Context-FFT: A Context Feed Forward Transformer Network for EEG-based Speech Envelope Decoding

Ximin Chen*‡, Yuting Ding*, Nan Yan†, Changsheng Chen‡, Fei Chen*
* Southern University of Science and Technology, Shenzhen, China
E-mail: 2020281106@email.szu.edu.cn, 12332166@mail.sustech.edu.cn, fchen@sustech.edu.cn
† Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
E-mail: nan.yan@siat.ac.cn
‡ Shenzhen University, Shenzhen, China
E-mail: cschen@szu.edu.cn

***Abstract*—Decoding speech envelope from electroencephalography (EEG) signals has been demonstrated to be useful for assessing speech intelligibility and boosting potential applications in neuroscience research as well as clinical diagnosis, which is also the focus of the ICASSP Auditory EEG 2023 Challenge. In order to further improve speech envelope decoding performance, this study proposes an end-to-end architecture based on multi-head attention mechanism called Context-FFT. Besides the transformer architecture, we also utilize a context layer to extract information and refine outputs based on given inputs. Notably, we decompose raw speech envelopes into envelopes in 12 frequency bands to model the relationship between 64-channel EEG signals and speech envelopes more precisely. Experiment results show that the proposed model achieves an average Pearson correlation value of 0.2148±0.1004 on held-out stories, outperforming the linear baseline by 51.65% and the VLAAI baseline by 23.17%, and 0.0701±0.0428 on held-out subjects. In terms of the final metric defined by the challenge, we obtain a final score of 0.1639, outperforming all submitted models of the ICASSP Auditory EEG 2023 Challenge. In the end, we explore the contributions of different brain regions to the process of continuous speech.**

I. Introduction

Electroencephalography (EEG) is a non-invasive neuroimaging technology that captures potential differences generated by electrical activities of the brain. Neurons exchange information between each other through synaptic connections in the form of electrical currents. When evoked, postsynaptic potentials generated by synchronized activities of a large number of neurons summate in the cortex and extend to the surface of the scalp, where they are recorded as EEG signals using a certain number of electrodes [1]. Due to its non-invasive and safe characteristics, EEG is widely used in fundamental research as well as clinical diagnosis [2]. EEG also has potential applications in smart hearing aids as researchers have found out that auditory attention can be decoded from EEG, which can improve outcomes of EEG-based smart hearing aids by determining the focus of a user's attention especially in noisy conditions [3].

To help understand how the brain processes continuous speech, recent studies have focused on neural tracking of speech features in EEG [4–7], which often refers to the time-locking effects of EEG to speech resources in a single-speaker scenario. One common paradigm to quantify the neural tracking of EEG is to reconstruct speech envelopes from EEG and calculate the similarity between the reconstructed envelope and the original envelope as the evaluation metric, which is also the focus of the ICASSP Auditory EEG 2023 Challenge [8] that calls for building the best model to relate speech to EEG. Linear models are commonly applied to reconstruct speech envelopes, but their reconstruction scores are low, ranging from 0.1 to 0.2 for subject-specific linear decoders [6, 9–12]. The limitation of linear models lies in assuming a linear relationship between speech and highly non-linear EEG [1].

Inspired by the success of deep learning architectures [13] in processing complex data such as text and audio signals [2], an increasing number of studies have been using deep learning models to relate speech to EEG. Accou et al. have proposed VLAAI, short for the Very Large Augmented Auditory Inference network, which stacks multiple convolutional blocks to enhance the model's non-linearity [6]. VLAAI has set the state-of-art performance by yielding an increase over the linear model by 52% and been chosen as the baseline of the challenge [8]. Notably, VLAAI utilizes an output context layer to take the output context into account, which contributes a 10% relative increase to the model performance, indicating that the context layer is capable to extract useful information from previous outputs. Transformer based on the multi-head attention mechanism is also widely used due to its significant success in fields such as natural language processing [14]. For instance, Yang et al. have proposed FastSpeech based on feed-forward Transformer (FFT) architectures to generate speech from text fast and robustly [15]. Drawing on its strong capability for fitting speech features, Piao et al. have proposed the HappyQuokka model consisting of FFT blocks and won the first place in the challenge [16], which indicates that incorporating the multi-head attention mechanism can effectively capture the dynamic variation characteristics of EEG signals. This motivates the present work to utilize the FFT architecture as well as the context layer to learn and fuse useful feature representations.

