BME FACULTY OF ELECTRICAL ENGINEERING AND INFORMATICS

MŰEGYETEM 1782

Brain-to-Speech Synthesis Project

Deep Learning

Cuadros Rivas, Alejandra Paola - KK5459

Konyakin, Dmitriy - JHGBFA

Xu, Yang - LYQU2F

Abstract

The goal of the Brain-Computer Interface (BCI) is to provide a natural or near-natural communication channel for people who are unable to speak due to physical or neurological impairments. Speech serves as the primary means of human communication; however, many people lose this ability due to disease or health problems. By synthesising speech in real time from measured neural activity (BRAIN-TO-SPEECH), it will be possible to achieve natural speech and significantly improve quality of life, especially for individuals with severe communication limitations. The speech production dataset from intracranial EEG (SingleWordProductionDutch) is used in this project.

In this project, we will delve into the field of BRAIN-TO-SPEECH, where we plan to replace the linear regression model with a deep neural network, which will in turn lead to the development and training of novel neural network architectures. This study delves into the field of audio processing by comparing the performance of linear regression and deep learning models, highlighting the significant advantages of deep learning in the spectrogram reconstruction task. Hyperparameter optimization via the Optuna library improves the deep learning model performance, and the introduction of a convolutional neural network structure further enhances the performance. Comparison of the results with other models shows that the optimised deep learning model outperforms in terms of both MAE and R^2, clearly demonstrating its superior performance in the spectrogram reconstruction task with respect to linear models. Overall, this study provides a feasible solution for the application of deep learning in audio processing, and also lays a deep foundation for the application of neural networks in different fields, expanding the interest in the field of deep learning.

Literature Review

1. The development of BRAIN2SPEECH

Brain-speech interfaces (BRAIN2SPEECH) are a brain-computer interface research direction that has received much attention in recent years. By synthesising speech in real time, BRAIN2SPEECH aims to provide a natural or near-natural means of communication for those who are unable to speak due to physical or neurological impairments[1]. Speech occupies a central place in human communication, yet many people have lost this essential ability due to disease or health problems. The BRAIN2SPEECH technology, which aims to achieve natural speech, is expected to significantly improve the quality of life of people with severe communication limitations.

2. BRAIN2SPEECH research methods in recent years

The human brain is considered to be the most complex human organ, comparable to a very powerful and sophisticated computer, and until today no one has succeeded in recreating and modelling its entire structure [2]. Recently, rapid advances in medicine and information technology have opened the era of the brain-computer interface (BCI), especially its non-invasive version, based on electroencephalography (EEG) [3][4]. Nowadays, the acquisition of EEG is basically performed in a minimally invasive way, without causing damage to the experimenter's brain, since they do not require any surgical intervention and their implementation is neither difficult nor dangerous [5]. In the BRAIN2SPEECH study, previous work has focused on the use of linear regression models to decode neural signals for speech synthesis. These studies have attempted to reconstruct speech features by analysing signals such as brainwaves to achieve conversion from neural activity in the brain to speech [1]. However, these linear models have limitations in capturing the complexity of speech. Some of the literature will also implement brainwave-to-speech conversion through neural networks [6].

3. Advances in research using neural networks

In order to overcome the limitations of linear models, the latest research trend is to introduce deep neural networks to decode neural signals more accurately and enable more natural speech synthesis. These techniques are architectural designs that take into account the characteristics of brain signals [7]. They are often used to decode human intentions through motor imagery or event-related latent potentials and have shown superior performance to traditional machine learning methods. Recently, several studies have attempted to find the optimal features of EEG by deep neural networks based on the three main features of EEG (temporal, spectral and spatial features) [8]. In addition, EEG-based speaker recognition studies have actively applied machine learning or deep learning techniques [9]. Deep learning may be effective in capturing individual features from brain signals to validate individual features. Therefore, the introduction of deep neural networks provides the BRAIN2SPEECH technique with more flexible and powerful tools to better capture complex patterns in neural signals and improve the quality of speech reconstruction.

4. Our Research Direction

In this study, we will further explore the BRAIN2SPEECH domain by trying to enhance the linear regression model with a deep neural network using datasets already collected by others. By using speech generation datasets from intracranial EEG, we will develop and train novel neural network architectures and improve the efficiency of the models by tuning some hyperparameters to improve the accuracy and naturalness of speech synthesis. We believe this will improve our knowledge and ability to apply deep learning in different domains.

Data

1. Data description

This study included a total of 10 participants with a mean age of 32 years (range 16 to 50 years, from Table 1), 5 males and 5 females. The participant data is derived from the study conducted by Verwoert et al. in 2022 [1]. These participants were implanted with stereo electroencephalography (sEEG) electrodes as part of the clinical management of their epilepsy. The location of the electrodes was determined solely on the basis of clinical need. Data recording was performed under the supervision of experienced medical personnel. In addition, it was noted that the native language of all participants was Dutch. In order to maintain participant anonymity, participants' voices were pitch-shifted with an offset of between 1 and 3 semitones to maintain constancy throughout the recording.

