Trainable EEG Interpolation and Structure-Sharing Dual-Path Encoders for Brain-Assisted Target Speaker Extraction

Zhao Lv[1], Haoran Zhou[1], Ying Chen[1], Youdian Gao[1], Xinhui Li[1], Ruibo Fu[2], Cunhang Fan[1]*

[1]State Key Laboratory of Opto-Electronic Information Acquisition and Protection Technology, (School of Computer Science and Technology), Anhui University, Hefei, 230601, Anhui, P. R. China
[2]Institute of Automation, Chinese Academy of Sciences
{kjlz,xinhuili,cunhang.fan}@ahu.edu.cn, {e24201017,e23201035,e23301143}@stu.ahu.edu.cn, ruibo.fu@nlpr.ia.ac.cn

Abstract

Brain-assisted target speaker extraction (TSE) isolates a target speaker's voice from a mixture by leveraging task-specific representations in Electroencephalogram (EEG) signals. However, existing methods rely on fixed interpolation for EEG-audio alignment, introducing redundant computations. They also employ single-path encoders that extract only target-relevant features while neglecting complementary, irrelevant ones, limiting discriminability. To address these limitations, this paper proposes a **T**rainable EEG **I**nterpolation and Structure-sharing **D**ual-path **E**ncoders network (TIDENet). The proposed Trainable EEG Interpolation (TEI) uses a neural network module to leverage cross-sample EEG information during resampling by parameters updating, thereby overcoming the limitations of fixed interpolation. The Structure-sharing Dual-path Encoders (SSDPE) extend existing speech and EEG encoders by introducing dual paths that separately process features relevant and irrelevant to the target speaker and incorporates interactive fusion between them, which enhances the encoder's ability to capture task-relevant information. Experimental results on public datasets demonstrate that TIDENet achieves relative improvements of up to **20.47%**, **22.22%**, **2.91%**, **6.20%**, and **15.84%** in signal-to-distortion ratio (SDR), scale-invariant SDR (SI-SDR), short-time objective intelligibility (STOI), extended STOI (ESTOI), and perceptual evaluation of speech quality (PESQ), respectively, compared to the state-of-the-art. These significant gains validate the effectiveness of the proposed TEI method and SSDPE architecture.

Code — https://github.com/segmentFT/TIDENet

Introduction

Selective auditory attention allows humans to focus on a target speaker and isolate their voice from a mixture of competing voices, which is crucial in complex auditory environments such as social gatherings or meetings. Target Speaker Extraction (TSE) aims to simulate this ability by extracting the target speaker's speech from mixed audio signals.

To distinguish target speaker from mixed speech, TSE models traditionally rely on various external cues, such as pre-recorded speech (Žmolíková et al. 2019; Xu et al. 2020; Ge et al. 2021; Liu et al. 2023; Hao, Li, and Zheng 2024), visual signals (Ephrat et al. 2018; Pan, Qian, and Li 2022; Lin et al. 2023; Li et al. 2024; Tao et al. 2025), or textual chat logs (Kim et al. 2025; Huo et al. 2025). However, these cues pose practical challenges in real-world applications like hearing aids. Speech-based cues require prior recordings of the target speaker, and users must manually indicate whom the cue refers to, as the model cannot autonomously identify the speaker. Visual cues, including lip or body movements, demand continuous precise camera alignment, proving unreliable in occlusion-prone dynamic environments. Textual cues from chat logs require integration with messaging platforms, limited by privacy and availability constraints.

Limitations of conventional TSE methods have driven alternative solutions. Mesgarani and Chang (2012) discovered pronounced neural responses to attended speech in the auditory cortex compared to unattended speech. Building on this, O'sullivan et al. (2015) successfully decoded attentional focus Electroencephalogram (EEG) signals to identify the attended speaker, suggesting that EEG can serve as a viable cue for TSE, which is known as Brain-Assisted TSE. Early approaches (Han et al. 2019; Ceolini et al. 2020) followed a two-stage pipeline: first separating the mixture, then estimating the target speaker's speech envelope from EEG to match the correct stream. Subsequently, to simplify the two-stage pipeline, Hosseini, Maryam and Celotti, Luca and Plourde, Éric (2021) proposed a fully end-to-end framework that treats EEG and speech as distinct modalities and performs feature-level fusion, enabling the direct extraction of target speech guided by neural activity. This multimodal fusion strategy has since become the prevailing paradigm in brain-assisted TSE research (Hosseini, Maryam and Celotti, Luca and Plourde, Éric 2022; Zhang et al. 2023; Fan et al. 2024; Pan et al. 2024; Fan et al. 2025).

Despite the superior performance of this modality fusion approach, it still faces two key challenges. First, there exists a significant discrepancy in temporal resolution between EEG and speech signals, with EEG exhibiting a much lower effective bandwidth, and this necessitates resampling to ensure temporal alignment for feature fusion (Broderick et al. 2018). Most existing approaches employ simple fixed interpolation, which incurs computational redundancy due to repeated or uninformative inputs. Second, current single-path encoders focus solely on extracting task-relevant features,

*Corresponding author.

neglecting the potential benefits of modeling irrelevant information. Prior work in image and speech processing (Chen et al. 2018; Zheng et al. 2021; Tao et al. 2025) shows that explicitly representing both relevant and irrelevant components and enabling their interaction can enhance feature discrimination. The absence of such mechanisms limits the encoder's ability to fully leverage available information.

To address these limitations, this paper proposes the **T**rainable EEG **I**nterpolation and Structure-Sharing **D**ual-path **E**ncoder Network (TIDENet). Trainable EEG Interpolation (TEI) method employs a neural network module for EEG resampling. Although the module's interpolation behavior is still fixed during a forward pass for a given batch, neural network's parameters updated via backpropagation cause the interpolation behavior for subsequent batches to *evolve*, which allows the module to implicitly leverage information across EEG samples during training in parallel with new useful information brought in for each interpolation step. Structure-Sharing Dual-Path Encoders (SSDPE) extends conventional speech/EEG encoders to a dual-path structure. This enables the encoder to explicitly extract both target-relevant and target-irrelevant features in parallel. Therefore, the model can actively amplify the discrepancy between them, and enhancing the discriminative power of the target-relevant features. A fusion module at the end of the SSDPE facilitates interaction between these paths, increasing their discriminability and further enhancing relevant features. Crucially, as relevant and irrelevant information within speech/EEG signals are homogeneous (both are fundamentally speech/EEG signals), the SSDPE employs shared structure between paths. This architectural constraint encourages the extracted features to reside in the same feature space.

