Interspeech 2024
1-5 September 2024, Kos, Greece

# Toward Fully-End-to-End Listened Speech Decoding from EEG Signals

*Jihwan Lee, Aditya Kommineni, Tiantian Feng, Kleanthis Avramidis, Xuan Shi, Sudarsana Reddy Kadiri, Shrikanth Narayanan*

University of Southern California, USA

jihwan@usc.edu, akommine@usc.edu, tiantiaf@usc.edu, avramidi@usc.edu, xuanshi@usc.edu, skadiri@usc.edu, shri@usc.edu

## Abstract

Speech decoding from EEG signals is a challenging task, where brain activity is modeled to estimate salient characteristics of acoustic stimuli. We propose FESDE, a novel framework for **F**ully-**E**nd-to-end **S**peech **D**ecoding from **E**EG signals. Our approach aims to directly reconstruct listened speech waveforms given EEG signals, where no intermediate acoustic feature processing step is required. The proposed method consists of an EEG module and a speech module along with a connector. The EEG module learns to better represent EEG signals, while the speech module generates speech waveforms from model representations. The connector learns to bridge the distributions of the latent spaces of EEG and speech. The proposed framework is both simple and efficient, by allowing single-step inference, and outperforms prior works on objective metrics. A fine-grained phoneme analysis is conducted to unveil model characteristics of speech decoding. The source code is available here: `github.com/lee-jhwn/fesde`.

**Index Terms**: speech decoding, speech synthesis, EEG, neural activity, brain signals

## 1. Introduction

Brain-computer interfaces (BCI), particularly those targeting decoding listened or articulated speech from neural (brain activity) signals such as electroencephalography (EEG) and electrocorticography (ECoG), hold promise for improving life quality and rehabilitation outcomes for patients with communication disorders. Speech decoding technologies from brain activity also offer possibilities for seamless, immersive, and interactive entertainment, where command and control are coordinated without compromising sensitive information accessible to others. For instance, a research group recently successfully prototypes a digital avatar of a paralyzed patient relying on understanding and interpretation of ECoG signals [1] enabling the patient to communicate more freely.

Several approaches have been proposed to decode speech or text from neural activities such as EEG or ECoG signals [1–6]. For example, many efforts have proposed the use of convolution neural networks (CNN) to decode speech from EEG [7, 8] and ECoG signals [9] especially to reconstruct speech signals or representations, like mel-spectrogram, from brain activity signals. Furthermore, several methods to decode imagined speech from neural activities have also been proposed by [10, 11].

Decoding speech from EEG signals rather than ECoG signals is more practical and economical in real world scenarios. Acquiring ECoG signals is challenging and limited to certain types of patients, as a surgical intervention is necessary to implant the sensing electrodes. On the other hand, obtaining EEG signals only requires participants to wear a cap-like device. Hence, speech decoding frameworks based on EEG signals can easily be applied to larger group of users, while ECoG based approaches are primarily focused on patients with specific clinical conditions such as epilepsy or certain movement disorders.

However, modeling and decoding information from EEG recordings has been a significant signal processing challenge, due to the low signal-to-noise ratio (SNR) that is inherent in EEG measurements and sparsity of measurements. EEG devices attempt to measure electrical activity in the brain through non-invasive electrodes, spatially arranged on the scalp. Hence, the sensed signals are weakened, with low spatial resolution, and contaminated by various biological, environmental noises, such as muscle activity, cardiac activity, eye movements, power lines, or nearby electronic devices [12, 13]. Various approaches have been proposed to deal with these sources of variability, ranging from sophisticated pre-processing and decomposition [12] algorithms to representation learning frameworks that are trained in a label-agnostic, self-supervised regime [14, 15].

Decoding speech waveforms *directly* from EEG signals has numerous advantages when compared to multi-step approaches that usually incorporate signal-to-text estimation as an intermediate step. Speech contains rich information such as prosody including rhythm, intonation, emotion, and speaker identity that are lost in the typical written lexical form. Hence, a pipelined approach to processing EEG signals that first goes through text decoding, followed by text-to-speech (TTS), faces challenges in decoding high-fidelity speech from the loss of important prosodic and other paralinguistic information.

