Decoding Neural Patterns for Naturalistic Speech Perception

A Project Report Submitted

in Partial Fulfilment of the Requirements

for the Degree of

Bachelor of Technology

in

Computer Science and Engineering

by

Ashutosh Rai (2020BCS0020)

Roshin Nishad (2020BCS0019)

Pratik Raj (2020BCS0112)

Sai Teja (2020BCS0145)

Indian Institute of
Information Technology
Kottayam

to

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

KOTTAYAM-686635, INDIA

November 2023

DECLARATION

We, Ashutosh Rai (2020BCS0020), Roshin Nishad (2020BCS0019), Pratik Raj (2020BCS0112) and Sai Teja (2020BCS0145), hereby declare that, this report entitled **"Decoding Neural Patterns for Naturalistic Speech Perception"** submitted to Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor of Technology** in **Computer Science and Engineering** is an original work carried out by us under the supervision of **Dr. Suchithra M S** and has not formed the basis for the award of any degree or diploma, in this or any other institution or university. I have sincerely tried to uphold the academic ethics and honesty. Whenever an external information or statement or result is used then, that have been duly acknowledged and cited.

Ashutosh Rai (2020BCS0020)
Roshin Nishad (2020BCS0019)
Pratik Raj (2020BCS0112)
Sai Teja (2020BCS0145)

Kottayam-686635
November 2023

CERTIFICATE

This is to certify that the work contained in this project report entitled **"Decoding Neural Patterns for Naturalistic Speech Perception"** submitted by **Ashutosh Rai (2020BCS0020), Roshin Nishad (2020BCS0019), Pratik Raj (2020BCS0112) and Sai Teja (2020BCS0145)** to the Indian Institute of Information Technology Kottayam towards partial requirement of **Bachelor of Technology** in **Computer Science and Engineering** has been carried out by them under my supervision and that it has not been submitted elsewhere for the award of any degree.

Kottayam-686635
November 2023

Dr. Suchithra M S
Project Supervisor

ABSTRACT

Traditional Brain-Computer Interface (BCI) research predominantly concentrates on deciphering neural signals during active speech or writing to generate text or speech outputs. In these established methodologies, individuals actively engage by speaking or typing, with BCI technology translating their neural signals into text or speech. This project takes a distinctive departure from this conventional approach, directing its attention towards speech perception rather than production, signifying a substantial paradigm shift in BCI studies. The central challenge addressed by this project is the accurate decoding of neural activity related to the passive perception of speech. While considerable progress has been made in decoding neural signals associated with speech production, there exists a substantial gap in comprehending how the brain processes speech during listening, in contrast to active speaking or typing. The primary objective of this project is to bridge this knowledge gap by concentrating on the neural patterns and representations intertwined with perceived speech. This will, in turn, enrich our comprehension of speech perception and offers the potential to revolutionize the field of BCIs and augment the quality of life for individuals who rely on assistive technologies.

Contents

List of Figures

Chapter 1

Introduction

1.1 What is BCI Technology

Brain-Computer Interface (BCI) technology is a system that enables communication between the human brain and various machines. It works by collecting brain signals, interpreting those signals, and then outputting commands to a connected machine based on the interpreted brain signals. BCIs can be applied to a variety of tasks such as restoring motor function to paralyzed patients, improving sensory processing, and even allowing communication with locked-in patients. BCI technology can be categorized into three types based on the method used to collect brain signals: Non-Invasive, Semi-invasive, and Invasive.

Non-Invasive BCIs: These systems collect the EEG (Electroencephalography) signal by placing electrodes on the scalp.

Semi-Invasive BCIs: These systems collect the ECoG (Electrocorticography) signal from electrodes placed on the dura or arachnoid, which are layers of the brain.

Invasive BCIs: These systems collect the Intraparenchymal signal by implanting electrodes directly into the cortex of the brain.

1.2 Features of BCI technology

Brain-Computer Interface (BCI) technology has several remarkable features that enable it to facilitate communication between the human brain and machines. Here are some of the major features of BCI:

- **Signal Acquisition:** BCI systems capture brain signals using different methods such as Electroencephalography (EEG), Electrocorticography (ECoG), and Intraparenchymal signals.
- **Signal Processing:** After acquiring the signals, the BCI system processes these signals to interpret the user's intent. This involves noise filtering, feature extraction, and feature translation.
- **Control Signals:** The control signals generated by the BCI can be categorized into evoked signals, spontaneous signals, and hybrid signals.
- **BCI Classifiers:** The BCI system uses various classification algorithms to convert brain activity patterns into commands.

