计算机科学
COMPUTER SCIENCE

面向电台通信的CLU-Net语音增强网络

姚瑶，杨吉斌，张雄伟，李毅豪，宋宫琨琨

引用本文

姚瑶，杨吉斌，张雄伟，李毅豪，宋宫琨琨. 面向电台通信的CLU-Net语音增强网络[J]. 计算机科学, 2024, 51(9): 338-345.

YAO Yao, YANG Jibin, ZHANG Xiongwei, LI Yihao, SONG Gongkunkun. CLU-Net Speech Enhancement Network for Radio Communication [J]. Computer Science, 2024, 51(9): 338-345.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230700200

面向电台通信的 CLU-Net 语音增强网络

姚　瑶　杨吉斌　张雄伟　李毅豪　宋宫琨琨
中国人民解放军陆军工程大学　南京 210007
(speech_11@163.com)

摘　要　为了消除电台系统中的环境噪声和信道噪声对语音通信质量的不利影响,提升电台语音通信的质量,提出了一种基于联合通道注意力与长短时记忆网络(Long Short Term Memory,LSTM)的深度可分离 U 形网络 CLU-Net(Channel Attention and LSTM-based U-Net)。该网络采用深度可分离卷积实现低复杂度的特征提取,联合利用注意力机制和 LSTM 同时关注语音通道特征和长时上下文联系,在参数量较少的情况下实现对干净语音特征的关注。在公开与实测数据集上进行多组对比实验,仿真结果表明,所提方法在 VoiceBank-DEMAND 数据集上的 PESQ 和 STOI 等指标得分优于同类语音增强模型。实测实验结果表明,所提 CLU-Net 增强框架能够有效抑制环境噪声与信道噪声,在低信噪比条件下的增强性能优于其他同类型的增强网络。

关键词:电台通信;语音增强;深度可分离卷积;注意力机制
中图分类号　TP391

CLU-Net Speech Enhancement Network for Radio Communication

YAO Yao,YANG Jibin,ZHANG Xiongwei,LI Yihao and SONG Gongkunkun
School of Command and Control Engineering,Army Engineering University of PLA,Nanjing 210007,China

Abstract　In order to overcome the adverse effects of environmental and channel noise on speech communication quality in radio systems and improve the speech quality of radio communication,this paper proposes a deep separable network called CLU-Net (channel attention and LSTM-based U-Net),which adopts the deep U-shape architecture and long short-term memory(LSTM). In the network,deep separable convolution is used to implement low-complexity feature coding. The combination of attention mechanisms and LSTM can pay attention to the relationship between different convolution channels and the context of clean speech simultaneously and obtain the clean speech characteristic with fewer parameters. Varieties of noisy speech datasets are tested,including public and self-built sets using noise collected in different environments and radio systems. The results of the simulation experiment on the VoiceBank-DEMAND dataset indicate that the proposed method outperforms similar speech enhancement models in terms of objective metrics such as PESQ and STOI. Field experimental results show that the enhancement scheme can effectively suppress different environmental and radio noise types. The performance under low signal-to-noise ratios is superior to that of the same kind of enhancement networks.

Keywords　Radio communication,Voice enhancement,Deep separable convolution,Attention mechanism

1　引言

电台通信相比固定电话、移动通信等其他通信方式,具有成本低廉、抗毁性好和对基础设施要求低等优点,在应急抢险、危险艰苦地域作业等场景中应用广泛。然而,这些典型应用场景中通常存在着较强的背景噪声,导致语音通信质量较差。同时,通信过程中引入的信道噪声会进一步恶化语音通信效果[1]。

语音增强可以实现从带噪语音信号中提取干净语音,其中单通道语音增强技术因具有更好的通用性与经济性被广泛应用。然而,传统单通道语音增强方法基于噪声平稳、语音和噪声不相关等先验假设,无法适应电台通信真实应用场景的复杂声学变化,增强效果无法满足高质量的应用需求。

近年来,基于深度学习的语音增强方法相比传统方法呈现出较好的增强性能。卷积神经网络(Convolutional Neural Networks,CNN)能够刻画语音序列输入的局部约束,高效地

