Decoding Neural Patterns for Naturalistic Speech Perception

BTP Phase II Review

Ashutosh Rai (2020BCS0020)

Roshin Nishad (2020BCS0019)

Pratik Raj (2020BCS0112)

Sai Teja (2020BCS0145)

Guided By,

Dr. Suchithra M S

Assistant Professor

Computer Science and Engineering

Progress Overview

- Data Acquisition:
  - Acquired a dataset comprising combined iEEG and fMRI recordings during an audiovisual movie stimulus.
  - The movie consisted of 13 interleaved blocks of speech and music, each lasting 30 seconds.

- Preprocessing Data:
  - iEEG data cropped based on annotations to match stimulus duration.
  - Feature extraction involved linear detrending, high gamma bandpass filtering, noise harmonics attenuation, and hilbert transform for feature space creation.
  - Split dataset into train, validation, and test sets, followed by normalization.

Progress Overview

- Model Building and Training:
  - Constructed a Fully Connected Deep Neural Network (FC-DNN) with a single hidden layer (3000 neurons, ReLU activation) and an output layer (80 neurons, linear activation).
  - Compiled model with Adam optimizer and Mean Squared Error (MSE) loss function, trained on the training set.

- Results & Insights:
  - Identified subtle patterns between neural activity and speech perception.
  - 2D CNN excelled in capturing intensity variations, while FC DNN captured finer details.

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.

Prime Subjects

- The selection of subjects was meticulously guided by a rigorous methodology aimed at identifying individuals demonstrating a high correlation with the speech envelope during the movie stimuli. This criterion, established by the dataset compilation team, is pivotal as it signifies the intricate relationship between neural activity and speech dynamics, crucial for the development of effective brain-computer interfaces for speech synthesis.

- Grounded in a 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 underlines our focus on individuals whose neural patterns reflect robust engagement and synchronization with auditory stimuli, essential qualities for our research objectives.

Prime Subjects

- Selected participants exhibited notably high correlation values, indicating a strong neural connection with the speech envelope. Intracranial electrodes were strategically placed over key speech areas such as the Broca's area, motor cortex, and superior temporal gyrus, ensuring comprehensive data capture [6]. Subject 38 was chosen for exceptional electrode coverage over critical speech areas. This strategic selection enhances the richness of our dataset, offering opportunities for more accurate speech reconstructions.

- Our approach combines quantitative and qualitative criteria, leveraging statistical metrics and a deep understanding of brain mechanisms. This not only enhances speech decoding and reconstruction but also furthers our understanding of neural processes. Ultimately, these advancements will lead to improved brain-computer interfaces for speech synthesis, benefiting both neuroscience and technology.

