TFGA-NET: TEMPORAL-FREQUENCY GRAPH ATTENTION NETWORK FOR BRAIN-CONTROLLED SPEAKER EXTRACTION

Youhao Si, Yuan Liao, Qiushi Han, Yuhang Yang, Rui Dai and Liya Huang†

College of Electronic and Optical Engineering & College of Flexible Electronics (Future Technology),
Nanjing University of Posts and Telecommunications, Nanjing, China

ABSTRACT

The rapid development of auditory attention decoding (AAD) based on electroencephalography (EEG) signals offers the possibility EEG-driven target speaker extraction. However, how to effectively utilize the target-speaker common information between EEG and speech remains an unresolved problem. In this paper, we propose a model for brain-controlled speaker extraction, which utilizes the EEG recorded from the listener to extract the target speech. In order to effectively extract information from EEG signals, we derive multi-scale time–frequency features and further incorporate cortical topological structures that are selectively engaged during the task. Moreover, to effectively exploit the non-Euclidean structure of EEG signals and capture their global features, the graph convolutional networks and self-attention mechanism are used in the EEG encoder. In addition, to make full use of the fused EEG and speech feature and preserve global context and capture speech rhythm and prosody, we introduce MossFormer2 which combines MossFormer and RNN-Free Recurrent as separator. Experimental results on both the public Cocktail Party and KUL dataset in this paper show that our TFGA-Net model significantly outper-forms the state-of-the-art method in certain objective evaluation metrics. The source code is available at: https://github.com/LaoDa-X/TFGA-NET.

Index Terms— Speaker extraction, EEG signals, Multi-modal fusion, Cocktail party, Multi-talker environment

1. INTRODUCTION

Selective auditory attention enables listeners to focus on a single talker in multi-speaker environments such as a cocktail party [1], while actively suppressing competing sources. However, individuals with hearing impairment often struggle with this task. Although modern hearing aids integrate front-end algorithms such as noise reduction and speech enhancement [2], they still cannot infer whom the wearer actually intends to listen to. Consequently, endowing machines with human-level selective listening remains a fundamental challenge.

With the advent of deep learning, speech separation (SS) has made continuous progress, from early deep clustering to Conv-TasNet [3], DPRNN [4], and more recent SepFormer [5] and TF-GridNet [6]. Under the assumption that the number of talkers is known, these systems can decompose a mixture into multiple independent channels. Nevertheless, they must separate all potential speakers, resulting in high computational complexity, and the separated outputs are not aligned with the listener's attentional focus. Downstream modules such as attention detection are still required to identify the target stream, which further increases consumption.

To reduce redundant computation and concentrate on the listener's object of attention, speaker extraction (SE) has been proposed. It exploit reference cues, such as enrolled target speech, lip movements [7], or spatial orientation [8] to directly extracts the target speech from the mixture. Although this strategy performs well when reliable priors are available, the practicality of both acoustic and visual reference cues is limited. When reference cues are missing or inaccurate, the practicality of speaker extraction degrades substantially. This limitation motivates the search for alternative modalities that can more robustly reflect the listener's true focus of attention.

Recent studies have demonstrated a strong association between brain activity and the speech being attended [9]. Electroencephalography (EEG), a non-invasive and low-cost technique, allows researchers to decode auditory attention of listeners and identify the target speaker. Early work commonly adopted a "blind separation + auditory attention decoding (AAD) [10]" cascade: EEG was first used to estimate the target speech envelope, which was then compared with each separated source to identify the target talker. However, the overall performance was highly dependent on the accuracy of the auditory attention decoding. Moreover, cascaded approaches are prone to error propagation, limiting system reliability.

