基于多分辨率并行特征提取的多声源分离方法

高 尚,贾懋珅*
(北京工业大学信息科学技术学院,北京 100124)

摘 要: 随着万物互联、智能感知及人机交互等技术的快速发展,复杂声场环境下的多声源分离已成为语音信号处理领域的重要的前端问题。然而,非平稳语音信号在不同时间和频率尺度呈现出不同的能量分布特性,其中既包括快速变化的共振峰结构,也包含相对平稳的谐波与周期信息。传统的单一时频分辨率分析方法在该场景下面临根本性约束:当分析窗较短时,频率分辨率不足,难以区分多个声源的谐波结构;而当窗长较长时,时间分辨率下降,又难以捕获语音快速变化的瞬态特征。因此,当前多声源分离方法在复杂声环境下往往表现出时频结构解析不足、语音细节丢失与分离失衡等问题。现有基于固定分辨率的分离方法在真实复杂声学环境中,常出现时频结构模糊、语音细节损失及分离后信号失真等问题,限制了系统在实际场景中的鲁棒性与可用性。为解决上述问题,所提方法实现了一种多分支并行的深度神经网络,每个分支独立处理由不同窗长生成的时频谱,并采用嵌套的层次化递归单元进行特征建模。具体而言,每个分支内部设计了两级递归模块:频率-空间建模单元(Frequency Long Short-Term Memory,F-LSTM)沿频带方向递归,提取跨通道的空间相关性与频域结构;时间-空间建模单元(Time Long Short-Term Memory,T-LSTM)沿时间轴递归,捕捉语音信号的长期动态演化与时序依赖性。此外,所提方法将不同分析窗生成的多组不同分辨率的时频谱并行输入网络,实现网络对于时间与频率分辨率的互补。在训练过程中,各分支通过共享的时域重建损失进行联合优化,推动网络学习跨分辨率的一致性表示与互补特征。每一个分支均设置嵌套结构以增强跨分辨率特征的交互与融合能力。在网络输出端,各分支估计的复数谱掩蔽经融合层集成,通过逆短时傅里叶变换重建时域信号,最终在时域和频域双重约束下进行端到端训练。所提多分辨率融合方案在高混响、多说话人环境下均能显著提升语音分离的客观指标与主观听感,且具备良好的结构灵活性,可迁移至其他基于时频分析的网络框架中,为未来面向复杂声场的多源分离模型设计提供了可扩展的思路与方法基础。

关键词: 多分辨率时频分析;稀疏成分分析;声源分离;深度神经网络
基金项目: 国家自然科学基金(No.62471012)
中图分类号: TN912.3 **文献标识码:** A **文章编号:** 0372-2112(2026)01-0183-12
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250937

Multi-Sound Source Separation Method Based on Multi-Resolution Parallel Feature Extraction

GAO Shang, JIA Maoshen*
(*School of Information Science and Technology*, *Beijing University of Technology*, *Beijing* 100124, *China*)

Abstract: With the rapid advancement of technologies such as the Internet of Everything, intelligent sensing, and human-machine interaction, multi-source separation in complex acoustic environments has become a crucial front-end challenge in speech signal processing. However, non-stationary speech signals exhibit distinct energy distribution characteristics across different temporal and frequency scales, encompassing both rapidly changing formant structures and relatively stable harmonic and periodic information. Traditional single-resolution time-frequency analysis methods face fundamental constraints in such scenarios: short analysis windows yield insufficient frequency resolution, hindering the distinction of harmonic structures across multiple sources; conversely, longer windows degrade temporal resolution, compromising the capture of rapidly changing transient features in speech. Consequently, existing multi-source separation techniques often exhibit inadequate time-frequency structure analysis, loss of speech details, and separation imbalance in complex acoustic environments. Therefore, existing fixed-resolution separation methods frequently suffer from blurred time-frequency structures, loss of speech detail, and distorted separated signals in real complex acoustic environments, which limits the robustness and practicality of the system in real-world scenarios. To address these challenges, the proposed method implements a multi-branch parallel deep neural network. Each branch independently processes time-frequency spectra generated with different

收稿日期:2025-10-19;录用日期:2026-01-23;责任编委:赵博
*通信作者:贾懋珅

window lengths and employs nested hierarchical recurrent units for feature modeling. Specifically, each branch incorporates a two-stage recursive module: a frequency-spatial modeling unit (Frequency Long Short-Term Memory, F-LSTM) that operates along the frequency axis to extract cross-channel spatial correlations and spectral structures, and a time-spatial modeling unit (Time Long Short-Term Memory, T-LSTM) that recurs over time to capture the long-term dynamic evolution and temporal dependencies of speech signals. Furthermore, the approach feeds multiple sets of time-frequency spectra—generated from different analysis windows and featuring varying resolutions—into the network in parallel. During training, all branches are jointly optimized through a shared time-domain reconstruction loss, promoting the learning of consistent and complementary representations across resolutions. Each branch incorporates a nested architecture to enhance the interaction and fusion of cross-resolution features. At the output stage, the complex spectral masks estimated by each branch are integrated via a fusion layer, and the time-domain signal is reconstructed through inverse short-time Fourier transform, ultimately enabling end-to-end training under both time-domain and spectral constraints. Through multi-resolution joint optimization, the model simultaneously captures transient details and periodic harmonic structures within the speech spectrogram. The proposed multi-resolution fusion scheme significantly improves both objective metrics and subjective listening quality in highly reverberant and multi-speaker environments, and demonstrates structural flexibility, making it transferable to other time-frequency analysis-based network frameworks, thereby providing a scalable design approach and methodological foundation for future multi-source separation models targeting complex acoustic fields.

Keywords: multi-resolution time-frequency analysis; sparse component analysis; sound source separation; deep neural network

Foundation Item(s): National Natural Science Foundation of China (No.62471012)

0 引言

近二十年来,音频物联网(Internet of Audio Things, IoAuT)[1]与音乐物联网(Internet of Musical Things, IoMusT)[2]的提出,标志着声音数据处理迈入网络化与协同智能的新阶段。以远程音乐合奏为例,为了实现沉浸式临场体验,系统需精确分离各演奏者的独立声源,并结合三维声场重建算法对声源方位、传播路径及混响特性进行动态建模,从而逼真再现舞台空间感与氛围[3-7]。然而,居家场景与专业录音室在声学条件上差异显著,用户常面临多源噪声耦合问题,尤其是小型电子设备的脉冲性噪声在时频特性上与乐器信号高度重叠,导致传统滤波方法失效。因此,在复杂声学环境中实现目标声源的高保真分离成为高质量声学重构的核心技术挑战。

近年来,深度学习驱动的多声源分离技术因强大的信号建模能力而受到广泛关注[8-11]。该类方法通常将分离任务视作从混合信号到各源信号的回归映射问题,可按输入通道数分为单通道与多通道两类,并各自包含监督与非监督策略。监督学习依赖大规模标注样本,典型方法包括深度聚类(DeeP CLustering, DPCL)与置换不变训练(Permutation Invariant Training, PIT)[12-13],用于缓解盲源分离中的“排序问题”。单通道监督方法多借助时频域掩蔽估计[14-16],而端到端时域模型(如TasNet与Conv-TasNet)[13-17]通过直接估计波形实现低延迟与高质量分离。其后提出的双路径RNN(Dual-Path Recurrent Neural Network, DPRNN)[18]、状态空间模型(如Convolutional Mamba)[19]及混合时频结构(如Hybrid Spectrogram Time-domain Audio Separation Network、LadderNet)[20-22]进一步提升了长序列建模与模型效率。多通道监督分离方法在联合学习空间与时频特征方面展现出优势,能够降低噪声与混响的影响。典型代表如基于子带长短时记忆(Long Short-Term Memory, LSTM)网络的多通道增强方法[23]以及利用双耳特征(如能级差与相位差)保持空间一致性的结构[24]。与此同时,非监督学习通过设计代理任务或利用预训练伪标签以降低数据获取成本。代表性方法包括Mixture Invariant Training(MixIT)[25-26]、基于混响先验的RAS(Reverberation As Supervision)框架[27],以及结合置信度筛选机制的域自适应方法[28-29],在真实环境下均表现出较强的泛化能力。