Recent development in TTS technology may facilitate direct speech decoding from EEG signals. Fully-end-to-end TTS systems, such as VITS [16], are TTS systems that directly synthesize speech in waveform, rather than intermediate acoustic features like mel-spectrograms or mel-frequency cepstral coefficients (MFCC) [17]. Fully-end-to-end TTS systems have numerous advantages over other TTS systems that require any additional intermediate acoustic feature mapping step. As the intermediate step is skipped, fully-end-to-end TTS systems are faster, simpler, and less vulnerable to error accumulation, compared to TTS systems that require an extra step of converting acoustic features to final waveforms. With such recent developments in the TTS domain, we believe the intermediate acoustic feature extraction step can be omitted.

In this paper, we propose **FESDE**, a novel framework for **F**ully-**E**nd-to-end **S**peech **D**ecoding from **E**EG signals. Our proposed framework directly generates waveform of listened speech given EEG signals, without any additional intermediate acoustic feature mapping step, such as to mel-spectrogram. To the best of our knowledge, this is the first framework that directly generates speech waveform given respective EEG signals. FESDE consists of three parts: the EEG module, the speech



10.21437/Interspeech.2024-223

module, and the connector. The EEG module learns to provide a better representation of EEG signals. The speech module aims to generate the speech waveform from the speech embeddings. The connector learns to map and convert the distribution of EEG embedding to speech embedding. Figure 1 illustrates the overall structure of the proposed framework. During inference, only the EEG encoder and the speech decoder are utilized, along with the connector.

The proposed framework has various advantages over previous approaches. It enables single-step inference and does not require any additional pipelines such as a vocoder. Hence, it is faster and more straightforward compared to previous multi-step approaches, while outperforming the previous approach in terms of objective metrics.

The main contributions of the paper are as follows:

- We propose a fully-end-to-end speech decoding framework for EEG signals, by incorporating EEG module and speech module, reducing multiple steps into one. To our best knowledge, this is the first attempt to directly reconstruct listened speech waveform from EEG signals.
- The proposed framework allows single-step inference, which is simpler and faster, while also outperforming the previous approaches in objective measures.
- We conduct phoneme analysis and present the characteristics of phonemes that are easy or challenging to decode by this approach.

# 2. Proposed Method

## 2.1. Model Architecture

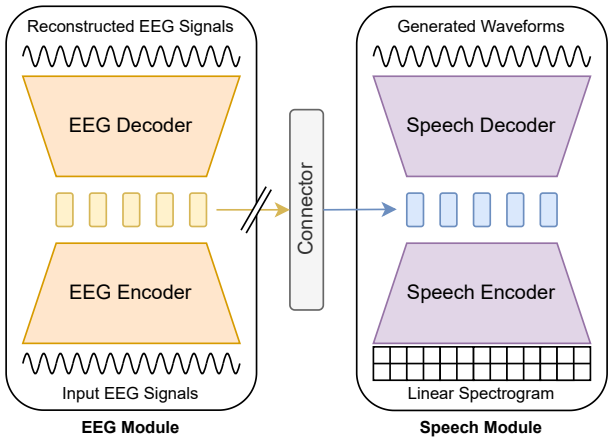



Figure 1: *The overall schematic of the proposed method. The EEG module is trained to produce descriptive representations of EEG signals. The speech module aims to generate speech waveform from the speech embeddings. The connector converts the distribution of the EEG embedding into the speech embedding. During inference, only the EEG encoder and the speech decoder are utilized, along with the connector.*

As shown in Figure 1, the proposed framework consists of three parts: the EEG module, the speech module, and the connector. The EEG and speech signals are handled by their respective modules. The connector bridges the two intermediate embeddings from EEG and speech. During inference, only the EEG encoder, the connector, and the speech decoder are used. For detailed implementation, refer to the source code.