Table 1. Basic information about participants

participant ID	**Age[1]**	**Sex[2]**
sub-01	20	F
sub-02	43	M
sub-03	24	M
sub-04	46	F
sub-05	50	F
sub-06	16	M
sub-07	47	M
sub-08	22	F
sub-09	20	F
sub-10	36	M

[1] Age of the participant at time of testing, units: years.

[2] Biological sex of the participant, 'F': 'female', 'M': 'male'

about their data, participants were asked to read aloud words displayed in front of them. A random word from the stimulus bank (Dutch IFA corpus 62, expanded word forms for the numbers one to ten) was displayed on the screen for 2 seconds, during which participants read the word aloud once. This relatively long time window takes into account differences in word length and speed of pronunciation. After the word, a cross was displayed for 1 second. This process was repeated a total of 100 times for a total recording time of 300 seconds per participant. The presented stimuli and times were saved for subsequent processing, referred to later as stimulus data.

Participants were implanted with platinum-iridium sEEG electrode spindles. Neural data were recorded using two or more Micromed SD LTM amplifiers with 64 channels each. Electrode contacts were referenced to a common white matter contact. Data were recorded using 1024 Hz or 2048 Hz and subsequently downsampled to 1024 Hz.Audio data were recorded at a sampling rate of 48 kHz via the inbuilt microphone of a recording laptop (HP Probook). To ensure participant anonymity, audio data were subsequently pitch-converted, processed using LibRosa63 and synchronised with neural, audio and stimulus data via LabStreamingLayer64. The electrodes were mainly located in the superior temporal sulcus, hippocampus and inferior occipital sulcus. The locations of the electrodes are shown in Figure 1.

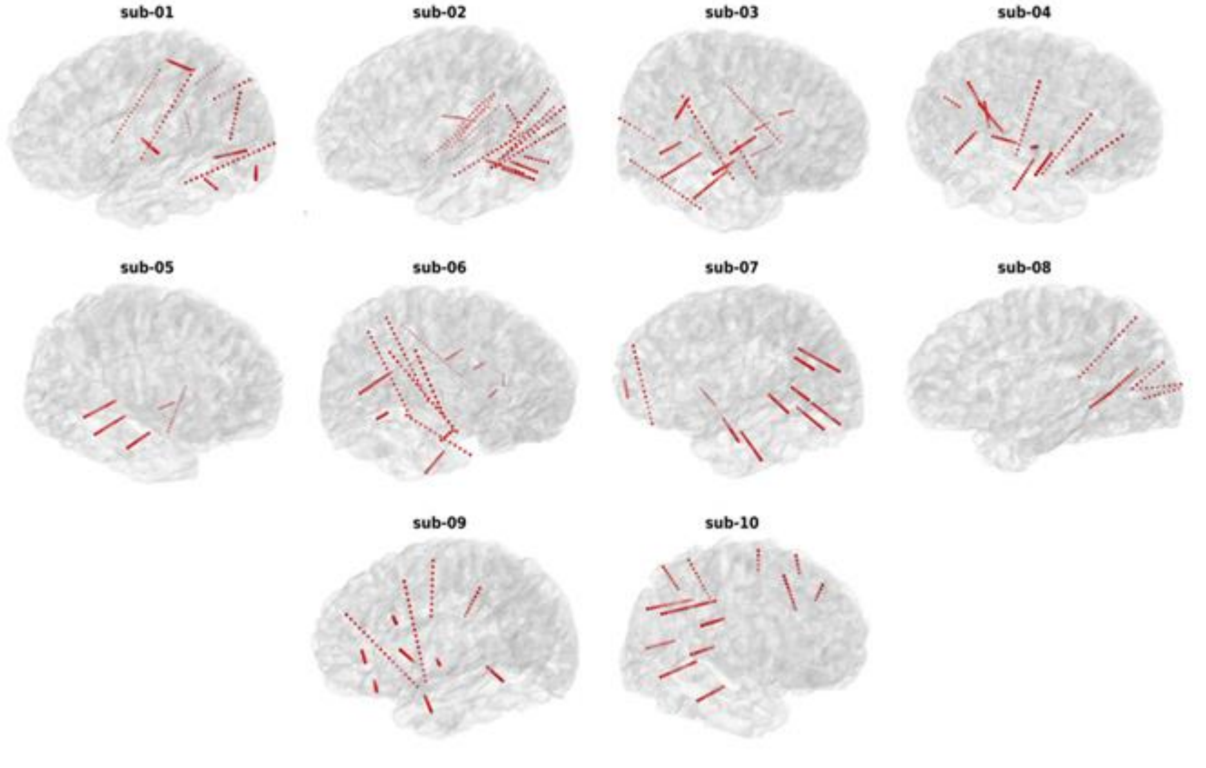


Fig. 1 Electrode locations of each participant in the surface reconstruction of their native anatomical MRI. Each red sphere represents an implanted electrode channel.