1.3 Background and Motivation

1.3.1 Challenges in BCI

Brain-Computer Interface (BCI) technology poses several challenges that need to be addressed for its successful application. These challenges include:

- **User Training:** BCI systems require significant user training to achieve accurate class discrimination, which can be a barrier to widespread adoption.

- **Signal Processing:** Despite advancements, there is a need for more resilient, accurate, and speedy algorithms to control BCI.

- **Performance Evaluation Metrics:** Lack of uniformity in performance evaluation metrics makes it challenging to compare different systems and establish benchmarks.

- **Privacy and Autonomy:** Ethical considerations arise due to the reading and interpreting of brain signals, raising concerns about potential misuse and the need for regulations to protect user rights.

1.3.2 Motivation

While much research has delved into the neural aspects of speech production, the passive perception of speech remains less explored. Our project seeks to understand these neural patterns, utilizing iEEG data and a FC DNN (Fully-Connected Deep Neural Network). By venturing into this understudied area of speech processing, we hope to contribute valuable insights to the scientific community. Our broader aspiration is to enhance Brain-Computer Interfaces (BCIs) for speech synthesis. By aiming to translate neural signals from passive speech perception into clear speech output, we envision potential applications that could assist those facing communication challenges, making communication more accessible and effective.

Chapter 2

Releated and Proposed Work

2.1 Literature Survey

Luo S. et al. [6] delve into the exploration of Brain-Computer Interfaces (BCIs) with a specific focus on speech decoding and synthesis. Their research is primarily aimed at enhancing communication capabilities, particularly for individuals suffering from locked-in syndrome (LIS). The authors underscore the crucial role of machine learning and neural recording technologies in this field. They also discuss the recent advancements in neural decoding strategies, which include the use of deep learning models and the direct concatenation of speech units. The significance of state-of-the-art vocoders for achieving natural-sounding speech synthesis is also highlighted in their study. However, they acknowledge the challenges that come with direct speech synthesis for LIS patients. These challenges encompass the need for a safe and effective chronically implanted ECoG array that provides sufficient cortical coverage. They also point out the real-time system requirements for decoding covert or attempted speech in the absence of acoustic output.

Brumberg et al. [4] present a robust framework for decoding neural signals with the aim of facilitating artificial speech production. They demonstrate the potential of brain-computer interfaces (BCIs) in controlling speech synthesizers, thereby enabling communication for individuals with severe speech impairments. Although their study is limited to a single subject, the findings provide a significant reference for understanding and decoding neural patterns associated with speech perception. The authors suggest that their methodology could serve as a foundational reference for future studies and development in this field.

Herff et al. [5] delve into the investigation of the possibility of communication between humans and machines based on natural speech-related cortical activity. They present the development and implementation of the "Brain-to-Text" system, which decodes continuously spoken speech into text from brain activity. This system models individual phones, which are the shortest contrastive units in the phonology of a language, and employs techniques from automatic speech recognition (ASR) to transform brain activity while speaking into corresponding textual representation. The researchers use intracranial electrocorticographic (ECoG) recordings to capture brain activity, a process that involves placing electrodes directly on the exposed surface of the brain to record electrical activity from the cerebral cortex. The results of the study demonstrate that the Brain-to-Text system can achieve word error rates as low as 25% and phone error rates below 50%. Additionally, their approach contributes to the understanding of the neural basis of continuous speech production by identifying cortical regions that contain substantial information about individual phones.