尽管现有方法在结构设计与学习范式上取得显著进展,但其普遍仅依赖单一时频分辨率下的信号表示,难以充分挖掘多尺度声学特征的互补信息。当单一分辨率的时频谱出现严重能量泄露时,模型提取的特征易受干扰,导致分离性能受限,尤其在混响或多声源同时发声的场景中更为突出。基于此,本文面向多声源分离问题,提出一种基于多分辨率并行特征提取的深度分离框架。该方法通过设置不同窗长生成多组互补的时频分辨率表示,在时间与频率分辨率间形成协同优势,使模型同时捕获快速变化的瞬态特征与缓变的周期性特征。网络采用多层递归结构与嵌套式LSTM以强化跨分辨率特征交互,训练目标为最小化融合输出与目标源信号之间的时域重建误差。

模型通过融合多分辨率复数谱掩蔽实现声源信息互补,从而显著提升在复杂声学场景下的分离性能。

1　多分辨率接收信号模型

对于D个声源信号同时发声的场景,采用多个时频分析窗口进行时频变换,则第k个时频分析窗口生成的时频信号模型可以表示为

$$X(\omega,t,k)=\sum_{d=1}^{D}\boldsymbol{A}_d\boldsymbol{S}_d(\omega,t,k)+\boldsymbol{S}_{\mathrm{R}}(\omega,t,k)+\boldsymbol{N}(\omega,t,k)\tag{1}$$

其中,$\boldsymbol{S}_d(\omega,t,k)$为声源$d$的信号成分;$\boldsymbol{S}_{\mathrm{R}}(\omega,t,k)$为多个声源的晚期混响成分;$\boldsymbol{N}(\omega,t,k)$为噪声成分,其定义如下:

$$\boldsymbol{S}_{\mathrm{R}}(\omega,t,k)=\sum_{m=1}^{M_k-1}\sum_{d=1}^{D}\sum_{l=1}^{L(d)}\alpha_d^l\cdot s_d(m+t\cdot H-\tau_d^l)\cdot g_k(m)\cdot \mathrm{e}^{\frac{-\mathrm{j}\omega m}{M_k}}\tag{2}$$

$$\boldsymbol{N}(\omega,t,k)=\sum_{m=1}^{M_k-1}n(m+t\cdot H)\cdot g_k(m)\cdot \mathrm{e}^{\frac{-\mathrm{j}\omega m}{M_k}}\tag{3}$$

其中,M_k为第k个STFT(Short-Time Fourier Transform)窗口的窗长;α_d^l代表声源d在第l条反射路径上的路径衰减;g_k表示第k个时频窗口;$L(d)$为声源d到达麦克风阵列的反射路径总数;l为反射路径索引;τ_d^l为声源d的第l个反射路径到达麦克风的时延;H为STFT的平移点数;$n(t)$为零均值宽平稳噪声信号的时域表示。

根据上述公式可知,采取不同窗口进行变换时,直达信号、噪声与晚期混响成分均具有不同的能量分布,而一般认为噪声与混响成分在不同分辨率的时频谱的相关性远低于直达信号成分在不同分辨率的相关性,因此进行多分辨率融合时,直达信号成分更趋于相互叠加,而其余干扰成分则相对被抑制。以噪声成分为例,验证其在不同分辨率下的不相关特性,从而说明多分辨率整合对于噪声抑制的作用。

基于式(3)可知,对于分辨率k_1和k_2,噪声成分在ω处的互相关函数可以表示为

$$R_{k_1k_2}(\omega)=\sum_{m,v}R_n(m-v)g_{k_1}(m)g_{k_2}(v)\mathrm{e}^{-\mathrm{j}\omega\left(\frac{m}{M_{k_1}}-\frac{v}{M_{k_2}}\right)}\tag{4}$$

其中,$R_n(\tau)$为噪声的自相关函数,基于噪声在时序上的不相关假设,其在$\tau\neq0$位置处的数值迅速衰减,因此,式(4)可以近似为

$$R_{k_1k_2}(\omega)\approx\sigma_n^2\sum_{m,v}g_{k_1}(m)g_{k_2}(m)\mathrm{e}^{-\mathrm{j}\omega m\left(\frac{1}{M_{k_1}}-\frac{1}{M_{k_2}}\right)}\tag{5}$$

从式(5)中可以观察到,$R_{k_1k_2}(\omega)$可以近似为针对窗函数乘积$g_{k_1}(m)\cdot g_{k_2}(m)$的离散傅里叶变换,套用其对应的连续傅里叶变换模式即可以表示为

$$R_{k_1k_2}(\omega)\approx\sigma_n^2\int f_{k_1k_2}(u)\mathrm{e}^{-\mathrm{j}\beta u}\mathrm{d}u\tag{6}$$

其中,$f_{k_1k_2}(u)=g_{k_1}(u)\cdot g_{k_2}(u)$,$\beta=\omega\left(1/M_{k_1}-1/M_{k_2}\right)$。根据黎曼–勒贝格定理,若$f(x)$为可积函数,则其对应的傅里叶变换$F(x)=\int_{-\infty}^{+\infty}f(x)\mathrm{e}^{-\mathrm{j}\omega x}\mathrm{d}x$在$\omega$趋近于无穷时,$F(x)$趋近于0,而可以观察到$R_{k_1k_2}$的对应系数$\beta$取决于两个分量,其一是频率参量$\omega$,其二则是两个分析窗口的窗长差异。因此可以得出结论,随着分析频率的提升与不同分辨率窗口差异的增加,噪声成分的相关性随之下降。另外,对于信号成分与混响成分,其在分辨率k_1和k_2上的频谱互相关可以表示为

$$\mathbb{E}\left[\sum_{d_1=1}^{D}A_{d_1}S_{d_1}(\omega,t,k_1)\cdot\sum_{d_2=1}^{D}A_{d_2}^{*}S_{d_2}^{*}(\omega,t,k_2)\right]=\sum_{d=1}^{D}\left|A_d\right|^2\sum_{m_1=0}^{M_{k_1}-1}\sum_{m_2=0}^{M_{k_2}-1}s_d(m_1+tH)s_d^{*}(m_2+tH)\,g_{k_1}(m_1)g_{k_2}^{*}(m_2)\mathrm{e}^{-\frac{\mathrm{j}\omega}{M_{k_1}}m_1+\frac{\mathrm{j}\omega}{M_{k_2}}m_2}\tag{7}$$

而将混响信号带入相关性判据,得到结果如下:

$$\mathbb{E}\left[S_R(\omega,t,k_1)\cdot S_R^{*}(\omega,t,k_2)\right]=\sum_{l=1}^{L(d)}\sum_{d=1}^{D}\mathbb{E}\left[\left|\alpha_d^l\right|^2\right]\sum_{m_1=0}^{M_{k_1}-1}\sum_{m_2=0}^{M_{k_2}-1}s_d(m_1+tH-\tau_d^l)s_d^{*}(m_2+tH-\tau_d^l)\,g_{k_1}(m_1)g_{k_2}^{*}(m_2)\,\mathrm{e}^{-\frac{\mathrm{j}\omega}{M_{k_1}}m_1+\frac{\mathrm{j}\omega}{M_{k_2}}m_2}\tag{8}$$

从上述公式可以看出,信号与混响成分不同分辨率互相关的核心差异在于二者的幅度项以及相位项,其中,混响信号互相关的幅度项为$\left|\alpha_d^l\right|^2$,而信号成分的幅度项为$\left|A_d\right|^2$,由于反射成分经过墙面吸收,且传播路径衰减往往大于信号成分,因此其幅度项往往相对较低。另一方面,由于传输路径时延各有不同,因此反射信号成分的相位分布往往具有更高的随机性,最终导致多分辨率融合时极有可能出现相互抵消的情况,因此反射成分的多分辨率相关性低于信号成分,经过多分辨率整合后的信号往往可以优化直达信号同时抑制干扰成分。

2　基于多分辨率并行特征提取的多声源分离网络框架

2.1　参数测试

根据第1节的描述可知,不同分辨率的时频谱中,语音信号的直达声成分相关性较强,而噪声与混响分量由于相位的随机性或能量泄漏而呈现低相关性特征。因此,在深度神经网络结构中引入多种时频分辨率的互补信息则有望强化直达语音成分同时抑制噪声与混响。基于这一思路,本文提出一种基于多分辨率时频分析的联合非线性滤波网络。在预处理阶段,首先通过不同窗长对多通道信号进行STFT,生

