计 算 机 研 究 与 发 展
Journal of Computer Research and Development

https://crad.ict.ac.cn
63(6): 1534−1547, 2026

GALENet: 融合门控注意力与局部特征增强的神经导向说话人提取网络

范存航[1] 李骏凌[1] 张晶晶[1] 高佑典[1] 陈 颖[1] 易江燕[2] 吕 钊[1]

[1](光电信息获取与防护技术全国重点实验室(安徽大学) 合肥 230601)

[2](清华大学自动化系 北京 100084)

(cunhang.fan@ahu.edu.cn)

GALENet: Neuro-Oriented Speaker Extraction Network Fusing Gated Attention and Local Feature Enhancement

Fan Cunhang[1], Li Junling[1], Zhang Jingjing[1], Gao Youdian[1], Chen Ying[1], Yi Jiangyan[2], and Lü Zhao[1]

[1] (*State Key Laboratory of Opto-Electronic Information Acquisition and Protection Technology* (*Anhui University*), *Hefei* 230601)

[2] (*Department of Automation*, *Tsinghua University*, *Beijing* 100084)

Abstract Neuro-oriented speaker extraction is an intelligent speech processing technique that emulates human auditory attention mechanisms, enabling target speaker extraction from mixed speech by decoding the attentional focus embedded in listeners' EEG (electroencephalogram) signals. This technology offers novel insights into resolving the "cocktail party problem" and holds significant implications for developing intelligent hearing-assistive devices. However, existing approaches confront three critical challenges: inefficient multimodal feature fusion, inadequate local speech feature extraction, and high computational complexity, which collectively constrain practical application efficacy. We propose GALENet, a novel end-to-end time-domain speaker extraction model, featuring three innovative modules: 1) A gated cross-attention fusion mechanism that establishes cross-modal correlations between EEG and speech through bidirectional cross-attention, while adaptively modulating modal contributions via dynamic gating weights to overcome the information interaction limitations of conventional linear fusion; 2) An E-DPRNN network incorporating Conv-U modules within the dual-path architecture to reinforce local short-term feature capture, synergized with residual connections to enhance gradient propagation efficiency; 3) A single-path global modulation (SPGM) block that replaces traditional inter-block modeling with parameter-free pooling and feature modulation, preserves global semantic modeling capability while reducing parameters. These modules collectively achieve efficient multimodal fusion and precise speaker extraction. Experimental validation demonstrates that the proposed method attains average SI-SDR scores of 13.59 dB, 10.44 dB, and 10.17 dB on the Cocktail Party, AVED, and MM-AAD datasets respectively, outperforming the baseline model MSFNet by 5.4%−8.2% while reducing parameters by 49.7%−55.6%. These results substantiate the model's application potential and effectiveness in neuro-oriented speaker extraction.

Key words neuro-oriented speaker extraction; EEG signals; multimodal fusion; gated cross-attention; speech separation; end-to-end learning

收稿日期: 2025-06-09; **修回日期**: 2026-01-16

基金项目: 科技创新 2030-重大项目(2021ZD0201500); 国家自然科学基金项目(62571002, 62201002, 6247077204, 62322120, U21B2010); 安徽省优秀青年基金项目(2408085Y034)

This work was supported by the Science and Technology Innovation 2030 Major Project (2021ZD0201500), the National Natural Science Foundation of China (62571002, 62201002, 6247077204, 62322120, U21B2010), and the Excellent Young Scientists Fund of the Natural Science Foundation of Anhui Province (2408085Y034).

通信作者: 吕钊(kjlz@163.com)

摘　要　神经导向说话人提取是一种模拟人类听觉注意机制的智能语音处理技术，通过解码听者脑电信号（electroencephalography, EEG）中的注意力指向，从混合语音中提取目标说话人语音。该技术为解决鸡尾酒会问题提供了新的解决思路，对开发智能助听设备具有重要意义。然而，现有方法面临多模态特征融合效率低、局部语音特征提取不足、模型计算复杂度高等挑战，制约了实际应用效果。提出新型端到端时域神经导向说话人提取模型 GALENet，其创新性体现在 3 个核心模块：1）门控交叉注意力融合机制，通过双向交叉注意力建模 EEG 与语音的跨模态关联，引入动态门控权重自适应调节模态贡献，解决传统线性融合的信息交互不足问题；2）E-DPRNN 网络，在双路径架构中嵌入 Conv-U 模块强化局部短时特征捕捉能力，配合残差连接提升梯度传递效率；3）单路径全局调制块（single-path global modulation, SPGM），采用无参数池化与特征调制替代传统块间建模，在减少参数量的同时保持全局语义建模能力。这 3 个模块协同实现了高效的多模态特征融合与精准的说话人提取。实验验证表明，该方法在 Cocktail Party, AVED, MM-AAD 这 3 个数据集上的平均 SI-SDR 分别达到 13.59 dB, 10.44 dB, 10.17 dB，相较基线模型 MSFNet 性能提升 5.4%~8.2%，参数量减少 49.7%~55.6%，验证了该模型在神经导向说话人提取中的应用潜力与有效性。

关键词　神经导向说话人提取；EEG 信号；多模态融合；门控交叉注意力；语音分离；端到端学习

中图法分类号　TP391

DOI: 10.7544/issn1000-1239.202550467　　**CSTR:** 32373.14.issn1000-1239.202550467

在复杂声学环境中，人类大脑会展现出卓越的听觉注意选择能力，即当多个说话人同时发声时，听者能通过选择性注意机制从混合语音中准确筛选出感兴趣的目标说话人，这一现象被称为“鸡尾酒会效应”[1]。其神经机制最早可追溯至 Cherry[2] 于 1953 年的双耳分听实验，该研究证实大脑会对专注侧的语音进行深度处理，而抑制非注意侧的干扰信号。然而，听力受损人群会因听觉系统功能障碍，难以在嘈杂环境中准确分辨出目标说话人，导致社交沟通受阻。为解决这一困难，说话人提取技术通过信号处理手段从混合声源中提取目标说话人语音，已广泛应用于助听设备和语音处理系统[3-5]。但传统的说话人提取方法往往缺乏对人类选择性听觉注意机制的建模能力，在多说话人环境中面临环境干扰，进而难以准确分辨并提取目标说话人的语音。而神经科学研究发现，听者的注意力指向可通过大脑活动的脑电信号（electroencephalography, EEG）解码[6-8]，这为解决“鸡尾酒会”问题提供了新思路——通过对 EEG 中的听觉注意信息进行听觉注意解码（auditory attention detection, AAD）[9]，建立混合语音与脑神经活动间的映射关系，实现基于主观注意指向的说话人提取。

当前目标说话人提取主流方法中比较高效的有 BASEN 模型[10]、NeuroHeed 网络[11] 和 MSFNet 模型[12]，前两者采用时序卷积网络（temporal convolutional network, TCN）[13] 和注意力机制的组合对语音序列进行有效建模，实现 EEG 与语音的时序对齐和特征交互，增强对目标说话人提取的针对性；后者采用双路径递归神经网络（dual-path recurrent neural network, DPRNN）[14] 对语音序列进行降维，通过分块建模策略实现对时序特征的有效提取。但是，这些目标说话人提取方法仍受制于 2 个关键缺陷：首先，在局部精细特征提取方面，基于递归机制的建模过程易导致语音瞬时特征的过度平滑化，如瞬时变化、细节纹理等。原因在于时间步间的输出存在递归依赖机制，这会削弱高分辨率时间细节的保持能力，同时梯度传播过程中的衰减效应进一步限制了长时间上下文建模的有效性。其次，全局建模效率方面，Subakan 等人[15] 通过系统的实验研究发现，DPRNN 用于建模全局关系的块间模块（inter-block）占模型整体近一半的参数量，但对整体性能提升的贡献非常有限。再者用于多模态特征交互的交叉注意力融合模块本身参数量也非常庞大，十分影响长序列语音处理的实时性。上述缺陷共同制约了语音波形端到端语音分离系统的性能，使得长序列语音信号处理成为当前技术的瓶颈。

为应对上述问题，本文在基线模型 MSFNet 基础上，提出了融合门控注意力与局部特征增强的神经导向说话人提取网络（gated attention and local feature enhancement neuro-orientedspeaker extraction network, GALENet），以解决当前研究方法面对长序列语音时无法在保证轻量化的前提下有效结合脑电信号进行建模提取特征的问题。该模型通过引入构建 3 个创

新性模块来协同优化实现性能突破：门控机制确保跨模态信息的对齐融合，E-DPRNN 提升对长序列语音的特征提取能力，SPGM 模块优化全局信息建模的效率。为全面验证方法的有效性，本研究在 3 个多模态听觉注意解码数据集（cocktail-party，AVED，MM-AAD）上开展对比实验，并复现多个主流基线模型作为对照，系统地评估 GALENet 在神经导向说话人提取任务中的优势。实验结果表明，在这 3 个公开数据集上我们提出的模型在各项指标上均优于基线模型和其他方法，同时可训练参数量相较基线模型降低 50%，为神经导向说话人提取技术的实用化提供了新的解决方案。

