会议场景下基于注意力机制的语音分离识别算法

赵成宝[1*] 张中海[1]
ZHAO Chengbao ZHANG Zhonghai

摘 要 在会议场景下，语音信号往往处于叠加状态，这会导致语音信号序列出现梯度爆炸或梯度消失的问题，进而降低语音识别的准确性。为此，文章开展了基于注意力机制的会议场景语音分离识别算法研究。首先，利用短时傅里叶变换将原始会议场景下的混合语音信号转换为语谱图形式；然后，通过包络检测提取语音信号的调制幅度谱，并将其作为编码器的输入。在分离器中，对编码器输出进行分割和 SepFormer 处理，以分离得到独立分布的语音信号。引入自注意力机制，通过多头学习处理输入的独立语音信号，并联合利用独立语音信号在不同空间的信息，从而在避免梯度爆炸或梯度消失问题的同时，准确识别并输出语音信息。在测试数据集上，所设计的算法分离输出的语音信号频域分布误差稳定在 0.05 Hz 以内，且对语音信息的准确识别率达到了 99.9%。

关键词 会议场景；注意力机制；语音分离识别；短时傅里叶变换；调制幅度谱；SepForme；多头学习

doi：10.3969/j.issn.1672-9528.2025.12.021

0 引言

在实际的会议场景下，会议现场的客观环境使得语音信号中包含大量噪声[1]，在极大程度上降低了音频的质量，这也导致对多种声音构成的混合音频信号进行识别的难度大大增加[2]。针对电力会议中多个角色同时发言的情况，对语音进行有效分离识别成为会议记录、会议总结以及会议关键信息提取的重要技术手段[3]。在相关研究中，刘子寒等[4]针对复杂场景中语音信号易受噪声干扰、多人语音相互重叠等特点，提出一种基于支持向量机的复杂场景中多人对话语音智能识别方法。首先，对采集到的复杂场景语音信号进行降噪和端点提取处理。然后，利用支持向量机构建适合多人对话语音识别的分类模型，以对语音特征进行识别。测试结果表明，设计方法在复杂场景中对多人对话语音具有较高的识别准确率，能够有效区分不同说话人的语音内容，但是语音信号序列较长时，对于语音信息的准确识别率较低。江楠等[5]开展了基于声纹识别的电力会议多角色语音的分离和识别研究，利用收集到的电力会议中不同角色的语音样本，提取声纹特征，并建立对应的角色模型。采用基于声纹识别的分离算法，将混合语音信号分离为不同角色的独立语音，并在角色模型中拟合识别分离后的语音。测试结果表明，设计方法能够有效分离和识别电力会议中的多角色语音，但是对于语音信息的准确识别率受语音序列状态的影响较为明显。吴亮等[6]以从包含多种声音的混合信号中分离出目标语音为目标，开展了基于多尺度自适应注意力机制的视听语音分离研究，结合音频和视频信息，通过多尺度处理捕捉不同层次的语音特征，包括音频信号的时频特征和视频的唇部运动特征。然后，利用多尺度自适应注意力机制融合音视频特征，以实现对目标语音的准确分离。测试结果表明，设计方法能够有效应对复杂环境下的视听语音分离任务，但是对于语音信息的准确识别率波动较为明显。屠彦辉等[7]开展了基于多模态波束方向特征的多模语音分离及识别研究，利用麦克风阵列采集语音信号，获取不同方向的语音波束信息，并提取多模态波束方向特征。然后，基于这些特征构建语音分离模型，将混合语音信号分离为不同声源的独立语音，并对分离后的语音进行识别。测试结果表明，设计方法能够有效提高语音分离的准确性和语音识别的性能，但是对于语音信息的准确识别率存在进一步提升的空间。

综合上述，本文开展了会议场景下基于注意力机制的语音分离识别算法研究，并通过对比测试的方式，分析了设计算法的性能。

1 语音分离识别算法设计

1.1 语音信号调制幅度谱提取

对于会议场景下的语音信号而言，其具有明显的多人语音信号叠加的特性，这也是导致语音识别面临的主要挑战[8]。基于此，本文在对其进行分解处理时，利用原始信号频率调制的窄带信号表征载波，利用原始信号幅度调制的低频信号

1. 甘肃同兴智能科技发展有限责任公司 甘肃兰州 730050

表征调制信号，以此实现对语音信号调制幅度谱的精准提取[9]。

利用短时傅里叶变换将原始会议场景下的混合语音信号转化为算语谱图的形式时[10]，其处理方式可以表示为：

$$\mathrm{STFT}f(t)=F(p,k)=\sum_{n=-\infty}^{\infty}h(p-t)f(t)W_K^{kt} \tag{1}$$

式中：$f(t)$ 表示原始会议场景下的混合语音信号；$F(p, k)$ 表示短时傅里叶变换输出的语音信号算语谱图；p 表示帧数参数；k 表示频率参数；$W_K=\mathrm{e}^{-\mathrm{j}(\frac{2\pi}{K})}$表示短时傅里叶变换尺度。