成多个具有互补特征的时频谱表示作为网络输入。每个分辨率的谱分别对应一个结构相同的子网络分支,其内部通过频域与时间域的双层递归单元联合建模语音的空间、频谱与时间特征,从而估计对应的复数掩蔽。最后在统一的时域重建框架中对各分辨率输出结果进行融合,并以端到端的时域损失函数进行参数更新。

所提方法的整体流程框图如图1(a)中可以看到,为了进一步增强网络的特征提取能力并改善复杂声场下的非线性拟合效果,所提方法在每个分辨率的子网络中均采用多级递归结构,即在第一组F-LSTM和T-LSTM之后再堆叠组用于高层特征细化与重建的F-LSTM与T-LSTM模块组,形成层次化的时频递归建模框架,每级模块分别学习不同层次的频谱特征与时间动态特征,以保证网络既能拟合语音信号的局部模式,又能刻画其全局结构的依赖关系。在每个分辨率对应的网络分支输出端均通过全连接层与tanh激活函数输出复数谱掩蔽,并以此得到对应的多声源分离结果。随后,将所有分辨率的语音谱进行融合得到融合谱表示,利用ISTFT变换重建为时域信号并计算损失函数,通过将其反向传播至所有分辨率分支实现跨分辨率的协同学习。

图1(b)则对网络的核心结构,多分辨率并行特征提取网络部分进行了更细致的解析。对于每一个分辨率下的网络分支结构,其输入均由各通道的复数谱对应的实部与虚部拼接而成,数据维度为$\boldsymbol{X}_{\text{input}}\in\mathbb{R}^{B\times T\times F\times 2C}$,其中$B$为批尺寸,$T$为帧数,$F$为频点总数,$2C$表示每个通道的实部与虚部拼接结果。网络先后通过频率-空间建模单元(F-LSTM)和时间-空间建模单元(T-LSTM)进行特征提取,两模块之间的特征数据经维度转换操作以适配不同的循环方向。F-LSTM模块主要学习不同通道上的瞬时频率特性、相位差异与谱局部结构;而T-LSTM模块则聚焦于语音的连续时间变化特征,实现空间-时间联合表征。下文将分别对这两个建模单元的计算过程进行详细说明。

2.2 频率-空间建模单元

语音信号在邻近频率之间往往呈现较强的相关性,而麦克风阵列各输出通道的相位和幅度差异则反映了声源信号的空间特性,因此F-LSTM的主要目标是同时学习输入信号的空间特性和频率特征,通过在频谱维度上展开递归循环结构,实现对跨频率、跨通道信息的联合建模。F-LSTM对每一帧的信号成分均进行独立处理,沿频率维度递归,学习跨频带的频谱结构与多通道空间特征的联合表示。

如图1(b)中左侧第一个立方体所示,其将单帧全频带的信号成分作为输入,尺寸为$\boldsymbol{X}_{\text{F-LSTM}}\in\mathbb{R}^{(BT)\times F\times 2C}$。该建模单元的递归关系可表示为

$$\boldsymbol{h}_f=\text{LSTM}_F(\boldsymbol{x}_f,\boldsymbol{h}_{f-1}) \tag{9}$$

其中,$\boldsymbol{x}_f$为第f个频点输入;$\boldsymbol{h}_f\in\mathbb{R}^{D_F}$为该频点对应的隐状态向量;$\boldsymbol{h}_{f-1}\in\mathbb{R}^{D_F}$是上一个频率点$f-1$的隐状态;$D_F$是F-LSTM的隐藏单元维度。F-LSTM是一个标准的LSTM单元,其沿频率索引f递归更新,其内部计算过程包括三个门控信号和候选细胞状态,对于频率点$f(1\leqslant f\leqslant F)$,内部迭代过程可以表示为

$$\begin{cases}\boldsymbol{f}_f=\sigma\left(\boldsymbol{W}_{\text{forget}}\cdot[\boldsymbol{h}_{f-1};\ \boldsymbol{x}_f^{(i)}]+\boldsymbol{b}_{\text{forget}}\right)\\ \boldsymbol{i}_f=\sigma\left(\boldsymbol{W}_{\text{input}}\cdot[\boldsymbol{h}_{f-1};\ \boldsymbol{x}_f^{(i)}]+\boldsymbol{b}_{\text{input}}\right)\\ \boldsymbol{o}_f=\sigma\left(\boldsymbol{W}_{\text{output}}\cdot[\boldsymbol{h}_{f-1};\ \boldsymbol{x}_f^{(i)}]+\boldsymbol{b}_{\text{output}}\right)\\ \tilde{\boldsymbol{c}}_f=\tanh\left(\boldsymbol{W}_{\text{cell}}\cdot[\boldsymbol{h}_{f-1};\ \boldsymbol{x}_f^{(i)}]+\boldsymbol{b}_{\text{cell}}\right)\end{cases} \tag{10}$$

其中,$\boldsymbol{f}_f$为遗忘门;$\boldsymbol{i}_f$为输入门;$\boldsymbol{o}_f$为输出门;$\tilde{\boldsymbol{c}}_f$为候选细胞状态;[;]表示向量拼接;$\boldsymbol{W}$为权重矩阵;$\boldsymbol{b}$为偏置向量;σ为Sigmoid激活函数,代表网络存储的频谱结构与空间特征的上下文信息。随后更新细胞状态:

$$\boldsymbol{c}_f=\boldsymbol{f}_f\otimes\boldsymbol{c}_{f-1}+\boldsymbol{i}_f\otimes\tilde{\boldsymbol{c}}_f \tag{11}$$

其中,$\boldsymbol{c}_{f-1}$是上一个频率点的细胞状态;$\otimes$表示逐元素相乘。计算当前隐状态输出:

$$\boldsymbol{h}_f=\boldsymbol{o}_f\otimes\tanh(\boldsymbol{c}_f) \tag{12}$$

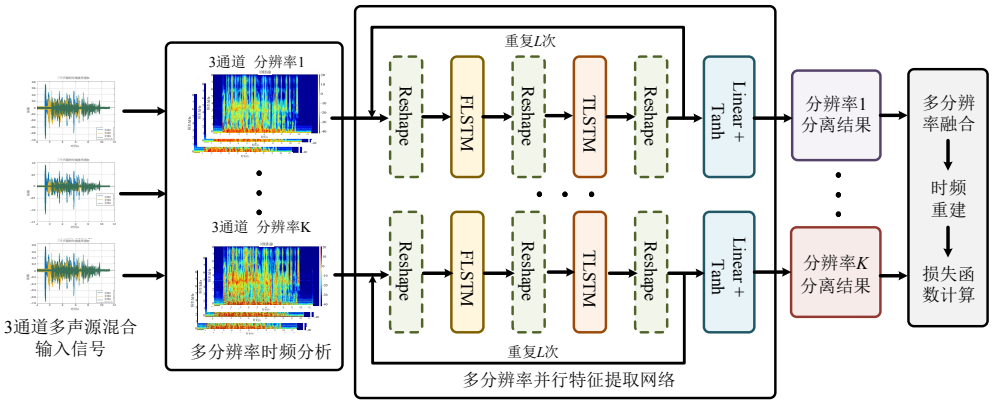
F-LSTM的输出结构为$\boldsymbol{H}_{\text{F-LSTM}}=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_F]$,其中$D_F$表示F-LSTM的隐藏单元数。F-LSTM模块能够在保留多通道语音信息的空间特征的同时捕捉多通道信号在频域上的变化特征与共振结构。

2.3 时间-空间建模单元

语音作为非平稳信号,其短时谱形态随时间连续变化。通过在时间维度上展开递归序列,T-LSTM模块学习每个频率点随时间的动态演化关系用以建模语音信号的长期时序特性(如节奏、共振峰轨迹周期性等),同时保持空间特征相位一致性。为了进行以频带为中心的时间建模,需要对F-LSTM的输出向量进行重排作为T-LSTM的输入。具体来说,对于每个固定的频率点f,我们提取它在所有时间帧上的F-LSTM特征,形成一条时间序列$\boldsymbol{X}_f=\left[\boldsymbol{h}_1^{(f)},\boldsymbol{h}_2^{(f)},\cdots,\boldsymbol{h}_T^{(f)}\right]^{\text{T}}$。在每一频带上,沿时间维度递归,学习语音时序动态与空间相位一致性的长期依赖。其递归关系定义为

