Deep Learning Methods For Audio EEG Analysis

A THESIS

SUBMITTED FOR THE DEGREE OF

Master of Technology (Research)

IN

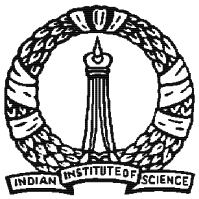
SYSTEM SCIENCE AND SIGNAL PROCESSING

by

Jaswanth Reddy Katthi

under the guidance of

Dr. Sriram Ganapathy



Electrical Engineering

Indian Institute of Science

BANGALORE – 560 012

APRIL 2021

TO

Every curious brain trying to figure out itself

Acknowledgments

This would not be possible if Dr. Sriram Ganapathy did not see the potential in me. All credits goes to him and his patience. Along with his guidance and mentoring, his encouragement and principles have made me a better researcher and person. Many thanks to him for hanging there with my mistakes and shaping me towards where I am now. I am grateful to him for encouraging me in my lows and lending me the space whenever I required it.

I am grateful to all my course instructors for helping me to explore this realm. My LEAP family played an important role by constantly pouring in the perspectives. Technically and personally, they played a huge role in making this work happen. I thank Dr. P S Sastry for the fruitful suggestions they gave me on my first day. I thank Dr. Chandrasekhar Seelamanthula, Dr. K V S Hari, Dr. Prasanta Kumar Ghosh and Dr. A G Ramakrishnan for their hospitality and the helpful insights provided throughout my courses. I am thankful to Dr. Malcolm Slaney and Sandeep Kothinti for the initial work and making me a part of this project. I thank Dr. Blair Kaneshiro and Dr. Edmund C Lalor for providing the datasets.

I thank my family (amma, nanna, arjun and Tiger), Chandana and my friends for their continuous love and support. I thank TSS for being there whenever I needed.

I owe my career to the visionaries and situations led the establishment of this amazing institute and my chance of pursuing research here. I am thankful to the

creators across the world for inspiring me time to time. Finally, I would like to thank all the giants providing me a space to stand on their shoulders.

Publications based on this Thesis

1. Katthi, Jaswanth Reddy, Sriram Ganapathy, Sandeep Kothinti, and Malcolm Slaney. "Deep Canonical Correlation Analysis For Decoding The Auditory Brain." In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 3505-3508. IEEE, 2020.

2. Katthi, Jaswanth Reddy, and Sriram Ganapathy. "Deep Multiway Canonical Correlation Analysis for Mulit-Subject EEG Normalization." In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

3. Katthi, Jaswanth Reddy, and Sriram Ganapathy. "Deep Correlation Analysis for Audio-EEG Decoding." IEEE Transactions on Neural Systems and Rehabilitation Engineering (2021).

Abstract

The perception of speech and audio is one of the defining features of humans. Much of the brain's underlying processes as we listen to acoustic signals are unknown, and significant research efforts are needed to unravel them. The non-invasive recordings capturing the brain activations like electroencephalogram (EEG) and magnetoencephalogram (MEG) are commonly deployed to capture the brain responses to auditory stimuli. But these non-invasive techniques capture artifacts and signals not related to the stimuli, which distort the stimulus-response analysis. The effect of the artifacts becomes more evident for naturalistic stimuli. To reduce the inter-subject redundancies and amplify the components related to the stimuli, the EEG responses from multiple subjects listening to a common naturalistic stimulus need to be normalized. The currently used normalization and pre-processing methods are the canonical correlation analysis (CCA) models and the temporal response function based forward/backward models. However, these methods assume a simplistic linear relationship between the audio features and the EEG responses and therefore, may not alleviate the recording artifacts and interfering signals in EEG. We propose novel methods using machine learning advances to improve the audio-EEG analysis.

We propose a deep learning framework for audio-EEG analysis in intra-subject and inter-subject settings. The deep learning based intra-subject analysis methods are trained with a Pearson correlation-based cost function between the stimuli and

EEG responses. This model allows the transformation of the audio and EEG features that are maximally correlated. The correlation-based cost function can be optimized with the learnable parameters of the model trained using standard gradient descent-based methods. This model is referred to as the deep CCA (DCCA) model. Several experiments are performed on the EEG data recorded when the subjects are listening to naturalistic speech and music stimuli. We show that the deep methods obtain better representations than the linear methods and results in statistically significant improvements in correlation values.

Further, we propose a neural network model with shared encoders that align the EEG responses from multiple subjects listening to the same audio stimuli. This inter-subject model boosts the signals common across the subjects and suppresses the subject-specific artifacts. The impact of improving stimulus-response correlations are highlighted based on multi-subject EEG data from speech and music tasks. This model is referred to as the deep multi-way canonical correlation analysis (DMCCA). The combination of inter-subject analysis using DMCCA and intra-subject analysis using DCCA is shown to provide the best stimulus-response in audio-EEG experiments.

We highlight how much of the audio signal can be recovered purely from the non-invasive EEG recordings with modern machine learning methods, and conclude with a discussion on future challenges in audio-EEG analysis.

Contents

List of Tables

List of Figures

Keywords

Electroencephalogram, Cochlear implants, Naturalistic stimuli decoding, Deep Learning, Auditory Neuroscience, Canonical Correlation Analysis, Deep CCA, Multiway CCA, Deep MCCA

Chapter 1

Introduction

1.1 Motivation

Since scientific methods started logging into human's activities, reverse-engineering the human brain has been a fascinating topic for research and understanding.

The brain performs multiple jobs, conscious and subconscious, and processes multimodal information at hierarchical levels of abstractions simultaneously. All the sensory information are processed and combined to extract information from the environment. However, very little is known about these hierarchical processing streams of data. The existing black box model of the brain combined with the need for advancements in artificial systems have placed the study into brain processes as a key area of research.

The brain's properties like energy efficiency, robustness in data processing and lack of data hunger are the major requirements of today's data-driven machine learning models. The auditory system plays a prominent role in exploring the world around us. Not being error-prone to reverberation effects, focusing on a particular sound amidst multiple sound sources, robust semantic mapping of phonemes, being

able to enjoy music are few examples which prove the efficiency of our auditory systems. These properties of our auditory systems attract more research into decoding them.

Recent technological advancements allow capturing the electromagnetic signals directly from the brain. The neurons are the basic processing elements in the brain. Each neuron generally fires at a rate of 100 Hz and the composition of these firings form higher frequency signals. The current brain recording techniques either provide considerable temporal resolution or spatial resolution. The techniques like electroencephalography (EEG) and magnetoencephalography (MEG) can record the brain activity at sampling rates as high as 8 kHz. The temporal resolution allows us to capture processes with high precision. However, as they are captured from the scalp, they lack the spatial resolution needed to locate the brain regions involved. Techniques like functional MRI (fMRI) provide considerable spatial resolution in the order of millimeters but lack the requisite temporal resolution with sampling rate ranging in seconds. The invasive techniques like electrocorticography (ECoG) provides high temporal resolution upto 10 kHz and spatial resolution in the order of millimeters, but it needs electrodes to be inserted into the brain. Depending on the purpose, suitable techniques to capture the brain signals are deployed.

To study the brain's response for an auditory stimulus, an auditory stimulus is provided to the subject and its corresponding brain signals are recorded to model the relationship between the two signals. As the stimuli and responses are temporal in nature, we need to use recording techniques with significant temporal resolution. The EEG and MEG are the two prominent brain signals recording techniques used in these experiments.

The EEG is preferred to MEG as EEG is more portable and easy to record. The EEG recording is inexpensive and takes less time to set up. But, as the recordings

are captured from the scalp, the response component is collected along with other artifacts. While the brain is processing a presented stimulus, it simultaneously performs various other tasks like maintaining the subject's internal systems. All these processes are unrelated to the presented stimulus, but still get captured by the EEG. And some stimuli induced responses also get smeared out while propagating to the scalp. Therefore, the EEG recordings contain significant amount of signals that are not related to the stimuli presented to the subject. These signals in the EEG recordings are considered as noise. In this regard, the EEG recordings have SNR < -20dB[1]. Another shortcoming of EEG is being sensitive to artifacts like eye blinking and muscular movement.

Initial studies have focused on event related potentials (ERPs) [2]. An ERP tries to model the brain's response to short term auditory stimuli, stimuli that last less than 2 seconds. They alleviate the effect of noise by repeating the same experiment multiple times. By repeating the experiment, all the brain recordings collected for a particular stimuli can be aggregated and averaged. It helps to remove all the processes unrelated to the stimuli from the EEG recordings. This facilitates the study of sensory and cognitive processes of the auditory brain [3]. The short span of the stimuli makes it possible to repeat the experiment multiple times for multiple subjects. The ERP methods have become the standard technique for scientific and medical studies [4, 5].

However, the ERP methods assume a highly simplistic environment. The necessity to repeat an experiment multiple times make the ERP methods inefficient for a natural long stimuli. Hence, developing single-trial based decoding algorithms are of profound interest.

Studying the relationship between various stimuli features and their corresponding EEG recordings help us to decode the brain computations on naturalistic stimuli. An auditory stimulus contains diverse amounts of information. It contains acoustic

information like pitch, rhythm, timber and spectral information. A speech audio's acoustic information contains speaker's voice, rate of speech, accent and ambience. Semantic data like context, purpose, emotional state, dialect and the speaker's vocabulary are also embedded in a speech audio. A musical audio contains details about the vocals, genre, instruments and tempo. The brain perceives all these details from the sounds in real time and evokes myriads of responses. Understanding how the brain processes all these information is the underlying theme of auditory neuroscience.

The successful single trial decoding techniques assume a linear and time-invariant impulse response between the stimulus and response. These earliest methods in this direction are referred to as the temporal response function (TRF) [6]. They rely on a reverse correlation/system identification framework. The linear TRF models are prominently used for two types of models. A "forward model" predicts the EEG response from the audio, whereas a "backward model" uses the neuronal response to predict the features of the audio signal. As forward models describe the encoding action of brain from stimulus to EEG, they come under the "encoder" models. Similarly backward models try to decode the stimulus for a particular EEG, hence they come under the "decoder" models.

The TRF models' performance is typically quantified using the Pearson correlation between the predicted signal and the true signal. The low SNR in the EEG gives rise to correlation values in the range of 0.1 - 0.2 [6]. EEG signals contain all the brain's activity along with the stimuli effects. Thus, only a fraction of the variance in the EEG can be explained by the external stimuli.

Apart from the encoder and decoder models, there exist hybrid models. Let the stimuli be represented as $\mathbf{S}$ and the corresponding response as $\mathbf{R}$. Let $\mathbf{f}$ represents a linear transform on $\mathbf{S}$. And $\mathbf{g}$ represents a linear transform on $\mathbf{R}$. The modelling process attempts to find the two transforms such that the final representations have

optimal performance metric. They aim to bring $\mathbf{f}(\mathbf{S})$ and $\mathbf{g}(\mathbf{R})$ correlate with each other. They follow a data-driven approach (ridge linear regression). A linear forward model learns the optimal $\mathbf{f}$ assuming $\mathbf{g}$ to be an identity function. A linear backward model assumes $\mathbf{f}$ as an identity function and tries to find the optimal $\mathbf{g}$. A hybrid model tries to find the both functions $\mathbf{f}$ and $\mathbf{g}$ that maintain only the related components in both signals. The hybrid models' transformed stimuli $\mathbf{f}(\mathbf{S})$ can also be used to predict the transformed response $\mathbf{g}(\mathbf{R})$. But the downside is that they are difficult to interpret. These models describe the sensory-dependent parts of brain activity, and the stimuli information they encode [7, 8, 9, 10].

Canonical correlation analysis (CCA) is a prominent hybrid model used in the naturalistic stimuli setting. It projects two signals to a domain that maximizes the correlation between the two signals [11, 12]. It finds a linear transform on each of the signals that maximizes the variability relevant to the other signal. Recently, the linear CCA method has been successfully applied in forward and backward models in auditory EEG analysis using a combination of linear transforms and convolutions [13, 14]. However, the model is still based on linear assumptions.

All these models can be trained specifically for each subject separately (subject-specific models), or trained on some subjects and tested on other subjects (subject-independent models). The subject-independent models do not need the tiresome process of collecting the ground-truth EEG data for each new subject. But, since every person's brain responses are different, the subject-independent models perform poorly compared to the subject-specific models [15]. A drawback to data-driven single trial analysis methods for naturalistic stimuli is the lack of data. As it is tedious to repeat the experiment multiple times for a single subject, aggregating information from multiple subjects is one way to increase the available data.

The linear CCA can be performed only on two signals at a time. In order to aggregate the EEG responses from multiple subjects, multiway CCA (MCCA) or generalized CCA [16, 17, 18] has been proposed. As all the EEG responses correspond to the same auditory stimulus, some components must be common across the EEG responses [19]. The application of multiway CCA (MCCA) for EEG mapping has shown improvements over the intra-subject linear CCA [14]. But, both the models, CCA for each subject and MCCA normalization of multiple subjects' EEG, assume a simple linear relationship between the stimuli and responses.

1.2 Related Prior Work

One of the earlier efforts to extend the decoding to longer naturalistic stimuli conditions is performed by Lalor et al. [20]. The AESPA (Auditory Evoked Spread Spectrum Analysis) [6] method stochastically modulates the amplitude of an auditory carrier stimulus and tries to estimate the linear impulse response (of the brain) from the recorded EEG recordings. Using a sliding window of auditory amplitude values and the measured neural data, the impulse response is determined using least-squares estimation (linear regression). Aiken et al. [21] have shown that the stimuli features from 4 - 16 Hz play major role for speech intelligibility.

Recent studies [22, 23] have shown that the EEG recordings clearly track the attended speaker's speech envelope in a listening experiment with more than one speakers. The problem of extracting the attention related information directly from the brain is generally referred to as the auditory attention decoding (AAD) task [10]. Most of the AAD algorithms follow a stimulus reconstruction approach, i.e., backward models.

Previous research has shown auditory space encoding in the subcortical neurons [24],

but less is known about the cortical representation of the auditory space. Being able to locate and attend to a particular auditory source among multiple sources is a complex process managed by the brain. It is important to consider the effects of hierarchial attentional mechanisms on cortical responses. In the context of cocktail party, the low-frequency cortical oscillations predominantly synchronize with the temporal structure of the attended sound stimulus, and less with the unattended source [25]. Decoding these processes is highly useful for developing cognitively steered devices.

Lauteslager et al. [26] has shown that the decoding of attention in a cocktail party setting, from single-trial EEG, is robust and pertinent to the task. It also addresses the usage of all the 128 channels data from EEG recordings. All the 128 channels are considered as their relative contribution is weighted by the model automatically.

Bednar et al. [27] showed that the attended source's position trajectory can be reliably reconstructed from both delta signals' phase and alpha signals' power of EEG. It is even found to be robust to distracting stimuli. In EEG recordings, the delta waves are the signals with a frequency of 3 Hz or below. Alpha waves have a frequency between 7.5 and 13 Hz.

Though the position of unattended source is not tracked using the cortical representation, delta phase of the EEG tracks it weakly [27]. It is also shown that the trajectory reconstruction method can also be used to decode the selective attention in a single-trial context. However, its performance is found to be inferior to envelope-based decoders.

It has been shown that the EEG recordings below 16 Hz robustly track the speech dynamics [20, 28]. Though invasive methods like ECoG reflect the same properties in high gamma power (HGP) waves (70 - 150 Hz), it is unclear whether the HGP from EEG show similar or complementary properties to that of the low frequency waves.

Typically, the high-frequency content of scalp-recorded EEG is filtered out because

they are low pass filtered by the skull [29]. They get smeared out by the dura and cerebrospinal fluid [30]. Hence, the high-frequency content has low signal-to-noise ratio. This poor SNR, low spatial resolution and high sensitivity to muscle artifacts [31] resulted in relatively few studies focused on HGP in EEG.

Synigal et al. [32] shows that HGP also offers speech tracking and attention decoding in the context of cocktail party. It also shows that the HGP and low frequency signals are sensitive to different characteristics of the stimuli. And combining them improves speech tracking for several subjects. It proves that tracking the HGP, along with the low frequency content, is beneficial for cognitively steered hearing devices.

O'Sullivan et al. [33] have successfully shown that the attention to both congruent and incongruent audiovisual speech can be decoded. It is mentioned that the parieto-occipital alpha power can be used to determine whether a subject is listening to a speaker's face or not.

Broderick et al. [34] has shown that the EEG recordings reflect the semantically surprising elements in the stimuli in the order of milliseconds. As we have evolved to live among natural sounds, Zuk et al. [35] shows that the synthesized sounds' responses get poor classification accuracy compared to natural speech or music stimuli.

The stimulus-response modelling offers insight into perceptual processes within the brain, giving it the potential for practical use in Brain Computer Interfaces (BCI). Cheveigné et al. [9] try to quantify such models' performance using metrics like match-mismatch, correlation, sensitivity and classification error rate. The match-mismatch task quantifies the classification efficiency of the models. Thus, it is directly applicable to BCI applications.

In an AAD problem, match-mismatch classification serves as a considerable metric. Final representations are obtained for the EEG response, and the correlation coefficient between each speaker's representation and the EEG's representations is calculated. This is estimated over a decision window length of τ seconds. Generally, the model's performance depends on the decision window length.

Machine learning methods for the extraction of information from EEG can have a significant impact on both understanding and applications like BCI . Therefore, it is quite important to extend the linear models using the recent advancements in machine learning (like deep learning [36]) for brain signal decoding and single-trial analysis.

Identifying the P300 wave in EEG signals using Convolutional Neural Networks (CNNs) is one of the first works in this direction [37]. The recent years have seen the use of deep learning for several brain mapping tasks like computational memory prediction [38], driver's cognitive state prediction [39], and the brain activity reconstruction for visual stimuli [40].

A review of several efforts in decoding brain activity using deep learning techniques is given in Zheng et al. [41]. Kriegeskorte et al. [7] discuss various aspects in the interpretation of the encoder and decoder models. They discuss the simplistic linear assumption of the models and the single-model-significance fallacy. The single-model-significance fallacy argues that evidence of variance explainability must not be interpreted in the favor of the model. It says that it does not represent the brain computation, but only a statistical tool to measure the dependency between the stimuli and their responses.

In auditory tasks, EEG recordings have shown to contain rhythm information in music perception using classifiers based on deep networks [42]. A recent work by Das et al. [43] has shown that auditory attention decoding in the perception of noisy

speech can also be improved by deep learning techniques. In multi-speaker cocktail party scenarios, Deckers et al. [44] showed that neural networks are capable of identifying the attended speaker. A Dense Neural Network (DNN) based model for EEG-based speech stimulus reconstruction was proposed by Taillez et al. [45]. The deep learning models are able to capture the non-linear relationship between the stimulus and response. Thus, it has been showed that deep learning is a feasible alternative to linear decoding methods. Liu et al. [46] proposed a deep version of the linear MCCA for image-EEG data which is similar to the DGCCA model [47] in the literature.

