one of the main differences is that the reconstruction does not drop to 0 during moments that are supposed to be silent. The general mean for both remains at approx. 50Hz but the reconstruction deviates much less, with smaller peaks and troughs. While there does seem to be similarities between the two, it is clear that the reconstruction is not fully accurate and there is definite noise.

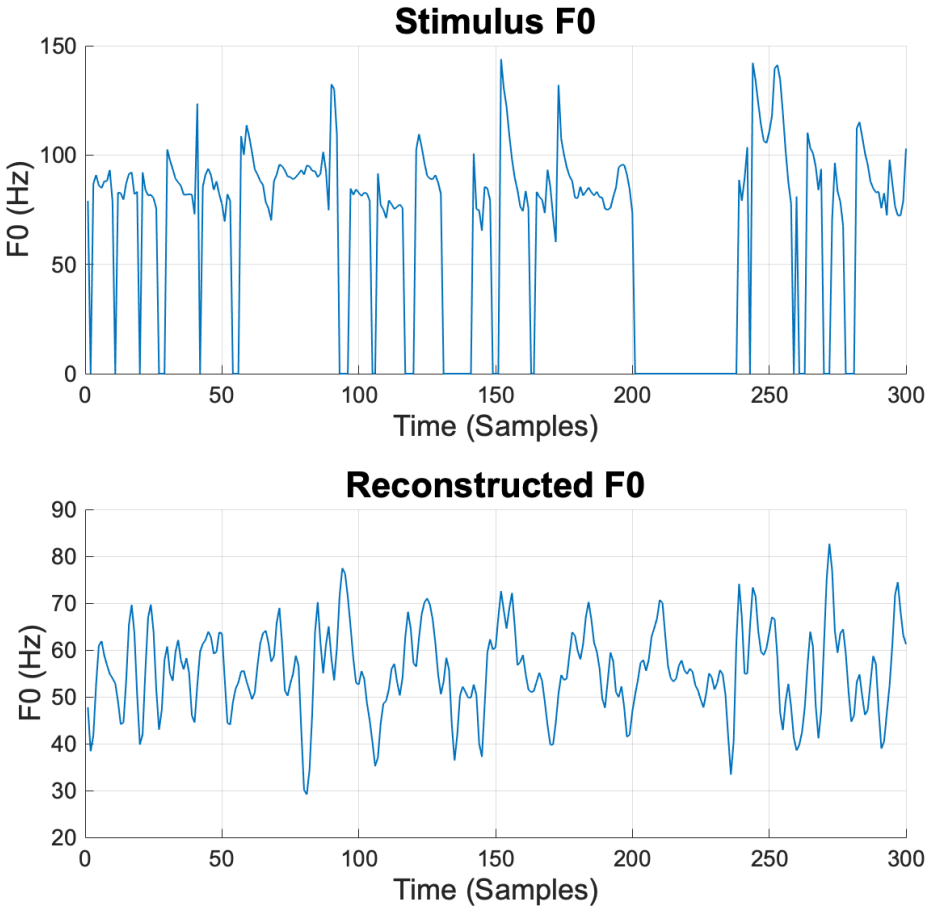

Figure 4.4: Comparing stimulus F0 with reconstructed F0.

A sample of the stimulus spectrogram along with its reconstructed counterpart can also be seen in Figure 4.5 below. It can be seen that the reconstruction follows the general flow of the audio, with higher amplitudes shown when the stimulus is non-silent. However, it can also be seen that the finer details of the stimulus are not being reconstructed. It should also be a point of note that negative values in the reconstruction were removed before creating this figure. The amplitude of a frequency cannot go below zero and does not do so in the stimulus. However, the model itself does not know this so can produce negative values. This can be remedied by either making all negative values zero or increasing every value in the reconstruction so that the most negative value is now zero.

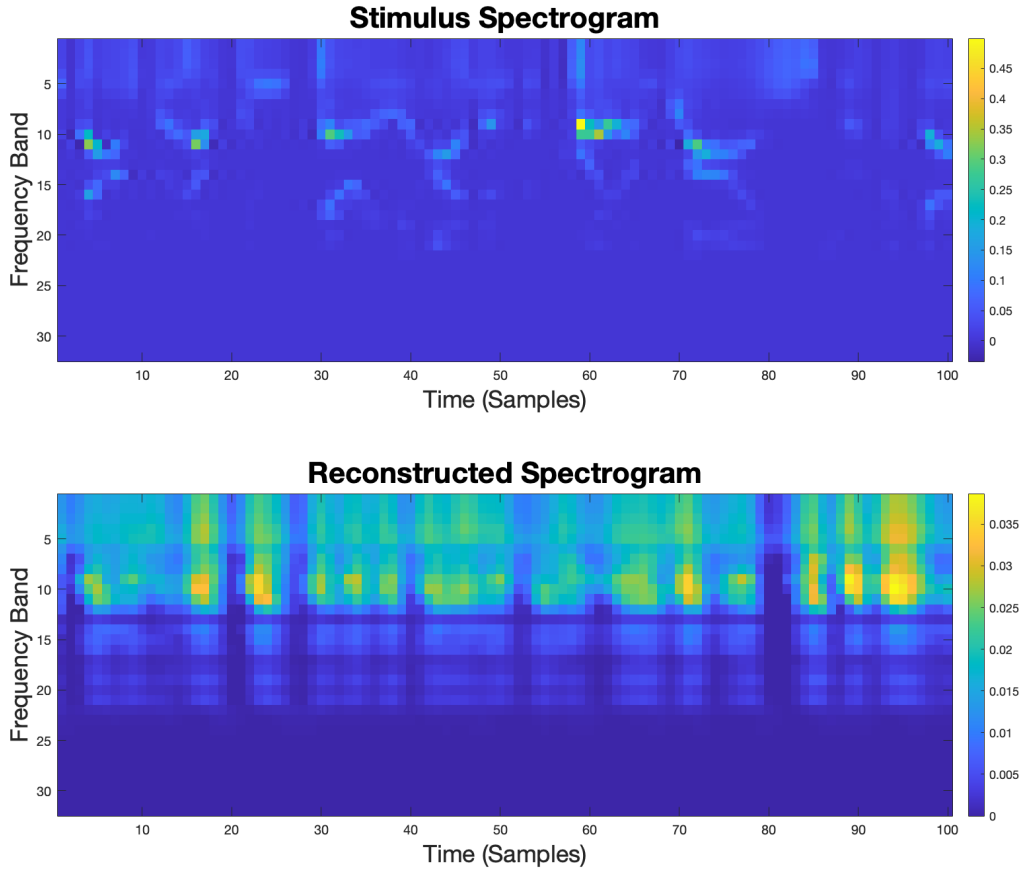

Figure 4.5: Comparing stimulus spectrogram with reconstructed spectrogram.

Finally, a sample of the stimulus and reconstructed aperiodicity can be seen in Figure 4.6. While not fully accurate, it is easy to see that many lines in the stimulus are recreated in the reconstruction. In particular, the higher frequency bands are quite accurate. However, the reconstruction tends to predict higher values more often than it should, such as between samples 60 and 80. This is interesting as it means unvoiced or silenced sound is being predicted too often, while the F0 and spectrogram reconstructions predict sound where there shouldn't be. As we will see in the following section, the eventual reconstructed audio is noisy and is not always silent when it should, once again showing that the aperiodicity does not have as large of an impact on the audio as the other features.

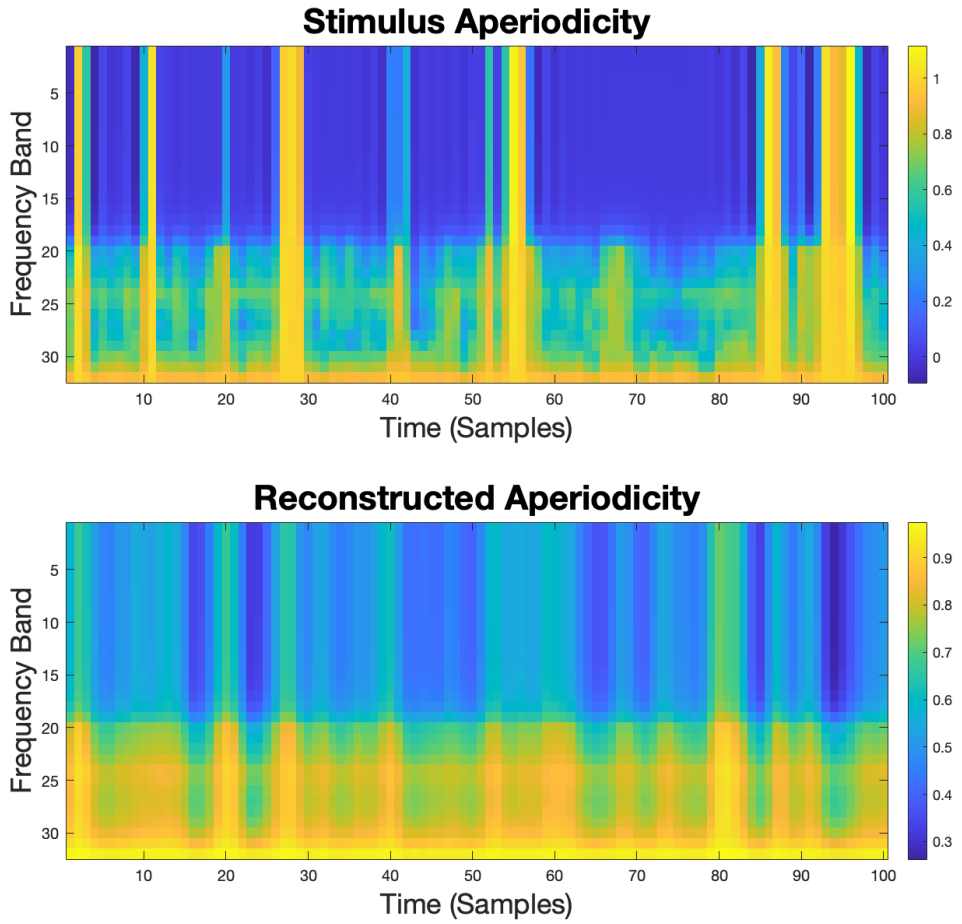

Figure 4.6: Comparing stimulus aperiodicity with reconstructed aperiodicity.