本文提出用于神经导向的目标说话人提取模型具体设计的 3 个核心创新点如下：1）门控交叉注意力机制，针对脑电信号与语音信号在时序动态和语义表征层面的异构性，提出双向跨模态交互范式。不同于传统线性投影的静态融合方式，该机制通过建立语音主导与脑电主导的双向注意力通路，实现跨模态特征的互补性重构。引入的动态权重分配机制突破固定比例融合的局限性，使网络能够依据语音场景复杂度自主调节模态贡献度，在抑制噪声干扰的同时强化目标说话人的声纹特征。2）E-DPRNN 架构，针对语音信号短时瞬态特征易被递归网络平滑化的问题，创新性地将深度可分离卷积与扩展线性层相结合的 Conv-U 单元嵌入残差序列中。通过深度卷积路径捕获帧级别的声学细节，配合扩展线性层的非线性映射能力，形成局部-全局特征的双向增强机制，有效缓解梯度消失问题，使网络在长语音处理中仍能保持稳定的特征传播。3）单路径全局调制模块，针对 DPRNN 中对分离性能贡献较小的块间处理网络，提出使用单路径全局调制模块（single-path global modulation，SPGM）替代传统 inter-BLSTM 模块。SPGM 模块由无参数的全局池化层与线性调制子模块构成，在确保整体语音分离性能提升的同时，也有效降低了块间信息在网络中建模传递的计算复杂度。

1 相关工作

说话人提取技术[16]作为解决“鸡尾酒会”问题的关键第 1 步，旨在从复杂声学环境中的混合语音信号中分离出每个说话人的纯净语音，使每个声源都能够被独立处理[17]。目标说话人提取[18]方法通过引入辅助参考信息，仅提取目标说话人单一特定的语音通道，避免了传统语音分离任务——盲源分离（blind signal separation，BSS）中常见的通道排序歧义、需预先设定说话人数等问题，显著降低了系统复杂度，因此，成为近年来国内外语音处理领域的研究热点[19]，受到广泛关注。而基于脑电信号，通过解码听者的 EEG 信号中的听觉注意模式、结合混合语音特征进行多模态融合[20-21]、提取听者关注的目标语音的此类方法则被称为神经导向说话人提取[22]，融合了脑机接口（brain computer interface，BCI）与语音处理技术[23]，是近年来兴起的前沿交叉研究方向，也是本文的研究内容。此类研究方向的理论依据源自于认知听觉神经科学的研究发现：当听者在特定环境中集中注意力时，其所关注的语音信号会在大脑中诱发出显著相关的神经活动模式[24-25]。

Han 等人[26]最早提出将盲源分离与听觉注意解码相结合的策略，从脑电信号中估计目标语音包络，并与各分离声源进行匹配识别。然而，该方法需要先分离出全部声源，计算成本较高。为降低复杂度，Ceolini 等人[27]提出了一种基于频域的脑启发说话人提取模型 BISS，结合 AAD 进行说话人选择与语音提取，允许神经信号直接引导单通道语音分离过程。然而，BISS 需分别训练指导网络与提取网络，并受限于频域方法固有的相位估计误差[28]。为克服上述不足，Hosseini 等人[29-30]提出 2 种端到端时域脑控语音增强方法（BESD 和 UBESD），在 EEG 与混合语音之间引入特征层面线性调制（feature-wise linear modulation，FiLM）[31]进行多模态融合，从特征层面动态调整网络响应，从而更有效地提取目标说话人的特征表示。为进一步提升跨模态特征融合效果，Zhang 等人[10]在 BASEN 模型中引入了卷积式多层交叉注意力模块（CMCA），基于 TCN 结构同时建模语音与 EEG，显著改善分离语音的清晰度和可懂度。Pan 等人[11]则提出了一种自回归说话人编码网络 Neuro-Heed，通过建模 EEG 与注意语音之间的时间关联，将历史语音动态纳入“听觉吸引子”，实现在线语音推理，模拟真实听觉过程中注意力的逐步累积。为解决神经导向语音增强中常见的说话人混淆问题，Pan 等人[32]在后续工作中进一步优化了 EEG 编码器的判别能力，并增强了语音提取网络对 EEG 与混合语音中目标语音之间映射关系的建模能力，从而提升系统的准确性和实用性。为突破单尺度特征融合的局限性，Fan 等人[12]提出的 MSFNet 创新性地构建多分支融合架构，在语音编码阶段并行采用不同时间窗长的卷积核，获取涵盖语音短时瞬态特征与长时全局模式的多尺度嵌入表示。通过多层交叉注意力模

块将 EEG 空间特征动态注入各尺度语音表示中，并采用 DPRNN 时序建模增强跨模态对齐能力。相较于 BASEN 的固定尺度融合策略，该方法可以降低有效说话人混淆错误率的同时提高推理速度。

本文提出 GALENet 模型，对现有的神经导向说话人提取方法在多模态融合和捕捉语音的短时瞬态特征与局部声学细节存在的局限进行完善，通过实现 EEG-语音的深度特征交互，优化网络架构在保证跨模态语义对齐与全局建模能力的同时，显著提升了特征表达效率与模型轻量化水平。

2 提出的方法

2.1 模型整体架构

本文提出的 GALENet 模型在整体架构上借鉴了时域多模态说话人提取方法 MSFNet 的设计理念，保留了其高效的 EEG 编码器结构，并在单尺度网络框架下进行了进一步优化。GALENet 主要由 5 个功能模块构成：语音编码器、EEG 编码器、注意力融合模块、说话人提取网络和语音解码器。整体框架如图 1 所示。

给定一段多说话人混合语音信号 $\boldsymbol{x}(t)$ 及同步采集的 EEG 脑电信号 $\boldsymbol{e}$，GALENet 的处理流程如下：

首先，语音编码器将时域混合语音信号 $\boldsymbol{x}(t)$ 编码为高维语音嵌入序列 $\boldsymbol{X}(t)$，以提取丰富的语音特征表示，可表示为

$$\boldsymbol{X} = SpeechEncoder(\boldsymbol{x}) \text{ 。} \tag{1}$$

EEG 分支利用高效的 EEG 编码器，将所采集记录的 EEG 信号 $\boldsymbol{e}$ 编码为 EEG 嵌入向量 $\boldsymbol{E}$ 以提取脑信号模态的特征信息，将其表述为

$$\boldsymbol{E} = EEGEncoder(\boldsymbol{e}) \text{ 。} \tag{2}$$

分别在语音嵌入 $\boldsymbol{X}$ 与脑电嵌入 $\boldsymbol{E}$ 上应用交叉注意力（cross-attention）机制，实现模态间的信息交互与补充，构建潜在共享的表征空间。随后，引入门控注意力（gated attention）融合模块，根据语音与脑电模态的重要性而进行动态加权，最终得到融合特征 $\boldsymbol{Z}$，这个过程可表示为

$$\boldsymbol{Z} = GatedAttention(CrossAttention(\boldsymbol{X},\boldsymbol{E})) \text{ 。} \tag{3}$$

将融合特征 $\boldsymbol{Z}$ 输入所提出的 E-DPRNN 模块，并通过 R 次迭代，有效提取短时局部与长时全局的目标语音信息，学习到源特定掩码 $\boldsymbol{M}$，用于分离目标说话人的语音特征：

$$\boldsymbol{M} = Extractor(\boldsymbol{Z}) \text{ 。} \tag{4}$$

最后，通过源特定掩码 $\boldsymbol{M}$ 对原始语音嵌入 $\boldsymbol{X}$ 进行筛选，得到掩蔽的语音嵌入 $\hat{\boldsymbol{S}}$，以此重构时域语音波形，完成波形到波形预测任务，表示为

$$\hat{\boldsymbol{s}} = Decoder(\boldsymbol{X} \otimes \boldsymbol{M}) \text{ 。} \tag{5}$$

2.2 语音编解码器

语音编码器由 1 个 1 维卷积层和 1 个 ReLU 激活函数构成，用于从原始时域语音信号中提取帧级别的高维特征。具体而言，对于输入的长时序语音样

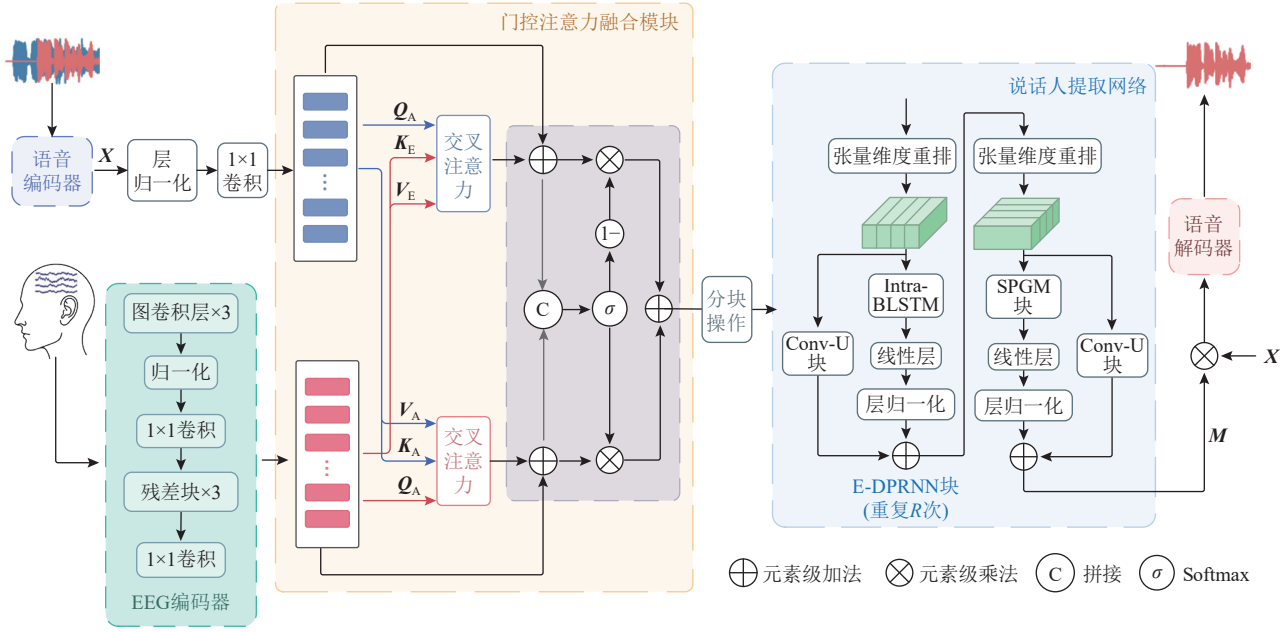


Fig. 1 Overall block diagram of the proposed GALENet model

图 1 所提出的 GALENet 模型整体框图

本$\boldsymbol{x}(t)\in\mathbb{R}^{1\times T}$，编码器首先将其切分为长度为$T_t$的短时语音片段，并通过1维卷积操作将其转换为潜在空间中类似频谱的帧级嵌入序列表示$\boldsymbol{X}(t)$，如式(6)所示。

$$\boldsymbol{X}(t)=ReLU\left(conv1D\left(\boldsymbol{x}(t),1,N_t,L\right)\right)\in\mathbb{R}^{N_t\times T_t}, \tag{6}$$

