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
April 2024

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. We 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
April 2024

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	**Dr. Suchithra M S**
April 2024	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 (Electrocorticogra-

phy) 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. [10] 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. [8] 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. [11] 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. [12] 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.

Cheah, K. H. et al. [5] explore the utilization of Convolutional Neural Networks (CNNs) in the classification of EEG data during music listening, a study with implications for Brain-Computer Interface (BCI) technologies. Their approach, comparing the efficacy of 1D and 2D convolutional kernels, highlights the cerebral laterality of musical influence on brain activity. The research showcases a notable accuracy in EEG signal classification, achieving 98.94% in binary and 97.68% in three-class validations through tenfold cross-validation, with test accuracies maintaining high levels of 97.46% and 95.71%, respectively. Despite these promising results, the paper notes a gap in its methodology, particularly the absence of a benchmark comparison with leading-edge methods, which limits a comprehensive evaluation of its performance. This study underscores the potential of CNNs in deciphering EEG signals related to music, presenting a significant stride toward advanced BCI applications.

Eger, S. et al. [7] engage in a detailed examination of various activation functions within the realm of deep learning, focusing particularly on their applicability to natural language processing (NLP) tasks. Their research, which juxtaposes 21 distinct activation functions across a suite of eight NLP challenges, earmarks the Swish function, characterized by the formula f(x) $= x \cdot \sigma(x)$, for its exceptional performance attributed to smoother gradient transitions. This attribute has positioned Swish as a frontrunner in numerous tasks, especially within technologies like sequence tagging and recurrent neural networks (RNNs). While the study offers a broad analysis of activation functions, it stops short of exploring the theoretical foundations that could explain the varying efficacies of these functions in different scenarios. Nevertheless, the results from Eger et al.'s investigation underscore the efficacy of Swish in potentially enhancing generalization capabilities in neural network models for NLP, thereby supporting its integration into future neural network architectures aimed at optimizing NLP tasks.

Choi, D. et al. [6] delve into the empirical evaluation of optimizers for deep learning, with a specific lens on the efficacy of Adam in comparison to traditional methods such as SGD and Momentum. Their findings underscore the pivotal role of hyperparameter tuning protocols in optimizing the performance of Adam, suggesting that, under ideal conditions, Adam may equal or surpass the efficiency of its counterparts across a diverse array of tasks. The research leverages first-order optimization techniques to highlight Adam's proficiency in managing gradients and enhancing model generalization, positioning it as a favorable choice for neural network optimization. However, the study also acknowledges the constraints imposed by the requisite for extensive hyperparameter tuning and substantial computational resources, which may limit Adam's applicability in environments with constrained optimization capabilities.

Zheng, J. et al. [13] explore the potential of a data-driven approach for the time-frequency analysis of scalp EEG signals through their study, utilizing the Hilbert-Huang Transform (HHT) to better account for individual variability in brainwave frequencies that conventional techniques may overlook. The research emphasizes the integration of HHT with deep learning methodologies to enhance the precision of EEG signal classification, thereby offering advancements towards more accurate and reliable diagnostic tools in the field. Despite the promising outcomes, the authors note the computational demands of the HHT, which pose challenges to its feasibility for real-time EEG analysis applications. This acknowledgment underscores the balance between innovative analytical depth and practical implementation constraints within the domain.

Lee, S. H. et al. [9] explore the cognitive representations of imagined and overt speech, alongside speech perception, through the lens of electroencephalogram (EEG) data analysis. Utilizing a deep neural network (DNN) capable of capturing temporal-spectral-spatial features, their research effectively classifies subjects based on EEG activities associated with these speech tasks. The study notably underscores the feasibility of subject identification via single-channel EEG in both imagined and overt speech contexts. A critical aspect of their methodology is the strategic electrode placement within key speech processing areas, particularly the Broca's and Wernicke's regions, and extending to the superior temporal gyrus. This focus, however, raises considerations about the potential underrepresentation of other cerebral areas in contributing to speech cognition. The findings accentuate the crucial role of precise electrode positioning in advancing the decoding of speech signals for innovative neural-based communication systems, despite the acknowledged limitation of overlooking broader neural contributions to speech processes.

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 Introduction

Our study utilizes the 'Open multimodal iEEG-fMRI dataset' [1], which contains intracranial electroencephalography (iEEG) brain activity data from 51 patients with medication-resistant epilepsy admitted to the University Medical Center Utrecht for diagnostic procedures. This dataset includes recordings of both iEEG and functional magnetic resonance imaging (fMRI) during a naturalistic task. The combination of these modalities provides valuable insights into neural mechanisms involved in multimodal perception and language comprehension. This enables a comprehensive investigation of neural correlates related to speech and language processing with high spatial and temporal resolutions.

2.3.2 Experimental Procedures

The patients took part in two types of experiments: movie-watching and resting state. In the movie-watching experiment, patients watched a short film as part of routine clinical tasks for presurgical functional language mapping. The resting state experiment involved patients resting for three minutes. Patients without a separate resting state task had a 3-minute 'natural rest' period selected from their 24/7 clinical iEEG recordings. Electrode types were implanted based on clinical requirements.