到稿日期:2023-07-26　返修日期:2023-11-04
基金项目:国家自然科学基金(62071484);陆军工程大学基础前沿项目(KYZYJKQTZQ23001)
This work was supported by the National Natural Science Foundation(62071484) and Basic Frontier Project of Army Engineering University of PLA(KYZYJKQTZQ23001).
通信作者:杨吉斌(yjbice@sina.com)

逐层提取不同尺度上的语音特征表示[2]。此外，研究者还提出了采用跳跃连接的卷积编解码网络 U-Net，能够有效融合各个层次的语音特征[3-6]，减少信息损失。

随着语音增强网络深度的加深，无论采用一维卷积还是二维卷积，CNN 网络的参数量都随之提升，内存消耗逐渐增加。为了缓解这一问题，ResNeXt 系列网络[7]中采用了分组的思想，通过分组卷积的方式高效实现数据的局部相关性计算，从而大幅减少网络参数量。在此基础之上，Xception 网络提出了卷积通道相关性与空间相关性充分解耦的思想，并提出了由深度卷积与逐点卷积构成的深度可分离卷积[8]，可显著提高卷积参数的利用效率 。

然而，即使是高效的卷积网络，其仍然存在感受野有限的问题，难以建模语音长时相关性。而递归神经网络（Recurrent Neural Networks，RNN）具有长序列表示能力，有益于语音特征的分析，但其在训练时容易产生梯度消失或爆炸[9]。为了缓解这个问题并建模语音长时依赖关系，长短时记忆网络[10][11]通过元胞状态和门机制控制信息的流动。相比传统的 RNN，LSTM 有更好的记忆性能，利于处理长序列信息。

此外，为了更好地关注语音的重要性差异，Zhang 等[12]在语音时频分析域中引入注意力机制，通过对时频域中的重要区域加权来实现对语音信号的关注。然而，卷积网络编码后通道维度特征却常常被忽略。近年来，研究表明对卷积通道维度的关注能够有效提升特征质量，文献[13-14]引入了通道注意力机制，在参数量较少的条件下充分利用通道维度中存在的语音特征相关性，改善了语音表示的提取效果。

为了改善电台通信语音质量，消除环境和信道不同噪声对语音的影响，同时为了提高语音增强网络的参数利用效率，本文提出了基于联合通道注意力与 LSTM 的深度可分离 U 形网络 CLU-Net。本文研究的主要贡献如下：

1）优化设计了基于深度可分离卷积编解码模块，在解码模块中提出了深度可分离反卷积结构，减少了卷积编解码的参数量。

2）提出了一种 CLU-Net 语音增强网络，利用低参数量的深度可分离卷积和通道注意力，实现了语音深层特征通道间的关注，并采用 LSTM 充分建模语音长短时相关性，使得 CLU-Net 网络以较少的参数量准确建模干净语音的特征。

3）基于公开的语音数据集与自建的中文电台语音数据集进行了仿真与实测实验。结果表明，所提 CLU-Net 在 Voice-Bank-DEMAND 数据集上的各项指标得分均优于近年来的同类语音增强模型，在中文语音仿真及实测实验中均能够同时有效抑制环境噪声和信道噪声。

2　相关工作

蒸馏学习、剪枝、量化、低秩分解等典型方法可以减少网络的参数量[15]。针对卷积网络的轻量化，MobileNet[16]采用了深度卷积与逐点卷积构成的深度可分离卷积 DSConv（Depth-wise Separable Convolution），将每个卷积通道分别卷积输出并对由多个通道构成的完整特征图进行逐点卷积，从而提高卷积效率。ShuffleNet[17]采用了将组卷积与通道重组结合的思想，提高了卷积分组间的关联性，在降低模型参数量的同时能达到较好的模型精度。Zeng 等[18]提出了一种使用多尺度扩展的深度可分离卷积轻量网络，能够从音频记录中进行家庭活动分类，参数量较少，适合部署在便携式终端中。

卷积网络与循环神经网络结合能够更有效地建模语音长时相关性与上下文联系。Tan 等[19]提出了一种由卷积编码器和 LSTM 构成的卷积循环网络，采用两层 LSTM 层对潜在特征序列进行建模。Le 等[20]提出了双路径卷积循环网络（Dual-Path Convolution Recurrent Network，DPCRN），采用块内 RNN 与块间 RNN 分别建模语音帧内的频谱模式与帧间的相关性。DEMUCS[21]基于 U-Net 卷积网络，采用双向 LSTM 层（Bi-LSTM）同时获得过去和未来的信息，但是其网络参数量较大，且忽略了各通道间的特征相关性。