In this paper, we introduce TFGA-Net, that approach directly models listeners' attentional focus from the recorded EEG signals to extract the target speech. It consists of four components: speech encoder, EEG encoder, Speaker Extraction module, and speech decoder. The EEG encoder captures multi–scale time–frequency signatures and embeds task-selective cortical topology. The speaker extraction integrates MossFormer [11] with an RNN-free Recurrent, enabling it to retain global context and capture the rhythm and prosodic characteristics of speech. By combining local feature modeling with long-range contextual information, this architecture provides a balanced mechanism to enhance target speech while suppressing irrelevant sources. Experiments on the Cocktail Party and KUL datasets show that the TFGA-Net model achieves state-of-the-art performance across multiple evaluation metrics, with improvements of 14.1% and 15.8% in terms of Scale-Invariant Signal-to-Distortion Ratio (SI-SDR).

The main contributions of this paper are summarized as follows:
(1) We introduce a novel EEG encoder, which not only extracts multi-scale time–frequency representations of EEG signals but also integrates cortical topological structures that are selectively recruited during the task.
(2) We introduce a new speaker extraction module, which preserves the global context of fused representations and, at the same time, captures the periodicity and prosodic patterns of speech.
(3) We validate the proposed TFGA-Net model through a series of experiments on the Cocktail Party and KUL datasets, which show significant improvements over the baselines.

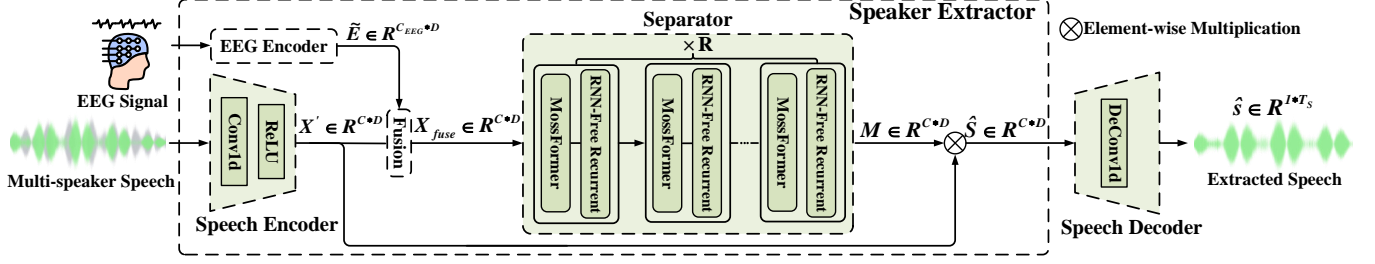
† This is the Corresponding Author.

Fig. 1: The overall block diagram of the proposed TFGA-Net model.

2. METHODS

2.1. Problem Formulation

Let $x(t)$ be a noisy multi-speaker mixture in the time domain:

$$x(t) = s_{\text{target}}(t) + \sum_{i=1}^{I} s_{\text{other},i}(t) \in \mathbb{R}^T \quad (1)$$

where $s_{\text{target}}(t)$ denotes the speech signal of the user-selected target speaker, $s_{\text{other},i}(t)$ denotes the speech of I interfering speakers, and T denotes the time length of mixture speech segments.

2.2. Overall Architecture

Fig.1 presents the overall structure of TFGA-Net, consisting of the the Speech Encoder, EEG Encoder, the Speaker Extraction Network, and the Speech Decoder.

The following sections provide a detailed explanation of each component.

Speech Encoder. The speech encoder consists of a one-dimensional convolutional layer(Conv1D) followed immediately by a ReLU activation function, ensuring that the encoded features remain non-negative. For an input sequence $X \in \mathbb{R}^{B \times 1 \times T_S}$ (where B is the batch size and T_S is the input length), the encoder applies a kernel size K_1 with a stride of $\frac{K_1}{2}$, producing an encoded output X', which can be defined as:

$$X' = \text{ReLU}(\text{Conv1D}(X)) \in \mathbb{R}^{B \times C \times D} \quad (2)$$

where N denotes the number of filters, and $S = \frac{2(T-K_1)}{K_1} + 1$ represents the reduced temporal dimension.