Pei X. et al. [7] explore the possibility of inferring spoken or even thought words from brain signals. They specifically focus on decoding vowels and consonants from spoken or imagined monosyllabic words. The authors utilize electrocorticographic (ECoG) signals, which are recorded from the surface of the brain, to discriminate between different vowels and consonants in spoken and imagined words. Techniques such as cortical discriminative mapping are used to identify which cortical locations contain the most information about the discrimination of vowels or consonants. However, the research is primarily focused on monosyllabic words, which may not fully represent the complexity of natural language. The study participants were patients with specific medical conditions (intractable epilepsy), which may limit the generalizability of the findings to the broader population. Despite these limitations, the paper offers a comprehensive approach to decoding neural signals related to specific speech elements, namely vowels and consonants. The results of the study shed light on the distinct mechanisms associated with the production of vowels and consonants, and could potentially provide the basis for brain-based communication using imagined speech. The average classification accuracies for decoding vowels were 40.7% for overt speech and 37.5% for covert speech. For consonants, the average classification accuracies were 40.6% for overt speech and 36.3% for covert speech. These classification accuracies were significantly better than those expected by chance.

Akbari H. et al. [2] conducted research on the reconstruction of intelligible speech from the human auditory cortex. The authors propose a deep neural network architecture consisting of two stages: feature extraction and feature summation. This architecture is utilized to calculate a high-dimensional representation of the input, which is then used to regress the output of the model. The framework incorporates technologies such as deep neural networks, auditory cortex analysis, and vocoder representation. The study acknowledges the limited diversity of the neural responses in their recordings, which restricts the additional information that can be obtained from additional electrodes. However, the deep neural network architecture proposed in this paper can serve as a blueprint for the fully connected deep neural network used in similar research. The authors have also made the codes for performing phoneme analysis, calculating high-gamma envelope, and reconstructing the auditory spectrogram available for further research.

Schirrmeister et al. [8] present a study that introduces a deep learning model capable of effectively decoding and visualizing EEG data. The authors utilize convolutional neural networks (CNNs) to understand and represent brain signals in the EEG data. They highlight the utility of deep learning methodologies, particularly CNNs, for EEG decoding. However, they also acknowledge that the flexibility of CNNs may be a limitation in certain brain-signal decoding scenarios. The research results demonstrate that their deep learning model achieves a mean decoding accuracy of 84.0%, which is at least as good as the widely used filter bank common spatial patterns (FBCSP) algorithm with a mean decoding accuracy of 82.1%. Additionally, the study introduces novel methods for visualizing the learned features, showing that CNNs have indeed learned to utilize spectral power modulations in the alpha, beta, and high gamma frequencies. The paper emphasizes the potential of deep ConvNets combined with advanced visualization techniques for EEG-based brain mapping.

2.2 Problem Statement

The core challenge addressed by this project is the precise decoding of neural activity associated with the passive perception of speech. Significant advancements have been made in the field of decoding neural signals related to speech production. However, there is still a considerable gap in understanding how the brain processes speech during listening, as opposed to active speaking or typing. This project aims to address this gap by focusing on the neural patterns and representations associated with perceived speech, thereby enhancing our understanding of speech perception and potentially offering new avenues for assistive technologies.

The specific problem this project seeks to solve is the development of a robust and accurate method for converting neural signals related to passive speech perception into coherent and natural speech output.

2.3 Dataset

2.3.1 Dataset Overview

Our dataset [1] stands as a valuable and extensive repository of intracranial electroencephalography (iEEG) brain activity data, meticulously gathered from a diverse cohort of 51 human participants. This dataset presents an unparalleled amalgamation of iEEG and functional magnetic resonance imaging (fMRI) recordings, all captured during a naturalistic task. It serves as an abundant source of information, facilitating in-depth exploration into the neural mechanisms underlying multimodal perception and language comprehension.

2.3.2 Data Collection

The dataset [1] was acquired during a movie-watching experiment tailored for presurgical functional language mapping [3]. Participants were immersed in a 6.5-minute short audiovisual film comprising segments from the timeless ”Pippi on the Run” (Pårymmen med Pippi Långstrump, 1970) production. Significantly, the movie incorporated 13 interleaved blocks of speech and music, each spanning 30 seconds. Participants were explicitly instructed to view the movie in an unaltered, naturalistic environment, devoid of any fixation cross or other contrived distractions.