近期的研究表明，引入通道注意力机制有助于语音增强性能的改善。Hu 等[22]提出了压缩-激励网络 SENet（Squeeze-and-Excitation Networks），通过全连接层学习通道间的联系；FU 等[23]提出了结合通道注意力与位置注意力的 DANet（Dual attention Net）网络，通过协方差矩阵分析特征两两之间的关联性，提高了模型的精度。Park 等[24]提出了多视图注意力网络 MANNER，引入通道维度的关注，提高了语音增强的性能。然而，MANNER 模型在每层一维卷积编解码中加入复杂注意力模块增加了网络负担，产生大量参数。Li 等[25]提出了具有通道注意力的深度可分离卷积网络的说话人识别模型，提高了识别精度。这些工作都说明，在卷积网络中引入通道注意力有助于在减少参数的同时缓解少样本学习中的过拟合问题。

3　基于 CLU-Net 的语音增强网络

面向电台语音通信的应用需求，在 U 形编解码网络的基础上，联合通道注意力与 Bi-LSTM 层，提出了 CLU-Net 语音增强网络，并基于深度可分离卷积实现网络的轻量化，以适应计算能力、存储能力受限的便携式应用场景。

3.1　网络框架

CLU-Net 网络的整体框架如图 1 所示，由深度可分离卷积编解码器、通道注意力层和 Bi-LSTM 层构成。其中，通过在编码器的最顶层利用通道注意力与双向 LSTM 的级联进行特征映射。

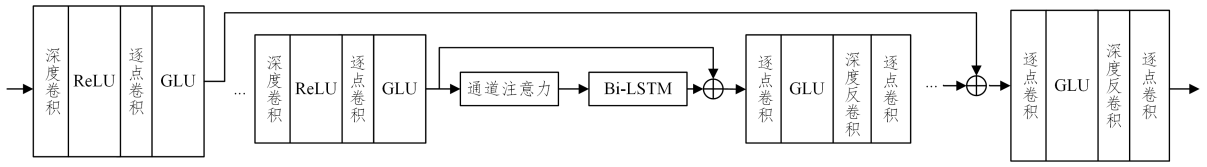


图 1　CLU-Net 网络结构图

Fig. 1　CLU-Net network architecture diagram

3.1.1 基于深度可分离卷积的编解码器

深度可分离卷积假设卷积神经网络中特征图的空间维和通道维具有可解耦性。在卷积操作中将通道和空间维度进行分离，可以减少模型参数和计算量，提高模型的效率和速度，过程如图 2 所示。

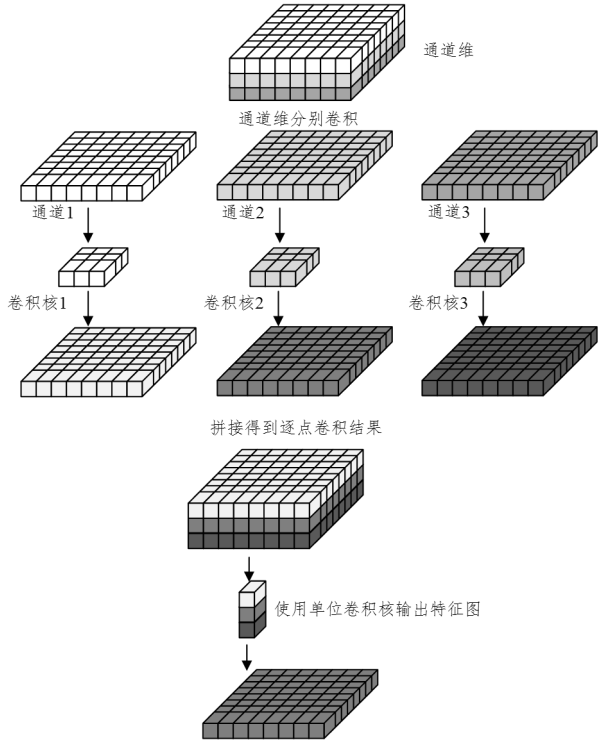


图 2 深度可分离卷积示意图

Fig. 2 Illustration of depth-wise separable convolution

深度可分离卷积将标准卷积操作分解为两个步骤：深度卷积和逐点卷积。深度卷积在每个输入通道上进行卷积，而逐点卷积在每个通道的输出上进行卷积。深度卷积只需要执行一次，而逐点卷积也仅需要执行少量的卷积运算，因此深度可分离卷积的计算量和参数量较少。