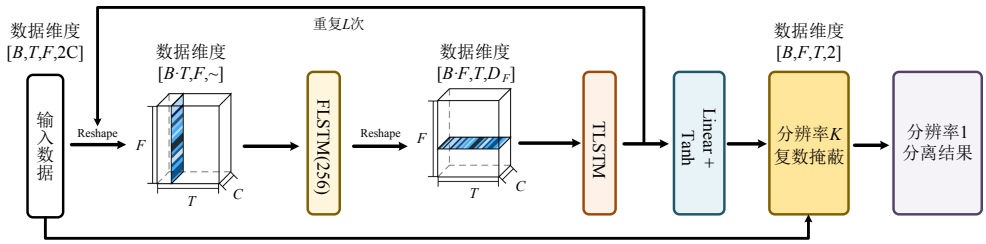
$$\boldsymbol{h}_t=\text{LSTM}_T(\boldsymbol{x}_t,\boldsymbol{h}_{t-1}) \tag{13}$$

其输出维度为$\boldsymbol{H}_{\text{T-LSTM}}\in\mathbb{R}^{(BF)\times T\times D_T}$,具体迭代过程与F-LSTM模块具有相似性,以固定的频率点f为例,对于时间帧$i(1\leqslant i\leqslant T)$网络计算门控与候选状态



(a) 基于多分辨率并行特征提取的多声源分离方法整体框架

(a) Overall framework of multi-source speech separation method based on multi-resolution parallel feature extraction



(b) 基于多分辨率并行特征提取的多声源分离方法整体框架

(b) Network architecture for multi-resolution parallel feature extraction

图 1　基于多分辨率并行特征提取的多声源分离网络框架

Figure 1 Multi-source speech separation network framework based on multi-resolution parallel feature extraction

的过程可以表示为

$$\begin{cases} \boldsymbol{f}_i^{(f)} = \sigma\left(\boldsymbol{W}_{\text{forget}} \cdot [\boldsymbol{h}_{i-1}^{(f)};\ \boldsymbol{x}_i^{(f)}] + \boldsymbol{b}_{\text{forget}}\right) \\ \boldsymbol{i}_i^{(f)} = \sigma\left(\boldsymbol{W}_{\text{input}} \cdot [\boldsymbol{h}_{i-1}^{(f)};\ \boldsymbol{x}_i^{(f)}] + \boldsymbol{b}_{\text{input}}\right) \\ \mathbf{o}_i^{(f)} = \sigma\left(\boldsymbol{W}_{\text{output}} \cdot [\boldsymbol{h}_{i-1}^{(f)};\ \boldsymbol{x}_i^{(f)}] + \boldsymbol{b}_{\text{output}}\right) \\ \tilde{\mathbf{c}}_i^{(f)} = \tanh\left(\boldsymbol{W}_{\text{cell}} \cdot [\boldsymbol{h}_{i-1}^{(f)};\ \boldsymbol{x}_i^{(f)}] + \boldsymbol{b}_{\text{cell}}\right) \end{cases} \tag{14}$$

其中，$\boldsymbol{h}_{i-1}^{(f)} \in \mathbb{R}^{D_T}$ 是该频率点下上一个时间帧的隐状态；D_T 是 T-LSTM 的隐藏单元维度，随后需要更新细胞状态与计算隐状态：

$$\boldsymbol{c}_i^{(f)} = \boldsymbol{f}_i^{(f)} \otimes \boldsymbol{c}_{i-1}^{(f)} + \boldsymbol{i}_i^{(f)} \otimes \tilde{\boldsymbol{c}}_i^{(f)} \tag{15}$$

$$\boldsymbol{h}_i^{(f)} = \boldsymbol{o}_i^{(f)} \otimes \tanh(\boldsymbol{c}_i^{(f)}) \tag{16}$$

其中，$\boldsymbol{c}_{i-1}^{(f)}$ 是上一个时间帧的细胞状态。T-LSTM 为网络提供了跨时间尺度的信息保留与特征平滑能力，使分离模型在强混响或非平稳噪声场景下仍能保持输出的连续性与鲁棒性。

在时间-空间建模单元之后，网络需要将 T-LSTM 输出的高维时频特征映射回与输入维度一致的复数谱空间，以获得每个时频点对应的掩蔽系数。为此，本文在 T-LSTM 模块后端引入全连接层与 tanh 激活函数的组合模块，用于完成对复掩蔽的非线性映射。全连接层在每个时频位置上独立执行线性映射，将其投影至二维复谱空间，输出为复值掩蔽 $M(\omega,t,k) \in \mathbb{R}^{B\times T\times F\times 2}$，维度中的 2 分别对应于复值掩蔽的实部与虚部。tanh 函数的输出范围限定在 [−1, 1]，以保证生成掩蔽的幅度稳定并符合物理约束。

2.4　多分辨率融合与损失函数计算

考虑到不同窗长对应的时频分辨率各异，短窗侧重于捕获语音的瞬态细节而长窗更能反映稳态谐波结构，在获得每个分辨率谱的复数谱掩蔽估计后，需要融合多时间尺度特征以生成最终的语音估计输出。本节直接采用各分辨率输出谱的均值作为融合策略，对于第 k 个分辨率分支，多分辨率融合的计算过程可表示为

$$\hat{S}_{\text{Fusion}}(\omega,t) = \frac{1}{K}\sum_{k=1}^{K}\hat{S}(\omega,t,k) = \frac{1}{K}\sum_{k=1}^{K}X(\omega,t,k)\cdot M(\omega,t,k) \tag{17}$$

其中，K 为分辨率总数。融合过程能够实现不同分辨率谱的互补平滑，有效削弱随机噪声和混响成分对语音结构的破坏。随后，所提方法将融合谱进行 ISTFT 得到时域整合谱 $\hat{s}$，网络以端到端方式训练，其损失函数定义为

$$\mathcal{L} = \left\| s - \hat{s} \right\|_1 + \left\| |S| - \left|\hat{S}_{\text{fusion}}\right| \right\|_1 \tag{18}$$

其中，第一项为时域L_1损失，用于约束输出波形与目标语音的一致性；第二项为幅度谱一致性损失，确保模型能准确恢复频域能量分布。在反向传播时，将损失函数平均分配至各分辨率子网络的Linear+tanh、T-LSTM和F-LSTM模块，实现共享目标下的协同优化。通过以上设计，所提网络在不增加额外参数的情况下即可充分利用多分辨率特征互补，实现对多声源语音信号的分离和干扰抑制。

3 实验与分析

为验证所提方法的有效性，本节基于WSJ0（Wall Street Journal）语料库构建实验数据集，共计生成20 000余条混合语音样本。所有样本均由两个同时发声的声源构成，单个数据集内部的语音内容和说话人均不重复，以保证样本的多样性和统计独立性。

混合信号在三种典型环境下进行模拟，以全面评估模型在不同声场条件下的表现能力。第一个场景对应于WSJ0-2mix数据集，仅叠加两路干净语音信号，不添加额外背景噪声或混响，用于评估网络在最理想条件下的语音分离上限性能；第二类场景在干净语音混合信号中添加通道等效白噪声，信噪比固定为20 dB，用于考查模型在随机背景噪声干扰下的分离稳定性；第三类场景则在噪声场景基础上进一步加入房间混响效应。混响信号通过镜像源法[30]生成，通过将实际声源关于房间各表面进行镜像，生成一系列虚拟声源，每个虚拟声源的直达路径对应一条从实际声源到麦克风经过一次或多次反射的传播路径，通过调整各墙面的吸收率，将混响时间控制在[0.2 s，0.8 s]区间内随机取值。所有场景的房间长、宽、高分别从[3 m，9 m]、[2.5 m，5 m]、[2.2 m，3.5 m]的范围内随机抽取，以生成不同的声学空间特性。声源水平方位角间隔在大于15°的条件下随机取值。在数据生成过程中，三通道麦克风阵列的直径固定为0.10 m，阵列平面高度保持在1.5 m。多分辨率时频分析部分采用三种不同长度的Hanning窗进行STFT，窗长分别为16、24和32 ms，帧移为窗长的一半，采样率设为16 kHz。下面将详细介绍实验结果。