2.3.3 Dataset Acquisition

The dataset [1] used in this study was acquired during a movie-watching experiment designed for presurgical functional language mapping [3]. Participants engaged in a 6.5-minute audiovisual film, featuring segments from "Pippi on the Run" (Pårymmen med Pippi Långstrump, 1970), with 13 interleaved blocks of speech and music lasting 30 seconds each. Participants were instructed to watch the movie naturally, without any artificial distractions such as fixation crosses. Intracranial EEG (iEEG) data were recorded using a 128-channel system during the tasks. Data were sampled at either 512 Hz or 2048 Hz and filtered accordingly (0.15–134.4 Hz or 0.3–500 Hz). A reference electrode was placed externally on the mastoid temporal bone. Additionally, HD ECoG data were recorded for six patients, either concurrently with clinical channels or in separate sessions.

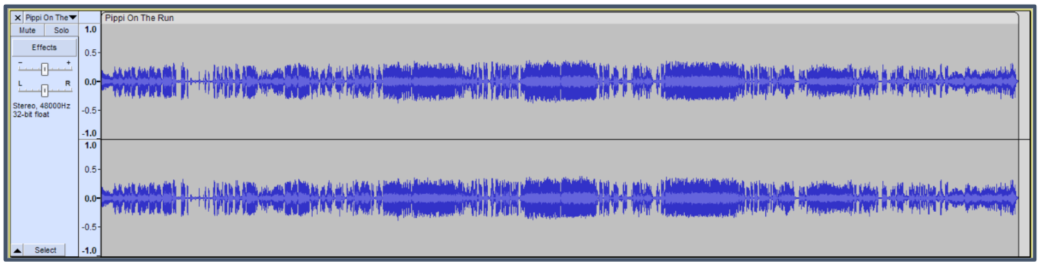

Figure 2.1: Data Collection

2.3.4 Prime Subjects

The methodology for selecting subjects for our project was meticulously crafted, guided by a rigorous process aimed at identifying individuals demonstrating a high correlation with the speech envelope during movie stimuli.

This criterion, established by our dataset compilation team, is crucial as it signifies the intricate relationship between neural activity and speech dynamics, pivotal for the development of effective brain-computer interfaces for speech synthesis. Based on the hypothesis suggesting that subjects exhibiting close alignment with the speech envelope would offer optimal neural responses, the selection process yielded four individuals (Subjects 43, 46, 55, and 60). This methodology underscores our focus on individuals whose neural patterns reflect robust engagement and synchronization with auditory stimuli, which are essential qualities for our research objectives.

The selected participants showed notably high correlation values, indicating a strong neural connection with the speech envelope. Intracranial electrodes were strategically placed over key speech areas such as Broca's area, the motor cortex, and the superior temporal gyrus, ensuring comprehensive data capture [9]. Subject 38 was chosen for exceptional electrode coverage over critical speech areas. This strategic selection enhances the richness of our dataset, providing opportunities for more accurate speech reconstructions. Our approach combines quantitative and qualitative criteria, leveraging statistical metrics and a deep understanding of brain mechanisms. This integrated approach not only improves speech decoding and reconstruction but also advances our comprehension of neural processes. Ultimately, these advancements will lead to improved brain-computer interfaces for speech synthesis, benefiting both neuroscience and technology.

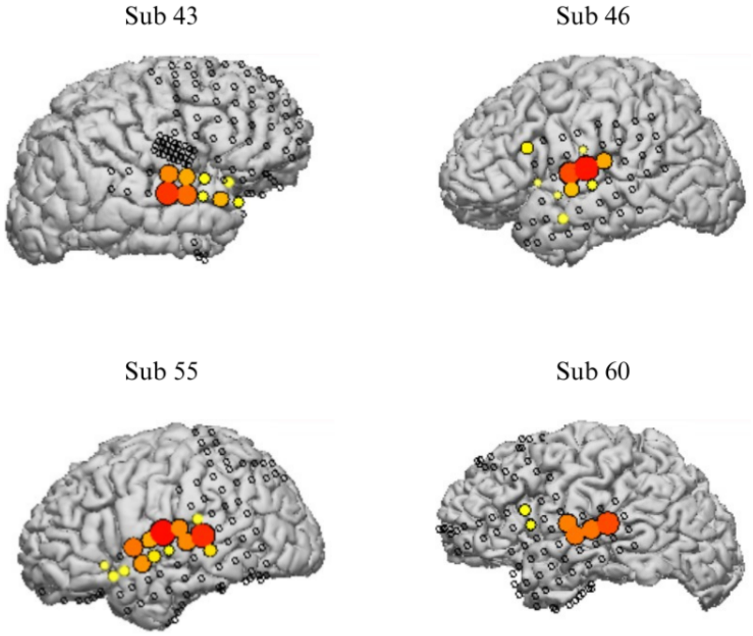

Figure 2.2: Four subjects with highest correlation with speech envelope. [3]

2.3.5 Data Preparation

During the data preparation phase, several crucial steps were taken to align and extract important information from the EEG and mel spectrogram data. Initially, the EEG data was trimmed to match the duration of the stimulus presentation during the experiment, using embedded annotations to mark the start and end points of the stimulus. This alignment, based on the 6.5-minute movie duration, ensures subsequent analysis focuses on task-relevant neural activity, providing a more accurate training basis. Additionally, cross-correlation between the electrode's high-frequency band signal and the sound envelope revealed an average delay of approximately 150 milliseconds between neural response and audio stimulus. To address this, the audio was shifted backward by 150 milliseconds to better align with the original auditory stimulus, thereby enhancing the naturalness and intelligibility of synthesized speech.