其中1维卷积的输入通道数量为1，输出通道数量为N_t，卷积核大小为L，步长设为$L/2$，以保证帧间重叠。输出的$\boldsymbol{X}(t)$为具有N_t维特征的T_t帧嵌入序列，用于表征语音的时频特性。为了降低计算复杂度并加深网络结构的深度，在后续处理过程中不直接使用$\boldsymbol{X}\in\mathbb{R}^{N_t\times T_t}$，而是通过一个1×1的逐点卷积将通道数由$N_t$映射至更小的瓶颈特征维度$C$，得到更紧凑的表征表示$\boldsymbol{X}\in\mathbb{R}^{C\times T_t}$，其中$N_t>L$。这种设计在保持语音重要特征的同时，有效降低了计算复杂度。

在说话人提取模块中，网络会生成1个与语音特征维度一致的掩码$\boldsymbol{M}\in\mathbb{R}^{N_t\times T_t}$，用于从混合语音特征$\boldsymbol{X}$中筛选出目标说话人的特征表示$\hat{\boldsymbol{S}}$。最后，语音解码器执行编码器的逆操作，通过将与编码器结构对称的1维转置卷积层还原为时域上的目标语音波形信号$\hat{\boldsymbol{s}}\in\mathbb{R}^{1\times T}$。该解码器的卷积核大小和步长仍设为$L$和$L/2$，确保输出与原始输入的时间序列长度对齐：

$$\hat{\boldsymbol{s}}(t)=Deconv1D\left(\hat{\boldsymbol{S}}(t),N_t,1,L\right)\in\mathbb{R}^{1\times T}。 \tag{7}$$

2.3 EEG编码器

EEG编码器旨在从与目标语音时序对齐的多个通道EEG信号中提取高区分度嵌入表征。如图2所示，编码器前端采用3层图卷积网络(graph convolutional network, GCN)建模脑电通道间空间依赖关系[33]：将电极抽象为图节点，通过邻接矩阵定义通道拓扑连接关系，采用对称归一化拉普拉斯矩阵实现图卷积运算。GCN通过多层特征传播机制聚合相邻节点信息，利用切比雪夫多项式简化卷积核计算，有效捕获与听觉注意相关的空间模式。

经GCN提取的时空特征经通道层归一化后，通过1×1卷积进行初步维度压缩，随后输入到由3个残差块构成的深度特征提取模块。如图3所示，每个ResNet块包含2个1×1卷积层、1个卷积核大小为1×3的最大池化层、2个批归一化(batch normalization, BN)层和1个参数化修正线性单元PReLU(parametric ReLU)激活函数。2个1×1卷积层分别用于对输入特征进行升维和降维处理，用于匹配特征维度和减少模型参数，随后引入批归一化层进行归一化处理，稳定训练；1×3最大池化实现3倍时域下采样以强化关键信息并抑制冗余特征；跳跃连接通过元素相加融合浅层细节与深层语义，实现残差学习以缓解梯度消失问题。PReLU激活函数引入可学习参数提高网络的非线性变换的灵活性和特征表达能力。

最终，经过ResNet模块处理的特征将通过1个1×1卷积层映射到预设的固定维度，生成下采样后的EEG嵌入表示$\boldsymbol{E}$。整个过程不仅实现了对脑电信号时序特征的有效压缩，也保留了重要的时域结构

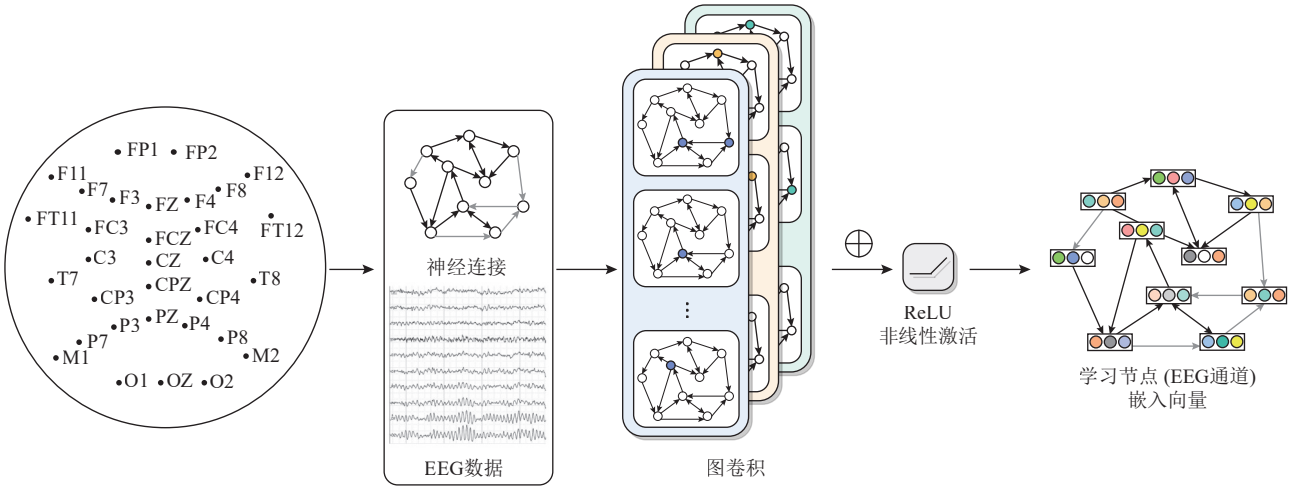


Fig. 2 Basic principles of multi-layer GCN

图2 多层GCN基本原理

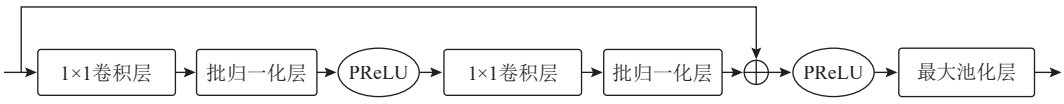


Fig. 3 Composition structure of ResNet module

图3 ResNet模块组成结构

信息，为后续的目标说话人提取模块提供了高质量的脑电特征支撑。

2.4 门控注意力融合模块

在完成语音和 EEG 这 2 个单模态的独立特征提取之后，本文引入多头交叉注意力（multi-head cross attention, MCA）机制，以建模 EEG 模态和音频模态之间的语义依赖关系，并在跨模态的交互之后通过非线性变换进一步挖掘更具判别性的特征。多头交叉注意力结构由 1 个线性全连接层与多个注意力头组成，注意力头数量设为 A_S。该模块的核心思想是以目标模态 m_1 为主导，从源模态 m_2 中引入需要对目标模态进行补充的信息，实现信息从 m_2 到 m_1 的跨模态流动与融合，得到信息增强的特征表示 $\boldsymbol{Z}_{m_2\to m_1}$，从而增强目标模态的特征表达能力。其具体如式（8）所示。

$$MCA\left(\boldsymbol{Z}_{m_1},\boldsymbol{Z}_{m_2}\right)=Concat\left(head_1,head_2,\cdots,head_{A_S}\right)\boldsymbol{W}\text{。} \tag{8}$$

每个头的注意力计算公式如式（9）（10）所示。

$$head_i=Softmax\left(\frac{\boldsymbol{Q}_i^{m_1}\times(\boldsymbol{K}_i^{m_2})^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V}_i^{m_2},\ i=1,2,\cdots,A_S\text{，} \tag{9}$$

$$\begin{cases}\boldsymbol{Q}_i^{\mathrm{m}}=\boldsymbol{Z}_{\mathrm{m}}\times\boldsymbol{W}_i^{\boldsymbol{Q}}\text{，}\\ \boldsymbol{K}_i^{\mathrm{m}}=\boldsymbol{Z}_{\mathrm{m}}\times\boldsymbol{W}_i^{\boldsymbol{K}}\text{，}\\ \boldsymbol{V}_i^{\mathrm{m}}=\boldsymbol{Z}_{\mathrm{m}}\times\boldsymbol{W}_i^{\boldsymbol{V}}\text{，}\end{cases} \tag{10}$$

其中 $\boldsymbol{W}$，$\boldsymbol{W}_i^{\boldsymbol{Q}}$，$\boldsymbol{W}_i^{\boldsymbol{K}}$，$\boldsymbol{W}_i^{\boldsymbol{V}}$ 均为可学习的参数矩阵。