3.1 无噪声、混响场景下的多声源分离结果评估

为验证所提多分辨率联合网络在理想条件下的基础分离能力，本节首先在无噪声、无混响场景下进行实验评估，此时混合信号仅由两个纯净语音源构成，旨在排除外界干扰因素，观察模型在单纯语音叠加情形下的波形分离精度。评估的所有方法可以分为三大类，其一是基于单一分辨率时频谱的方法，包括经典的Conv TasNet方法[17]、DPRNN方法[18]、SepFormer方法[31]以及 JNF方法[32]。第二类方法则是本文所提的联合多分辨率谱分析的多声源分离方法，将不同窗长下的时频谱组合后并行输入多分支JNF网络进行联合学习。具体包括三种组合形式，融合短窗（16 ms）与中窗（24 ms）生成的时频谱，融合中窗（24 ms）与长窗（32 ms）的时频谱，融合短窗（24 ms）与长窗（32 ms）的时频谱，在后续客观指标统计中，计算三种组合方式的平均指标，并称为“双分辨率JNF”。第三类方法则是三分辨率并行融合模型，通过多尺度特征协同学习语音的短时变化与长时结构，在后续被称为“三分辨率JNF”。

为了突出所提方法的优势，本小节运用四个评估指标以全方位评价各种方法，评估结果如表1所示。

表1 无噪声、无混响场景下的双声源分离结果评估

Table 1 Evaluation of the separation results of two sound sources in a noise-free and reverberation-free environment

方法名称		评估标准			
		SI-SDR	SAR	STOI	PESQ
单分辨率方法	Conv-Tas-Net	12.649 5	11.895 2	0.897 55	2.801 0
	DPRNN	12.975 1	11.917 0	0.907 60	2.807 0
	Sepformer	14.036 1	13.614 0	0.935 83	3.944 8
	单分辨率JNF	14.909 2	14.851 6	0.934 00	4.038 6
双分辨率方法	双分辨率JNF	16.703 4	17.066 3	0.957 69	**4.366 5**
三分辨率方法	三分辨率JNF	**18.893 9**	**18.874 4**	**0.962 46**	4.328 3

注：加粗表示当前指标最优表现。

从表1可以看出，在无噪声、无混响的场景下，各模型均能取得较高的语音分离性能。传统的时域分离模型Conv-TasNet仅依赖卷积时域特征进行掩蔽估计，其评分略低于其他方法。相比之下，联合时频域特征的JNF模型在相同条件下获得了显著改善。而在输入JNF网络的三个分辨率中，16 ms短窗口生成的时频谱输入对应的分离效果较好，表明在无干扰环境下较高的时间分辨率有助于模型准确分离瞬态语音成分。在引入多分辨率语音信息后，模型的语音分离性能得到进一步提升。对于两个分辨率并行特征提取的多声源分离结果，其所有指标均高于任意单分辨率模型，这验证了不同分辨率下的时频特征在分离过程中具有明显互补作用。最后，当使用三个分辨率的时频谱输入网络并进行同步优化时，分离结果在各项指标上均可获得最优的表现，这一结果说明多分辨率特征融合不仅能改善分离谱的时频一致性，还能在主观听感上提高语音流畅性。

3.2 带噪场景下的多声源分离结果评估

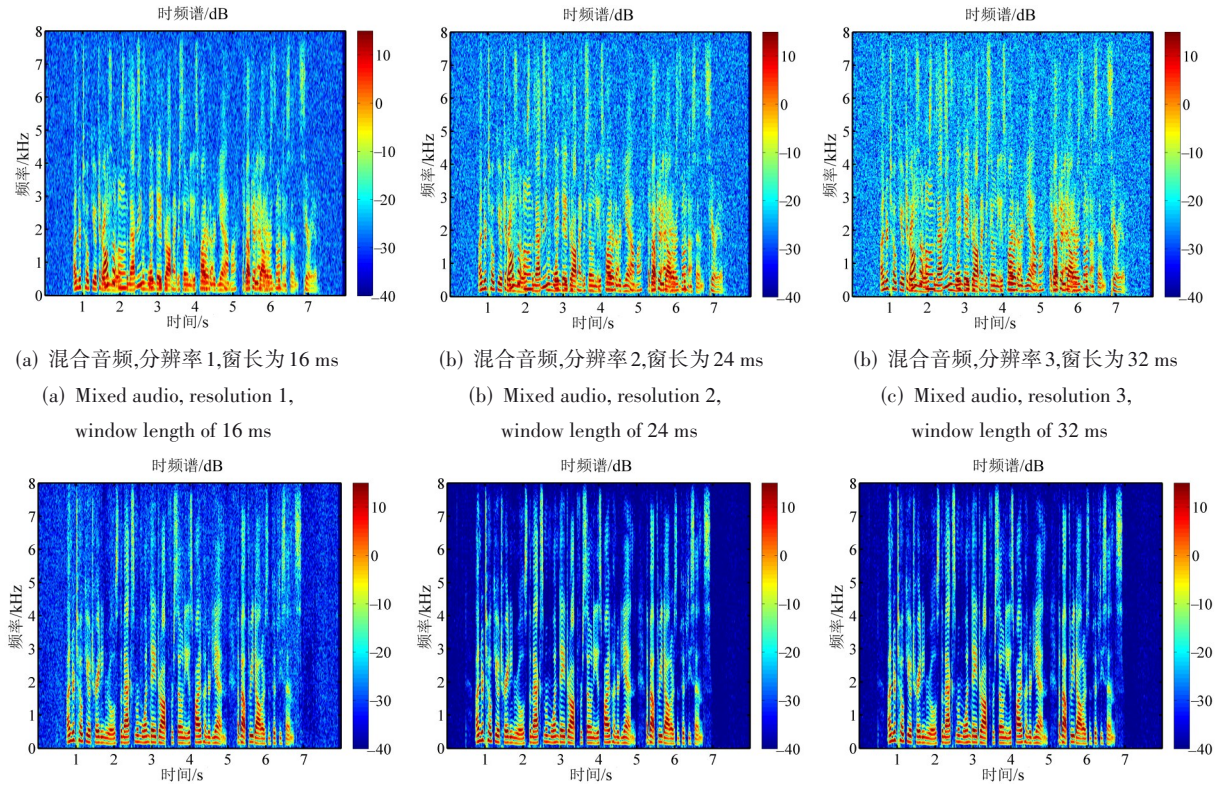
本小节实验为验证所提方法在带噪环境中的有效性，在所有麦克风通道中均加入了20 dB的等效白噪声，其他参数和训练配置保持一致。为更直观地展示不同分辨率下的噪声抑制效果，图2绘制各信号的

语谱图以进行说明。

在图2中,图(a)、(b)、(c)展示了对应分辨率的含噪混合语谱,也是网络的输入语谱;图(d)展示了基线方法JNF模型在输入单一分辨率(分辨率2)的语谱后,其输出分离结果的声源之一对应的语谱图;图(e)表示将两个分辨率(分辨率1和3,对应窗口16 ms和32 ms)输入网络后的输出结果;图(f)表示选取三个分辨率混合信号谱并行特征提取网络后,对应声源的分离结果。

从图2(d)中可以看出,在噪声场景中,虽然单分辨率JNF方法能够实现部分语音分离,但由于其仅利用单一尺度的STFT特征,因此在整体时频域范围内仍有部分噪声成分残留,导致无能量区域以及弱能量区域中可以看到明显的噪声能量残留,而基于第二节所提出的理论验证,即不同分辨率场景下的噪声呈现不相关趋势,因此进行多分辨率融合时可以有效抑制噪声的影响。另外,从图2(e)中也可以看出,经过多分辨率联合训练后,所提方法中的噪声成分相互抵消,而声源信号成分则因其较强的相关性而相互叠加,达到抑制噪声的目的。

图2(f)展示了带噪声场景下采用三分辨率并行特征提取网络的输出结果与对应的纯净语音时频谱对比情况。可以观察到,虽然仍旧可以发现部分噪声残留,但三分辨率并行特征提取的输出信号在各频带范围内均已大致与纯净语音近乎一致,且噪声抑制效果相较于双分辨率场景更为优秀,为了进一步说明所提方法的优势,表2展示了各个方法的量化指标及分析。