4.2.2 Audio Produced

To the subjective human-ear, the speech reconstructed by the vocoder using our method was not intelligible. It was unlikely that the initial results of this methodology would be successful in that regard, seeing as we are still trying to find out whether EEG can even encode enough information on speech for it to be possible. However, certain speech intelligibility and quality metrics can be used to analyse what is and is not working reconstruction-wise.

The main metrics used were the Short-time Objective Intelligibility (STOI), the Frequency weighted Segmental SNR (fwSNRSeg), and the Cepstrum Distance Objective Speech Quality Measure (CD). STOI is a speech intelligibility metric based on the linear correlation between speech temporal envelopes. This means that the larger the value is for the reconstructed speech, the better. CD and fwSNRSeg, on the other hand, are

speech quality metrics which measure things such as the signal-to-noise ratio. A positive fwSNRseg value means that the signal level is greater than the noise level, so a high ratio would indicate better signal quality. CD, being a distance metric between the two audios being compared, gives low values for similar audios and higher values for audios 'distant' from each other.

First, it is important to note that due to the downsampling of the data to 32Hz and then subsequent upsampling back to a higher frequency, data is lost and noise is introduced. The best possible result will never be of the same quality as the original audio because the stimulus itself is no longer at that standard. In order to have a fair base value to compare our results to, the goal stimulus features themselves were upsampled back to 200Hz and put into the vocoder to produce speech. Subjectively, this speech was still completely intelligible but there was an obvious difference, with an extra 'fuzzy' sound to the audio as well as some slurring of speech. The objective metrics were also retrieved for the speech that was produced and can be seen in Table 4.1 below.

STOI	fwSNRSeg	CD
0.86723	5.35199	6.9537

Table 4.1: Objective metrics for 'The Standard' of reconstructed speech. These results comes from getting stimulus features from the vocoder, downsampling them to 32Hz, resampling them to 200Hz, then putting them back through the vocoder again as inputs.

While the reduction in the number of bands in the spectrogram and aperiodicity also add to the reduction in intelligibility and quality, the sampling frequency has a significantly bigger impact. This was proven by creating stimulus features with (1) no downsampling or band reduction, (2) just downsampling, and (3) just band reduction. Each was put back into the vocoder to produce speech and evaluated. It can be seen in the table below that while the band reduction does slightly decrease the quality, downsampling is the root cause. Unfortunately, downsampling is necessary due to the time and resource constraints mentioned previously.

	STOI	fwSNRSeg	CD
No Compression	0.94505	14.84775	2.24497
Reduced Bands	0.92519	11.78009	4.97064
Reduced FS	0.87966	6.12806	5.62898

Table 4.2: Comparing the reduction in speech intelligibility and quality of compressing the stimulus bands and sampling frequency. No compression means that nothing is done to the stimulus, it is created from the vocoder and immediately put back through it to produce speech.

Now that we have the goal values from Table 4.1, the speech produced by this project's models can be evaluated in context. These models include the individual, average, normalised-stimulus MCCA, and non-normalised-stimulus MCCA. A dummy model, created by training an individual model on the trials put in a random order, was also used as a sanity check baseline. It can be seen from the values in Table 4.3 below that, with the exception of MCCA, there does not seem to be any significant difference between the majority of the results. A summary of these results can be seen in Figure 4.7, where the results have been normalised to range between 0 and 1 for ease of comparison. The dummy baseline has better STOI and CD results than the individual and average models, and has a better fwSNRSeg result than all of the models except for non-normalised MCCA. It is clear that keeping the original stimulus values, i.e. not normalising them, matters.

	STOI	fwSNRSeg	CD
Random	0.38096	0.39607	7.72078
Individual	0.37229	0.1967	7.79697
Average	0.37353	0.22224	7.84664
Normalised MCCA	0.39105	0.18571	7.7501
MCCA	0.40497	0.6048	7.78162

Table 4.3: Comparison of model predictions in regards to speech intelligibility and quality metrics. The difference between Normalised MCCA and MCCA is that the latter did not involve normalising the stimulus training values.

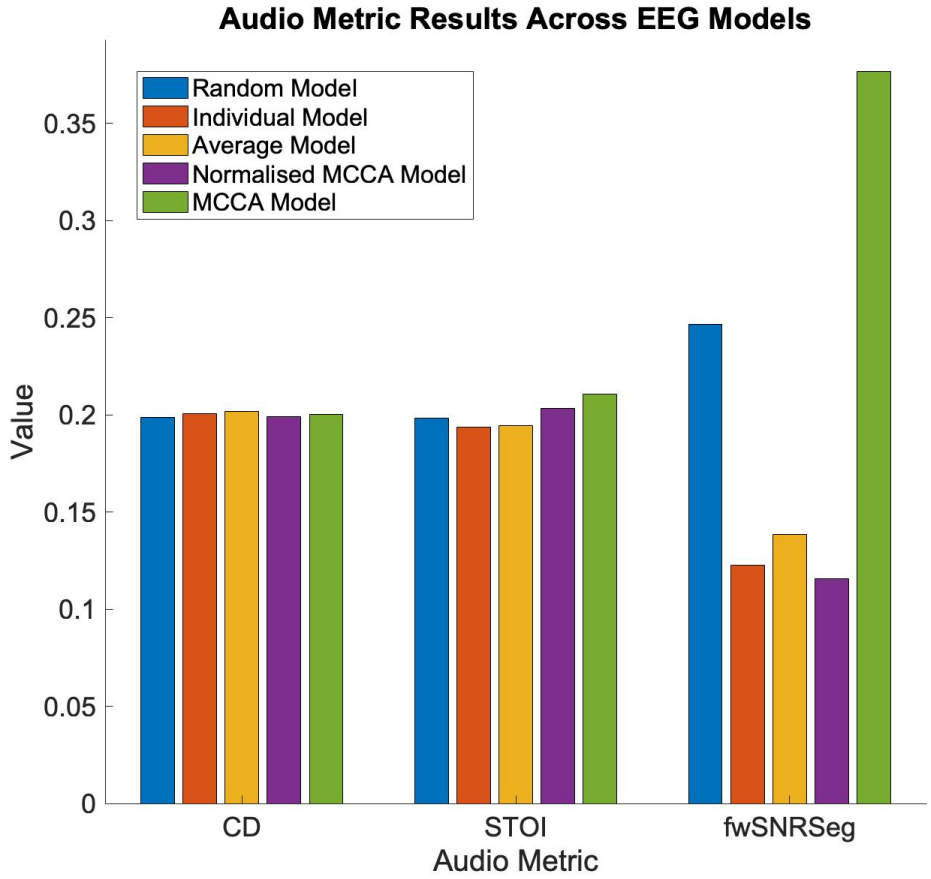

Figure 4.7: Comparison of model predictions in regards to normalised speech intelligibility and quality metric results.

These results imply that the reconstructed speech may not be good enough or close enough to real speech to be accurately evaluated using these metrics. These metrics are meant to be used on degraded speech but if the inputs do not even sound like speech then it is difficult to evaluate them in this regard. The lack of normalisation for the MCCA model makes a significant difference. It is the only model that performs better than the dummy in two metrics. There is also a distinct difference in subjective sound in comparison to the rest of the models' reconstructed speech. While the other models produce sound that is faint/vague, similar to the sound of, for example, ocean waves, the non-normalised MCCA audio has a noticeable buzzing-esque sound that seems to represent the speech. The wave-esque sound is also still faintly there but is no longer the main focus. The audio files are included in the appendix of this project for further listening.

Taking this non-normalised MCCA result as a baseline, some further processing was applied to see if we could improve on this, i.e. how good could the results be in optimal

circumstances. First, the F0 prediction was scaled according to the range of the real stimulus' F0. We also tried modulating the output sound by the real stimulus' amplitude envelope. The results, seen in Table 4.4, generally improve the speech intelligibility and quality metrics but can also make some worse.

	STOI	fwSNRSeg	CD
MCCA	0.40497	0.6047	7.78162
MCCA + Scale	0.41734	0.59059	7.72
MCCA + Env	0.47832	1.25655	8.02426
MCCA + Scale + Env	0.4888	1.21121	8.005

Table 4.4: Effects of scaling F0 and modulating the audio with the stimulus envelope of speech intelligibility and quality.

Again, it is important to note that with the speech so degraded, these metrics may not be the most reliable. However, assuming these results are accurate, the scaling and envelope modulation have different effects on the output speech. Both improve the STOI metric. Scaling the F0 values also improves the CD while making the fwSNRSeg slightly worse. It makes sense that fwSNRSeg does not improve as how high or low F0 is is not related to how pure or denoised the signal is. The envelope modulation, contrastingly, improves the fwSNRSeg but increases the CD value. This improvement in fwSNRSeg also makes sense as a modulated 'cheat' envelope should silence some of the extra noise in the audio. A combination of both somewhat rounds out the effects of both overall, but still leads to an improvement in STOI and fwSNRSeg and a degradation in CD results. Overall, a combination of both seems to give slightly better results than the other iterations.

An attempt to evaluate only the specific best segment of the best audio in terms of high R values was also made. However, the slightly elevated R values did not seem to translate into better speech intelligibility and quality metric results. The values ranged from similar to the MCCA results to slightly worse. Again, slight improvements like these may not be picked up using these metrics as the metrics themselves are too high-level.

Finally, we want to know which feature has the biggest impact on the speech vocoding, as well as which features are being the most and least well-predicted. In order to investigate these questions, we used the real stimulus features in combination with the predicted features. An audio was reconstructed focusing on each feature, with the current feature being replaced by the real stimulus feature and the other two features remaining as the prediction data. If this causes a huge improvement it would show that this feature is important and it being accurate and correct has a significant impact on the speech result.

Subjectively, the audio with the real spectrogram seemed to show the biggest improvement. With the real spectrogram and predicted F0 and aperiodicity, the speech reconstruction is intelligible. Inserting the real aperiodicity, on the other hand, did not seem to change much other than removing extra noise, or 'wave', sounds from the audio to make the buzzing sound the only thing audible. Replacing F0 was also noticeable but was still not intelligible, even though the intonation and pauses in speech were there.

	STOI	fwSNRSeg	CD
MCCA	0.40497	0.6047	7.78162
Real F0	0.44672	0.62096	7.45779
Real Spectrogram	0.76783	4.63118	7.38002
Real Aperiodicity	0.39763	0.56595	7.82307

Table 4.5: Speech intelligibility and quality results when replacing predicted features with real stimulus data.