Prime Subjects

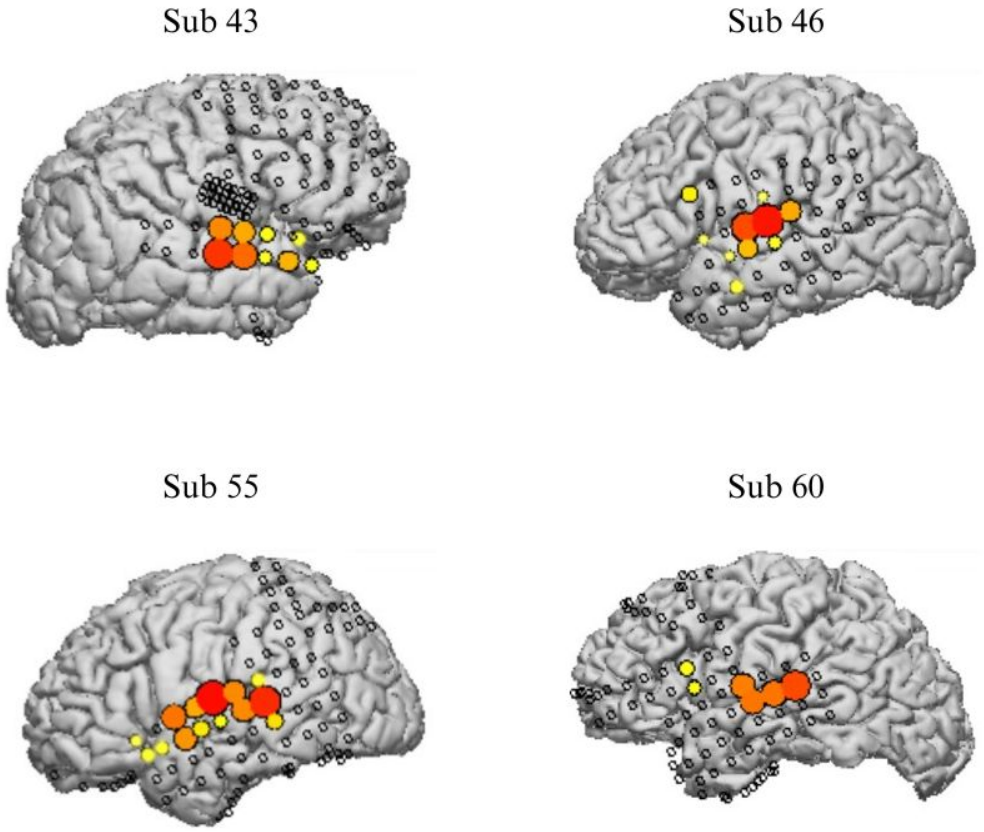

Figure 1.1: The 4 subjects with highest correlation with speech envelope. [1]

Literature Review

Indian Institute of Information Technology Kottayam

S no.	Author(s) and Pub. Year	Study Title	Key Findings	Limitations	Relevance to Project
1.	Cheah, K. H., Nisar, H., Yap, V. V., & Lee, C. Y. (2020).	Convolutional neural networks for classification of music-listening EEG: comparing 1D convolutional kernels with 2D kernels and cerebral laterality of musical influence [2]	The CNN model achieved a 98.94% accuracy in tenfold cross-validation for binary classification and 97.68% for three-class classification during validation phases. Test accuracies were slightly lower but remained impressively high at 97.46% for binary classification and 95.71% for three-class classification. Technologies: 2D-CNN, Spatiotemporal Convolution, FC-DNN	The paper lacks a comparison of its EEG classification results with state-of-the-art methods on a public benchmark, hindering the assessment of its relative performance.	The paper demonstrates the effectiveness of CNNs in classifying EEG data associated with music listening, which aligns with BCI applications in decoding brain signals.
2.	Eger, S., Youssef, P., & Gurevych, I. (2019).	Is it time to swish? Comparing deep learning activation functions across NLP tasks [3]	The study conducted a comprehensive comparison of 21 activation functions across eight NLP tasks. Swish, defined as $f(x)=x \cdot \sigma(x)$, was noted for its smoother gradients and was among the top performers in several tasks. Technologies: Sequence Tagging, RNN	While the study provided a broad comparison of activation functions, it did not delve deeply into the theoretical underpinnings of why certain functions, performed better or worse in specific contexts.	The findings from the paper validate the choice of using Swish due to its smoother gradients and potential for better generalization, aligning with our objectives in enhancing neural network performance for NLP tasks.
3.	Choi, D., Shallue, C. J., Nado, Z., Lee, J., Maddison, C. J., & Dahl, G. E. (2019)	On empirical comparisons of optimizers for deep learning. [4]	The paper demonstrates the significant impact of hyperparameter tuning protocols on the performance of optimizers like Adam, highlighting that with optimal tuning, Adam can perform as well as or better than SGD and Momentum across various tasks. Technology: First order optimization, Adam, SGD	The study's conclusions are heavily dependent on the hyperparameter tuning protocol and computational budget, which might not be replicable in all practical scenarios, especially for those with limited resources to extensively tune Adam's parameters.	The paper highlights Adam's advantages in gradient management and model generalization, underscoring its suitability for neural network optimization.