### *2.1.1. EEG Module*

The EEG module is based on [15], a self-supervised learning framework for EEG signal representation. The EEG module consists of an encoder, that learns to encode input EEG signals into intermediate EEG embeddings, and a decoder, that learns to reconstruct EEG signals from those embeddings. The EEG encoder consists of convolution blocks and Structured State Space Sequence (S4) layers [18]. Each convolution block is composed of a 1D convolution layer, a dropout layer, layer normalization, and GELU activation. Each S4 layer consists of an S4 kernel estimator, gated linear unit (GLU) activation, dropout, and layer normalization. The S4 layers show superior performance in encoding long range temporal dependencies which make them suitable for encoding EEG signal [15]. The EEG decoder is composed of deconvolution blocks, where each block contains a 1D transpose convolution layer along with a dropout layer, layer normalization, and GELU activation.

### *2.1.2. Speech Module*

We base the speech module on VITS [16], one of the well performing fully-end-to-end TTS frameworks. The speech module consists of two parts: the speech encoder and the speech decoder. The speech encoder takes linear spectrograms as input and outputs the intermediate speech embeddings. It is composed of non-causal WaveNet [19] residual blocks, as in [16, 20, 21], and a projection layer that outputs the mean and the variance of the distribution of the speech embeddings. We adopt HiFi-GAN V1 [22] as the speech decoder, as in [16].

### *2.1.3. Connector*

The connector consists of two parts: the prenet and the flow. As in [16], the prenet consists of the transformer encoder [23] and a linear projection layer. It takes the intermediate EEG embeddings as input and outputs mean and variance of the distribution. The flow is identical to the normalizing flow network in [16], with a stack of affine coupling layers [24]. The flow works as an invertible function between the two distributions of the EEG and speech embeddings. There is a gradient stop between the intermediate EEG embeddings and the connector. We empirically observed that the gradient stop stabilizes the training of the EEG module, as it is not affected by the performance of the speech module in early stages of training.

## 2.2. Training Objectives

Employing the approach in [15], the cosine similarity loss is used for training the EEG module, as in Eq. (1):

$$L_{\text{EEG}}(x, \hat{x}) = 1 - \frac{1}{N_{\text{ch}}} \sum_{i=1}^{N_{\text{ch}}} \frac{x_i^T \cdot \hat{x_i}}{\|x_i\| \|\hat{x_i}\|} \tag{1}$$

where $N_{\text{ch}}$ is the number of EEG channels, and $x$ and $\hat{x}$ represent the input and the reconstructed EEG signals, respectively.

We adopt training objectives for the speech module from [16, 22]. The reconstruction loss $L_{\text{mel}}$ is defined as the L1 loss between the mel-spectrograms of generated and actual speech waveform. The KL-divergence loss $L_{\text{KL}}$ helps map the distributions of the intermediate embeddings of speech and EEG to one another, as in Eq. (2).

$$L_{\text{KL}}(p, q) = \log q(z|y) - \log p(z|x) \tag{2}$$

where $y$ is the input speech and $z$ is the intermediate speech embedding. Hence, $q(z|y)$ and $p(z|x)$ represent the distribution of the intermediate speech embeddings given $y$ (speech) and $x$ (EEG signals), respectively. The total loss for the speech module $L_{\text{speech}}$ is as follows:

$$L_{\text{speech}} = L_{\text{mel}} + L_{\text{KL}} + L_{\text{GAN}} \tag{3}$$

where $L_{\text{GAN}}$ consists of the adversarial loss and the feature matching loss from HiFi-GAN V1 [22]. Note that gradient stop is applied between the EEG module and the connector, the training of the EEG module is not affected by the speech module.

# 3. Experiments

## 3.1. Dataset

We conducted experiments on the N400 dataset [25], which contains EEG signals that were recorded from 24 subjects, while each subject was listening to 440 sentences in English. The speech samples are around 2 - 3 seconds long and contain 5 to 8 words. The 128 channel EEG was recorded at a sampling rate of $512$ Hz, while the participants listened to the sentences. Four subjects[1] were excluded owing to unreliable data, as suggested by [25]. The test set consists of two subjects[2] and 40 sentences. As a result, the train set contains 7,200 pairs[3] and the remaining 1,600 pairs were selected for the test set, where 720 and 800 of them are unseen audio and subject, respectively, and 80 are unseen in both audio and subject.