Ciccarelli et al. [48] implements an end-to-end Convolutional Neural Network (CNN) network classification approach to decode the speaker the subject is trying to attend. This approach outperformed linear methods for a decision window of 10 seconds. Vandecappelle et al. [49] have used CNN layers to extract the locus (left/right) of the auditory attention for the scenarios without access to each speaker's stimuli. CNNs have shown encouraging results in the domain of EEG classification for seizure detection [50, 51], sleep stage classification [52] and depression detection [53, 54].

1.3 Key contributions and contrast with prior literature

The linear CCA is one of the popular approaches in the context of EEG-audio data as SRC (Stimulus Response Correlation). All the popular analysis methods rely on the highly simplistic assumption of a linear relationship between stimulus and response.

The linear CCA [13] has been extensively used for modelling a subject's stimulus-response relationship. To address the problem of lack of data, linear MCCA [14] is introduced. The linear MCCA aggregates the EEG responses from multiple subjects to a common stimulus and linearly transforms them such that each subject's stimulus-response correlations improve.

We propose deep learning based models for the existing linear models in this context of audio-EEG data. The deep models for the CCA and MCCA are proposed in the context of audio-EEG data. Andrew et al. [55] has proposed a deep model of CCA that outperforms the linear CCA on image data under low noise conditions. The significant amount of noise in EEG recordings makes it complicated to directly deploy the deep CCA for audio-EEG data. The dropout strategy [56] tries to alleviate the impact of noise partly. We show that leaky-ReLU based non-linearity at the output of the networks to be more robust to noise. We address the usage of CCA models (linear and deep) in Chapter 3 on "Intra-Subject Analysis".

The linear MCCA [14], as a denoising step for multiple subjects EEG data, is shown to provide better representations compared to the standard linear CCA model alone for each subject. We propose a deep variant of the MCCA for the denoising step. A deep MCCA model is developed such that it denoises the EEG recordings significantly better than the linear and other existing deep variants of MCCA. Our proposed deep MCCA model is a generalized version of the DGCCA model [47], and it is tested in the context of audio-EEG data. We use a reconstruction approach with a shared hidden representation to derive the deep transform that aligns multiple EEG recordings. These analyses is referred to as "Inter-Subject Analysis", and detailed in Chapter 4.

We also illustrate the combinations of linear/deep MCCA with the linear/deep CCA methods for audio-EEG relationship analysis in speech and music listening tasks.

1.4 Organization of the thesis

The thesis is organised as follows. Chapter 2 sets the mathematical background for the models discussed in this work along with the datasets these models are experimented

with. Chapter 3 discusses linear and non-linear intra-subject stimulus-response methods. Chapter 4 explores the linear and non-linear inter-subject methods. Chapter 5 discusses an extension to the work which includes the temporal information of the signals. This chapter concludes the discussion with a summary, limitations of the study and future directions for this work.

Chapter 2

Background and Setup

As discussed earlier, audio-EEG analysis can be performed using one of the three models : forward, backward and hybrid. The most popular mathematical models for forward and backward analysis come under the name Temporal Response Function (TRF). The linear CCA is the most prevalent hybrid model for each subject's audio-EEG response analysis. The linear MCCA is a technique to aggregate multiple subjects' EEG responses to a common subspace such that each subject's intra-subject analysis benefits from the accumulated EEG responses for a common stimulus.

2.1 Mathematical Background

2.1.1 Linear Regression and Temporal Response Function

Let a stimulus $\mathbf{x}[t]$ be provided to a subject and its corresponding EEG response is recorded as $\mathbf{y}[t, c]$. The stimulus $\mathbf{x}[t]$ represents the 1D temporal audio envelope and the response $\mathbf{y}[t, c]$ represents the C channel EEG signals. The stimulus, $\mathbf{x}[t]$, and the response, $\mathbf{y}[t, c]$, are represented at same sampling frequency.

The temporal response function (TRF) is a prominently used linear model for modelling the forward and backward relationships between the stimulus and response. A TRF model approximates the relation between the stimulus and response as a convolution with an impulse response.

The TRF considers a context window of $[\tau_{\text{min}}, \tau_{\text{max}}]$ for the input at each instant. Let $\tau_{\text{max}} - \tau_{\text{min}} = \tau_{\text{window}}$. Hence, if the signal is of $T \times d$ dimensions, then the input is transformed to a signal of shape $T \times d\tau_{\text{window}}$, and the modelling is performed on this new input representations [57]. This can be called as a time-lagged version of the input.

The stimulus can be represented either as the temporal envelope, $\mathbf{x}[t]$ or any processed features. Therefore, the stimulus is denoted as $\mathbf{X} \in \mathbb{R}^{T \times d_S}$ where $d_S = 1$ for the audio envelope representations and differs if any other processed features are utilized. The EEG response data can also be represented using either the same number of channels used for their collection or can be processed and projected onto a different subspace. Therefore, $\mathbf{Y} \in \mathbb{R}^{T \times d_R}$ where $d_R = C$ if the EEG is not dimensionality reduced.

Forward Model

A forward model tries to predict the EEG response from given audio stimulus. The input to the model is the time-lagged version of the stimulus $\mathbf{x}[t]$. The time-lagged stimuli is represented as $\mathbf{X} \in \mathbb{R}^{T \times d_S}$ where $d_S = \tau_{\text{window}}$, and the response is represented as $\mathbf{Y}$.

The relation between the stimulus $\mathbf{X}$ and the response $\mathbf{Y}$ is assumed to be linear,

$$\mathbf{Y} = \mathbf{XF} \tag{2.1}$$

The model is learnt using ridge linear regression as :

$$\mathbf{F} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{Y} \tag{2.2}$$

where $\mathbf{I} \in \mathbb{R}^{d_S \times d_S}$ is an identity matrix and λ is the regularization parameter.

Backward Model

A backward model tries to reconstruct the provided audio stimulus from the EEG response. The input to the model is the time-lagged version of the response $\mathbf{y}\,[t, c]$. It can be represented as $\mathbf{Y} \in \mathbb{R}^{T \times d_R}$ where $d_R = C\tau_{\text{window}}$, C is the number of EEG channels and τ_{window} is the context window. The stimulus is represented as $\mathbf{X}$.

The relation between the stimulus $\mathbf{X}$ and the response $\mathbf{Y}$ is assumed to be

$$\mathbf{X} = \mathbf{Y}\mathbf{G} \tag{2.3}$$

The model is learnt using ridge linear regression as :

$$\mathbf{G} = (\mathbf{Y}^\top \mathbf{Y} + \lambda \mathbf{I})^{-1} \mathbf{Y}^\top \mathbf{X} \tag{2.4}$$

where $\mathbf{I} \in \mathbb{R}^{d_R \times d_R}$ is an identity matrix and λ is the regularization parameter.

These models for relating the continuous stimuli to their neural signals are built into a publicly available MATLAB toolbox named as "mTRF Toolbox" [57].

2.1.2 Linear Canonical Correlation Analysis

For a pair of multi-variate datasets, Canonical Correlation Analysis (CCA) [11] finds the optimal linear transforms, for both the stimulus and response, that maximize the Pearson correlation between the transformed vectors.

Let, $\mathbf{x}$ and $\mathbf{y}$ denote d_S and d_R dimensional vectors respectively. Let, d denote the dimension of the desired canonical subspace that maximizes the correlation between transformed vectors. For example, if $d = 1$, let $\mathbf{u}_1$, $\mathbf{v}_1$ denote the pair of vectors which project $\mathbf{x}$ and $\mathbf{y}$ respectively into 1-dimensional space. Now, the problem is to find $\mathbf{u}_1$ and $\mathbf{v}_1$ such that the correlation coefficient ρ between $\mathbf{x}' = \mathbf{u}_1^\top \mathbf{x}$ and $\mathbf{y}' = \mathbf{v}_1^\top \mathbf{y}$ is maximized. The problem can be formulated as maximizing

$$\rho(\mathbf{u}_1, \mathbf{v}_1) = \frac{\mathbb{E}[(\mathbf{x}' - \mathbb{E}[\mathbf{x}'])(\mathbf{y}' - \mathbb{E}[\mathbf{y}'])]}{\sqrt{\mathbb{E}[(\mathbf{x}' - \mathbb{E}[\mathbf{x}'])^2] \cdot \mathbb{E}[(\mathbf{y}' - \mathbb{E}[\mathbf{y}'])^2]}} \tag{2.5}$$

Without loss of generality, let $\mathbb{E}[\mathbf{x}] = 0$ and $\mathbb{E}[\mathbf{y}] = 0$. Then, the correlation coefficient becomes

$$\begin{aligned} \rho(\mathbf{u}_1, \mathbf{v}_1) &= \frac{\mathbb{E}[\mathbf{x}'\mathbf{y}']}{\sqrt{\mathbb{E}[(\mathbf{x}')^2] \cdot \mathbb{E}[(\mathbf{y}')^2]}} \\ &= \frac{\mathbb{E}[\mathbf{u}_1^\top \mathbf{x}\mathbf{y}^\top \mathbf{v}_1]}{\sqrt{\mathbb{E}[\mathbf{u}_1^\top \mathbf{x}\mathbf{x}^\top \mathbf{u}_1] \cdot \mathbb{E}[\mathbf{v}_1^\top \mathbf{y}\mathbf{y}^\top \mathbf{v}_1]}} \\ &= \frac{\mathbf{u}_1^\top \mathbf{C}_{xy} \mathbf{v}_1}{\sqrt{\mathbf{u}_1^\top \mathbf{C}_{\mathbf{xx}} \mathbf{u}_1 \mathbf{v}_1^\top \mathbf{C}_{\mathbf{yy}} \mathbf{v}_1}} \\ (\mathbf{u}_1^*, \mathbf{v}_1^*) &= \underset{\mathbf{u}_1, \mathbf{v}_1}{\operatorname{argmax}} \rho(\mathbf{u}_1, \mathbf{v}_1) \end{aligned} \tag{2.6}$$

where, $\mathbf{C}_{\mathbf{xy}} = \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y})^\top]$ and $\mathbf{C}_{\mathbf{xx}}$, $\mathbf{C}_{\mathbf{yy}}$ are the auto-correlation matrices of $\mathbf{x}$, $\mathbf{y}$ respectively.

Generalizing the problem for $d := k(> 1)$ and constraining the denominator to 1, we can calculate the optimum transforms $(\mathbf{U}_d, \mathbf{V}_d)$ as:

$$\begin{aligned} (\mathbf{U}_\mathbf{d}^*, \mathbf{V}_\mathbf{d}^*) &= \underset{\mathbf{U_d}, \mathbf{V_d}}{\operatorname{argmax}} \operatorname{Trace}(\mathbf{U_d}^\top \mathbf{C_{xy}} \mathbf{V_d}) \\ \text{subject to } & \mathbf{U_d}^\top \mathbf{C_{xx}} \mathbf{U_d} = \mathbf{V_d}^\top \mathbf{C_{yy}} \mathbf{V_d} = \mathbf{I_d} \end{aligned} \tag{2.7}$$

where $\mathbf{I}_d$ is identity matrix of shape $d \times d$, the matrices $\mathbf{U}_d = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_d] \in \mathbb{R}^{d_S \times d}$ and $\mathbf{V}_d = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_d] \in \mathbb{R}^{d_R \times d}$ are the linear transforms for $\mathbf{x}$ and $\mathbf{y}$ respectively.

Let,

$$\mathbf{T} \triangleq \mathbf{C}_{\mathbf{xx}}^{-1/2} \mathbf{C}_{\mathbf{xy}} \mathbf{C}_{\mathbf{yy}}^{-1/2} \tag{2.8}$$

The solution to the CCA problem ($\mathbf{U}_d^*$ and $\mathbf{V}_d^*$) are given as the first d left and right singular vectors of the matrix $\mathbf{T}$ and the maximum correlation is given by the sum of the top d singular values of $\mathbf{T}$ [55].

2.1.3 Linear Multiway Canonical Correlation Analysis

The multiway CCA (MCCA) generalizes the linear CCA to multiple (more than two) multivariates. The linear MCCA finds a linear transform for each random variable, such that all the projections are maximally correlated to each other [14, 19].

Let us consider N random multivariates, $\mathbf{x}_n \in \mathbb{R}^{d_n}$, for $n = 1$ to N. And $D_N = \sum_{n=1}^{N} d_n$. Let us project all the N random variables $\mathbf{x}_n$ onto a 1D subspace. Let $\mathbf{v}_n \in \mathbb{R}^{d_n}$ denote the transform vector that projects $\mathbf{x}_n$ onto the common subspace.

The MCCA finds the transform vectors $\{\mathbf{v}_n\}_{n=1}^{N}$ such that the inter-set correlation (ISC) among the projections is maximum. The ISC is defined as

$$\rho_{\text{ISC}} = \frac{1}{N-1} \frac{r_B}{r_W} \tag{2.9}$$

where r_B is the between-set covariance and r_W is the within-set covariance. The factor $N-1$ scales the correlation to be $\rho_{\text{ISC}} \leq 1$. The between-set and within-set covariances are obtained as

$$r_B = \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \mathbf{v}_i^\top \mathbf{R}^{ij} \mathbf{v}_j \ , \ r_W = \sum_{i=1}^{N} \mathbf{v}_i^\top \mathbf{R}^{ii} \mathbf{v}_i$$

where $\mathbf{R}^{ij} \in \mathbb{R}^{d_i \times d_j}$ is the cross-covariance matrix between $\mathbf{x}_i$ and $\mathbf{x}_j$. Each element of the cross covariance matrix is obtained as $[\mathbf{R}^{ij}]_{kl} = \left(\mathbf{x}_i^k - \overline{\mathbf{x}}_i^*\right)^\top \left(\mathbf{x}_j^l - \overline{\mathbf{x}}_j^*\right)$ with $\overline{\mathbf{x}}_i^*$ as the mean of $\mathbf{x}_i$.

The cross-covariance matrices among all the views are aggregated to form the block matrix $\mathbf{R} \in \mathbb{R}^{D_N \times D_N}$ such that $[\mathbf{R}]_{ij} = \mathbf{R}^{ij}$. By considering only the autocovariance matrices, a block-diagonal matrix $\mathbf{D} \in \mathbb{R}^{D_N \times D_N}$ is formed such that $[\mathbf{D}]_{ii} = \mathbf{R}^{ii}$.

The optimum transform vectors $\{\mathbf{v}_n\}_{n=1}^N$ are obtained by solving the eigen equation [19]:

$$\mathbf{R}\mathbf{v} = \lambda \mathbf{D}\mathbf{v} \tag{2.10}$$

The eigenvector $\mathbf{v} \in \mathbb{R}^{D_N \times 1}$ with the maximum eigenvalue contains the optimum transform vectors $\{\mathbf{v}_n\}_{n=1}^N$.

To solve the eigen equation 2.10, we can either find the eigenvectors for the matrix $\mathbf{D}^{-1}\mathbf{R}$, or follow the two-step solution as discussed by Parra [19]. The two-step solution first decomposes the block-diagonal matrix $\mathbf{D}$ into an orthonormal matrix $\mathbf{U}$ and diagonal matrix Λ as

$$\mathbf{D} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \tag{2.11}$$

Since the block-diagonal matrix $\mathbf{D}$ elements are covariance matrices of each multivariate, this decomposition is similar to performing PCA on each $\mathbf{x}_n$ separately. It is

analogous to whitening each of them. The eigen equation can be rewritten as:

$$\mathbf{R}\mathbf{v} = \lambda\mathbf{D}\mathbf{v} \tag{2.12}$$

$$\mathbf{R}\mathbf{v} = \lambda\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{1/2}\mathbf{U}^\top\mathbf{v} \tag{2.13}$$

$$\mathbf{\Lambda}^{-1/2}\mathbf{U}^\top\mathbf{R}\mathbf{v} = \lambda\mathbf{\Lambda}^{1/2}\mathbf{U}^\top\mathbf{v} \tag{2.14}$$

$$\left(\mathbf{\Lambda}^{-1/2}\mathbf{U}^\top\mathbf{R}\right)\left(\mathbf{U}\mathbf{\Lambda}^{-1/2}\mathbf{\Lambda}^{1/2}\mathbf{U}^\top\right)\mathbf{v} = \lambda\left(\mathbf{\Lambda}^{1/2}\mathbf{U}^\top\mathbf{v}\right) \tag{2.15}$$

$$\left(\mathbf{\Lambda}^{-1/2}\mathbf{U}^\top\mathbf{R}\mathbf{U}\mathbf{\Lambda}^{-1/2}\right)\tilde{\mathbf{v}} = \lambda\tilde{\mathbf{v}} \tag{2.16}$$

$$\tilde{\mathbf{R}}\tilde{\mathbf{v}} = \lambda\tilde{\mathbf{v}} \tag{2.17}$$

Multiplying the concatenated whitening transforms $\mathbf{U}\mathbf{\Lambda}^{-1/2}$ on either side of the matrix $\mathbf{R}$ produces the matrix $\tilde{\mathbf{R}}$. Hence, $\tilde{\mathbf{R}}$ is the covariance matrix of concatenated whitened multivariates. Therefore, to solve the eigen equation 2.10, the eigenvectors $\mathbf{U}$ and the eigenvalues $\mathbf{\Lambda}$ are obtained by eigen decomposition of $\mathbf{D}$ (or PCA of each $\mathbf{x}_n$), and $\tilde{\mathbf{V}}$ by eigen decomposition of $\tilde{\mathbf{R}}$. And the final eigenvectors are obtained as $\mathbf{V} = \mathbf{U}\mathbf{\Lambda}^{-1/2}\tilde{\mathbf{V}}$.

The first column of this transformation matrix $\mathbf{V}$ is considered for projecting the multivariates onto a 1D subspace. For a higher dimensional subspace d (> 1), the first d columns of $\mathbf{V}$ (because they correspond to the top d eigenvalues) are considered.

Our work primarily focuses on the hybrid CCA models. We discuss the linear hybrid models extensively used for audio-EEG data and the deep models we propose which outperform the linear models.

2.2 Datasets

We use audio-EEG recordings from two datasets, speech and music. The speech-EEG dataset consists of subjects' EEG recordings while they were listening to speech

stimuli (an American audiobook). The music-EEG dataset contains EEG recordings of the subjects listening to music (Hindi popular songs).

2.2.1 Speech - EEG dataset

This dataset contains pairs of speech stimuli and their corresponding recorded EEG responses. This is an open dataset and was recorded by Liberto et al. [25]. The subjects were presented with snippets of a popular American novel read by a male speaker. All the snippets were of same length (≈ 155 seconds). All the stimuli were presented monophonically in dichotic fashion at a sample rate of $44,100$ Hz. While the subjects were listening to the audiobook, their EEG data were recorded from 128 electrodes at a sampling rate of 512 Hz using Biosemi Active Two system. The EEG data were preprocessed to remove the phase-distortions [25] and downsampled to 128 Hz. This dataset is chosen for experimenting our methods for a speech-EEG setting.