The contributions of this paper are summarized as follows:

- This paper proposes a novel Brain-Assisted Target Speaker Extraction model with Trainable EEG Interpolation method, which enables the model to introduce new, effective information during resampling, improving the resampled EEG features.
- This paper proposes a Shared-Structure Dual-Path Encoder architecture for both speech and EEG encoders, which simultaneously extracts target-relevant and target-irrelevant features and facilitates interaction between them.
- On the Cocktail Party and AVED datasets, our proposed model surpasses existing SOTA models, demonstrating the effectiveness of the proposed improvements.

Related Works

Early brain-assisted target speaker extraction methods split the process into stimulus reconstruction and speech separation. In the first stage, a linear regression approach (Mesgarani et al. 2009; Akbari et al. 2019) and its deep learning extensions map EEG signals elicited by speech stimuli to amplitude envelopes; in the second stage, conventional speech separation unmixes the auditory mixture, and each separated output is compared against the reconstructed envelope to identify the target speaker (Han et al. 2019). Recognizing the limited expressiveness of linear models, Ceolini et al. (2020) simulated EEG inputs using noisy clean speech and trained a deep network on the corresponding envelope spectra, but the distribution mismatch between training and inference degraded extraction quality. To enhance multimodal fusion, Hosseini, Maryam and Celotti, Luca and Plourde, Éric (2021) introduced FiLM (Feature-wise Linear Modulation) (Perez et al. 2018)-based affine modulation in their Brain-Enhanced Speech Denoiser (BESD), later improved by UBESD (Hosseini, Maryam and Celotti, Luca and Plourde, Éric 2022) with an optimized separation network. Zhang et al. (2023) replaced FiLM with a cross-attention mechanism, developing the convolutional multi-layer cross-attention module and the Brain-Assisted Speech Enhancement Network (BASEN) architecture, which outperformed prior methods. Fan et al. (2024) further advanced BASEN by incorporating multi-scale convolutions for speech feature extraction, substituting its temporal convolutional network (Luo and Mesgarani 2019) with a dual-path recurrent neural network (Luo, Chen, and Yoshioka 2020), and employing graph convolutions to strengthen the EEG encoder. Pan et al. (2024) simplified multimodal fusion through linear projections and stacked self-attention layers, shifting complexity to the EEG encoder and introducing a pseudo-autoregressive scheme for online processing. The latest M3ANet (Multi-scale and Multi-Modal Alignment Network) (Fan et al. 2025) integrates the Mamba (Gu and Dao 2024) model to enhance the speech encoder, and uses contrastive learning to improve the temporal alignment between features of speech and EEG.

However, existing networks suffer from two critical limitations: fixed EEG resampling compromises cross-modal alignment precision, while single-path encoders neglect target-irrelevant features, preventing relevant feature enhancement through interaction with irrelevant one.

Model Architecture

TIDENet comprises five core components: a speech encoder, EEG encoder, fusion module, speech extractor, and decoder (architecture shown in Figure 1). The speech encoder extracts features from mixed audio, while the EEG encoder derives stimulus-aligned features from elicited EEG signals. The fusion module synthesizes these into new speech features embedding target speaker information. The speech extractor then isolates target speaker features, and the decoder reconstructs monaural waveforms.

Speech Encoder

The speech encoder employs our proposed **Structure-Sharing Dual-Path Encoder (SSDPE)** architecture (Figure 1), whose core innovation lies in explicitly modeling interaction mechanisms between complementary features. Each path comprises a 1D convolution layer (kernel size K_{in}, stride S_{in}, 1 input and C output channels) followed by PReLU activation. Processing monaural mixture $\mathbf{x} \in \mathbb{R}^{1\times T}$

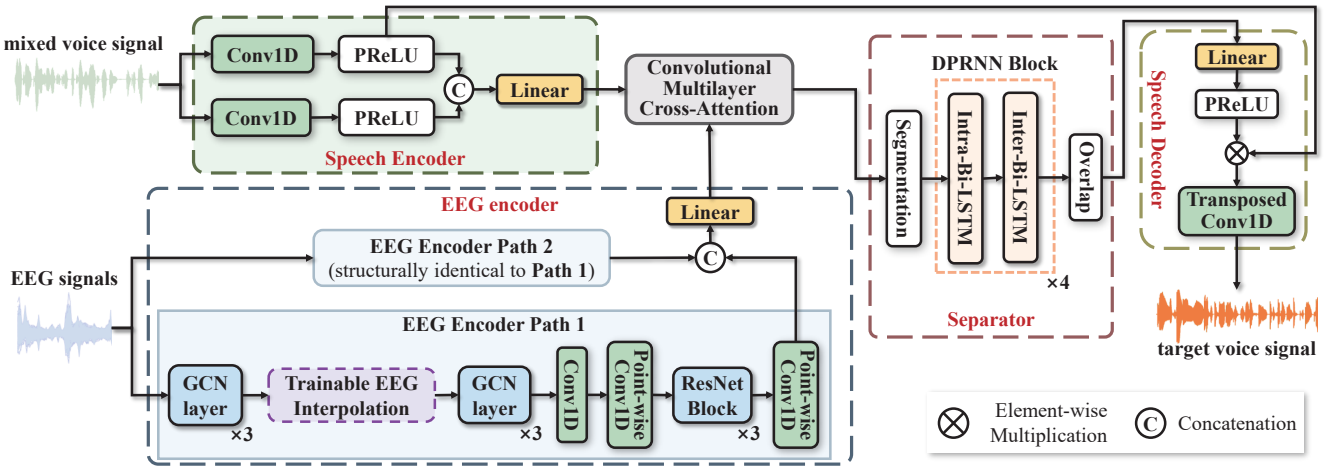

Figure 1: The overall architecture of the model. Dashed boxes delineate distinct components in the diagram. According to our proposed structure-sharing strategy, the second EEG encoder path has the same architecture as that of the first.

