Decoding Neural Patterns for Naturalistic Speech Perception

Ashutosh Rai (2020BCS0020)

Roshin Nishad (2020BCS0019)

Pratik Raj (2020BCS0112)

Sai Teja (2020BCS0145)

Guided By,

Dr. Suchithra M S

Introduction

- This research presents a novel approach to study the cognitive processes of speech perception, with the aim to establish connections between brain's electrical activity and speech sounds.

- Investigates speech perception mechanisms and decodes the brain's neural activity for speech synthesis in the context of Brain computer interface (BCI) technology.

- Employs deep learning techniques on iEEG and mel spectrogram data corresponding to the brain activity and speech respectively, identifying the shared features.

Problem Statement

The challenge at hand involves decoding speech from brain activity during passive listening, utilizing advanced deep learning techniques. The goal is to address this intricate task and advance the synthesis of speech in Brain-Computer Interfaces (BCI) for improved accessibility.

Motivation

- Diverging from prior research, we concentrate on decoding during passive listening, presenting unique challenges. Unlike active tasks, passive listening lacks explicit signals, complicating the decoding process.

- The absence of overt motor and auditory responses in passive scenarios makes developing effective decoding algorithms a formidable task.

- Contrasting with studies like Anumanchipalli et al. **[1]**, our focus on passive listening differs significantly. Their success in decoding relied on active tasks involving overt speech, emphasizing the impact of task nature on decoding outcomes.

- Through translating neural signals from passive speech perception to clear speech output, our goal is to enhance communication accessibility, especially for those with communication needs.

Review Summary
Phase - I

1. **Data acquisition:** acquired a dataset with brain activity data. The dataset contains a combination of iEEG and fMRI recordings acquired through a audiovisual movie stimulus. The movie featured 13 interleaved blocks of speech and music, each lasting 30 seconds.

2. **Preprocessing data:** iEEG data was cropped based on annotations, aligning with the experiment's stimulus duration. Feature extraction involved linear detrending, high gamma bandpass filtering, attenuation of line noise harmonics, and hilbert transform for creating the feature space. Data was split into train, validation, and testing sets and normalized.

Review Summary
Phase - I

3. **Model building and training:** A Fully connected Deep Neural Network (FC-DNN) was constructed with a single hidden layer and 3000 neurons and ReLU activation, and an output layer with 80 neurons and a linear activation. Model was compiled with the Adam optimizer and MSE loss and fit on the training data.

4. **Model evaluation:** Quantitatively, mean squared error (MSE) served as the primary metric for evaluating the model's ability to predict. Qualitatively, the models outputs were visualized in the form of reconstructed Mel Spectrograms. The model achieved a best validation MSE of 1.1899 and a minimum training loss of 0.0226. Additional metrics were used to compare visual similarity.

Literature Review

Simistira Liwicki, F. et al. (2021) [3] present a pioneering study in the paper "Bimodal pilot study on inner speech decoding reveals the potential of combining EEG and fMRI." This research introduces the first publicly available dataset for bimodal EEG/fMRI inner speech decoding and demonstrates that combining these two modalities significantly enhances decoding accuracy compared to using each modality separately. The study highlights the synergy between EEG's high temporal resolution and fMRI's high spatial resolution, offering a promising direction for future research in inner speech decoding. However, the study's limitations include a small sample size and the need for further validation in larger, more diverse populations. The findings are highly relevant to the project, suggesting that a multimodal approach could substantially improve the decoding of inner speech, potentially benefiting applications in brain-computer interfaces and neurorehabilitation.

Literature Review

Katthi, J.R., and Ganapathy, S. (2021) [4] introduce an innovative approach in "Deep Correlation Analysis for Audio-EEG Decoding," focusing on the integration of deep learning techniques to enhance the correlation analysis between audio stimuli and EEG responses. Their methodology significantly surpasses traditional linear models, particularly in auditory tasks involving speech and music, by directly optimizing a correlation-based loss function within a neural network framework. This advancement promises improvements in auditory EEG decoding, with potential applications in brain-computer interfaces and auditory neuroscience. The study, however, acknowledges the limitations inherent in the noise-prone nature of EEG data and the challenges of generalizing findings across diverse subject samples. The relevance of this research to the project lies in its potential to refine EEG-based auditory processing models, contributing to more accurate interpretations of neural responses to complex stimuli like speech and music.