For the experiments, we have collected 20 trials of 8 subjects from the dataset. The preprocessing steps proposed by Cheveigné et al. [13] are followed. The EEG data are downsampled to 64 Hz. They are further detrended to exclude outliers using a robust detrending routine [58]. Channel-specific noise are suppressed using the STAR algorithm [59]. Both the detrending and denoising algorithms are used from the noise suppression tools [58]. To suppress the 50 Hz and its harmonics, the EEG data are convolved with a boxcar window of duration 20ms. Then, they are finally passed through a band-pass filter of passband $0.1 - 12$ Hz.

The stimuli envelopes are squared and smoothed by passing through a square window filter of width 15.6 milliseconds. Then, they are downsampled to 64 Hz and followed by a cubic-root compression.

These final preprocessed 1D stimuli and 128D EEG data are used for the further intra-subject and inter subject analysis methods.

2.2.2 Music - EEG dataset (NMED-H)

The Naturalistic Music EEG Dataset - Hindi (NMED-H) dataset is used for the analysis on a music-EEG dataset. It is an open source dataset, and an extension to the NMED-T dataset [60]. 4 versions of 4 full length Hindi pop songs are the music stimuli provided to subjects. The four versions are normal, reversed, phase-scrambled and measure-shuffled. In the normal version, the songs are played in their natural order. In the reversed version, each song is played in temporally reverse order. The samples are shuffled and played in the measure-shuffled version. Each stimulus is of approximately 4.5 minutes in length.

The EEG recordings are recorded from 48 adults listening to a subset of 4 stimuli from 16 naturalistic music stimuli. Each subject is provided with 2 trials of 4 stimuli, and each stimulus is provided to 12 subjects. The stimuli are presented with a sampling frequency of $44,100$ Hz. The EEG are recorded at a sampling frequency of 1 kHz with 125 electrodes at the scalp. The EEG data are recorded using the Electrical Geodesics, Inc. (EGI) GES 300 platform. Each recording is filtered between 0.3-50 Hz using EGI Net Station zero-phase filters and downsampled to 125 Hz. NMED-H consists of three differently preprocessed EEG data. We use the "Clean EEG" recordings. Here, the EEG data are cleaned and aggregated on a per-stimulus, per-listen basis. The details of data acquisition and preprocessing are given in Blair Kaneshiro [61].

From the 1D stimuli envelopes, acoustic features are extracted as discussed by Gang et al. [62]. Music Information Retrieval (MIR) toolbox, Version 1.7.2 [63] is used to extract the acoustic features. As proposed by Alluri et al. [64], 20 stimuli features are extracted in 25ms analysis windows with a 50% overlap between frames [64, 65]. This yields a final sampling rate of 80 Hz for the stimuli features. The 20 stimuli features are: zero crossing rate, spectral centroid, high/low energy ratio, spectral spread, spectral roll-off, spectral entropy, spectral flatness, roughness, RMS energy,

broadband spectral flux, and spectral flux for 10 octave-wide sub-bands.

The principal component analysis (PCA) is performed on the 20D features of each stimulus to obtain a 1D representation (PC1). The two individual features, root mean square (RMS) and spectral flux, along with the PC1 are chosen to obtain a 3D representation for the stimuli. The RMS and spectral flux features reflect the amplitude and timbre of the stimuli. As a result, the EEG responses are also re-sampled to the sampling rate of the acoustic features (80 Hz).

2.3 Experiments Setup

For the speech-EEG dataset, the preprocessed 1D stimuli envelopes and 128D EEG responses are used for the experiments. For the NMED-H dataset, the 1D stimuli envelopes and each dimension of the 3D preprocessed stimuli are used with the 125D Clean EEG recordings.

From the speech dataset, stimulus-response data of 8 subjects are considered to perform the experiments. Each subject's data contains 20 sessions with each session of approximately 160 seconds. For all the methods, 20 cross-validation experiments are performed with 18 sessions used for training, one session for validation and the remaining session for testing. As the sampling rate is 64 Hz, the approximate number of instances for the linear/deep model training per subject is about $18{*}64{*}160 \approx 185$k. These 6 sets of stimulus-response data with a common stimulus is used for both the intra-subject and inter-subject analyses.

Two set of experiments, that is, intra-subject and inter-subject experiments, are performed on the NMED-H dataset. For the intra-subject experiments, each subject's EEG responses are aggregated resulting in 48 sets of stimulus-response data. For the

inter-subject experiments, EEG data are aggregated based on the common stimulus, and it results in 16 sets of stimulus-response data. As each stimulus is presented to 12 subjects, each pair of stimulus-response data, for a given stimulus, has EEG readings from 12 subjects for the inter-subject analysis. Later, each subject's intra-subject analysis is performed separately.

For both the experiments, the data are split into $90 - 5 - 5$ for training, validation and test respectively. It results in about 155k samples for training and 8.5k samples for testing and validation, for each subject in the intra-subject analysis, and 38k samples for training and 2k samples for testing and validation for each subject per stimulus for the inter-subject analysis.

2.4 Performance Metric

The performance of the analysis methods in our experiments are evaluated using two standard metrics. Primarily, we use the Pearson correlation coefficient between the transformed signals (EEG recordings and the stimuli) on the held-out test data.

For intra-subject analysis, the performance of linear CCA is measured using the Pearson correlation coefficient ρ (as defined in equation 2.6), and that of deep CCA is measured using the equation 3.1 (defined in Chapter 3) which measures the Pearson correlation coefficient between the new representations. For inter-subject analysis, the performance of linear MCCA is measured by the inter-set correlation of the new representations, ρ_{ISC} (as defined in equation 2.9), and that of deep MCCA (proposed method) is measured using the sum of Pearson correlation coefficients between pairs of the new representations (defined as ρ_{total} in equation 4.1 of Chapter 4).

The second performance metric is based on a classification efficiency of the aligned versus misaligned EEG-audio segments [13]. For the classification task, fixed-length

segments of the stimuli and corresponding EEG data are randomly selected, and the correlation between them is measured. If the audio and the EEG segments are from the same time instances, they must be aligned and generate a higher correlation score than the signals that are misaligned. These correlation values are used for determining the Cohen's d' statistic. It serves as a sensitivity index. It quantifies the new representations' ability to be matched with the corresponding stimulus-response pair based on the single value, ρ.

Let, the means of the matched and mismatched segments' correlation coefficients be μ_1 and μ_2 respectively. Let, σ_1^2 and σ_2^2 be their respective variances. Then, the Cohen's d' statistic is measured as :

$$d' = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{1}{2}\left(\sigma_1^2 + \sigma_2^2\right)}} \tag{2.18}$$

In order to quantify the improvements provided by our proposed methods, we perform statistical significant tests. One-tailed pairwise t-tests are performed on the correlation scores from baseline and proposed work. This is used to infer whether the two methods' results are intrinsically different.

In the next two chapters, we discuss the intra-subject and inter-subject analysis on both the speech-EEG and music-EEG datasets. Their performances are compared to the respective baselines using the metrics discussed above. And effects of various architecture choices are also studied.

Chapter 3

Intra-Subject Analysis

This chapter deals with the audio-EEG stimulus response analysis for each subject. Typically, the linear CCA is used as a hybrid model for stimulus response correlation (SRC) analysis. A CCA model is learnt for each subject separately.

Let, an auditory stimulus $\mathbf{S} \in \mathbb{R}^{T \times d_S}$ is provided to a subject, and the corresponding EEG $\mathbf{R} \in \mathbb{R}^{T \times d_R}$ response is recorded. The linear CCA model projects both the signals onto a common subspace of d dimensions such that the final representations are highly correlated to each other.

The deep variant of the CCA has two deep Neural Networks (DNNs). One DNN gets the stimulus $\mathbf{S}$ as the input, whereas the second DNN gets the response $\mathbf{R}$. Both DNNs project their inputs onto a common subspace (of d dimensions) while they are trained to a cost function which maximizes the correlation coefficient between the two outputs.

In this chapter, we discuss the mathematical background of deep CCA model. Later, we discuss our proposed deep CCA methods, and compare the results with their linear counterparts.

3.1 Deep Canonical Correlation Analysis

Andrew et al. [55] first proposed the extension to the linear transformation learning based CCA analysis using deep learning based CCA. The two input sets of vectors are passed through a pair of feed-forward connections to undergo a set of non-linear transformations. The outputs of each network are the final representations on which the correlation coefficient is computed. The neural networks are trained to maximize the correlation cost.

Let $f_1(\cdot)$ denotes the series of non-linear transforms performed by the first neural network on $\mathbf{x}$. Similarly, let $f_2(\cdot)$ denotes the second network that non-linearly transforms $\mathbf{y}$. Let, all trainable parameters of the first neural network be denoted as $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be that of the second network.

We need to find the optimal neural networks $f_1(\cdot)$, $f_2(\cdot)$ with trainable parameters $\boldsymbol{\theta}_1^*$, $\boldsymbol{\theta}_2^*$ respectively, such that their outputs $f_1(\mathbf{x}\ ;\ \boldsymbol{\theta}_1)$, $f_2(\mathbf{y}\ ;\ \boldsymbol{\theta}_2)$ are highly correlated. It can be formulated as:

$$(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*) = \underset{(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\operatorname{argmax}}\ \rho\left(f_1\left(\mathbf{x}\ ;\ \boldsymbol{\theta}_1\right), f_2\left(\mathbf{y}\ ;\ \boldsymbol{\theta}_2\right)\right) \tag{3.1}$$

Let, d be the the dimensionality of the outputs of the two neural networks and a batch of m examples from each of the ($\mathbf{x}$, $\mathbf{y}$) are used in training. Let $\mathbf{H}_\mathbf{x}, \mathbf{H}_\mathbf{y} \in \mathbb{R}^{d \times m}$ denote the matrices whose columns are the feed-forward network outputs from the first and second network respectively.

Let, $\bar{\mathbf{H}}_\mathbf{x} = \mathbf{H}_\mathbf{x} - \frac{1}{m}\mathbf{H}_\mathbf{x}\mathbf{1}$ and similarly, $\bar{\mathbf{H}}_\mathbf{y} = \mathbf{H}_\mathbf{y} - \frac{1}{m}\mathbf{H}_\mathbf{y}\mathbf{1}$ denote the centred data matrices, where $\mathbf{1}$ is an all-1s matrix of dimension $m \times m$. Now, the covariance matrices

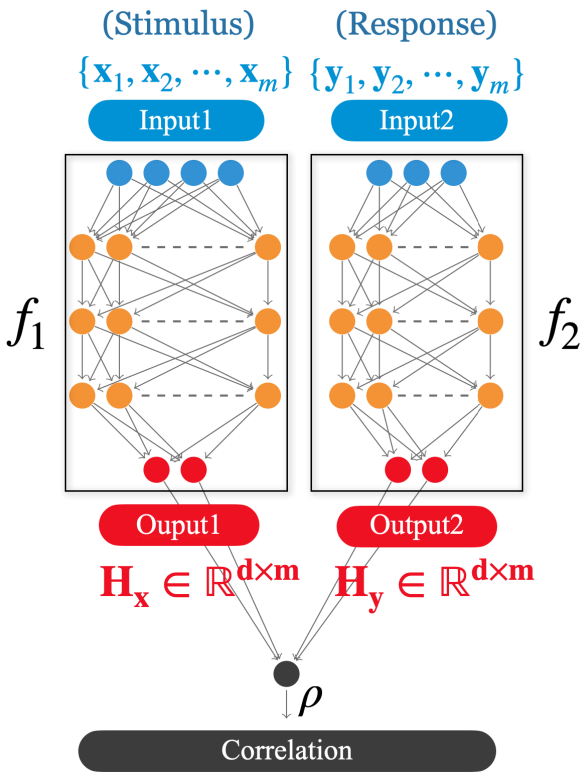


Figure 3.1: The deep CCA model. f_1 takes $\mathbf{x}$ as input and f_2 takes $\mathbf{y}$ as the input. They obtain final representations as the columns of the matrices $\mathbf{H_x}$ and $\mathbf{H_y}$ trained to be highly correlated. When applied for stimulus-response data, one network can take stimulus features as input and the other takes response features.

of $\bar{\mathbf{H}}_\mathbf{x}$ and $\bar{\mathbf{H}}_\mathbf{y}$ are given as

$$\begin{aligned}
\hat{\mathbf{C}}_{\mathbf{xx}} &= \frac{1}{m}\bar{\mathbf{H}}_\mathbf{x}\bar{\mathbf{H}}_\mathbf{x}^\top + r_1\mathbf{I} \\
\hat{\mathbf{C}}_{\mathbf{yy}} &= \frac{1}{m}\bar{\mathbf{H}}_\mathbf{y}\bar{\mathbf{H}}_\mathbf{y}^\top + r_2\mathbf{I} \\
\hat{\mathbf{C}}_{\mathbf{xy}} &= \frac{1}{m}\bar{\mathbf{H}}_\mathbf{x}\bar{\mathbf{H}}_\mathbf{y}^\top
\end{aligned} \tag{3.2}$$

where $r_1, r_2 > 0$ are the small regularization parameters so that the covariance matrices are positive definite and $\mathbf{I}$ is the identity matrix.

The total correlation between the outputs $\mathbf{H_x}$ and $\mathbf{H_y}$ can be formulated as [55]:

$$\rho\left(\mathbf{H_x}, \mathbf{H_y}\right) = \text{trace}\left(\mathbf{T}^\top \mathbf{T}\right)^{1/2} \text{ where } \mathbf{T} \triangleq \hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2} \hat{\mathbf{C}}_{\mathbf{xy}} \hat{\mathbf{C}}_{\mathbf{yy}}^{-1/2} \tag{3.3}$$

It can be shown [55] that the gradient of $\rho\left(\mathbf{H_x}, \mathbf{H_y}\right)$ is given by,

$$\frac{\partial \rho\left(\mathbf{H_x}, \mathbf{H_y}\right)}{\partial \mathbf{H_x}} = \frac{1}{m-1}\left(2\nabla_{\mathbf{xx}} \bar{\mathbf{H}}_x + \nabla_{\mathbf{xy}} \bar{\mathbf{H}}_{\mathbf{y}}\right) \tag{3.4}$$

where

$$\begin{aligned} \nabla_{\mathbf{xy}} &= \hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2} \mathbf{U} \mathbf{V}^\top \hat{\mathbf{C}}_{\mathbf{yy}}^{-1/2} \\ \nabla_{\mathbf{xx}} &= -\frac{1}{2} \hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2} \mathbf{U} \mathbf{D} \mathbf{U}^\top \hat{\mathbf{C}}_{\mathbf{xx}}^{-1/2} \end{aligned} \tag{3.5}$$

where $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{D}$ are obtained from the singular value decomposition of $\mathbf{T}$ as $\mathbf{T} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$. Similar expression can be obtained for gradient with respect to $\mathbf{H_y}$. These gradients are backpropagated to learn the model parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ of the two neural networks $f_1(\cdot)$ and $f_2(\cdot)$. The gradient ascent update for each network's trainable parameters can be represented as

$$\boldsymbol{\theta}_j^{t+1} = \boldsymbol{\theta}_j^t + \eta \frac{\partial \rho\left(\mathbf{H_x}, \mathbf{H_y}\right)}{\partial \boldsymbol{\theta}_j^t} \tag{3.6}$$

where η is the learning rate of the parameters' updates.

Andrew et al. [55] showed that the deep CCA model improved the correlation between left and right halves of MNIST images can be increased significantly over the linear CCA model. This experiment is under low noise condition. We analyze the impact of noise on the deep CCA performance. Specifically, we train and test the

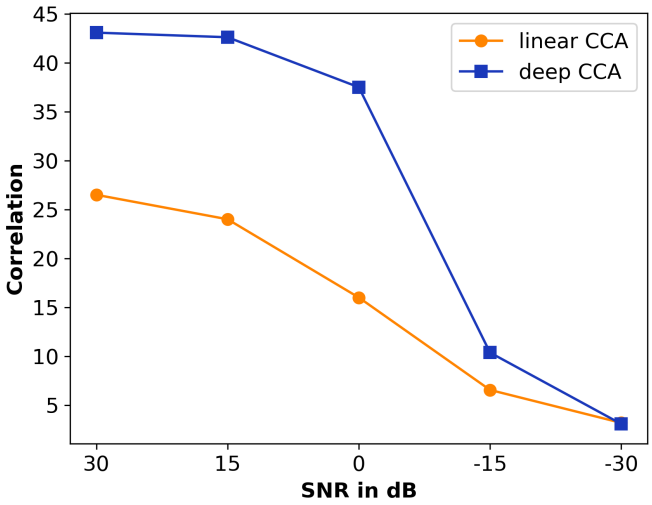


Figure 3.2: Linear and Deep CCA performance on the MNIST task for SNR varying from 30dB to -30dB. The outputs dimension is 50.

linear CCA and deep CCA for varying amounts of noise on the left half of the image (similar to the presence of noise in EEG recordings) and measure the performance. This analysis in presented in Figure 3.2.

Here, we use a 50D outputs for both the linear and deep CCA. As the amount of noise increases, the correlation drops significantly for both the models in the presence of noise. The deep models no longer have an advantage over the linear models under noisy conditions below -15dB. The difficulty in modeling noisy data proves to be challenging for the DCCA methods when the deep CCA model is applied on the EEG data. This is tackled by applying the dropout strategy in the deep CCA models. By incorporating various levels of dropouts in the deep CCA models, we show that the deep CCA outperforms linear CCA in the noisy conditions of EEG.

3.2 LCCA and DCCA Methods

The LCCA methods are performed on the preprocessed 1D stimulus and d_R dimensional EEG recordings. The stimuli are preprocessed as proposed by the baseline methods [13, 66]. Cheveigné et al. [13] have proposed three LCCA methods: LCCA1, LCCA2 and LCCA3. We propose deep methods homologous to the three LCCA methods.

1) **DCCA1** The 1D preprocessed stimuli feature are delayed by a time lag of 40. This converts the stimuli features to 40D. The d_R dimensions EEG data are provided to a Principal Component Analysis (PCA) that projects them onto a 40D subspace. The 40D stimuli and response are provided to the Deep CCA model as inputs.

2) **DCCA2** Here, the d_R dimensions EEG data are provided to a PCA to get projected onto a 60D subspace. Then, delays of time lag 60 are applied to the 60D EEG data, yielding 600 dimensions. Then, a PCA transformation to 80D is applied. The 80D time delayed stimuli features and the 80D EEG data are provided to the Deep CCA model.