Within the participant's personal data, each subject-specific folder contains .tsv files regarding implanted electrode coordinates, recording montages, and event markers. In addition, the _ieeg.nwb file contains raw data streams for iEEG, audio, and stimuli. Detailed descriptions of the recording aspects and .tsv columns can be found in the corresponding .json files (e.g. participants.json). Based on the previous data research [1], and compared to the results of the other participants, the choice of the participant is informed by the mean correlation coefficient results obtained by sub-06. We can select some of them to check from Table 2.

Table 2 Data about some sample from participant sub-6

	name	x	y	z	size
0	LFA1	-40.608156	30.370011	9.628015	5
1	LFA2	-43.844691	29.366746	10.646079	5
2	LFA3	-47.081227	28.363481	11.664143	5
3	LFA4	-50.317763	27.360216	12.682207	5
4	LFA5	-53.554298	26.356951	13.70027	5
5	LFB1	-40.381965	8.461976	19.005244	5
6	LFB2	-43.913651	8.67062	18.776345	5
7	LFB3	-47.445337	8.879265	18.547445	5
8	LFB4	-50.977024	9.087909	18.318545	5
9	LFB5	-54.50871	9.296553	18.089646	5
10	LFB6	-58.040396	9.505198	17.860746	5

Table 2 describes the spatial coordinate information for a series of electrodes labelled LFA (Left Frontal Anterior) and LFB (Left Frontal Back)[3]. Each electrode has corresponding 3D coordinates (x, y, z), and a numerical value indicating the size of the electrode (size). After that we will prepare the data for model training using the information collected through the above mentioned ways.

This study contains 3 types of data, that is Audio, Stimulus and iEEG (iEEG stands for intracranial electroencephalogram). From Fig.2, we can see a sample waveform of audio data. Fig.3 is a sample visualization of iEEG.

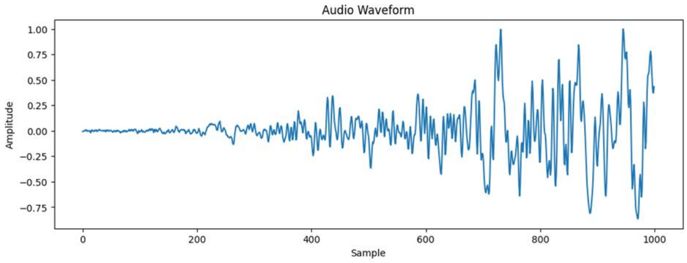


Fig.2 Sample waveform

[3] In the actual data, there are different positions such as LFC, LO, LT, RA, RF, and others. Details can be seen in the output of the code.

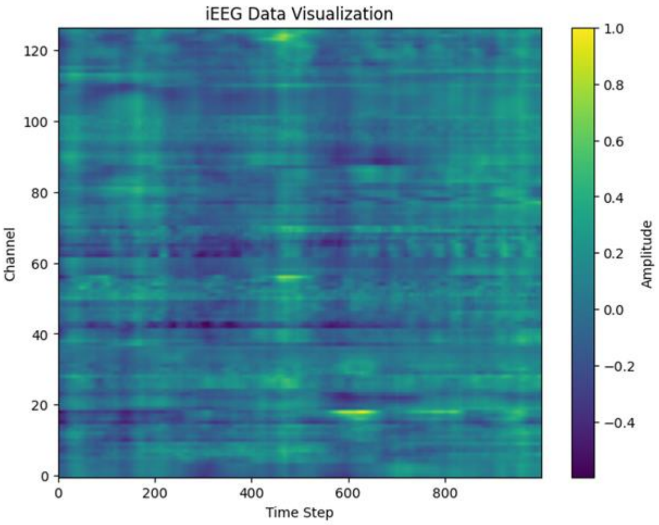


Fig.3 sample visualization of iEEG

2. Data preparation

In this phase, we performed fine-grained processing and feature extraction of electroencephalography (EEG) and audio data. By importing the necessary Python libraries, implementing the Hilbert transform auxiliary functions and developing multiple feature extraction functions, we defined the parameters required for the processing in the preparation phase. The main workflow consists of reading participant data from Neurodata Without Borders (NWB) files, calling feature extraction functions for EEG and audio data processing, and finally saving the feature-extracted and processed data to disk. This sequence of steps lays the foundation for subsequent deep learning model training and evaluation, allowing us to efficiently load and utilise a single speaker's dataset.

We first performed the temporal smoothing of the EEG feature values to reduce the noise, Fig.4 shows the processed EEG. The horizontal axis indicates the sequential time window, and the vertical axis shows the smoothed EEG feature values. The figure contains multiple overlapping lines of different colours, indicating multiple EEG features plotted simultaneously. The richly coloured lines indicate that these features often change simultaneously, while some distinct spikes indicate moments of higher values. Overall, the smoothing process helps to reveal underlying trends in the EEG data by suppressing short-term fluctuations.