Literature Review

Sun, S. et al. (2016) [5] explore the impact of depth in deep neural networks (DNNs) from a theoretical standpoint, particularly focusing on margin bounds. They find that while increased depth can enhance a network's representation power, leading to lower empirical margin errors, it also raises the Rademacher Average (RA), potentially worsening test performance. This suggests a trade-off, where increasing depth beyond a certain point may not always be beneficial. Inspired by these insights, they propose large margin DNN algorithms (LMDNNs) that achieve significant performance improvements over standard DNNs by incorporating margin-based penalty terms to reduce empirical margin errors without increasing depth. This work provides a valuable theoretical foundation for understanding and optimizing the depth of DNNs, with practical implications for improving DNN design and training methodologies.

Current Work

1. Based on the lagplots of the cross correlation of the electrodes high frequency band signal and the sound envelope there was an average delay of 150 ms discovered. This represented the delay between the neural response and the audio stimulus which influences the timing of speech synthesis. To account for this observed lag, the audio was shifted backwards by 150 ms, ensuring the adjusted audio aligns more accurately with the original auditory stimulus.

2. We introduced an additional step of data preprocessing, extracting the EEG data corresponding to the segments where speech was present in the movie, with the aim of aligning the EEG data with the speech specific segments of the movie thereby focusing the analysis on the the brain's response to the auditory speech stimuli. To achieve this, the EEG and mel spectrogram data were selectively sliced based on the annotations provided in the events array, indicating the start and end times of each speech segment.

Current Work

3. Increased the complexity of the FCDNN model by increasing the number of layers and neurons. There was no substantial degradation in accuracy, however there was noticeable issue with overfitting. The generated melspectograms exhibited aberrant characteristics significantly deviating from what one would anticipate from a typical melspectrogram. This suggested the model was overly complex learning even the noise. And therefore the previous model with one hidden layer and 3000 neurons was preferred, providing better performance, and capacity for generalization.

Current Work

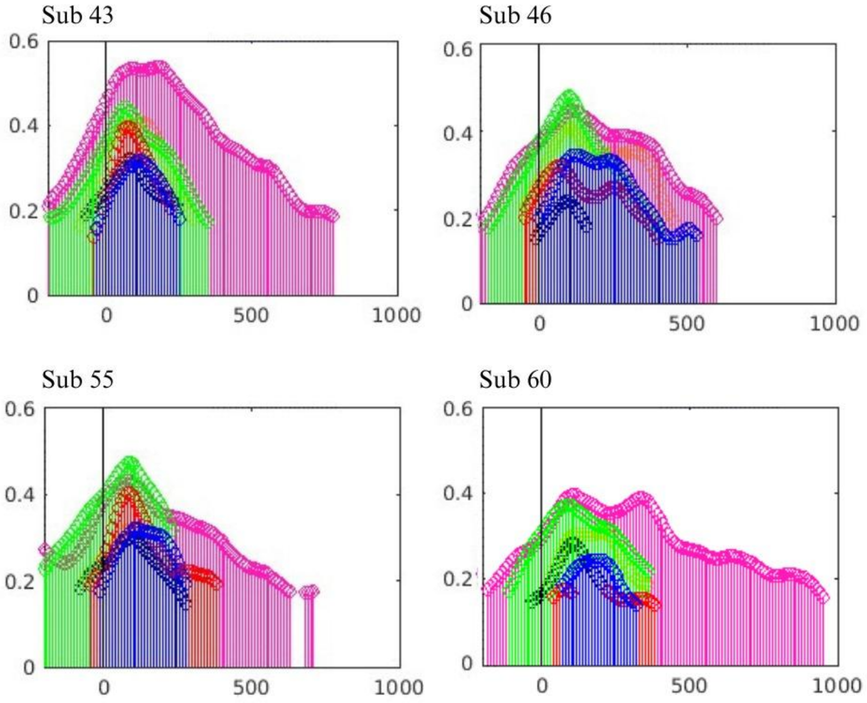


Fig 1.1: Lagplots of the cross-correlation of the electrode's high frequency band signal and the sound envelope **[2]**