The pre-processing pipeline for EEG is as follows: First, a notch filter at $60$ Hz was employed in order to remove the powerline noise. Then, bandpass filtering with frequency limits of low $0.5$ Hz and high $50$ Hz was applied in order to preserve the bands relevant to EEG spectral information, followed by eye blink removal using independent component analysis (ICA). The signals were then resampled to $256$ Hz.

All of the speech samples were down-sampled to $22,050$ Hz, to match the sampling rate of pre-trained speech models. For linear spectrograms, the following short-time Fourier transform (STFT) parameters were used: $1024$ for both FFT and window size, and $256$ for the hop size. For mel-spectrograms, the same parameters were utilized with $80$ mel-bands.

## 3.2. Experimental Setup

Five different training configurations were considered as below:

- VANILA: all of the modules were trained from scratch, without any pre-trained parameters.
- PT-AUDIO: only the speech module was initialized with pre-trained parameters for training.
- PT-AUDIO-FZ: the pre-trained parameters were used for the speech module, but frozen, that is, the speech module was not trained.
- PT-AUDIO-EEG: both of the EEG module and the speech module were initialized with pre-trained parameters.
- PT-AUDIO-EEG-FZ: the EEG module and the speech module adopted the pre-trained parameters, but they were not trained.

For the pre-trained speech module, `pretrained_ljs.pth`[4] was adopted. In PT-AUDIO-EEG and PT-AUDIO-EEG-FZ, the EEG module, that had been pre-trained for 300 epochs, was later frozen when combined with the speech module. The compared baseline model is VLAAI [7], which consists of a stack of convolution layers with skip connections. The final layer of the baseline model was modified to generate 80 band mel-spectrograms. One Nvidia A40 GPU was utilized for each training configuration. All of the FESDE training configurations were trained for 100k iterations with the AdamW optimizer [26]. The baseline model was trained for 68 epochs.

[1] subject # 5, 10, 15, and 18.
[2] subject # 23 and 24.
[3] 18 subjects with 400 sentences each.
[4] `https://github.com/jaywalnut310/vits`

# 4. Results and Discussion

## 4.1. Evaluation

The performance was measured in the following two objective metrics: mel-cepstral distortion (MCD) [27] and mel-spectrogram correlation (Mel-Corr). MCD measures the distance between two mel-cepstral coefficients as Eq (4):

$$\text{MCD} = \alpha \sqrt{\sum_{i=1}^{N_{\text{MCC}}} (\text{MCC}_i - \widehat{\text{MCC}_i})^2} \tag{4}$$

where $\alpha = \frac{10\sqrt{2}}{\ln 10}$ and $N_{\text{MCC}}$ is the number of mel-cepstral coefficients. Mel-Corr is calculated as the Pearson correlation coefficient between two mel-spectrograms. For convenience, the reported numbers are multiplied with 100. Lower MCD and higher Mel-Corr values indicate better performance.

Table 1 shows the MCD and Mel-Corr values for the baseline model and the proposed five models. From the results in the table, the best performance was achieved when both of the EEG and speech modules are pre-trained separately (i.e., PT-AUDIO-EEG-FZ). Also, almost all of the proposed models perform better than the baseline model.

## 4.2. Phoneme Analysis

In order to analyze the effectiveness of proposed models in decoding speech at the level of phoneme, a fine-grained analysis was carried out by calculating some objective metrics. This analysis will reveal which of the category of phonemes are difficult or easy to decode. The Montreal forced aligner (MFA) ver3 [28] was used to acquire the timestamps of each phoneme. The pre-trained `english_us_arpa` configuration was used for both acoustic model and the dictionary. As MFA does not perform accurately for reconstructed speech, the timestamps acquired from the ground-truth speech are assumed to be identical to the reconstructed speech.