Besides model architectures, recent studies have also in-

corporated inherent characteristics of the auditory system to help model the complex relationship between EEG signals and speech stimuli. Thornton et al. have combined two decoders which exploit different EEG responses to speech: slow neural tracking of the speech envelope and high-frequency speech-related frequency following responses, achieving a significant improvement over the linear baseline [17]. Wu et al. have decomposed speech signals into envelopes in 12 frequency bands for more intelligible speech reconstruction [18]. This motivates the present work to utilize multi-band envelope information to guide the model towards more accurate envelope decoding.

To sum up, the aim of this work was to improve the decoding accuracy of reconstructing the speech envelope from the 64-channel EEG signal. Specifically, we propose a context feed forward Transformer network called Context-FFT, which combines multi-head attention mechanism and the context layer. Moreover, the raw speech is decomposed into envelopes of 12 frequency bands for model optimization. For model evaluation, we employ the SparrKULee dataset [19] as designated by the ICASSP Auditory 2023 Challenge [8].

II. METHODS

The proposed architecture, Context-FFT, is modified based on the system proposed by [16] as shown in Fig. 1. The input EEG signals are first fed into a convolutional layer for input embedding, then are passed through the modified FFT blocks to model the relationship between given EEG signals and speech features using the multi-head attention mechanism. The attention-weighted features are further refined through a context layer. Finally, the outputs of stacked blocks are fused by a linear layer to summarize extracted information into a single speech envelope. Key components and modifications are illustrated as follows.

A. Multi-head Attention Module

The multi-head attention module is based on Transformer's self attention mechanism [14], which can attend to specific tokens according to their attention scores. The normalization layer is placed inside the residual blocks instead of putting it between the residual block to help the model converge faster [20].

B. Context Layer

WaveNet [21] used a context module to expand the receptive field of the model. Similarly, VLAAI [6] employed a context layer to refine outputs by taking existing outputs into account. Inspired by these works, a context layer is introduced in the pre-LN FFT block, referred as the modified FFT block. Inside the context layer, attention-weighted features are passed through a convolutional layer with a kernel size of 9 and padding of 4, which is then followed by a nonlinear transformation using the ReLU activation function and a 1D convolutional layer. Finally, resulting features are normalized through layer normalization and adjusted using residual connections to obtain the refined features.

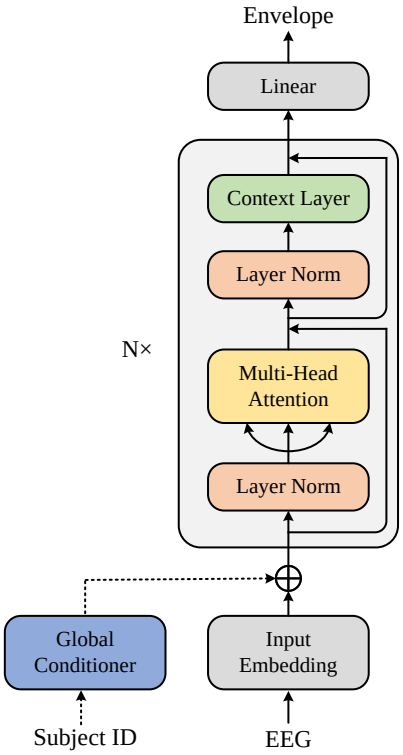

Fig. 1. Architecture of the proposed Context Feed Forward Transformer network, Context-FFT. Dash lines through the global conditioner block indicate that the block is only used in held-out stories test.

C. Auxiliary Global Conditioner

According to [16], the auxiliary global conditioner is only for within-subject generation, which provides the model with additional information of seen subjects. Specifically, it encodes a subject's identity into an embedding matching the dimension of the input embedding. This embedding is then added to the input embedding and fed into the modified FFT blocks.

III. EXPERIMENTS

For comparative studies, we perform experiments in SparrKULee dataset [19]. Details about the dataset and preprocessing procedure are provided as follows.