(a) 混合音频,分辨率1,窗长为16 ms　(a) Mixed audio, resolution 1, window length of 16 ms

(b) 混合音频,分辨率2,窗长为24 ms　(b) Mixed audio, resolution 2, window length of 24 ms

(b) 混合音频,分辨率3,窗长为32 ms　(c) Mixed audio, resolution 3, window length of 32 ms

(d) 分离声源,分辨率1,窗长为16 ms　(d) Separated source 1, resolution 2, window length of 24 ms

(e) 分离声源,分辨率1+3,窗长为16 ms和32 ms　(e) Separated source 1, resolution 1+3, window length of 16 ms and 32 ms

(f) 分离声源,分辨率1+2+3,窗长为16、24和32ms　(f) Separated source 1, resolution 1+2+3, window length of 16, 24, and 32 ms

图2　带噪场景下,多分辨率语谱图对比

Figure 2　Comparison of multi-resolution spectrograms in noisy environments

当在通道中叠加20 dB的等效白噪声后,所有模型的分离性能均出现不同程度的下降,但性能变化趋势具有明显的可比性。首先,经典分离方法Conv-TasNet的SI-SDR指标仅为8.51 dB,SAR为8.03 dB,STOI与PESQ分别下降至0.795与2.31,显示其无法完全摆脱噪声的影响,分离表现有所下降。相比之下,基于时频域建模的单分辨率JNF表现出更好的噪声抑制能力。三种窗长下的SI-SDR分别达到10.89、10.76和11.10 dB,表现出其时频递归结构在噪声条件下相对稳定,纵向对比表1和表2可以看出,单分辨

率JNF输出的结果仍旧保有噪声残留成分，导致其相较于无噪声场景下的表现平均下降2~3 dB。另外，还可以观察到不同窗长下JNF性能差异并不明显，其中分辨率1略优于分辨率2和3。当引入多分辨率联合学习后，性能提升趋势更加明显。最后，可以观察到三分辨率并行特征提取网络的输出指标优于其他情况，验证了在有噪环境下，多尺度网络可同时建立对短期瞬态噪声和宽带平稳噪声的双层抑制机制，优化网络在噪声影响下的多声源分离表现。为进一步验证模型在更极端噪声条件下的鲁棒性与泛化能力，我们在信噪比为–10~10 dB的实录语音混合噪声场景下进行了补充实验，结果如表3所示。当信噪比降至–10 dB时，所有方法性能均显著下滑，但所提出的多分辨率JNF架构依然保持了最优的综合性能。其中，三分辨率JNF在SI-SDR、SAR与STOI上均优于所有对比方法，尤其在低信噪比下较单分辨率JNF的SI-SDR提升约8.4%，表明多分辨率特征能有效增强模型对强噪声的抑制能力。

表2 20 dB带噪场景下的声源分离评估

Table 2 Evaluation of the separation results of two sound sources in a noisy environment of 20 dB

方法名称		评估标准			
		SI-SDR	SAR	STOI	PESQ
单分辨率方法	Conv-Tas-Net	8.512	8.030 0	0.795 04	2.306 1
	DPRNN	8.793	8.305 0	0.805 44	2.420 2
	Sepformer	10.692	10.683 0	0.851 60	2.754 1
	单分辨率JNF	10.758	10.779 0	0.883 82	3.314 4
双分辨率方法	双分辨率JNF	12.969	13.151 0	0.916 80	3.738 0
三分辨率方法	三分辨率JNF	**13.322**	**13.200 3**	**0.918 93**	**3.764 8**

注：加粗表示当前指标最优表现。

表3 不同信噪比场景下的声源分离评估

Table 3 Evaluation of the separation results of two sound sources in environments with different SNR

方法名称		评估标准							
		SI-SDR		SAR		STOI		PESQ	
信噪比		–10 dB	10 dB	–10 dB	10 dB	–10 dB	10 dB	–10 dB	10 dB
单分辨率方法	Conv-Tas-Net	4.125 0	6.231 0	4.730 0	6.722	0.715 04	0.799 8	1.906 1	2.183 0
	DPRNN	4.154 0	6.351 0	4.808 0	6.813	0.754 40	0.809 1	2.020 2	2.254 0
	Sepformer	8.516 5	9.997 0	8.175 0	10.025	0.801 00	0.849 5	2.154 1	2.520 0
	单分辨率JNF	8.818 2	10.074 0	8.395 0	10.208	0.815 00	0.848 4	2.814 4	3.144 0
双分辨率方法	双分辨率JNF	9.316 5	11.425 7	9.485 1	11.317	0.819 60	0.867 2	3.242 7	3.542 7
三分辨率方法	三分辨率JNF	**9.554 5**	**11.661 2**	**9.581 3**	**11.722**	**0.827 20**	**0.877 8**	**3.271 2**	**3.571 2**

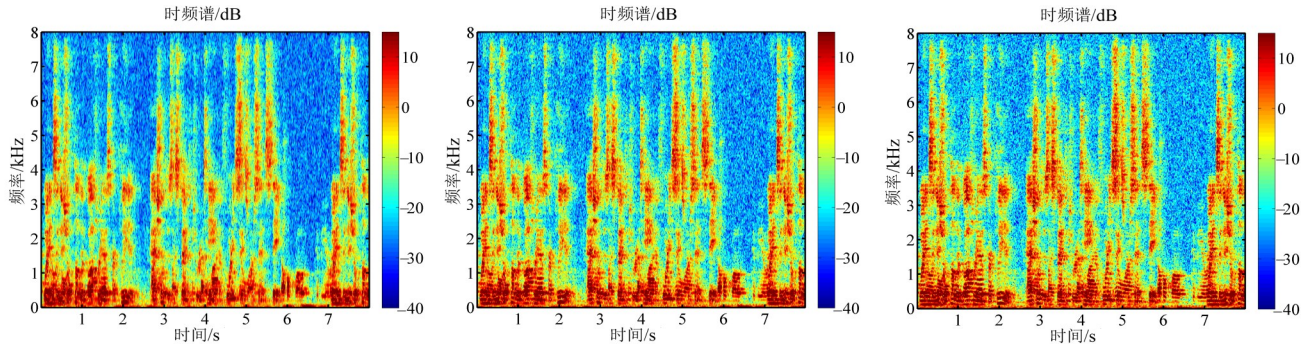
注：加粗表示当前指标最优表现。

需要说明的是，在–10 dB时，三分辨率JNF方法的PESQ相较于双分辨率JNF方法的提升有限，出现这一现象的原因可能在于多分辨率方法在整合极低信噪比的掩蔽信号时引入了部分噪声残余，影响了主观听感评分，但在更客观的SI-SDR与可懂度指标上，三分辨率方法仍具有优势。随着信噪比提升至10 dB，三分辨率JNF的各项指标进一步改善，尤其在PESQ与STOI上达到最优，说明其在中等噪声环境下能同时保障语音质量与可懂度。以上结果表明，多分辨率并行学习机制在不同强度噪声环境中均能保持稳定且优越的分离性能，验证了该方法在复杂真实场景中的泛化潜力。

3.3 带噪场景下的多声源分离结果评估

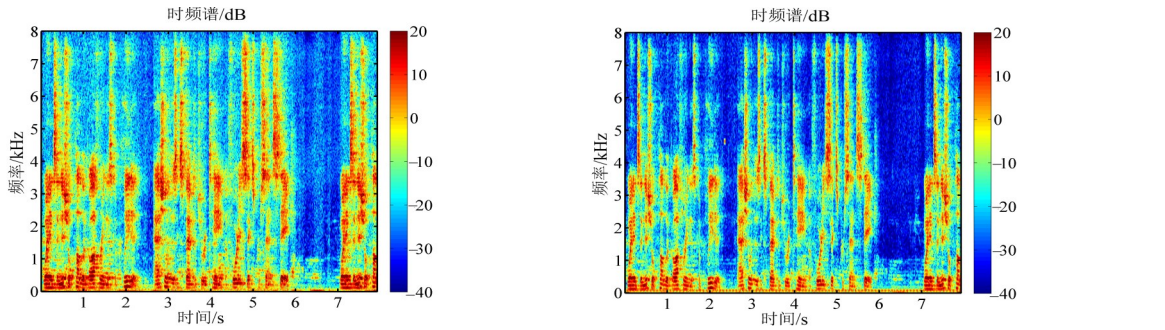
为进一步验证所提多分辨率并行特征提取分离网络在复杂声学环境中的有效性，本节在前述带噪实验的基础上引入房间混响效应，构建更符合真实录音条件的多声源混合场景。实验中所有参数与前两节保持一致，除噪声外，在声学仿真阶段对每组样本随机设置混响时间，其取值范围为$RT_{60} \in [0.2\,\text{s}, 0.8\,\text{s}]$。需要特别说明的是，为更贴近现实录音条件，麦克风通道的等效噪声计算中已包含混响回声成分。这意味着相对于带噪但无混响的环境，当前实验中的“等效噪声”不仅包括背景白噪声，还包含由房间反射产生的晚期混响能量，因此整体噪声能量更高。混响的引入会在时频域上造成语音能量的跨帧扩散和相位畸变，显著增加分离任务的难度。在这种条件下，传统的时域分离方法和单分辨率JNF模型往往难以同时消除加性噪声与晚期混响成分，输出的语谱在中高频区仍会出现能量拖影和频带模糊。因此，本节的目标通过比较单分辨率与多分辨率网络模型在随机混响时间下的表现，可以进一步验证多分辨率特征并行提取机制对于抑制混响，提升语音分离质量的有效性。