对于 $F(p, k)$，从时频域对其进行分解处理，调制信号和载波信号与之的关系可以表示为：

$$F(p,k)=M(p,k)C(p,k) \tag{2}$$

式中：$M(p, k)$ 表示调制信号；$C(p, k)$ 表示载波信号。

利用包络检波对调制信号 $M(p, k)$ 进行计算的方式可以表示为：

$$M(p,k)\triangleq DF(p,k)=\left|F(p,k)\right| \tag{3}$$

式中：D 表示执行调制信号 $M(p, k)$ 识别的包络检测器。

则由语音信号调制频谱分布输出的调制幅度谱可以表示为：

$$X(k,i)=\left|\mathrm{FFTM}(p,k)\right| \tag{4}$$

式中：$X(k, i)$ 表示语音信号的调制幅度谱；i 表示调制频率参数；$\mathrm{FFTM}(p, k)$ 表示语音信号的调制频谱。

按照上述所示的方式，提取输出语音信号的调制幅度谱，将其作为后续语音分离的执行基础。

1.2 调制幅度谱分离

结合上述输出的语音信号调制幅度谱，本文在开展对语音的分离处理时[11]，通过分离模型对调制幅度谱分型计算实现。

将语音信号调制幅度谱作为编码器的输入，在构建的分离器中，包含分割和 SepFormer，编码器的输出作为其输入，利用一个维度为 m 的线性层对输入信息进行层归一化处理，其计算方式可以表示为：

$$(X,G)(i)=\sum_{j=0}^{m-1}z(r\cdot i+j)*G(j) \tag{5}$$

式中：G 表示大小为 m 的卷积核；z 表示线性滑动窗口；r 表示卷积核的步长参数；j 表示卷积通道的数量。

经分割操作将线性层的输出分割为长度固定，且重叠程度为定值的子信号。将所有子信号的调制幅度谱堆叠成的三维向量作为 SepFormer 的输入，利用 DPRNN 的双尺度方法对输入进行处理时，分别采用子信号内 Transformer 进行短期依赖关系建模，利用子信号间 Transformer 进行长期依赖关系建模，其处理方式可以表示为：

$$o_i=f\left(\sum_i\omega_i x_i+\prod(X,G)(i)\right) \tag{6}$$

式中：o 表示 SepFormer 的整体运算输出，即语音信号的分离结果；ω_i 表示子信号内 Transformer 系数；f 表示子信号间 Transformer 函数。

按照上述所示的方式，实现对语音信号的分离处理，为后续的语音识别提供基础。

1.3 分离语音信号自注意力识别

在处理分离输出的独立语音信号时，本文采用了自注意力机制来进行识别。这种方法通过有效避免语音信号序列长度问题所带来的梯度爆炸或梯度消失现象，从而确保了识别结果的可靠性和准确性。具体来说，自注意力机制能够使模型在处理长序列数据时，更好地捕捉到各个时间步之间的依赖关系，从而提高了识别的精度和稳定性。通过这种方式，本文不仅解决了传统序列模型在处理长语音信号时可能遇到的梯度问题，还进一步提升了语音识别系统的整体性能。

设置自注意力机制对输入独立语音信号的处理方式可以表示为：

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{soft\,max}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{o_i}}\right)\boldsymbol{V} \tag{7}$$

$$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}(o_1,o_2,...,o_n)W^{o_i} \tag{8}$$

式中：$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ 表示独立语音信号的自注意力识别输出；$\boldsymbol{Q}$ 表示独立语音信号的查询矩阵；$\boldsymbol{K}$ 表示独立语音信号的键矩阵；$\boldsymbol{V}$ 表示独立语音信号的值矩阵；softmax 表示防止 softmax 函数，利用其实现对独立语音信号梯度的计算；$\frac{1}{\sqrt{o_i}}$表示缩放因子，利用其避免 softmax 进入极小的梯度区域，进而防止梯度消失问题的出现；$\mathrm{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ 表示多头注意力函数，利用其联合利用独立语音信号不同空间信息的参数，确保识别的泛化效果能够覆盖信号的调制幅度谱；Concat 表示关联函数；W^{o_i}表示独立语音信号的权重矩阵。

按照上述所示的方式，通过自注意力机制实现对语音信号的识别，保障识别结果的可靠性。

2 测试分析

2.1 数据集

在测试本文设计语音分离识算法的性能时，选用开源语音数据库 THCHS-30 中的部分语音数据作为测试数据集。对测试数据集的具体构成情况进行分析，其为安静办公室环境由单个碳粒麦克风录取的音频数据，数据集中包含语音的总时长为 30.6 h。对语音数据的对象来源进行统计，主要为大学生。在语音信号采样阶段，采样频率为 16.0 kHz，大小为 16.0 bit，输出的音频格式为 WAV。根据语音数据的内容构成对其进行分类，可以分为表 1 所示的四部分。