yields feature maps:

$$\begin{aligned}\mathbf{F}_{\text{in}}^{(1)} &= \text{Conv1D}\left(\mathbf{x} \mid K_{\text{in}}, S_{\text{in}}, 1, C\right) \in \mathbb{R}^{C\times T'} \\ \mathbf{F}_{\text{in}}^{(2)} &= \text{Conv1D}\left(\mathbf{x} \mid K_{\text{in}}, S_{\text{in}}, 1, C\right) \in \mathbb{R}^{C\times T'}\end{aligned} \tag{1}$$

where $\mathbf{F}_{\text{in}}^{(1)}$ and $\mathbf{F}_{\text{in}}^{(2)}$ encode features relevant/irrelevant to target speaker respectively. The dual-path design fundamentally decouples feature extraction processes to enable subsequent feature interaction. The fusion module concatenates them along channel dimension:

$$\begin{bmatrix}\mathbf{F}_{\text{in}}^{(1)} \\ \mathbf{F}_{\text{in}}^{(2)}\end{bmatrix} \in \mathbb{R}^{2C\times T'}$$

After layer normalization, a bias-free linear layer with parameter matrix $\mathbf{W}_{\text{fusion}_1} \in \mathbb{R}^{C\times 2C}$ reduces channel depth:

$$\mathbf{F}_{\text{in}} = \mathbf{W}_{\text{fusion}}^{(1)} \cdot \text{LayerNorm}\left(\begin{bmatrix}\mathbf{F}_{\text{in}}^{(1)} \\ \mathbf{F}_{\text{in}}^{(2)}\end{bmatrix}\right) \in \mathbb{R}^{C\times T'} \tag{2}$$

With nonlinearity of layer normalization, the linear layer intrinsically learns nonlinear interactions between effective/ineffective features: By adaptively adjusting channel-wise weighting coefficients, $\mathbf{W}_{\text{fusion}}^{(1)}$ suppresses interference components while amplifying target-specific attributes, yielding $\mathbf{F}_{\text{in}}$ with enhanced target relevance and robustness.

Trainable EEG Interpolation Module

The trainable EEG interpolation (TEI) module is implemented as a one-dimensional grouped transposed convolution layer without bias, where the number of groups equals the input channel count. Given an input length T_{EEG} matching the speech duration, the output length T'_{EEG} of the transposed convolution satisfies

$$T'_{\text{EEG}} = (T_{\text{EEG}} - 1) \cdot S_{\text{rs}} + K_{\text{rs}} + P_{\text{out}}, \tag{3}$$

where S_{rs}, K_{rs}, and P_{out} denote the stride, kernel size, and output padding, respectively. By selecting appropriate hyperparameter of S_{rs}, K_{rs} and P_{out}, $T'_{\text{EEG}} = T$ should be ensured. Since the input signals are not padded, the formula omits an input-padding term.

The principal advantage of the TEI module resides in its trainable nature. Let $\mathbf{B}_{\text{EEG}}^{(\tau)}$ and $\mathbf{B}_{\text{speech}}^{(\tau)}$ denote the EEG and speech input tensors for the τ-th iteration batch, with $\mathbf{\Theta}_{\text{TEI}}^{(\tau)}$ representing the TEI parameter tensor at this iteration. According to backpropagation and gradient descent, the updated parameters at iteration $\tau+1$ are given by:

$$\mathbf{\Theta}_{\text{TEI}}^{(\tau+1)} = \mathbf{\Theta}_{\text{TEI}}^{(\tau)} - \eta \cdot \nabla_{\mathbf{\Theta}_{\text{TEI}}^{(\tau)}} \mathcal{L}\left(\mathbf{B}_{\text{EEG}}^{(\tau)}, \mathbf{B}_{\text{speech}}^{(\tau)}\right) \tag{4}$$

where η is the learning rate, $\mathcal{L}$ denotes the loss function (dependent on $\mathbf{B}_{\text{EEG}}^{(\tau)}$ and $\mathbf{B}_{\text{speech}}^{(\tau)}$), and ∇ is the gradient operator. By definition, $\nabla_{\mathbf{\Theta}_{\text{TEI}}^{(\tau)}} \mathcal{L}$ constitutes a function of $\mathbf{B}_{\text{EEG}}^{(\tau)}$, implying that $\mathbf{\Theta}_{\text{TEI}}^{(\tau+1)}$ can be expressed as a mapping of $\mathbf{B}_{\text{EEG}}^{(\tau)}$.

The resampled EEG features $\mathbf{O}_{\text{TEI}}^{(\tau+1)}$ at iteration $\tau+1$ are computed from input features and updated parameters $\mathbf{\Theta}_{\text{TEI}}^{(\tau+1)}$, establishing $\mathbf{O}_{\text{TEI}}^{(\tau+1)}$ as a function of $\mathbf{\Theta}_{\text{TEI}}^{(\tau+1)}$. Through function composition, $\mathbf{O}_{\text{TEI}}^{(\tau+1)}$ is consequently a function of the input $\mathbf{B}_{\text{EEG}}^{(\tau)}$ at iteration τ.

Critically, Equation (4) forms a recurrence relation over iterations τ. Sequential application of function composition reveals that $\mathbf{O}_{\text{TEI}}^{(\tau+1)}$ inherently depends on all EEG samples processed prior to iteration τ. When τ exceeds the number of iterations per training epoch, $\mathbf{O}_{\text{TEI}}^{(\tau+1)}$ effectively incorporates information from the entire EEG training set. This demonstrates that through backpropagation and gradient descent, the TEI module leverages cross-sample global information to enhance feature representation. This distinguishes it from fixed interpolation relying solely on current EEG inputs without useful information from historical data.

The parameter $\mathbf{\Theta}_{\text{TEI}}^{(\tau)}$ evolves with iteration τ. During training, this evolution should monotonically improve overall model performance while preventing degradation of the TEI module itself. This implies that TEI's interpolation undergoes continuous optimization until convergence. Such inherent self-optimization capability constitutes a fundamental advantage over fixed interpolation methods.

EEG Encoder

The EEG encoder also follows the **SSDPE** architecture (Figure 1), each path adopting the structure from (Fan et al. 2024). The dynamic graph convolutional network (DGCN) encoder (Song et al. 2020) models C_{EEG}-channel EEG signals as directed weighted complete graphs: performing channel-wise batch normalization first, then transforming adjacency matrix $\mathbf{A}_{\text{EEG}}$ into Laplacian matrix, followed by 3rd-order Chebyshev polynomial filtering (Defferrard, Bresson, and Vandergheynst 2016). Trainable elements in $\mathbf{A}_{\text{EEG}}$ dynamically capture inter-channel correlations (element (i, j) denotes correlation between channels i and j), aggregating channels to extract preliminary features with output dimension $\mathbb{R}^{C_{\text{EEG}} \times T_{\text{EEG}}}$.