Literature Review

S no.	Author(s) and Pub. Year	Study Title	Key Findings	Limitations	Relevance to Project
4.	Zheng, J., Liang, M., Sinha, S., Ge, L., Yu, W., Ekstrom, A., & Hsieh, F. (2021)	Time-frequency analysis of scalp EEG with Hilbert-Huang transform and deep learning [5]	This study introduces a data-driven approach utilizing Hilbert-Huang Transform (HHT) for the time-frequency analysis of scalp EEG signals, aiming to capture the individual variability in brainwave frequencies which traditional methods might overlook. Technology: HHT, Time frequency analysis	This study acknowledges the computational complexity associated with the Hilbert-Huang Transform (HHT), which may limit its application in real-time EEG analysis.	The paper presents evidence that HHT-based method and metrics can significantly improve the accuracy of EEG signal classification, contributing to more reliable and effective diagnostic tools.
5.	Lee, S. H., Lee, Y. E., & Lee, S. W. (2021)	Voice of your brain: Cognitive representations of imagined speech, overt speech, and speech perception based on EEG [6]	The study employed a deep neural network (DNN) that captures temporal-spectral-spatial features from EEG data to classify nine subjects based on their EEG during imagined speech, overt speech, and speech perception tasks. This approach demonstrated the possibility of subject identification using single-channel EEG of imagined and overt speech. Technology: iEEG, Electrode Placement	The analysis of functional connectivity was focused on channels located in the Broca's and Wernicke's areas, known for their involvement in speech processing. This may overlook the contributions of other brain regions to the cognitive representations of speech.	This research highlights the pivotal role of electrode placement in areas like the Broca’s area and superior temporal gyrus for enhanced speech signal decoding, pertinent to projects focusing on advanced neural-based communication systems.

Current Work

- Introduction to 2D CNN: For this study, we employed a 2D Convolutional Neural Network (2D-CNN) to analyze EEG data [9]. Two-Dimensional Convolutional Neural Networks (2D-CNNs) are specialized neural networks tailored for grid-like data, such as images. They excel in capturing both local and global patterns, making them ideal for tasks like image and audio processing.

- In this study, 2D-CNNs were used to process spectrogram data derived from EEG recordings. The data was split into training and test sets, normalized, and reshaped to fit the required 3D structure. A sliding window approach was employed to convert the data into 3D blocks.

- The 2D-CNN model consisted of three convolutional layers with (13, 13) kernel sizes and Swish activation functions. Dropout layers were integrated to prevent overfitting, and max pooling layers were used for dimensionality reduction. The architecture included flatten and densely connected layers with Swish activation. A final output layer matched the shape of the training spectrogram.

Current Work

- We utilized the Swish activation function [3] instead of ReLU due to its smoother gradients and potential for better generalization. Mathematically, Swish is defined as f(x) = x * sigmoid(x), where sigmoid is the logistic sigmoid function.

- The model was compiled with the 'Adam' optimizer [4] and 'mean squared error' loss function. Techniques such as early stopping and learning rate reduction were employed to control overfitting. Training occurred over 100 epochs with a batch size of 128.

- After training, the model's predicted spectrogram was saved and inverse transformed by scaling it back using the standard deviation and adding the mean. This process restored the original scale of the spectrogram data for further evaluation.

Current Work

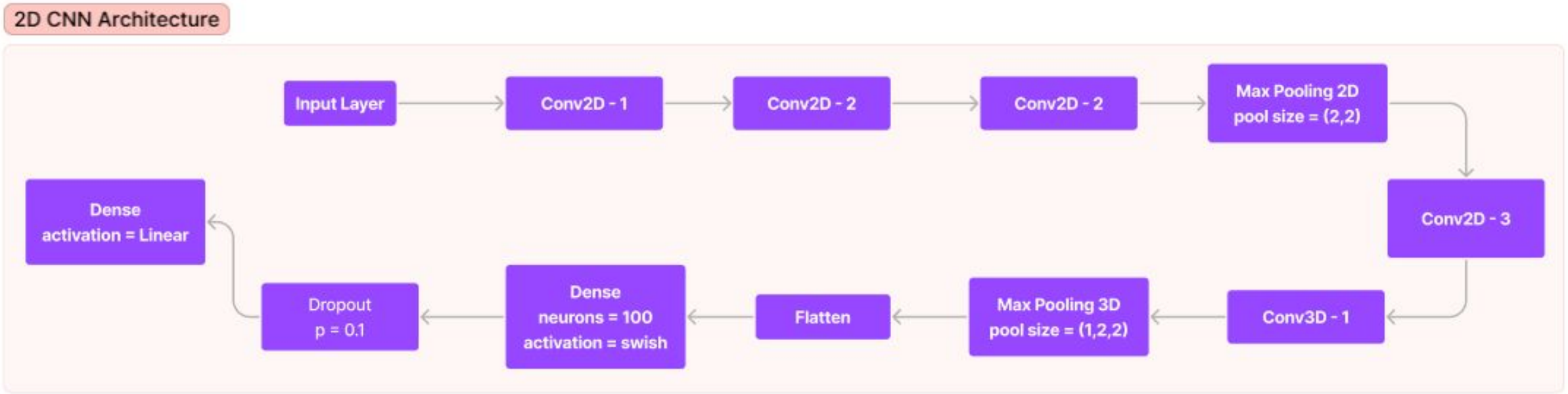

Figure 2.1: Figure illustrates the 2D-CNN architecture.

