基于双路注意力循环网络的轻量化语音分离

杨弋[1,2] 胡琦[1,2†] 张鹏远[1,2]

(1 中国科学院语言声学与内容理解重点实验室(声学研究所) 北京 100190)

(2 中国科学院大学 北京 100049)

摘要 提出了基于双路注意力循环网络（Dual-Path Attention and Recurrent Network, DPARNet）的轻量化语音分离模型。该模型由编码器、分离网络和解码器三部分组成。编码器使用子带处理的方法降低计算量；分离网络在各子带内使用双路注意力机制和双路循环网络结构对语音信号进行建模，提取深层次的特征信息并获得丰富的频谱细节；解码器将各子带信息进行合并还原出分离的目标信号。用尺度不变信损比（SI-SDR）提升、语音质量感知评估（PESQ）、短时客观可懂度（STOI）等多个客观指标对分离语音的质量和可懂度进行了评价。在仿真测试集和 LibriCSS 中句子级别测试集上的结果表明，提出的方法能够在保持语音分离性能的前提下，大幅减小模型复杂度。

关键词 语音分离, 轻量化模型, 深度神经网络, 双路网络, 自注意力网络

PACS 数 43.60, 43.72

Light-weight speech separation based on dual-path attention and recurrent neural network

YANG Yi[1,2] HU Qi[1,2†] ZHANG Pengyuan[1,2]

(1 *Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences* Beijing 100190)

(2 *University of Chinese Academy of Sciences* Beijing 100049)

Abstract A light-weight model based on dual-path attention and recurrent network (DPARNet), which is composed of an encoder, a separation network and a decoder, is proposed for speech separation. To alleviate the computation burden, sub-band processing approach is leveraged in the encoder. Dual-path attention mechanism and recurrent network structure are introduced in the separation network to model the speech signals in each sub-band, which facilitate extraction of deep feature information and rich spectrum details. The quality and intelligibility of separated speech are evaluated by several objective indexes including scale-invariant signal-to-distortion ratio (SI-SDR), perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI). It demonstrates that the proposed method largely reduces model complexity while maintains good separation

performance on a simulated test set and utterance-wise LibriCSS test set.

Keywords Speech separation, Light-weight model, Deep neural network, Dual path network, Self-attention network

PACS number 43.60, 43.72

† 通讯作者: 胡琦, huqi@hccl.ioa.ac.cn

引言

语音分离是指从一段包含背景噪声、混响和若干干扰说话人的混合信号中提取出目标说话人的过程。作为语音信号处理的前端技术，语音分离被广泛应用于听觉辅助、车载、多说话人会议转录等实际场景中[1-3]。为了设备的便携和实用性，语音分离算法通常运行在低计算资源平台。因此，分离性能和模型复杂度是衡量相关算法优劣的重要指标。

几十年来，学者们提出了多种语音分离方法。根据使用的传声器数目，语音分离任务可以分为单通道语音分离和多通道语音分离。对于单通道语音分离，经典的方法有计算听觉场景分析（Computational Auditory Scene Analysis, CASA）[4]和基于语音增强的方法，如非负矩阵分解（Non-negative Matrix Factorization, NMF）[5]等；对于多通道语音分离，通常使用波束形成算法[6]。然而在真实环境下，由于噪声、混响等不确定性因素，这些方法取得的效果都十分有限[3,7]。近年来，基于深度神经网络（Deep Neural Network, DNN）的语音分离取得了显著的进步。与传统方法相比，DNN 具有更强的表示学习能力，在不同的声学环境下，能学习到对任务更有利的深层特征[3,8]。

深度长短时记忆（Long Short-Term Memory, LSTM）网络可以对语音序列的上下文信息进行建模[9]，是主流语音分离模型中的一个重要结构[9-12]。其中，时域音频分离网络（time-domain audio separation network, TasNet）[12]直接对时域混合信号进行处理，取得了很好的分离效果。TasNet 由编码器模块、分离模块和解码器模块组成，在分离模块中，使用深度 LSTM 网络对时间依赖性进行建模。然而，由于每层 LSTM 都有较长的输入特征序列，传统的单路 LSTM 网络会带来上千万的参数量[13]。为了解决这一问题，双路神经网络被提出并获得广泛的应用[14-17]。双路循环网络（Dual-Path Recurrent Neural Network, DPRNN）[14]将二维特征序列分割为若干个带有交叠的块，首先使用块内 RNN（intra-chunk RNN）对块内信息进行建模，然后使用块间 RNN（inter-chunk RNN）对全局信息进行建模。假设语音序列的长度为 L，与单路网络相比，DPRNN 将输入特征长度由 $O(L)$ 减小到 $O(\sqrt{L})$ [14]，从而减小了模型参数量。

近年来，双路处理的思想被进一步应用于 Transformer 模型，并展示了它在语音分离任务上的有效性[15-17]。其中，文献 17 将双路 Transformer 网络与全带和子带处理相融合（Dual-Path Transformer based Full-band and Sub-band fusion Network, DPT-FSNet），在 2020 深度降噪（Deep Noise Suppression, DNS）数据集[18]上证明了双路网络可以对语音序列的全带信息和子带信息进行更好的建模。然而，DPT-FSNet 存在以下三点问题：

第一，其编码器和解码器模块中均使用了密集卷积层[19]。虽然密集卷积层可以起到缓解梯度消失、对特征进行重复利用的作用[19]，但会带来大量的参数量和计算量。以文献 17 中的实验设置为例：密集卷积层共包含 4 层卷积，每层卷积的输入维度分别为$C,2C,3C,4C$，输出维度为C，卷积核的大小为2×3，则其参数量$\Omega=(C\times C+C\times 2C+C\times 3C+C\times 4C)\times(2\times3)$，当$C=64$时，$\Omega\approx 246\text{k}$；第二，DPT-FSNet 的输入特征为短时傅里叶变换（Short-Time Fourier Transform, STFT）后的全频谱，其计算复杂度和 STFT 频率点数目相关：频率点数越大，计算复杂度就越高；第三，为了学习到更好的位置编码信息，DPT-FSNet 将 Transformer 中前馈网络的第一个线性层替换为循环网络层[17]，然而，它依然将双路 Transformer 模块作为整体设置迭代次数，没有进一步探究自注意力机制和循环网络结构给模型性能带来的影响。

为进一步提升语音分离性能、减小模型复杂度，在 DPT-FSNet 模型[17]的基础上，提出了基于双路注意力循环网络（Dual-Path Attention and Recurrent Network, DPARNet）的轻量化语音分离模型[1]。DPARNet 模型由编码器模块、双路处理模块、门控单元模块和解码器模块四部分组成，并具有如下特点：第一，在编/解码器中不使用密集卷积层以减小模型复杂度；第二，相关研究[20-21]表明，子带分析通过对频谱沿频率轴降维，能够有效降低计算量，并维持系统性能。因此，在 DPARNet 中，通过子带处理的方法，降低计算复杂度。具体而言，编码器将全频带 STFT 谱分割为若干子带，双路处理模块和门控单元模块对各子带分别进行处理，解码器则将分离后的所有子带进行合并还原为全频带谱；第三，使用交替连接的双路多头自注意力网络和双路循环网络结构。由于卷积和非线性激活函数的存在，语音信息的细节容易丢失，然而语音分离需要恢复时域信号，对细节的要求很高[22]。注意力机制可以在提取目标说话人语音时有选择性地使用编码特征并生成加权特征[22-23]，循环网络结构可以提取特征序列的上下文信息[9]，进而获得更丰富的语音细节，提升语音分离性能。