本文中的 MCA 在 2 条特征路径上分别进行交互建模：

1）语音主导路径（EEG 为辅助模态）。当目标模态为语音特征 $\boldsymbol{A}$ 且源模态为 EEG 特征 $\boldsymbol{E}$ 时，MCA 模块的输出表示如式（11）所示。

$$\boldsymbol{Z}_{E\to A}=FC(MCA(\boldsymbol{A},\boldsymbol{E}))+\boldsymbol{A}\text{。} \tag{11}$$

2）EEG 主导路径（语音为辅助模态）。当目标模态为 EEG 特征 $\boldsymbol{E}$ 且源模态为语音特征 $\boldsymbol{A}$ 时，MCA 模块的输出表示如式（12）所示。

$$\boldsymbol{Z}_{A\to E}=FC(MCA(\boldsymbol{A},\boldsymbol{E}))+\boldsymbol{E}\text{。} \tag{12}$$

经过上述双路径交叉注意力机制后，分别得到 2 组初步融合的特征表示 $\boldsymbol{Z}_{E\to A}$ 与 $\boldsymbol{Z}_{A\to E}$，其中均融合了对方模态的上下文信息。

为了进一步提升模态融合的灵活性与有效性，本文创新型地引入了一种门控注意力（gated attention）机制，用于自适应地调整 2 个模态在最终融合中的权重。该机制首先将 $\boldsymbol{Z}_{E\to A}$ 和 $\boldsymbol{Z}_{A\to E}$ 进行拼接，并输入至线性变换层和 Tanh 激活函数，获得原始注意力得分，随后通过 Softmax 归一化生成融合权重，如式（13）所示。

$$[\alpha,1-\alpha]=Softmax(Tanh(FC(\boldsymbol{Z}_{E\to A},\boldsymbol{Z}_{A\to E})))\text{。} \tag{13}$$

最终，融合后的特征表示 $\boldsymbol{Z}$ 可由以下加权求和得到：

$$\boldsymbol{Z}=\alpha\boldsymbol{Z}_{E\to A}+(1-\alpha)\boldsymbol{Z}_{A\to E}\text{。} \tag{14}$$

这种交叉注意力与门控机制的联合设计，突破了固定模态权重融合的限制，使网络能够根据任务需求动态调节各模态的贡献比例，增强关键信息的表达能力，有效抑制冗余特征，从而显著提升神经导向说话人提取的性能。

2.5 E-DPRNN 网络

本文创新型地提出了一种基于 E-DPRNN 模块的新型说话人提取网络架构，旨在更高效地对语音信号进行时序建模与特征融合。该架构的核心思想是将长时间的语音序列划分为若干小块，并分别在块内（intra-chunk）和块间（inter-chunk）2 个层面进行特征提取，从而兼顾局部信息的精细建模与全局语义的整合。

在块内特征提取阶段，采用了双分支结构以提升局部特征的表达能力。一条分支引入 Conv-U 模块，通过卷积操作捕捉每个块内的短时动态特征，增强局部结构信息的感知能力；另一条分支继承了 DPRNN 中的 intra-BLSTM 设计，利用双向 LSTM 模型挖掘更长时间跨度内的时序依赖，提升块内语义建模能力。2 条路径的输出在特征维度上进行相加操作，以获得更精细化的块内特征表示。在块间特征提取阶段，网络同样采用双路径设计，以平衡建模能力与计算效率。一条路径继续使用 Conv-U 模块，用于捕捉语音信号在块与块之间的动态变化，强化全局时间结构的建模；另一条路径引入了轻量级的单路径全局调制块（single-path global modulation, SPGM）模块，通过低参数设计有效提取块间的全局语义信息，从而减少冗余计算并提升整体建模效率。通过上述设计，网络能够在局部与全局 2 个层次上实现对语音信号的多粒度建模，既保证了说话人特征的精细提取，又控制了模型的参数规模与计算复杂度，达成性能与效率的良好平衡。

2.5.1 Conv-U 模块

为在保留 BLSTM 全局建模能力的基础上进一步增强局部特征的刻画能力，并缓解训练过程中的梯度消失或爆炸问题，本文在 E-DPRNN 框架中引入了 Zhao 等人[34-35]提出的 Conv-U 模块，如图 4 所示。该模块首先将输入序列经过层归一化（LayerNorm）

层进行标准化处理,以提升特征分布的一致性;接着,通过带有扩展因子的线性变换层,结合 SiLU(sigmoid linear unit)激活函数,实现对特征的非线性映射,从而提升模型的表达能力;随后,使用深度 1 维卷积(depthwise 1D convolution)提取序列中的短时局部模式,充分捕捉语音中的细粒度短时动态信息。同时为增强模型的鲁棒性与训练稳定性,模块引入跳跃连接机制并叠加 Dropout 层进行正则化,有效防止过拟合并加快收敛速度。

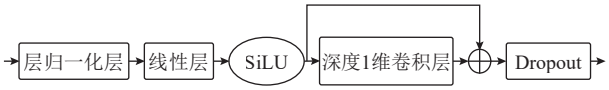


Fig. 4 Composition structure of Conv-U module

图 4 Conv-U 模块组成结构

2.5.2 单路径全局调制块 SPGM

Subakan 等人[36]针对 Sepformer 模型中跨块建模效率问题展开研究,发现 inter-block 结构虽占据模型近半参数量,但对性能提升贡献有限,表明全局信息建模需求可通过更高效机制实现。基于此,本文在 GALENet 模型中引入 Yip 等人[37]提出的 SPGM 模块,结构如图 5 所示,能通过全局池化与特征调制的方式有效建模块间信息,同时减少计算复杂度。该模块由全局池化模块和特征调制模块 2 个核心子模块组成。

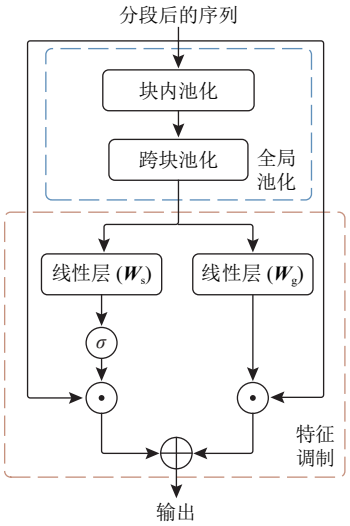


Fig. 5 Composition structure of SPGM module

图 5 SPGM 模块组成结构

全局池化模块采用层级压缩策略,首先通过最后元素选择法(last element selection, LE)进行块内池化(intra block pooling),直接选取各时间块末位时间步特征作为块级向量。由于 E-DPRNN 采用 50% 重叠分块策略,末位元素在相邻块间重复出现,可以将其重新用作全局向量的代表,无需额外的可训练参数即可实现有效建模。块池化将输入特征维度由 $[K,S,N]$压缩至$[1,S,N]$,其中K为块长,S为块数,N为特征维度。随后执行跨块池化(inter block pooling),沿块维度S对各通道特征进行平均池化操作,最终生成一个维度大小为$[1,1,N]$的全局向量。这个全局向量随后作为输入传递给 SPGM 的调制模块。该方法确保所有块在最终全局表示中具有相同的权重,使得全局向量能够综合所有块的信息,而不会偏向某个特定的块。

特征调制模块基于特征线性调制思想,将全局向量与局部特征动态融合。输入包含块内 BLSTM 提取的局部特征$\boldsymbol{x}_{\mathrm{f}} \in \mathbb{R}^{K\times S\times N}$和通过全局池化模块生成的全局向量$\boldsymbol{x}_{\mathrm{ebm}} \in \mathbb{R}^{1\times 1\times N}$。模块设计双路径融合机制:1)线性路径通过可训练矩阵$\boldsymbol{W}_{\mathrm{g}} \in \mathbb{R}^{N\times N}$对$\boldsymbol{x}_{\mathrm{ebm}}$进行线性变换映射后,与$\boldsymbol{x}_{\mathrm{f}}$执行逐元素相乘,实现特征通道的全局缩放;2)非线性路径利用$\boldsymbol{W}_{\mathrm{s}} \in \mathbb{R}^{N\times N}$矩阵与 Sigmoid 函数生成门控权重,再与$\boldsymbol{x}_{\mathrm{f}}$相乘,动态增强关键局部信息。2 条路径输出经加和运算得到调制后特征$\boldsymbol{x}_0 \in \mathbb{R}^{R\times S\times N}$,作为后续处理的输入。

3 实验设置

为了全面验证所提出的 GALENet 模型的有效性,本文在 3 个数据集 Cocktail Party, AVED, MM-AAD 上进行了广泛的对比实验。表 1 总结了所有实验中使用的数据集的设置和参数。

Table 1 Summary of Detailed Settings for Each Dataset

表 1 各个数据集的详细设置总结

参数	Cocktail Party 数据集	自采 AVED 数据集	MM-AAD 数据集
受试者数量	33	20	50
受试者性别	28 名男性、5 名女性	14 名男性、6 名女性	34 名男性、16 名女性
演讲者性别	不同男性	男性和女性	男性和女性
EEG 采集通道数	128	32	32
每个受试者参与试验次数	30	16	20
每个 EEG 试验持续时间/s	60	152	165
演讲语言	英语	普通话	普通话

3.1 Cocktail Party 数据集

2018 年 Broderick 等人[38]开展了一项基于鸡尾酒会问题的听觉注意机制研究。实验招募 33 名听力正常受试者(28 名男性及 5 名女性,平均年龄 27.3 岁),