图3展示了同时存在噪声和混响的声学环境下，不同分辨率语音时频谱表示。图(a)、(b)、(c)为在相同三种分辨率下的混合语音谱，可以明显观察到混响和噪声共同作用带来的失真特征，即在时间与频率方向都出现模糊。图(d)则是以其中一个声源为例，



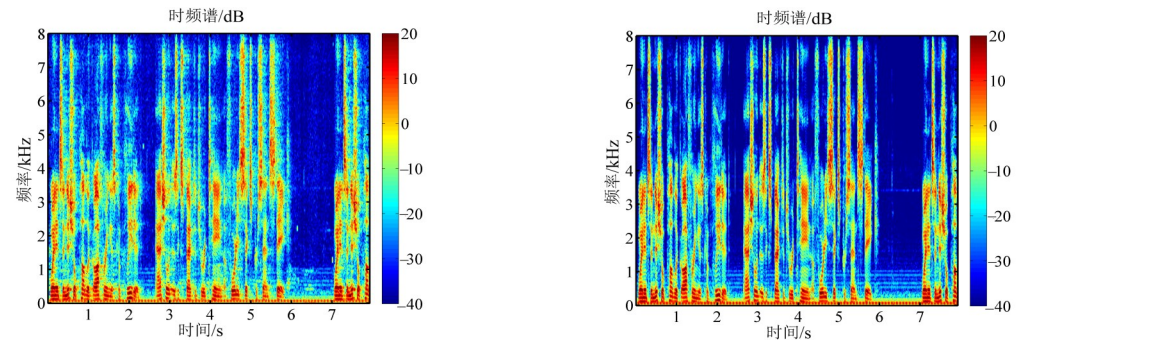
(a) 混合音频,分辨率1,窗长为16 ms　(b) 混合音频,分辨率2,窗长为24 ms　(b) 混合音频,分辨率3,窗长为32 ms

(a) Mixed audio, resolution 1, window length of 16 ms　(b) Mixed audio, resolution 2, window length of 24 ms　(c) Mixed audio, resolution 3, window length of 32 ms



(d) 分离声源1,分辨率2,窗长为24 ms　(e) 分离声源1,分辨率1+3,窗长为16 ms和32 ms

(d) Separated source 1,resolution 2,window length of 24 ms　(e) Separated source 1, resolution 1+3,window length of 16 ms and 32 ms



(f) 分离声源1,分辨率1+2+3,窗长为16、24和32 ms　(g) 纯净声源1,分辨率2,窗长为24 ms

(f) Separated source 1, resolution 1+2+3,window length of 16, 24, and 32 ms　(g) Clean source 1, resolution 2, window length of 24 ms

图3　混响与噪声场景下,多分辨率语谱图对比

Figure 3　Comparison of multi-resolution spectrograms in reverberant and noisy environments

展示了单分辨率谱输入网络后的输出分离结果,可以观察到网络能够大致恢复语音的主体结构,但是仍旧残留大量干扰成分,同时语音的部分细节也仍旧损失。相比之下,图3(e)显示,若采用两个分辨率的时频谱输入网络并进行并行特征提取与多声源分离,则可以获得相对更好的分离结果,其中一个最明显的结果即是在声源能量较低的区域,噪声的能量显著下降。低频位置处的语音脉络更为清晰,说明在多分辨率融合过程中网络学习了如何提取高分辨率频率信息,另一方面,同样可以观察到因拖尾而连接到一起的语音成分也几乎消失,这说明多分辨率联合网络在时间维度上能更有效地区分直达声与反射声,使得分离结果具有更为明确的语音起止边界。

总体而言,图3的结果进一步说明,在噪声与混响同时存在时,语音的时频失真更加复杂,单一尺度难以同时兼顾时域与频域特征,而所提方法通过融合不同窗长的特征输入,实现了多分辨率联合优化,在较为复杂的声学场景中能更准确地重现目标语音的结构。

图3(f)和(g)进一步展示了在带噪声与混响环境

下,将所有的三个分辨率一同输入网络进行并行特征提取与分离的输出结果与纯净语音时频谱的对比情况,可以观察到语谱结构得到更全面地恢复。然而,与纯净语音仔细比较仍可发现轻微差异,首先是噪声的残余比上一场景更多,另外,也出现了大量音频拖尾情况。这种现象表明,尽管联合多分辨率时频谱可以有效抑制各分辨率间相关性较低的噪声成分,但对于混响这种跨分辨率仍保有一定相干性的干扰成分,即便是使用了多分辨率进行分析,仍难以将其完全消除。另一方面,需要指出的是,本研究采用的分辨率范围主要集中在16~32 ms之间,而语音分离任务的有效分析窗口通常限定在20~30 ms左右。这一限制意味着多分辨率分支之间的窗口尺度跨度相对较小,融合特征虽在细节上起到补偿作用,但其整体处理能力依然受限。

表4展示了同时存在噪声和混响的场景下,各方法分离的评估指标,从中可以直观看到,各个模型的分离效果均有所下降。从经典时域分离网络来看,Conv TasNet和DPRNN在该复杂声场中的SI-SDR分别仅为3.51 dB和3.59 dB,说明单纯依靠时域卷积结构难以同时解决噪声叠加与混响拖尾干扰,而基于时频域递归建模的JNF方法则相对而言可以取得更好的效果,其单分辨率处理结果可以获得6.1 dB的SI-SDR评分,而将输入的分辨率增加至2时,则可以得到更优的声源分离结果,这源于多分辨率谱中噪声成分的不相干性与混响成分的弱相干性。

表4 带噪带混响场景下的声源分离评估

Table 4 Evaluation of the separation results of two sound sources in reverberant and noisy environments

方法名称		评估标准			
		SI-SDR	SAR	STOI	PESQ
单分辨率方法	Conv-Tas-Net	3.512 0	3.581 8	0.689 42	1.649 3
	DPRNN	3.596 8	3.597 0	0.680 50	1.631 0
	Sepformer	8.961 0	9.195 0	0.794 10	2.251 0
	单分辨率JNF	6.102 9	6.012 5	0.743 00	2.163 9
双分辨率方法	双分辨率JNF	9.087 0	9.427 0	**0.816 80**	2.924 1
三分辨率方法	三分辨率JNF	**9.336 7**	**9.605 9**	0.805 80	**2.946 4**

注:加粗表示当前指标最优表现。

进一步将输入的分辨率增加到3时,可以观察到所有指标的评分略有上升,但部分指标不再占据最优地位,这与图3展现出的现象一致。由于混响信号在不同窗长下的相干性较低,因此多分辨率整合后仍会残留少量混响成分,导致可懂度指标略微下降。也就是说,多分辨率模型在去噪和总体纯净度上的效果更好,但对混响反射的消除仍然存在一定限制。由于面向多声源分离场景的分析窗长仅在20~30 ms范围内,在分辨率数量大于或等于三个时,多分辨率时频谱上的冗余信息数量增加,最终导致整体算法的优化程度受限,甚至可能出现因为多分辨率混响残留的融合导致可懂度指标下降的情况,这一点仍需后续进行完善。