Within each path of the EEG encoder, the input EEG signals are first processed by the DGCN encoder to extract low-level features. These features are then resampled along the time dimension by our TEI module so that the output length matches the sample count T of the mixed speech. The resampled EEG features $\mathbf{F}'_{\text{EEG}} \in \mathbb{R}^{C_{\text{EEG}} \times T}$ undergo a second DGCN-based aggregation, yielding $\mathbf{F}'_{\text{EEG}_{\text{rs}}} \in \mathbb{R}^{C_{\text{EEG}} \times T}$. A subsequent one-dimensional convolution

$$\mathbf{F}''_{\text{EEG}} = \text{Conv1D}\left(\mathbf{F}'_{\text{EEG}_{\text{rs}}} \mid K_{\text{in}}, S_{\text{in}}, C_{\text{EEG}}, C/2\right) \in \mathbb{R}^{\frac{C}{2} \times T'} \tag{5}$$

reduces the channel count to $C/2$, aligning the temporal dimension with the mixed speech features. The final EEG feature extractor consists of two pointwise convolution layers and a stack of three ResBlocks, as shown in Figure 1. The first pointwise convolution and the first ResBlock preserve the channel count of $\mathbf{F}''_{\text{EEG}}$, while the second and third ResBlocks output C channels. The final pointwise convolution restores the channel count to $C/2$. All pooling operations are removed to maintain temporal alignment. Denoting the outputs of the two paths $\mathcal{E}_{\text{EEG}_1}$ and $\mathcal{E}_{\text{EEG}_2}$ as

$$\begin{aligned} \mathbf{F}_{\text{EEG}}^{(1)} &= \mathcal{E}_{\text{EEG}}^{1}(\mathbf{Y}) \in \mathbb{R}^{\frac{C}{2} \times T'} \\ \mathbf{F}_{\text{EEG}}^{(2)} &= \mathcal{E}_{\text{EEG}}^{2}(\mathbf{Y}) \in \mathbb{R}^{\frac{C}{2} \times T'} \end{aligned} \tag{6}$$

with input $\mathbf{Y} \in \mathbb{R}^{C_{\text{EEG}} \times T_{\text{EEG}}}$ representing the raw EEG signals, $\mathbf{F}_{\text{EEG}}^{(1)}$ and $\mathbf{F}_{\text{EEG}}^{(2)}$ are interpreted as speaker-related/unrelated features, respectively, which means that $\mathbf{F}_{\text{EEG}}^{(1)}$ encodes target-relevant features (effective components) and $\mathbf{F}_{\text{EEG}}^{(2)}$ captures irrelevant characteristics (ineffective components). These are fused by a bias-free linear layer

$$\mathbf{F}_{\text{EEG}} = \text{PReLU}\left(W_{\text{fusion}}^{(2)} \cdot \begin{bmatrix} \mathbf{F}_{\text{EEG}}^{(1)} \\ \mathbf{F}_{\text{EEG}}^{(2)} \end{bmatrix}\right) \in \mathbb{R}^{C \times T'} \tag{7}$$

where the parameter matrix is $W_{\text{fusion}}^{(2)} \in \mathbb{R}^{C \times C}$, preserving the channel count to match the speech encoder output. This linear layer dynamically regulates the contribution ratio of effective/ineffective features via learnable weights, suppressing interference while enhancing target-specific representations-constituting the core mechanism for feature interaction.

Fusion Module

TIDENet employs the convolutional multi-layer cross-attention (CMCA) module (Zhang et al. 2023) for feature fusion. In CMCA, EEG features enhance target speaker information in speech features and speech features filter EEG components irrelevant to auditory stimuli. This interaction amplifies target speaker-related representations, optimizing subsequent speech extraction. The CMCA architecture stacks three identical layers, each containing two multi-head cross-attention modules ($\text{CrossAttn}_j^{(i)}, j = 1, 2$) and two layer normalization units ($\text{LayerNorm}_j^{(i)}, j = 1, 2$). For layer i with inputs $\mathbf{G}_{\text{in}}^{(i)}$ (speech) and $\mathbf{G}_{\text{EEG}}^{(i)}$ (EEG):

$$\begin{aligned} \mathbf{G}_{\text{in}}^{(i+1)} &= \text{LayerNorm}_1^{(i)}\left(\text{CrossAttn}_1^{(i)}\left(\mathbf{G}_{\text{EEG}}^{(i)}, \mathbf{G}_{\text{in}}^{(i)}\right)\right) \\ \mathbf{G}_{\text{EEG}}^{(i+1)} &= \text{LayerNorm}_2^{(i)}\left(\text{CrossAttn}_1^{(i)}\left(\mathbf{G}_{\text{in}}^{(i)}, \mathbf{G}_{\text{EEG}}^{(i)}\right)\right) \end{aligned} \tag{8}$$

where $\text{CrossAttn}_j^{(i)}$ generates query matrices from its first input and key/value matrices from its second input. Initialized with $\mathbf{G}_{\text{in}}^{(1)} = \mathbf{F}_{\text{in}}$ and $\mathbf{G}_{\text{EEG}}^{(1)} = \mathbf{F}_{\text{EEG}}$, the final output concatenates initial $(\mathbf{F}_{\text{in}}, \mathbf{F}_{\text{EEG}})$ and processed features $\left(\mathbf{G}_{\text{in}}^{(4)}, \mathbf{G}_{\text{EEG}}^{(4)}\right)$ along channel dimension, compressed to C channels via bias-free linear layer followed by PReLU.

Speech Extractor

The dual-path recurrent neural network (DPRNN) (Luo, Chen, and Yoshioka 2020) is employed as the speech extractor. This time-domain separation network segments 1D feature maps into overlapped chunks (50% overlap) and reshapes them into a 3D tensor $\mathcal{T} \in \mathbb{R}^{P \times N \times C}$ (P: chunk length, N: chunk count, C: channels). The architecture stacks R DPRNN blocks, each containing two bidirectional LSTMs: intra-chunk and inter-chunk LSTM, processing along P-dimension and N-dimension respectively. This design reduces time steps by $1/P$ for inter-chunk LSTM, mitigating gradient explosion risks. Chunk overlap ensures temporal continuity for intra-chunk LSTM, with final output reconstructed via overlap-add. In TIDENet, fused features feed into DPRNN for end-to-end target speaker separation.