EEG Encoder. EEG signals encompass rich temporal and frequency characteristics, and the task-selective responsiveness of distinct cortical regions makes spatial topology equally important. However, current temporal convolution models (TCN [12] capture short-range patterns but ignore EEG functional connectivity, whereas graph-convolution models (GCN) enhance task-relevant regional activity, yet neglect long-range temporal dependencies. To address this, we introduce a temporal-frequency graph attention EEG encoding framework. This would allow us to characterize the hierarchical processing of the brain of target speech and provide top-down cues for speaker extraction.

The EEG encoder is designed to learn EEG embedding $\tilde{E}$ from the input EEG signal E that exhibit correlations with the interested speech.

Specifically, for EEG data $E \in \mathbb{R}^{B \times C \times T_e}$ (where B is the batch size, C is the number of electrode channels, and T_e is the input sequence length), the signal is sent to two components: a multi-scale temporal convolution module and a multi-frequency feature extraction module.

In the temporal convolution module, EEG data is processed by one-dimensional convolutional kernels with different receptive fields. We employ five convolution kernels whose lengths decay exponentially while remaining proportional to the sampling rate: $S_T^k = (1, 0.5^k f_s)$, where $k \in \{1, \ldots, 5\}$. Let the output of the kth temporal kernel be $E_T^k \in \mathbb{R}^{B \times C \times T \times f_k}$, where T is the number of temporal kernels, and f_k denotes the feature length:

$$E_T^k = \text{ELU}\Big(\text{BN}\big(\text{Conv1d}(E, S_T^k)\big)\Big) \quad (3)$$

We concatenate the five outputs along the feature dimension and apply a 1×1 convolution to obtain E_T.

In the frequency module, a short-time Fourier transform (STFT) is applied to each channel. Band-limited power is used to extract PSD and DE features. The signal is split into the five canonical bands: δ (0–4 Hz), θ (4–8 Hz), α (8–12 Hz), β (12–30 Hz), and γ (30–50 Hz). Averaging within each band yields $E_p \in \mathbb{R}^{C \times D_F}$ (PSD) and $E_D \in \mathbb{R}^{C \times D_F}$ (DE), where C is the number of channels and $D_F = 5$. Then, we combine the PSD and DE features to represent the EEG information in the frequency domain, denoted as $E_F \in \mathbb{R}^{C \times D_F}$.

In the next part of this module, we model multi-channel EEG features using a graph; each electrode in the EEG data is regarded as a node. To explore the implicit relationships among nodes, we employ Graph Convolutional Networks (GCNs). The adjacency matrix A represents the long–short distance brain network G and is symmetric because the graph is undirected. The initial adjacency matrices for the two views, A_T^{initial} and A_F^{initial}, are set identically to A. Using this construction, we obtain features from both views. The temporal graph-convolution branch (T-GCN) and the frequency branch (F-GCN) are defined as follows:

$$\tilde{E}_i = \varepsilon\big(D_i^{-\frac{1}{2}} A_i D_i^{-\frac{1}{2}} \varepsilon(E_i W_{i1}) W_{i2} + E_i\big), \, i \in \{T, F\} \quad (4)$$

Where $\tilde{x}_i \in \mathbb{R}^{C \times D_i}$ are hidden features of each view, with D_i denoting the degrees of A_i. $W_{i1}, W_{i2} \in \mathbb{R}^{D_i \times D_i'}$, are weight matrices, where D_i' is adjustable hyper-parameters, and $\varepsilon(\cdot)$ denote batch normalization followed by ELU non-linear functions. Finally, the temporal and frequency features are concatenated along the feature dimension and then sent into a self-attention [13] mechanism to capture global features:

$$\tilde{E} = \text{SA}\Big(\text{PE}_C\big(\text{Concat}(\tilde{E}_T, \tilde{E}_F)\big)\Big) \in \mathbb{R}^{B \times C_{EEG} \times D} \quad (5)$$