Current Work

Experimental Results

- In speech synthesis and enhancement systems, mel spectrograms play a vital role in accurately representing acoustic features. However, generated spectrograms might be over-smoothed, leading to a compromise in the quality of synthesized speech.

- Current modelling methods may neglect correlations present in both time and frequency domains during the generation of speech mel spectrograms. This oversight often results in the production of blurry and over-smoothed spectrograms.

- Due to these limitations, comparing the visual similarity of predicted and ground truth mel spectrograms isn't appropriate. Metrics like SSIM may not effectively capture the required precision and correlation in acoustic representations.

- Consequently, we rely on quantitative evaluation using Mean Squared Error (MSE) to assess model performance. We'll compare the lowest training error and best validation MSE for the FC-DNN and 2D-CNN models applied to the selected prime subjects.

Experimental Results

Subject	Minimum 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

Figure 3.1: Performance of the Fc-DNN for each subject

Experimental Results

Subject	Minimum 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	0.8576

Figure 3.2: Performance of the 2D-CNN for each subject

Experimental Results

- The FC-DNN model was trained on data from five different subjects, and the results are summarized in Table 3.1.

- Among the subjects, subject 43 exhibited the lowest training loss, indicating the model's effective fit to the training data. Furthermore, subject 43 also showed the best generalization performance, with the lowest validation MSE, suggesting the model's ability to perform well on unseen data.

- Similar to the FC-DNN, the 2D-CNN model was trained on data from the same five subjects, and the results are detailed in Table 3.2.

- Subject 60 recorded the lowest training loss among the subjects, indicating a strong fit of the model to the training data. However, subject 55 demonstrated the best generalization performance, with the lowest validation MSE, implying superior performance of the 2D-CNN model on unseen data.