采用 128 通道 BioSemi ActiveTwo 系统以 512 Hz 采样率采集 EEG 数据，每名完成 30 次 60 s 的试验。双耳听觉刺激为 2 名男性朗读的独立英文小说片段（Sennheiser HD650 耳机播放，44.1 kHz 采样率），要求受试者专注指定耳侧内容并忽略对侧语音。受试者分为 2 组，17 人持续关注左耳，17 人持续关注右耳（编号 6 被试因噪声干扰被剔除）。为确保试验连续性，各次试验的音频内容首尾衔接，避免故事重复。每次试验后通过 4~6 项选择题评估注意力集中程度。

实验采用严格参数控制，音频经幅度归一化处理至统一均方根（RMS）水平，静音间隔不超过 0.5 s；要求受试者注视屏幕中央十字准线以抑制运动伪影。数据处理阶段，原始音频下采样至 14.7 kHz 并执行 RMS 归一化，混合信号作为神经网络输入。EEG 信号经预处理后统一降采样至 128 Hz。最终数据划分策略为：每位被试随机抽取 5 次试验构成测试集，其中 2 次作为验证集，其他用于训练集构建。

3.2 AVED 数据集

AVED 数据集是由张宏玉等人[39]于 2024 年创建的新型脑电-音频数据集，该数据集专为听觉注意解码任务设计，所有受试者均签署知情同意书。实验招募 20 名听力正常受试者（14 名男性和 6 名女性，平均年龄 20 岁），每名完成 16 次持续 152 s 的试验（含实验初始 2 s 静默时段）。听觉刺激选自中文短篇小说集，每次试验同步播放男女讲述者朗读的不同故事，要求受试者依据指令定向注意且确保关注故事无内容重叠。每次试验后通过 3 项多选题来验证注意力集中程度。音频信号统一设定为 44.1 kHz 采样率并保持相同响度水平。32 通道脑电信号以 1 kHz 原始采样率采集后降采样至 128 Hz。为避免注意力切换引起的实验不稳定性，仅选取关注男性讲述者的试验数据用于评估。数据划分采用 75% 训练集、12.5% 验证集与 12.5% 测试集的标准化比例。

3.3 MM-AAD 数据集

MM-AAD（multimodal auditory attention decoding）数据集是由 Fan 等人[40]于 2025 年提出的一项最新研究成果。该数据集包含来自 50 名正常听力、无神经系统疾病的在校大学生（34 名男性和 16 名女性）的 EEG 数据。每名受试者参与了 20 次试验，每次试验持续约 165 s，受试者平均年龄为 22.4 岁。所有实验均遵循伦理规范，确保受试者在知情同意的前提下自愿参与。

MM-AAD 使用 32 通道 EEG 设备以 2 048 Hz 的原始采样率进行记录，并将数据下采样至 128 Hz，以提高处理效率。混合音频刺激信号的采样率为 44.1 kHz，实验采用 E-prime 软件实现音频/音视频刺激与 EEG 采集设备之间的精确同步。与 AVED 数据集类似，MM-AAD 同样设有音频刺激与音视频联合刺激 2 种试验条件。所有受试者均在这 2 种条件下完成了实验采集，而本研究中仅使用音频刺激条件下的数据进行分析与建模。

在实验过程中，受试者需在每个试验中交替将注意力集中于左右耳不同的说话人。考虑到注意目标频繁切换可能导致模型训练不稳定，为提升训练质量，本文仅选取受试者在注意男性说话人情境下的数据进行训练与评估，与 AVED 数据集的设置保持一致。最终，针对每名受试者选取的 10 个有效试验，按照如下比例划分数据集：2 个试验用于测试集，1 个用于验证集，其余 7 个用于训练集。

3.4 脑电信号预处理

本研究为 EEG 信号构建了一套系统的预处理框架，旨在高效提取神经活动特征。针对 Cocktail Party 数据集，在进行 0.1~45 Hz 带通滤波并应用独立成分分析（ICA）技术[41]抑制眼动、肌电等生理伪影后，还引入了频带耦合模型[42]。该模型通过联合解析高频 γ 波段的功率谱与低频 δ 波段的相位动态，估算出与听觉皮层神经放电高度相关的多单元神经活动（MUA）特征，并以此作为模型的最终 EEG 输入，显著增强了信号与语音刺激的神经关联性。对于 AVED 数据集，预处理的重点是采用 50 Hz 陷波滤波器消除工频干扰，并结合 ICA 分离残留伪影。而对于 MM-AAD 数据集，则主要采用带通滤波和通道级 Z-score 归一化来提升数据的稳定性。

最后，为确保 EEG 与语音特征能在同一时间尺度上进行融合，研究对 2 类数据的采样率进行了调整：对语音信号进行降采样以降低计算复杂度，同时对 EEG 信号进行上采样，最终实现了多模态特征的精确时序对齐，为构建后续的神经解码模型奠定了坚实基础。

3.5 实验环境设置

GALENet 模型的实现基于 PyTorch 深度学习框架，构建环境为 Python3.8.12，CUDA 版本为 11.3，并在 NVIDIA 3090Ti GPU 上完成训练任务。模型训练总共进行 60 个 epoch，初始学习率设定为 8E−4，批量大小为 8。为提升训练效率与稳定性，采用线性预热与余弦退火策略对学习率进行动态调节。优化器选用 Adam（adaptive moment estimation）优化算法[43]，并使用负 SI-SDR（scale-invariant signal-to-distortion ratio）

作为损失函数，用于衡量语音重构质量并进行反向传播优化。在多数据集适配方面，不同 EEG 数据集对应的编码器输入通道数略有差异，Cocktail Party 数据集使用 128 通道 EEG 输入；而 AVED 与 MM-AAD 数据集则使用 32 通道 EEG 输入。为缓解不同数据集带来的过拟合风险以及抑制训练过程中的梯度爆炸问题，实验设置了针对性权重衰减参数：AVED 数据集训练时设定权重衰减为 1E−2，MM-AAD 数据集则设为 5E−3，其余参数保持一致。

模型结构参数详见表 2，其中，超参数包括卷积核大小K_s、卷积步长S_t、多头交叉注意力机制中的注意头数量A_s、E-DPRNN 模块中的分段长度K与块数量S。最终模型通过合理的训练配置、有效的优化策略以及对不同数据集的适配性设计，在保障计算效率的同时，有效提升了神经导向说话人提取任务的整体性能表现。

Table 2 Overall Structure Parameters of GALENet Model

表 2 GALENet 模型整体结构参数

模块	模型层	输入尺寸	超参数	输出尺寸
语音编码器	Conv1D	（B，1，29 184）	K_s=16 S_t=8	（B，128，3 653）
EEG 编码器	GCN	（B，128/32，29 184）		（B，128/32，29 184）
	Conv1D	（B，128/32，29 184）	K_s=2 S_t=1	（B，64，29 183）
	通道层归一化	（B，64，29 183）		（B，64，29 183）
	Conv1D	（B，64，29 183）	K_s=1 S_t=1	（B，64，29 183）
	ResBlock×3	（B，64，29 183）	K_s=1 S_t=1	（B，128，1 080）
	Conv1D	（B，128，1 080）	K_s=1 S_t=1	（B，64，1 080）
注意力融合	语音 Conv1D	（B，128，3 653）	K_s=1 S_t=1	（B，64，3 653）
	EEG 线性插值	（B，64，1 080）		（B，64，3 653）
	多层交叉注意力	（B，3 653，64）	A_s=4	（B，3 653，64）
	门控注意力融合	（B，3 653，64）		（B，64，3 653）
说话人提取 E-DPRNN×5	分段	（B，64，3 653）	K=250 S=32	（B，64，250，32）
	块内 BLSTM	（B*32，64，250）		（B，64，250，32）
	块间 SPGM	（B*250，64，32）		（B，64，250，32）
	重叠相加	（B，64，250，32）		（B，64，3 653）
	Conv1D	（B，64，3 653）	K_s=1 S_t=1	（B，128，3 653）
语音解码器	Deconv1D	（B，128，3 653）	K_s=16 S_t=8	（B，1，29 232）
	切片操作	（B，1，29 232）		（B，1，29 184）

注：混合语音为 1×21 984；EEG 信号为 128×29 184（Cocktail Party）/32×29 184（AVED/MM-AAD）；B 表示实验时的 Batch 大小。

4 实验结果与分析

4.1 与基线模型的对比实验

为全面评估所提出的 GALENet 方法在神经导向说话人提取任务中的性能表现，本研究选取了目前神经导向说话人提取的主流基线模型进行对比实验，涵盖 BESD, UBESD, BASEN, MSFNet。

所有模型均在 Cocktail Party, AVED, MM-AAD 这 3 个数据集上进行训练与测试，性能评估主要基于测试集中所有受试者语音片段的 SI-SDR（中位数）指标。同时，为验证模型的轻量化效果，还对各方法的参数量进行了对比分析，以检验 GALENet 是否能在保持或提升性能的同时显著降低模型复杂度。特别说明，为了保证实验结果的公平性与可比性，表 3~5 中类似 BASEN*的形式的结果均为本研究在统一配置下的复现结果。其中，Cocktail Party 数据集采用未知目标说话人提取设定，以更贴近真实应用场景；而 AVED 与 MM-AAD 数据集由于实验范式相似，均在已知目标说话人提取条件下进行训练与推理，从而确保结果的可重复性与稳定性。