Decoder

The decoder comprises a linear layer followed by a 1D transposed convolutional layer. The linear layer transforms speech features into target speech masks, which are element-wise multiplied with the encoder output $\mathbf{F}_{\text{in}}^{(1)}$ to produce intermediate features $\mathbf{F}_{\text{out}} \in \mathbb{R}^{C \times T'}$. The transposed convolution then reconstructs target speech:

$$\widehat{\mathbf{x}} = \text{DeConv}\left(\mathbf{F}_{\text{out}} \mid K_{\text{in}}, S_{\text{in}}, C, 1\right) \in \mathbb{R}^{1 \times T}, \tag{9}$$

where K_{in} and S_{in} maintain identical kernel size and stride as the encoder's convolutional layer.

Loss Function

The negative SI-SDR between target speech $\mathbf{s} \in \mathbb{R}^{1\times T}$ and estimated speech $\widehat{\mathbf{x}}$ serves as the loss function:

$$\mathcal{L}\left(\mathbf{s}, \widehat{\mathbf{x}}\right) = -20 \cdot \log_{10} \frac{\left\| \frac{\langle \widehat{\mathbf{x}}, \mathbf{s} \rangle}{\|\mathbf{s}\|_2^2} \cdot \mathbf{s} \right\|_2}{\left\| \widehat{\mathbf{x}} - \frac{\langle \widehat{\mathbf{x}}, \mathbf{s} \rangle}{\|\mathbf{s}\|_2^2} \cdot \mathbf{s} \right\|_2} \in \mathbb{R} \tag{10}$$

This function is differentiable w.r.t. both $\widehat{\mathbf{x}}$ and $\mathbf{s}$.

Experiments

Datasets

Cocktail Party. The dataset (Broderick et al. 2018) comprises 33 normal-hearing subjects (28 male, 5 female). Each subject completed 30 trials with monaural auditory stimuli (44.1 kHz) presented via headphones. Subjects were divided into two attention groups: one attending exclusively to left-ear stimuli, the other to right-ear stimuli. EEG signals were recorded synchronously using a 128-electrode cap at 512 Hz sampling rate.

AVED. The dataset (ZHANG et al. 2024) comprises 20 normal-hearing subjects (14M/6F; mean age 20). Participants completed 16 trials (152s each) using in-ear headphones, attending to gender-specific narratives (male/female voice to left/right ear). All audio stimuli maintained constant amplitude (44.1 kHz), with synchronized EEG recorded via 32-channel system at 1 kHz.

Data Preprocessing

Audio stimuli downsampled from 44.1 kHz to 14.7 kHz. Left/right ear signals RMS-normalized and superimposed to create mixed speech samples. Data are partitioned at trial level and training/validation sets segmented into non-overlapping 2s clips (60s trials); test set into 20s clips. EEG signals re-referenced to mastoid average, artifacts removed via ICA in EEGLAB (Delorme and Makeig 2004), downsampled to 128 Hz.

In the Cocktail Party dataset, 5 trials per subject are randomly selected for testing, 2 are for validation, and remainder for training. Bandpass-filtered EEG (0.1-45 Hz) fed into frequency-band coupling (FBC) model (Moinnereau et al. 2020; Whittingstall and Logothetis 2009) estimating cortical multi-unit activity (MUA):

$$N(t) \;=\; a_\gamma \times P_\gamma(t) + a_\delta \times \angle\delta(t)$$

where $N(t)$ estimates neural activity at t, $P_\gamma(t)$ and $\angle\delta(t)$ denote gamma-band amplitude and delta-band phase respectively, and a_γ and a_δ predefined coefficients respectively.

In the AVED dataset, 1 trial per subject is randomly assigned to test/validation sets each, and the remainder is for training. EEG channels are bandpass-filtered (1-50 Hz) with average re-reference. Only trials where subjects attended to male narrator analyzed to avoid attention shift interference.

Implementation Details

The 1D convolutional layer adopts kernel $K_{\text{in}} = 8$, stride $S_{\text{in}} = 4$, with feature map channels $C = 128$. To reduce parameter overhead in CMCA, standard multi-head attention linear layers are replaced by depthwise separable convolutions (depthwise kernel size 3), using single attention head. The separator network sets hyperparameter $R = 4$ with LSTM hidden dimension equal to C. Since input speech/EEG signals have variable lengths, resampling parameters $S_{\text{rs}}/K_{\text{rs}}/P_{\text{out}}$ are dynamically configured to synchronize EEG and speech sampling rates.

Evaluation Metrics

Scale-invariant signal-to-distortion ratio (SI-SDR) serves as the core metric for speech physical fidelity, with higher values indicating superior signal quality (Le Roux et al. 2019). Four complementary metrics are reported: signal-to-distortion ratio (SDR) measures physical signal quality (scale-sensitive), where elevated values denote better speech quality (Vincent, Gribonval, and Févotte 2006); short-time objective intelligibility (STOI, 0-1) assesses speech clarity, with higher scores positively correlating with human auditory comprehension (Taal et al. 2011); extended STOI (ESTOI, 0-1) provides enhanced precision, where increased scores reflect improved intelligibility (Jensen and Taal 2016); perceptual evaluation of speech quality (PESQ, 0.5-4.5) predicts subjective audio quality, with higher ratings signifying superior perceptual quality (Rix et al. 2001).

Training Details

Implemented in PyTorch, all models trained for 400 epochs with two NVIDIA GeForce RTX 3090 GPUs. Both full and ablated variants use Adam optimizer with cyclic learning rate (2,000 iterations): linear warmup over initial 40 iterations per cycle followed by cosine decay. Peak learning rate 3.5×10^{-4}, trough $1/25$ of peak, fixed batch size 8.

To resolve channel mismatch (AVED: 32 vs encoder: 128), quadruplicate and concatenate 32-channel segments along channel axis to form 128-channel inputs, enabling accelerated convergence via Cocktail Party-pretrained model fine-tuning. Unlike prior encoder reduction approaches, this work maintains unified 128-channel architecture across datasets.