主要贡献如下：（1）将单通道 DPT-FSNet 降噪模型进行扩展，提出轻量化的多通道语音分离模型 DPARNet；（2）对密集卷积层的有无和子带处理中不同子带个数对 DPARNet 模型性能的影响进行了探究，使模型在保持语音分离性能的前提下，大幅减小其复杂度；（3）使用交替连接的双路多头自注意力网络和双路循环网络结构，并详细分析了这两个网络单元个数的选取对 DPARNet 模型性能的影响，进一步提升语音分离性能。

1 信号模型

考虑在远场环境中，使用P个通道的传声器阵列对包含C个说话人的语音进行分离，在 STFT 域，不同通道上的观测信号可以表示为：

[1] 核心代码位于 https://github.com/yangyi0818/DPARNet

$$\begin{aligned}\boldsymbol{Y}(t,f) &= \sum_{c=1}^{C}(\boldsymbol{S}(c,t,f)*\boldsymbol{h}_{\mathrm{d}}(c,t,f)+\boldsymbol{S}(c,t,f)*\boldsymbol{h}_{\mathrm{r}}(c,t,f))+\boldsymbol{N}(t,f)\\ &= \sum_{c=1}^{C}(\boldsymbol{X}_{\mathrm{d}}(c,t,f)+\boldsymbol{X}_{\mathrm{r}}(c,t,f))+\boldsymbol{N}(t,f)\\ &= \sum_{c=1}^{C}\boldsymbol{X}(c,t,f)+\boldsymbol{N}(t,f)\,,\end{aligned} \tag{1}$$

其中，$\boldsymbol{Y}(t,f)\in\mathbb{C}^{P\times T\times F}$ 是观测信号向量，$\boldsymbol{S}(c,t,f)$ 是说话人 c 的声源信号向量，$\boldsymbol{h}_{\mathrm{d}}(c,t,f)$ 是房间冲激响应（Room Impulse Response, RIR）的直达和早期混响部分向量，$\boldsymbol{h}_{\mathrm{r}}(c,t,f)$ 是 RIR 的晚期混响部分向量，$*$ 表示卷积运算，$\boldsymbol{X}(c,t,f)$，$\boldsymbol{X}_{\mathrm{d}}(c,t,f)$ 和 $\boldsymbol{X}_{\mathrm{r}}(c,t,f)$ 分别表示传声器接收到的第 c 个说话人的信号向量、信号向量中的直达和早期混响部分、信号向量中的晚期混响部分，$\boldsymbol{N}(t,f)$ 是噪声信号向量。$t\in 1,2,...,T$ 表示帧数，$f\in 1,2,...,F$ 表示频率点数，T 和 F 分别是语音的总帧数和总频率点数。

实验的目标是对观测到的多通道远场混合语音 $\boldsymbol{Y}(t,f)$ 进行分离，得到每一个说话人的早期混响语音 $\hat{\boldsymbol{X}}_{\mathrm{d}}(c,t,f)$。前期研究表明，当早期混响设置为 RIR 中前 50ms 信号时，会得到更好的分离语音质量[24]和语音识别性能[25]，因此本文使用的早晚期混响划分标准与其相同。为了清晰表述，下文将对 c,t,f 符号进行省略。

2 DPARNet 语音分离模型

图 1(a)展示了 DPT-FSNet[17]和提出的 DPARNet 的模型架构。它们均由编码器模块、双路处理模块、门控单元模块和解码器模块四部分组成。图 1(b)展示了 DPARNet 模型中双路注意力-循环网络单元的细节。

2.1 MIMMO

传统的多通道输入、多源单通道输出（Multi-channel Input, Single-channel Multi-source Output, MISMO）结构只对参考通道信号的分离结果进行估计。文献 26 提出了多通道输入、多源多通道输出（Multi-channel Input, Multi-channel Multi-source Output, MIMMO）结构，对所有通道的分离结果同时进行估计。与MISMO相比，由于可以充分利用不同通道间的相关性信息，MIMMO结构取得了更好的分离效果[26]。因此，DPARNet使用相同的MIMMO结构[26]。DPARNet模型的输入 $\boldsymbol{Y}^{\mathrm{in}}\in\mathbb{R}^{B\times 2P\times T\times F}$ 由混合语音 $\boldsymbol{Y}$ 的实部和虚部沿通道维度拼接组成，其中 B 为批处理大小。编码器模块（Encoder）将输入的复数谱映射到维度为 L 的特征空间：

$$\boldsymbol{R}=\mathrm{Encoder}(\boldsymbol{Y}^{\mathrm{in}})\in\mathbb{R}^{B\times L\times T\times F}, \tag{2}$$