3) **DCCA3** The 1D preprocessed stimuli are provided to a dyadic bank of 21 FIR band-pass filters. The filters' characteristics (center frequency, bandwidth) are approximately uniformly distributed on a logarithmic scale. The impulse response duration of the 21 filters range from 2 to 128 samples (2 seconds) [13]. This transforms the 1D stimulus features to 21D representations.

   The d_R dimensional EEG recordings are provided to a Principal Component Analysis (PCA) that projects them onto a 60D subspace. The dyadic FIR filterbank is applied on these 60D features. It transforms the 60D EEG data to 1260D.

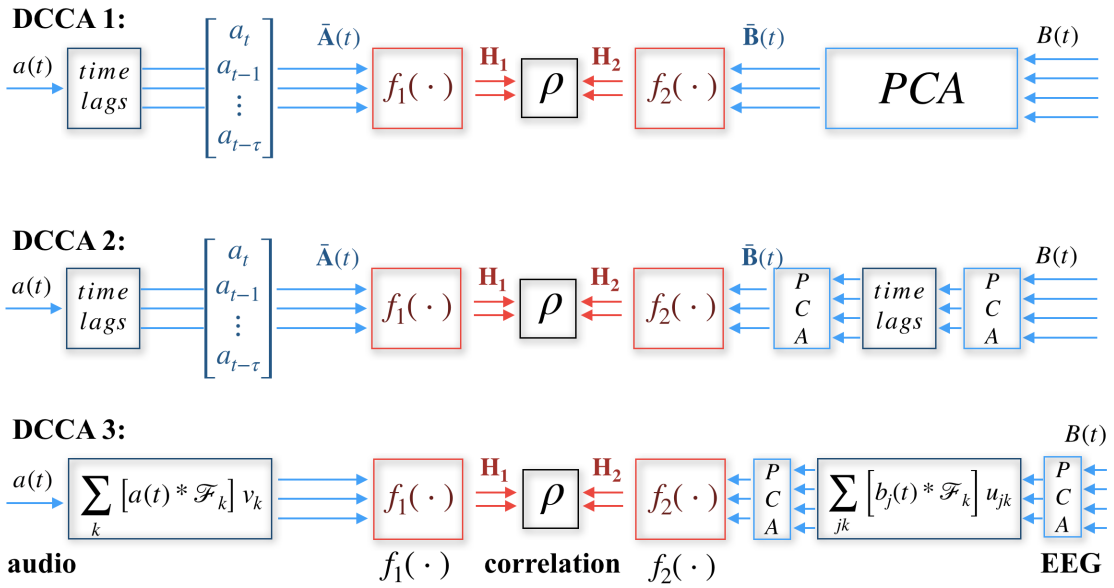


Figure 3.3: In DCCA1 method, the time lagged stimulus audio is provided to the Deep CCA module whereas EEG response is provided after performing PCA. In the DCCA2 method, the outputs of the EEG data passed through the PCA go through a set of time lags and one more PCA before being provided to the Deep CCA module. In DCCA3 method, both the audio inputs and the EEG outputs go through the filterbank of 21 FIR filters instead of delays.

A second PCA is applied to project them onto a 139D subspace. These 139D features are the final representations for the EEG data. The 21D stimuli data and the 139D EEG data are provided to the deep CCA model.

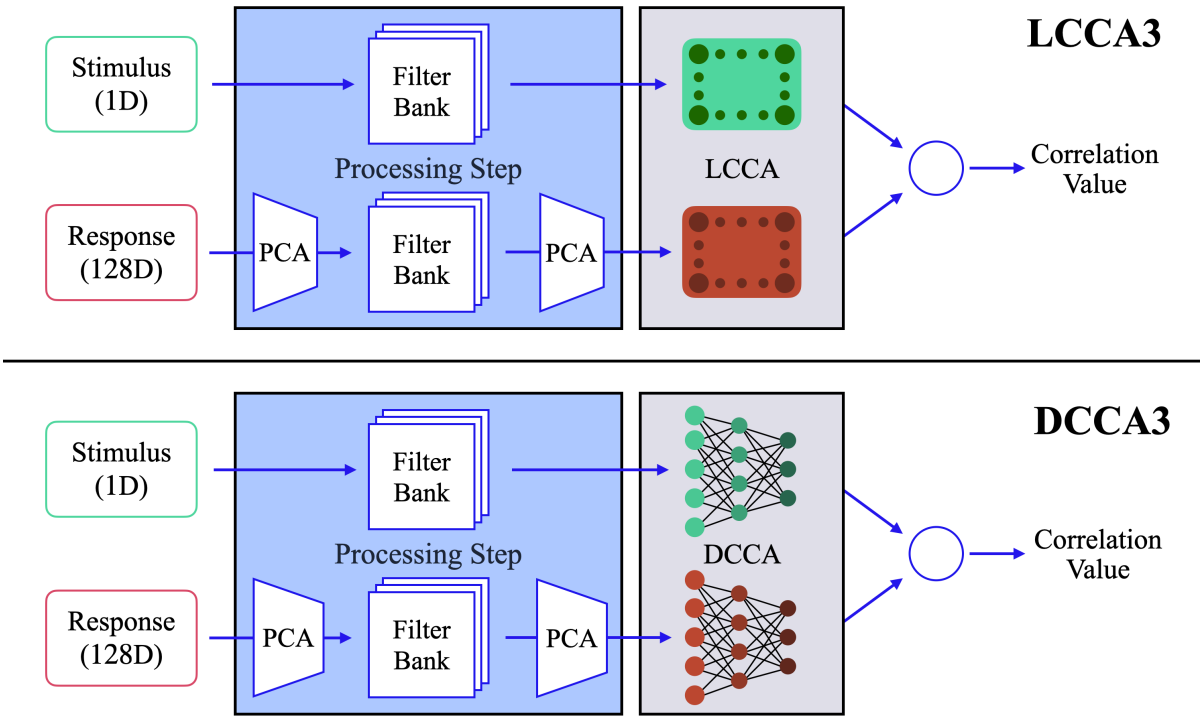


Figure 3.4: LCCA3 and DCCA3 methods for the speech-EEG dataset. The responses would be of 125D for the NMED-H dataset.

3.2.1 The Deep CCA Model Architecture

We have tried various architectures for the deep CCA model. A comparison of their performances is detailed in the later sections. The architecture with the best results is as follows. It has two identical DNN networks with 2 hidden layers. The first layer has 2048 units and the second hidden layer contains 1608 units. The output layer has d units, which varies with the dimension of the final representations. It is similar to the deep CCA architecture proposed for MNIST data [55]. The two hidden layers have "sigmoid" as their non-linear activation function. The output layer's non-linearity as "leaky-ReLU" [67] with a negative slope of 0.1 proved to be empirically better than the other activation functions like linear and sigmoid.

3.3 Results

The preprocessing and training-validation-test division for each subject data from the two datasets are performed as discussed in the Chapter 3. Now, for each subject, the CCA methods are trained and tested for the 20 cross-validation sessions. For overall performance for the 20 sessions, instead of direct averaging the Pearson correlation values which is mathematically incorrect, we perform a z-score based averaging as implemented in the Statsoft software [68].

3.3.1 Speech-EEG Dataset

All three DCCA methods are compared with their linear counterparts for a randomly chosen subject, and the results are presented in Figure 3.5. The LCCA3 method gives the highest correlation value among all the linear models (as reported by Cheveigné et al. [13]). Comparing the linear and deep CCA methods for the three configurations, all the deep variants outperform the linear methods consistently for all the 20 sessions. The DCCA3 method has shown the best correlation values (the average correlation is around 0.4). The absolute improvement in the Pearson correlation over the best linear model, LCCA3, for the DCCA3 method is about 9%.

The comparison of the LCCA3 and the DCCA3 methods for the 8 subjects randomly chosen from the speech-EEG dataset is shown in Figure 3.6. The DCCA3 method consistently improves over the linear method in all the evaluations. The absolute improvements in the correlation ranges from $3 - 9\%$ for these subjects.

We have tested the statistical significance of the improvements in correlations for the DCCA3 method over the LCCA3 method using a pairwise t-test on each subject. The improvement in correlation values are found to be statistically significant for 6 out of 8 subjects ($p < 0.05$).

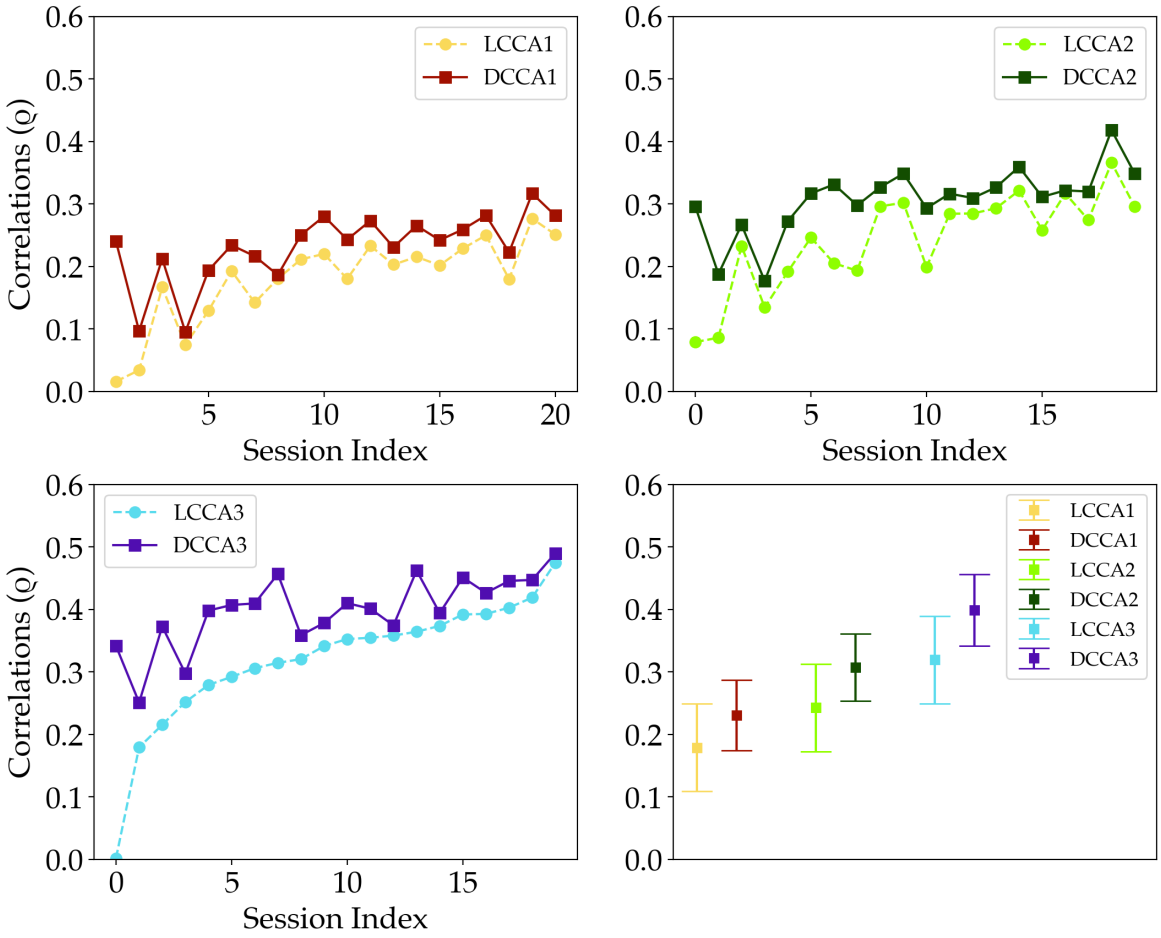


Figure 3.5: Comparison of linear and deep methods in the CCA1, CCA2 and CCA3 configurations, for a subject from speech-EEG dataset. The session indices (x-axis) are arranged in the non-decreasing order of the correlations obtained for the LCCA3 method. The first three plots show the results for 20 sessions. The last plot shows the average of the 6 configurations over the 20 sessions. A dropout of 10% is used in the DCCA methods for these experiments.

3.3.2 Music-EEG Dataset (NMED-H)

For the LCCA3/DCCA3 methods, the average correlation values for the 48 subjects from the NMED-H dataset is reported in Table 3.1. The results are reported for different stimuli versions - normal, shuffled, time-reversed and phase-scrambled; and

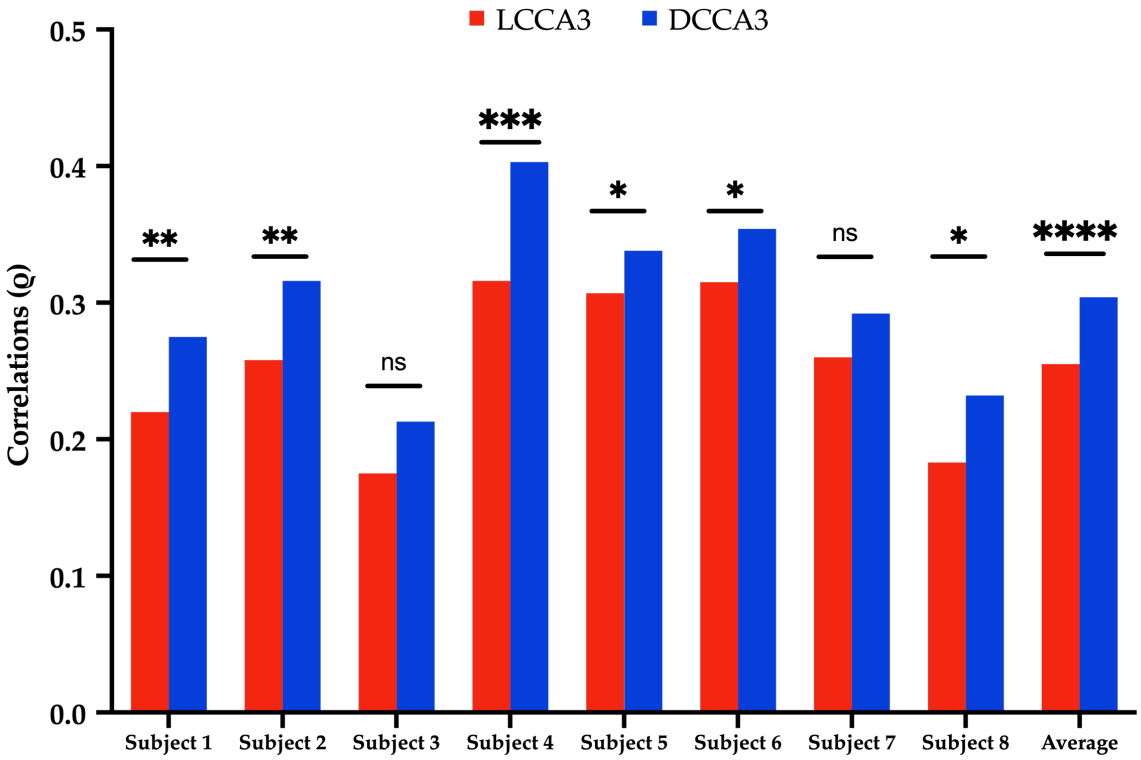


Figure 3.6: Comparing the average correlations of LCCA3 vs DCCA3 for 8 subjects randomly chosen from the speech-EEG dataset. A pairwise t-test is used to calculate the statistical significance (ns implies no significance ($p > 0.05$), $*$ implies $p \leq 0.05$), $**$ implies $p \leq 0.01$), $***$ implies $p \leq 0.001$), $****$ implies $p \leq 1e-4$)

stimuli features - envelope, PC1, RMS and spectral flux. The performance of LCCA3 and DCCA3 methods on the PC1 features of 48 subjects from the NMED-H dataset is shown in Figure 3.7. The DCCA3 method provides an average absolute improvements of 11% over the LCCA3 method.

To test the statistical significance of the improvements, a pair-wise t-test is performed on the NMED-H dataset. It shows that the improvements provided by the DCCA3 are statistically significant ($p < 0.05$).

Stimulus feature	Normal			Reversed		
CCA Method	LCCA3	DCCA3	[t-test]	LCCA3	DCCA3	t-test
Envelope	0.007	**0.118**	{1e-4}[3.7]	-0.003	**0.117**	{3e-5}[4.1]
PC1	-0.020	**0.077**	{9e-4}[3.2]	0.012	**0.105**	{1e-3}[3.0]
RMS	-0.004	**0.087**	{2e-4}[3.6]	0.008	**0.101**	{3e-4}[3.5]
Spectral Flux	0.008	**0.102**	{3e-4}[3.5]	-0.004	**0.113**	{9e-6}[4.5]

Stimulus feature	Phase-Scrambled			Shuffled		
CCA Method	LCCA3	DCCA3	[t-test]	LCCA3	DCCA3	t-test
Envelope	-0.052	**0.095**	{4e-7}[5.3]	-0.013	**0.134**	{3e-5}[4.2]
PC1	-0.016	**0.072**	{1e-3}[3.1]	0.030	**0.135**	{1e-3}[3.0]
RMS	-0.042	**0.091**	{4e-7}[5.2]	-0.025	**0.100**	{3e-5}[4.2]
Spectral Flux	-0.034	**0.107**	{3e-7}[5.3]	0.005	**0.123**	{3e-5}[4.2]

Table 3.1: Average correlation values for 48 subjects from the NMED-H Dataset in intra-subject analysis. A pairwise t-test between the LCCA3 and DCCA3 methods is reported as {p-value}[t-value].

3.4 Hyperparameters

In this section, we analyze the impact of the hyperparameters involved in the deep CCA models and their effect on the correlation metric.

The parameters analyzed are: dropout percentage, outputs' dimensionality and batchsizes. We have also tried various deep CCA model architectures. A learning rate of $1e{-}3$ and a batch size of 2048 are used in experiments where these hyperparameters are not explicitly mentioned.

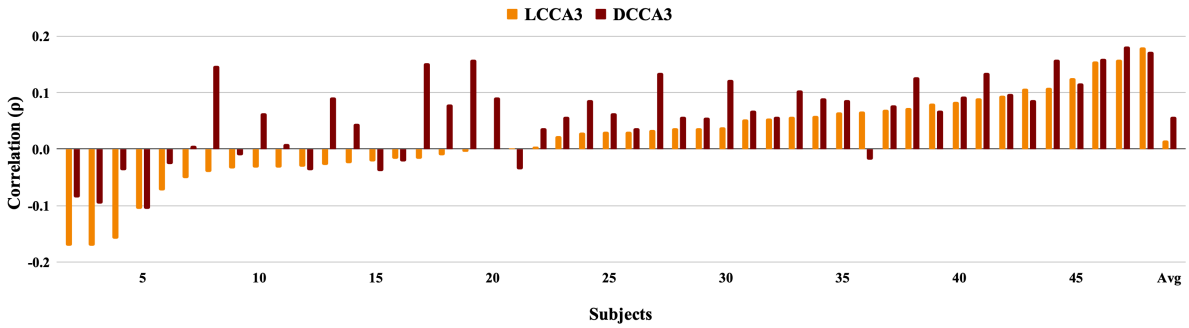


Figure 3.7: Comparing the LCCA3 and DCCA3 methods for PC1 stimuli features, of the 48 subjects from the NMED-H dataset. The correlations are arranged in the increasing order of the LCCA3 correlation values. The last column shows the average of the 48 subjects.