Results

Comparison with Baseline Models

This paper presents a systematic comparative analysis of TIDENet against classical EEG-based speech extraction models (BESD, UBESD, NeuroHeed) and state-of-the-art architectures (MSFNet, M3ANet), as detailed in Table 1. On the Cocktail Party dataset, TIDENet demonstrates significant performance improvements over M3ANet: SI-SDR **+1.68 dB**, SDR **+1.65 dB**, STOI **+3.61%**, ESTOI **+4.75%**, and PESQ **+0.39**. On the AVED dataset, the performance gains are: SI-SDR **+2.42 dB**, SDR **+2.28 dB**, STOI **+2.64%**, ESTOI **+5.09%**, and PESQ **+0.35**. These results indicate that TIDENet achieves the highest SDR and SI-SDR scores on both datasets, signifying minimal distortion and optimal reconstruction quality of the separated speech. Concurrently, TIDENet also attains the highest STOI and ESTOI scores, highlighting its advantage in enhancing speech intelligibility. Furthermore, its superior performance on the PESQ met-

Datasets	Methods	SDR(dB)	SI-SDR(dB)	STOI	ESTOI	PESQ
Cocktail Party	Mixture	0.47	0.45	74.00%	55.00%	1.61
	BESD (2021)	-	5.75	79.00%	-	1.79
	UBESD (2022)	-	8.54	83.00%	-	1.97
	BASEN (2023)	-	12.23	86.00%	-	2.24
	NeuroHeed (2024)	-0.09	-0.11	71.48%	54.79%	1.45
	MSFNet (2024)	13.03	12.89	88.00%	77.00%	2.51
	M3ANet (2025)	14.11	13.95	89.23%	78.36%	2.58
	TIDENet(ours)	**15.76**	**15.63**	**92.84%**	**82.93%**	**2.97**
AVED	Mixture	1.54	1.52	75.83%	60.57%	1.50
	UBESD (2022)	8.1	7.89	85.00%	72.00%	1.75
	BASEN (2023)	8.68	8.46	86.00%	75.00%	1.91
	NeuroHeed (2024)	8.77	8.61	88.11%	77.81%	1.82
	MSFNet (2024)	9.84	9.65	89.00%	79.00%	2.07
	M3ANet (2025)	11.14	10.89	90.60%	82.06%	2.21
	TIDENet(ours)	**13.42**	**13.31**	**93.24%**	**87.15%**	**2.56**

Table 1: Comparisons with baseline models on the Cocktail Party and AVED datasets. Experimental results for the NeuroHeed on both the Cocktail Party and AVED datasets, along with results of all other models on the AVED dataset, are reproduced from (Fan et al. 2025). The remaining experimental results in the tables are sourced from their respective original publications.

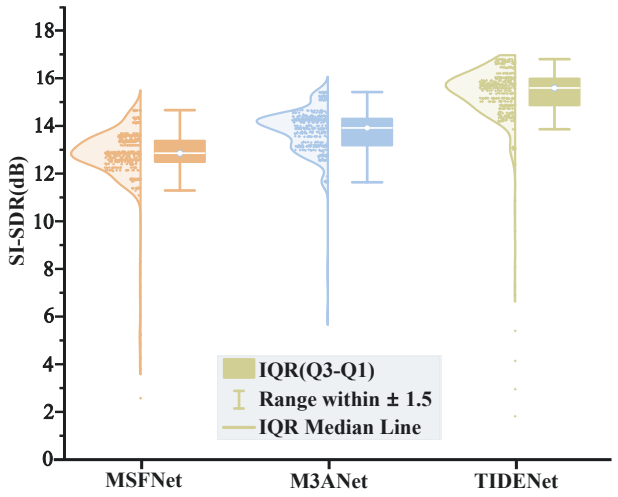

Figure 2: Illustration of SI-SDR (dB) distributions of different models on the Cocktail Party dataset test set. Solid color markers show SI-SDR values per test sample. Symmetrical half-violin plots depict sample density with adjacent box plots displaying median, interquartile range (IQR), data range ($\pm 1.5 \times$IQR), and potential outliers. The x-axis identifies three models while the y-axis quantifies SI-SDR (0-18 dB).

ric indicates better subjective auditory quality of the separated speech.

Figure 2 systematically visualizes the SI-SDR distribution characteristics of MSFNet, M3ANet, and TIDENet on the Cocktail Party dataset test set through an integrated half-violin and box plot representation. Compared to MSFNet and M3ANet, TIDENet exhibits a pronounced rightward shift in both probability density curves and box plot medians, with median SI-SDR values increasing by approximately 2.7 dB and 1.6 dB respectively, which demonstrates its superior average speech separation accuracy. Crucially, TIDENet displays significantly narrower interquartile ranges (IQR) and whisker lengths (defined by the $1.5 \times$IQR range), indicating reduced performance variance across diverse test samples and enhanced algorithmic consistency and robustness. Collectively, these results confirm that TIDENet achieves not only improvements in separation quality but also surpasses MSFNet and M3ANet in stability, validating its efficacy in complex acoustic mixing scenarios.

Ablation Study

Table 2 presents the results of 8 ablation experiments conducted on the Cocktail Party dataset. This study systematically evaluates the individual contributions and synergistic effects of the Trainable EEG Interpolation (TEI) module and the Structure-sharing Dual-Path Encoders (SSDPE) architecture (applied separately to the speech encoder and EEG encoder) on the overall model performance. The baseline model (#1), representing the configuration without the TEI module or SSDPE architecture, which uses fixed interpolation algorithm for resamping EEG features and both of the EEG and speech encoders are single-path, establishes the lower performance bound for the TIDENet model.

Model #2 incorporating the TEI module, demonstrates significant performance improvements over the baseline (#1) across all metrics: **+1.38 dB** in SI-SDR, **+1.31 dB** in SDR, **+2.15%** in STOI, **+3.84%** in ESTOI, and **+0.28** in PESQ. This indicates that the TEI module enhances the model's performance by refining the interpolation for EEG features, thereby improving the temporal alignment precision between EEG and speech features.