4 结论

本文围绕多通道语音分离任务,提出了一种基于多分辨率并行特征提取的多声源分离网络,通过引入多个不同长度的时频分析窗口构建混合信号的多分辨率时频谱,随后联合空间、时间与频率维度的深度神经网络模型生成具有互补特性的掩蔽,从而实现更优的多声源分离结果。

具体来说,所提网络通过为每个分辨率的时频谱输入设置独立的F-LSTM与T-LSTM分支,并在输出端以统一时域损失进行反向传播,使得多分辨率网络能够在训练过程中形成多分辨率联合优化机制。其中,短窗生成的高时域分辨率语谱分支旨在强化模型的瞬态特征提取能力,而长窗生成的高频域分辨率语谱分支则增强了稳态谐波还原,中等长度的窗对应的网络分支保证了整体能量的一致性,三者在融合层互为补偿,使模型在时频结构上获得最优平衡。损失函数采用时域与频域联合L1范数形式,不仅保证波形的重建精度,也提高了复谱幅相的一致性,从而在全局上增强了网络分离性能。实验证明,所提多分辨率融合网络通过结合多个维度的特征实现在复杂环境下的语音结构恢复与噪声抑制,为未来其他基于时频分析的多声源分离模型打下基础。

参考文献

[1] Turchet L, Fazekas G, Lagrange M, et al. The Internet of audio things: State of the art, vision, and challenges[J]. IEEE Internet of Things Journal, 2020, 7(10): 10233-10249.

[2] Turchet L, Fischione C, Essl G, et al. Internet of musical things: Vision and challenges[J]. IEEE Access, 2018, 6: 61994-62017.

[3] Turchet L, Casari P. The Internet of musical things meets satellites: Evaluating starlink support for networked music performances in rural areas[C]//2024 IEEE 5th International Symposium on the Internet of Sounds. Piscataway: IEEE, 2024: 10704207.

[4] Gabrielli L, Principi E, Turchet L. Sustainability and the Internet of sounds: Case studies[J]. IEEE Transactions on Technology and Society, 2025, 6(2): 165-180.

[5] Bosi M, Servetti A, Chafe C, et al. Experiencing remote

classical music performance over long distance: A JackTrip concert between two continents during the pandemic[J]. Journal of the Audio Engineering Society, 2021, 69(12): 934-945.

[6] Zlabinger T. Managing telematic pain: Migrating a student ensemble online during COVID[PP/OL]. V2. arVix (2024-10-15)[2025-12-20]. https://aes2.org/publications/elibrary-page/ id=21105.

[7] Chen Xi, Mo Yefei, Ouyang Kang, et al. Internet streaming audio based speech reception threshold measurement in cochlear implant users[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2022: 9012-9016.

[8] Hershey J R, Chen Zhuo, Le Roux J, et al. Deep clustering: Discriminative embeddings for segmentation and separation[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2016: 31-35.

[9] 张亚洲, 刘祈蒙, 戎璐, 等. 语音大模型: 架构、训练与挑战分析[J]. 电子学报, 2025, 53(9): 3454-3472.
Zhang Yazhou, Liu Qimeng, Rong Lu, et al. Speech large language models: Architecture, training and challenges analysis[J]. Acta Electronica Sinica, 2025, 53(9): 3454-3472. (in Chinese)

[10] 苏兆品, 周晓琳, 张国富, 等. 基于对抗学习和增强优化的深度转换语音还原方法[J]. 电子学报, 2025, 53(6): 1815-1828.
Su Zhaopin, Zhou Xiaolin, Zhang Guofu, et al. Adversarial learning and enhanced optimization based restoration method for VC-generated speeches[J]. Acta Electronica Sinica, 2025, 53(6): 1815-1828. (in Chinese)

[11] 周静, 鲍长春, 张旭. 基于聚焦信号子空间估计导向矢量的干扰声源抑制方法[J]. 电子学报, 2023, 51(1): 76-85.
Zhou Jing, Bao Changchun, Zhang Xu. Suppression method of the interference sound sources by estimated steering vector based on the focusing signal subspace[J]. Acta Electronica Sinica, 2023, 51(1): 76-85. (in Chinese)

[12] Yu Dong, Kolbæk M, Tan Zhenghua, et al. Permutation invariant training of deep models for speaker-independent multi-talker speech separation[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2017: 241-245.

[13] Luo Yi, Mesgarani N. TaSNet: Time-domain audio separation network for real-time, single-channel speech separation[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2018: 696-700.

[14] Wang Deliang, Chen Jitong. Supervised speech separation based on deep learning: An overview[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(10): 1702-1726.

[15] Delfarah M, Wang Deliang. Deep learning for talker-dependent reverberant speaker separation: An empirical study[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(11): 1839-1848.

[16] Wang Xianyun, Bao Changchun, Cheng Rui. IRM with phase parameterization for speech enhancement[C]//2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. Piscataway: IEEE, 2019: 209-213.

[17] Luo Yi, Mesgarani N. Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation[J]. ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(8): 1256-1266.

[18] Luo Yi, Chen Zhuo, Yoshioka T. Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation[C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2020: 46-50.

[19] Liu Debang, Zhang Tianqi, Wei Ying, et al. Speech convmamba: Selective structured state space model with temporal dilated convolution for efficient speech separation[J]. IEEE Signal Processing Letters, 2025, 32: 2015-2019.

[20] Siddiqua A, Basha C H, Abbas H M, et al. Real-time audio noise reduction and speech enhancement using LadderNet with hybrid spectrogram time-domain audio separation network[C]//2024 4th International Conference on Mobile Networks and Wireless Communications. Piscataway: IEEE, 2024: 10872040.

[21] Lin Jingru, Ge Meng, Ao J Y, et al. SA-WavLM: Speaker-aware self-supervised pre-training for mixture speech[C]//Interspeech 2024. ISCA, 2024: 597-601.

[22] Hsieh T A, Choi H, Kim M. Multimodal representation loss between timed text and audio for regularized speech separation[C]//Interspeech 2024. ISCA, 2024: 592-596.

[23] Li Xiaofei, Horaud R. Multichannel speech enhancement based on time-frequency masking using subband long short-term memory[C]//2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. Piscataway: IEEE, 2019: 298-302.

[24] Wood S U N, Stahl J K W, Mowlaee P. Binaural codebook-based speech enhancement with atomic speech presence probability[J]. IEEE/ACM Transactions on Audio,

Speech, and Language Processing, 2019, 27(12): 2150-2161.

[25] Sivaraman A, Wisdom S, Erdogan H, et al. Adapting speech separation to real-world meetings using mixture invariant training[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2022: 686-690.

[26] Tzinis E, Adi Y, Ithapu V K, et al. RemixIT: Continual self-training of speech enhancement models via bootstrapped remixing[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1329-1341.

[27] Aralikatti R, Boeddeker C, Wichern G, et al. Reverberation as supervision for speech separation[C]//ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2023: 10095022.

[28] Maciejewski M, Wichern G, McQuinn E, et al. WHAMR!: Noisy and reverberant single-channel speech separation [C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2020: 696-700.

[29] Saijo K, Ogawa T. Self-remixing: Unsupervised speech separation VIA separation and remixing[C]//ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2023: 10095596.

[30] Schimmel S M, Muller M F, Dillier N. A fast and accurate "shoebox" room acoustics simulator[C]//2009 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2009: 241-244.

[31] Subakan C, Ravanelli M, Cornell S, et al. Attention is all you need in speech separation[C]//ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2021: 21-25.

[32] Tesch K, Gerkmann T. Spatially selective deep non-linear filters for speaker extraction[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2023: 10096098.

作者简介

高 尚 男,1997年9月出生于北京市。于2019年获北京工业大学通信工程学士学位、2026年1月获北京工业大学电子科学与技术博士学位。主要研究方向为语音与音频信号处理,内容涉及多声源定位与分离。

E-mail: gaoshang9795@163.com

贾懋珅 男,1982年9月出生于河北省张家口市。于2010年7月获北京工业大学电路与系统博士学位,目前为北京工业大学教授、博士生导师。主持多项国家自然科学基金项目和北京市自然科学基金项目。主要研究方向为三维音频信号处理、多声道音频编码、声源定位与分离及声场重建技术。中国电子学会会员编号:E190011303S。

E-mail: jiamaoshen@bjut.edu.cn