A. Dataset

The SparrKULee dataset [19] contains 85 participants with normal hearing. EEG data were recorded from 64 channels using a BioSemi ActiveTwo system at a sampling rate of 8192 Hz. All the stimuli are single-speaker stories spoken in Flemish by a native Flemish speaker, stored at a sampling rate of 48 kHz. Each subject listened to between 8 and 10 trials, each of approximately 15 minutes in length, totaling 168 hours of EEG data and 157 hours of stimuli.

For all our experiments, we use the preprocessed dataset provided by the challenge [8], which has already been normalized and split into train, validation and test set in a 1:1:1 ratio. Specifically, the training set includes EEG data from 71

subjects, numbered from sub-01 to sub-71. The test set consists of held-out stories (test set 1) and held-out subjects (test set 2). Test Set 1 includes EEG data from 71 subjects seen in training, while the corresponding stimuli are never used in the training set, thus referred as held-out stories, amounting to a total of 944 minutes. Test Set 2 includes EEG data from the remaining 14 subjects, numbered from sub-72 to sub-85 who are not in the training set, thus referred as held-out subjects, amounting to a total of 1260 minutes.

B. Data Preprocessing

The preprocessing procedure follows the setting of [8, 19]. For EEG data, firstly raw EEG recordings are downsampled from 8192 Hz to 1024 Hz and a multichannel Wiener filter is applied to remove eye blink artifacts. Then, the EEG signal is re-referenced to a common average. Finally, the EEG signal is downsampled to 64 Hz. As for stimuli, a sixth-order Butterworth filter is applied to band-pass filter raw speech signals to obtain 12 frequency band signals with center frequencies between 80 and 3000 Hz following the setting in [9, 18]. Next, each frequency band signal is filtered using a GammaTone filter bank with 28 subbands at equivalent bandwidth intervals with center frequencies ranging from 50 Hz to 5 kHz. Then, the absolute value of each sample in the filter is raised to the power of 0.6 and averaged to obtain a single speech envelope. Finally, the obtained envelope is downsampled to 64 Hz. During the training phase, we use the 12 subband estimated envelopes and the original signal's 12 subband envelopes as inputs of the loss function for optimization.

To improve computational efficiency, EEG signals and speech features are randomly cropped into segments of 5 seconds in the training phase. During testing, EEG signals are split into 5-second segments and are fed into the model, whose outputs are then concatenated together to make the entire envelope.

C. Channel Groups of Cortical Regions

To investigate the impact of different cortical regions on processing continuous speech, the cortical regions are divided into eight channel groups (CG) following the setting from [22], as shown in Table I. Broca's and Wernicke's areas are considered to be related to auditory production [23], and speech stimuli may also activate the auditory cortex. Moreover, the motor cortex, prefrontal cortex, and sensory cortex regions incorporate together in language production and processing [22].

D. Network Configuration

This work uses the TensorFlow framework to implement the linear decoder and VLAAI models for comparison, and the proposed model is implemented using the PyTorch framework. The number of modified FFT blocks is set to 8 and we use 2 heads for multi-head attention. The negative Pearson correlation is used as the loss function as required by the challenge [8]. We use the Adam optimizer with an initial learning rate of 0.0005 and a StepLR scheduler with a learning rate decay factor of 0.9.

IV. RESULTS AND DISCUSSION

In this section, we demonstrate the performance of Context-FFT on the SparrKULee dataset [19]. It is then compared with linear decoder [9] and VLAAI [6], which are both baselines provided by the challenge [8], as well as the HappyQuokka System [16], which placed the first in the challenge. Subsequently, an ablation study is conducted to evaluate effects of key components on model performance. In the end, we present the results of decoding performance of different brain regions and discuss its potential applications in smart hearing aids.