在神经导向说话人提取任务中，性能评估指标主要包括主观评价与客观评价两大类，其中客观评价可通过公式算法实现快速、可重复的定量分析，已广泛应用于本领域研究当中。常用的客观指标包括源信号失真比（signal-to-distortion ratio, SDR），其通过

Table 3 Comparison of Various Methods on Cocktail Party Dataset

表 3 在 Cocktail Party 数据集上各种方法的比较

方法	SI-SDR/dB↑	SDR/dB↑	STOI↑	ESTOI↑	PESQ↑	参数量
Mixture	0.45	0.47	0.71	0.55	1.61	
BESD[29]	5.75		0.79		1.79	0.54×10^6
UBESD[30]	8.54		0.83		1.97	1.84×10^6
BASEN*[10]	11.56	11.66	0.86	0.72	2.21	0.65×10^6
MSFNet[12]	12.89	13.03	0.88	0.77	2.51	1.53×10^6
GALENet（本文）	**13.59**	**13.74**	**0.89**	**0.78**	**2.54**	0.77×10^6

注：黑体数值表示最优值。“↑”表示该值越高越好。

衡量目标语音与失真成分之间的能量比来评估分离质量，虽效果直观但对噪声较敏感。尺度不变信号失真比（scale-invariant SDR, SI-SDR）在 SDR 基础上引入尺度因子，增强了鲁棒性与稳定性，是当前语音增强任务中的主流评估方式。短时客观可懂度（short-time objective intelligibility, STOI）指标通过分析参考语音与预测语音之间短时频谱包络的相关性来量化语音可懂度，得分范围为 [0, 1]。为进一步提升在噪声环境中的评估精度，扩展短时客观可懂度（extended short-time objective intelligibility, ESTOI）在建模过程中融入了听觉感知的非线性特性。语音质量感知评估指标（perceptual evaluation of speech quality, PESQ）则模拟人类听觉系统对语音质量的感知过程，输出与平均意见得分（mean opinion score, MOS）相关的量化结果，广泛用于评估语音增强与分离系统的感知质量表现。上述指标共同构成了神经导向语音分离系统性能评估的核心工具体系。

4.1.1 Cocktail Party 数据集上的实验结果

在模拟未知目标说话人提取的复杂语音环境下，如表 3 所示，GALENet 在主要性能指标上均取得最佳结果，SI-SDR 达到 13.59 dB, STOI 为 0.89, PESQ 为 2.54, 相较当前最优的 MSFNet 分别提升了 0.7 dB, 0.01 dB, 0.03 dB, 验证 GALENet 在多模态融合和目标语音建模方面的有效性。同时，在参数量方面，GALENet 也展现出良好的轻量化优势。与 BASEN 模型相比，GALENet 在 SI-SDR 上实现了 2.03 dB 的显著提升，而模型参数仅小幅增加 0.12×10^6。图 6 总结了 MSFNet 模型和 GALENet 模型在测试受试者上的 SI-SDR, STOI, PESQ 评估结果。可以清晰地看到，所提出的 GALENet 在所有评估指标和全部受试者中均优于 MSFNet 方法。另外，这两者的性能在不同受试者之间表现出高度的一致性，这充分表明 DPRNN 对于长序列语音特征提取的有效性和在听觉动态变化下卓越的鲁棒能力。进一步对比 MSFNet, 尽管 GALENet 性能提升幅度较小，但模型参数量从 1.53×10^6 减少至 0.77×10^6, 降幅近 50%, 显示出在性能与复杂度之间达成了更优的平衡。

4.1.2 AVED 与 MM-AAD 数据集上的实验结果比较

在已知目标说话人提取的实验设定下，本研究进一步在 AVED 和 MM-AAD 这 2 个数据集上进行了对比实验，以全面评估 GALENet 方法在不同任务条件下的泛化能力与鲁棒性。由于 EEG 输入通道数的差异，GALENet 在 Cocktail Party 与 AVED 数据集之间的模型参数量存在一定差异，但整体性能表现和趋势保持一致。如表 4 所示，在 AVED 数据集上，GALENet 在仅使用 0.63×10^6 参数量的条件下，分别在 SI-SDR 和 SDR 这 2 项指标上相较 MSFNet 模型取得了 0.79 dB 和 0.76 dB 的性能提升，进一步凸显了 GALENet 在提升语音质量和减少语音失真的能力。虽然 BASEN 模型在参数量上与 GALENet 相近，但 GALENet 在 SI-SDR 指标上实现了约 23.4% 的性能提升，验证了其更优的说话人提取能力。

在 MM-AAD 数据集上的评估结果也同样印证了 GALENet 的卓越表现。如表 5 所示，GALENet 在主要评估指标上均取得当前最优值：SI-SDR 达到 10.17 dB, STOI 为 0.91, PESQ 为 2.19。特别是在 SI-SDR 指标上，相较于 BASEN 方法提升了约 26.5%，并且相比 MSFNet 方法进一步提升了 6.3%。同时，GALENet 的模型参数量较 MSFNet 减少了 0.79×10^6, 且与 BASEN 模型保持在同一数量级，展现出在保持高性能的前提下显著压缩模型规模的能力。

为了进一步评估模型在实际部署场景中的可行性，本研究还使用 PyTorch-OpCounter 工具对训练过程中的乘积累加运算量（multiply-accumulate operations, MAC）进行了统计。结果表明，GALENet 的计算量为 47.89 GMAC, 相比于 MSFNet 的 89.58 GMAC, 计算复杂度减少近一半，表明 GALENet 不仅在性能上优于现有方法，也具备更强的资源效率与实际应用潜力。

4.2 消融实验

为系统评估 GALENet 模型中各关键组成部分（包括输入模态及关键模块）的有效性，本节设计了多组消融实验，并基于 Cocktail Party 数据集开展对比分析，实验结果详见表 6 和表 7。

首先，我们分析了独立模态的作用，测试了仅使用语音（speech-only）的提取效果。实验结果显示，模

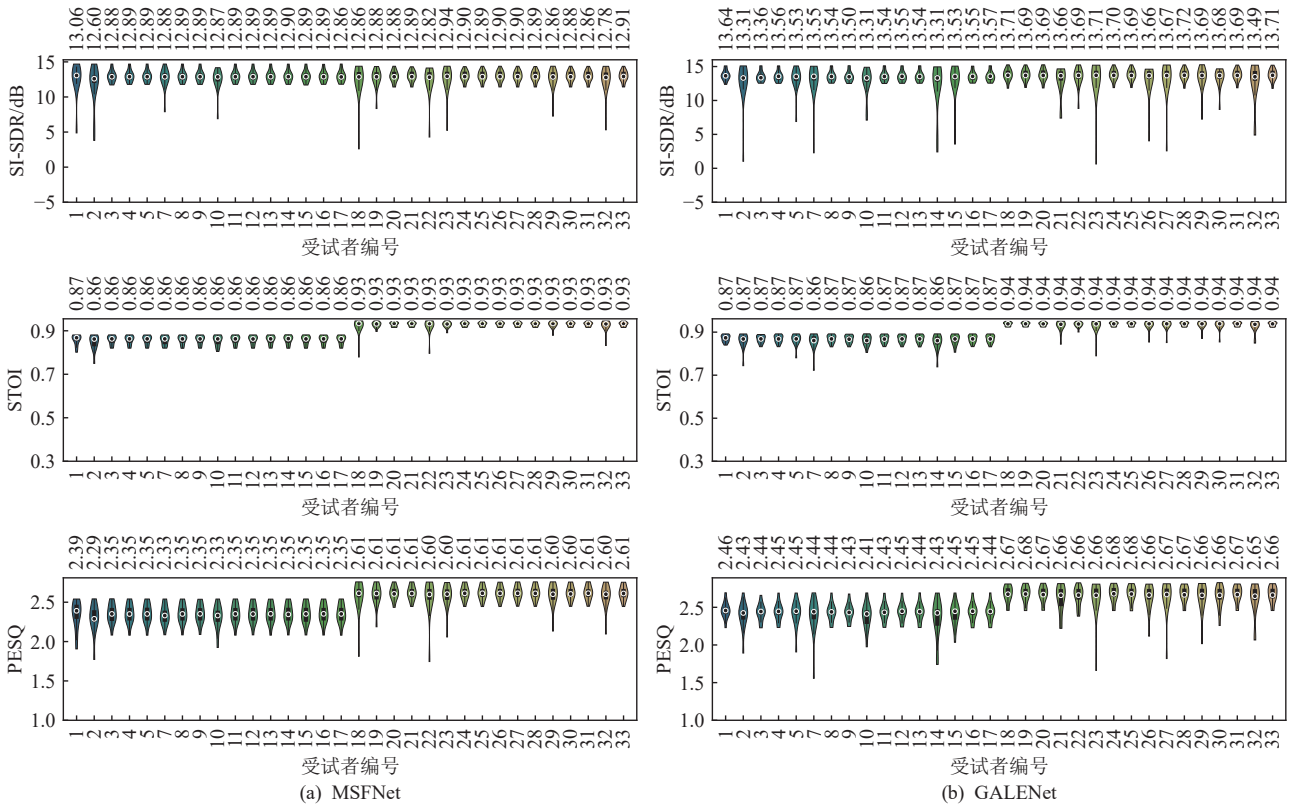


Fig. 6 Comparison of the proposed model and MSFNet method on Cocktail Party dataset