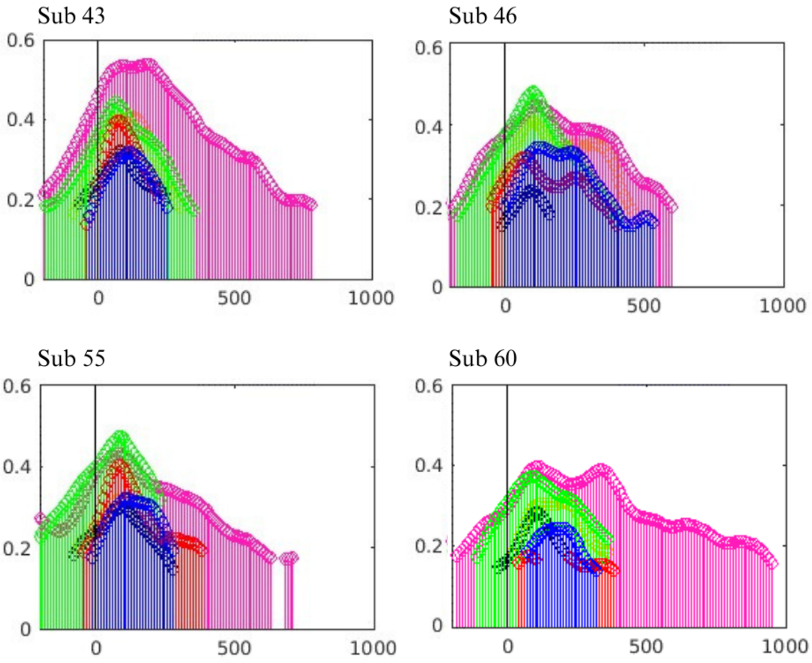

Figure 2.3: Lag plots illustrating the cross-correlation between the high-frequency band signal of electrodes and the sound envelope.

The subsequent data preparation phase included extracting features from the iEEG and mel spectrogram data using the Hilbert transform. This process involved linear detrending, determining window numbers based on window length and frame shift parameters, high-gamma bandpass filtering, and computing absolute values of the Hilbert transform to create the feature space. Feature vectors were generated for each window through window-based feature extraction and stacking of features. Additionally, EEG data corresponding to speech segments in the movie was meticulously extracted to align with speech-specific segments. This involved selectively slicing raw EEG and mel spectrogram data based on provided annotations, converting start and end times into corresponding indices, and refining arrays by cutting and appending data. The resulting refined dataset enhances the relevance and accuracy of subsequent deep learning model training.

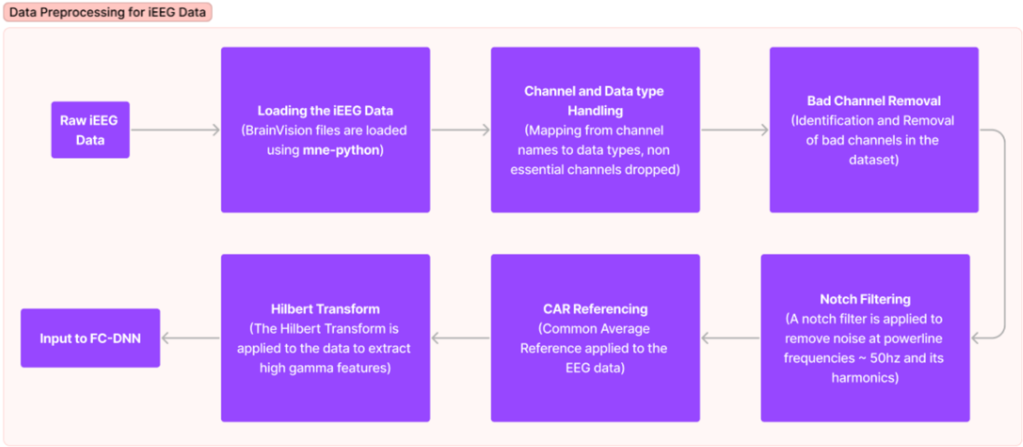
Data Preprocessing for iEEG Data
Raw iEEG Data
Loading the iEEG Data (BrainVision files are loaded using mne-python)
Channel and Data type Handling (Mapping from channel names to data types, non essential channels dropped)
Bad Channel Removal (Identification and Removal of bad channels in the dataset)
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 2.4: Data Preprocessing for iEEG Data.