双路处理模块和门控单元模块（Separator）提取编码特征的上下文信息和长时依赖信息，并进行进一步处理：

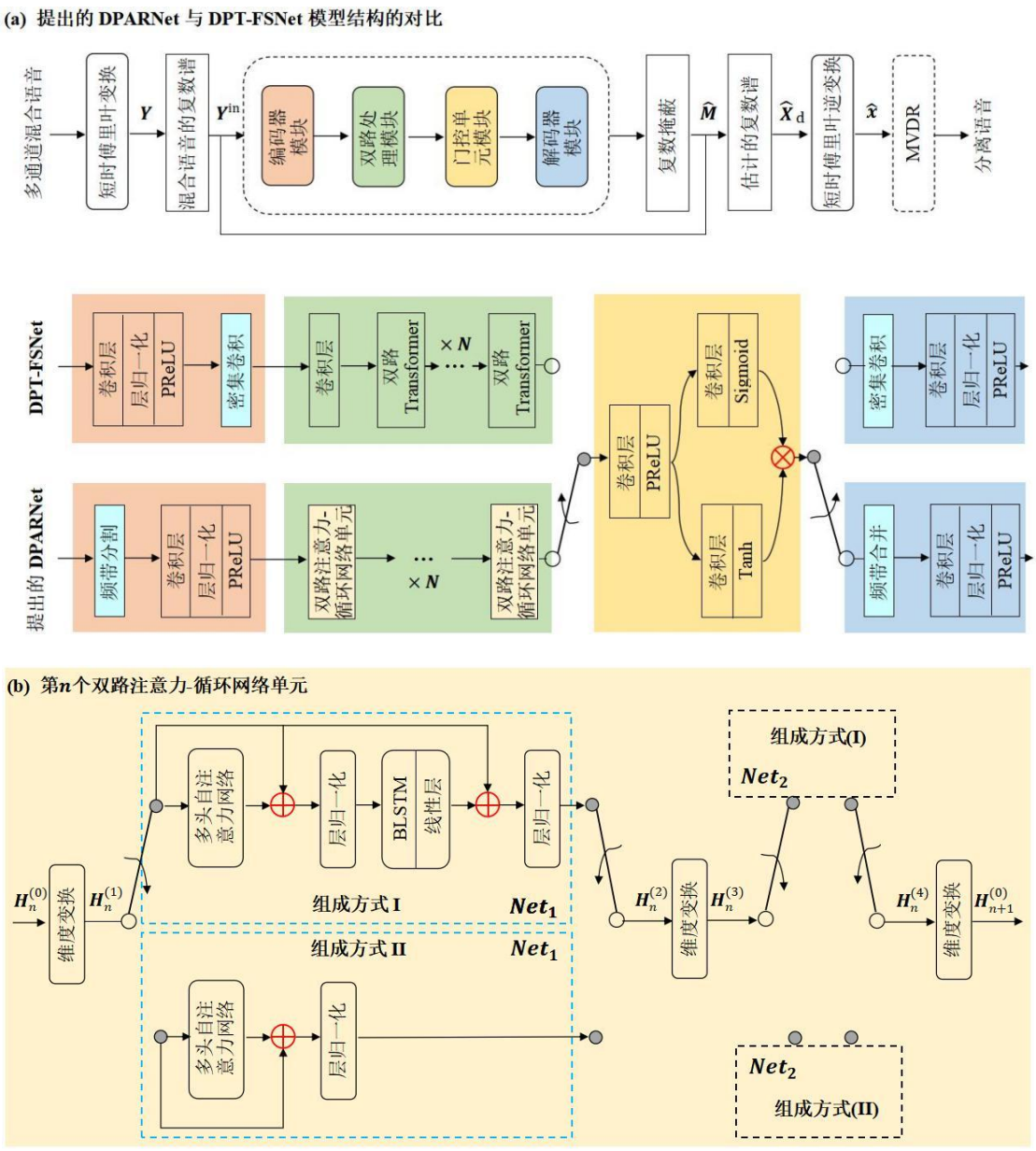

图 1 DPARNet模型架构，⊕和⊗分别表示加操作和逐元素相乘

$$\boldsymbol{H} = \text{Separator}(\boldsymbol{R}) \in \mathbb{R}^{B\times L\times T\times F}, \tag{3}$$

解码器模块（Decoder）估计所有说话人在所有通道上的复值掩蔽（Complex Ratio Mask, CRM）[27]：

$$\hat{\boldsymbol{M}} = \text{Decoder}(\boldsymbol{H}) \in \mathbb{C}^{B\times (P\times C)\times T\times F}, \tag{4}$$

CRM网络可以对语音谱的实部和虚部分别进行处理，同时实现目标语音幅度和相位的重构[27]。由CRM得到目标语音的复数谱：

$$\hat{\boldsymbol{X}}_{\text{re}} = \hat{\boldsymbol{M}}_{\text{re}}\boldsymbol{Y}_{\text{re}} - \hat{\boldsymbol{M}}_{\text{im}}\boldsymbol{Y}_{\text{im}}, \tag{5}$$

$$\hat{\boldsymbol{X}}_{\text{im}} = \hat{\boldsymbol{M}}_{\text{re}}\boldsymbol{Y}_{\text{im}} + \hat{\boldsymbol{M}}_{\text{im}}\boldsymbol{Y}_{\text{re}}, \tag{6}$$

其中下标 re 和 im 分别代表实部和虚部符号。对估计的复数谱进行短时傅里叶逆变换（Inverse Short-Time Fourier Transform, ISTFT）可以得到目标语音的时域信号 $\hat{\boldsymbol{x}}$。

2.2 子带处理

语音频谱在不同子带范围内的特性通常是不同的[28]，因此可以对不同子带分别进行处理以减小模型计算量。对于提出的 DPARNet 模型，编码器模块将输入的复数谱 $\boldsymbol{Y}^{\text{in}}(t,f)$ 沿频率维度分割为 K 个带宽相同的子带，每个子带可以表示为：

$$\boldsymbol{Y}^{(k)}(t,f)=\sum_{f_k=(F/K)\cdot(k-1)}^{(F/K)\cdot k}\boldsymbol{Y}^{\text{in}}(t,f_k), \tag{7}$$

其中 $k\in\{1,2,...,K\}$。双路处理模块和门控单元模块对这 K 个子带分别进行处理。解码器模块将不同子带的特征序列 $\boldsymbol{H}_k$ 合并还原为全频带的特征序列，即：

$$\boldsymbol{H}=\text{CAT}(\boldsymbol{H}_1,\boldsymbol{H}_2,...,\boldsymbol{H}_k), \tag{8}$$

其中 CAT 表示拼接操作。值得注意的是，为了模型的轻量化，DPARNet 没有对这 K 个子带间的相关性信息进行建模，因此当 K 增大时，模型计算速度会得到极大的提升，但其性能可能会降低，后续实验会对这两方面进行权衡。

2.3 双路注意力-循环网络单元

在双路处理模块中，DPARNet 使用 N 个串联的双路注意力-循环网络单元。其中前 N'（$N'\le N$）个网络单元由多头自注意力网络和循环网络交替连接组成（如图 1(b)中的组成方式 I），后 $(N-N')$ 个网络单元仅由多头自注意力网络组成（如图 1(b)中的组成方式 II），因此整个双路处理模块共包含 N 个双路多头自注意力单元和 N' 个双路循环网络单元。注意力机制可以在提取目标说话人语音时有选择性地使用编码特征并生成加权特征[22-23]；循环网络结构可以提取特征序列的上下文信息[9]，进而获得更丰富的语音细节，提升语音分离性能。本文使用的循环网络结构包括一层双向长短时记忆（Bi-directional Long Short-Term Memory, BLSTM）网络和一层线性（Linear）网络。与 DPT-FSNet 使用的双路 Transformer 相比，DPARNet 网络单元的结构更小，通过设计不同的 N 和 N' 值，可以清晰地研究注意力机制和循环网络结构对模型性能的影响，在减小网络结构冗余的同时，将性能调整到最优。

如图 1(b)所示，用 Net_1 和 Net_2 分别表示双路网络中的每一路网络。首先对第 n（$n\in\{1,2,...,N\}$）个网络单元的输入特征 $\boldsymbol{H}^{n,0}\in\mathbb{R}^{B\times L\times T\times F'}$（其中 $F'=F/K$）进行维度变换，得到 $\boldsymbol{H}^{n,1}\in\mathbb{R}^{(B\times T)\times F'\times L}$。$Net_1$ 对每一帧上的不同频率点进行操作，可以有效提取特征序列的全频带信息：

$$\boldsymbol{H}^{n,2}=Net_1(\boldsymbol{H}^{n,1})\in\mathbb{R}^{(B\times T)\times F'\times L}, \tag{9}$$

对 $\boldsymbol{H}^{n,2}$ 进行维度变换，得到 $\boldsymbol{H}^{n,3}\in\mathbb{R}^{(B\times F')\times T\times L}$。$Net_2$ 对每一频率点上的不同帧进行操作，可以有效提取特

征序列的子带信息：

$$\boldsymbol{H}^{n,4} = Net_2(\boldsymbol{H}^{n,3}) \in \mathbb{R}^{(B\times F')\times T\times L}, \tag{10}$$

对 $\boldsymbol{H}^{n,4}$ 进行维度变换，可以得到第 $n+1$ 个网络单元的输入 $\boldsymbol{H}^{n+1,0} \in \mathbb{R}^{B\times L\times T\times F'}$ 。以第一路网络 Net_1 为例，假设多头自注意力网络的输出向量为 $\boldsymbol{A}$ （多头自注意力网络的计算方法可以参考文献 29），则 Net_1 的输出可以表示为：

$$\boldsymbol{H}^{n,2} = \mathrm{LN}(\mathrm{Linear}(\mathrm{BLSTM}(\mathrm{LN}(\boldsymbol{A}+\boldsymbol{H}^{n,1}))) + \boldsymbol{H}^{n,1}), \quad \text{当组成方式为 I 时，} \tag{11}$$

$$\boldsymbol{H}^{n,2} = \mathrm{LN}(\boldsymbol{A}+\boldsymbol{H}^{n,1}), \quad \text{当组成方式为 II 时，} \tag{12}$$