Experimental Results

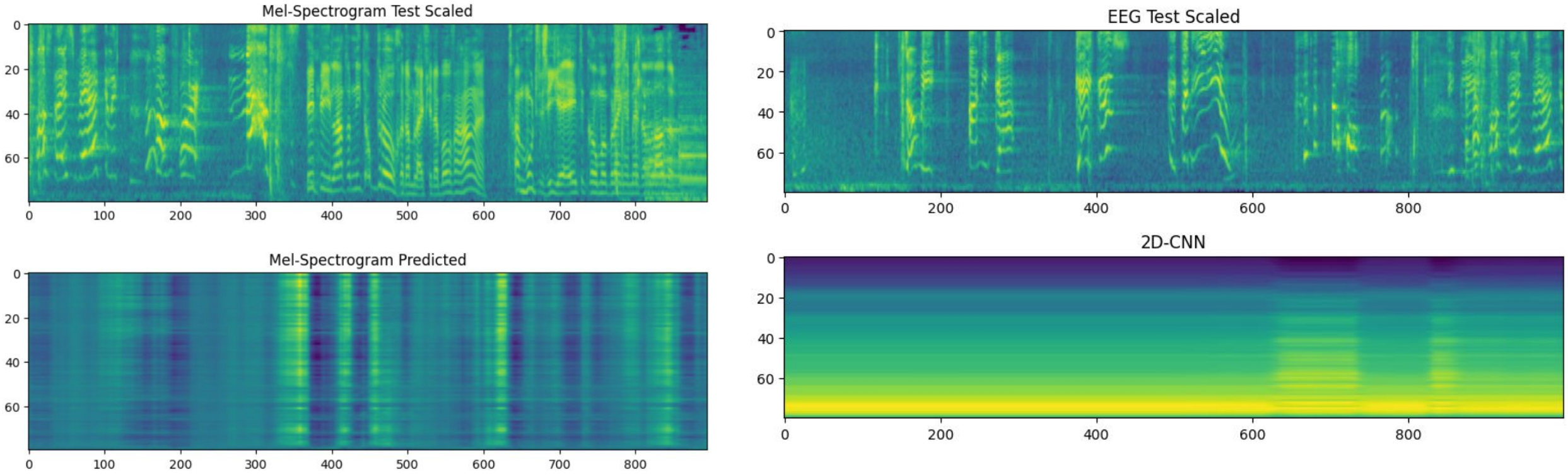

Fig 3.3: Mel Spectrograms for sub 38
[L] FC-DNN and [R] 2D-CNN

Experimental Results

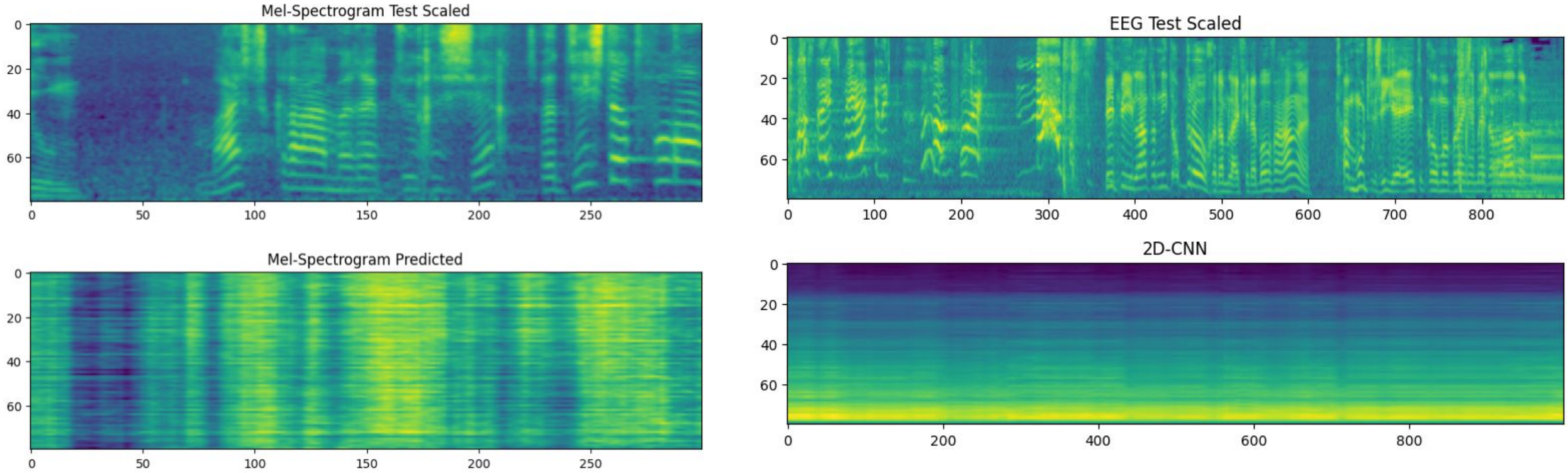

Fig 3.3: Mel Spectrograms for sub 43
[L] FC-DNN and [R] 2D-CNN

Experimental Results

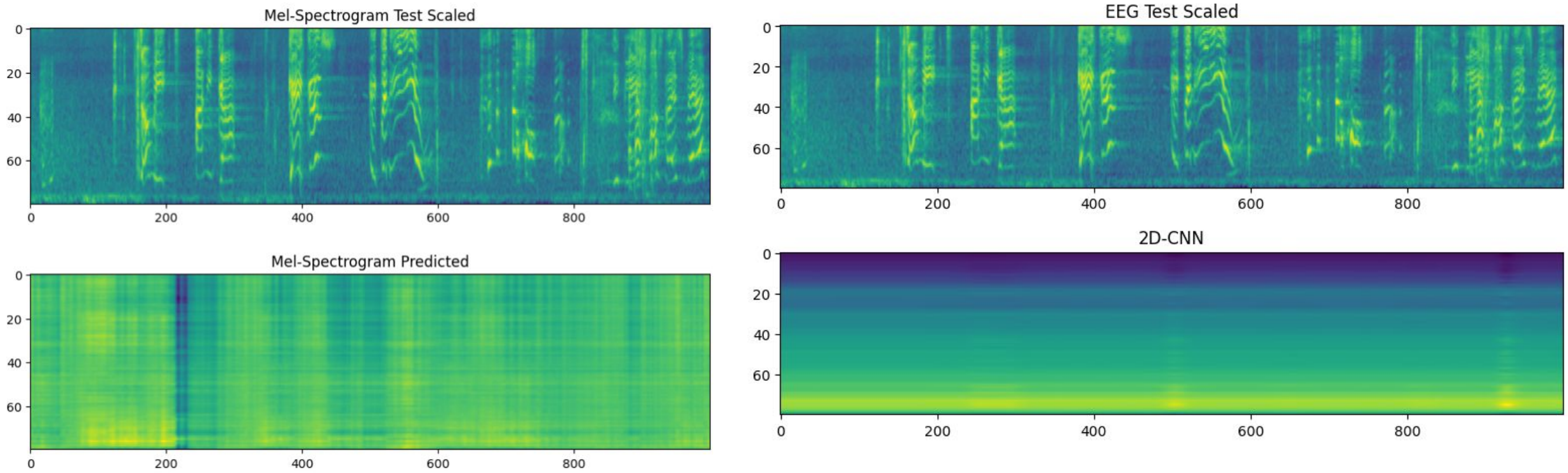

Fig 3.3: Mel Spectrograms for sub 46
[L] FC-DNN and [R] 2D-CNN

Experimental Results

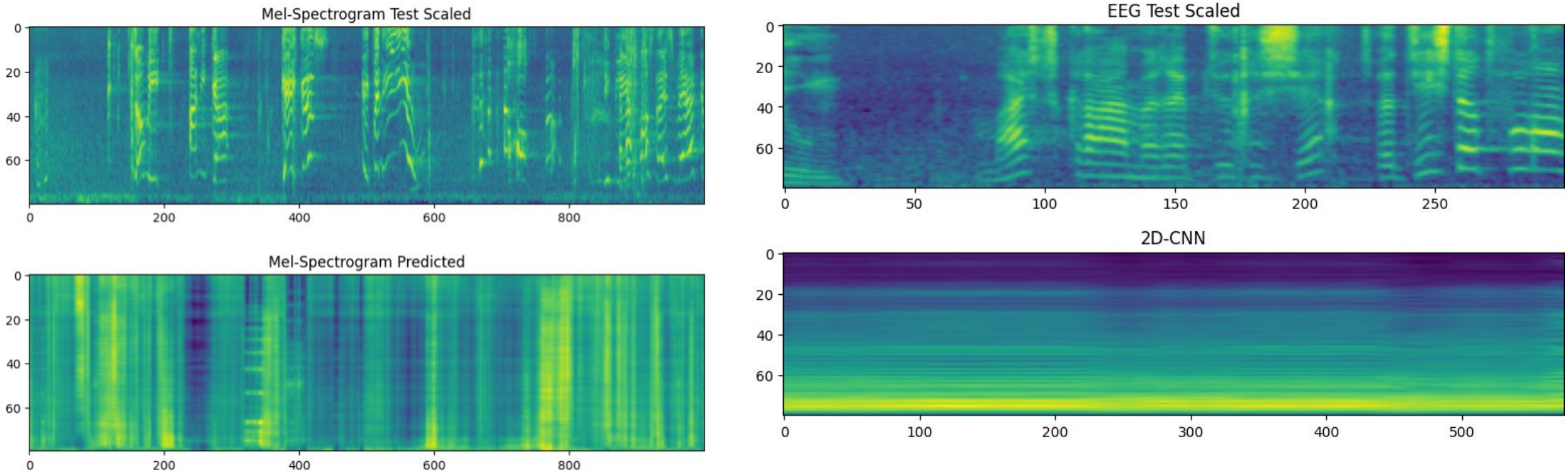

Fig 3.3: Mel Spectrograms for sub 55
[L] FC-DNN and [R] 2D-CNN

Experimental Results

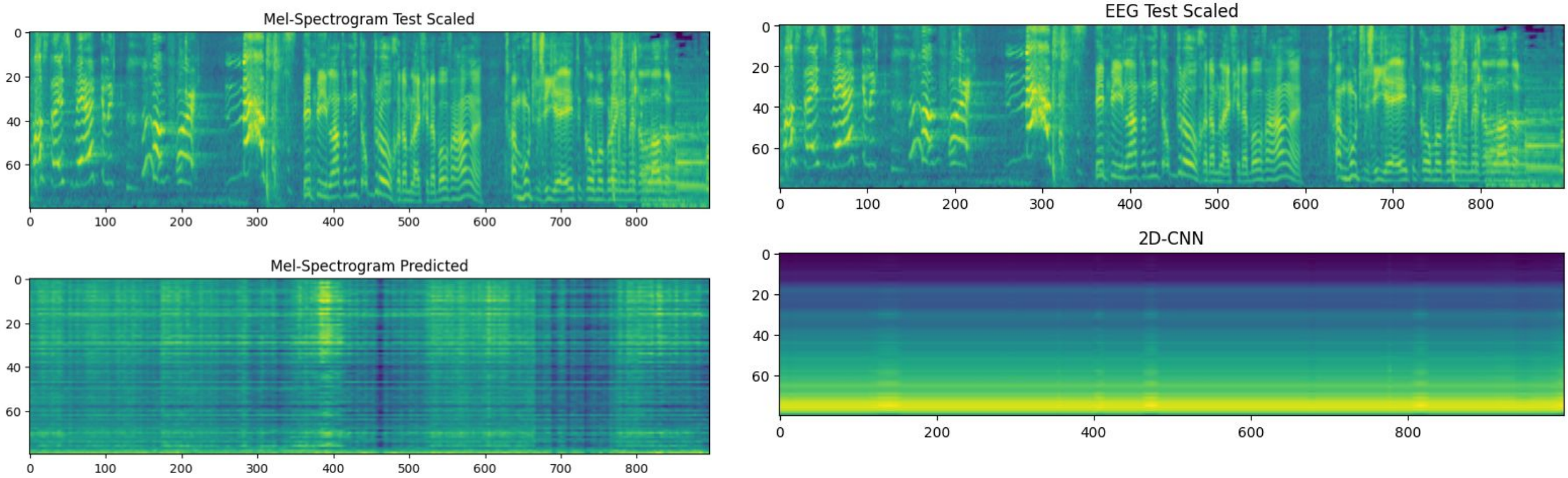

Fig 3.3: Mel Spectrograms for sub 60
[L] FC-DNN and [R] 2D-CNN

DEMONSTRATION

Conclusion

- Our investigation into deep learning models for decoding brain activity during passive listening and spoken speech reveals promising potential, particularly with the Fully Connected Deep Neural Network (FC-DNN) and 2D Convolutional Neural Network (2D-CNN) architectures. These models consistently reduce test loss, with some instances showing a decrease in validation loss, albeit limited by early stopping mechanisms.