Experimental Results

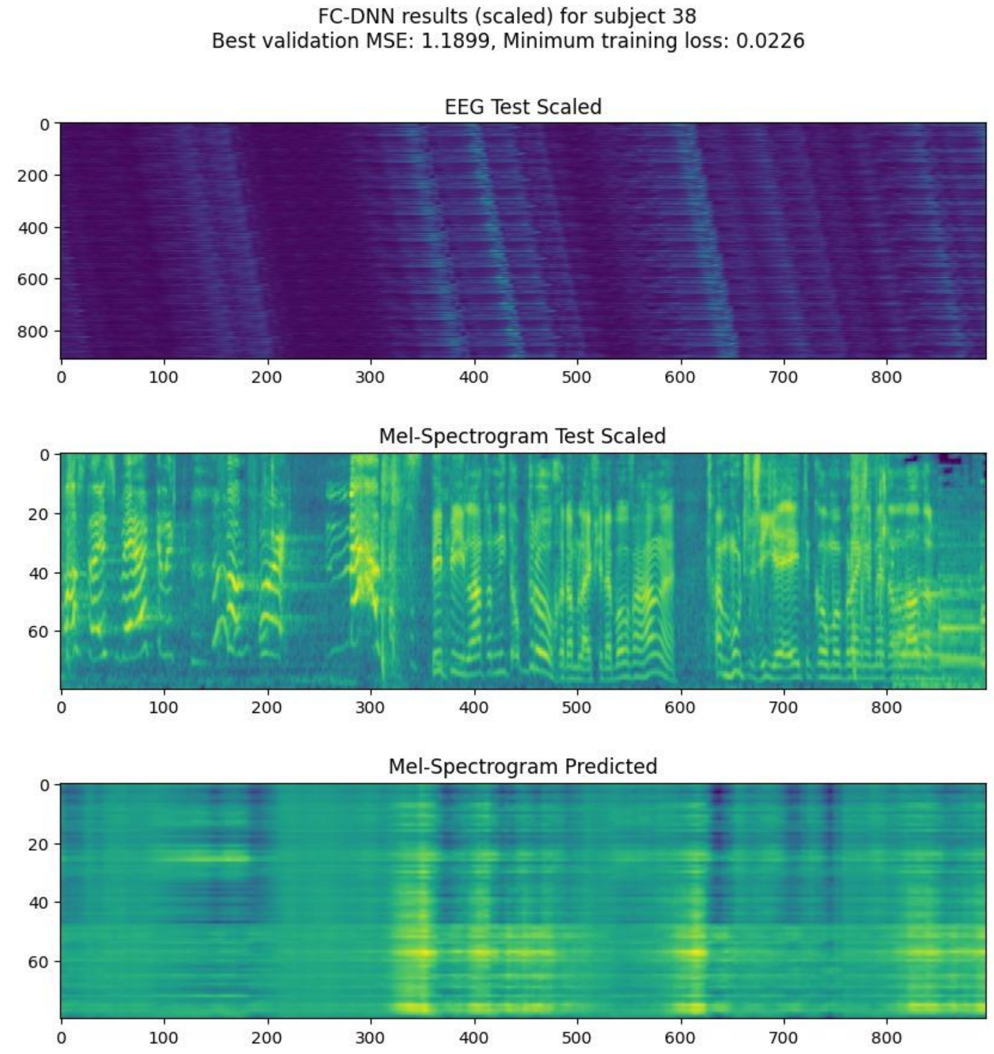


Figure 2.1: Figure illustrating the predicted Mel Spectrogram compared to the ground truth in the test set for FC-DNN with 3 hidden layers