3.4.1 Dropouts

Given that the deep models are prone to over-fitting, particularly with the significant amount of noise in the EEG data, it is found that incorporating dropouts in the model training provides significant boost in the correlation performance (Figure 3.8). A dropout of 5% is found to provide the best average correlation (for the 20 cross-validation sessions). Hence, all the experiments use dropout in the deep CCA models training.

3.4.2 DCCA3 with 5D outputs

A CCA model can be trained to obtain multiple canonical components dimensions from the data. Similarly, the deep CCA model also can be trained for multiple output dimensions. To study the effectiveness of the DCCA methods, a comparison of the LCCA3 and the DCCA3 methods for 5 canonical correlation components is performed. These results are presented in Figure 3.9. It shows the improvements in the

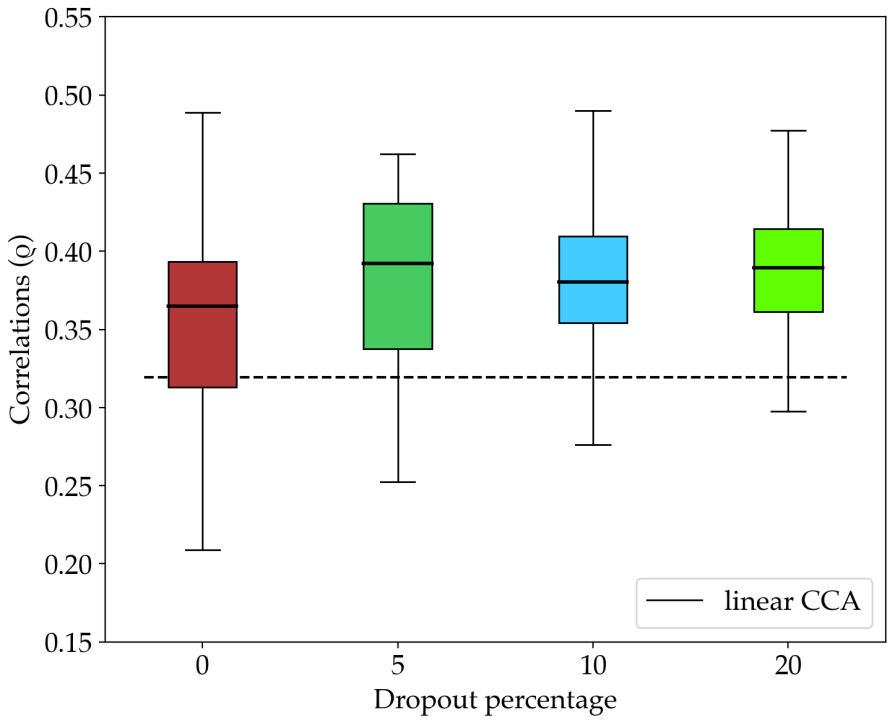


Figure 3.8: For a subject from speech-EEG dataset, the average correlation as function of the dropout regularization in the neural network. The horizontal dotted line is of the LCCA3 model.

correlation values per dimension, for the DCCA3 for each subject in the speech-EEG dataset. As seen in the figure, the DCCA3 method improves over the linear counterpart for 7 out of 8 subjects.

3.4.3 Batchsize

The effect of the batch size is also analyzed for the DCCA3 model. The average correlation values of 6 subjects from the speech dataset, for 20 cross-validation trials is reported in Figure 3.10. Given the noisy nature of the data, we find that the higher batchsizes (compared to typical choices of few hundreds in supervised classification

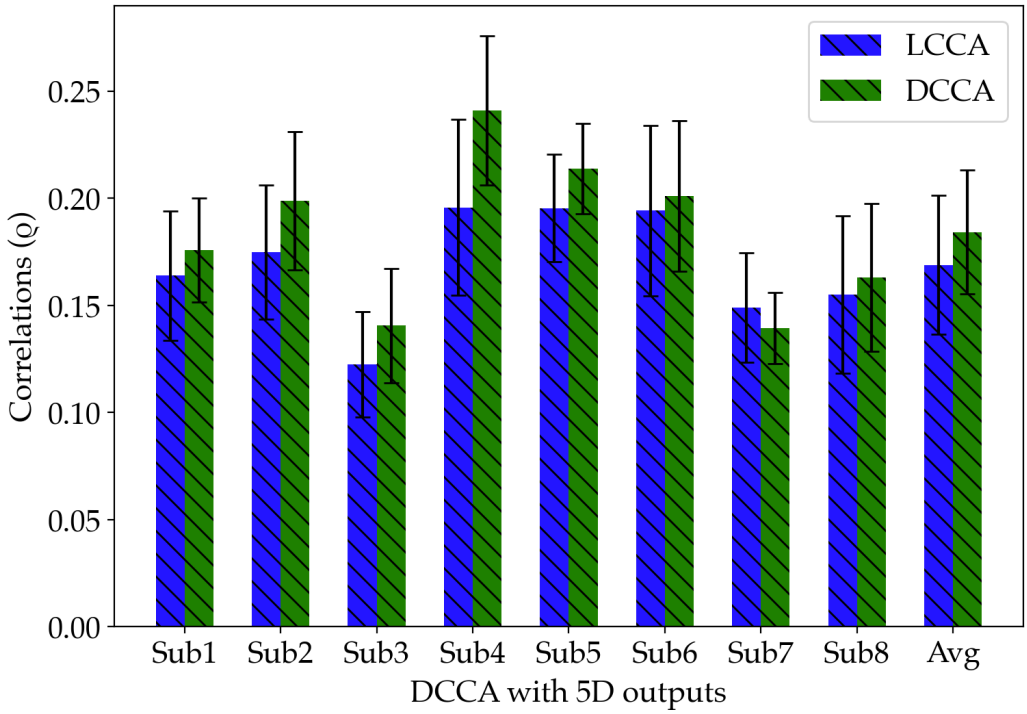


Figure 3.9: Comparison between the correlation per dimension of the final representations from LCCA3 and DCCA3 with outputs of 5D.

setting) are found to improve the final correlation value. The optimal batch size on the validation data is 2048.

3.5 Various DCCA Architectures

Various architectures for the deep CCA model are tried to study the performance of each deep model. Number of hidden layers (L) is varied from 2 to 5 and number of units (n_L) in each layer is also varied from 256 to $10,240$. A total of 9 architectures are explored for the deep CCA model. Their corresponding average correlation values are shown in Figure 3.11. As seen in this plot, increasing the number of layers

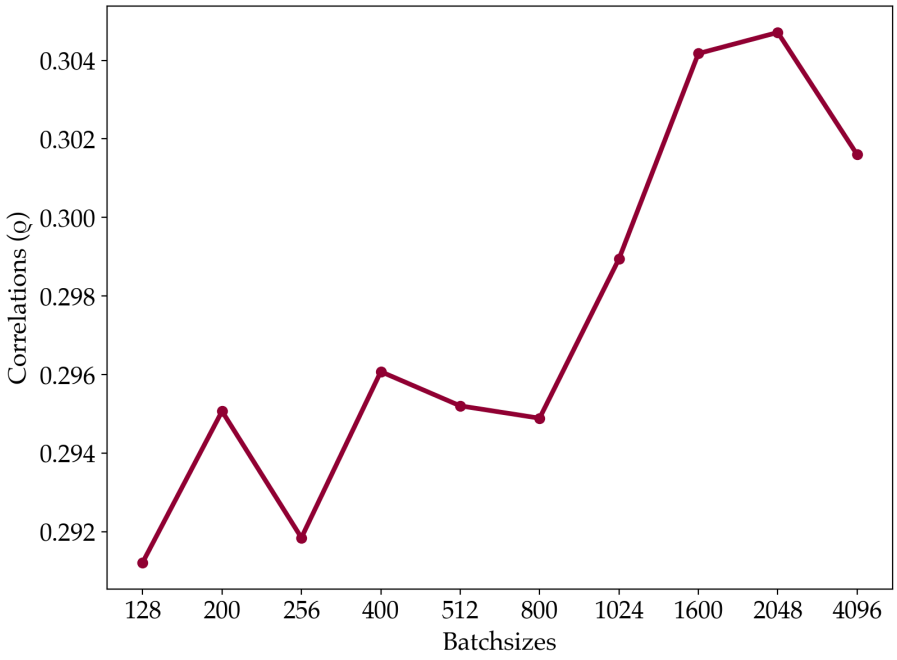


Figure 3.10: Impact of the batchsize on the average DCCA3 correlation value of all the 6 subjects from speech-EEG dataset.

degrades the correlation, as the model tends to over-fit. Overall, the two hidden layer architectures provided the best results.

This trend of decrease in performance with increase in the deep models' depth highlights the lack of sufficient audio-EEG data for each subject. The depth of a neural network increases its capacity, and given the lack of large amounts of audio-EEG data along with significant noise in the EEG, the neural networks tend to get overfit to the noise.

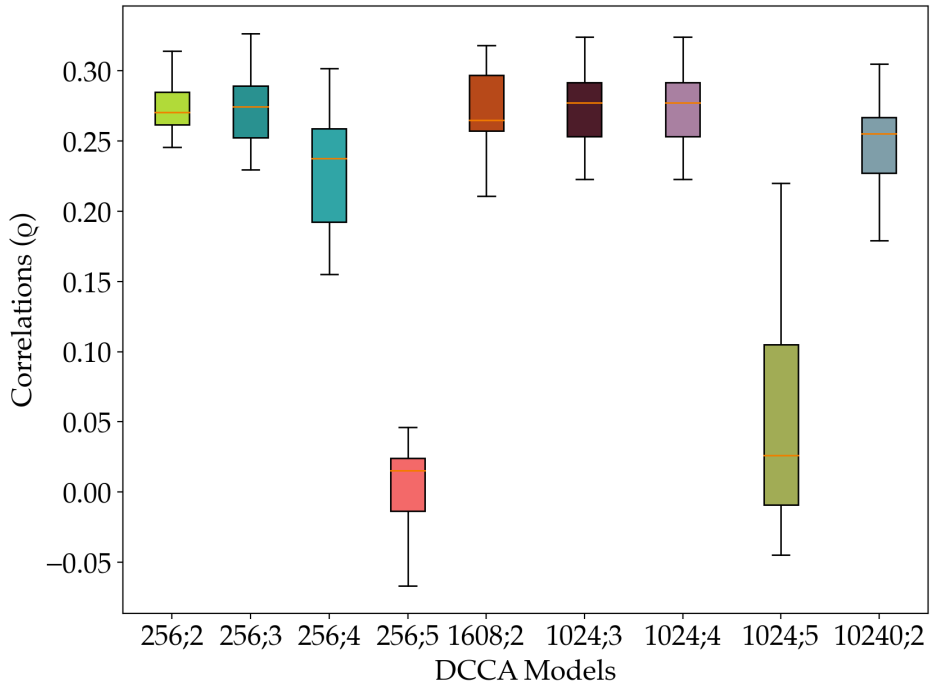


Figure 3.11: Different architectures are explored in the deep CCA models. The x-axis denotes "number of units per layer; number of layers"

.

3.6 Remarks

Hence, we compare the DCCA methods with their linear counterparts for intra-subject analysis on the two types of the datasets. The discussion shows the efficiency of the deep methods to align the auditory stimuli and their corresponding EEG data. The improvements provided by the deep methods are also shown to be statistically significant for both the datasets. Hence, the discussion shows that deep methods have the potential to become the de-facto standard for auditory attention decoding.

From hereon, the LCCA3 and the DCCA3 methods are used for the intra-subject analysis, and are referred to as LCCA and DCCA methods.

Chapter 4

Inter-Subject Analysis

This chapter deals with the aggregating EEG recordings from multiple subjects. The linear and deep CCA models discussed in previous chapter try to obtain better representations for each subject independently. As the machine learning models are data-driven and obtaining more EEG data for each subject is burdensome, we attempt to aggregate the common signals from EEG data of multiple subjects.

Linear MCCA method has proved to be successful to obtain improved representations [14]. Since all the responses are for the same stimuli, the MCCA tries to extract the signals related to the stimuli and suppress the components that are unrelated to the stimuli. Hence, the aggregation of multiple subjects' EEG data and extracting common signals among them can be referred to as a denoising step.

Let, a stimulus $\hat{\mathbf{S}} \in \mathbb{R}^{T \times d_s}$ be presented to N subjects, and their corresponding EEG responses are recorded as $\mathbf{R}_n \in \mathbb{R}^{T \times d_n}$ for $n = 1, \cdots, N$. We can safely assume $d_n = d_R$ for $n = 1, \cdots, N$ because generally all subjects are recorded under similar conditions. However, even if all d_n are not equal, the corresponding transforms can be modified accordingly and the final denoised representations are going to be of same dimension for all the EEG data.

To provide the temporal information, the stimulus $\hat{\mathbf{S}}$ can be transformed into a time-lagged version $\mathbf{S} \in \mathbb{R}^{T \times d_S}$.

4.1 Deep Multiway Canonical Correlation Analysis

The deep version of multiway CCA [47] attempts to generalize the CCA without the linearity assumptions of the linear MCCA.

The goal of the deep version of the MCCA is to derive optimal non-linear transforms for multiple (more than two) multivariates such that the transformed vectors are highly correlated. Let N multivariates be $\{\mathbf{x}_n \in \mathbb{R}^{d_n}\}_{n=1}^{N}$ and $f_n(\cdot)$ represent neural network, with trainable parameters $\boldsymbol{\theta}_n$, that transforms $\mathbf{x}_n$. The architecture is shown in Figure 4.1.

The N neural networks are trained to maximize the inter-set correlations defined as:

$$\rho_{\text{Total}} = \sum_{j=1}^{N} \sum_{k>j}^{N} \rho \left(f_j \left(\mathbf{x}_j \; ; \; \boldsymbol{\theta}_j \right), f_k \left(\mathbf{x}_k \; ; \; \boldsymbol{\theta}_k \right) \right) \tag{4.1}$$

Comparing with Equation 2.9, the correlation cost here is the summation of pairwise correlation coefficients. The optimum parameters are obtained as:

$$(\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_N^*) = \underset{(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)}{\operatorname{argmax}} \rho_{\text{Total}}(\mathbf{x}_1, \ldots, \mathbf{x}_N \; ; \; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \tag{4.2}$$

The backpropagation for each network is similar to the deep CCA model, as described in the deep CCA section 3.1. This architecture is referred to as DGCCA (Deep Generalized CCA) [47]. In the context of EEG, we have found that our proposed model generalizes the DGCCA and helps to provide better representations.

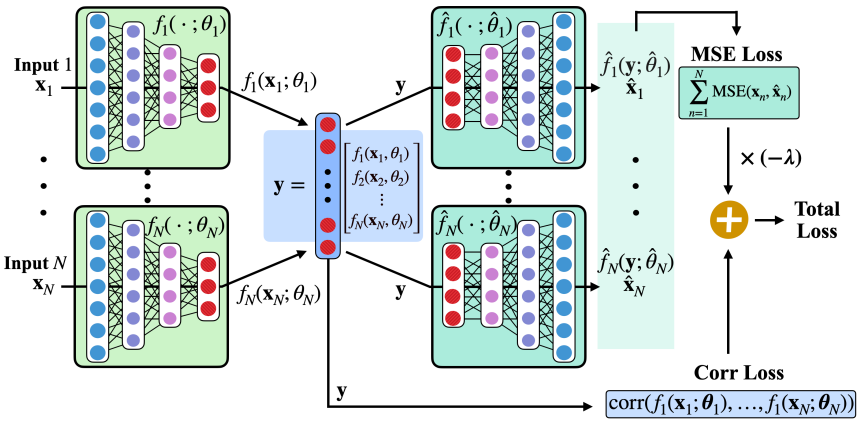


Figure 4.1: The DGCCA and the Proposed DMCCA Model. N inputs are provided to N encoders. All N encoder outputs are provided to the correlation loss and all N decoders. The decoders' outputs are provided to the reconstruction (MSE) loss. The model is trained to maximize the sum of the correlation loss and negative of the reconstruction loss.

The proposed model is referred to as DMCCA (Deep Multiway CCA). The DMCCA model has multiple autoencoders sharing encoded representations (N autoencoders for N dataviews respectively). Each random variable $\mathbf{x}_n$ forward propagates through the encoder part of an autoencoder, $f_n(\boldsymbol{\theta}_n, \cdot)$. All the encoded representations, $f_n(\mathbf{x}_n\ ;\ \boldsymbol{\theta}_n)$ are concatenated (denoted as $\mathbf{y}$), and provided to the decoders, $\hat{f}_n(\hat{\boldsymbol{\theta}}_n, \cdot)$. This shared encoder-decoder model allows the learning of non-linear transforms that align the multivariates.

The model is trained such that the joint cost function of correlation is maximized and the mean square error (MSE) in reconstruction is minimized. This cost function is formulated as,

$$E = \rho_{\text{Total}} - \lambda \sum_{n=1}^{N} \text{MSE}\left(\mathbf{x}_n, \hat{f}_n(\mathbf{y}; \hat{\boldsymbol{\theta}}_n)\right) \tag{4.3}$$

where ρ_{Total} is defined by Equation (4.1) and $\mathrm{MSE}(\cdot)$ is the average squared reconstruction loss. The hyperparameter λ balances the trade-off between maximizing the correlation metric and minimizing the MSE loss while learning the model parameters.

Training the autoencoders to maximize the cost function, E, obtains their optimum parameters ($\boldsymbol{\theta}_n, \hat{\boldsymbol{\theta}}_n$ for $n = 1, 2, \ldots, N$).

$$\left(\boldsymbol{\theta}_1^*, \ldots, \hat{\boldsymbol{\theta}}_N^*\right) = \underset{(\boldsymbol{\theta}_1, \ldots \hat{\boldsymbol{\theta}}_N)}{\operatorname{argmax}} E\left(\mathbf{x}_1, \ldots, \mathbf{x}_N \; ; \; \boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_1, \ldots, \boldsymbol{\theta}_N, \hat{\boldsymbol{\theta}}_N\right) \tag{4.4}$$

The model is trained using multiple multivariates with the cost metric defined. The correlation loss is independent of the decoder parameters $\hat{\boldsymbol{\theta}}_n$ while the $\mathrm{MSE}(\cdot)$ is a function of both the encoder parameters $\boldsymbol{\theta}_n$ and decoder parameters $\hat{\boldsymbol{\theta}}_n$.