本文在一维卷积的基础上改进了深度可分离卷积，设计了由深度卷积、ReLU 激活层、逐点卷积和 GLU 激活层构成的深度可分离编码器。各层的连接结构如图 3 所示。

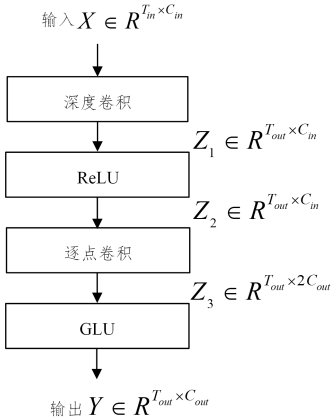


图 3 深度可分离卷积编码器

Fig. 3 Depth-wise separable convolution encoder

设输入特征为 $X \in R^{T_{in} \times C_{in}}$，输出特征为 $Y \in R^{T_{out} \times C_{out}}$，其中 C_{in} 和 C_{out} 分别是输入的通道数和输出的通道数。当采用卷积核大小为 k 的普通卷积时，需要 $k \times C_{in} \times C_{out}$ 个参数。而深度可分离卷积只需要空间维映射 $k \times C_{in}$ 个参数和通道维映射 $C_{in} \times C_{out}$ 个参数。可以明显看出 $k \times C_{in} + C_{in} \times C_{out} \leqslant k \times C_{in} \times C_{out}$，当编码器层数越深，$C_{in}$ 与 C_{out} 愈大，减少的参数量愈多，降低复杂度的效果越明显。

与编码器中使用卷积块不同，为了实现高效的卷积解码，本文设计了深度可分离的反卷积块。卷积块包含逐点卷积层、GLU 激活层和深度反卷积层，深度反卷积的示意图如图 4 所示。

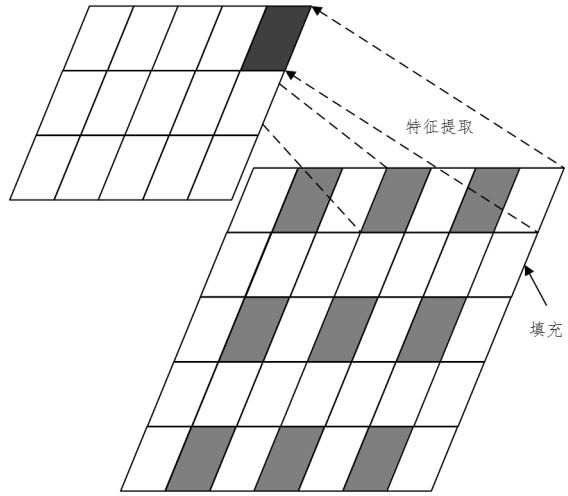


图 4 深度反卷积示意图

Fig. 4 Depthwise deconvolution diagram

对于输入特征 $X \in R^{T_{in} \times C_{in}}$，首先沿时间帧方向进行逐点卷积扩张通道维度，其次使用 GLU 对通道维度进行压缩，以丰富特征表示。接着，深度反卷积首先需要根据步幅 s 对解码器输入特征进行填充，以高效恢复压缩后的特征。若卷积填充数为 p，卷积核大小为 k。当 $s=1$ 时，对输入特征边缘填充 $k-p-1$ 个零；当 $s \neq 1$ 时，针对相邻时间维度特征，填充 $s-1$ 个零，对输入特征边缘填充 $k-p-1$ 个零。对填充后的特征进行深度卷积操作以减少其网络参数，再进行逐点卷积，对通道维度进行 GLU 操作，以保留更多的空间信息，最终得到尺度恢复后的特征图。

3.1.2 双向 LSTM

双向 LSTM 用于捕获音频深层特征的前后关系，提供更为精确的长时语音相关性的描述，结构如图 5 所示。前向传播层和反向传播层共同连接着输出层，网络权重在不同时刻共享。