The consonants are grouped according to their manner of articulation, place of articulation, and tenseness of articulation. Similarly, the vowels are grouped based on position of the tongue and tenseness. As illustrated in Figures 2, the consonants that are nasal, dental, or lenis tend to be relatively easily decoded. Also, an interesting tendency is observed wherein more closed vowels are easier to decode.

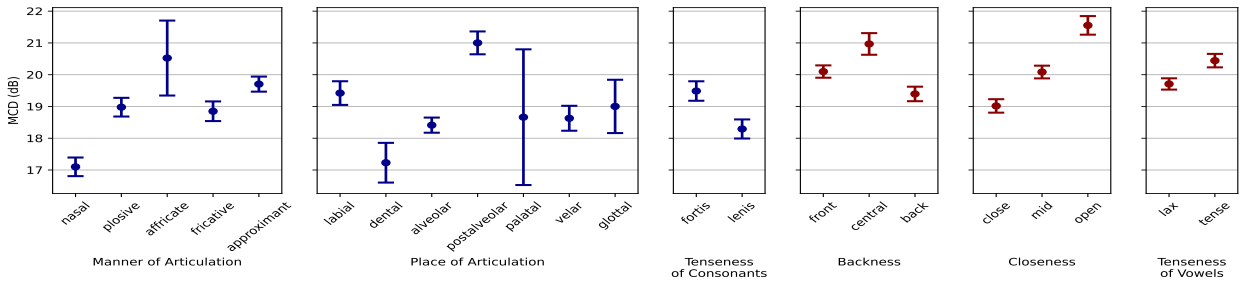
# 5. Ablation Study

## 5.1. EEG Channels

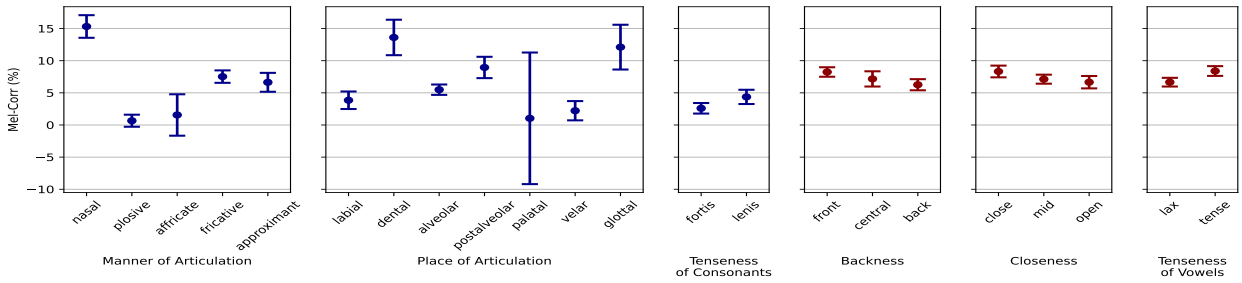
As recent studies [29] suggest, we explore the utilization of only the parietal and temporal regions of EEG that are known to contain rich auditory information. Instead of an increase in performance, a drop in performance was observed, as shown in Table 2, especially when the speech module was frozen during training. This result may suggest that non-auditory parts of the

Table 1: *MCD (dB) and Mel-Corr (%) of each configuration, with 95% confidence intervals.*