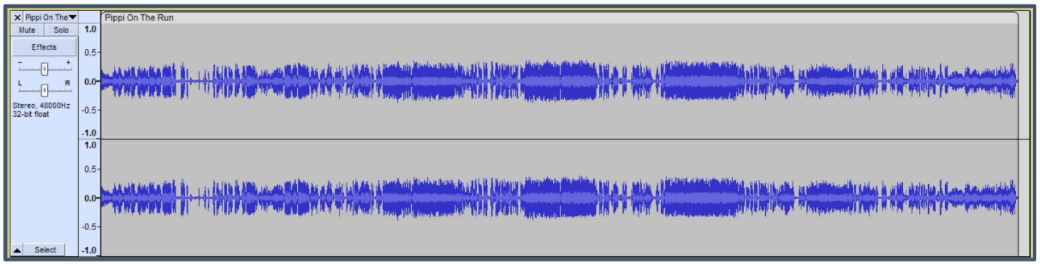


Figure 2.1: Data Collection

2.4 System Architecture

The aim of the project is to decode neural patterns associated with speech perception through intracranial electroencephalography (iEEG) data, leveraging a rich and unique dataset.

The key steps of the implementation:

1. **Data Collection:** We make use of a dataset containing iEEG recordings from 51 participants. The dataset is originally enriched with simultaneous fMRI (functional magnetic resonance imaging) data as well, extending the analysis to a multimodal level. The data was collected during participants' engagement with a short audiovisual film, providing a diverse and naturalistic stimulus.

2. **Data Preprocessing:** Intracranial EEG data is processed to remove noise and artifacts. Data is detrended, and a bandpass filter is applied to focus on frequencies between 70 Hz and 170 Hz.

3. **Feature Extraction:** High gamma (HG) features are extracted from the iEEG data to capture neural patterns associated with speech per-

ception. The HG extraction involves breaking down the data into time windows and applying the Hilbert transform, resulting in a time series of HG features.

4. **Audio Data Alignment:** The project also involves preprocessing of the audio stimulus. Audio data is time-shifted and resampled to match the iEEG sampling rate. Mel spectrogram features are extracted from the audio, allowing for further alignment with iEEG data.

5. **Model Training:** The project employs a Fully Connected Deep Neural Network (FC-DNN) architecture to establish a connection between the iEEG-derived HG features and mel spectrogram features. The model architecture consists of an input layer, multiple hidden layers, and an output layer. The network is trained to minimize mean squared error loss, optimizing the mapping between neural responses and speech perception.

6. **Model Validation and Selection:** Model performance is monitored using a validation dataset to ensure it generalizes well. Early stopping criteria are in place to prevent overfitting. The best-performing model is saved for further use.

7. **Predictions and Audio Synthesis:** The trained model is used to predict mel spectrogram features from iEEG data, providing a direct link between neural activity and speech perception. Predicted features are then transformed into audio, offering synthesized audio representations of perceived speech.

8. **Results Evaluation:** The quality and accuracy of the synthesized audio are assessed by comparing it to the original audio stimulus.

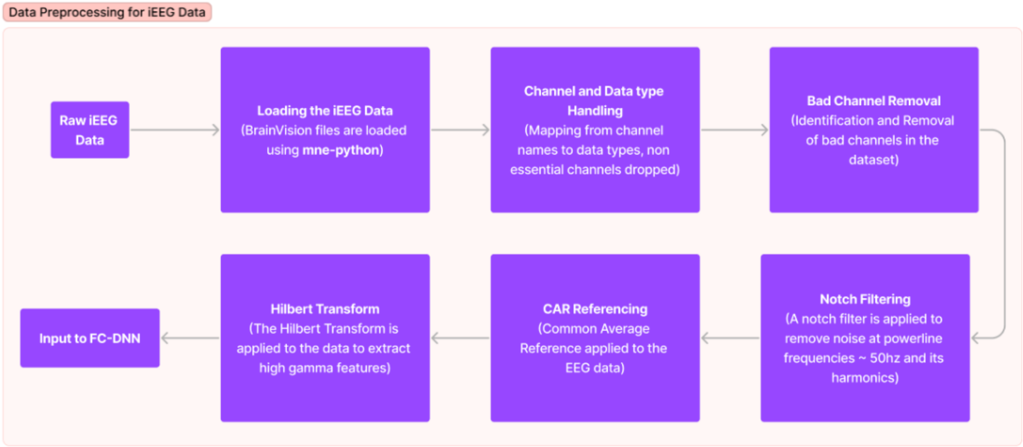

Data Preprocessing for iEEG Data
Raw iEEG Data
Loading the iEEG Data (BrainVision files are loaded using mne-python)
Channel and Data type Handling (Mapping from channel names to data types, non essential channels dropped)
Bad Channel Removal (Identification and Removal of bad channels in the dataset)
Notch Filtering (A notch filter is applied to remove noise at powerline frequencies ~ 50hz and its harmonics)
CAR Referencing (Common Average Reference applied to the EEG data)
Hilbert Transform (The Hilbert Transform is applied to the data to extract high gamma features)
Input to FC-DNN

Figure 2.2: Data Prepossessing for iEEG Data.