Once the model is trained, the each random variable $\mathbf{x}_n$ is projected using the encoder $f_n(\mathbf{x}_n \; ; \; \boldsymbol{\theta}_n)$. After training the autoencoders, the encoders' outputs are taken as the final representations.

Compared to the DGCCA model [47], the additional decoders are used for incorporating the MSE regularization to the network's cost function. It is found that the ISC, among the encoders' outputs, improves in the presence of MSE regularization. The DGCCA can be viewed as a variant of DMCCA model with $\lambda = 0$.

4.2 LMCCA Method

The preprocessed EEG responses from N subjects and the time-lagged version of their time-lagged common stimuli, $\mathbf{S}$, are provided to a linear MCCA model. The MCCA model outputs the denoised representations (of d_OD) for each subject's EEG response and the common stimuli. The MCCA model returns $N + 1$ linear transformation matrices for the $N + 1$ inputs it receives. They are $\mathbf{D}_1, \; \cdots, \; \mathbf{D}_N \in \mathbb{R}^{\mathbf{d}_R \times \mathbf{d}_O}$ for the N EEG

responses and $\mathbf{D}_S \in \mathbb{R}^{\mathbf{d}_S \times \mathbf{d}_O}$ for the common stimuli.

4.3 DMCCA Method

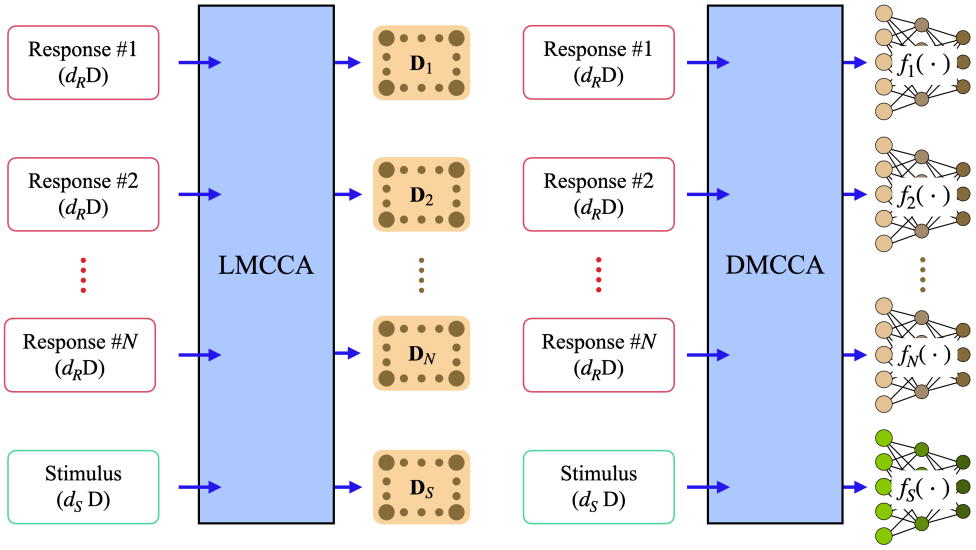


Figure 4.2: LMCCA and DMCCA models used for inter-subject EEG analysis. Here, D_1 to D_N are the linear transforms for N subjects respectively, and the D_S is the linear transform for the time-lagged stimulus. $f_1(\cdot)$ to $f_N(\cdot)$ and $f_S(\cdot)$ are the non-linear transforms for N subjects and the time-lagged stimulus respectively.

Our proposed DMCCA model follows a shared autoencoders style which generalizes the DGCCA model.

The preprocessed N EEG responses, along with the common stimuli $\mathbf{S}$ are provided to the DMCCA model. It obtains a d_OD denoised representation for all the $N + 1$ signals.

The DMCCA model has $N + 1$ autoencoders with shared encoder outputs. The architecture of the DMCCA model is shown in Figure 4.1. The encoder has two hidden

layers of 60 units each and the output layer has d_O units. The decoding part has two hidden layers of 60 and 110 units respectively. The decoders are absent in the DGCCA model.

The networks are trained to maximize the correlation among the encoder outputs and minimize the MSE between the decoders' outputs and its inputs. The MSE loss acts as a regularization term to decrease the effect of noise.

The amount of stimulus time-lag, d_S and the encoded outputs' dimension d_O are hyperparameters of the MCCA models. Their best values are selected for both the LMCCA and DMCCA methods separately. Figure 4.3 shows the block diagram for the inputs and outputs of the two MCCA methods. Both the methods take the N subjects' EEG responses and the time-lagged stimuli as inputs, and the linear MCCA returns $N + 1$ linear transforms ($\mathrm{D}_1,\ \cdots,\ \mathrm{D}_N,\ \mathrm{D}_S$) whereas the deep MCCA returns $N + 1$ non-linear transforms ($f_1,\ \cdots,\ f_N,\ f_S$).

4.4 Combinations of Inter- and Intra-Subject analyses

After the MCCA denoising, each subject's denoised EEG and the common stimuli (denoised) are considered for the intra-subject analysis. The intra-subject analysis can be performed using LCCA or DCCA method. This provides the final representations for each subject's EEG response and its stimulus.

Therefore, an MCCA method is used for the inter-subject analysis of aligning multiple subjects' EEG data followed by a CCA method to perform an intra-subject analysis of obtaining highly correlated EEG and stimuli representations for each subject separately. This results in four methods:

1. **LMLC**: LMCCA (on multiple subjects) + LCCA (for each subject)
2. **LMDC**: LMCCA (on multiple subjects) + DCCA (for each subject)

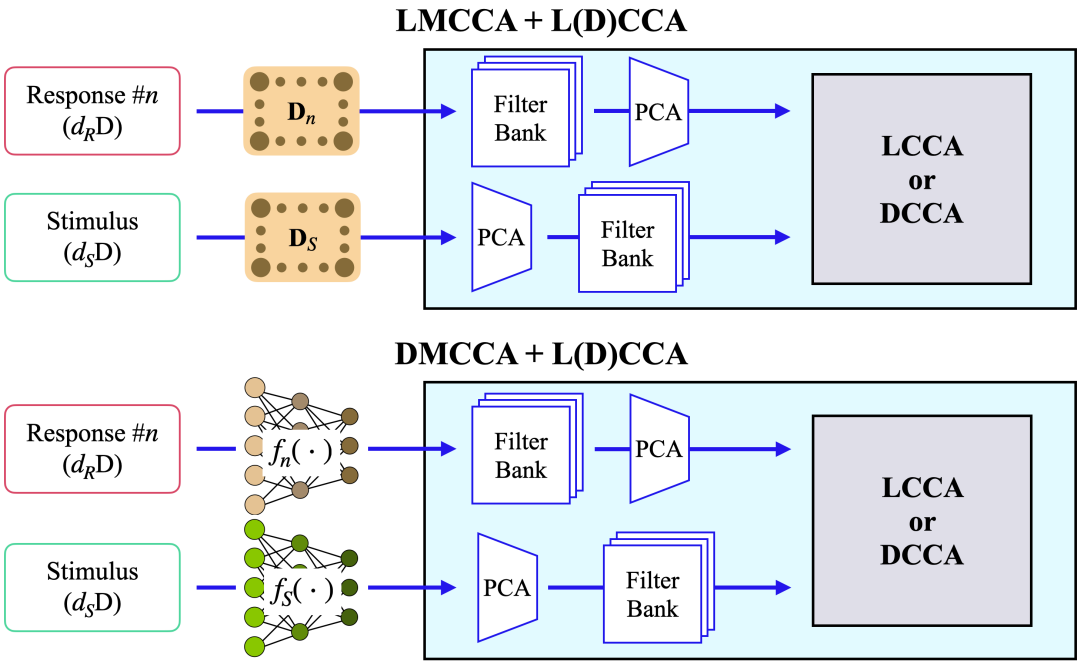


Figure 4.3: The four analysis methods - linear multiway CCA with linear CCA (LMLC), linear multiway CCA with deep CCA (LMDC), deep multiway CCA with linear CCA (DMLC) and deep multiway CCA with deep CCA (DMDC) methods.

3. **DMLC**: DMCCA (on multiple subjects) + LCCA (for each subject)

4. **DMDC**: DMCCA (on multiple subjects) + DCCA (for each subject)

The outputs are of d_R dimension for the LMCCA method and d_O dimension for the DMCCA method. After the MCCA step, the CCA step for each subject is performed as follows. The denoised EEG responses are provided to the dyadic FIR filterbank of 21 filters followed by a PCA to generate 139D vectors. The d_O D stimuli obtained are projected onto a 1D subspace using PCA, followed by the filterbank resulting in a 21D data. These steps make sure that the inputs to the CCA transforms are equivalent in both versions of MCCA (linear and deep).

4.5 Results

4.5.1 Speech Dataset

The performance of the combinations of inter-subject and intra-subject experiments on the speech-EEG dataset is shown in Table 4.1. The inter-subject alignment improves the correlation scores for both the intra-subject scores reported in Figure 3.6. The deep models consistently improve over the linear counterparts. The deep multiway CCA approach improves over the linear multiway CCA by an absolute correlation value of 8.8 % on the average. The improvements are also found to be statistically significant ($p < 0.05$) for all subjects.

4.5.2 Music Dataset

In the NMED-H dataset, each of the 16 stimuli is provided to 12 subjects for 2 trials. Hence, for each stimulus, the 12 subjects EEG data can be utilized to align themselves before performing the intra-subject analysis. The average correlation values after performing both the steps (MCCA and CCA methods) are reported in Table 4.2. The results are reported for different music conditions - normal, shuffled, time-reversed and phase-scrambled; and stimuli features - envelope, PC1, RMS and spectral flux. The DMDC improves over the LMLC method with an average absolute improvement of 29.3%.

4.5.3 Statistical Analysis : d-primes

The pairwise t-test results, comparing the linear and deep methods, are reported in the Table 4.1 for the speech task and in Table 4.2 for the music task.

Models	LMLC	LMDC	DMLC	DMDC	t-test
Sub1	0.262	0.271	0.375	**0.377**	{9.1e-7}[5.6]
Sub2	0.289	0.325	0.367	**0.374**	{5.1e-4}[3.6]
Sub3	0.160	0.177	0.258	**0.259**	{6.3e-5}[4.2]
Sub4	0.310	**0.378**	0.341	0.361	{3.6e-2}[1.8]
Sub5	0.309	0.354	0.389	**0.392**	{8.5e-5}[4.2]
Sub6	0.327	0.342	0.416	**0.420**	{4.6e-5}[4.4]
Sub7	0.275	0.289	**0.310**	**0.310**	{4.4e-2}[1.7]
Sub8	0.221	0.245	0.259	**0.272**	{2.8e-2}[2.0]
Average	0.270	0.299	0.339	**0.344**	{9e-14}[7.7]

Table 4.1: Comparison of the four methods - linear multiway CCA with LCCA (LMLC), linear multiway CCA with DCCA (LMDC), deep multiway CCA with LCCA (DMLC) and deep multiway CCA with DCCA (DMDC). A pairwise t-test between LMLC and DMDC methods (indicated as {p-value}[t-value]) is also reported where all the results are found to be significant ($p < 0.05$).

As mentioned in Section 2.2, a classification metric is also performed where stimulus-EEG segments are classified as aligned or misaligned based on the Pearson correlation measure. The LMLC and DMLC methods are performed on the stimulus-EEG segments and the corresponding correlation values are computed. Using the correlation score from the respective models, the Cohen's d-prime is computed on the correlation score. The d-prime statistics are presented in Figure 4.4.

This is performed separately for the speech-EEG and music-EEG datasets. The size of audio-EEG segments plays an important role in the classification task. The longer segments provide better match/mismatch classification. Figure 4.4 shows that

Stimulus feature	Normal				
MCCA Model	LMLC	LMDC	DMLC	DMDC	t-test
Envelope	0.076	0.146	0.344	**0.349**	{3e-26}[14]
PC1	-0.007	0.102	**0.384**	0.321	{1e-26}[14]
RMS	0.001	0.114	**0.341**	0.246	{3e-13}[8.5]
Spectral Flux	0.017	0.110	0.341	**0.343**	{2e-24}[13]

Stimulus feature	Reversed				
MCCA Model	LMLC	LMDC	DMLC	DMDC	t-test
Envelope	0.062	0.099	0.299	**0.384**	{1e-37}[21]
PC1	0.030	0.159	**0.360**	0.323	{2e-19}[11]
RMS	0.042	0.135	**0.318**	0.220	{1e-08}[6.2]
Spectral Flux	0.053	0.170	**0.340**	0.321	{1e-16}[10]

Stimulus feature	Phase-Scrambled				
MCCA Model	LMLC	LMDC	DMLC	DMDC	t-test
Envelope	0.042	0.092	**0.312**	0.299	{8e-26}[14]
PC1	0.012	0.166	0.262	**0.389**	{6e-21}[12]
RMS	0.020	0.108	0.176	**0.397**	{2e-07}[7.4]
Spectral Flux	0.038	0.207	0.340	**0.390**	{3e-22}[12]

Table 4.2: For NMED-H dataset, average correlation values for normal, time-reversed and phase-scrambled stimuli conditions in inter-subject analysis. A statistical significance test (t-test) between LMLC and DMDC methods is indicated as {p-value}[t-value].

Stimulus feature	Shuffled				
MCCA Model	LMLC	LMDC	DMLC	DMDC	t-test
Envelope	0.077	0.132	**0.341**	0.333	{4e-19}[11]
PC1	0.051	0.149	0.345	**0.347**	{5e-21}[12]
RMS	0.051	0.156	0.327	**0.345**	{1e-19}[11]
Spectral Flux	0.061	0.145	0.294	**0.322**	{8e-15}[9.2]

Table 4.2: For NMED-H dataset, average correlation values for the measure-shuffled stimuli condition in inter-subject analysis. A statistical significance test (t-test) between LMLC and DMDC methods is indicated as {p-value}[t-value].

the deep model improves over the linear model in all the cases except for 1 second segments in speech-EEG data. For longer segments, considerable improvements in the d-prime statistic are observed for the deep models.

4.6 Hyperparameters

We have discussed the effects of various hyperparameters involved in the deep CCA models for intra-subject analysis. Similarly, we analyze the impact of the hyperparameters on the correlation values of the deep MCCA models here.

We discuss the effect of dropout percentage, denoised outputs' dimension, context size of the stimuli and the MSE regularization. We have also tried various deep MCCA model architectures. A learning rate of $1e-3$ and a batch size of 2048 are used in the experiments. Unless specified otherwise, the stimuli time-lag is set to 60, encoders' outputs are set to 10D, dropout is set to 5%, and the MSE regularization parameter is set to 0.1.

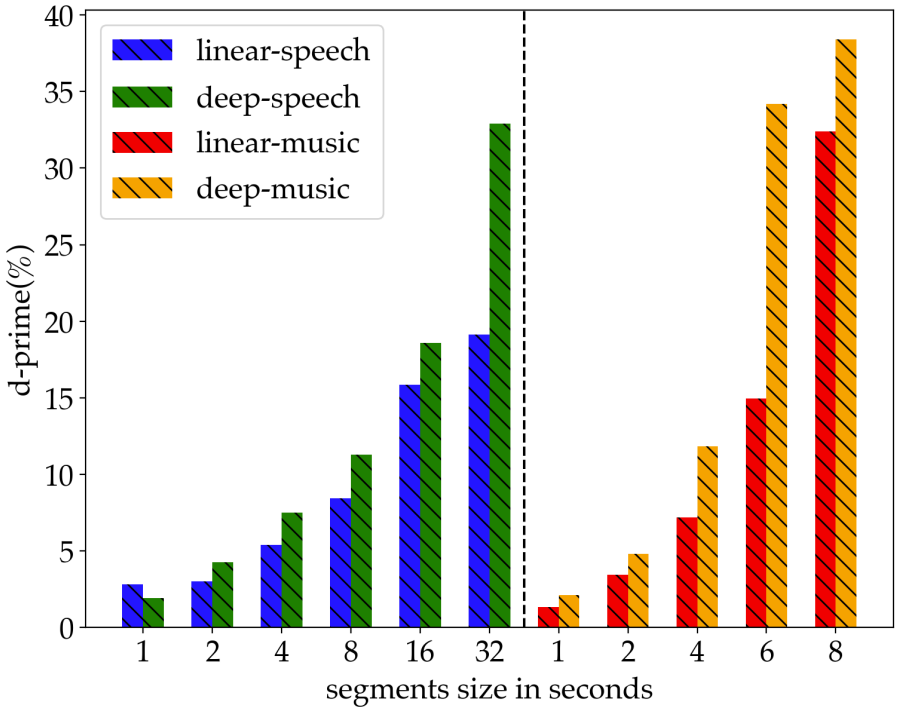


Figure 4.4: Comparing the d-prime metric for both the datasets for varying time length of the segments. The left half corresponds to speech-EEG dataset and the right half corresponds to music-EEG dataset. The linear-speech and linear-music correspond to the d-prime values for LMLC method's final representations of speech and music datasets respectively. Similarly, deep-speech and deep-music correspond to DMLC method for the two datasets.

4.6.1 Effect of Dropouts

Similar to the intra-subject analysis, we experiment with dropout percentage from $0 - 20\%$ in the deep MCCA model for the speech-EEG dataset. The average correlation values of the 6 subjects chosen from speech-EEG dataset for DMLC methods are shown in Figure 4.5. It shows the change in correlation values as the amount of dropouts inserted into the deep MCCA model of DMLC method. When there is no

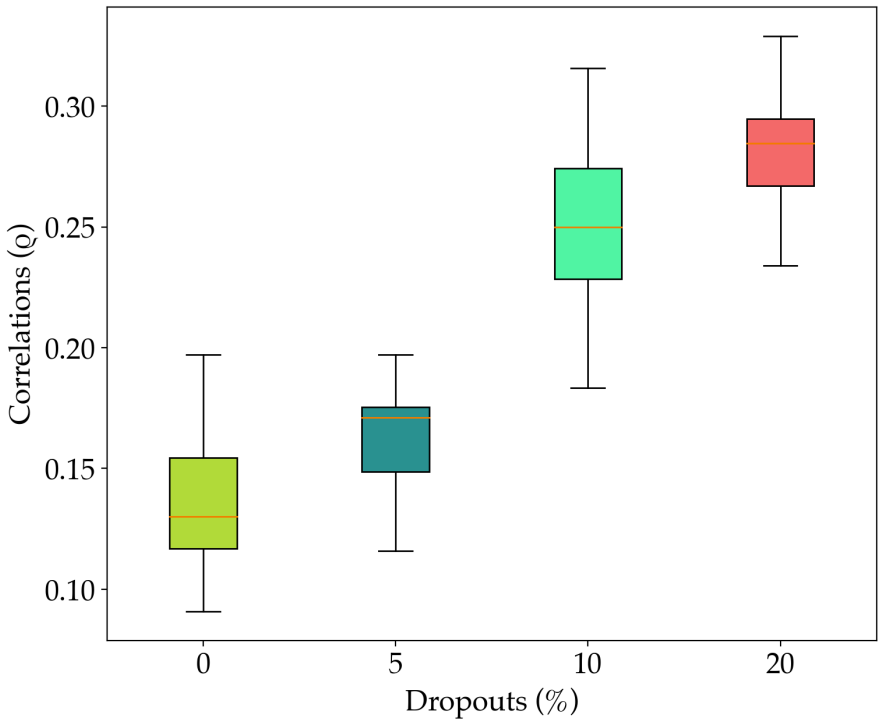


Figure 4.5: Effect of dropout on the DMLC method. For the 6 subjects from speech-EEG dataset, a DMLC method with the deep MCCA model as described in the section 4.3 is considered for the DMLC. The effect of dropout is compared on the average correlation of the final representations of all the subjects.

dropout, there is a tendency for the model to overfit. As the dropout increases, the effect of noise decreases on the deep MCCA model.