其中，LN表示层归一化（Layer Normalization）。层归一化和残差连接的目的是加速模型的收敛。

需要指出的是，本文关注的是离线情况下的语音分离，对于在线处理，可以将 Net_2 中的自注意力网络和BLSTM网络替换为因果自注意力网络和LSTM网络，这种情况下， Net_2 仅对语音的第一帧到当前帧进行操作。具体方法可以参考文献 30 和文献 14。在线处理的相关研究不在本文讨论的范围之内。

3 实验结果及分析

3.1 数据集

所有实验均在仿真训练集下进行训练，并分别在仿真测试集和LibriCSS[2]中句子级别的测试集下进行测试。数据集的简介如表 1 所示。

表 1 数据集简介

数据	仿真训练集	仿真验证集	仿真测试集	LibriCSS测试集
通道数	7	7	7	7
每条语音中说话人数	1~2	1~2	1~2	1~2
重叠比例	[0,10%,...,100%]	[0,10%,...,100%]	[0,10%,...,100%]	[0S,0L,10%,...,40%]
总时长 (h)	~800	~100	~3.3	~10
总语句数	480k	60k	2k	5023

仿真数据由训练集、验证集和测试集 3 部分组成。使用的语音和噪声分别来自Librispeech[31]中train-clean-{100,360}子集、dev-clean子集、test-clean子集和在不同大小房间录制的真实噪声。噪声类型括电视噪声、空调噪声、脚步噪声和室外车辆噪声等。采用镜像源[32]的方法来模拟房间脉冲响应：对于每条语

音，仿真房间的大小范围（长×宽×高）为[(3m×4m×2.6m), (8m×11m×3.4m)]，混响时间（T60）的范围为[150ms, 600ms]，随机生成传声器阵列和说话人的位置。仿真数据中，说话人能量比例范围为[-5dB, +5dB]，信噪比范围为[5dB, 25dB]，不同说话人语音的重叠比例在[0,10%,...,100%]中均匀分布。选择的传声器阵列结构与文献 2 中所述相同：半径为 4.25cm的环形传声器阵列，其中 6 个传声器均匀地分布在圆环上，最后一个传声器位于圆环中央。LibriCSS[2]是在真实会议室中录制得到的数据，录制的语音来自LibriSpeech[31]中test-clean子集，说话人总个数为 40（男性、女性各 20 名），不同说话人语音的重叠比例范围为[0S,0L,10%,...,40%]，其中 0S和 0L子集均为单说话人语音（即：重叠比例为 0%）。

3.2 实验设置

训练时，每条输入语音截取的时长为 6s，采样率是 16000Hz。使用 512 点STFT，帧长为 32ms，帧移为 8ms。特征维度、自注意力网络的头数均与文献 17 中的设置相同，即 $L=64, h=4$，BLSTM网络的dropout比例为 0.4，模型中所有卷积网络的卷积核大小均为1×1。

使用Adam优化器[33]优化网络参数，初始学习率为 0.001，Epoch轮次为 60，且网络均达到收敛。使用基于信号逼近（Signal Approximation, SA）[34]和置换不变训练（Permutation Invariant Training, PIT）[11]的信噪比（Signal-to-Noise Ratio, SNR）[35]作为训练目标函数。SA的主要思想是最小化参考语音和由估计掩蔽得到的估计语音之间的差异，在有效改善语音可懂度和感知质量的同时，提升模型在不同SNR范围数据集上的鲁棒性[34]。PIT可以解决不同说话人的标签置换问题[11]。目标函数 $\mathcal{L}_{\mathrm{SA}}$ 的计算过程可以表示为：

$$R_{\mathrm{SN}}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = 10\log_{10}(\|\boldsymbol{x}\|^2 / \|\boldsymbol{x}-\hat{\boldsymbol{x}}\|^2), \tag{13}$$

$$\mathcal{L}_{\mathrm{SA}} = -\max_{\pi\in\mathcal{P}} \frac{1}{C}\sum_{c=1}^{C} R_{\mathrm{SN}}(\boldsymbol{x}(c), \hat{\boldsymbol{x}}_{\pi}(c)), \tag{14}$$

其中，$\boldsymbol{x}$ 和 $\hat{\boldsymbol{x}}$ 分别表示参考语音和估计语音的时域信号，π 表示 $C!$种置换方式 $\mathcal{P}$ 中的一种。

3.3 评价指标

为了衡量语音分离效果，本文使用四种客观评价指标：

（1）尺度不变信损比（Scale-Invariant Signal-to-Distortion Ratio, SI-SDR）[35]提升(Δ SI-SDR)，Δ SI-SDR的值越高表示分离效果越好。估计语音的SI-SDR的计算方式如下：

$$\begin{aligned} \boldsymbol{e}_{\mathrm{target}} &= \frac{\hat{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{x}}{\|\boldsymbol{x}\|^2}\boldsymbol{x}, \\ \boldsymbol{e}_{\mathrm{res}} &= \hat{\boldsymbol{x}} - \boldsymbol{e}_{\mathrm{target}}, \\ \mathrm{SI\text{-}SDR}(\hat{\boldsymbol{x}}, \boldsymbol{x}) &= 10\log_{10}(\|\boldsymbol{e}_{\mathrm{target}}\|^2 / \|\boldsymbol{e}_{\mathrm{res}}\|^2), \end{aligned} \tag{15}$$

将式 15 中估计语音的时域信号 $\hat{\boldsymbol{x}}$ 替换为混合语音的时域信号 $\boldsymbol{y}$，可以得到混合语音的SI-SDR。将估计语音与混合语音的SI-SDR做差，可以得到Δ SI-SDR；

（2）语音质量感知评估（Perceptual Evaluation of Speech Quality, PESQ）[36]，范围为[-0.5,4.5]，分数越

高表示语音听感越好；

（3）短时客观可懂度（Short-Time Objective Intelligibility, STOI）[37]，范围为[0,1]，分数越高表示语音可懂度越高。

（4）词错误率（Word Error Rate, WER），WER越低表示语音分离对识别任务越有利。WER的定义如下：

$$\mathrm{WER} = (S + D + I) / N_{\mathrm{words}}. \tag{16}$$

其中，S, D, I 和 N_{words} 分别表示替换错、删除错、插入错和总词数。由于LibriCSS句子级别测试集没有对应的参考语音，因此只对分离语音的WER进行测试。

为了与基线模型进行更好的比较，本文中ΔSI-SDR、PESQ和STOI在BSS_EVAL[38]工具箱下进行计算，WER在文献 2 提供的Kaldi脚本[2]下进行计算。

3.4 实验结果

实验共分为 3 部分，分别是：（1）双路处理模块中双路多头自注意力单元的总个数 N 和双路循环网络单元的总个数 N' 对DPARNet模型性能影响的研究；（2）子带个数 K 和密集卷积层对DPARNet模型性能影响的研究；（3）提出的DPARNet模型与目前取得先进水平的分离模型的对比。

实验（1）：双路处理模块中双路多头自注意力单元的总个数 N 和双路循环网络单元的总个数 N' 对DPARNet 模型性能影响的研究。共设置五组实验参数：① $N=2, N'=1$；② $N=3, N'=1$；③ $N=4, N'=0$；④ $N=4, N'=1$；⑤ $N=4, N'=2$。这五组实验中，子带个数 $K=2$，编/解码器模块不使用密集卷积层。在仿真测试集上分离语音的ΔSI-SDR、PESQ 和 STOI 的对比分别如图 2(a)、(b)、(c)所示，在 LibriCSS 测试集上分离语音的 WER、模型参数量和计算量的对比如表 2 所示。可以得到以下 4 个结论：