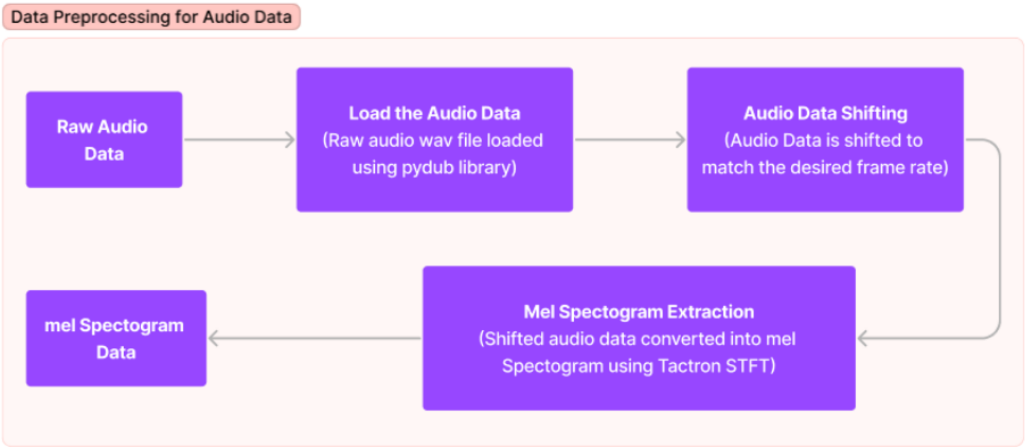

Data Preprocessing for Audio Data
Raw Audio Data
Load the Audio Data (Raw audio wav file loaded using pydub library)
Audio Data Shifting (Audio Data is shifted to match the desired frame rate)
Mel Spectogram Extraction (Shifted audio data converted into mel Spectogram using Tactron STFT)
mel Spectogram Data

Figure 2.3: Data Prepossessing for Audio Data.

Chapter 3

Experimental Results and Conclusions

3.1 Visual Results

The three subplots in the figure display the following information:

- **EEG Test Scaled:** TThe top subplot shows a transposed representation of a random sequence from the scaled test EEG data. Each row represents a channel, and the data is scaled for model input.

- **Test Scaled Mel-Spectogram:** The middle subplot illustrates the corresponding ground truth mel-spectogram associated with the EEG data. The mel-spectogram is computed from the original audio stimuli.

- **Predicted Mel-Spectogram:** The bottom subplot displays the mel-spectogram predicted by the trained FCDNN on the EEG sequence.

The prediction is a result of the learned patterns and relationships captured during training.

The **EEG Test Scaled** provides a visual representation of the input EEG data as it is fed into the model. Each row corresponds to a different electrode or channel, and the scaling ensures consistency in model input. The **Mel-Spectrogram Test Scaled** represents the actual mel-spectrogram derived from the EEG sequence. This spectrogram serves as the ground truth for comparison. The **Mel-Spectrogram Predicted** showcases the model's output, demonstrating its ability to predict the associated Mel-Spectrogram for the given EEG input.