表 1 测试数据集构成

分类	内容类型	语音数量 / 条	音源人数 / 人
A	经济与贸易	115	男：10 女：12
B	天气与气候	205	男：17 女：13
C	环境与环保	163	男：9 女：16
D	其他	143	男：14 女：11

根据表 1 所提供的详细信息，按照 8:2 的比例将数据集划分为训练组和验证组。在这两个组中，分别挑选了 4 种不同类别的语音数据，每种类别的语音数据各选取一条，以此来构建一个混合语音数据集。通过这种方式，总共组建了 10 组混合语音数据，用以测试和评估设计的算法在实际应用中的性能表现。图 1 展示了这些混合语音信号的重叠状态，以便更直观地理解它们的构成和相互之间的关系。

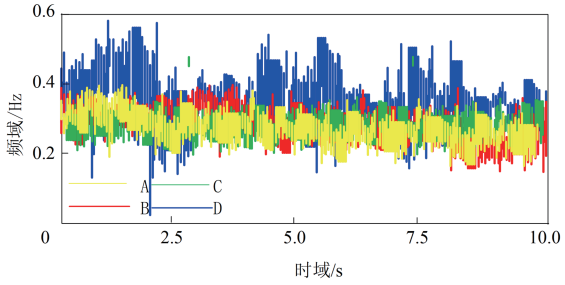

图 1 混合语音信号重叠状态

在对混合语音信号进行分离处理时，本文设置 SepFormer 的卷积核的步长参数 r 值为 0.2，卷积通道的数量 j 值为 4。

2.2 测试结果与分析

分别选取基于支持向量机的复杂场景中多人对话语音智能识别方法，以及基于声纹识别的电力会议多角色语音的分离和识别方法作为测试的对照组，测试 3 种不同方法和算法的性能。

首先对比分析了 3 种不同方法和算法对于语音信号的分离效果，将语音信号的频域分布误差作为评价指标，结果如图 2 所示。

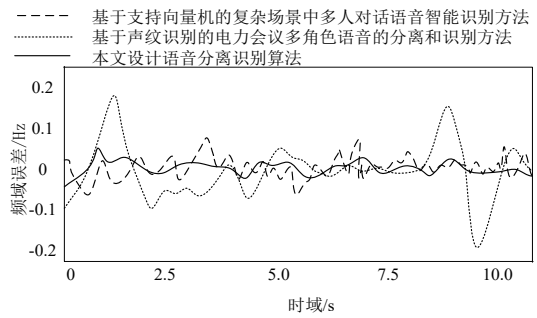

图 2 语音信号分离效果对比图

结合图 2 所示的测试结果对三种不同方法和算法对于语音信号的分离效果进行对比可以看出，本文设计算法的分离输出结果更加精准，对应的频域分布误差稳定在 0.05 Hz 以内，在稳定性和误差幅值方面均明显优于对照组，表明本文设计算法可以实现对混合语音信号的有效分离。这是因为设计算法通过提取输出语音信号的调制幅度谱，避免了语音信号叠加对分离的干扰，同时利用 SepFormer 分离调制幅度谱，输出了独立分布的语音信号，进而大大提高了分离的精准性。

其次，对比分析了不同算法对语音所属分类的识别结果，具体如表 2 所示。

表 2 识别结果对比表

应用方法 / 算法	分类	正确识别语音数量 / 条	错误识别语音数量 / 条	未识别语音数量 / 条
基于支持向量机的复杂场景中多人对话语音智能识别方法	A	8	2	0
	B	7	2	1
	C	8	1	1
	D	9	1	0
基于声纹识别的电力会议多角色语音的分离和识别方法	A	7	1	2
	B	9	1	0
	C	9	1	0
	D	6	2	2
本文设计语音分离识别算法	A	10	0	0
	B	9	0	1
	C	10	0	0
	D	10	0	0

结合表 2 所示的测试结果可以看出，在 10 组测试混合语音数据中，本文设计算法正确分类的语音数量最多，且基本达到了 100.0% 准确识别率。这是因为本文设计算法在对分离输出的语音信号进行识别时，引入了自注意力机制，有效避免了由于语音信号序列长度导致的梯度爆炸或梯度消失问题，进而保障了识别结果的可靠性。