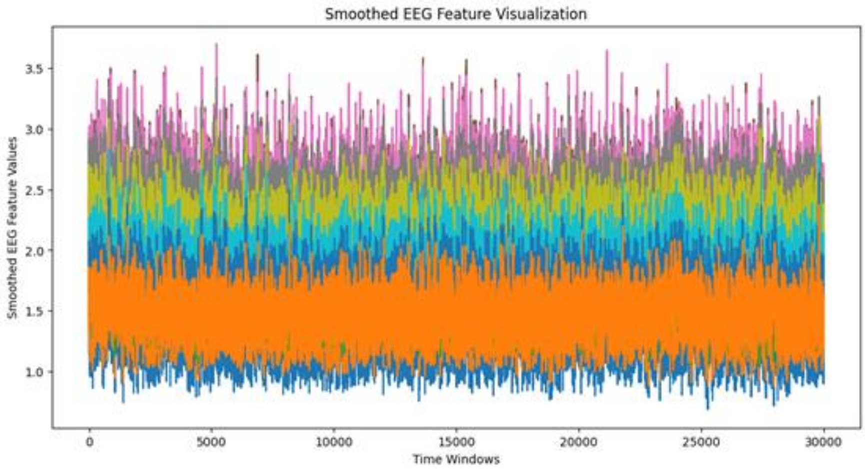


Fig.4 Smoothed EEG Feature Visualization

Then, Fig.5 shows a heat map of the electroencephalogram (EEG) channel data over time. The visualisation primarily demonstrates the homogeneity of the feature values, which are dominated by the blue colour, indicating that these values are predominantly lower. There are no apparent patterns or anomalies, suggesting that the EEG readings are relatively stable between channels and throughout the recording time.

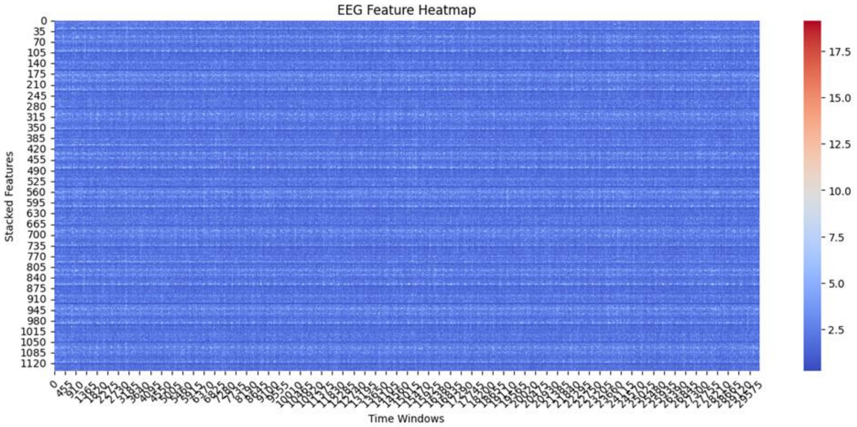


Fig.5 EEG Feature Heatmap

We also reconstruct the original audio signal from the spectrogram. The script converts the log Mel spectrogram values back into a spectrum by training a linear regression model, and uses PCA to reduce the dimensionality. We evaluated the quality of the reconstructed spectrograms, saving results such as correlation coefficients, explained variance of the reconstructed spectrograms and PCA components.

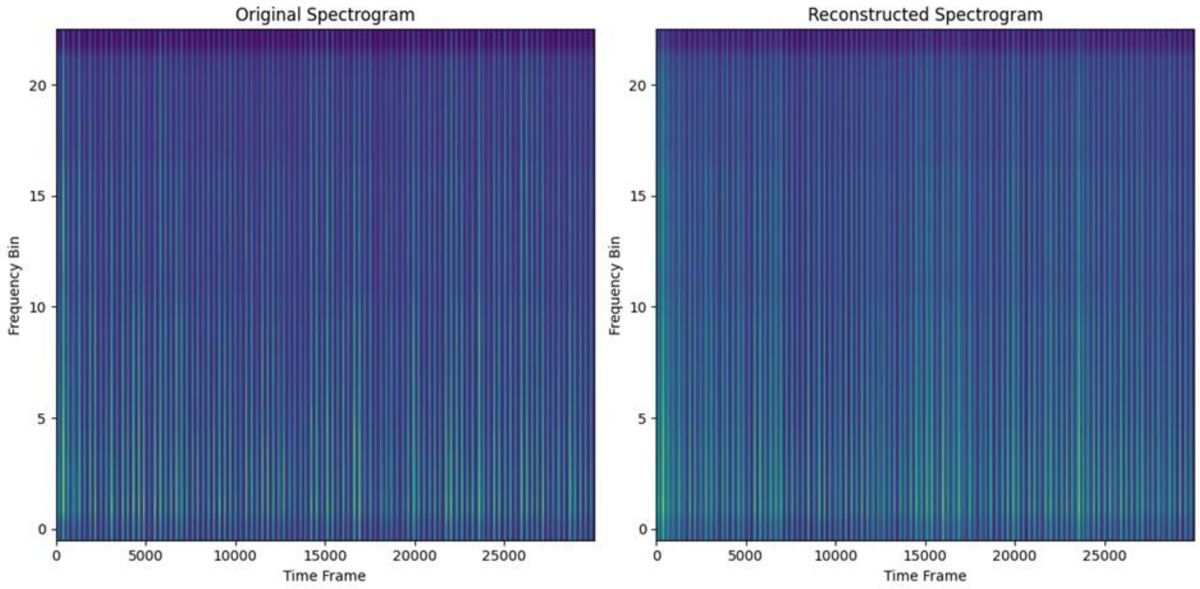


Fig.6 The compare of Original and Reconstructed Spectrogram

From Fig.6 we can find that the two spectra are visually extremely similar, suggesting that the reconstructed spectra successfully mimic the original spectra in terms of pattern and structure. We can also visualized these two spectrograms in Fig.7 and get the similar result.