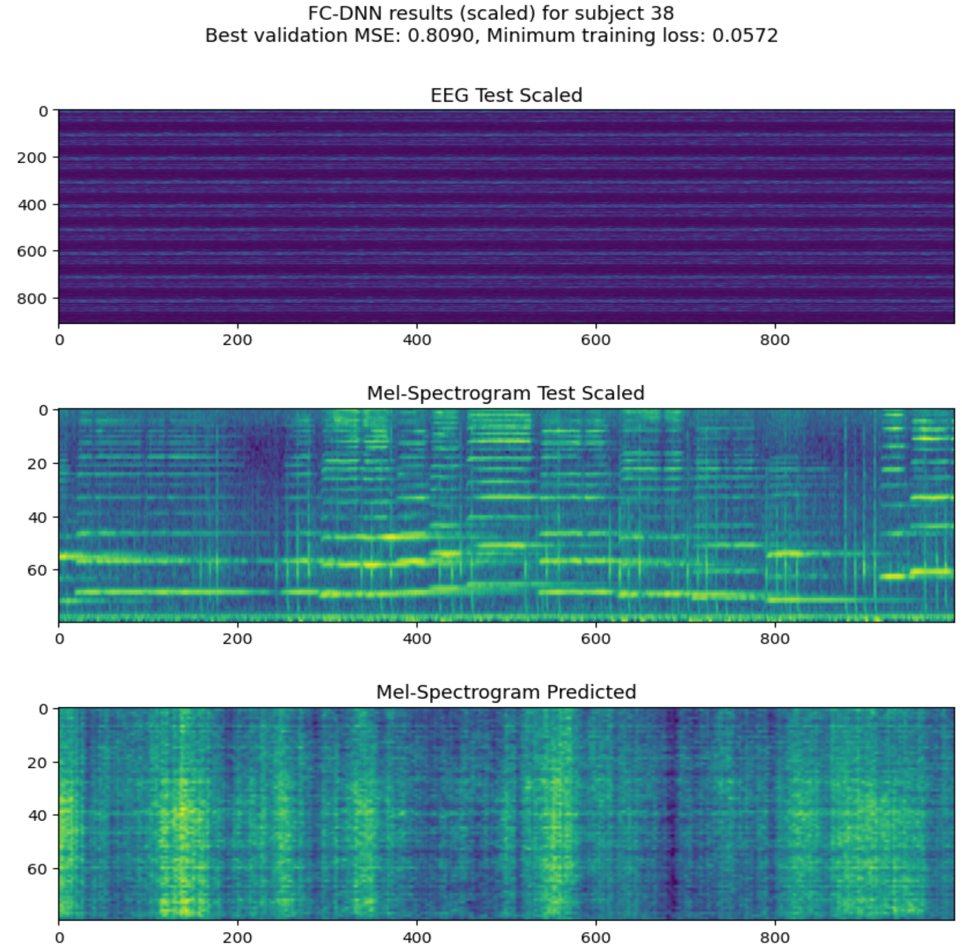


Figure 3.1: Results of the Fully Connected Deep Neural Network (FC-DNN) model on a random sequence of the test EEG data

3.2 Performance Measures

The following Metrics are being used for the performance analysis of the network:

1. **SSIM (Structural Similarity Index):** SSIM measures the similarity between two images, considering luminance, contrast, and structure. It evaluates how well the predicted mel-spectrogram preserves the structural information of the ground truth.

2. **MSE (Mean Squared Error):** MSE calculates the average squared difference between corresponding elements of the ground truth and predicted mel-spectrograms. It quantifies the overall magnitude of errors in the prediction, emphasizing larger errors.

3. **PSNR (Peak Signal-to-Noise Ratio):** PSNR represents the ratio between the maximum possible power of a signal and the power of the error signal. It measures the quality of the predicted mel-spectrogram, emphasizing higher values for better predictions.

4. **KL - Divergence (Kullback-Leibler):** KL Divergence measures the information lost when the predicted distribution is used to approximate the ground truth distribution. It quantifies the difference in information content between the predicted and actual mel-spectrograms.

5. **Pearson Correlation Coefficient:** Pearson Correlation calculates the linear correlation between two variables. Here, it assesses the linear relationship between the ground truth and predicted mel-spectrograms.

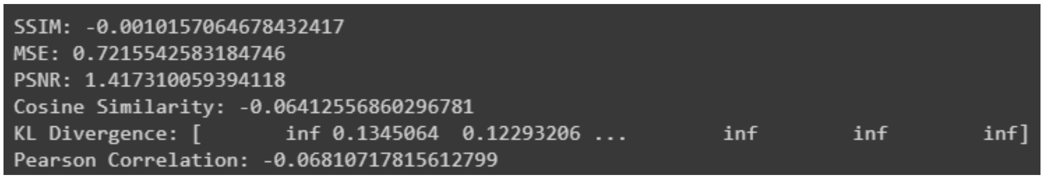
6. **Cosine Similarity:** Cosine Similarity computes the cosine of the angle between two non-zero vectors. It assesses the directional similarity between the ground truth and predicted mel-spectrograms.