（1）通过对比图 2 的①、②、④实验，发现增加双路多头自注意力单元的总个数 N 有助于提升分离语音的ΔSI-SDR、PESQ 和 STOI 指标，表明尽管循环网络可以对语音序列建模，但对于目标说话人语音分离，注意力机制可以对上下文信息进行重要性评估、强化模型对这部分信息的利用能力，从而提高模型分离性能；

（2）通过对比图 2 的③、④、⑤实验，发现增加双路循环网络单元的总个数 N' 对提高分离语音质量也具有一定的帮助，表明 BLSTM 网络在提取特征序列上下文信息上的有效性。然而，增加双路循环网络单元的总个数 N' 对分离语音质量的提升没有增加双路多头自注意力单元的总个数 N 那样显著：当 N'=1时，N 由 3 增加到 4（实验②→实验④），ΔSI-SDR、PESQ 和 STOI 分别提升 1.6dB、0.062 和 0.029；当 N=4 时，N' 由 1 增加到 2（实验④→实验⑤），ΔSI-SDR、PESQ 和 STOI 分别提升 0.2dB、0.027 和 0.010。这说明注意力机制可以通过少量的循环网络单元挖掘出更多有效的目标语音频谱细节，更多的循环网络单元只能提供有限的有效信息，同时增加了模型训练的困难度；

（3）在分离语音的识别性能上，可以得到与上述相似的结论。不同之处在于通过对比表 2 的④、⑤实验，发现将双路循环网络单元的总个数 N' 由 1 增加至 2，识别性能会降低，说明分离语音的识别性能与听

[2] https://github.com/chenzhuo1011/libri_css

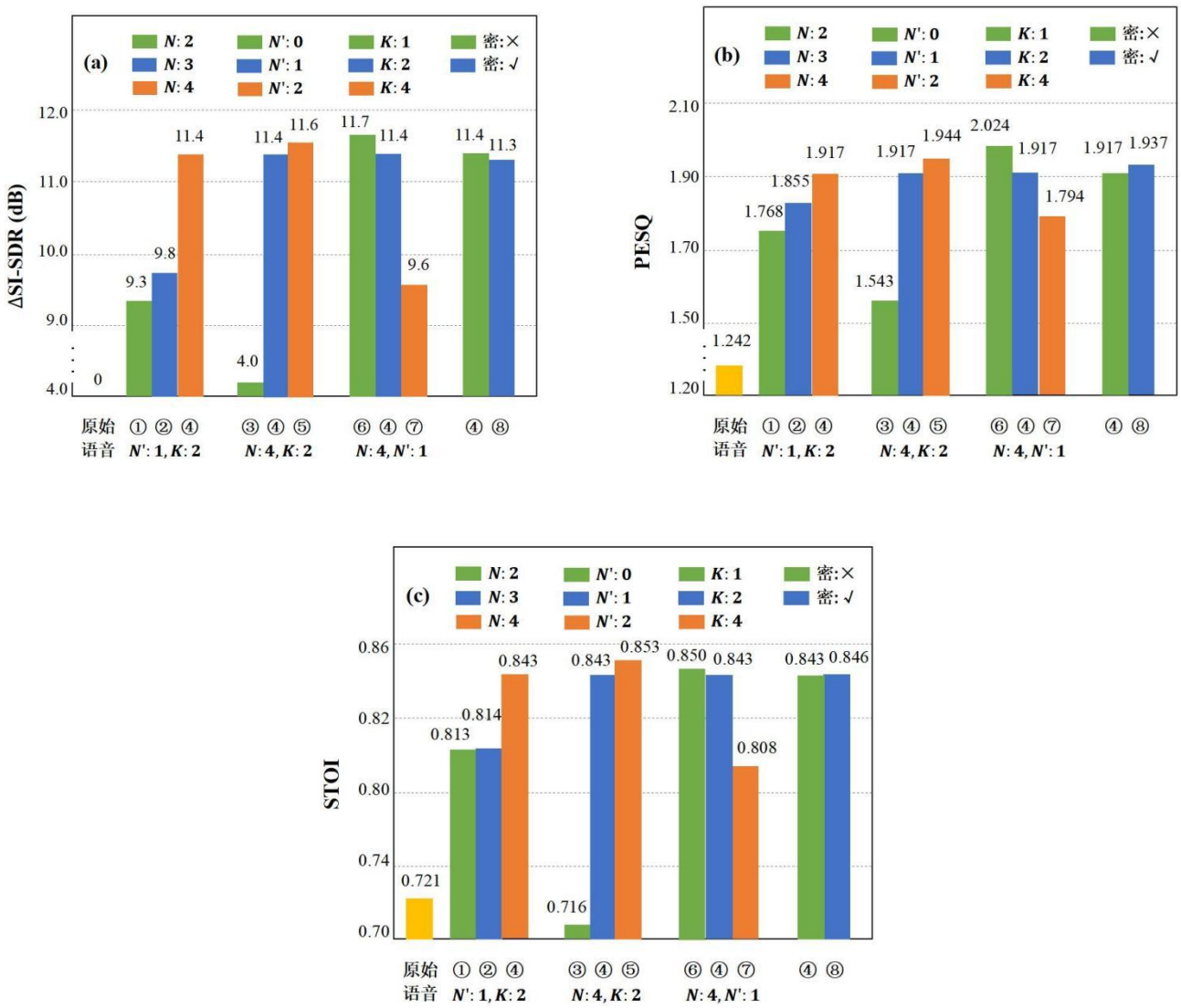

图 2：不同参数设置对DPARNet模型性能的影响：在仿真测试集上的

(a) Δ SI-SDR（dB）↑ (b)PESQ ↑ 和(c)STOI ↑

感质量等指标没有绝对意义上的相关性；

（4）通过对比表 2 的③、④、⑤实验，发现增加双路循环网络单元的总个数 N' 会带来很大的额外计算量：平均每层双路循环网络的计算量为 10.5G/6s（为了方便比较，这里指出单通道 DPRNN 降噪模型的计算量为 33.2G/6s[14]）。

为了更直观地比较不同网络单元个数对模型性能带来的影响，图 3 以 LibriCSS 的一条测试语音（女-女混合，重叠比例为 40%，说话人能量比例约为-2dB，SNR 约为 15dB）为例，对分离语音的语谱图进行对比。图 3(f)展示了图 3(b-e)红框部分的频谱，可以发现随着 N 和 N' 值的增大，模型可以去除更多的干扰信号，并恢复更多的频谱细节，表明循环网络和注意力机制能够获得更多的频谱有效信息[23]。

综合考虑模型的语音分离效果和复杂度，在后续研究中，N 和 N' 分别取值为 4 和 1。

表 2 不同参数设置对 DPARNet 模型性能的影响：在 LibriCSS 上的 WER（%）↓、模型参数量（MiB）和计算量（G/6s）

实验序号		参数设置				LibriCSS 测试集重叠比例 (%)						参数量	计算量
		N	N'	K	密集层	0S	0L	10	20	30	40		
原始语音		—	—	—	—	11.8	11.7	18.8	27.2	35.6	43.3	—	—
实验（1）	①	2	1	2	×	7.4	7.4	7.7	9.1	11.7	13.1	0.14	13.7
	②	3	1	2	×	7.3	7.3	7.7	9.0	11.1	11.2	0.14	14.4
	③	4	0	2	×	7.9	8.0	10.1	14.6	19.6	25.0	0.04	4.7
	④	**4**	**1**	**2**	×	**7.2**	**7.2**	**7.4**	**8.6**	**10.3**	**10.9**	**0.15**	**15.2**
	⑤	4	2	2	×	7.3	7.5	7.7	9.1	10.8	12.1	0.25	25.7
实验（2）	⑥	4	1	1	×	7.2	7.3	7.6	8.8	10.5	11.3	0.15	30.1
	⑦	4	1	4	×	7.4	7.4	8.4	10.4	12.5	13.7	0.15	7.8
	⑧	4	1	2	√	**7.1**	**7.2**	**7.3**	**8.4**	**9.9**	**10.8**	0.64	62.8