The subjective results seem to also be reflected in the objective speech metrics, as shown in Table 4.5 above. Replacing the spectrogram with the real stimulus' spectrogram has by far the biggest impact. All three metrics improve significantly. The results are close to that of the standard, optimal values, by far the closest of any model thus far. It is clear that the spectrogram is very important and should be the focus of future improvements. The aperiodicity, on the other hand, does not seem to have much of an impact on the speech metric results, with no metric improving. F0 does improve all three results but nowhere near as much as the spectrogram does.

Overall, it is clear that the current results are not intelligible. However, we have also shown that MCCA, along with some additions, is a definite improvement over existing baselines such as individual subject models and averaged models. In terms of future improvements, the spectrogram is the key to getting better speech reconstruction intelligibility and quality and clearly is not yet being accurately predicted.

4.3 MEG Dataset

4.3.1 Stimulus Predictions

When it came to the MEG dataset, we had two goals. The first was to repeat the same evaluation done on the EEG dataset in order to confirm that our methods work for different datasets. Following this, we wanted to explore how the model performs with imagined data in particular. As such, the same metrics as for the EEG dataset are used,

i.e. Pearson's linear correlation coefficient and paired t-tests.

We want to prove that using the MCCA model is significantly better than using an average or individual model. Like with the EEG dataset, the R correlation results for each model show that this is true. However, this time, we not only have MEG data for when the subjects listened to the audios, but also when they imagined the audios. The same process was followed for both the listened and imagined MEG data in order to also be able to compare the two. The MCCA model was evaluated along with the individual, average, and random models.

There is also a fifth model included in the evaluation, the MCCA model with a reduced lambda value. Lambda dictates how much regularisation is used in the model. After some analysis of the MEG dataset prediction results, we found that sometimes when too much regularisation is used, the predictions became very generic and unchanging. For example, the spectrogram could have a single frequency, or a small range of frequencies, constantly have high amplitudes with the rest of the frequencies having small amplitudes. In order to prevent the model from over-generalising and always using the highest lambda value, we reduced the maximum lambda value and re-ran all of the models. An example of the difference this makes can be seen in Figure 4.8.

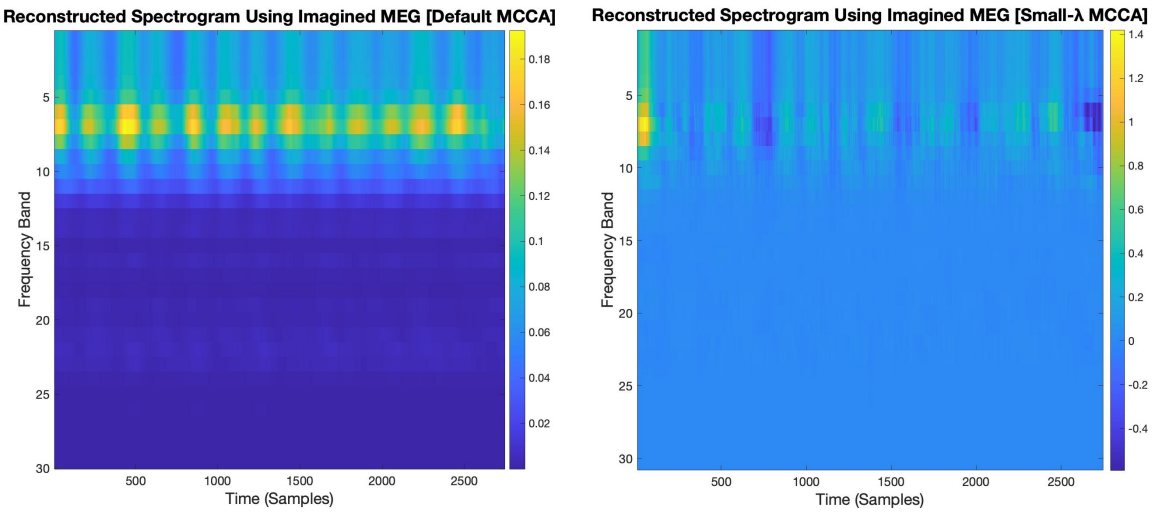

Figure 4.8: Spectrograms reconstructed using imagined MEG data with 32-component MCCA models. The first used a model with the highest lambda value possible. The frequencies with the most power are constantly within the approx. 6th frequency band. This rarely changes over time, showing an over-generalised model. The second used a model the second highest lambda value. While the predictions may not be as accurate in terms of R values, the frequencies with the most power at least change as time passes.

We included both versions of the MCCA model in the R value comparisons. The

MCCA models used in the final figures were those with the optimal number of components in terms of R value. It can be seen in Figures 4.9 and 4.10 below that the MCCA model performs the best out of all the models, and seems significantly better than the random, individual, and average models. Figure 4.9, with the exception of the reduced-lambda MCCA model, looks similar to 4.1, the results shown for the EEG dataset. The random model performs the worst, with the individual and average models performing better than this baseline but worse than the MCCA model. The main significant difference is that of the reduced-lambda model, which has a standard error much larger than the rest of the models. This makes sense as it is a less-generalised model, possibly predicting some trials a lot better than others. It still, however, performs better than the average model on average, which was confirmed by a paired t-test between the two rejecting the null hypothesis for all three features.

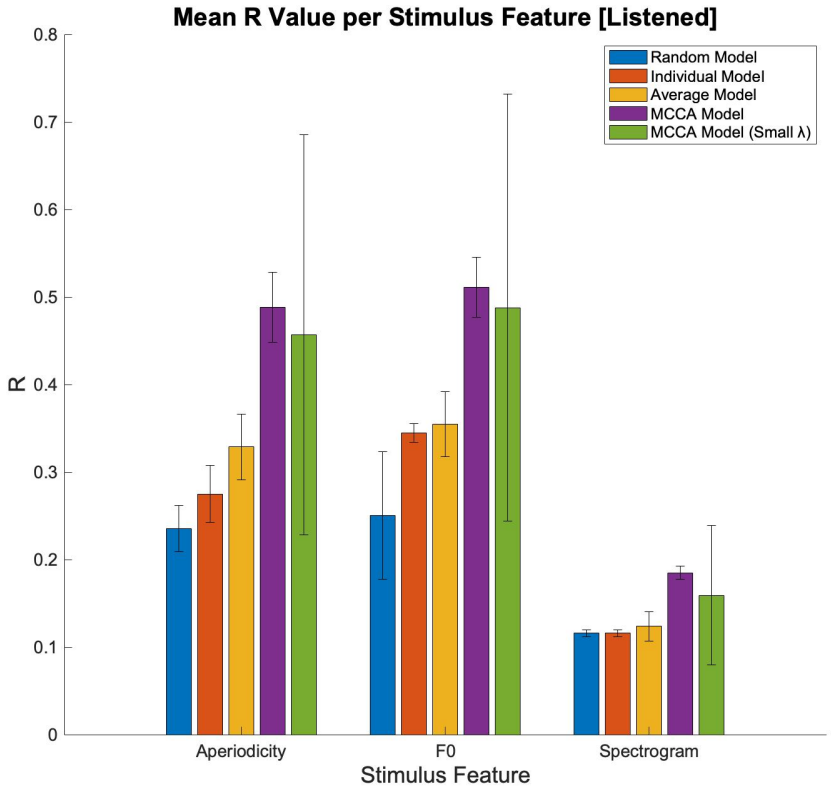

Figure 4.9: A comparison of listened MEG R values based on different models for each stimulus feature. The values are the mean R correlation over 20 trials. The standard error is also shown for each bar.

The results in Figure 4.10 for the imagined MEG models look somewhat different to those seen previously. Neither the individual, the reduced-lambda model, nor the average model performs better than the random model. The only model that improves on the

random model baseline is the initial MCCA model. A paired t-test between the MCCA model and the random model was used to confirm this for each feature. All tests rejected the null hypothesis, showing that there is a significant improvement with MCCA. It is clear that it is harder to create an accurate model using imagined brain data but using MCCA can help. The amount of regularisation used also makes a significant difference, with both the aperiodicity and the F0 results being significantly different according to paired t-tests which rejected the null hypothesis.

A point of note is that the random model was created in the same way as an individual model except with the order of the trials randomised. However, since there are only two audios the randomisation has half a chance to match the correct trial to its corresponding audio. This was done so as to match the process of the EEG evaluation, which had 20 different audios and thus less chance of this happening. For future evaluations, a different baseline dummy approach should be considered.

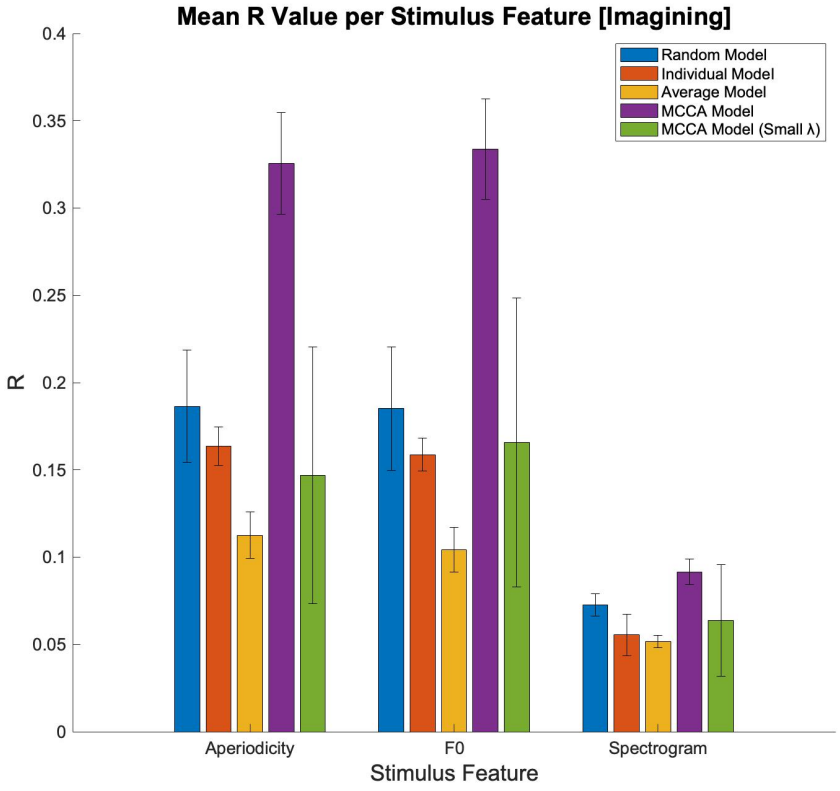

Figure 4.10: A comparison of imagined MEG R values based on different models for each stimulus feature. The values are the mean R correlation over 20 trials. The standard error is also shown for each bar.