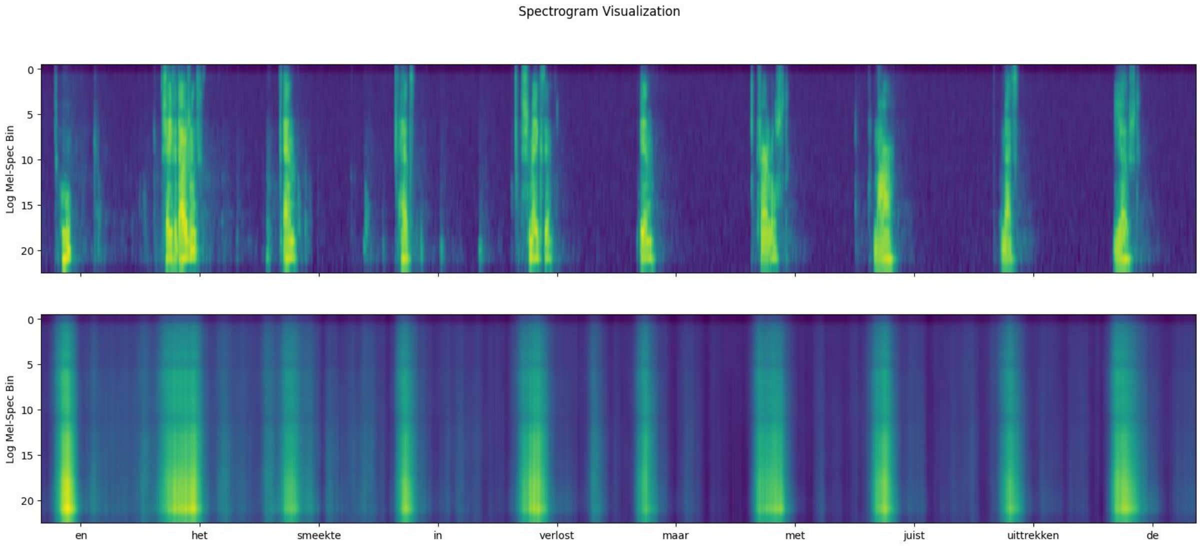


Fig.7 Spectrogram Visualization

Furthermore, for conclusive validation, a graph was generated to juxtapose the original audio waveform with its reconstructed counterpart. From Fig.8, it can be inferred that despite discrepancies between the two audio segments, they remain fundamentally alike. These disparities might arise from signal processing inaccuracies, data preprocessing nuances, and other factors. Further analysis and experimentation are warranted to pinpoint the precise origins of these variations.

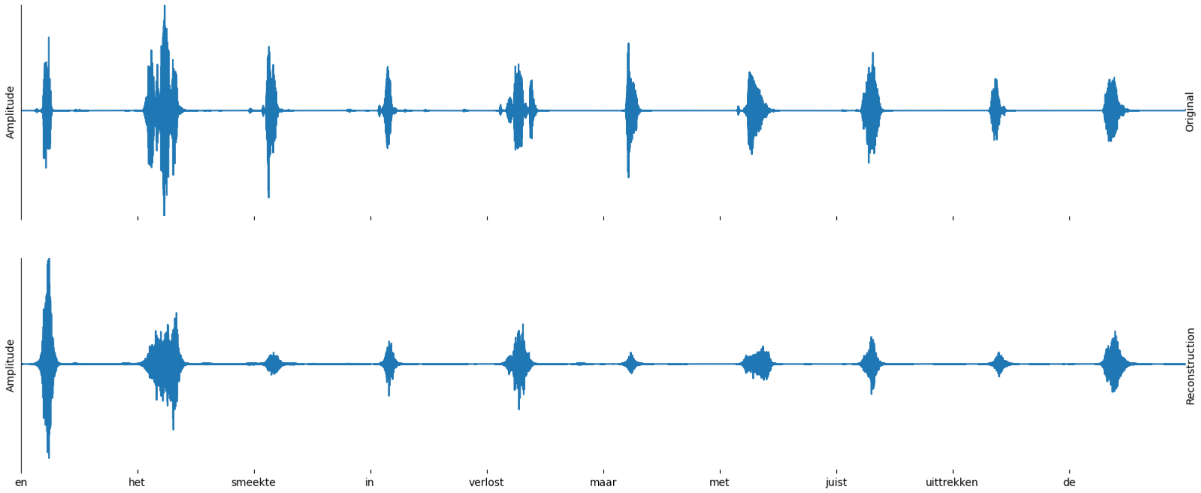


Fig.8 Waveform comparison

(The words is “en het smeekte in verlost maar met juist uittrekken de”)

We can also listen to the sound in the code. In the code after Fig.8, we output the original audio, and the comparison between the synthesised audio and the predicted audio, and although we can hear the difference(especially the predicted audio), there is a basic similarity. In conclusion, the synthesised audio and predicted audio closely match the original audio in terms of temporal alignment, but there appears to be an excess of low-frequency content, affecting the tonal quality.