实验(2)：子带个数 K 和密集卷积层对 DPARNet 模型性能影响的研究。共设置三组实验参数：⑥ $K=1$（即不使用子带处理方法）且不使用密集卷积层；⑦ $K=4$ 且不使用密集卷积层；⑧ $K=2$ 且使用密集卷积层。在仿真测试集上分离语音的 Δ SI-SDR、PESQ 和 STOI 的对比分别如图 2(a)、(b)、(c)所示，在 LibriCSS 测试集上分离语音的 WER、模型参数量和计算量的对比如表 2 所示。可以得到以下 3 个结论：

（1）通过对比图 2 的④、⑦实验，发现当子带个数 $K=4$ 时，与 $K=2$ 相比，分离语音的 Δ SI-SDR、PESQ 和 STOI 下降较多（1.8dB、0.123 和 0.035）；通过对比图 2 的④、⑥实验，发现当不使用子带处理方法时，与 $K=2$ 相比，Δ SI-SDR 和 STOI 的提升幅度较小（0.3dB 和 0.007），但 PESQ 提升较为显著（0.107）；通过对比表 2 的④、⑥、⑦实验，发现当 $K=2$ 时，分离语音的识别效果最好；

（2）通过对比表 2 的④、⑥、⑦实验，发现当 $K=1/2/4$ 时，处理一条长度为 6s 的语音，模型计算量分别为 30.1/15.2/7.8G，即子带个数每增加一倍，计算量会减小约一倍；

以上两点说明：在DPARNet模型中，当子带个数 $K=2$ 时，即使在没有对各子带相关性信息进行建模的情况下，分离语音的 Δ SI-SDR和STOI值的下降也并不明显，识别性能反而有所提升。同时，以牺牲较小的语音分离性能为代价，模型的计算速度得到了极大的提高。这无论是在低计算资源或是在线应用的真实场景下，都是十分必要的。值得指出的是，当使用子带处理方法时，PESQ的值会较大幅度减小，这可能是由于在未考虑各子带相关性的情况下，不同子带间的降噪效果存在差异。为了证明这一猜想，图 4 以一条仿真语音为例，对 $K=1/2/4$ 时得到的分离语音的语谱图进行了对比：观察图 4(c)可知，分离语音中高频子带（4kHz~8kHz）和低频子带（0~4kHz）有着明显不同的残留噪声信号分布；图 4(d)中，频带 0~2kHz、2kHz~4kHz、4kHz~6kHz、6kHz~8kHz的残留噪声信号分布也存在差异。文献 39 使用LSTM网络对各子带相

关性信息进行建模，在未来的研究中会考虑使用更轻量化的网络得到这一信息，增强模型对语音的建模能力。

（3）通过对比图 2 和表 2 的④、⑧实验，发现去除编/解码器模块中的密集卷积层对分离语音的 Δ SI-SDR、PESQ、STOI和WER没有明显影响（11.3dB→11.4dB, 1.937→1.917,0.846→0.843 和 10.9%→10.8%）。去除该层会使模型参数量和计算量分别相对减小 4.3 倍和 4.1 倍（0.64MiB→0.15MiB和 62.8G/6s→15.2G/6s），可以很大程度地节约模型的存储资源和计算资源。