Now that we know that MCCA is the optimal model out of the options evaluated, we can compare the models created from the subjects listening to the stimuli and those

created from subjects imagining the stimuli. As mentioned, there are two trials, or predictions, for each stimulus audio. We want to know if the spectrogram predictions made using imagined data and listened data are significantly different, in the hopes that they are not. Paired t-tests were conducted first between the trials of the listened data, then the imagined data, and then between both.

The average p value between two trials representing the same audio was 0.1629 for the listened MEG data and 0.1809 for the imagined MEG data. This means that the two trials for each audio are not significantly different, i.e. the data in the two trials follows a normal distribution with a mean equal to 0. On the other hand, the average p value between two trials representing different audios was 0.0048 for the listened MEG data and 0.0079 for the imagined MEG data. This rejects the null hypothesis, showing that the predictions from, for example, trial 1 representing audio 1 and trial 3 representing audio 2, are significantly different in terms of their means. Now that we have confirmed that the trials for the two audios work as expected using either the listened or imagined dataset, we can now compare trials between the two datasets.

When we used a paired t-test to compare trials from the listened dataset and imagined dataset that are trying to predict the same audio, the p value was 0.1060 which does not reject the null hypothesis. Contrastingly, the p value was 0.0185 when comparing trials across the listened and imagined datasets that were not predicting the same audio. We can conclude from all of these paired t-tests that the predictions made using the imagined MEG dataset are similar to those made using the listened MEG dataset.

An example that supports this conclusion is visualised in Figure 4.11 below. The first image shows a sample of a stimulus audio spectrogram, the second shows the MCCA prediction for that sample using the listened MEG data, and the final image shows the same except using the imagined MEG data. While the imagined data prediction is not as detailed and less 'strong' than the one using listened data, it is clear that they look very similar. Neither look very similar to the actual stimulus, but it can be seen that the areas with higher amplitudes generally match that of the stimulus, even if the finer details are missing. From this evaluation, we can see that using imagined MEG data is not extremely different from using listened MEG data. It may be possible, with more data and more advanced methods, to retrieve useful speech information from it.

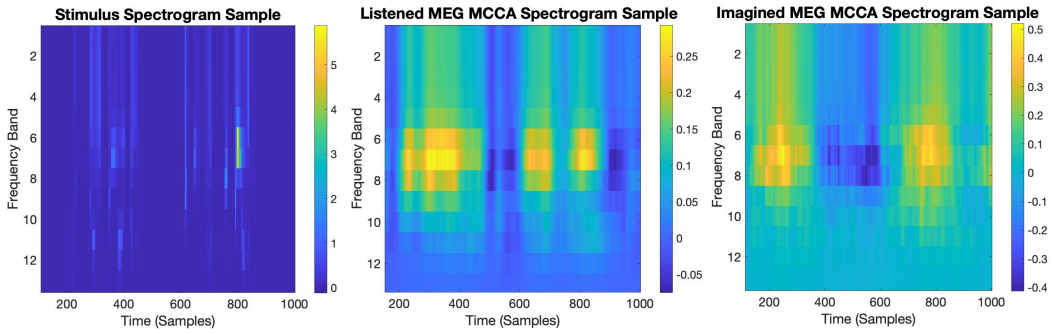

Figure 4.11: Spectrograms for a sample audio from the stimulus and MCCA models using listened MEG data and imagined MEG data.

Following this, it is important to try and evaluate whether the predictions made using the imagined MEG data actually contain any detail in them, rather than just showing frequencies with high amplitudes when someone is speaking and low amplitudes otherwise. One way to do this is to find words that are highly and lowly correlated in the stimulus spectrograms and see if they are similarly highly and lowly correlated in the predicted spectrograms. If the same words are seen as similar, this means some of the essence of the words is being encoded in the imagined MEG data.

We followed the same process for the most correlated and least correlated words. First, we calculate the correlations between all of the stimulus word pairs. The word onset data, i.e. at what sample each word begins, is included in the dataset so this is possible. The top 100 most/least correlated pairs across the two stimulus audios are recorded. We then find those same pairs of words in the MCCA predicted spectrograms and get their correlations. For this use case, the two trials of predictions for each stimulus audio have been averaged so that there is a single prediction for each audio. This is done both for the predictions created using listened MEG data and imagined MEG data. We can then compare the average correlation of these top 100 and bottom 100 most correlated words across the stimulus, the listened MEG predictions, and the imagined MEG predictions. On top of this, a similar evaluation was conducted except with the time dimension of the spectrogram 'squeezed' so that there is now only one value for each frequency band, i.e. the average of each frequency band over time. This was done in order to look at the spectrogram results from two different perspectives.

First, this evaluation was done using the 'best' models for each dataset which used the largest lambda for regularisation and the optimal number of components. The top row in Figure 4.12 shows the spectrotemporal results, with the bottom row showing results for when the time dimension was averaged away. On the right hand side, a zoomed in version of the figure on the left can be seen for extra clarity. The most obvious first impression is that both models have high correlation values for both the highly correlated and lowly

correlated words from the stimulus. As mentioned previously, a large lambda can lead to over-generalisation so it is likely that all of the words look somewhat similar, leading to any pair of words having a high correlation.

Despite this, however, the highly correlated words do have higher correlation values using both the imagined and listened MEG models. This is not the case once the time aspect is squeezed. While the imagined MEG model continues to show good results, the listened MEG model has a higher average correlation for lowly correlated words in the stimulus than highly correlated words in the stimulus. The 'best' listened MEG model uses just 8 components, in comparison to 32 components in the imagined MEG model, so it might be the case that there is not enough components used in the listened MEG model to accurately represent and differentiate words.

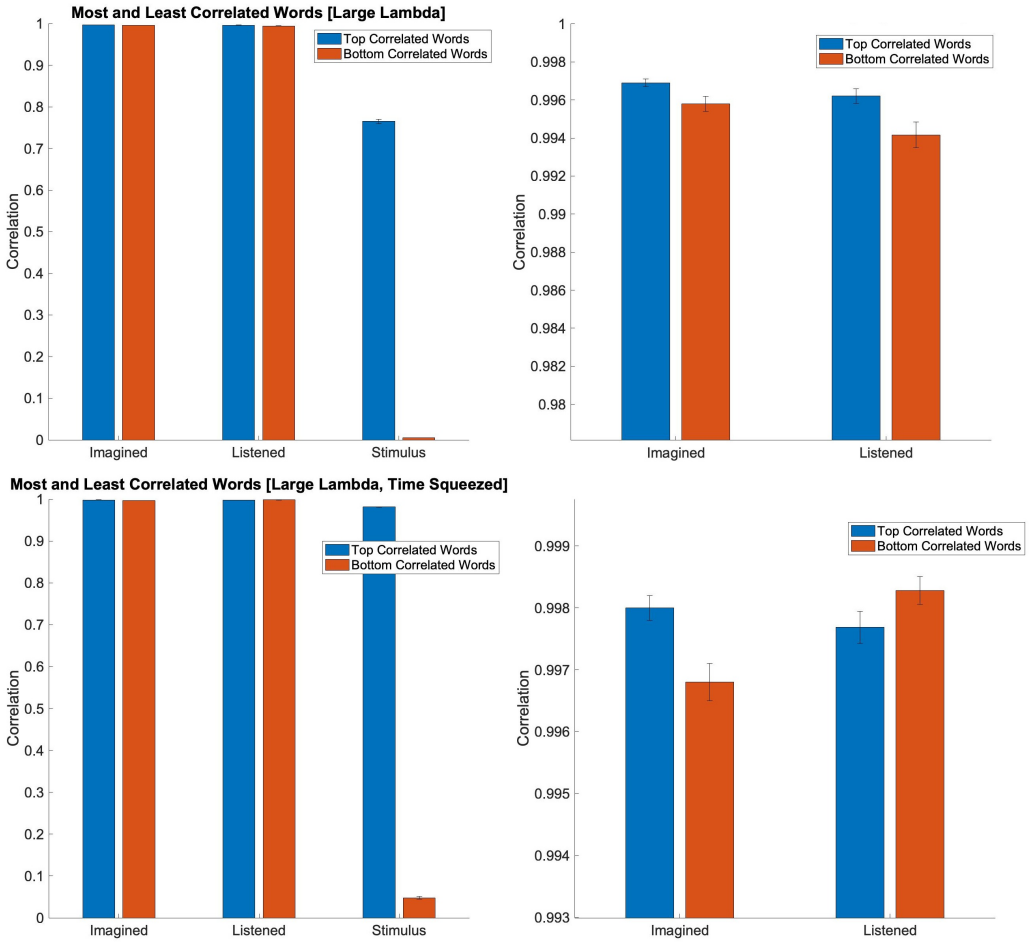

Figure 4.12: Average correlation of word pairs that are the top 100 most highly correlated and lowly correlated in the stimulus spectrogram.

To further investigate the two models without dealing with the over-generalisation leading to every word pair having a high correlation, we decided to also look into the

models created using smaller lambda values. These results were somewhat less conclusive. The 'best' small lambda models were used and the results can be seen in Figure 4.13. While the listened model performed in the way we had hoped, with highly correlated word pairings having higher correlations, the imagined MEG model did not. The results changed somewhat when the time aspect was removed, with the listened MEG model having a smaller gap between highly and lower correlated words and the imagined MEG model performing better than with the time aspect. It is also interesting to note that the listened MEG model still has an issue with having very high correlation values for most word pairings, while the average correlation values for the imagined MEG model are smaller. This could mean that the imagined MEG model is less accurate in general or that the listened MEG model is still overly generalised.