| Model | | MCD (dB) ↓ unseen audio | MCD (dB) ↓ unseen subject | MCD (dB) ↓ unseen both | Mel-Corr (%) ↑ unseen audio | Mel-Corr (%) ↑ unseen subject | Mel-Corr (%) ↑ unseen both |
|---|---|---|---|---|---|---|---|
| BASELINE [7] | | $13.00 \pm 0.12$ | $13.20 \pm 0.12$ | $12.82 \pm 0.39$ | $13.27 \pm 0.66$ | $12.66 \pm 0.74$ | $13.07 \pm 2.06$ |
| OURS | VANILLA | $11.67 \pm 0.12$ | $11.72 \pm 0.11$ | $11.56 \pm 0.36$ | $13.75 \pm 0.75$ | $12.76 \pm 0.68$ | $13.74 \pm 2.15$ |
| | PT-AUDIO | $11.58 \pm 0.12$ | $11.61 \pm 0.11$ | $11.68 \pm 0.34$ | $14.34 \pm 0.68$ | $13.84 \pm 0.68$ | $13.31 \pm 2.25$ |
| | PT-AUDIO-FZ | $11.71 \pm 0.11$ | $11.71 \pm 0.10$ | $11.57 \pm 0.31$ | $12.80 \pm 0.71$ | $12.97 \pm 0.68$ | $11.11 \pm 2.47$ |
| | PT-AUDIO-EEG | $11.67 \pm 0.12$ | $11.64 \pm 0.11$ | $11.69 \pm 0.34$ | $\mathbf{14.71 \pm 0.68}$ | $14.42 \pm 0.70$ | $\mathbf{14.46 \pm 2.42}$ |
| | PT-AUDIO-EEG-FZ | $\mathbf{11.45 \pm 0.10}$ | $\mathbf{11.46 \pm 0.10}$ | $\mathbf{11.43 \pm 0.32}$ | $14.56 \pm 0.70$ | $\mathbf{14.50 \pm 0.69}$ | $14.04 \pm 2.06$ |




(a) *MCD (dB) of each consonant (blue) and vowel (red) group.*




(b) *Mel-Corr (%) of each consonant (blue) and vowel (red) group.*

Figure 2: *MCD (dB) and Mel-Corr (%) of each phoneme group. The lower MCD and higher Mel-Corr indicate better performance. The consonants (blue) are clustered by three criteria: manner, place, and tenseness of articulation. The vowels (red) are clustered by its tongue position and tenseness.*

Table 2: *Performance drop (%) when only the parietal and temporal regions of EEG signals are used.*

| Model | Performance Drop (%) MCD | Performance Drop (%) Mel-Corr |
|---|---|---|
| PT-AUDIO-EEG | $-0.13$ | $-5.99$ |
| PT-AUDIO-EEG-FZ | $-1.44$ | $-14.64$ |

brain may be involved in speech decoding, however, further investigation is necessary for a clearer explanation.

### 5.2. Text-Spotting

Apart from evaluating the performance of speech reconstruction from EEG, we investigate whether our proposed EEG encoder captures textual information. A binarized word-spotting experiment was conducted to validate whether our proposed EEG presentation can be used to identify certain words in a speech sample. The 30 most frequently occurring nouns among all sentences are chosen as the set of keywords, and the task is to detect whether any of these keywords exist in a speech. The pre-trained EEG module was utilized. The results across different test conditions are presented in Table 3. We demonstrate the feasibility of text detection from EEG signals, and it remains a challenging and open endeavor.

Table 3: *Performance of predicting binarized word-spotting. The performance score is measured in the unweighted F1 score.*

| unseen audio | unseen subject | unseen both |
|---|---|---|
| 60.12 | 56.74 | 57.78 |

## 6. Conclusion

In this work, we introduce FESDE, a framework to decode listened speech waveforms directly from EEG signals. The proposed approach is faster and simpler by enabling single-step inference and it also outperforms the baseline model that involves an intermediate conversion to text representation. We also explore the characteristics of phonemes that are easily or challenging to be decoded by the proposed framework. Our approach is currently limited to listened speech, however, in the foreseeable future, we plan to extend our research scope to speech production tasks, such as imagined or phonated speech decoding.

## 7. Acknowledgements

This project was partially supported by a USC Annenberg Graduate Fellowship and also by the Defense Advanced Research Projects Agency (DARPA) under cooperative agreement No. N660012324006. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

## 8. References

[1] S. L. Metzger, K. T. Littlejohn, A. B. Silva, D. A. Moses, M. P. Seaton, R. Wang, M. E. Dougherty, J. R. Liu, P. Wu, M. A. Berger *et al.*, "A high-performance neuroprosthesis for speech decoding and avatar control," *Nature*, vol. 620, no. 7976, pp. 1037–1046, 2023.