图 6 本文模型在 Cocktail Party 数据集与 MSFNet 方法的比较

Table 4 Comparison of Various Methods on AVED Dataset

表 4 在 AVED 数据集上各种方法的比较

方法	SI-SDR/dB↑	SDR/dB↑	STOI↑	ESTOI↑	PESQ↑	参数量
Mixture	1.72	1.73	0.76	0.63	1.58	
UBESD*[29]	7.89	8.10	0.85	0.72	1.75	1.69×10^6
BASEN*[30]	8.46	8.68	0.86	0.75	1.91	0.56×10^6
MSFNet[12]	9.65	9.84	0.89	0.79	2.07	1.42×10^6
GALENet（本文）	**10.44**	**10.60**	**0.90**	**0.80**	**2.08**	0.63×10^6

注：黑体数值表示最优值。“↑”表示该值越高越好。

Table 5 Comparison of Various Methods on MM-AAD Dataset

表 5 在 MM-AAD 数据集上各种方法的比较

方法	SI-SDR/dB↑	SDR/dB↑	STOI↑	ESTOI↑	PESQ↑	参数量
Mixture	2.12	2.14	0.80	0.64	1.54	
UBESD*[29]	6.29	6.60	0.84	0.70	1.66	1.69×10^6
BASEN*[30]	8.04	8.22	0.88	0.74	2.01	0.56×10^6
MSFNet*[12]	9.57	9.69	0.90	0.79	2.18	1.42×10^6
GALENet（本文）	**10.17**	**10.32**	**0.91**	**0.81**	**2.19**	0.63×10^6

注：黑体数值表示最优值。“↑”表示该值越高越好。

Table 6 Comparison of Model Performance Under Different Modal Inputs

表 6 不同模态输入下的模型性能对比

方法	SI-SDR/dB↑	SDR/dB↑	STOI↑	ESTOI↑	PESQ↑
Speech-Only	0.47	0.46	0.71	0.55	1.61
GALENet（本文）	**13.59**	**13.74**	**0.89**	**0.78**	**2.54**

注：黑体数值表示最优值。“↑”表示该值越高越好。

Table 7 Ablation Experiments of E-DPRNN Network

表 7 E-DPRNN 网络的消融实验

Conv-U模块	SPGM模块	SI-SDR/dB↑	SDR/dB↑	STOI↑	ESTOI↑	PESQ↑	参数量
×	×	13.49	13.63	**0.89**	**0.78**	2.52	1.03×10^6
×	√	13.31	13.51	**0.89**	0.77	2.47	0.72×10^6
√	√	**13.59**	**13.74**	**0.89**	**0.78**	**2.54**	0.77×10^6

注：黑体数值表示最优值。“↑”表示该值越高越好。

型性能发生显著下降，实验结果详见表 6，其 SI-SDR 指标仅为 0.47 dB，PESQ 评分也低至 1.61。该结果与融合 EEG 信号后的 GALENet 性能形成鲜明对比，这有力地证实了 EEG 作为神经引导信号在本研究中的关键指导作用，凸显了多模态融合的绝对必要性。

其次，为验证 SPGM 和 Conv-U 这 2 个模块的有效性，我们构建了一个移除 2 者的基线模型（DPRNN，

inter-BLSTM)，其 SI-SDR 为 13.49 dB，参数量大小为 1.03×10^6。在此基础上，仅引入 SPGM 来替换 Inter-BLSTM，参数量显著降至 0.72×10^6，有效验证了 SPGM 模块通过全局池化和轻量调制机制在极低参数开销下仍能维持较强的块间建模能力。最后，在已引入 SPGM 的基础上，加入 Conv-U 模块构成完整 GALENet，SI-SDR 与 SDR 指标提升至整体最优，PESQ 升至 2.54，而模型参数量仅小幅增加至 0.77×10^6。表明 Conv-U 模块能有效加强语音局部特征的建模能力，并在不显著增加模型复杂度的前提下，提升说话人提取的整体性能。

4.3 不同融合方法对模型性能的影响

表 8 展示了在 4 种不同信息融合策略下，GALENet 系统的性能对比结果，旨在评估不同模态融合机制对神经导向说话人提取效果的影响。所比较的方法包括：元素相加 Add 方法（GALENet#1）、特征拼接 Concat 方法（GALENet#2）、多层交叉注意力 CMCA 方法（GALENet#3）以及本文提出的门控交叉注意力机制（MCA+GA）。其中，Add 的具体实现是通过线性变换调整 EEG 嵌入特征维度，使其与语音特征的维度匹配后进行加法融合，保留每种模态特征信息的同时，减少维度不一致带来的问题。Concat 的具体实现是将语音特征与 EEG 特征直接拼接，形成更丰富的特征表示。从实验结果来看，Add 和 Concat 方法虽结构简单、计算开销低，但未能有效建模语音与 EEG 模态间的复杂依赖关系，导致语音分离性能受限。相比之下，CMCA 方法通过引入多层交叉注意力机制，能够更深入地挖掘跨模态特征之间的相互作用关系，显著提升系统性能。但这一方法的代价在于其引入了大量参数和计算负担，影响模型的整体效率与可部署性。为兼顾性能与效率，本研究提出的 MCA+GA 融合方式在结构上采用单层多头交叉注意力以实现特征交互，同时通过引入门控机制自适应地调节模态间的融合权重。实验结果表明，MCA+GA 方法在参数量仅为 0.77×10^6 的情况下，超越其他方法取得最优性能，不仅有效压缩了模型规模，还充分保留了跨模态深层特征的表达能力。

Table 8 Comparative Study of Different Fusion Methods

表 8 不同融合方式的对比研究

方法	融合方式	SI-SDR/dB↑	SDR/dB↑	STOI↑	ESTOI↑	PESQ↑	参数量
Mixture		0.45	0.47	0.71	0.55	1.61	
GALENet#1	Add	12.83	12.99	0.88	0.75	2.38	0.73×10^6
GALENet#2	Concat	13.08	13.22	0.88	0.76	2.44	0.74×10^6
GALENet#3	CMCA	13.34	13.49	**0.89**	0.77	2.47	1.25×10^6
GALENet（本文）	MCA+GA	**13.59**	**13.74**	**0.89**	**0.78**	**2.54**	0.77×10^6

注：黑体数值表示最优值。“↑”表示该值越高越好。

5 总 结

在本文中我们提出了一种新型端到端时域神经导向说话人提取模型——GALENet，通过整合时域神经导向技术与 EEG 信号的深度解码，成功地解决了传统语音分离方法在复杂环境中面临的挑战。创新型地引入门控交叉注意力融合机制，模型能够在语音和 EEG 信号之间进行高效的信息交互，从而提升了目标语音的提取精度。同时，模型设计中创新性地引入和设计局部特征来增强模块和 E-DPRNN 网络结构，进一步提升了对说话人特征的建模能力。在多个基准数据集上的实验结果表明，GALENet 在神经导向说话人提取任务中展现出强大的优势，特别是在处理多说话人场景时。实验结果也表明，GALENet 在性能与效率之间找到了良好的平衡，其 SI-SDR 和 PESQ 等指标在多个数据集上均超越了基线 MSFNet 模型和其他各种先进的对比模型，验证了该模型的广泛适用性和潜在的工业应用价值。在未来的工作中我们将进一步优化模型结构，探索更高效的模态信息和跨模态特征融合方法，并考察不同语言、不同导联数等条件对模型性能的影响，并在实际应用中实现更低的延迟和更高的实时处理能力。这些改进有望为助听器、语音增强设备以及多模态智能助手等应用提供技术支持，从而推动个性化听觉系统的发展，并为听力受损群体提供更好的沟通支持。

作者贡献声明：范存航提出了模型思路和实验方案；李骏凌开展实验、实验分析并撰写和修改论文；张晶晶负责补充实验和实验分析；高佑典、陈颖、易江燕提出指导意见并修改论文；吕钊参与论文审阅并提出修改意见。

参 考 文 献

[1] Bronkhorst A W. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions[J]. Acta Acustica United with Acustica, 2000, 86(1): 117−128

[2] Cherry E C. Some experiments on the recognition of speech, with one and with two ears[J]. Journal of the Acoustical Society of America, 1953, 25: 975−979

[3] Zhao Yan, Wang Deliang, Johnson E M, et al. A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions[J]. The Journal of the Acoustical Society of America, 2018, 144(3): 1627−1637

[4] Liu Yuchen, Zong Chengqing. End-to-end speech translation by integrating cross-modal information[J]. Journal of Software, 2023, 34(4): 1837−1849 (in Chinese)
(刘宇宸, 宗成庆. 跨模态信息融合的端到端语音翻译[J]. 软件学报, 2023, 34(4): 1837−1849)

[5] Han Songshen, Guo Songhui, Xu Kaiyong, et al. Perturbation analysis of the vital region in speech adversarial example based on frame structure[J]. Journal of Computer Research and Development, 2024, 61(3): 685−700 (in Chinese)
(韩松莘, 郭松辉, 徐开勇, 等. 基于帧结构的语音对抗样本重点区域扰动分析[J]. 计算机研究与发展, 2024, 61(3): 685−700)

[6] Chen Xiaoyu, Du Changde, Zhou Qiongyi, et al. Auditory attention decoding with task-related multi-view contrastive learning[C]//Proc of the 31st ACM Int Conf on Multimedia. New York: ACM, 2023: 6025−6033

[7] Fan Cunhang, Zhang Hongyu, Huang Wei, et al. DGSD: Dynamical graph self-distillation for EEG-based auditory spatial attention detection[J]. Neural Networks, 2024, 179: 106580