Model #3, which integrates the SSDPE architecture solely in the EEG encoder, also exhibits noticeable gains compared to the baseline (#1): **+1.18 dB** in SI-SDR, **+1.20 dB**

No.	Methods			Metrics				
	Trainable EEG Interpolation	Structure-Sharing Dual-Path Encoders						
		Speech	EEG	SI-SDR(dB)	SDR(dB)	STOI	ESTOI	PESQ
#1	×	×	×	14.10	14.22	90.08%	78.93%	2.66
#2	✓	×	×	15.48	15.53	92.23%	82.77%	2.94
#3	×	×	✓	15.28	15.42	91.36%	82.34%	2.89
#4	×	✓	×	14.28	14.41	90.43%	79.93%	2.72
#5	×	✓	✓	15.31	15.43	91.90%	82.41%	2.90
#6	✓	×	✓	15.35	15.46	92.03%	82.53%	2.94
#7	✓	✓	×	15.30	15.44	91.86%	**85.07%**	2.93
#8(ours)	✓	✓	✓	**15.63**	**15.76**	**92.84%**	82.93%	**2.97**

Table 2: Structure of ablation studies on the Cocktail Party dataset. The **No.** column indicates experiment IDs. Under the **Methods** header, checkmarks (✓) in the **Trainable EEG Interpolation** column denote inclusion of the module in experiments, while crosses (×) indicate its removal, and the **Structure-Sharing Dual-Path Encoders** column contains two subcolumns (**Speech** and **EEG**). Checkmarks (✓) in these subcolumns signify dual-path encoder architectures for speech/EEG respectively, with crosses (×) indicating single-path structures.

in SDR, **+1.28%** in STOI, **+3.41%** in ESTOI, and **+0.23** in PESQ. These improvements demonstrate that the SSDPE within the EEG encoder effectively extracts both target-relevant/irrelevant features from the input EEG signals. The interaction between these features leads to more discriminative representations, which boosts model performance.

In contrast, model #4, which applies the SSDPE architecture only to the speech encoder, shows marginal gains over the baseline (#1) in SI-SDR (**+0.18 dB**), SDR (**+0.19 dB**), STOI (**+0.35%**), ESTOI (**+1.00%**), and PESQ (**+0.06**). This suggests that while the SSDPE architecture in the speech encoder contributes positively, its impact in isolation is limited. Its benefits are more effectively realized when combined with other enhancements, such as the TEI module or SSDPE applied to the EEG encoder.

The model #5 equipped with SSDPE in both speech and EEG encoders but without TEI, which achieves moderate gains over the baseline (#1): SI-SDR (**+1.21 dB**), SDR (**+1.21 dB**), STOI (**+1.28%**), ESTOI (**+3.48%**), PESQ (**+0.24**). However, it underperforms TEI-only model #2 (e.g., SI-SDR: **15.31 dB** vs. **15.48 dB**). This reveals that SSDPE deployment in both encoders provides additive benefits (surpassing model #4's marginal gains), confirming SSDPE's representational capacity, and the performance deteriorates without TEI module, highlighting TEI's essential role in enabling precise alignment for cross-modal fusion.

Model #6 and #7 augment model #3 (SSDPE in EEG encoder only) and #4 (SSDPE in speech encoder only) with the TEI module respectively. Results show that both models #6 and #7 outperform models #3 and #4 across all metrics respectively, confirming that the TEI module can improve performance provided by the SSDPE architecture. However, critically, the performance of models #6 and #7 on all metrics (except the PESQ's value of model #7) is slightly inferior to or merely matches that of model #2 with only the TEI module. This implies that applying the SSDPE architecture to either the speech or EEG encoder alone may interfere with the TEI module learning the optimal interpolation.

Crucially, when the model integrates both the TEI module and the SSDPE architecture (applied to both encoders) (model #8), it achieves the optimal performance across all evaluation metrics except ESTOI (SI-SDR **15.63 dB**, SDR **15.76 dB**, STOI **92.84%**, PESQ **2.97**). Although model #8 exhibits a lower ESTOI metric compared to model #7, it is still regarded as the optimal model among all variants based on the other four performance indicators. This pivotal result indicates that while the TEI module and the SSDPE architecture (applied to either encoder) can independently improve performance, there exists a significant non-linear synergistic effect among them. Their combined integration unlocks the model's full performance potential.

Conclusion

This paper introduces a novel time-domain brain-guided target speaker extraction model, where a trainable EEG interpolation module supersedes existing fixed methods only formally extending EEG signals without additional information brought in by learning optimal interpolation for EEG-speech temporal alignment, and dual-path speech/EEG encoders extract target-relevant/irrelevant features with fusion modules, overcoming the limitation of single-path encoders which can only actively process target-relevant features while ignoring the extraction and utilization of irrelevant ones. Evaluations on two public datasets demonstrate state-of-the-art performance. Future work will investigate sophisticated interpolation for cross-modal alignment and advanced fusion methods for utilization of irrelevant features.

Acknowledgments

This work is supported by the {STI 2030—Major Projects (No. 2021ZD0201500)}, the National Natural Science Foundation of China (NSFC) (No.62571002, 62201002, 6247077204), Excellent Youth Foundation of Anhui Scientific Committee (No. 2408085Y034), Cloud Ginger XR-1.

References

Akbari, H.; Khalighinejad, B.; Herrero, J. L.; Mehta, A. D.; and Mesgarani, N. 2019. Towards reconstructing intelligible speech from the human auditory cortex. *Scientific reports*, 9(1): 874.

Broderick, M. P.; Anderson, A. J.; Di Liberto, G. M.; Crosse, M. J.; and Lalor, E. C. 2018. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, 28(5): 803–809.

Ceolini, E.; Hjortkjær, J.; Wong, D. D.; O'Sullivan, J.; Raghavan, V. S.; Herrero, J.; Mehta, A. D.; Liu, S.-C.; and Mesgarani, N. 2020. Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception. *NeuroImage*, 223: 117282.

Chen, S.; Tan, X.; Wang, B.; and Hu, X. 2018. Reverse attention for salient object detection. In *Proceedings of the European conference on computer vision (ECCV)*, 234–250.

Defferrard, M.; Bresson, X.; and Vandergheynst, P. 2016. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Delorme, A.; and Makeig, S. 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1): 9–21.

Ephrat, A.; Mosseri, I.; Lang, O.; Dekel, T.; Wilson, K.; Hassidim, A.; Freeman, W. T.; and Rubinstein, M. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. arXiv:1804.03619.