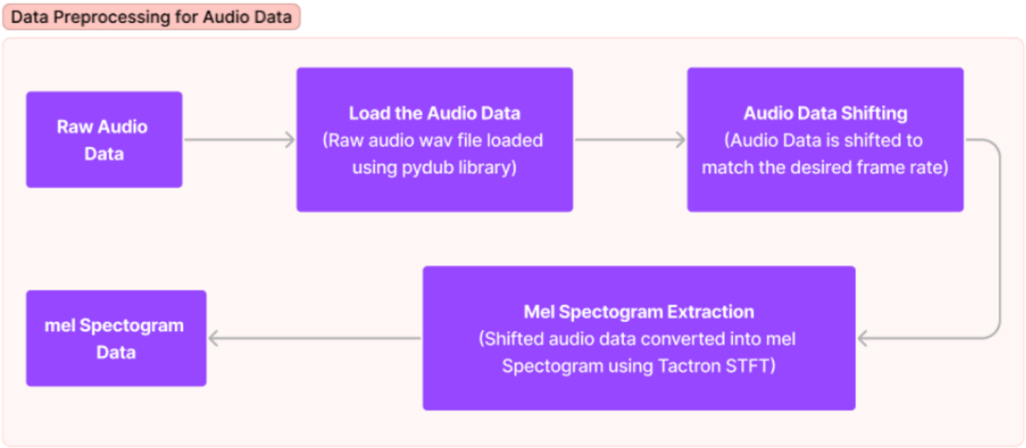
Data Preprocessing for Audio Data
Raw Audio Data
Load the Audio Data (Raw audio wav file loaded using pydub library)
Audio Data Shifting (Audio Data is shifted to match the desired frame rate)
Mel Spectogram Extraction (Shifted audio data converted into mel Spectogram using Tactron STFT)
mel Spectogram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

Figure 2.5: Data Preprocessing for Audio Data.

2.4 System Architecture

2.4.1 FC-DNN

In our study, we developed a fully connected Deep Neural Network (Fc-DNN) model featuring a single hidden layer containing 3000 neurons. This selection stemmed from a thorough examination of various layer configurations and neuron counts. The objective was not only to improve performance metrics like accuracy but also to establish a robust model that demonstrates effective generalization and prevents overfitting.

As we enhanced the model's complexity by adding layers and neurons, we faced challenges with overfitting. Although accuracy stayed mostly consistent, began exhibiting aberrant characteristics in its outputs, particularly in the generated melspectrograms. These irregularities suggested an excessively complex model that may have memorized the training data too closely, including noise and generating unrealistic outputs.

In order to find the right balance between model complexity and performance while mitigating the risk of overfitting, we settled on an optimal Fc-DNN configuration with a single hidden layer containing 3000 neurons. This architecture not only achieved commendable accuracy but also displayed superior generalization capabilities, as evidenced by the mel-spectrograms it produced, which closely aligned with the expectations.

We used Rectified Linear Unit (ReLU) as the activation function for the input layer to create sparse representations and address the vanishing gradient problem. For the output layer, we employed a linear activation function, suitable for regression tasks where the output can vary across real numbers. The Adam optimizer, chosen for its adaptive learning rate features and stable convergence, was utilized to facilitate the model's training process.

To address overfitting, a common issue in deep learning, where models may memorize training data excessively and struggle with unseen data, we applied early stopping and saved the best model weights during training.

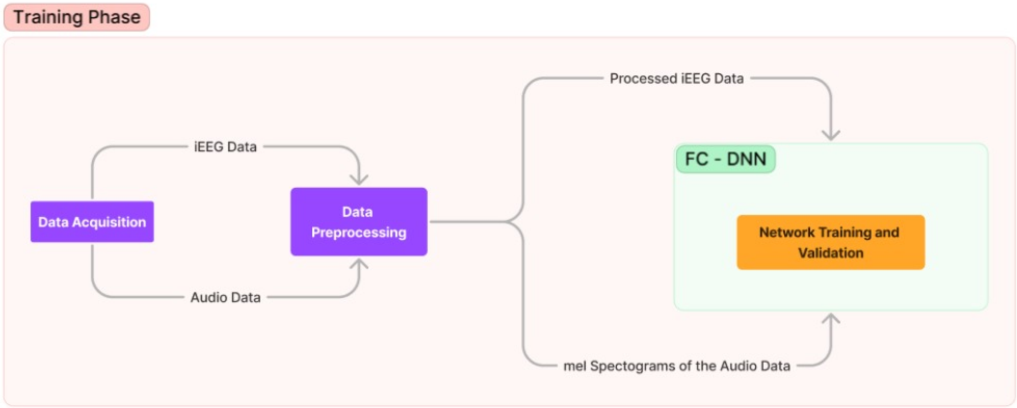

Figure 2.6: Training phase for Fc-DNN.

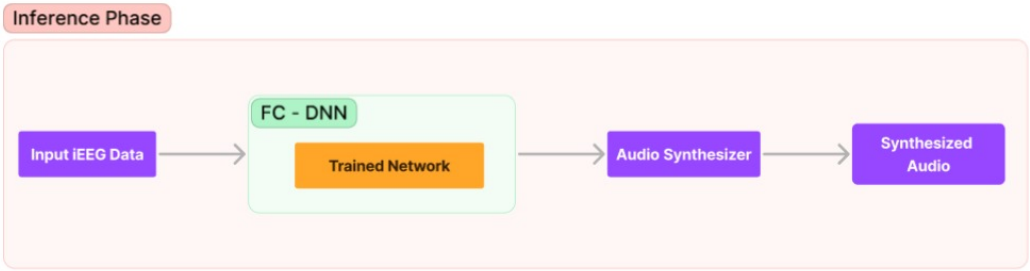

Figure 2.7: Inference phase for Fc-DNN.