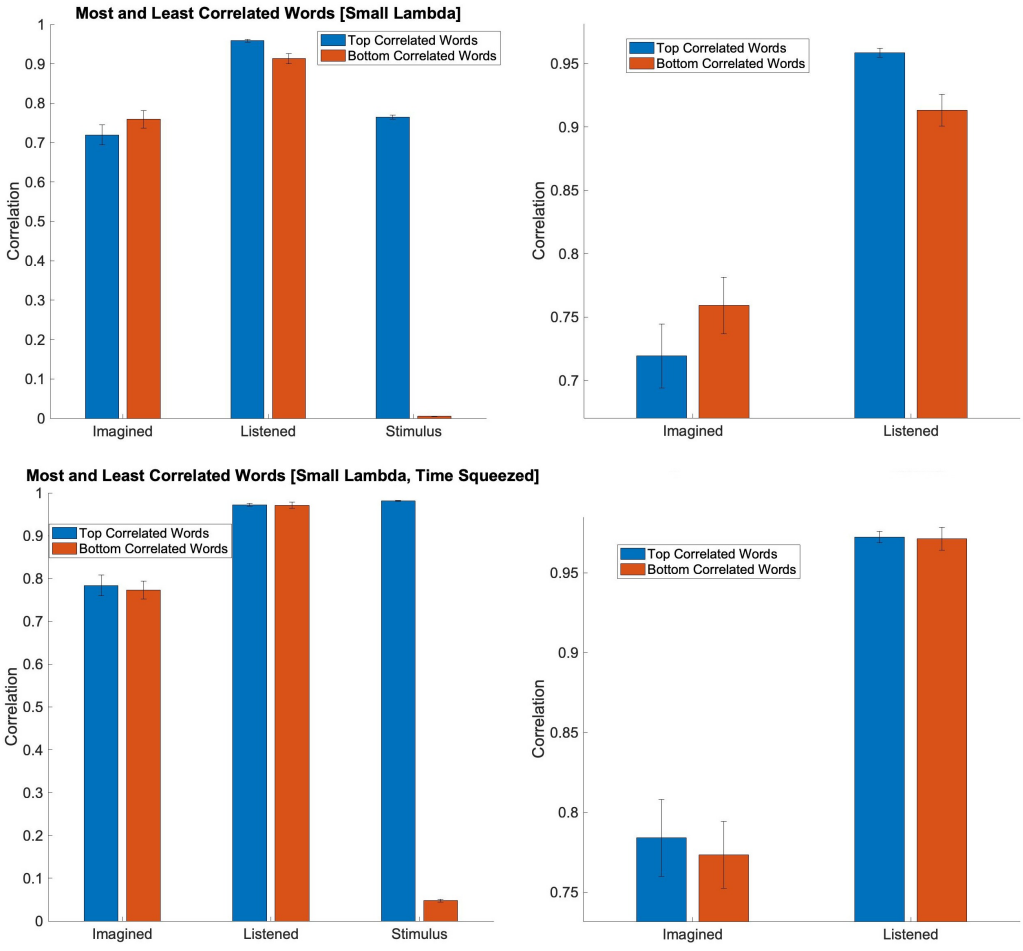

Figure 4.13: Average correlation of word pairs that are the top 100 most highly correlated and lowly correlated in the stimulus spectrogram when a smaller lambda is used in the MCCA models.

A final evaluation using this methodology was conducted by, rather than using the

'best' models for each dataset, using 64 components for both models. This was done to see how the two datasets fared in direct comparison and with a lot more components than in the 'best' small lambda models. The results of this evaluation look quite positive for the imagined MEG model, with the average correlation of highly correlated words in the stimulus being higher than lowly correlated words. This is the case for both when the time aspect is present and when the values are averaged over time. The listened MEG model also does well when the time is not squeezed, but performs poorly once it is. It is unclear as to why. It does, however, seem to be the case that using more components counteracts the regularisation somewhat as the predictions do not seem as generalised - the average correlation values are lower for both models. Using more components like this to create a more complex model may be necessary in order to encode enough detail in the spectrogram to differentiate words.

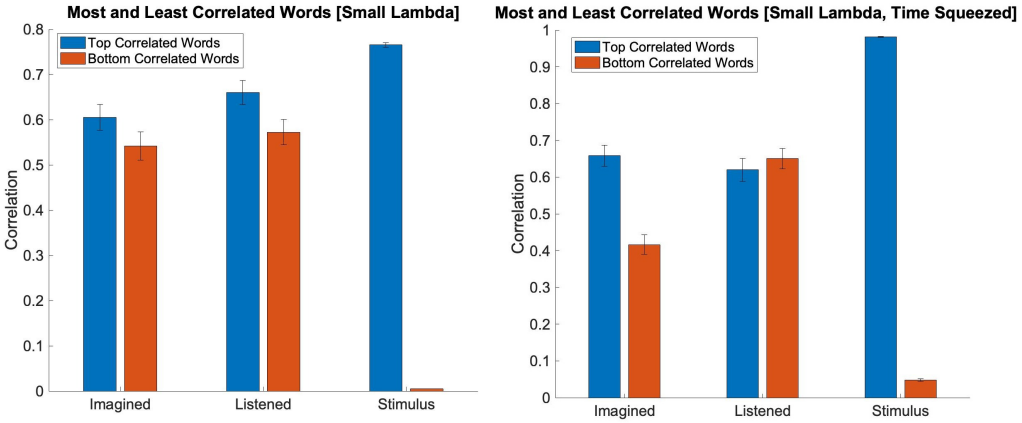

Figure 4.14: Average correlation of word pairs that are the top 100 most highly correlated and lowly correlated in the stimulus spectrogram when a large number of components are used in the MCCA model.

Finally, a sample word from the top 100 most correlated stimulus words was taken and we visualised the corresponding spectrogram predicted by the MCCA models using listened and imagined MEG data. The spectrograms in the first row represent one word, while those in the second row represent a second word that is highly correlated to it. As with Figure 4.11, Figure 4.15 shows us that the two models using different data predict somewhat similar results. There is a visual similarity between the stimulus, listened MEG prediction, and imagined MEG prediction, even if it gets 'blurrier' and less defined. However, we can also see that the models have a hard time predicting silence, shown by how the second word should have low 'dark' values at the beginning but does not in either prediction, both show a line of power throughout.

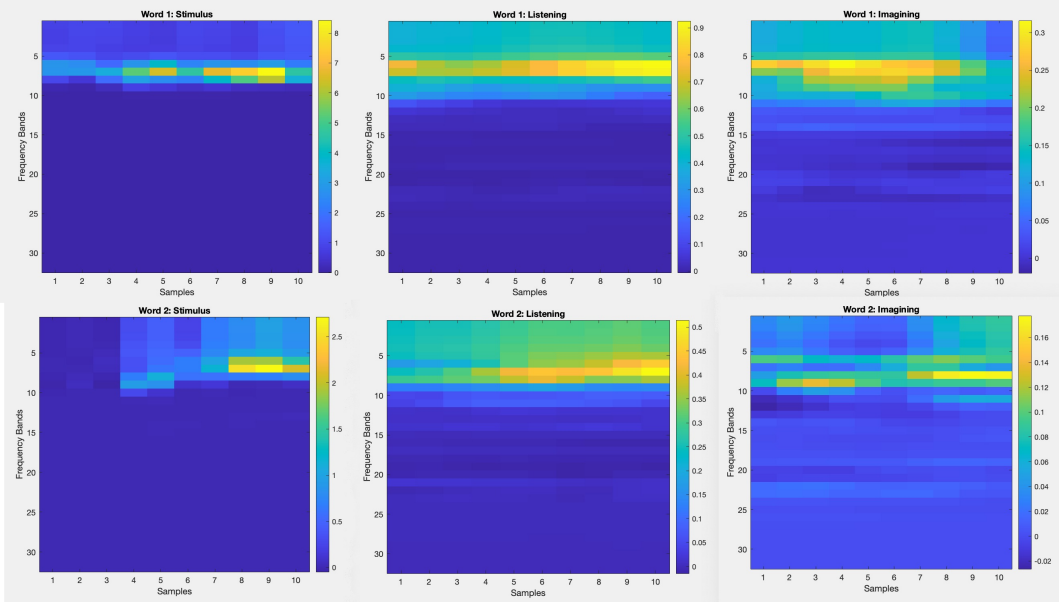

Figure 4.15: Spectrograms for two words that are highly correlated. The stimulus spectrogram is compared to the MCCA model predictions using listened and imagined MEG data.

4.3.2 Audio Produced

Once again, we use the same metrics as with the EEG dataset in order to objectively evaluate the audio reconstructions for the MEG dataset. Subjectively, the reconstructions using both the listened and imagined MEG data are not intelligible. However, there did seem to be a noticeable improvement in the prediction of intonation and rhythm, such as predicting when the speaker was and was not speaking. This makes sense as there was a consistent pattern to the stimulus audios due to them being poems. The reconstructed audio does seem somewhat closer to sounding like speech than when using the EEG dataset.

In order to have a point of comparison for the model reconstructions, we evaluated the best possible reconstruction that could be produced. This meant, as with the EEG stimulus, going through the process of downsampling and then upsampling the stimulus, and so on. To summarise, the higher the Short-time Objective Intelligibility (STOI) and the Frequency weighted Segmental SNR (fwSNRSeg) the better, while the lower the Cepstrum Distance Objective Speech Quality Measure (CD) is the better.

STOI	fwSNRSeg	CD
0.78943	5.87776	3.1094

Table 4.6: Objective metrics for 'The Standard' of reconstructed speech with the MEG dataset. These results comes from getting stimulus features from the vocoder, downsampling them to 100Hz, resampling them to 200Hz, then putting them back through the vocoder again as inputs.

The same disclaimer can be made for the following metric results as was made with the EEG dataset. The reconstructions themselves are intelligible and thus may not actually be good enough to be compared using these higher-level metrics. However, in the MEG dataset's case, the metric results do seem to make more sense and follow a pattern.

First, the reconstructions using MEG data where the subject listened to each stimulus audio were evaluated. While our main focus is the MEG data where the subject imagined each stimulus audio, it is useful to have direct points of comparison. Each model described in the previous section was used to predict each stimulus feature, which was then put through the vocoder to create a reconstruction of each audio. The results can be seen in the table below, as well as a normalised version of the results in Figure 4.16.