Experimental Results

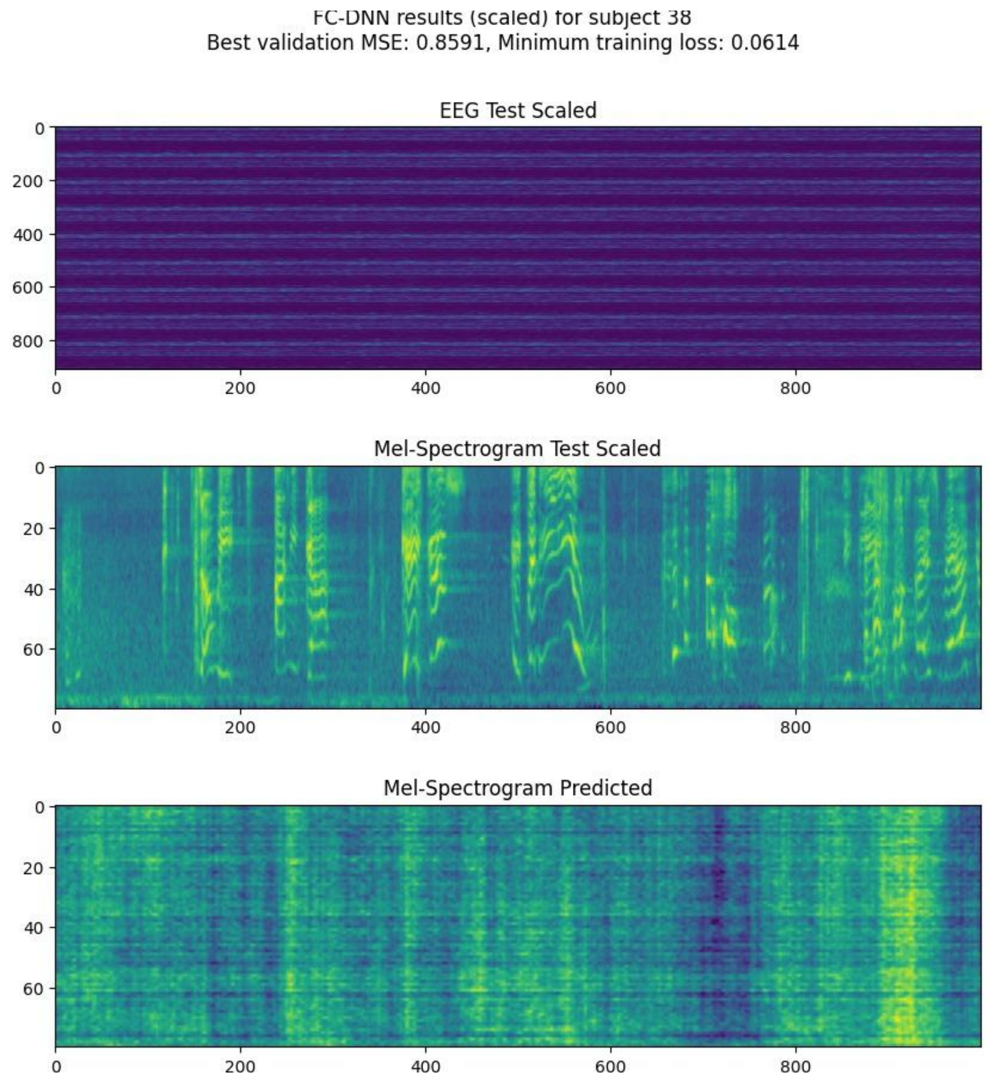
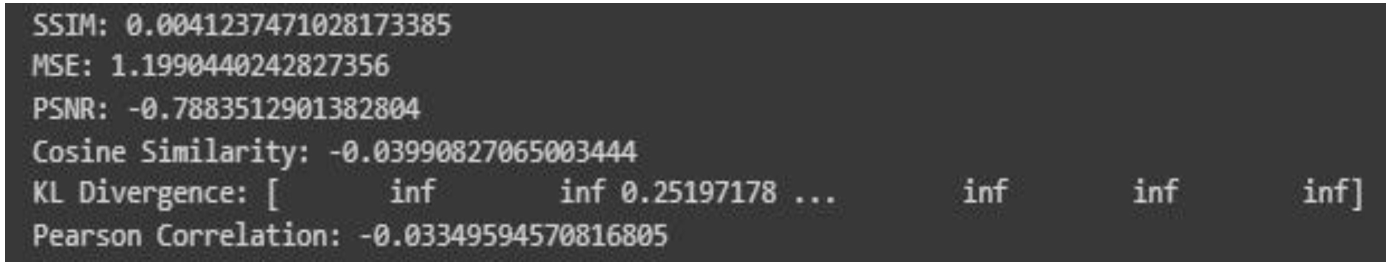