2.4.2 2D-CNN

In this study, we employed the 2D-CNN to analyze spectrogram data obtained from EEG recordings. The data was divided into training and test sets, with 80% allocated to training. We normalized both input and output data using the mean and standard deviation from the training set. Reshaping the input data to a 3D structure of (9, 127) was necessary to fit the 2D-CNN's requirements. Moreover, a sliding window method was used to convert the data into 3D blocks.

The 2D-CNN model's architecture included three convolutional layers with kernel sizes of (13, 13) and 'swish' activation function. Dropout layers were added to prevent overfitting, and a max pooling layer was used to reduce input dimensionality and highlight important features.

After the convolutional layers, the model transitioned to a flatten layer, transforming the input into a one-dimensional array. Next, a densely connected layer, activated by 'swish', followed. Before the final output layer, which had a dense layer with a linear activation function, another dropout layer was added. This ensured that the output layer matched the shape of the training spectrogram.

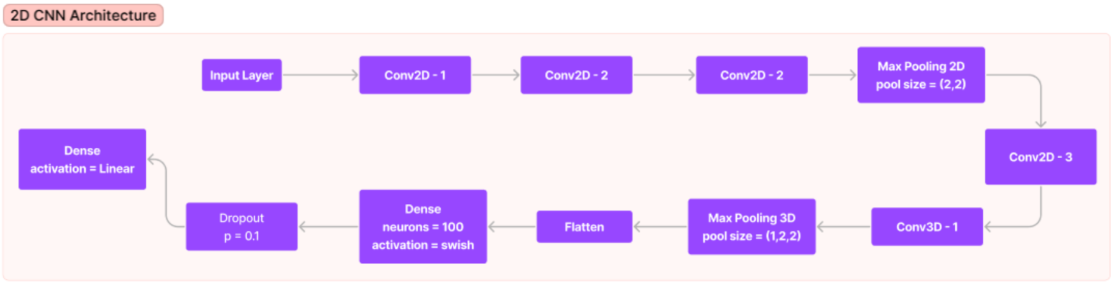

Figure 2.8: Architecture of 2D-CNN.

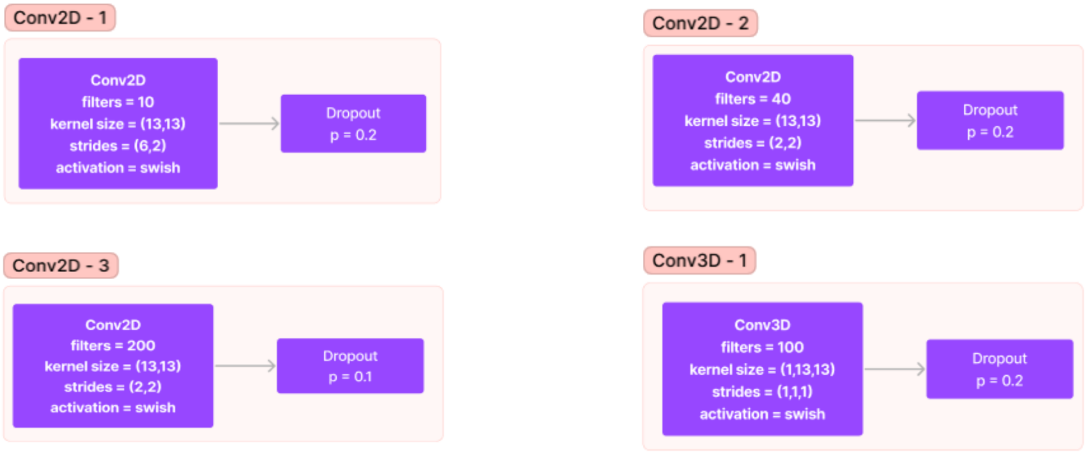

Figure 2.9: Convolution blocks present in 2D-CNN.

2.5 Evaluation Methods

In speech synthesis and enhancement systems, mel spectrograms are crucial for accurately representing acoustic features. However, generated spectrograms may suffer from over-smoothing, compromising the quality of synthesized speech. Additionally, current modeling methods often overlook correlations present in both time and frequency domains, resulting in blurry and over-smoothed spectrograms. Given these limitations, comparing the visual similarity of predicted and ground truth mel spectrograms using metrics like SSIM may not effectively capture the required precision and correlation in acoustic representations.

As a result, we rely on quantitative evaluation using Mean Squared Error (MSE) to assess model performance. Specifically, we compare the lowest training error and best validation MSE for the FC-DNN and 2D-CNN models applied to the selected prime subjects. This approach allows us to objectively evaluate the models' abilities to predict Melspectrogram data from EEG signals, providing insight into their effectiveness and generalization capabilities.