Speaker Extraction Network. The speaker extraction module is designed to estimate a mask M that allows only the attended
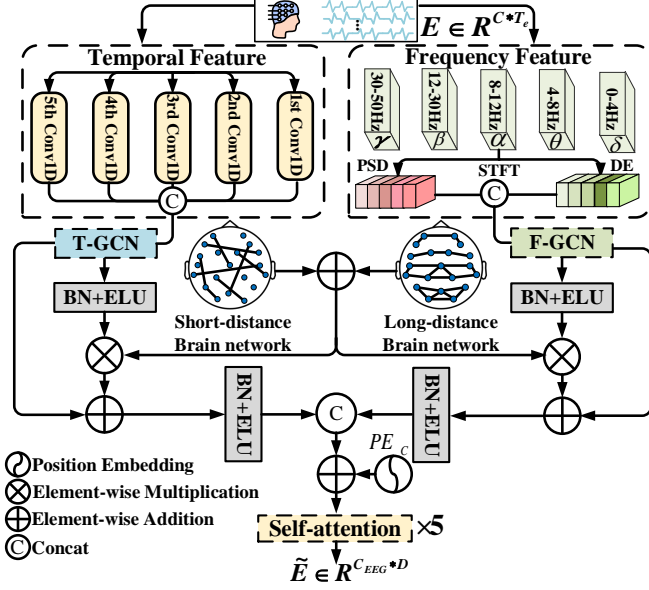
Fig. 2: The overall block diagram of the proposed EEGEncode.

speaker's voice to pass through X', and the masked speech embedding $\hat{S}$ is obtained by:

$$\hat{S} = X' \otimes M \in \mathbb{R}^{B \times C \times D} \tag{6}$$

We concatenate the mixed audio features X' and EEG features $\tilde{E}$ along the channel dimension, and then apply a 1D convolution to merge these features:

$$X'_{\text{fuse}} = \text{Conv1D}\big(\text{concat}(X', \tilde{E})\big) \in \mathbb{R}^{B \times C \times D} \tag{7}$$

The fused features are then fed into the speech separation models to extract the mask M.

Inspired by speech separation models based on Temporal Convolutional Networks (TCN) [12] and Dual-Path RNNs (DPRNN) [4], we adopt the more advanced MossFormer2 [14] as our separator. Unlike TCN and DPRNN, which are limited in long-term dependency modeling and computational efficiency, MossFormer2 combines local feature modeling with long-range contextual information. This balanced architecture enhances the extraction of target speech while suppressing irrelevant sources.

The MossFormer2 consists of two modules: the MossFormer Module and the RNN-Free Recurrent Module, which together model global context and local temporal patterns.

For the MossFormer module, it applies local full attention within non-overlapping chunks and linearized attention over the entire sequence. The fused output is refined by a gated convolutional unit:

$$O = X'' + \text{ConvM}\big(\sigma\big((U \otimes AV)\otimes AU\big)\big) \tag{8}$$

where X'' is the block input, U, V are projected features, A is the attention map, $\sigma(\cdot)$ denotes the element-wise sigmoid gating, $\text{ConvM}(\cdot)$ is a convolutional gating module, and $\otimes$ indicates element-wise multiplication.

For the RNN-free recurrent module, it employs a dilated feed-forward sequential memory network (FSMN) with gated convolutional units to model temporal recurrence:

$$U = \text{ConvU}(X), \quad V = \text{ConvU}(X) \tag{9}$$

$$Y = \text{DilatedFSMN}(V), \quad O = X + (U \otimes Y) \tag{10}$$

Here, X is the projected input to the recurrent block; $\text{ConvU}(\cdot)$ denotes a point-wise convolutional unit used to form gates; DilatedFSMN() is a dilated feed-forward memory block with dense connections that aggregates contexts. This parallel design enlarges the receptive field and avoids sequential dependencies, enabling efficient real-time separation.

Speech Decoder. The separated feature sequence is finally decoded into a waveform by the decoder:

$$\hat{s} = \text{deconv1D}(\hat{S}) \in \mathbb{R}^{B \times 1 \times T_s} \tag{11}$$

The decoder is a 1D transposed convolutional layer, and it uses the same kernel size and stride as the encoder.