3 结语

本文开展了会议场景下基于注意力机制的语音分离识别算法研究，利用短时傅里叶变换将原始会议场景下的混合语音信号进行转化，输出了算语谱图，并对其进行包络检测，提取输出语音信号调制幅度谱。在分离器中对调制幅度谱进行了分割和 SepFormer 处理，得到了独立分布的语音信号。在自注意力机制下通过多头学习输出了输出语音信息的识别结果。在测试数据集上，分离输出的语音信号频域分布误差稳定在较低水平，且准确识别率明显高于对照组。结合本文对于混叠模态语音的分离识别研究，希望为复杂场景下语音数据的处理与应用分析提供有价值的参考。

基于深度学习 – 强化学习联合算法的电力调度信息化择优决策

白桂君[1] 张 静[1]
BAI Guijun ZHANG Jing

摘 要 针对现有决策方法在应用到电力调度中，存在消纳率过低，造成能源过度浪费的问题，文章开展了一种基于深度学习 - 强化学习联合算法的电力调度信息化择优决策研究。通过构建双向长短期记忆网络负荷预测模型精准捕捉源荷时序特征，结合改进近端策略优化算法动态优化调度策略，实现电力系统安全约束、经济运行与可再生能源消纳的多目标协同优化。通过专家策略预训练与强化学习在线调参的联合训练流程，提升算法收敛速度。通过对比实验证明，相较于现有决策方法，新的方法在不同可再生能源出力场景下均表现出更强的适应性，特别是在高波动工况中能有效平衡系统安全性与消纳率。

关键词 深度学习；联合算法；信息化；电力调度；强化学习

doi：10.3969/j.issn.1672-9528.2025.12.022

0 引言

电网运营中传统“源随荷动”的单向运行模式逐步向“源荷互动”的双向协作模式过渡，电网耦合关系更为复杂，运行控制难度大幅提升。在此背景下，传统调度方式已很难应对这种动态、不确定环境下的动态、不确定因素，迫切需要引入智能技术来提高决策水平。

目前，电力系统中的人工智能技术已经得到了很大的发展。在此基础上，李志华[1]采用数据挖掘、机器学习等方法，对电网调度方案进行优化，以提高电网运行效率。但目前的研究主要集中在静态数据上，不能很好地反映电网的动态演变特征，也不能适应实时互动决策的需求。在此基础上，刘冉等[2]研究新能源电力系统的可信计算与前瞻性决策，并在此基础上，引入可信评价机制，提升新能源电力系统的健壮性。然而，该方法过于依赖于离线优化，很难应对电网实时负荷变化及源 - 荷交互的随机特性，并且计算量大，难以用于在线调度。

1. 国网青海省电力公司信息通信公司 青海西宁 810000

参考文献：

[1] 余传旗，郭海燕，王婷婷，等 . 基于图注意力网络和门控网络的轻量级单通道语音分离方法 [J]. 信号处理，2025, 41(4): 706-717.

[2] 褚俊佟，魏爽 . 基于深度学习的交叉残差连接网络应用于语音分离 [J]. 上海师范大学学报 (自然科学版中英文), 2025, 54(2): 229-237.

[3] 王树斌 .NMF 声场分离技术在语音识别中的优化研究 [J]. 电声技术 ,2025,49(1):68-70.

[4] 刘子寒，沈力，奚梦婷，等 . 基于支持向量机的复杂场景中多人对话语音智能识别方法研究 [J]. 计算技术与自动化，2024, 43(4): 59-65.

[5] 江楠，陈洁，肖潘，等 . 基于声纹识别的电力会议多角色语音的分离和识别研究 [J]. 高电压技术 ,2023,49(S1):40-46.

[6] 吴亮，王甲祥，施汉琴，等 . 基于多尺度自适应注意力机制的视听语音分离 [J]. 人工智能 ,2024(3):1-14.

[7] 屠彦辉，霍伟明，高建清，等 . 基于多模态波束方向特征的多模语音分离及识别 [J]. 人工智能 ,2024(3):36-44.

[8] 孙林慧，王春艳，张蒙 . 基于全卷积神经网络多任务学习的时域语音分离 [J]. 信号处理 ,2024,40(12):2228-2237.

[9] 刘宏清，谢奇洲，赵宇，等 . 基于扩张卷积和 Transformer 的视听融合语音分离方法 [J]. 信号处理 ,2024,40(7):1208-1217.

[10] 李可 . 面向语言对话场景的智能语音交互关键技术研究 [J]. 自动化与仪器仪表 ,2023(8):295-299.

[11] 陈佳佳，张海剑，华光 . 基于 Conformer 的时域多通道语音分离方法 [J]. 无线电工程 ,2023,53(9):2054-2060.

【作者简介】

赵成宝（1989—），通信作者（email:male202209@163.com），男，甘肃灵台人，本科，研究方向：会议通信。

张中海（1995—），男，甘肃天水人，本科，研究方向：会议通信。

（收稿日期：2025-06-17 修回日期：2025-11-28）