Chapter 3

Experimental Results

In this chapter, we present the results obtained from our deep learning models, namely FC-DNN and 2D-CNN. These models were trained using brain activity data collected during passive listening and spoken speech tasks. We discuss the performance of each model, highlighting their potential applications in brain-computer interfaces for speech synthesis and communication.

3.1 Fully-connected Deep Neural Network (Fc-DNN)

The Fully Connected Deep Neural Network (Fc-DNN) underwent training using data from five distinct subjects. Table 3.1 outlines the model's performance for each subject, showcasing the best training loss and validation mean squared error (MSE) achieved.

Subject	Best Training Loss	Best Validation MSE
38	0.2577	1.8012
43	0.1261	0.6920
46	0.3625	0.8313
55	0.2069	0.9742
60	0.3160	0.8032

Table 3.1: Performance of the Fc-DNN on prime subjects.

The training loss values indicate how accurately the model predicts Mel-spectrogram data from EEG signals during training. A lower training loss suggests a better fit of the model to the training data. On the other hand, validation MSE measures the model's performance on unseen data, with lower MSE values indicating better generalization performance.

The results presented in Table 3.1 show that the model attained the lowest training loss with subject 43, suggesting effective fit to the training data. Additionally, subject 43 also exhibited the best generalization performance on unseen data, with the lowest validation MSE.

3.2 Two-Dimensional Convolutional Neural Network (2D-CNN)

Similar to the FC-DNN, the 2D-CNN model underwent training with data from five distinct subjects. Table 3.2 presents the performance metrics measures for the 2D-CNN, including the best training loss and validation mean squared error (MSE) for each subject.

Subject	**Best Training Loss**	**Best Validation MSE**
38	1.0119	0.7872
43	1.0015	0.8685
46	1.0258	0.7779
55	0.9713	0.7759
60	0.9114	0.9576

Table 3.2: Performance of the 2D-CNN on prime subjects.

The performance of the 2D-CNN model is assessed using the same metrics as the Fc-DNN model: training loss and validation MSE. The training loss reflects how accurately the model predicts Melspectrogram data from EEG signals during training. A lower training loss suggests a better fit to the training data, with subject 60 achieving the lowest training loss. Validation MSE measures the model's performance on new data, with smaller values indicating better generalization. Subject 55 had the lowest validation MSE (0.7759), while subject 46 had a marginally higher MSE at 0.7779.

3.3 Mel-Spectrogram samples

Figure 3.1,3.2,3.3 display an original speech stimuli sample at the top, followed by mel-spectrograms generated from iEEG input by the 2D-CNN in the middle and the Fc-DNN at the bottom. Upon visual examination, it's evident that the 2D-CNN result appears oversmoothed, whereas the FC-DNN produces more ”realistic” patterns. However, the resemblance between the original audio stimuli and the predicted spectrogram remains unsatisfactory.