- Despite successes, significant challenges persist. Difficulty in achieving satisfactory accuracy for both validation and test sets concurrently hampers the decoding process. While the 2D-CNN captures intensity accurately and the FC-DNN offers more detail, both fall short in generating realistic spectrograms. Consequently, synthesized speech clarity remains below anticipated levels.

References

1. Julia Berezutskaya, Mariska J. Vansteensel, Erik J. Aarnoutse, Zachary V. Freudenburg, et al. (2022). Open multimodal iEEG-fMRI dataset from naturalistic stimulation with a short audiovisual film.

2. Cheah, K. H., Nisar, H., Yap, V. V., & Lee, C. Y. (2020). 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.

3. Eger, S., Youssef, P., & Gurevych, I. (2019). Is it time to swish? Comparing deep learning activation functions across NLP tasks. arXiv preprint arXiv:1901.02671.

References

4. Choi, D., Shallue, C. J., Nado, Z., Lee, J., Maddison, C. J., & Dahl, G. E. (2019). On empirical comparisons of optimizers for deep learning.

5. Zheng, J., Liang, M., Sinha, S., Ge, L., Yu, W., Ekstrom, A., & Hsieh, F. (2021). Time-frequency analysis of scalp EEG with Hilbert-Huang transform and deep learning. IEEE Journal of biomedical and health informatics, 26(4), 1549-1559.

6. Lee, S. H., Lee, Y. E., & Lee, S. W. (2021). Voice of your brain: Cognitive representations of imagined speech, overt speech, and speech perception based on EEG. arXiv preprint arXiv:2105.14787.

References

7. 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.

8. Défossez, Alexandre, et al. "Decoding speech perception from non-invasive brain recordings." Nature Machine Intelligence 5.10 (2023): 1097-1107.

9. Schirrmeister, R. T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M. et al. Deep learning with convolutional neural networks for eeg decoding and visualization. Hum. Brain Mapp. 38, 5391–5420 (2017).