Method and Model

The training and optimisation of our models is performed on the Colab platform. The requirements and version of the module can be seen on Table 3.

Table 3 Requirements

Module	Version	Module	Version
gdown	4.6.6	pynwb	2.5.0
h5py	3.9.0	pytorch_lightning	2.1.2
matplotlib	3.7.1	scipy	1.11.4
nibabel	4.0.2	seaborn	0.12.2
numpy	1.23.5	sklearn	1.2.2
pandas	1.5.3	torch	2.1.0+cu118

1. Model

This study introduces SpectrogramReconstructionNet, a deep learning model implemented in PyTorch. Its primary function is to process scalp electroencephalography (sEEG) data, generating corresponding spectrograms through training. The model employs Mean Absolute

Error (MAE) as the loss function in training and assesses performance using R-squared (R^2). The training objective is to adeptly reconstruct spectrograms from raw sEEG signals, offering insights into the relationship between the original signals and their frequency domain representation depicted by spectrograms. This model contributes to our understanding of transforming sEEG data into its spectral counterpart.

2. Modified by Optuna

Initially, this study employed the initial version of the model and later refined it using Optuna, an open-source framework designed for hyperparameter optimization. We can also see the Spectrogram segment from Fig.9.

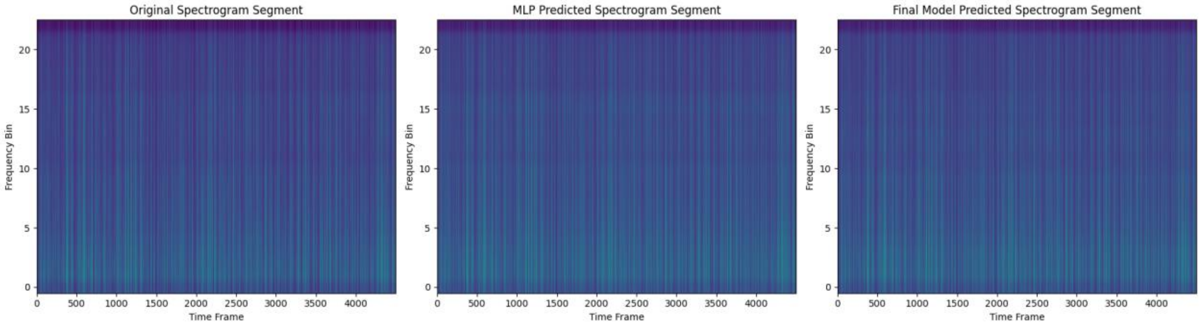


Fig.9 The spectrogram segment comparison

3. Modified by CNN

Convolutional Neural Networks (CNNs) excel in spectrogram reconstruction tasks due to their ability to efficiently extract hierarchical, local, spatial, and translationally invariant features. Through convolutional layers and pooling operations, CNNs are able to capture the complex hierarchical structure in spectrogram data, which is crucial for understanding the evolution of frequency patterns in time and space. Parameter sharing and nonlinear activation functions further enhance the generalisation ability of the model, enabling it to accurately reconstruct spectrograms. As a result, CNNs exhibit excellent performance in spectrogram data processing, providing an effective tool for pattern recognition and modelling of complex relationships in the task.

In our enhanced model, SpectrogramReconstructionNetCNN, we incorporate convolutional layers with max-pooling operations. This updated architecture comprises two convolutional layers followed by two fully connected layers. The convolutional layers are designed to capture hierarchical features within the input spectrogram data, while the fully connected layers process the flattened output, ultimately generating the final reconstruction. ReLU activation functions are applied after each convolutional layer and the first fully connected layer to introduce non-linearity. The model is trained using mean squared error loss and optimized with the AdamW optimizer. This refined architecture is particularly well-suited for tasks involving the reconstruction of spectrogram data. We also optimised the CNN model using Optuna to obtain the Final SpectrogramReconstructionNetCNN model. We can see the Spectrogram segment From Fig.10.