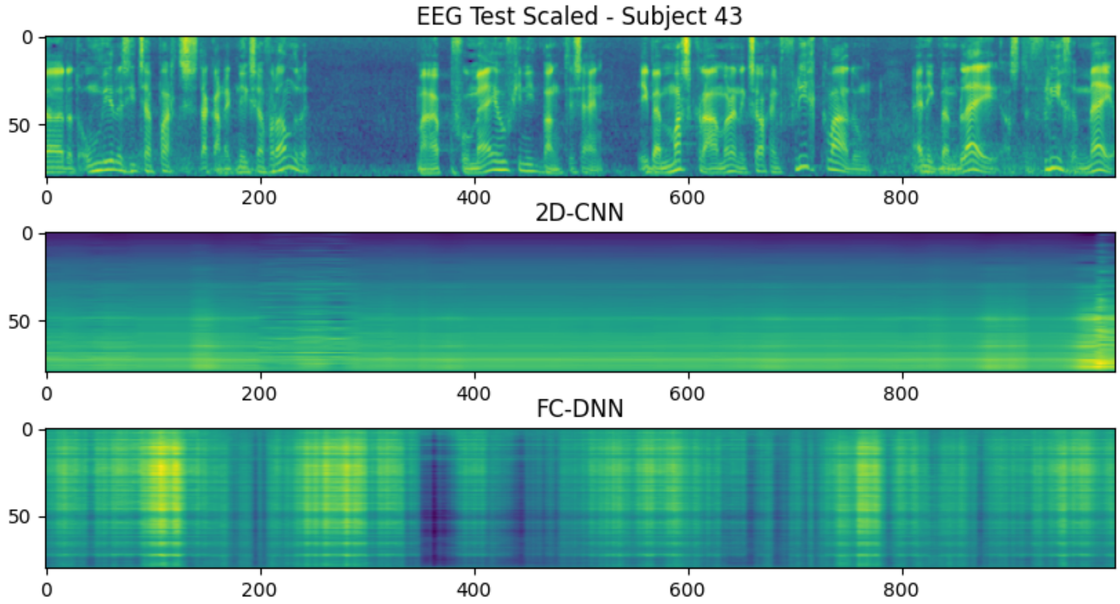

Figure 3.1: Mel-spectograms for sub 43

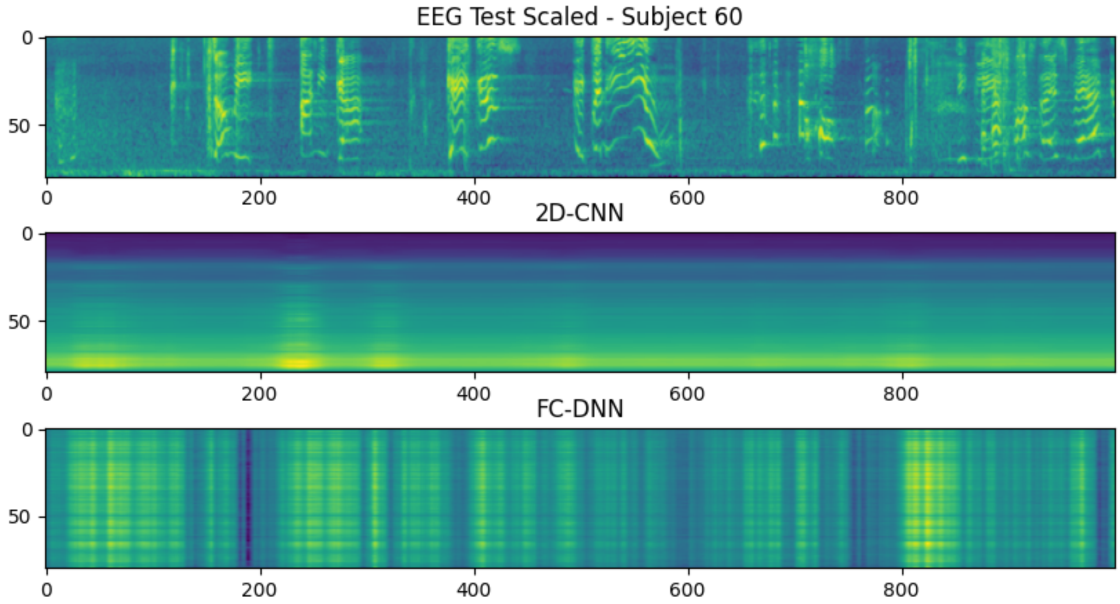

Figure 3.2: Mel-spectograms for sub 60

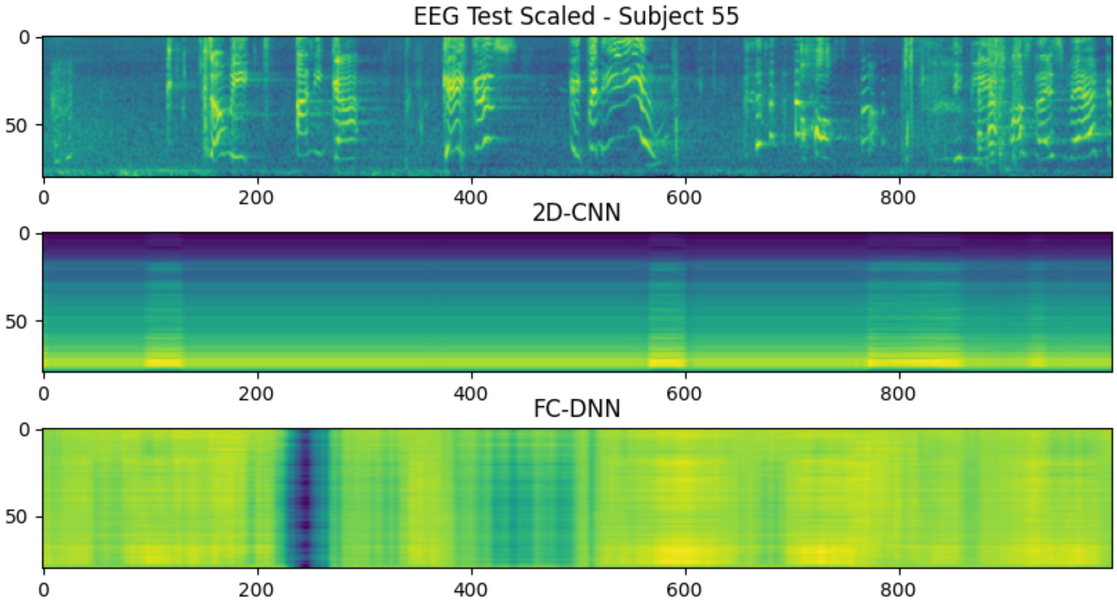

Figure 3.3: Mel-spectograms for sub 55

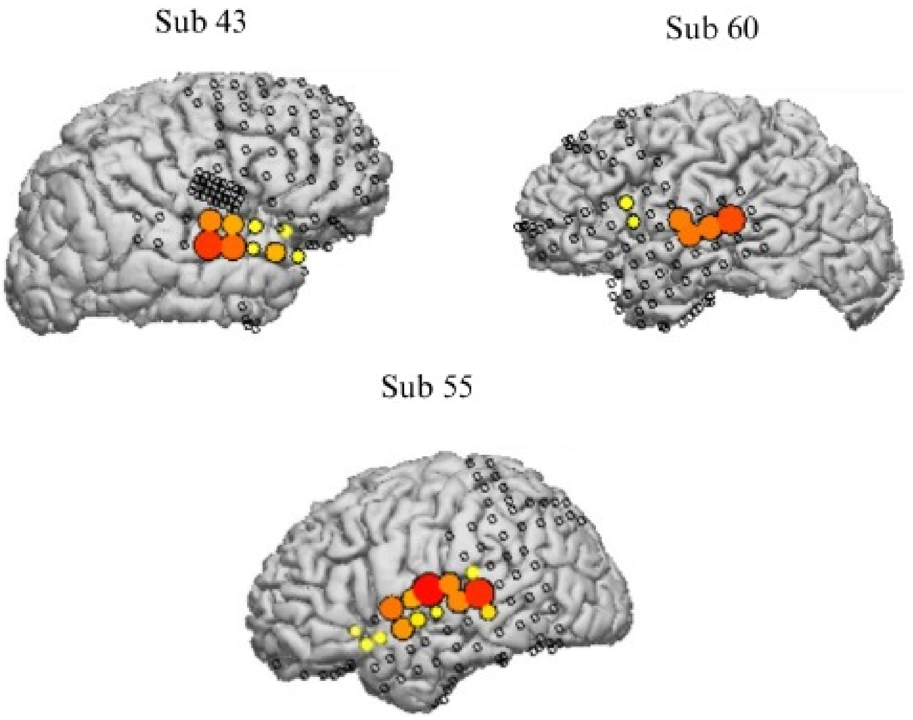

Figure 3.4: Electrode placement comparison for subjects 43, 60, 55. [3]

Drawing from Figure 3.4, depicting a comparison of electrode placements among subjects with top metric scores, subject 55's results emerge as the most realistic. This outcome likely stems from the electrodes positioning in areas of greater relevance within the subject's brain.

3.4 Audio Synthesis

The authors conducted informal human evaluation on the synthesized audio from both models to gauge quality and clarity. Although not fully intelligible, the audio wasn't entirely unclear either. The models effectively captured important auditory features, including pauses between speech segments, essential for determining rhythm and pace. However, accurately reconstructing speech content remains a significant challenge, indicating that the models have yet to decode the complex relationship between brain activity and speech production effectively.

Chapter 4

Conclusion

The study explored the utilization of deep learning models for decoding brain activity observed during passive listening, employing the FC-DNN and 2D-CNN architectures. Despite limited training data, both models consistently reduced test loss and, in certain instances, decreased validation loss as well, with the early stopping mechanism preventing overfitting.

While the models demonstrated relative success, significant challenges persisted. Notably, achieving satisfactory accuracy levels for both validation and test sets simultaneously proved difficult, as reflected in the predicted Mel-spectrograms. Despite capturing intensity areas and finer details, the models failed to generate realistic spectrograms, resulting in synthesized speech lacking expected audibility and clarity.

The study distinguishes itself by focusing on decoding during passive listening scenarios, a less-explored area presenting unique challenges. This complexity complicates the decoding process, posing challenges in the development of effective algorithms.