4.6.2 Effect of the Final Representations Dimension

Building the deep MCCA model architecture involves deciding the hyperparameter encoders' outputs dimensions d_O. The number of output dimensions, d_O, is varied from 2 to 128. The performance for different d_O values is shown in the Figure 4.6. The

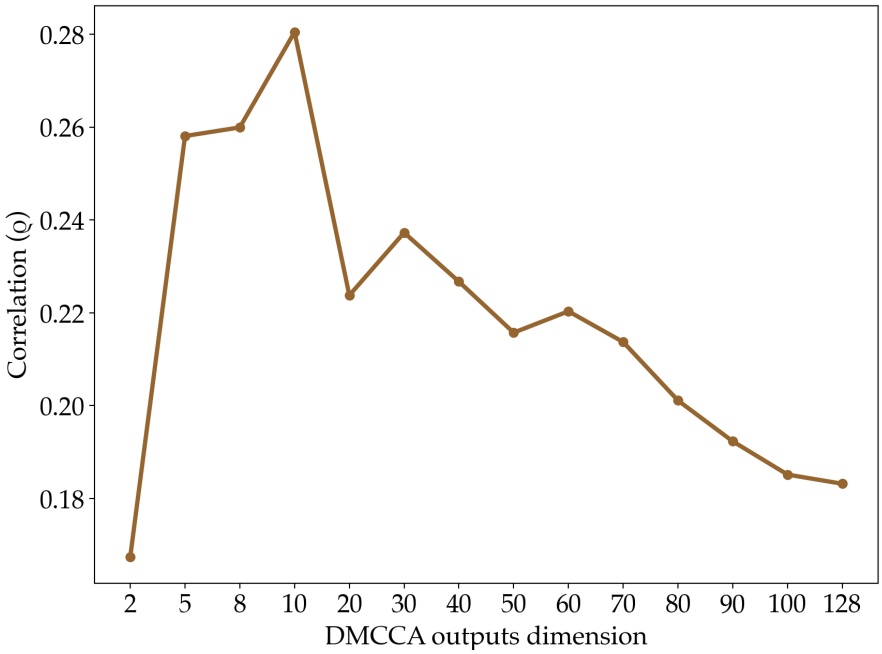


Figure 4.6: Effect of encoder output dimension on the DMLC method on the speech-EEG dataset. Changing the encoder outputs dimension for the deep MCCA model, the average correlation of the 6 subjects is compared. The deep MCCA model is as described in the section 4.3.

performance for each d_O is measured using the average correlation of the final representations after the DMLC method, for the 6 subjects from the speech-EEG dataset. The best performance is achieved for DMCCA model with encoder outputs of 10D. And this value of d_O is used in all the remaining deep MCCA experiments.

4.6.3 Effect of the Context size for the Stimuli Features

Introducing the time-lags to the stimulus features, d_S, helps the DMCCA model to capture the temporal information of the signals. While varying the d_S from 10 to 110, we have tested the DMLC method on the 6 subjects from the speech-EEG dataset. The

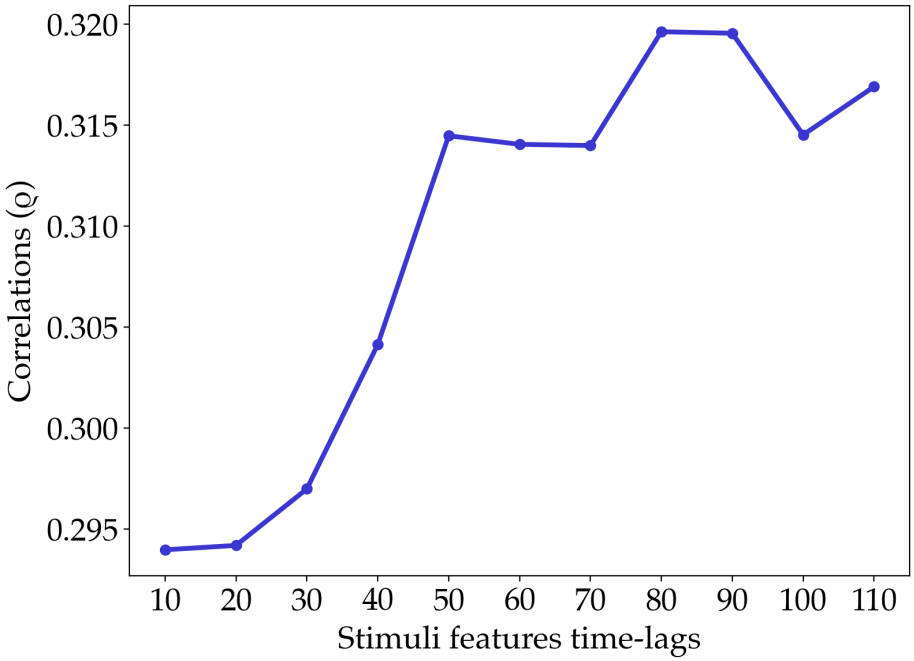


Figure 4.7: Effect of time-lags d_S on the stimulus features on the DMLC. The average correlation of the 6 subjects from the speech-EEG dataset is studied for different time-lags applied to the stimuli. The deep MCCA model used in the DMLC is as described in the section 4.3.

effect of d_S on the final correlation metric is shown in the Figure 4.7. The performance is measured using the average correlation of the 6 subjects' final representations. A stimulus time-lag of 80 gives the best performance for a deep MCCA model.

4.6.4 Effect of the MSE Regularization Parameter

In addition to the inter-set correlation loss present in the DGCCA model, a DMCCA model incorporates an MSE regularization loss. Varying the regularization coefficient (λ), we show that the additional loss helps the deep MCCA models to obtain better correlated representations. The effect of the λ on the average correlation of the DMLC

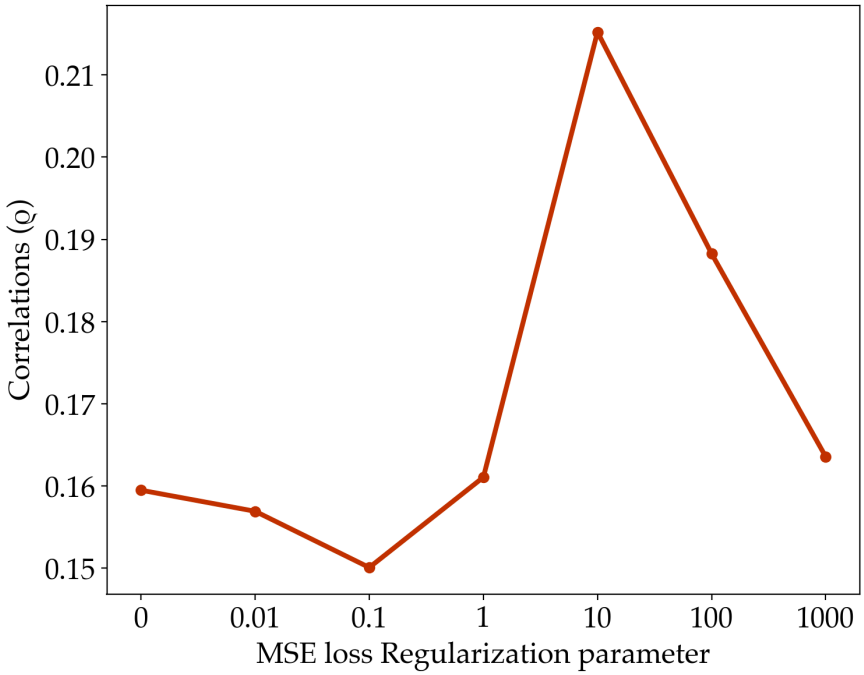


Figure 4.8: Effect of MSE regularization on the DMLC. The MSE regularization strength (λ) is varied from 0 to 1000 and the corresponding DMLC method's performance is measured. The deep MCCA model used in the DMLC is as described in the section 4.3.

final representations of 6 subjects from the speech-EEG dataset is studied in Figure 4.8. The value of λ is varied from from 0 to 1000. The results show that the regularization with $\lambda = 10$ gives the best performance for the speech-EEG dataset.

4.7 Remarks

Our experiments show that the DMCCA method can be used in the place of the linear MCCA method for intra-subject analysis on both the types of audio-EEG datasets. We have shown that the improvements are statistically significant and provide better

correlations for the classification task too. The combinations of inter-subject and intra-subject analyses show the effectiveness of the deep models to be used in the hybrid CCA models for audio-EEG.

Chapter 5

Extension and Conclusions

As an extension work, we have explored the efficiency of the speech reconstruction from EEG recordings.

5.1 Neural EEG-speech Translation

A backward model that reconstructs speech from the EEG recordings is studied in this section. The existing linear models, TRF, along with the recent machine learning architectures used for sequence-to-sequence translation, LSTM and transformers, are studied for this task. Experiments incorporating generative adversarial network (GAN) style objective function are also performed.

We have used the Speech-EEG dataset [25] for the task of speech reconstruction from EEG data. The EEG data are preprocessed to a sampling rate of 64 Hz, as discussed in Chapter 2. The EEG data are further processed similar to the intra-subject analysis (LCCA) methods, which project the EEG responses onto a 139D subspace. For the stimuli data, the speech signals are represented using three types of features: spectrogram, WORLD vocoder [69] and mel spectrogram.

For extracting the spectrogram features, we have obtained the spectrograms with 256 point FFT at a hop rate of 15.625 ms and 50% overlapping windows. Recent ECoG literature [70, 71] has shown that intelligible speech can be reconstructed from the ECoG recordings. These experiments by Akbari et al. [70] show that the WORLD vocoder [69] representations are suited for this task. Additionally, recent deep learning advancements have shown that audible speech can be reconstructed from the mel spectrograms using MelGANs [72]. The speech stimuli are represented using 80D features for mel spectrograms. Hence, we experimented with three kinds of stimuli representations: 129D spectrogram, $1027D$ $(1 + 513 + 513)$ WORLD vocoder [69] and 80D mel spectrogram features.

5.1.1 Transformer

A typical transformer [73] consists of a pair of encoder and decoder. The transformer takes a sequence as its input, generates a sequence from it, and is trained to predict the output sequence as close as possible to the target sequence. Each encoder layer has two sublayers: a multi-head self-attention and a feed-forward dense neural network. Both the sublayers contain a residual connection layer followed by layer normalization.

We have used a transformer encoder model for reconstructing the auditory stimuli using EEG responses. The transformer is trained with an objective function of MSE loss between the reconstructed stimuli features and the actual stimuli features. Figure 5.1 shows the transformer encoder model as the backward model for speech reconstruction from the EEG data. Further details about training are discussed in Section 5.1.4.

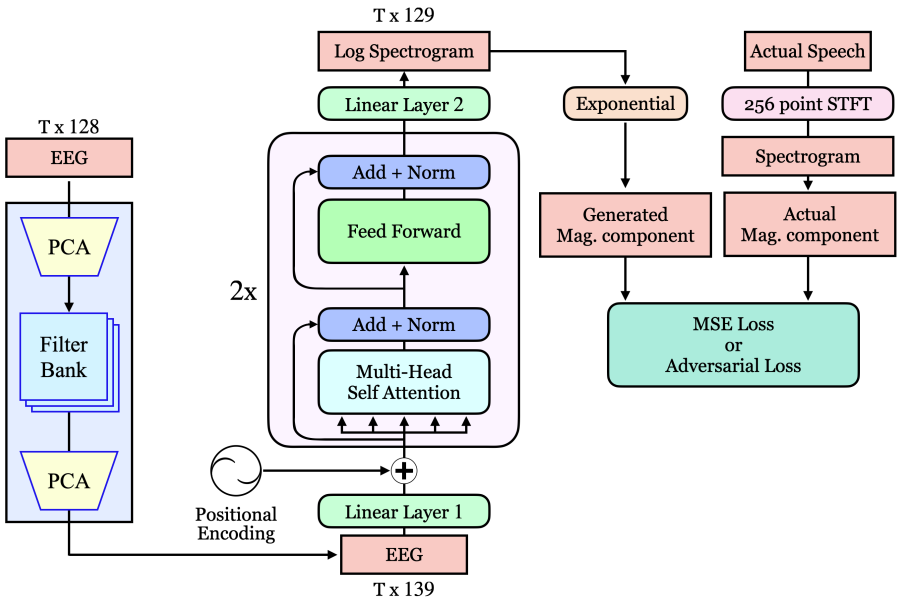


Figure 5.1: The pipeline of the backward model with the processed EEG as input and the objective function as the MSE loss or adversarial loss between the generated log spectrograms and the actual log spectrograms.

5.1.2 LSTM and Bi-LSTM

Long Short Term Memory (LSTM) models have been introduced by Hochreiter et al. [74] to tackle the problem of vanishing and exploding gradients. For the speech reconstruction from EEG task, we have explored LSTM and Bidirectional LSTMs which receive the EEG recordings as input, and generate the stimuli features as outputs.

5.1.3 Adversarial Loss Regularization

To reduce the noise and direct the output representations towards the natural stimuli features, an adversarial based regularization is incorporated into the objective function. To achieve this, a generative adversarial network (GAN) style model is built using the transformer / LSTM models.

A convolutional neural network is used as the discriminator. The generator produced stimuli features along with the actual stimuli features are used to train the discriminator, and the corresponding loss is used as a regularization to the MSE objective function of the transformer.

5.1.4 Architecture of the reconstruction model

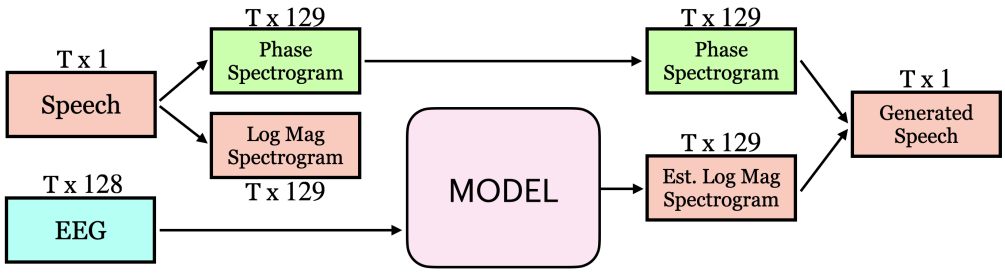


Figure 5.2: The pipeline of the backward model with the processed EEG as input which estimates the log magnitude spectrogram features of the speech.

The pipeline for the task of speech reconstruction task, using spectrogram features, from the EEG signals is as follows. The log-magnitude features are provided as targets and the inputs are 139D EEG features for the backward models. The speech spectrograms are divided into phase and log-magnitude features. The models are trained to estimate the log-magnitude speech features, which are multiplied with the actual phase features to reconstruct the speech. This is presented in the Figure 5.2.

When vocoder and log-mel spectrogram representations are utilized as the stimuli features, the model estimates the stimuli features which are directly provided to the synthesizer, i.e., the vocoder synthesizer or a pre-trained MelGAN [72], to reconstruct the speech.

The backward models are realized using transformers or LSTMs. For comparison,

TRF based linear model is also explored for the task. For the transformer based backward model, a 2 layer encoder with 2 heads self-attention and a 2 layer feedforward network with 512 units in each layer is used. And 2 layers LSTM and Bi-directional LSTM based models with hidden vectors of size 512D are studied.

For the GAN based training, the backward models act as the generator. A pretrained VGG-19 [75] is used as the discriminator, with all the parameters being trainable. A regularization parameter of $\lambda = 1e-3$ is used. Hence, the reconstruction model's objective function is designed as :

$$\mathrm{MSE}(\mathbf{G}(\mathbf{z}), \mathbf{x}) + \lambda E_{\mathbf{z}}[\log(1 - \mathbf{D}(\mathbf{G}(\mathbf{z})))] + E_{\mathbf{x}}[\log(\mathbf{D}(\mathbf{x}))] \tag{5.1}$$

LSD	Spectrogram		Vocoder		Mel	
GAN reg. (λ)	0	$1e-3$	0	$1e-3$	0	$1e-3$
TRF	22748.24	-	17337.9	-	15735.74	-
LSTM	6.304	6.275	4.606	4.622	8.543	8.568
Bi-LSTM	5.984	5.980	4.634	4.830	7.855	7.986
Transformer	6.397	6.421	4.957	5.043	8.568	8.687

Table 5.1: Performance of the four backward models (TRF, LSTM, Bi-LSTM, transformer) for the three stimuli features (Spectrogram, Vocoder, Mel), with and without GAN regularization.

5.1.5 Results

The performance of the backward models is measured as the log spectral distance (LSD) between the ground-truth and estimates speech signals. The experiments are

performed on one subject from the speech-EEG dataset and reported for the overall performance of the 20 cross-validation experiments. Four types of backward models are tried: TRF, LSTM, BiLSTM and transformer, for three types of stimuli features: spectrogram, WORLD vocoder features and mel spectrogram. A set of experiments with and without the GAN regularization is also studied. The three stimuli features are compared in the Table 5.1. The spectrograms of the estimated speech for the backward models with spectrograms as the stimuli features are shown in Figure 5.3.