2.3. Loss Function

In this study, we adopt the negative SI-SDR as the loss function, owing to its consistently good performance and its widespread use in target-speaker extraction. The SI-SDR is defined as:

$$\text{SI-SDR} = 10 \log_{10} \frac{\left\| \frac{\hat{s}^\top s}{\|s\|^2}\, s \right\|^2}{\left\| \frac{\hat{s}^\top s}{\|s\|^2}\, s - \hat{s} \right\|^2} \tag{12}$$

where $\hat{s}$ and s denote the extracted and true target-speaker signals, respectively.

3. EXPERIMENTS

3.1. Datasets

Cocktail Party Dataset. The first dataset [15] used in this experiment comprises 33 adults (28 male, 5 female; 27.3 ± 3.2 years) with normal hearing and no neurological disorders. Each subject undergoes 30 trials, each lasting 60 seconds, where they listen to two different stories–one in each ear–narrated by different male speakers. Subjects are divided into two groups: one focusing on the left ear (17 subjects) and the other on the right ear (16 subjects, with one subject excluded).

KUL Dataset. The KUL dataset [16] comprises 16 normal-hearing participants, each completing 20 dichotic trials. We use the first 8 trials in which each subject participated, in which subjects were presented with different speech in the left and right ears. Participants are asked to pay attention to the sounds in one ear and ignore the sounds in the other. The BioSemi ActiveTwo system is used to record 64-channel EEG signals at an 8196 Hz sample rate.

3.2. Data Processing

For the Cocktail Party data, our preprocessing steps remain consistent with UBESD: band-pass filtering (0.1–45 Hz), spline repair of noisy channels, mastoid-average rereferencing. For the KUL dataset, EEG data are first notch-filtered at 50 Hz to suppress power-line noise, then band-pass filtered (0.1–45 Hz, fourth-order Butterworth), re-referenced to the common average. For both datasets, the EEG data are downsampled to 128Hz and cleaned with ICA to remove ocular and muscular artefacts, while the speech sampling rate is set to 44.1kHz.

3.3. Implementation Details

For the Cocktail Party dataset, five trials are randomly selected as the test set, two trials are used for validation, and the remaining trials are used for training for each subject. For the KUL dataset, each subject's trials are divided into training, validation, and test sets with proportions of 75%, 12.5%, and 12.5%, respectively.

Experiments were conducted using the PYTORCH framework on an NVIDIA GeForce 4090 GPU. All models were trained for 60 epochs with a batch size of 1. The Adam optimizer was employed with a maximum learning rate of 0.0001. A StepLR scheduler reduced the learning rate by a factor of 0.5 every 20 epochs, producing a piecewise-constant decay throughout training.

For the model implementation, the kernel sizes are set to 20 for the Speech Encoder and Decoder. The encoder output dimension C is set to 128, and the number of separator blocks R is set to 6.

3.4. Evaluation Metrics

We assess our method with four metrics:SI-SDR(dB),PESQ [17], STOI [18], and ESTOI [19]. SI-SDR quantifies reconstruction fidelity; PESQ measures perceptual speech quality; STOI evaluates intelligibility via time–frequency correlations; ESTOI refines STOI for noisy conditions. Higher scores on all metrics indicate superior performance.

4. RESULTS

4.1. Comparative Analysis

To validate the effectiveness of the proposed algorithm, we perform experiments on the Cocktail Party and the KUL dataset. First, we compare our method with baseline models. Then we analyze SI-SDR improvement variations across different EEG Encoder and speaker extraction networks.