[2] C. Puffay, B. Accou, L. Bollens, M. J. Monesi, J. Vanthornhout, T. Francart *et al.*, "Relating eeg to continuous speech using deep neural networks: a review." *Journal of Neural Engineering*, 2023.

[3] D. Lopez-Bernal, D. Balderas, P. Ponce, and A. Molina, "A state-of-the-art review of eeg-based imagined speech decoding," *Frontiers in Human Neuroscience*, vol. 16, p. 867281, 2022.

[4] F. R. Willett, E. M. Kunz, C. Fan, D. T. Avansino, G. H. Wilson, E. Y. Choi, F. Kamdar, M. F. Glasser, L. R. Hochberg, S. Druckmann *et al.*, "A high-performance speech neuroprosthesis," *Nature*, vol. 620, no. 7976, pp. 1031–1036, 2023.

[5] J. Li, C. Guo, L. Fu, L. Fan, E. F. Chang, and Y. Li, "Neural2speech: A transfer learning framework for neural-driven speech reconstruction," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 2200–2204.

[6] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, and J.-R. King, "Decoding speech perception from non-invasive brain recordings," *Nature Machine Intelligence*, vol. 5, no. 10, pp. 1097–1107, 2023.

[7] B. Accou, J. Vanthornhout, H. V. Hamme, and T. Francart, "Decoding of the speech envelope from eeg using the vlaai deep neural network," *Scientific Reports*, vol. 13, no. 1, p. 812, 2023.

[8] X. Xu, B. Wang, Y. Yan, H. Zhu, Z. Zhang, X. Wu, and J. Chen, "Convconcatnet: a deep convolutional neural network to reconstruct mel spectrogram from the eeg," *arXiv preprint arXiv:2401.04965*, 2024.

[9] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz, "Speech synthesis from ecog using densely connected 3d convolutional neural networks," *Journal of Neural Engineering*, vol. 16, no. 3, p. 036019, apr 2019. [Online]. Available: https://dx.doi.org/10.1088/1741-2552/ab0c59

[10] Y.-E. Lee, S.-H. Lee, S.-H. Kim, and S.-W. Lee, "Towards voice reconstruction from eeg during imagined speech," in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. [Online]. Available: https://doi.org/10.1609/aaai.v37i5.25745

[11] T. Proix, J. Delgado Saa, A. Christen, S. Martin, B. N. Pasley, R. T. Knight, X. Tian, D. Poeppel, W. K. Doyle, O. Devinsky *et al.*, "Imagined speech can be decoded from low-and cross-frequency intracranial eeg features," *Nature communications*, vol. 13, no. 1, p. 48, 2022.

[12] J. A. Urigüen and B. Garcia-Zapirain, "EEG artifact removal: state-of-the-art and guidelines," *Journal of neural engineering*, vol. 12, no. 3, p. 031001, 2015.

[13] C. M. Michel and D. Brunet, "Eeg source imaging: a practical review of the analysis steps," *Frontiers in neurology*, vol. 10, p. 446653, 2019.

[14] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data," *Frontiers in Human Neuroscience*, vol. 15, p. 653659, 2021.

[15] A. Kommineni, K. Avramidis, R. Leahy, and S. Narayanan, "Knowledge-guided eeg representation learning," in *International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2024.

[16] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[17] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[18] A. Gu, K. Goel, and C. Re, "Efficiently modeling long sequences with structured state spaces," in *Proc. ICLR*, 2021.

[19] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," in *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016, p. 125.

[20] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[21] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[22] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[24] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=HkpbnH9lx

[25] K. K. Toffolo, E. G. Freedman, and J. J. Foxe, "Evoking the n400 event-related potential (erp) component using a publicly available novel set of sentences with semantically incongruent or congruent eggplants (endings)," *Neuroscience*, vol. 501, pp. 143–158, 2022.

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[27] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, 1993, pp. 125–128 vol.1.

[28] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[29] R. A. Sharon, S. Narayanan, M. Sur, and H. A. Murthy, "An empirical study of speech processing in the brain by analyzing the temporal syllable structure in speech-input induced eeg," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 4090–4094.