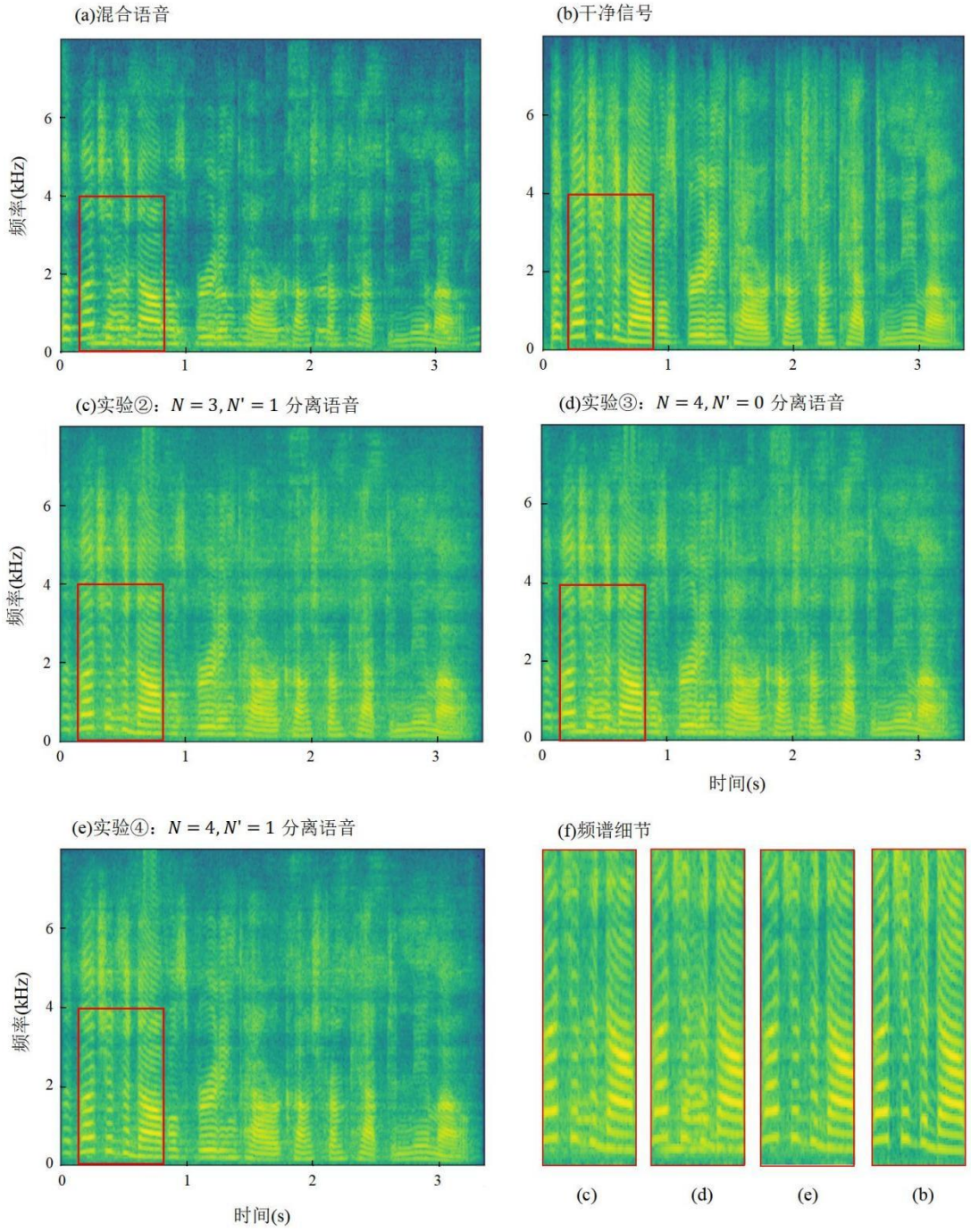

图 3：语谱图分析

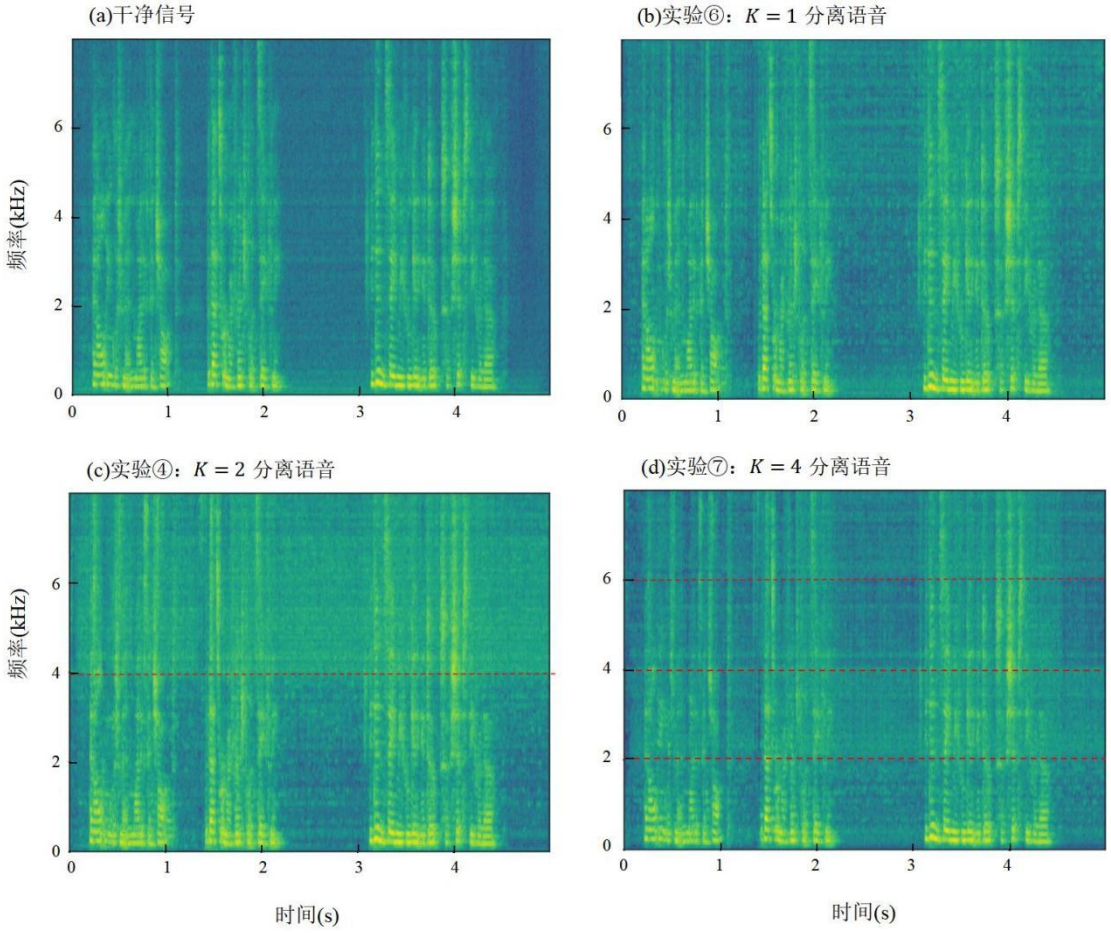

图 4：语谱图分析

因此，通过实验（1）和实验（2）的分析，综合考虑模型的语音分离效果和复杂度，本文提出的DPARNet模型的参数设置为：$N=4, N'=1, K=2$ 且不使用密集卷积层。

实验（3）：提出的DPARNet模型与目前取得先进水平的分离模型的对比。本文选择的基线模型包括：①三层BLSTM分离模型[40]；②基于成对窄带深度滤波（pairwise narrow band deep filtering, PW-NBDF）[41]的BLSTM分离模型。PW-NBDF方法将混合语音中参考通道的复数谱分别与其它 $P-1$ 个通道的复数谱组合成复数谱对，并分别输入同一个BLSTM网络进行特征提取，将得到的 $P-1$ 个输出的平均值作为BLSTM分离模型的输入特征。这一方法可以更好地提取传声器阵列的固有空间特征[41]；③Conformer-large模型[40]；④DPT-FSNet模型[17]；⑤使用Beam-Guided策略[26]的两阶段DPT-FSNet模型。该策略使用波束形成信息指导语音分离过程：第一阶段利用混合语音估计初始分离语音 $\hat{z}^{(1)}$，第二阶段利用混合语音和 $\hat{z}^{(1)}$ 经最小方差无失真响应（Minimum Variance Distortionless Response, MVDR）处理后的结果 $\hat{x}^{(1)}$ 估计第二阶段经一次迭代的分离语音 $\hat{z}^{(2:1)}$。同样地，利用混合语音和 $\hat{x}^{(2:n-1)}$ 可以估计第 n 次迭代的分离语音 $\hat{z}^{(2:n)}$，在本文中，训练阶段 n=2，测试阶段 n=1。为了与两阶段DPT-FSNet模型进行公平的比较，将其使用的策略[26]引入本文提出的

DPARNet模型中，并命名为Beam-Guided DPARNet。各模型在LibriCSS上分离语音WER的对比、在仿真测试集上分离语音PESQ和STOI的对比以及模型参数量和训练阶段计算量的对比如表 3 所示。

表 3 提出的 DPARNet 模型与目前取得先进水平的分离模型的对比：在 LibriCSS 上的 WER（%）↓、在仿真测试集上的 PESQ ↑ 和 STOI ↑、模型参数量（MiB）和训练阶段计算量（G/6s）

模型	年份	LibriCSS 测试集重叠比例 (%)						仿真测试集		参数量	计算量
		0S	0L	10	20	30	40	PESQ	STOI		
原始语音	2020	11.8	11.7	18.8	27.2	35.6	43.3	1.242	0.721	—	—
①BLSTM	2021	**7.0**	7.5	10.8	13.4	16.5	18.8	—	—	21.8	17.1
②PW-NBDF	2021	7.3	7.3	8.3	10.6	13.4	15.8	1.445	0.799	18.9	20.1
③Conformer-large	2021	7.2	7.5	9.6	11.3	13.7	15.1	—	—	58.7	43.6
④DPT-FSNet	2022	7.1	7.3	7.6	8.9	10.8	11.3	1.851	0.847	0.50	49.1
提出的 DPARNet	—	7.2	**7.2**	**7.4**	**8.6**	**10.3**	**10.9**	**1.917**	**0.843**	**0.15**	**15.2**
⑤两阶段DPT-FSNet[3]	2022	**7.1**	7.1	**7.1**	8.0	9.2	9.7	1.890	0.853	1.0	50.1
Beam-Guided DPARNet	—	7.3	**6.9**	7.2	**7.7**	**9.0**	**9.4**	**2.201**	**0.890**	**0.41**	**41.1**