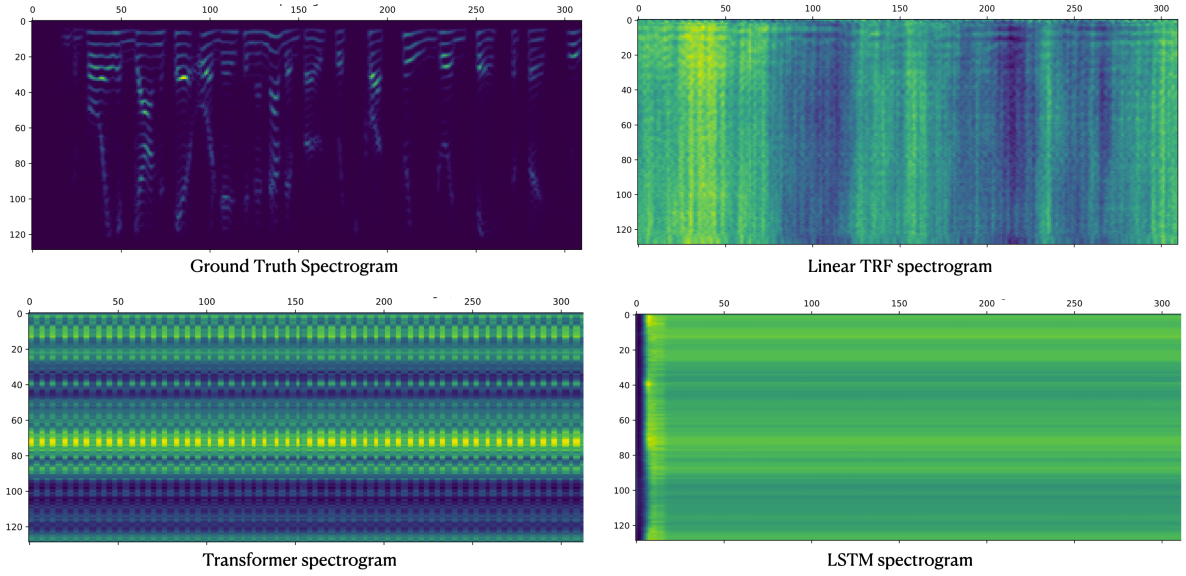


Figure 5.3: The magnitude spectrograms of the estimated speech for the backward models with spectrograms as the stimuli features. The top-left image shows the spectrogram of the ground truth speech. The top-right image shows the spectrogram of the TRF output, the bottom-left image shows that of the LSTM outputs and bottom-right shows that of the transformer outputs. A sample of almost 5 seconds is selected and its magnitude spectrogram is plotted.

5.1.6 Observations

The results show that the deep models offer a huge advantage over the linear model for the task of speech reconstruction. But, the spectrograms and the generated speech are not intelligible. To investigate the capacity of the transformer and LSTM models in analyzing the EEG recordings, we have trained an auto-encoder model using them and found that the LSTM autoencoder model had an advantage over the other models to capture the patterns in the EEG recordings. Next, we tried increasing the capacity of these models by increasing the depth. Increasing the number of layers decreased the speech reconstruction audibility (LSD increased). This points to the issue of limited datasets (similar to Section 3.5).

Hence, more investigation with larger amounts of data is required to build meaningful backward models.

5.2 Applications

Some of the applications of the study can be stated as:

1. **Understanding the Auditory Perception** Modelling the relationship between the acoustic stimuli and their responses helps us to understand how brains process the auditory information and what components of the stimuli are observable in the EEG data.

2. **Smart Hearing Aids** A cochlear implant provides acoustic information to the user's brain in the form of electrical signals. It provides electric signals that directly stimulate the auditory nerves near the ears. But once installed, it cannot adapt to the user requirements. Its processing systems are fixed, static and not flexible. It cannot recognize what the user is trying to attend. It cannot adapt to

the changes of the user's brain. The users cannot cognitively control the device. An EEG based hearing device can provide adaptive solutions.

3. **Other Medical Applications** The mere possibility of decoding what a coma patient might be listening to, is a fascinating application in itself. This study can also be extended to animals, especially our pet friends like dogs and cats.

4. **Brain-Computer Interfaces** The methods discussed in the study helps to build better Brain-Computer Interfaces for various applications.

5.3 Summary and Limitations

To summarize our work, we have explored the existing linear hybrid models for audio-EEG data, proposed their deep variants, and have also illustrated that the deep models statistically significant improvements in correlation. In the Chapter 3, we have proposed three deep CCA methods for intra-subject analysis which outperform their linear counterparts for both the speech-listening and music-listening tasks. In the Chapter 4, we have proposed the deep MCCA method, DMCCA, for normalizing the EEG recordings from multiple subjects listening to a common stimulus, which outperform the linear MCCA for both the speech-listening and music-listening tasks. We have discussed the four CCA combinations (LMLC, LMDC, DMLC and DMDC) for the intra-subject and inter-subject analyses and have shown that the deep versions perform better than the linear versions overall. In the last chapter, we have explored a backward model using transformers and an adversarial loss which generates the speech stimuli from the processed EEG response and stimuli's phase information. We have shown that it performs better than the linear TRF counterpart, while still not being able to generate audible output.

5.3.1 Limitations

The limitations of our work are as follows:

1. **Lack of data**: Exploring various architectures of the deep models has shown a common trend of decrease in the performance while increasing the depth of the neural networks. As discussed in the chapter 2, the datasets used in our experiments are very small compared to a typical machine learning setting. As the stimulus is a naturalistic audio signal and the field is still emerging, it is difficult to accumulate large datasets for the naturalistic listening task. The depth of a neural network increases its capacity, and the inadequate audio-EEG data with significant noise results in the neural networks getting overfit to the noise.

2. **EEG Noise**: The EEG readings are recorded from the scalp of the subjects. This results in collection of brain signals not only related to the auditory stimuli. These unrelated signals act as noise while studying the stimulus-response relationship. Though invasive methods suppress the noise, they need extensive surgical procedures to record the data. Hence, more noise suppression methods are needed to work with scalp recordings like EEG.

3. **Analytical Study**: This work have been an analytical study measuring the correlation values of the final representations (performance) for the existing and proposed deep methods. The neural and biological impact of the improved correlations analyzed with various speech representations are of substantial interest to understand what the brain encodes while listening to natural speech.

5.4 Conclusions and Future Directions

To conclude the work, we discussed the motivation for decoding the auditory brain and reviewed the existing methods for single-trial naturalistic audio-EEG data. We described the linear models like TRF and CCA, their limitation from the simplistic linear assumption, and the two datasets (speech and music) for the analysis. Later, we studied the intra-subject analysis and inter-subject analysis, and illustrated that the proposed deep methods offer an advantage over the baseline linear models for both the datasets. As an extension work, we explored the paradigm of speech stimulus reconstruction from the recorded EEG signals, and discussed the advantages and limitations of the deep models in this context.

The future directions for this work are:

1. **Data Collection** As the deep models' performance is limited by the amount of data in the datasets, the collection of more naturalistic audio-EEG data can significantly impact the single-trial audio-EEG analysis.

2. **Exploring robust ML models** As the EEG data contain significant levels of noise, deploying more robust machine learning models can improve the analysis.

3. **Transfer Learning** A common practice for training a complex model on small datasets is to transfer learn the weights from a large similar dataset. Incorporating such techniques can provide better initialization for the analysis to start.

4. **Expanding the study to fMRI and MEG** The study can be expanded to other brain responses collection techniques like fMRI and MEG to get a better understanding of the proposed methods.

5. **Interpretation and Neuroscience** Interpreting the learnt features by incorporating the knowledge from theoretical neuroscience can help us elicit more information about the models and the brain.

References

[1] S. Sanei and J. A. Chambers, *EEG signal processing*. John Wiley & Sons, 2013.

[2] M. G. Coles and M. D. Rugg, *Event-related brain potentials: An introduction.* Oxford University Press, 1995.

[3] T. W. Picton, S. A. Hillyard, H. I. Krausz, and R. Galambos, "Human auditory evoked potentials. i: Evaluation of components," *Electroencephalography and clinical neurophysiology*, vol. 36, pp. 179–190, 1974.

[4] R. F. Burkard, J. J. Eggermont, and M. Don, *Auditory evoked potentials: basic principles and clinical application*. Lippincott Williams & Wilkins, 2007.

[5] V. M. Leavitt, S. Molholm, W. Ritter, M. Shpaner, and J. J. Foxe, "Auditory processing in schizophrenia during the middle latency period (10-50 ms): high-density electrical mapping and source analysis reveal subcortical antecedents to early cortical deficits." *Journal of psychiatry & neuroscience*, 2007.

[6] E. C. Lalor, A. J. Power, R. B. Reilly, and J. J. Foxe, "Resolving precise temporal processing properties of the auditory system using continuous stimuli," *Journal of neurophysiology*, vol. 102, no. 1, pp. 349–359, 2009.

[7] N. Kriegeskorte and P. K. Douglas, "Interpreting encoding and decoding models," *Current opinion in neurobiology*, vol. 55, pp. 167–179, 2019.

[8] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A tutorial on auditory attention identification methods," *Frontiers in neuroscience*, vol. 13, p. 153, 2019.

[9] A. de Cheveigne, M. Slaney, J. Hjortkjaer, and S. Fuglsang, "Auditory stimulus-response modeling with a match-mismatch task," *bioRxiv*, 2020.

[10] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigné, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Neuro-steered hearing devices," *arXiv preprint arXiv:2008.04569*, 2020.

[11] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.

[12] B. Thompson, *Canonical correlation analysis: Uses and interpretation*. Sage, 1984, no. 47.

[13] A. de Cheveigné, D. D. Wong, G. M. Di Liberto, J. Hjortkjaer, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, 2018.

[14] A. de Cheveigné, G. M. Di Liberto, D. Arzounian, D. D. Wong, J. Hjortkjaer, S. Fuglsang, and L. C. Parra, "Multiway canonical correlation analysis of brain data," *NeuroImage*, vol. 186, pp. 728–740, 2019.

[15] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial eeg," *Cerebral cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[16] N. M. Correa, T. Eichele, T. Adalı, Y.-O. Li, and V. D. Calhoun, "Multi-set canonical correlation analysis for the fusion of concurrent single trial erp and functional mri," *Neuroimage*, vol. 50, no. 4, pp. 1438–1445, 2010.

[17] X. Fu, K. Huang, M. Hong, N. D. Sidiropoulos, and A. M.-C. So, "Scalable and flexible multiview max-var canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4150–4165, 2017.

[18] Q. Zhang, J. P. Borst, R. E. Kass, and J. R. Anderson, "Inter-subject alignment of meg datasets in a common representational space," Wiley Online Library, Tech. Rep., 2017.

[19] L. C. Parra, "Multi-set canonical correlation analysis simply explained," *arXiv preprint arXiv:1802.03759*, 2018.

[20] E. C. Lalor and J. J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *European journal of neuroscience*, vol. 31, no. 1, pp. 189–193, 2010.

[21] S. J. Aiken and T. W. Picton, "Envelope and spectral frequency-following responses to vowel sounds," *Hearing research*, vol. 245, no. 1-2, pp. 35–47, 2008.

[22] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[23] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11 854–11 859, 2012.

[24] B. Grothe, M. Pecka, and D. McAlpine, "Mechanisms of sound localization in mammals," *Physiological reviews*, vol. 90, no. 3, pp. 983–1012, 2010.

[25] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, 2015.

[26] T. Lauteslager, J. A. O'Sullivan, R. B. Reilly, and E. C. Lalor, "Decoding of attentional selection in a cocktail party environment from single-trial eeg is robust to task," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014, pp. 1318–1321.

[27] A. Bednar and E. C. Lalor, "Where is the cocktail party? decoding locations of attended and unattended moving sound sources using eeg," *NeuroImage*, vol. 205, p. 116283, 2020.

[28] H. Luo and D. Poeppel, "Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex," *Neuron*, vol. 54, no. 6, pp. 1001–1010, 2007.

[29] G. Pfurtscheller and R. Cooper, "Frequency dependence of the transmission of the eeg from cortex to scalp," *Electroencephalography and clinical neurophysiology*, vol. 38, no. 1, pp. 93–96, 1975.

[30] G. A. Light, L. E. Williams, F. Minow, J. Sprock, A. Rissling, R. Sharp, N. R. Swerdlow, and D. L. Braff, "Electroencephalography (eeg) and event-related potentials (erps) with human participants," *Current protocols in neuroscience*, vol. 52, no. 1, pp. 6–25, 2010.

[31] A. Llorens, A. Trébuchon, C. Liégeois-Chauvel, F. Alario *et al.*, "Intra-cranial

recordings of brain activity during language production," *Frontiers in psychology*, vol. 2, p. 375, 2011.

[32] S. R. Synigal, E. S. Teoh, and E. C. Lalor, "Including measures of high gamma power can improve the decoding of natural speech from eeg," *Frontiers in human neuroscience*, vol. 14, 2020.

[33] A. E. O'Sullivan, C. Y. Lim, and E. C. Lalor, "Look at me when i'm talking to you: Selective attention at a multisensory cocktail party can be decoded using stimulus reconstruction and alpha power modulations," *European Journal of Neuroscience*, vol. 50, no. 8, pp. 3282–3295, 2019.

[34] M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor, "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech," *Current Biology*, vol. 28, no. 5, pp. 803–809, 2018.

[35] N. J. Zuk, E. S. Teoh, and E. C. Lalor, "Eeg-based classification of natural sounds reveals specialized responses to speech and music," *NeuroImage*, vol. 210, p. 116558, 2020.

[36] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[37] H. Cecotti and A. Graser, "Convolutional neural networks for p300 detection with application to brain-computer interfaces," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 3, pp. 433–445, 2010.

[38] X. Sun, C. Qian, Z. Chen, Z. Wu, B. Luo, and G. Pan, "Remembered or forgotten?—an EEG-based computational prediction approach," *PloS one*, vol. 11, no. 12, p. e0167497, 2016.

[39] M. Hajinoroozi, Z. Mao, T.-P. Jung, C.-T. Lin, and Y. Huang, "EEG-based prediction of driver's cognitive performance by deep convolutional neural network," *Signal Processing: Image Communication*, vol. 47, pp. 549–555, 2016.

[40] Y. Yuan, G. Xun, Q. Suo, K. Jia, and A. Zhang, "Wave2vec: Deep representation learning for clinical temporal data," *Neurocomputing*, vol. 324, pp. 31–42, 2019.

[41] X. Zheng, W. Chen, M. Li, T. Zhang, Y. You, and Y. Jiang, "Decoding human brain activity with deep learning," *Biomedical Signal Processing and Control*, vol. 56, p. 101730, 2020.

[42] S. Stober, D. J. Cameron, and J. A. Grahn, "Using convolutional neural networks to recognize rhythm stimuli from electroencephalography recordings," in *Advances in neural information processing systems*, 2014, pp. 1449–1457.

[43] N. Das, J. Zegers, T. Francart, A. Bertrand *et al.*, "Linear versus deep learning methods for noisy speech separation for EEG-informed attention decoding," *Journal of Neural Engineering*, vol. 17, no. 4, p. 046039, 2020.

[44] L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the attended speaker and the locus of auditory attention with convolutional neural networks," *bioRxiv*, p. 475673, 2018.

[45] T. de Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *European Journal of Neuroscience*, vol. 51, no. 5, pp. 1234–1241, 2020.

[46] Q. Liu, Y. Jiao, Y. Miao, C. Zuo, X. Wang, A. Cichocki, and J. Jin, "Efficient representations of eeg signals for ssvep frequency recognition based on deep multiset cca," *Neurocomputing*, vol. 378, pp. 36–44, 2020.

[47] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora, "Deep generalized canonical correlation analysis," *arXiv preprint arXiv:1702.02519*, 2017.

[48] T. Francart, L. Deckers, N. Das, A. H. Ansari, and A. Bertrand, "Eeg-based detection of the locus of auditory attention with convolutional neural networks," in *Attention to sound (The Royal Society), Date: 2018/11/14-2018/11/15, Location: Buckinghamshire, England*, 2018.

[49] G. Ciccarelli, M. Nolan, J. Perricone, P. T. Calamia, S. Haro, J. O'Sullivan, N. Mesgarani, T. F. Quatieri, and C. J. Smalt, "Comparison of two-talker attention decoding from eeg with nonlinear neural networks and linear methods," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[50] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals," *Computers in biology and medicine*, vol. 100, pp. 270–278, 2018.

[51] A. H. Ansari, P. J. Cherian, A. Caicedo, G. Naulaers, M. De Vos, and S. Van Huffel, "Neonatal seizure detection using deep convolutional neural networks," *International journal of neural systems*, vol. 29, no. 04, p. 1850011, 2019.

[52] N. Liu, Z. Lu, B. Xu, and Q. Liao, "Learning a convolutional neural network for sleep stage classification," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2017, pp. 1–6.

[53] A. H. Ansari, O. De Wel, M. Lavanga, A. Caicedo, A. Dereymaeker, K. Jansen, J. Vervisch, M. De Vos, G. Naulaers, and S. Van Huffel, "Quiet sleep detection

in preterm infants using deep convolutional neural networks," *Journal of neural engineering*, vol. 15, no. 6, p. 066006, 2018.

[54] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, and D. P. Subha, "Automated eeg-based screening of depression using deep convolutional neural network," *Computer methods and programs in biomedicine*, vol. 161, pp. 103–113, 2018.

[55] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*, 2013, pp. 1247–1255.

[56] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[57] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli," *Frontiers in human neuroscience*, vol. 10, p. 604, 2016.

[58] A. de Cheveigné and D. Arzounian, "Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data," *NeuroImage*, vol. 172, pp. 903–912, 2018.

[59] A. de Cheveigné, "Sparse time artifact removal," *Journal of neuroscience methods*, vol. 262, pp. 14–20, 2016.

[60] S. Losorelli, D. T. Nguyen, J. P. Dmochowski, and B. Kaneshiro, "Nmed-t: A tempo-focused dataset of cortical and behavioral responses to naturalistic music." in *ISMIR*, 2017, pp. 339–346.

[61] B. B. Kaneshiro, "Toward an objective neurophysiological measure of musical engagement," Ph.D. dissertation, Stanford University, 2016.

[62] N. Gang, B. Kaneshiro, J. Berger, and J. P. Dmochowski, "Decoding neurally relevant musical features using canonical correlation analysis." in *ISMIR*, 2017, pp. 131–138.

[63] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *International conference on digital audio effects*. Bordeaux, 2007, pp. 237–244.

[64] V. Alluri, P. Toiviainen, I. P. Jääskeläinen, E. Glerean, M. Sams, and E. Brattico, "Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm," *Neuroimage*, vol. 59, no. 4, pp. 3677–3689, 2012.

[65] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[66] N. G. B. K. J. Berger and J. P. Dmochowski, "Decoding neurally relevant musical features using canonical correlation analysis," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, Souzhou, China*, 2017.

[67] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010.

[68] I. StatSoft, "Statistica (data analysis software system), version 6," *Tulsa, USA*, vol. 150, pp. 91–94, 2001.

[69] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[70] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[71] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex," *PLoS Biol*, vol. 10, no. 1, p. e1001251, 2012.

[72] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," 2019.

[73] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/pdf/1706.03762.pdf

[74] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[75] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.