	STOI	fwSNRSeg	CD
Random	0.28600	-0.49478	9.25354
Individual	0.29875	-0.42699	9.27131
Average	0.30163	-0.39527	9.24132
MCCA	0.29627	-0.24926	9.18599
MCCA (Norm F0)	0.30210	-0.74342	9.24578
MCCA (Small λ)	0.33639	-0.25055	9.24177
MCCA (Small λ, Norm F0)	0.32443	-0.57295	9.26079

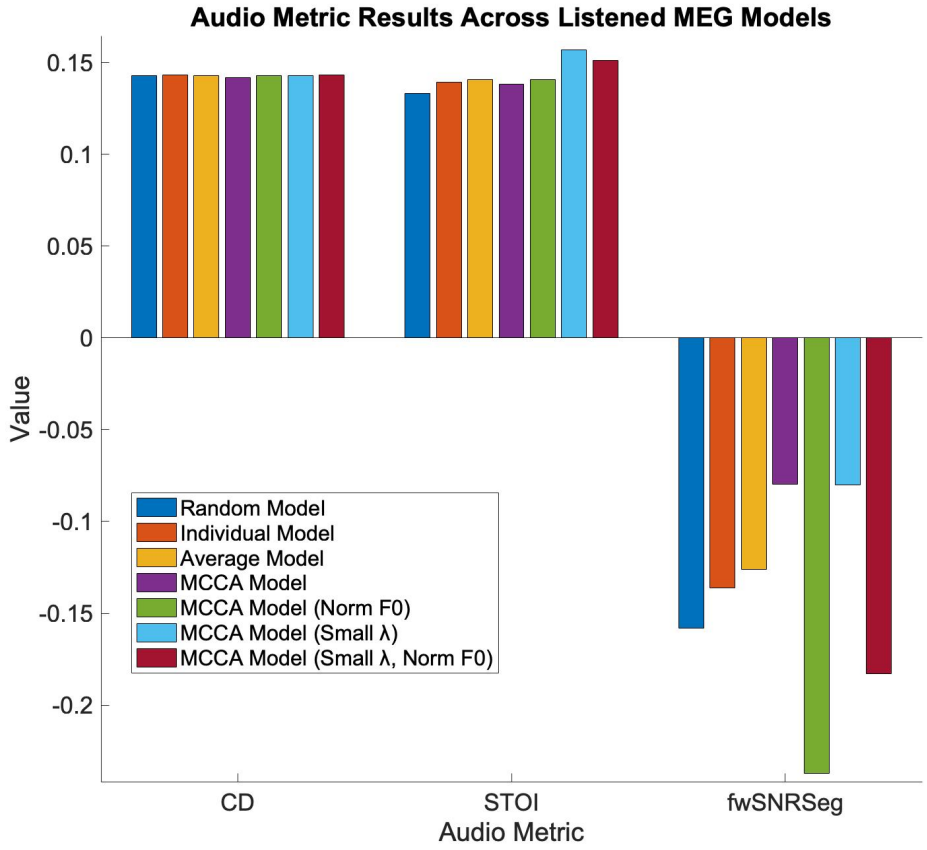

Figure 4.16: Comparison of model predictions in regards to normalised speech intelligibility and quality metric results with listened MEG data.

Overall, the regular MCCA model performs the best for listened MEG data, with the lowest CD and the least negative fwSNRSeg. All of the models outperform the random model in terms of STOI, as well as in terms of fwSNRSeg except those created using a normalised F0 stimulus. The CD values seem inconclusive, with the model results all being very similar and the random model outperforming half of the other models.

Based on these results, a normalised F0 stimulus seems to either weaken the signal or increase the amount of noise in the reconstructed audio. It is better to use the original F0 values so that the model predicts results that follow the natural range of the stimulus speaker's voice. While decreasing the lambda regularisation improves the STOI metric, it leads to worse results for the other two metrics. In general, using a larger lambda with non-normalised stimulus values seems optimal. All of the models produce results with bad signal-to-noise ratios, shown by the all-negative values for the fwSNRSeg metric.

When it comes to using the imagined MEG data, results are similar. Once again, the initial MCCA model performs the best overall. Some of the modified MCCA models outperform it in terms of the STOI metric, but then underperform with the fwSNRSeg

and CD metrics. It is also interesting to see that the values themselves do not get a lot worse compared to the listened MEG data. Using the imagined data does not seem to have a very noticeably negative effect on the reconstructions.

	STOI	fwSNRSeg	CD
Random	0.28311	-0.43119	9.19278
Individual	0.27425	-0.51578	9.22207
Average	0.26044	-0.50207	9.28602
MCCA	0.28737	-0.28433	9.17192
MCCA (Norm F0)	0.29530	-0.82428	9.25325
MCCA (Small λ)	0.28390	-0.29206	9.36828
MCCA (Small λ, Norm F0)	0.27046	-0.60517	9.38879

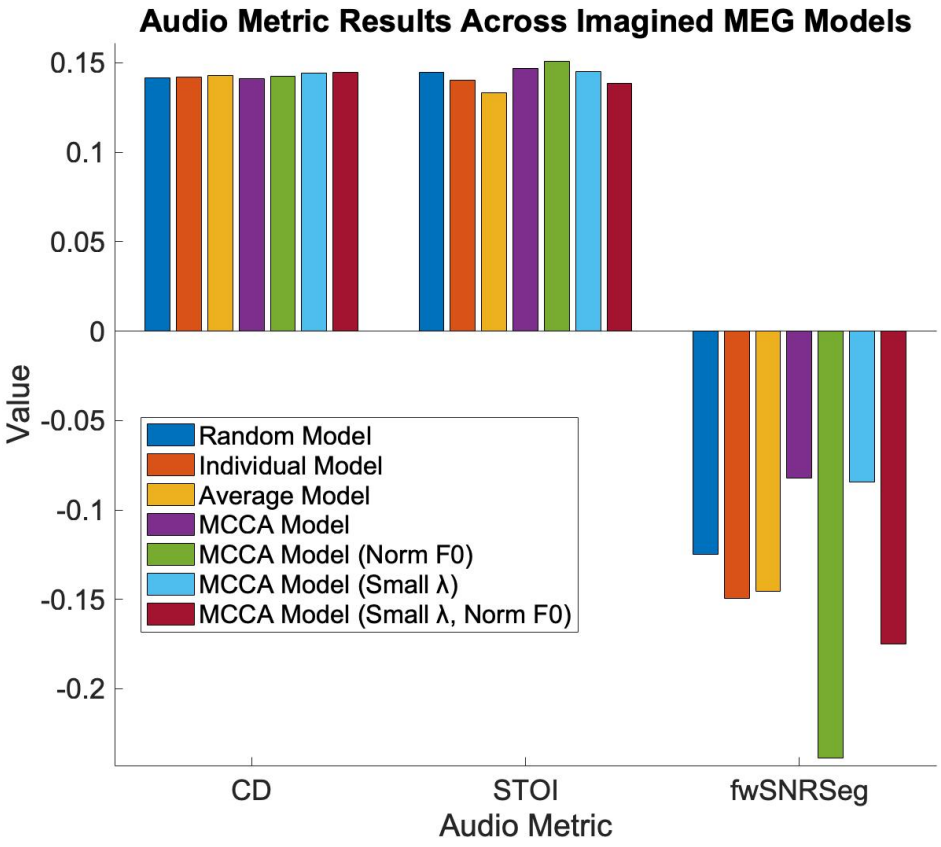

Figure 4.17: Comparison of model predictions in regards to normalised speech intelligibility and quality metric results with imagined MEG data.

With the MCCA model chosen as the overall best reconstruction, the next step is to try and improve its audio metric results. This is done not only to see how to realistically

improve the reconstruction, but also to deconstruct what exactly it is missing. We focused on the model created using the imagined MEG data as that is the topic of interest.

One method of improvement is to scale the predicted values to match the range of the stimulus. We can also average the two trials of predictions that are made for each audio. The way the model works is it has four trials, two for each audio. Averaging them could reduce some of the noise. Another option is to modulate the stimulus envelope on top of the predictions. While this is generally not realistic for all natural speech, considering the stimuli were poems which had a regular rhythm, it is conceivable to try this out. Finally, we replaced each predicted feature with the real stimulus, again using the same processes as with the EEG dataset. The results can be seen in the table below, as well as a normalised summary of the values in Figure 4.18.

	STOI	fwSNRSeg	CD
MCCA	0.28737	-0.28433	9.17192
Real F0	0.33490	-0.59949	9.24410
Real Spectrogram	0.65099	4.51074	3.71222
Real Aperiodicity	0.24191	-0.20560	9.21234
Average Trials	0.29538	-0.82564	9.25348
Scale Features	0.31281	0.59812	7.88534
Envelope Modulation	0.47845	0.39008	7.90165
Scale + Envelope	0.4774	0.88643	7.42453

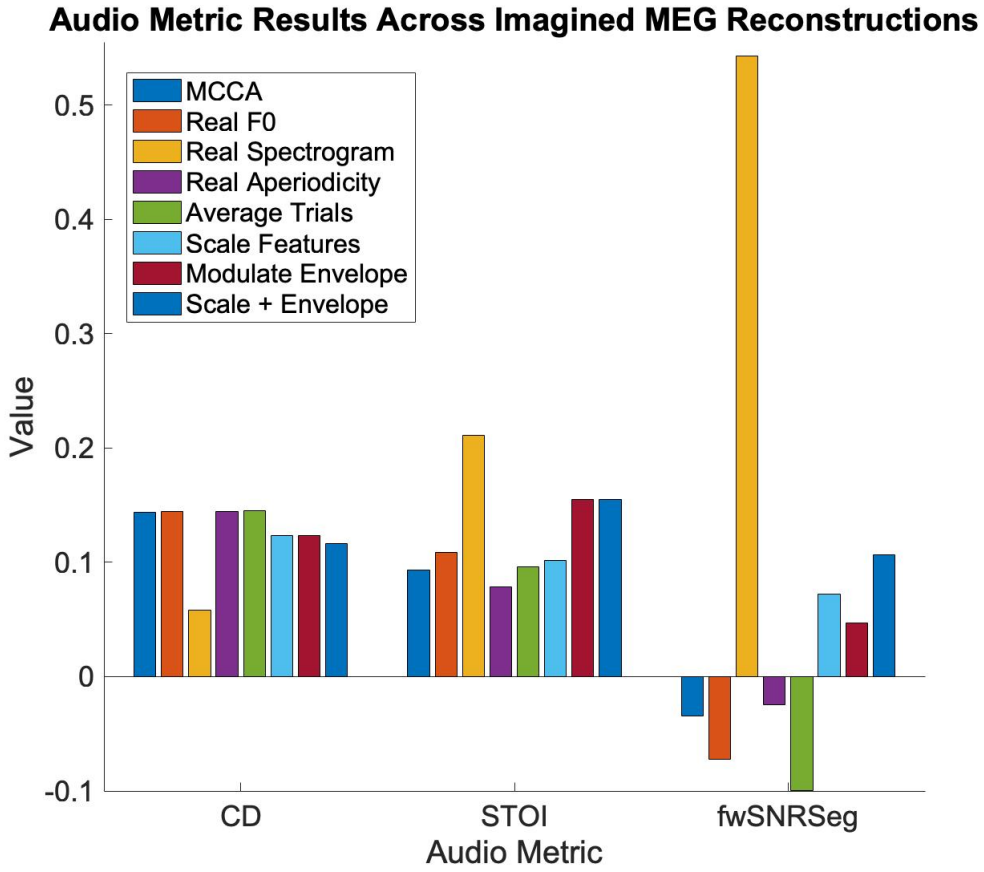

Figure 4.18: Comparison of audio metric results for reconstructions with an array of improvements on the initial MCCA model.