[8] O'Sullivan J, Chen Zhuo, Herrero J, et al. Neural decoding of attentional selection in multi-speaker environments without access to clean sources[J]. Journal of Neural Engineering, 2017, 14(5): 056001

[9] Cai Siqi, Zhang Ran, Li Haizhou. Robust decoding of the auditory attention from EEG recordings through graph convolutional networks[C]//Proc of the 49th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2024: 2320−2324

[10] Zhang Jie, Xu Qingtian, Zhu Qiushi, et al. BASEN: Time-domain brain-assisted speech enhancement network with convolutional cross attention in multi-talker conditions[C]//Proc of the 24th Interspeech. Kos Island, GR: ISCA, 2023: 3117−3121

[11] Pan Zexu, Borsdorf M, Cai Siqi, et al. NeuroHeed: Neuro-steered speaker extraction using EEG signals[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 4456−4470

[12] Fan Cunhang, Zhang Jingjing, Zhang Hongyu, et al. MSFNet: Multi-scale fusion network for brain-controlled speaker extraction[C]//Proc of the 32nd ACM Int Conf on Multimedia. New York: ACM, 2024: 1652−1661

[13] Zeghidour N, Grangier D. Wavesplit: End-to-end speech separation by speaker clustering[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 2840−2849

[14] Luo Yi, Chen Zhuo, Yoshioka T. Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation[C]//Proc of the 45th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2020: 46−50

[15] Subakan C, Ravanelli M, Cornell S, et al. Exploring self-attention mechanisms for speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2169−2180

[16] Huang Yating, Shi Jing, Xu Jiaming, et al. Research advances and perspectives on the cocktail party problem and related auditory models[J]. Acat Automatiac Sinica, 2019, 45(2): 234−251 (in Chinese)
(黄雅婷, 石晶, 许家铭, 等. 鸡尾酒会问题与相关听觉模型的研究现状与展望[J]. 自动化学报, 2019, 45(2): 234−251)

[17] Huang Xiangdong, Chen Honghong, Gan Lin. Speech enhancement method based on frequency-time dilated dense network[J]. Journal of Computer Research and Development, 2023, 60(7): 1628−1638 (in Chinese)
(黄翔东, 陈红红, 甘霖. 基于频率-时间扩张密集网络的语音增强方法[J]. 计算机研究与发展, 2023, 60(7): 1628−1638)

[18] Zmolikova K, Delcroix M, Ochiai T, et al. Neural target speech extraction: An overview[J]. IEEE Signal Processing Magazine, 2023, 40(3): 8−29

[19] Kong Yonghui, Fan Cunhang, Chen Yujie, et al. Bilateral masking with prompt for knowledge graph completion[C]//Proc of Findings of the 19th Association for Computational Linguistics (NAACL). Stroudsburg, PA: ACL, 2024: 240−249

[20] Li Zeyu, Zhang Xuhong, Pu Yuwen, et al. A survey on multimodal deepfake and detection techniques[J]. Journal of Computer Research and Development, 2023, 60(6): 1396−1416 (in Chinese)
(李泽宇, 张旭鸿, 蒲誉文, 等. 多模态深度伪造及检测技术综述[J]. 计算机研究与发展, 2023, 60(6): 1396−1416)

[21] Fan Cunhang, Gao Youdian, Pan Zexu, et al. Improved feature extraction network for neuro-oriented target speaker extraction[C]//Proc of the 50th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2025: 1−5

[22] Qiu Zelin, Yao Dingding, Li Junfeng. Two-stage brain-controlled speech enhancement with integrated dual-view features[J]. Acat Acustica, 2025, 50(2): 362−372 (in Chinese)
(邱泽林, 姚鼎鼎, 李军锋. 融合双视角特征的两阶段脑控语音增强[J]. 声学学报, 2025, 50(2): 362−372)

[23] Fan Cunhang, Zhang Sheng, Zhang Jingjing, et al. SSM2Mel: State space model to reconstruct mel spectrogram from the EEG[C]//Proc of the 50th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2025: 1−5

[24] Chen Yujie, Yi Jiangyan, Fan Cunhang, et al. Region-based optimization in continual learning for audio deepfake detection [C]//Proc of the 39th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2025: 23651−23659

[25] Fan Cunhang, Chen Yujie, Xue Jun, et al. Progressive distillation based on masked generation feature method for knowledge graph completion[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 8380−8388

[26] Han Cong, O'Sullivan J, Luo Yi, et al. Speaker-independent auditory attention decoding without access to clean speech sources[J]. Science Advances, 2019, 5(5): eaav6134

[27] Ceolini E, Hjortkjær J, Wong D D E, et al. Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception[J]. NeuroImage, 2020, 223: 117282

[28] Reju V G, Koh S N, Soon Y. Convolution using discrete sine and cosine transforms[J]. IEEE Signal Processing Letters, 2007, 14(7): 445−448

[29] Hosseini M, Celotti L, Plourde É. Speaker-independent brain enhanced speech denoising[C]//Proc of the 46th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2021: 1310−1314

[30] Hosseini M, Celotti L, Plourde E. End-to-end brain-driven speech

enhancement in multi-talker conditions[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 1718−1733

[31] Perez E, Strub F, De Vries H, et al. Film: Visual reasoning with a general conditioning layer[C]//Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 1

[32] Pan Z, Wichern G, Germain F G, et al. NeuroHeed+: Improving neuro-steered speaker extraction with joint auditory attention detection [C]//Proc of the 49th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2024: 11456−11460

[33] Zhu Chuanzhen, Wang Xuan, Zheng Qiang. Element separation and holistic attention based graph convolutional network framework[J]. Journal of Computer Research and Development, 2024, 61(8): 2008−2019 (in Chinese)
(祝传振, 王璇, 郑强. 基于元素分离与整体注意力的图卷积网络框架[J]. 计算机研究与发展, 2024, 61(8): 2008−2019)

[34] Zhao Shengkui, Ma Bin. Mossformer: Pushing the performance limit of monaural speech separation using gated single-head transformer with convolution-augmented joint self-attentions[C]//Proc of the 48th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2023: 1−5

[35] Zhao Shengkui, Ma Yukun, Ni Chongjia, et al. Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation[C]//Proc of the 49th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2024: 10356−10360

[36] Subakan C, Ravanelli M, Cornell S, et al. Attention is all you need in speech separation[C]//Proc of the 46th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2021: 21−25

[37] Yip J Q, Zhao Shengkui, Ma Yukun, et al. SPGM: Prioritizing local features for enhanced speech separation performance[C]//Proc of the 49th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE: 326−330

[38] Broderick M P, Anderson A J, Di Liberto G M, et al. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech[J]. Current Biology, 2018, 28(5): 803−809

[39] Zhang Hongyu, Zhang Jingjing, Dong Xingguang, et al. Based on audio-video evoked auditory attention detection electroencephalogram dataset[J]. Journal of Tsinghua University: Science and Technology, 2024, 64(11): 1919−1926 (in Chinese)
(张宏玉, 张晶晶, 董兴广, 等. 基于视听诱发的听觉注意检测脑电数据集[J]. 清华大学学报: 自然科学版, 2024, 64(11): 1919−1926)

[40] Fan Cunhang, Zhang Hongyu, Ni Qinke, et al. Seeing helps hearing: A multi-modal dataset and a mamba-based dual branch parallel network for auditory attention decoding[J]. Information Fusion, 2025, 118: 102946

[41] Delorme A, Makeig S. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis[J]. Journal of Neuroscience Methods, 2004, 134(1): 9−21

[42] Moinnereau M A, Rouat J, Whittingstall K, et al. A frequency-band coupling model of EEG signals can capture features from an input audio stimulus[J]. Hearing Research, 2020, 393: 107994

[43] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint, arXiv: 1412.6980, 2014



Fan Cunhang, born in 1991. PhD, associate professor, master supervisor. His main research interests include speech information processing, brain-computer interface technology, and brain-inspired auditory computing.

范存航, 1991 年生。博士, 教授, 博士生导师。主要研究方向为语音信息处理、脑机接口技术、类脑听觉计算。



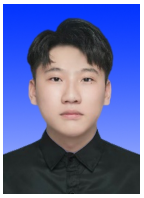
Li Junling, born in 2002. Master candidate. His main research interests include speech separation and brain-controlled target speaker extraction.

李骏凌, 2002 年生。硕士研究生。主要研究方向为语音分离、脑控目标说话人提取。



Zhang Jingjing, born in 2000. Master. Her main research interests include speech separation and brain-controlled target speaker extraction.

张晶晶, 2000 年生。硕士。主要研究方向为语音分离、脑控目标说话人提取。



Gao Youdian, born in 2001. Master candidate. His main research interests include audio-visual speech separation and target speaker extraction.

高佑典, 2001 年生。硕士研究生。主要研究方向为视听语音分离、目标说话人提取。



Chen Ying, born in 2001. Master candidate. Her main research interests include speech separation and brain-controlled target speaker extraction.

陈　颖, 2001 年生。硕士研究生。主要研究方向为语音分离、脑控目标说话人提取。



Yi Jiangyan, born in 1984. PhD, assistant professor. Her main research interests include speech recognize, and speech synthesis and transfer learning.

易江燕, 1984 年生。博士, 助理教授。主要研究方向为语音识别、语音合成和迁移学习。



Lü Zhao, born in 1979. PhD, professor, PhD supervisor. His main research interests include emotion recognition, speech enhancement, and motor imagery.

吕　钊, 1979 年生。博士, 教授, 博士生导师。主要研究方向为情绪识别、语音增强、运动想象。