Table I: Performance comparison on Cocktail Party and KUL datasets

Dataset	Model	SI–SDR (dB)	STOI	ESTOI	PESQ
Cocktail Party dataset	Mixture	0.45	0.71	0.55	1.61
	UBESD [20]	8.54	0.83	–	1.97
	BASEN [21]	11.56	0.86	0.72	2.21
	M3ANet [22]	13.95	0.89	0.78	2.58
	TFGA-Net(Ours)	15.91	0.92	0.82	2.36
KUL	Mixture	0.25	0.69	0.52	1.17
	UBESD [20]	6.1	0.73	0.75	1.09
	BASEN [21]	11.5	0.82	0.76	1.76
	NeuroHeed [23]	14.6	0.83	0.76	2.12
	TFGA-Net(Ours)	16.9	0.87	0.78	2.17

As shown in Table I, the proposed TFGA-Net model attains state-of-the-art performance on the Cocktail Party dataset, achieving 15.91 dB SI–SDR. Relative to UBESD, BASEN, and M3ANet, SI–SDR improves by 7.37, 4.35, 3.02, and 1.96 dB, respectively. Compared with the state-of-the-art M3ANet method, our model achieves relative improvements of 0.03 and 0.04 in STOI and ESTOI, respectively, while reaching a modest PESQ score of 2.36. Against NeuroHeed on the KUL dataset, the TFGA-Net model delivers relative improvements of 2.3 dB, 0.04, 0.02, and 0.05 in SI–SDR, STOI, ESTOI, and PESQ, respectively. Therefore, the TFGA-Net model offers competitive performance compared with existing brain-controlled speaker-extraction approaches.

4.2. Ablation Study

To validate the contribution of each key module in TFGA-Net, we conduct ablation experiments on the Cocktail Party dataset.

Table II: Ablation experiments on the Cocktail Party dataset

Model	EEG Encoder	SI–SDR (dB)	STOI	ESTOI	PESQ
Mixture	-	0.45	0.71	0.55	1.61
TFGA-Net(Envelope)	Envelope	10.24	0.78	0.69	1.66
TFGA-Net(T–GCN)	T–GCN	14.78	0.86	0.73	1.91
TFGA-Net(F–GCN)	F–GCN	14.72	0.86	0.72	1.90
TFGA-Net(ours)	TF–GCN	15.91	0.92	0.82	2.36

Ablation Study on EEGEncoder. To validate the performance of the temporal-frequency graph attention EEG-encoding framework, we conducted a controlled experiment in which only the EEG encoder varied. Specifically, the Envelope model feeds raw EEG signals directly without feature extraction; the T-GCN and F–GCN models extract temporal and spectral features, respectively. The experimental results are summarized in Table II. Relative to the Envelope, T–GCN, and F–GCN based encoders, the TF-GCN model yields SI–SDR gains of 5.67 dB, 1.13 dB, and 1.19 dB, respectively, confirming its superiority.

Ablation Study on Speaker Extraction Network.

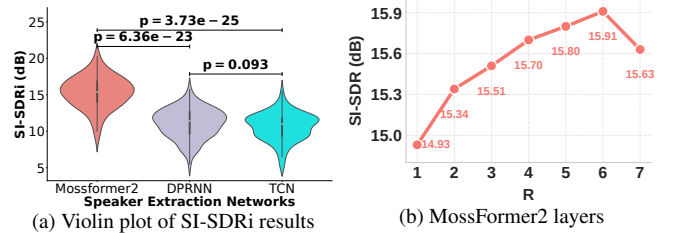(a) Violin plot of SI-SDRi results   (b) MossFormer2 layers

Fig. 3: Performance on Speaker Extraction Network

Fig.3(a) presents the violin plot of SI-SDRi distributions for the different speaker separation modules.The results demonstrate that the TFGA-Net model achieves superior overall performance.To fine-tune the optimal number of Mossformer2 layers, we evaluated the impact of varying the depth from 1 to 7. The results are summarized in Fig.3(b) Overall, increasing the number of layers significantly improves model performance, and the 6-layer MossFormer2 achieves the best results.