As was the case with the EEG dataset, the spectrogram has the biggest impact on the reconstruction. When the real spectrogram is used instead of the predicted spectrogram, the metric results are almost as good as the optimal values. Replacing the predicted F0 or aperiodicity values with the stimulus ones does not have as positive of an effect. Scaling the features and modulating the envelope over the audio also had positive results, especially when the two improvements are combined. Subjectively, the reconstructed audio when both scaling and envelope modulation is used sounds the closest to real speech than any of the others, of course with the exception of substituting in the real spectrogram. The intonation and rhythm of the speech can clearly be heard, even if the details such as phonemes or words cannot be heard. This is a hopeful result, especially considering this is using the imagined MEG data.

Chapter 5

Conclusions

5.1 Overview

The initial goal of our project, which was to use MCCA among other methods like TRFs to create a baseline for linear EEG decoding, was achieved. The MCCA model clearly and significantly outperforms the average model, its closest competitor. This was true not only when using the EEG dataset but also the MEG dataset. Furthermore, the WORLD vocoder was used with this model to create reconstructions of the audio stimuli. While the resulting audio files were not intelligible, it sets a working baseline for future research to build off of. We set out to follow some of the steps taken by Akbari et al. (2019) except with non-invasive data and linear models, which we did. While the spectrogram reconstructions leave much to be desired when it comes to details, we have shown that some speech information can be gleaned from EEG and there is definite room for improvement.

Along this process, some useful insights were also discovered. Some of these include the importance of how many MCCA components are used in the linear model and how much of an impact the regularisation makes. Both are worth fine-tuning. Moreover, we know that the focus of improvements should be the spectrogram as it affects the reconstruction the most of the three features. The model predictions can definitely be improved, even with small post-processing steps such as scaling the results to match the stimulus. This does not have to be a 'cheat' and the average range of a male voice, for example, could be used instead of the actual stimulus.

Furthermore, from the evaluations conducted there does not seem to be a large difference between MEG data recorded while the subject listens to an audio and MEG data recorded while the subject imagines an audio. While the listened MEG data does seem to perform somewhat better, the imagined MEG data still generally picks up on a lot of the same information, just in a 'blurrier' or weaker form. This is a very hopeful sign

for future research into brain data recorded while imagining speech. While linear models using this type of imagined data may not ever be able to reproduce intelligible speech, we have learned a lot about what it can and cannot do, what it does and does not contain using the current methods and data.

5.2 Future Work

Now that this baseline has been set, it leads the way for future research to be done in this area. The methodologies used, such as MCCA and TRFs, can be reused with more improvements and innovation. There is also the possibility of using the methodology developed for this project with deep learning methods. This may help find more detailed information possibly encoded in the brain data.

While we cannot confirm it with certainty, our research has also shown that it is a definite possibility that imagined stimulus data can contain the type of speech information that listened stimulus data can. One avenue of future work could be to continue on to definitively prove this.

Finally, as usual, another way to improve upon this research is to use more data. The MEG data used, for example, had five subjects and two audio stimuli. If more larger-scale research can be conducted in the area of, for example, recording brain data while imagining continuous speech, then we will have more data to work with. This will improve the models and make them more robust. Furthermore, it will allow for some of the more complex deep learning methods to be a possibility as they usually are very data-hungry and require larger datasets in order to work well.

Bibliography

Akbari, H., Khalighinejad, B., Herrero, J., Mehta, A., and Mesgarani, N. (2019). Towards reconstructing intelligible speech from the human auditory cortex. *Scientific Reports*, 9:874.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L. (2021). Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *J Big Data 8*, 54.

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):036019.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., and Lalor, E. C. (2017). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *bioRxiv*.

Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C. (2016). The multivariate temporal response function (mtrf) toolbox: A matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, 10.

da Silva, F. L. (2013). Eeg and meg: Relevance to neuroscience. *Neuron*, 80(5):1112–1128.

Dash, D., Ferrari, P., and Wang, J. (2020a). Decoding imagined and spoken phrases from non-invasive neural (meg) signals. *Frontiers in Neuroscience*, 14.

Dash, D., Wisler, A., Ferrari, P., Davenport, E., Maldjian, J., and Wang, J. (2020b). Meg sensor selection for neural speech decoding. *IEEE access : practical innovations, open solutions*, 8.

de Cheveigné, A., Di Liberto, G. M., Arzounian, D., Wong, D. D., Hjortkjær, J., Fuglsang, S., and Parra, L. C. (2019). Multiway canonical correlation analysis of brain data. *NeuroImage*, 186:728–740.

Destoky, F., Philippe, M., Bertels, J., Verhasselt, M., Coquelet, N., Vander Ghinst, M., Wens, V., De Tiège, X., and Bourguignon, M. (2019). Comparing the potential of meg and eeg to uncover brain tracking of speech temporal envelope. *NeuroImage*, 184:201–213.

Di Liberto, G. and Nidiffer, A. (2021). Cnd format: Cnsp2022.

Guan, C., Thulasidas, M., and Wu, J. (2004). High performance p300 speller for brain-computer interface. In *IEEE International Workshop on Biomedical Circuits and Systems, 2004.*, pages S3/5/INV–S3/13.

Herff, C., Heger, D., de Pesters, A., Telaar, D., Brunner, P., Schalk, G., and Schultz, T. (2015). Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in Neuroscience*, 9.

Herff, C. and Schultz, T. (2016). Automatic speech recognition from neural signals: A focused review. *Frontiers in Neuroscience*, 10.

Luck, S. J. (2005). *An Introduction to the Event-Related Potential Technique*. MIT press.

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

Morise, M. (2015). Cheaptrick, a spectral envelope estimator for high-quality speech synthesis. *Speech Communication*, 67:1–7. Funding Information: This work was supported by JSPS KAKENHI Grant Nos. 24300073 and 26540087 and the Research Institute of Electrical Communication, Tohoku University (H25/A08).

Morise, M., Kawahara, H., and Katayose, H. (2009). fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech. *journal of the audio engineering society.*

Morise, M. and Watanabe, Y. (2018). Sound quality comparison among high-quality vocoders by using re-synthesized speech. *Acoustical Science and Technology*, 39(3):263–265.

Morise, M., Yokomori, F., and Ozawa, K. (2016). World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, E99.D(7):1877–1884.

Panachakel, J. T. and Ramakrishnan, A. G. (2021). Decoding covert speech from eeg-a comprehensive review. *Frontiers in Neuroscience*, 15.

Price, C. J. (2012). A review and synthesis of the first 20years of pet and fmri studies of heard speech, spoken language and reading. *NeuroImage*, 62(2):816–847. 20 YEARS OF fMRI.

Proix, T., Delgado Saa, J., Christen, A., Martin, S., Pasley, B., Knight, R., Tian, X., Poeppel, D., Doyle, W., Devinsky, O., Arnal, L., Mégevand, P., and Giraud, A.-L. (2022). Imagined speech can be decoded from low- and cross-frequency intracranial eeg features. *Nature Communications*, 13.

Razaeizadeh, M., Shamma, S., and Di Liberto, G. (TBD). Meg dataset. In Preparation.

Souza, P. and Rosen, S. (2009). Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech. *The Journal of the Acoustical Society of America*, 126:792–805.

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., and Francart, T. (2018). Speech intelligibility predicted from neural entrainment of the speech envelope. *Journal of the Association for Research in Otolaryngology*, 19(2):181–191.

Värbu, K., Muhammad, N., and Muhammad, Y. (2022). Past, present, and future of eeg-based bci applications. *Sensors*, 22:3331.

Zhuang, X., Yang, Z., and Cordes, D. (2020). A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping*, 41(13):3807–3833.

Appendix

.1 Appendix A

The source code used in this project can be found on GitHub here. It is also available in the electronic resources submitted with this report, along with the reconstructed audio files.

.2 Appendix B

An evaluation was also conducted as to how the calculated MCCA components relate to the initial EEG electrodes. It can be seen in Figure 1 below that certain MCCA components, such as the 56th component, have noticeably higher correlations with the original EEG electrodes. Considering how much of an impact the first MCCA component has on the model, it was expected that it would have a significantly high correlation with the EEG electrodes. However, this was not the case.