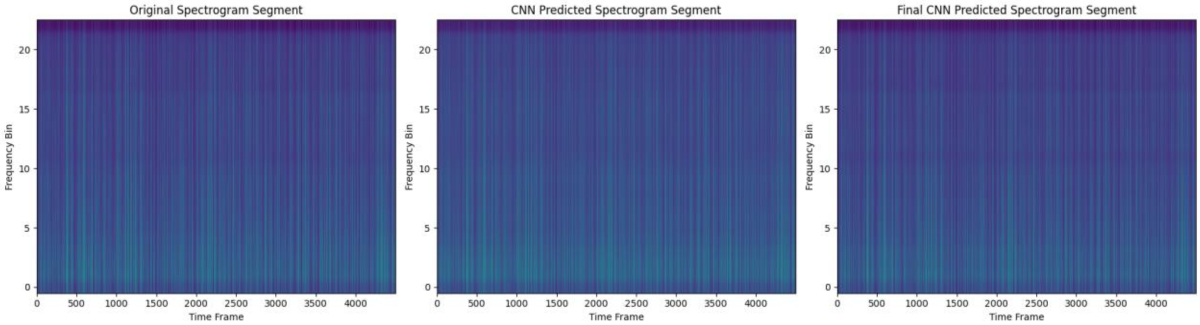


Fig.10 The spectrogram segment comparison

Result

Firstly, we will check the Linear Regression Model, by Mean Absolute Error (MAE), Mean Squared Error (MSE) and R-squared (R²). The result is in Table 4.

Table 4 Evaluation result for Linear Regression Model

Model	MAE	MSE	R2 Score
Linear Regression Model	0.671038	0.816182	0.798712

The model evaluation metrics indicate a satisfactory level of predictive accuracy. The MAE of approximately 0.671 signifies the average magnitude of prediction errors, with deviations around 0.671 units. The MSE of 0.816, though moderate, reflects the average of squared errors, considering larger errors more significantly. The R² score of around 0.799 demonstrates that the model explains approximately 79.87% of the variability in the outcome variable, indicating a strong fit and alignment with the observed data. Overall, these metrics affirm the model's effectiveness in making predictions.

Then we will check the results of deep learning models. The model's performance was assessed using metrics such as Mean Absolute Error (MAE) and R-squared (R²). We can see the result from Table 5.

Table 5 Evaluation result for Deep Learning Model

Model	MAE	R2 Score
SpectrogramReconstructionNet	0.303593	0.957334
SpectrogramReconstructionNet Optimized with Optuna	0.288412	0.961205
SpectrogramReconstructionNetCNN	0.347463	0.941735
SpectrogramReconstructionNetCNN optimized with Optuna	0.299689	0.953703

In the evaluation of four distinct models for spectrogram reconstruction, the initial performance metrics of each model were considered. The SpectrogramReconstructionNet exhibited a MAE of 0.303593 and an R-squared (R^2) score of 0.957334. Following optimization with Optuna, notable improvements were observed, resulting in a reduced MAE of 0.288412 and an enhanced R^2 score of 0.961205. On the other hand, the SpectrogramReconstructionNetCNN, tailored for convolutional neural network applications, displayed a comparatively higher MAE of 0.347463 and a slightly lower R^2 score of 0.941735. However, optimization with Optuna led to a more favorable performance, yielding a decreased MAE of 0.299689 and an improved R^2 score of 0.953703. Importantly, the optimized versions of both models outperformed their initial counterparts, highlighting the efficacy of the optimization process. It is noteworthy that all models, including the linear regression baseline, demonstrated superior performance, emphasizing the effectiveness of advanced models in enhancing predictive accuracy for spectrogram reconstruction in comparison to a linear regression approach.

Then we use the first model for Reconstruction Spectrogram, from Fig.11 and Fig.12 we can see the comparison between the original Spectrogram with the Spectrogram that was reconstructed by the best deep learning model.

Fig.11 Original Spectrogram

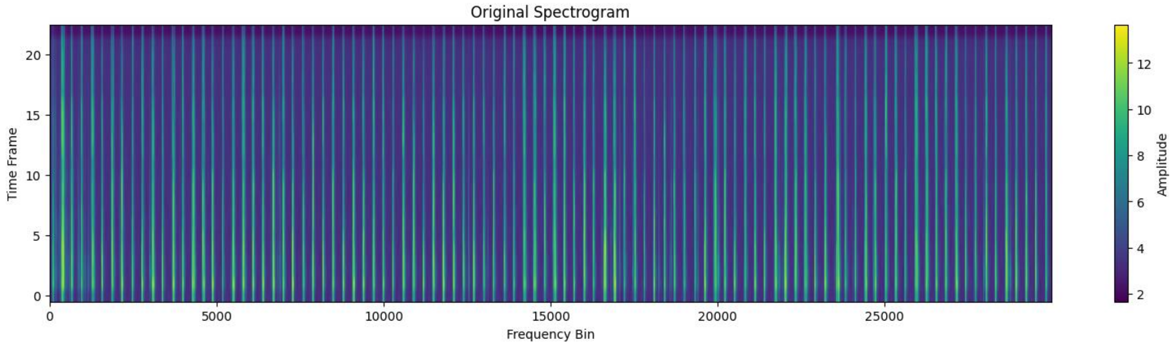
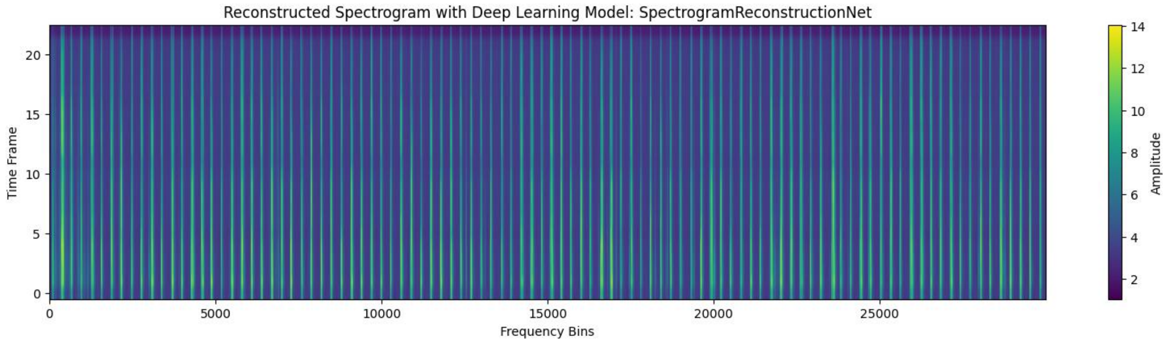