Fan, C.; Chen, Y.; Zhou, J.; Pan, Z.; Zhang, J.; Gao, Y.; Yang, X.; Wen, Z.; and Lv, Z. 2025. M3ANet: Multi-scale and Multi-Modal Alignment Network for Brain-Assisted Target Speaker Extraction. arXiv:2506.00466.

Fan, C.; Zhang, J.; Zhang, H.; Xiang, W.; Tao, J.; Li, X.; Yi, J.; Sui, D.; and Lv, Z. 2024. MSFNet: Multi-scale fusion network for brain-controlled speaker extraction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1652–1661.

Ge, M.; Xu, C.; Wang, L.; Chng, E. S.; Dang, J.; and Li, H. 2021. Multi-stage speaker extraction with utterance and frame-level reference signals. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6109–6113. IEEE.

Gu, A.; and Dao, T. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752.

Han, C.; O'Sullivan, J.; Luo, Y.; Herrero, J.; Mehta, A. D.; and Mesgarani, N. 2019. Speaker-independent auditory attention decoding without access to clean speech sources. *Science advances*, 5(5): eaav6134.

Hao, F.; Li, X.; and Zheng, C. 2024. X-TF-GridNet: A time-frequency domain target speaker extraction network with adaptive speaker embedding fusion. *Information Fusion*, 112: 102550.

Hosseini, Maryam and Celotti, Luca and Plourde, Éric. 2021. Speaker-Independent Brain Enhanced Speech Denoising. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1310–1314.

Hosseini, Maryam and Celotti, Luca and Plourde, Éric. 2022. End-to-end brain-driven speech enhancement in multi-talker conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 1718–1733.

Huo, M.; Jain, A.; Huynh, C. P.; Kong, F.; Wang, P.; Liu, Z.; and Bhat, V. 2025. Beyond speaker identity: Text guided target speech extraction. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Jensen, J.; and Taal, C. H. 2016. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11): 2009–2022.

Kim, M.; Mira, R.; Chen, H.; Petridis, S.; and Pantic, M. 2025. Contextual Speech Extraction: Leveraging Textual History as an Implicit Cue for Target Speech Extraction. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Le Roux, J.; Wisdom, S.; Erdogan, H.; and Hershey, J. R. 2019. SDR–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 626–630. IEEE.

Li, K.; Xie, F.; Chen, H.; Yuan, K.; and Hu, X. 2024. An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10): 6637–6651.

Lin, J.; Cai, X.; Dinkel, H.; Chen, J.; Yan, Z.; Wang, Y.; Zhang, J.; Wu, Z.; Wang, Y.; and Meng, H. 2023. Av-sepformer: Cross-attention sepformer for audio-visual target speaker extraction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Liu, K.; Du, Z.; Wan, X.; and Zhou, H. 2023. X-sepformer: End-to-end speaker extraction network with explicit optimization on speaker confusion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Luo, Y.; Chen, Z.; and Yoshioka, T. 2020. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 46–50. IEEE.

Luo, Y.; and Mesgarani, N. 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8): 1256–1266.

Mesgarani, N.; and Chang, E. F. 2012. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397): 233–236.

Mesgarani, N.; David, S. V.; Fritz, J. B.; and Shamma, S. A. 2009. Influence of Context and Behavior on Stimulus

Reconstruction From Neural Activity in Primary Auditory Cortex. *Journal of Neurophysiology*, 102(6): 3329–3339. PMID: 19759321.

Moinnereau, M.-A.; Rouat, J.; Whittingstall, K.; and Plourde, E. 2020. A frequency-band coupling model of EEG signals can capture features from an input audio stimulus. *Hearing Research*, 393: 107994.

O'sullivan, J. A.; Power, A. J.; Mesgarani, N.; Rajaram, S.; Foxe, J. J.; Shinn-Cunningham, B. G.; Slaney, M.; Shamma, S. A.; and Lalor, E. C. 2015. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral cortex*, 25(7): 1697–1706.

Pan, Z.; Borsdorf, M.; Cai, S.; Schultz, T.; and Li, H. 2024. NeuroHeed: Neuro-steered speaker extraction using EEG signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Pan, Z.; Qian, X.; and Li, H. 2022. Speaker extraction with co-speech gestures cue. *IEEE Signal Processing Letters*, 29: 1467–1471.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Rix, A. W.; Beerends, J. G.; Hollier, M. P.; and Hekstra, A. P. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, 749–752. IEEE.

Song, T.; Zheng, W.; Song, P.; and Cui, Z. 2020. EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks. *IEEE Transactions on Affective Computing*, 11(3): 532–541.

Taal, C. H.; Hendriks, R. C.; Heusdens, R.; and Jensen, J. 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on audio, speech, and language processing*, 19(7): 2125–2136.

Tao, R.; Qian, X.; Jiang, Y.; Li, J.; Wang, J.; and Li, H. 2025. Audio-visual target speaker extraction with selective auditory attention. *IEEE Transactions on Audio, Speech and Language Processing*.

Vincent, E.; Gribonval, R.; and Févotte, C. 2006. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4): 1462–1469.

Whittingstall, K.; and Logothetis, N. K. 2009. Frequency-band coupling in surface EEG reflects spiking activity in monkey visual cortex. *Neuron*, 64(2): 281–289.

Xu, C.; Rao, W.; Chng, E. S.; and Li, H. 2020. Spex: Multi-scale time domain speaker extraction network. *IEEE/ACM transactions on audio, speech, and language processing*, 28: 1370–1384.

ZHANG, H.; ZHANG, J.; DONG, X.; LÜ, Z.; TAO, J.; ZHOU, J.; WU, X.; and FAN, C. 2024. Based on audio-video evoked auditory attention detection electroencephalogram dataset. *Journal of Tsinghua University (Science and Technology)*, 64(11): 1919–1926.

Zhang, J.; Xu, Q.-T.; Zhu, Q.-S.; and Ling, Z.-H. 2023. BASEN: Time-Domain Brain-Assisted Speech Enhancement Network with Convolutional Cross Attention in Multi-talker Conditions. arXiv:2305.09994.

Zheng, C.; Peng, X.; Zhang, Y.; Srinivasan, S.; and Lu, Y. 2021. Interactive speech and noise modeling for speech enhancement. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14549–14557.

Žmolíková, K.; Delcroix, M.; Kinoshita, K.; Ochiai, T.; Nakatani, T.; Burget, L.; and Černockỳ, J. 2019. Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, 13(4): 800–814.