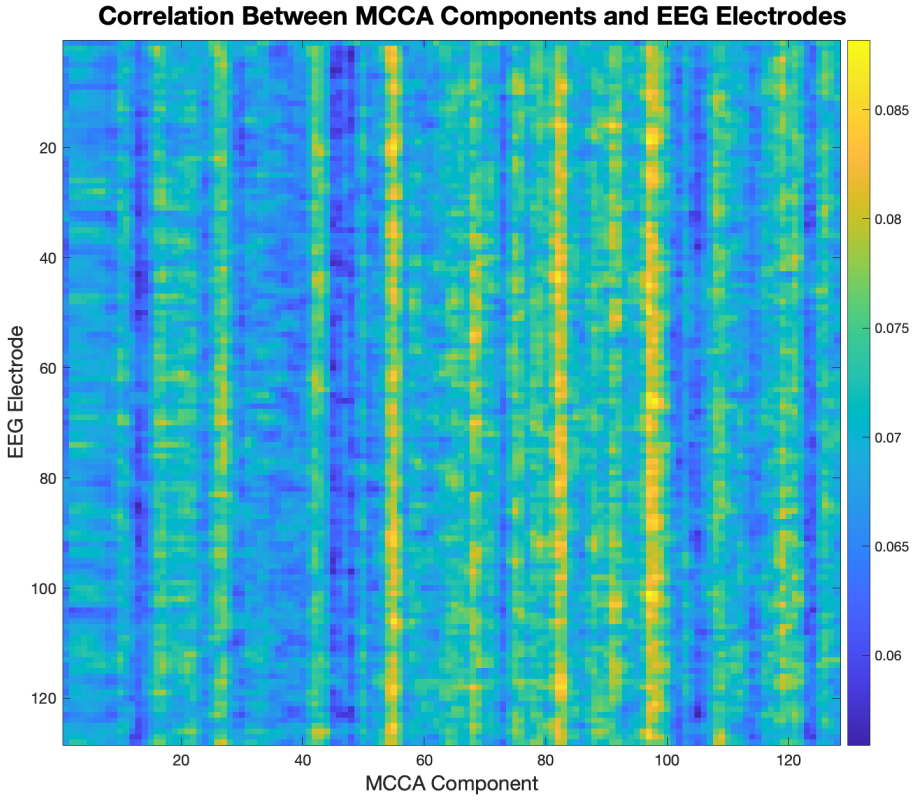

Figure 1: Correlation of MCCA Components to EEG Electrodes. The absolute values of the correlations were averaged over the 19 EEG subjects and the 20 audio trials.

We also looked into how well the MCCA model with EEG data predicted the individual bands in a spectrogram. For each frequency band, the average R value over time for the predicted spectrogram was retrieved. It can be seen in Figure 2 that the lower frequency bands are best predicted by the model. The overall pattern is the higher the frequency the lower the R value. In particular, the uncommon very high frequencies has a significantly lower R value than the rest of the bands.

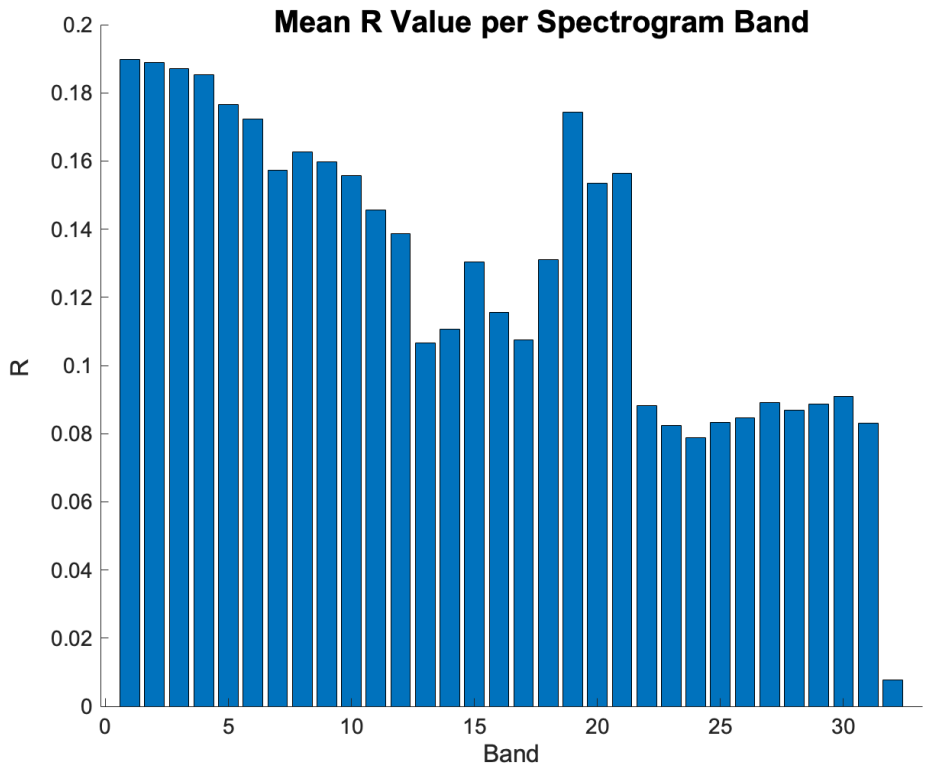

Figure 2: R Value of Individual Spectrogram Bands for MCCA Model.

.3 Appendix C

As with the EEG dataset, an evaluation as to how many components should be used in the MCCA models was also completed for the MEG datasets. For each feature of both the listened and imagined MEG data, the R value results for models containing 8, 16, 32, 64, and 128 respectively were retrieved. This is shown in Figures 3 and 4. The optimal number of components chosen for the listened MEG data were the following: 64 components for F0, 16 components for the spectrogram, and 64 components for the aperiodicity. For the imagined MEG data, 32 components proved to be optimal for all three features.

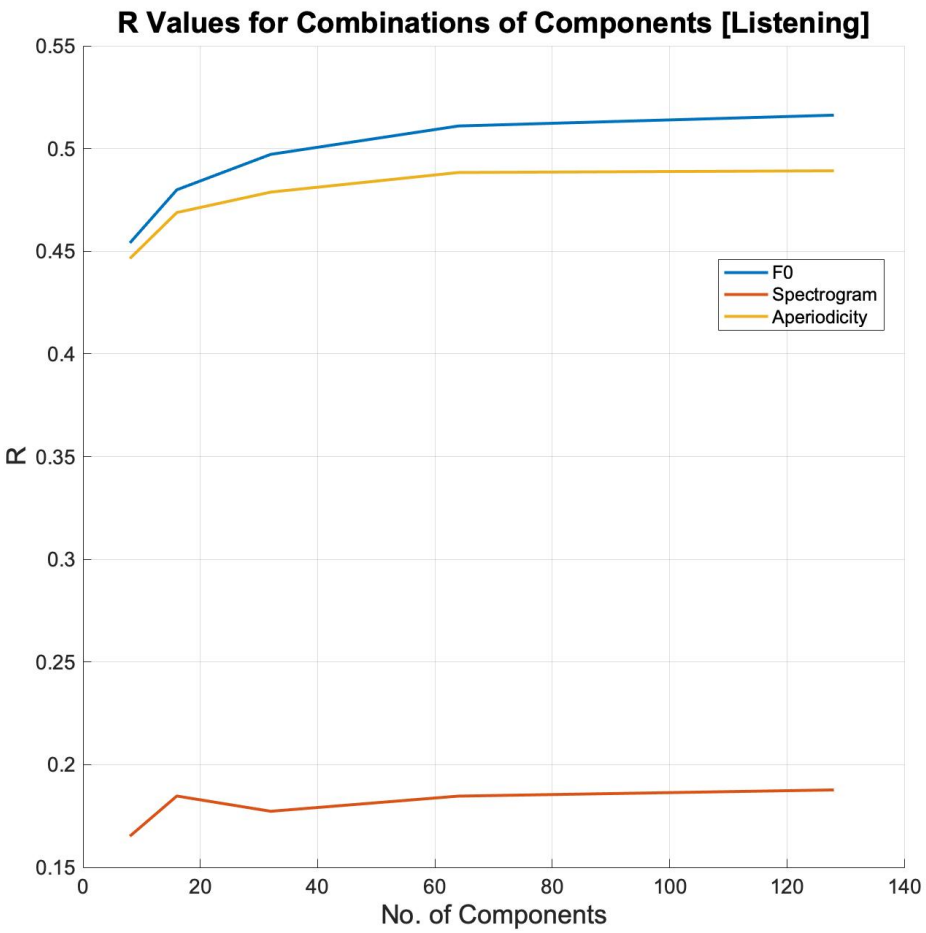

Figure 3: R values for different numbers of components used in an MCCA model with listened MEG data. This is used to choose the optimal number of components to include in each feature's final MCCA model.

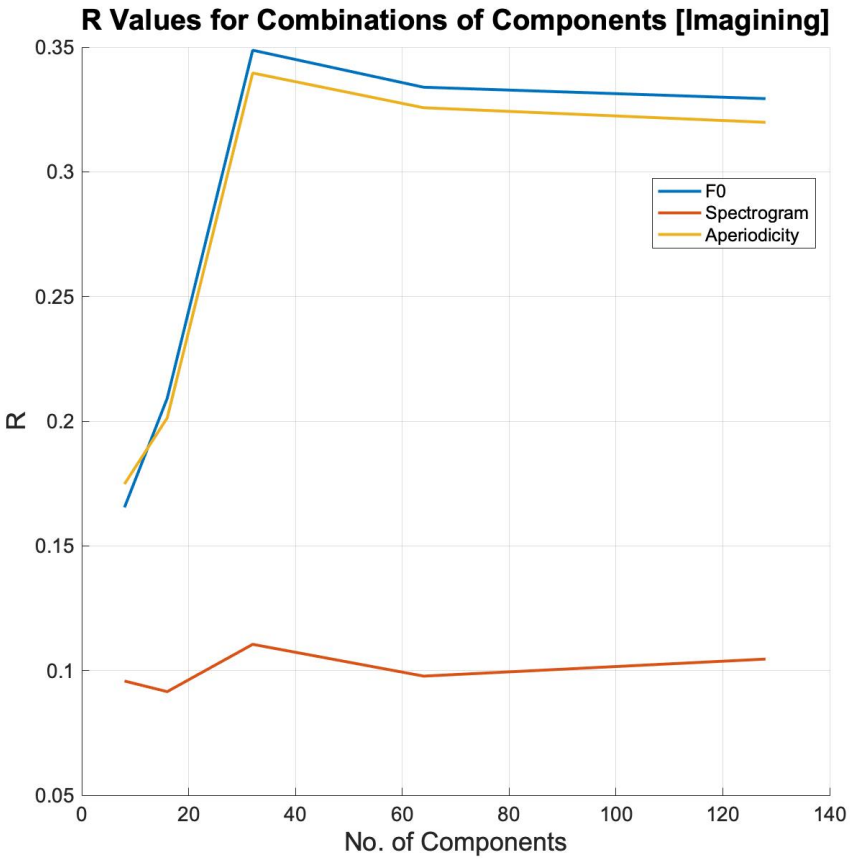

Figure 4: R values for different numbers of components used in an MCCA model with imagined MEG data. This is used to choose the optimal number of components to include in each feature's final MCCA model.