A. Model Performance

Table II shows the results of different models on the test set. Pearson correlation (Pearson *r*) is employed as evaluation metric between estimated envelopes and actual envelopes. Reconstruction scores of test set 1 (S_1) and test set 2 (S_2) are averaged across stimuli and subjects. Final score is a weighted sum of the average Pearson *r* of both test subsets defined by the challenge [8], computed as

$$\text{Final Score} = \frac{2}{3}S_1 + \frac{1}{3}S_2. \tag{1}$$

From Table II, we observe that our proposed model Context-FFT outperforms both baseline models and the state-of-art HappyQuokka model. Specifically, Context-FFT achieves an average Pearson *r* of 0.2180 ± 0.1004 for held-out stories and 0.0701 ± 0.0428 for held-out subjects, achieving the highest final score of 0.1639. Notably, our model has significantly improved in held-out stories by 51.65% and 23.17% compared with the linear decoder and VLAAI respectively. As for held-out subjects, the decoding performance slightly declines compared with other three models, possibly because the model's ability to learn unique characteristics of seen subjects has improved due to feature enhancement and context layer, however, feature patterns learned from seen subjects can't fit well to unseen subjects, thus leading to a decrease in decoding performance.

B. Ablation Study

To gain insight into what parts of the model are responsible for improvements on decoding performance, we have conducted ablation studies on the basis of the HappyQuokka model [16], which is referred as the benchmark model. Since our proposed model Context-FFT has outperformed baseline models by a large margin on held-out stories, while results of ablation studies on held-out subjects have minimal changes and are all below 0.1, we only present ablation studies on held-out stories. Specifically, key modules are added to the original stacked FFT block in the following steps:

1. Benchmark+FB: decomposing speech signals into 12 subbands and extracting their envelopes respectively on the basis of the benchmark model. FB is short for Filter Band.
2. Context-FFT: adding a context layer to the FFT blocks of the model in step 1.

Figure 2 shows the decoding performance of baseline models and each model variant in held-out stories. First of all,

TABLE I
CHANNEL GROUPS FOR CORTICAL REGIONS

Channel Groups	Cortical Regions	Channels
CG1	Broca's and Wernicke's areas	AF3, F3, F5, FC3, FC5, T7, C5, TP7, CP5, P5
CG2	Auditory cortex	FT7, FT8, T7, T8, TP7, TP8
CG3	Motor cortex	FZ, F1, F2, F3, F4, FC1, FC2, FC3, FC4, CZ
CG4	Prefrontal cortex	FPZ, FP1, FP2, AF3, AF4, F5, F6, F7, F8
CG5	Sensory cortex	CZ, C1, C2, C3, C4, CPZ, CP1, CP2, CP3, CP4
CG6	Left brain	FP1, AF3, F1, F3, F5, F7, FC1, FC3, FC5, FT7, C1, C3, C5, T7, CP1, CP3, CP5, TP7, P1, P3, P5, P7, PO3, PO5, PO7, CB1, O1
CG7	Right brain	FP2, AF4, F2, F4, F6, F8, FC2, FC4, FC6, FT8, C2, C4, C6, T8, CP2, CP4, CP6, TP8, P2, P4, P6, P8, PO4, PO6, PO8, CB2, O2
CG8	Whole brain	All 64 channels

TABLE II
PERFORMANCE OF MODELS ON TEST SETS OF SPARRKULEE DATASET

Model	Test Set 1	Test Set 2	Final Score
Linear Decoder [9]	0.1054±0.0538	0.0960±0.0387	0.1023
VLAAI [6]	0.1675±0.0732	**0.1139±0.0410**	0.1496
HappyQuokka [16]	0.1895±0.0869	0.0976±0.0444	0.1589
Context-FFT	**0.2180±0.1004**	0.0701±0.0428	**0.1639**

the median Pearson correlation values of all model variants have gradually increased. The decoding performance of Benchmark+FB has improved by 9.3% compared to the benchmark model, suggesting that the model can better find the correspondence between EEG signals and speech stimuli from the envelopes of different sub-bands than from the entire envelope. It also demonstrates that combining inherent characteristics of auditory-EEG processing can fit the relationship between EEG signals and speech stimuli more precisely, further enhancing the model's feature fusion capability. Furthermore, the performance of Context-FFT has also significantly improved by 13.1% compared with the benchmark model, indicating that expanding the model's receptive field can enhance its ability to extract useful context information.

C. Contributions of Cortical Regions

Figure 3 shows the comparison of decoding performance of different cortical regions in decoding speech envelopes from EEG signals.