Several challenges highlighted critical areas for improvement in deep learning models for speech decoding. These include precise synchronization of EEG and audio data, dataset size and diversity limitations, and the need for more targeted data collection.

Moving forward, extending stimulus duration, exploring advanced neural network architectures, and incorporating scenarios where subjects audibly reproduce heard speech could enhance model performance and deepen understanding. Interdisciplinary research, as showcased here, bridges gaps between machine learning, cognitive science, and neuroscience. This demonstrates the effectiveness of integrating perspectives from diverse fields to improve our understanding of complex phenomena like speech.

In summary, this study underscores the efficiency of deep learning models in interpreting brain activity during speech-related tasks, offering valuable insights into cognitive and neural processes. These findings pave the way for future research and the exploration of human cognition.

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] Kit Hwa Cheah, Humaira Nisar, Vooi Voon Yap, and Chen-Yi Lee. Convolutional neural networks for classification of music-listening eeg: comparing 1d convolutional kernels with 2d kernels and cerebral laterality of musical influence. *Neural Computing and Applications*, 32(13):8867–8891, 2020.

[6] Dami Choi, Christopher J Shallue, Zachary Nado, Jaehoon Lee, Chris J Maddison, and George E Dahl. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446*, 2019.

[7] Steffen Eger, Paul Youssef, and Iryna Gurevych. Is it time to swish? comparing deep learning activation functions across nlp tasks. *arXiv preprint arXiv:1901.02671*, 2019.

[8] 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.

[9] Seo-Hyun Lee, Young-Eun Lee, and Seong-Whan Lee. Voice of your brain: Cognitive representations of imagined speech, overt speech, and speech perception based on eeg. *arXiv preprint arXiv:2105.14787*, 2021.

[10] 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.

[11] 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.

[12] 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.

[13] Jingyi Zheng, Mingli Liang, Sujata Sinha, Linqiang Ge, Wei Yu, Arne Ekstrom, and Fushing Hsieh. Time-frequency analysis of scalp eeg with

hilbert-huang transform and deep learning. *IEEE Journal of biomedical and health informatics*, 26(4):1549–1559, 2021.