通过对比，可以得到如下结论：第一，提出的DPARNet模型在LibriCSS测试集上的识别性能和在仿真测试集上的分离性能都优于基线DPT-FSNet模型，证明本文使用的方法具有优越性；第二，与两阶段DPT-FSNet模型⑤相比，Beam-Guided DPARNet在LibriCSS测试集重叠比例为 40%子集上的WER相对降低了 3.1%（9.7%→9.4%）。此外，与单阶段DPARNet模型相比，Beam-Guided DPARNet在仿真测试集上的PESQ和STOI指标也得到了进一步提升（1.917→2.201，0.843→0.890）；第三，提出的DPARNet模型参数量只有 0.15MiB，远小于其它基线模型，这有利于后续的低内存下的在线语音分离研究。

4 结论

针对多通道远场语音分离，将单通道 DPT-FSNet 降噪模型进行扩展，提出了基于双路注意力循环网络的轻量化语音分离模型 DPARNet，并对减小模型复杂度的方法进行了探究；在分离网络中使用交替连接的双路多头自注意力网络和双路循环网络结构：循环网络结构可以有效提取特征序列的上下文信息，注意力机制对上下文信息进行重要性评估并强化模型对这部分信息的利用能力。实验结果表明，在保持语音分离性能的同时，DPARNet 模型的参数量只有 0.15MiB。与目前取得先进水平的分离模型相比，DPARNet 模型

[3] https://github.com/hangtingchen/Beam-Guided-TasNet/blob/main/INTERSPEECH_2022_Appendix.pdf

在分离语音的 PESQ、STOI 和识别性能上均具有一定的优越性。在未来的工作中，将研究 DPARNet 模型在在线语音分离上的应用。

参考文献

1 王泽林，陈锴，卢晶. 车载场景结合盲源分离与多说话人状态判决的语音抽取. 声学学报，2020; **45**(5): 696-706

2 Chen Z, Yoshioka T, Lu L *et al*. Continuous speech separation: dataset and analysis. Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2020: 7284-7288

3 Wang D L, Chen J T. Supervised Speech Separation Based on Deep Learning: An Overview. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2018; **26**(10): 1702-1726

4 Cooke M. Modelling auditory processing and organisation. Cambridge University Press, 2005:7

5 Schmidt M N, Olsson R K. Single-channel speech separation using sparse non-negative matrix factorization. Proc. Interspeech, 2006

6 Van Veen B D, Buckley K M. Beamforming: a versatile approach to spatial filtering. *IEEE ASSP Magazine*, 1988; **5**(2): 4-24

7 Cooke M, Hershey J R, Rennie S J. Monaural speech separation and recognition challenge. *Comput. Speech Lang.*, 2010; **24**(1): 1-15

8 刘文举，聂帅，梁山等. 基于深度学习语音分离技术的研究现状与进展. 自动化学报，2016; **42**(6): 819-833

9 Heymann J, Drude L, Haeb-Umbach R. Neural network based spectral mask estimation for acoustic beamforming. Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2016: 196-200

10 Chen J T, Wang D L. Long short-term memory for speaker generalization in supervised speech separation. Proc. Interspeech, 2016: 3314-3318

11 Kolbaek M, Yu D, Tan Z-H *et al*. Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2017; **25**(10): 1901-1913

12 Luo Y, Mesgarani N. TasNet: Time-domain audio separation network for real-time, single-channel speech separation. Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2018: 696-700

13 Luo Y, Han C, Mesgarani N. Ultra-Lightweight Speech Separation Via Group Communication. Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2021: 16-20

14 Luo Y, Chen Z, Yoshioka T. Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation. Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2020: 46-50

15 Chen J J, Mao Q, Liu D. Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation. Proc. Interspeech, 2020: 2642-2646

16 Zhang Z N, He B S, Zhang Z J. Transmask: A compact and fast speech separation model based on transformer. Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2021: 5764-5768

17 Dang F, Chen H T, Zhang P Y. DPT-FSNet: Dual-path Transformer Based Full-band and Sub-band Fusion Network for Speech Enhancement. Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2022: 6857-6861

18 Reddy C K, Beyrami E, Dubey H *et al*. The interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework. arXiv preprint:2001.08662, 2020

19 Huang G, Liu X, Maaten L *et al*. Densely connected convolutional networks. Proc. IEEE conference on computer vision and pattern recognition., 2017: 4700-4708

20 Li X F, Horaud R. Online Monaural Speech Enhancement Using Delayed Subband LSTM. Proc. Interspeech, 2020: 2462-2466

21 Yang G, Yang S, Liu L *et al*. Multi-Band Melgan: Faster Waveform Generation For High-Quality Text-To-Speech. IEEE Spoken Language Technology Workshop (SLT), 2021: 492-498

22 武瑞沁，陈雪勤，俞杰等. 结合注意力机制的改进 U-Net 网络在端到端语音增强中的应用. 声学学报，2022; **47**(2): 266-275

23 蓝天，惠国强，李萌等. 采用上下文相关的注意力机制及循环神经网络的语音增强方法. 声学学报，2020; **45**(6): 897-905

24 Li X, Li J F, Yan Y H. Ideal Ratio Mask Estimation Using Deep Neural Networks for Monaural Speech Segregation in Noisy Reverberant Conditions. Proc. Interspeech, 2017: 1203-1207

25 Sehr A, Habets E, Maas R *et al*. Towards a better understanding of the effect of reverberation on speech recognition performance. Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC), 2010

26 Chen H T, Yang Y, Dang F *et al*. Beam-Guided TasNet: An Iterative Speech Separation Framework with Multi-Channel Output. Proc. Interspeech, 2022: 866-870

27 Williamson D S, Wang Y X, Wang D L. Complex ratio masking for monaural speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2015; **24**(3): 483-492

28 Takahashi N, Mitsufuji Y. Multi-Scale multi-band densenets for audio source separation. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017; 21-25

29 Vaswani A, Shazeer N, Parmar N *et al*. Attention is all you need. Advances in neural information processing systems, 2017; 5998-6008

30 Pandey A, Wang D L. Dense CNN With Self-Attention for Time-Domain Speech Enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2021; **29**: 1270-1279

31 Panayotov V, Chen G, Povey D *et al*. Librispeech: An ASR corpus based on public domain audio books. Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2015: 5206-5210

32 Alien J B, Berkley D A. Image method for efficiently simulating small - room acoustics. *J. Acoust. Soc. Am.*, 1976; **60**(S1): S9-S9

33 Kingma D P, Ba J. Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR, 2015

34 Zhang X-L, Wang D L. A Deep Ensemble Learning Method for Monaural Speech Separation. *IEEE/ACM Trans.*

Audio Speech Lang. Process., 2016; **24**(5): 967-977

35 Roux J L, Wisdom S, Erdogan H *et al*. SDR - Half-baked or Well Done?. Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2019: 626-630

36 Rix A W, Beerends J G, Hollier M P *et al*. Perceptual evaluation of speech quality (PESQ): a new method for speech quality assessment of telephone networks and codecs. Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2001; 2: 749-752

37 Taal C H, Hendriks R C, Heusdens R *et al*. A short-time objective intelligibility measure for time-frequency weighted noisy speech. Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2010: 4214-4217

38 Vincent E, Gribonval R, Fevotte C. Performance measurement in blind audio source separation. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2006; **34**(3): 85-93

39 Li J, Mohamed A, Zweig G *et al*. LSTM time and frequency recurrence for automatic speech recognition. Automatic Speech Recognition and Understanding (ASRU), 2015: 187-191

40 Chen S Y, Wu Y, Chen Z *et al*. Continuous Speech Separation with Conformer. Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2021; 5749-5753

41 Zhang S Y, Li X F. Microphone Array Generalization for Multichannel Narrowband Deep Speech Enhancement. Proc. Interspeech, 2021: 666-670