First of all, we can observe that the Pearson r of CG8 is the highest among all groups, indicating the entire brain has involved in processing speech stimuli. Among the remaining groups, CG6 has the highest reconstruction score, followed by CG1 and CG2. Specifically, CG1 corresponds to Broca's

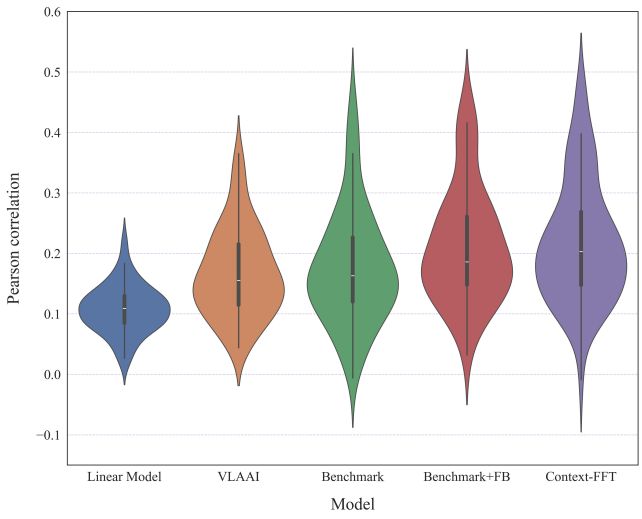

Fig. 2. Results of the ablation studies. Each point in the violin plot represents the average Pearson correlation for a subject averaged across stimuli.

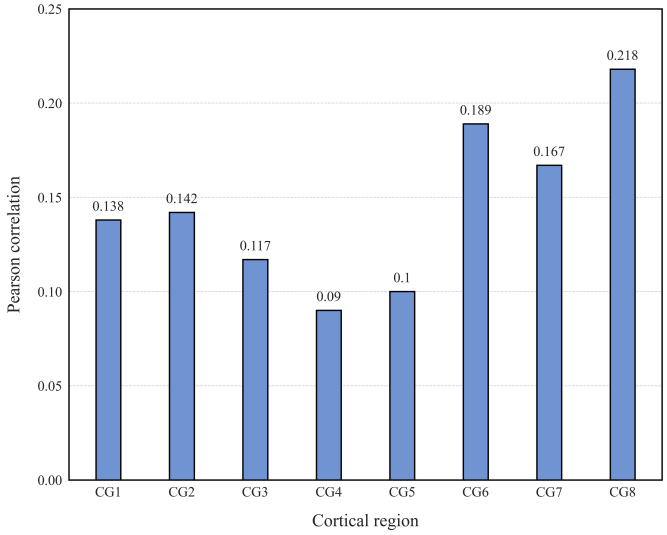

Fig. 3. Decoding performance of different cortical regions.

and Wernicke's areas which are located in the left brain region (CG6). Reconstruction scores of CG1 and CG6 are relatively close, suggesting that Broca's and Wernicke's areas play crucial roles in auditory processing of the left brain, which is consistent with conclusions obtained in [22, 23].

Moreover, we can observe from results of CG6 and CG7 that the involvement of the left brain is higher than that of the right brain in processing speech stimuli. Results of CG3, CG4, and CG5 are the lowest among all groups, indicating that when the brain processes speech stimuli, the corresponding motor cortex, prefrontal cortex and sensory cortex are also activated, but their contributions are relatively small compared to the auditory cortex and Broca's and Wernicke's areas.

Although CG2 covers only 6 channels, its Pearson r is relatively close to scores of CG1 and CG6, which cover 10 channels and 27 channels separately. Therefore, the above results have demonstrated that using fewer channels can achieve similar performance obtained by using full-channel EEG signals, which is essential for EEG-based smart hearing aids that require low computational costs and short computation times, allowing users to receive real-time feedback from the devices.

V. CONCLUSIONS

This work proposes a new model called Context-FFT for decoding speech envelopes from EEG signals and demonstrates that the performance of this model surpasses the SOTA model of the ICASSP Auditory 2023 Challenge. On average, Context-FFT achieves a Pearson r of 0.2180 on the held-out-stories and 0.0701 on the held-out subjects. Furthermore, through ablation experiments, it is determined that utilizing the context layer and incorporating the inherent characteristics of EEG signals play a crucial role in improving the model performance. Although our model achieves better results on seen subjects, the generalization performance on unseen subjects is not significantly improved, indicating that improving the generalization ability of the model remains challenging. Additionally, the decoding performance of the Broca's and Wernicke's areas containing 10 channels and the auditory cortex containing 6 channels are close to the performance obtained using 64-channel EEG signals, providing a theoretical basis for achieving a balance between computation speed and decoding performance in portable EEG-based cochlear implants and other hearing devices.