The selected metrics provide a comprehensive evaluation of the performance of the FC-DNN on the task of predicting mel-spectrograms from EEG data. These metrics cover various aspects such as structural similarity, error magnitude, signal quality, directional similarity, information content, and linear correlation. The obtained metric values indicate a suboptimal performance of the FC-DNN model:

- **SSIM:** -0.0010 (Poor)
- **MSE:** 0.7216 (High error magnitude)
- **PSNR:** 1.4173 (Low signal quality)
- **Cosine Similarity:** -0.0641 (Poor directional similarity)
- **KL Divergence:** [Inf, 0.1345, 0.1229, ..., Inf] (High information difference)
- **Pearson Correlation:** -0.0681 (Weak linear correlation)

Despite the observed poor performance, it's essential to acknowledge that the primary objective was to explore the application of deep learning methods for the given task. The obtained results serve as a baseline for future work, especially in the context of applying 2D Convolutional Neural Networks (2DCNN), which is expected to enhance the model's ability to capture spatial features in the data. In future iterations, further optimizations and architectural adjustments will be pursued to improve predictive accuracy and overall model performance.

Additional Training Metrics:
Best Validation MSE: 0.8090
Minimum Training Loss: 0.0572
These values provide an insight into the model's performance during training, guiding future refinements.



```
SSIM: -0.0010157064678432417
MSE: 0.7215542583184746
PSNR: 1.417310059394118
Cosine Similarity: -0.06412556860296781
KL Divergence: [       inf 0.1345064  0.12293206 ...        inf        inf        inf]
Pearson Correlation: -0.06810717815612799
```

Figure 3.2: Performance metric values obtained

3.3 Conclusion and Future Works

In summary, our project is a significant endeavor involving the implementation of a Fully Connected Deep Neural Network (FC-DNN) for the precise decoding of neural activity linked to passive speech perception. This project has yielded promising results, with the FC-DNN trained successfully using iEEG data and audio-derived mel spectrograms. We have compared the predicted mel spectrogram with the scaled mel spectrogram of the testing dataset, revealing striking similarities. Finally, we've applied an audio synthesizer to convert the predicted mel spectrogram into audio. As we move forward, our future work will involve the implementation of similar techniques using 2D Convolutional Neural Networks (CNNs). This shift towards the integration of 2D Convolutional Neural Networks (CNNs) heralds a significant advancement in the domain of Brain-Computer Interfaces (BCIs) dedicated to speech perception.

Bibliography

[1] Open multimodal ieeg-fmri dataset from naturalistic stimulation with a short audiovisual film - openneuro.org. https://openneuro.org/datasets/ds003688/versions/1.0.7. [Accessed 3-Nov-2023].

[2] H Akbari, B Khalighinejad, JL Herrero, et al. Towards reconstructing intelligible speech from the human auditory cortex. *Sci Rep*, 9:874, 2019.

[3] J. Berezutskaya, M.J. Vansteensel, E.J. Aarnoutse, et al. Open multimodal ieeg-fmri dataset from naturalistic stimulation with a short audiovisual film. *Sci Data*, 9:91, 2022.

[4] J. S. Brumberg, P. R. Kennedy, and F. H. Guenther. Artificial speech synthesizer control by brain–computer interface. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[5] C. Herff, D. Heger, A. de Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Front. Neurosci.*, 9:217, 2015.

[6] S. Luo, Q. Rabbani, and N.E. Crone. Brain-computer interface: Applications to speech decoding and synthesis to augment communication. *Neurotherapeutics*, 19(2):263–273, 2022.

[7] X Pei, DL Barbour, EC Leuthardt, and G Schalk. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *J Neural Eng*, 8(4):046028, 2011.

[8] RT Schirrmeister, JT Springenberg, LDJ Fiederer, et al. Deep learning with convolutional neural networks for eeg decoding and visualization. *Hum Brain Mapp*, 38(11):5391–5420, 2017.