5. CONCLUSION

In this paper, we propose a network that efficiently extracts multi-scale time–frequency features and incorporates cortical topological structures selectively engaged during the task. During the post-fusion feature processing stage , the network preserves the global context of the fused representations and, under EEG guidance, captures speech periodicity and prosody. In two-speaker scenarios, the results suggest that TFGA-Net achieves higher signal fidelity and better perceptual quality than current state-of-the-art methods. Experiments show that temporal-frequency graph attention network with MossFormer2 for EEG extraction outperforms T–GCN and F–GCN approaches. The MossFormer2 module further improves local–global dependency modeling under EEG guidance, achieving higher SI–SDR scores and lower variance than TCN and DPRNN.

6. REFERENCES

[1] Adelbert W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[2] Sriram Srinivasan, Niels Henrik Pontoppidan, and Jesper Jensen, "A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions," *The Journal of the Acoustical Society of America*, vol. 146, no. 2, pp. 1372–1388, 2019.

[3] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[4] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 46–50.

[5] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong, "Attention is all you need in speech separation," in *Proc. Interspeech*, 2021, pp. 1636–1640.

[6] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe, "Tf-gridnet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[7] Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Lianwu Chen, Yuexian Zou, and Dong Yu, "Multi-modal multi-channel target speech separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 530–541, Mar. 2020.

[8] Rongzhi Gu, Lianwu Chen, Shi-Xiong Zhang, Jimeng Zheng, Yong Xu, Meng Yu, Dan Su, Yuexian Zou, and Dong Yu, "Neural spatial filter: Target speaker speech separation assisted with directional information," in *Proc. Interspeech*, Graz, Austria, 2019, pp. 4290–4294.

[9] Enea Ceolini, Jens Hjortkjær, Daniel D. E. Wong, James O'Sullivan, Vinay S. Raghavan, Jose Herrero, Ashesh D. Mehta, Shih-Chii Liu, and Nima Mesgarani, "Brain-informed speech separation (biss) for enhancement of target speaker in multitalker speech perception," *NeuroImage*, vol. 223, pp. 117282, 2020.

[10] S. Van Eyndhoven, T. Francart, and A. Bertrand, "Eeg-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 5, pp. 1045–1056, 2017.

[11] Shengkui Zhao and Bin Ma, "Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2023, pp. 10096–10100.

[12] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, vol. 30.

[14] Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jia Qi Yip, Dianwen Ng, and Bin Ma, "Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2024, pp. 10357–10360.

[15] Michael P. Broderick, Andrew J. Anderson, Giovanni M. Di Liberto, Michael J. Crosse, and Edmund C. Lalor, "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech," *Current Biology*, vol. 28, no. 5, pp. 803–809, Mar. 2018.

[16] Wouter Biesmans, Neetha Das, Tom Francart, and Alexander Bertrand, "Auditory-inspired speech envelope extraction methods for improved eeg-based auditory attention detection in a cocktail party scenario," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 5, pp. 402–412, May 2017.

[17] Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra, "Perceptual evaluation of speech quality (pesq): A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2001, pp. 749–752.

[18] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 4214–4217.

[19] Jesper Jensen and Cees H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[20] Maryam Hosseini, Luca Celotti, and Éric Plourde, "End-to-end brain-driven speech enhancement in multi-talker conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1718–1731, 2022.

[21] Jie Zhang, Qing-Tian Xu, Qiu-Shi Zhu, and Zhen-Hua Ling, "Basen: Time-domain brain-assisted speech enhancement network with convolutional cross attention in multi-talker conditions," in *Proc. Interspeech*, 2023, pp. 3117–3121.

[22] Cunhang Fan, Ying Chen, Jian Zhou, Zexu Pan, Jingjing Zhang, Youdian Gao, Xiaoke Yang, Zhengqi Wen, and Zhao Lv, "M3anet: Multi-scale and multi-modal alignment network for brain-assisted target speaker extraction," *arXiv preprint arXiv:2506.00466*, 2025.

[23] Zexu Pan, Marvin Borsdorf, Siqi Cai, Tanja Schultz, and Haizhou Li, "Neuroheed: Neuro-steered speaker extraction using eeg signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4456–4468, 2024.