ACKNOWLEDGMENTS

This work was supported by Guangdong Basic and Applied Basic Research Foundation (Grant No. 2022B1515120056), and the Basic Research Foundation of Shenzhen (Grant No. JCYJ20220818101217037). Part of this study was the basis for the Bachelor's thesis of the first author (X.M.C.).

REFERENCES

[1] D. P. Subha, P. K. Joseph, R. Acharya U, and C. M. Lim, "EEG signal analysis: A survey," *Journal of Medical Systems*, vol. 34, no. 2, pp. 195–212, 2010.

[2] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of Neural Engineering*, vol. 16, no. 5, p. 051001, 2019.

[3] M. Thornton, D. Mandic, and T. Reichenbach, "Robust decoding of the speech envelope from EEG recordings through deep neural networks," *Journal of Neural Engineering*, vol. 19, no. 4, p. 046007, 2022.

[4] S. J. Aiken and T. W. Picton, "Human cortical responses to the speech envelope," *Ear and Hearing*, vol. 29, no. 2, p. 139, 2008.

[5] D. A. Abrams, T. Nicol, S. Zecker, and N. Kraus, "Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech," *Journal of Neuroscience*, vol. 28, no. 15, pp. 3958–3965, 2008.

[6] 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.

[7] C. Puffay, B. Accou, L. Bollens, M. J. Monesi, J. Vanthornhout, H. V. hamme, and T. Francart, "Relating EEG to continuous speech using deep neural networks: A review," *Journal of Neural Engineering*, vol. 20, no. 4, p. 041003, 2023.

[8] M. J. Monesi, L. Bollens, B. Accou, J. Vanthornhout, H. Van Hamme, and T. Francart, "Auditory EEG decoding challenge for icassp 2023," *IEEE Open Journal of Signal Processing*, 2024.

[9] E. C. Lalor and J. J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *European Journal of Neuroscience*, vol. 31, no. 1, pp. 189–193, 2010.

[10] J. Vanthornhout, L. Decruy, J. Wouters, J. Z. Simon, and T. Francart, "Speech intelligibility predicted from neural entrainment of the speech envelope," *Journal of the Association for Research in Otolaryngology*, vol. 19, pp. 181–191, 2018.

[11] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biology: CB*, vol. 25, no. 19, pp. 2457–2465, 2015.

[12] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mtrf) toolbox: A matlab toolbox for relating neural signals to continuous stimuli," *Frontiers in Human Neuroscience*, vol. 10, 2016.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[16] Z. Piao, M. Kim, H. Yoon, and H.-G. Kang, "Happyquokka system for icassp 2023 auditory EEG challenge," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–2, IEEE, 2023.

[17] M. Thornton, D. Mandic, and T. Reichenbach, "Relating EEG recordings to speech using envelope tracking and the speech-FFR," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–2, IEEE, 2023.

[18] H. Wu and F. Chen, "A temporal envelope-based speech reconstruction approach with EEG signals during speech imagery," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 894–899, 2020.

[19] B. Accou, L. Bollens, M. Gillis, W. Verheijen, H. Van hamme, and T. Francart, "Sparrkulee: A speech-evoked auditory response repository of the ku leuven, containing EEG of 85 participants," *bioRxiv*, pp. 2023–07, 2023.

[20] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T.-Y. Liu, "On layer normalization in the transformer architecture," in *International Conference on Machine Learning*, pp. 10524–10533, PMLR, 2020.

[21] A. v. d. 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," *arXiv preprint arXiv:1609.03499*, 2016.

[22] M. Li, S. H. Pun, and F. Chen, "Impacts of cortical regions on EEG-based classification of lexical tones and vowels in spoken speech," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1–4, IEEE, 2023.

[23] M. N. I. Qureshi, B. Min, H.-j. Park, D. Cho, W. Choi, and B. Lee, "Multiclass classification of word imagination speech with hybrid connectivity features," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 10, pp. 2168–2177, 2018.