Fig.12 the Spectrogram that was reconstructed by the first deep learning model



From the comparison of the original spectrograms with the spectrograms reconstructed by the best deep learning model and find significant similarity between them. This finding indicates that our deep learning model performs well in reconstructing spectrograms efficiently.

Conclusion

In the field of Brain2Speech, more and more studies are beginning to employ neural networks for speech generation. This study delves into the application of deep learning in audio processing, especially the performance in the task of spectrogram reconstruction.

Firstly, we evaluate the efficacy of the linear regression model and highlight the significant superiority of the deep learning model in the spectrogram reconstruction task by comparing the performance of other deep learning models. Second, we successfully improved the performance of the deep learning model by performing hyperparameter optimisation using the Optuna library. This optimisation significantly reduces the mean square error (MAE) and improves the R^2 score, which enhances the model's ability to capture spectrogram features and improves the reliability of the model. Subsequently, we introduced a convolutional neural network (CNN) architecture to create the SpectrogramReconstructionNetCNN model, designed for spectrogram reconstruction. The model captures the hierarchical features in the spectrogram through convolutional layers and maximum pooling operations, and performs well compared to linear models, demonstrating more robust performance.

Finally, we compare the performance of all models, including the linear model (results derived from other studies). The results show that the deep learning model optimised with hyperparameters outperforms the other models, including the linear model, in both MAE and R^2. This clearly demonstrates the superior performance of deep learning in spectrogram reconstruction tasks relative to linear models.

In conclusion, our study digs deep into the application of deep learning models in audio processing, providing a range of viable solutions for the spectrogram reconstruction task. Through performance comparisons, we clearly demonstrate the significant advantages of deep learning models over linear models for this task. This project has established a deep foundation for applying neural networks in different fields and has expanded our interest in deep learning.

Reference

[1] Verwoert M, Ottenhoff M C, Goulis S, et al. *Dataset of speech production in intracranial electroencephalography*[J]. Scientific data, 2022, 9(1): 434.

[2] Kawala-Sterniuk A, Browarska N, Al-Bakri A, et al. *Summary of over fifty years with brain-computer interfaces—a review*[J]. Brain Sciences, 2021, 11(1): 43.

[3] Martins, N.R.; Angelica, A.; Chakravarthy, K.; Svidinenko, Y.; Boehm, F.J.; Opris, I.; Lebedev, M.A.; Swan, M.; Garan, S.A.; Rosenfeld, J.V.; et al. *Human brain/cloud interface*. Front. Neurosci. 2019, 13, 112.

[4] Kawala-Sterniuk, A.; Podpora, M.; Pelc, M.; Blaszczyszyn, M.; Gorzelanczyk, E.J.; Martinek, R.; Ozana, S. *Comparison of smoothing filters in analysis of EEG data for the medical diagnostics purposes*. Sensors 2020, 20, 807.

[5] Miller, K.J.; Hermes, D.; Staff, N.P. *The current state of electrocorticography-based brain–computer interfaces. Neurosurg*. Focus 2020, 49, E2.

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

[7] Qian W, Tan J, Jiang Y, et al. *Deep learning with convolutional neural networks for EEG-based music emotion decoding and visualization*[J]. Brain-Apparatus Communication: A Journal of Bacomics, 2022, 1(1): 38-49.

[8] Bhatti M H, Khan J, Khan M U G, et al. *Soft computing-based EEG classification by optimal feature selection and neural networks*[J]. IEEE Transactions on Industrial Informatics, 2019, 15(10): 5747-5754.

[9] Moctezuma L A, Torres-García A A, Villaseñor-Pineda L, et al. *Subjects identification using EEG-recorded imagined speech*[J]. Expert Systems with Applications, 2019, 118: 201-208.