Figure 2.2: Figure illustrating the predicted Mel Spectrogram compared to the ground truth in the test set for FC-DNN with a single hidden layer

Note: both the versions of the model are trained on data after shifting by 150ms and performing the selective slicing to align with the speech stimuli

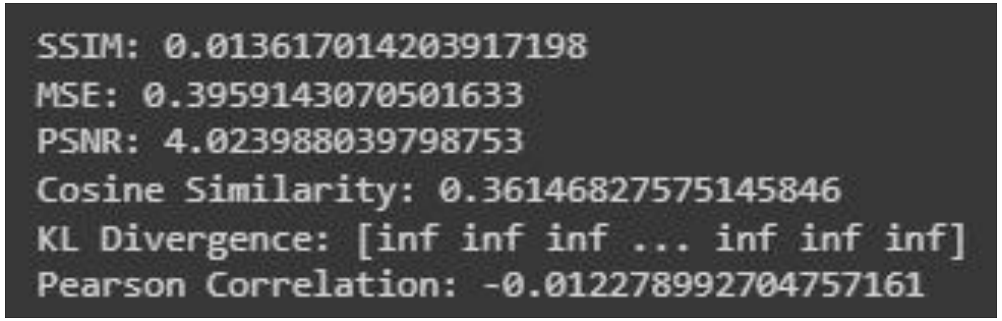
Experimental Results



```
SSIM: 0.0041237471028173385
MSE: 1.1990440242827356
PSNR: -0.7883512901382804
Cosine Similarity: -0.03990827065003444
KL Divergence: [       inf        inf 0.25197178 ...        inf        inf        inf]
Pearson Correlation: -0.03349594570816805
```

Figure 2.3: Metrics assessing the visual similarity between the predicted and ground truth Mel Spectrograms for the model with three hidden layers

SSIM - Structural Similarity Index
PSNR - Peak Signal to Noise Ratio



```
SSIM: 0.013617014203917198
MSE: 0.3959143070501633
PSNR: 4.023988039798753
Cosine Similarity: 0.36146827575145846
KL Divergence: [inf inf inf ... inf inf inf]
Pearson Correlation: -0.012278992704757161
```

Figure 2.4: Metrics assessing the visual similarity between the predicted and ground truth Mel Spectrograms for the model with one hidden layer

Conclusion

- Increasing model complexity by adding additional layers decreased generalizability.

- Shifting audio back by 150 ms and selectively slicing data to include only the speech segments had a positive impact.

- The FCDNN approach shows promising potential, especially considering the limited training data. The identification of shared features in neural activity associated with both passive listening and spoken speech is a compelling revelation. This not only confirms the existence of these shared patterns but also demonstrates their successful identification and decoding by deep learning models.

Conclusion

- Future work involves adapting these methods for use with a 2D-CNN. The goal is to achieve better metric scores and overall performance, further advancing the capabilities of the model.

References

1. G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," in Nature, vol. 568, no. 7753, pp. 493-498, 2019.

2. J. Berezutskaya, M. J. Vansteensel, E. J. Aarnoutse, Z. V. Freudenburg, G. Piantoni, M. P. Branco, and N. F. Ramsey, "Open multimodal iEEG-fMRI dataset from naturalistic stimulation with a short audiovisual film," in Scientific Data, vol. 9, no. 1, 91, 2022.

3. F. Simistira Liwicki et al., 'Bimodal pilot study on inner speech decoding reveals the potential of combining EEG and fMRI'.

4. J. Reddy Katthi and S. Ganapathy, "Deep Correlation Analysis for Audio-EEG Decoding," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 2742-2753, 2021,

5. Sun, S., Chen, W., Wang, L., Liu, X., & Liu, T.-Y. (2016). On the Depth of Deep Neural Networks: A Theoretical View. Proceedings of the AAAI Conference on Artificial Intelligence, 30(1).