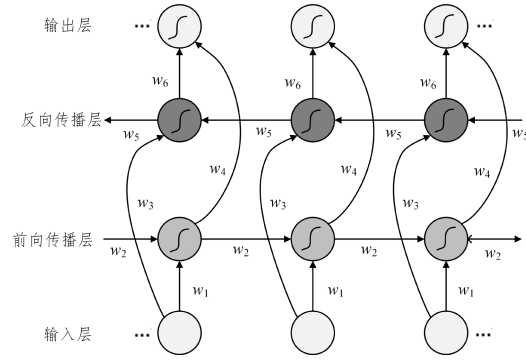


图 5 双向 LSTM 网络(Bi-LSTM)结构图

Fig. 5 Structure diagram of bidirectional LSTM

双向 LSTM 将潜在表示进行序列建模，并输出相同大小的非线性变换，如式(1)－式(3)所示：

$$h_t = f(w_1 x_t + w_2 h_{t-1}) \tag{1}$$

$$h_t' = f(w_3 x_t + w_5 h_{t+1}') \tag{2}$$

$$O_t = f(w_1 h_t + w_2 h_t') \tag{3}$$

其中，h_t 表示 t 时刻的前向隐含层输出特征，h_t' 表示反向隐含层输出特征，O_t 表示输出层输出特征，x_t 表示 t 时刻的输入特征。首先，前向传播层中通过从 1 时刻到 t 时刻正向计算，得到并保存每个时刻向前隐含层的输出。其次，在反向传播层，沿着时刻 t 到时刻 1 反向计算，得到并保存每个时刻向后隐含层的输出。最后，在每个时刻结合前向传播层和反向传播层的相应时刻输出的结果，得到最终的输出特征。

3.1.3　通道注意力

采用文献[23]中使用的通道注意力来获得通道特征表示。为了聚合信号信息，将平均和最大池化应用于输入特征 $X \in R^{C \times T}$，C 为通道数，T 为时间帧长度，分别得到两类不同的池化特征。

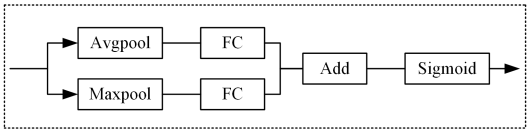


图 6　通道注意力

Fig. 6　Channel attention

接着，每个池化输出都通过全连接层，将得到的输出特征相加，再经过 sigmoid 激活层得到通道注意力权重 $\alpha_c \in R^{C \times 1}$，计算式如式(4)所示：

$$\alpha_c = \sigma(FC(AvgPool(X)) + FC(MaxPool(X))) \tag{4}$$

其中，FC 表示全连接层，σ 表示 sigmoid 激活。

最后使用注意力权重对输入特征进行加权，得到变换后的输出特征图。

3.2　损失函数

在网络中，使用时域损失函数 $loss\_t$ 和频域损失函数 $loss\_f$ 训练模型。

时域损失函数 $loss\_t$ 定义为干净语音信号与增强语音信号之间的均方误差(Mean Squared Error，MSE)，具体如式(5)所示：

$$loss\_t = \frac{1}{T}\sum_{t=0}^{T-1}(s_t - \hat{s}_t)^2 \tag{5}$$

其中，s_t 和 $\hat{s}_t$ 分别表示干净语音信号和增强语音信号，T 表示时间帧长度。

时频域损失函数 $loss\_f$ 定义为干净语音信号与增强语音信号幅度谱分量之间的平均绝对误差(Mean Absolute Error，MAE)，具体如式(6)所示：

$$loss\_f = \frac{1}{T \cdot F}\sum_{t=0}^{T-1}\sum_{f=0}^{F-1}[(|S_r(t,f)| + |S_i(t,f)|) - (|\hat{S}_r(t,f)| + |\hat{S}_i(t,f)|)] \tag{6}$$

其中，$S(t,f)$ 和 $\hat{S}(t,f)$ 分别表示干净语音信号和增强语音信号的幅度谱，r 和 i 分别表示复数变量的实部和虚部，T 和 F 分别表示时间帧长度和频率段数。实验中，结合了上述两种类型的损失函数，具体定义如式(7)所示：

$$loss\_sum = (1-\alpha)loss\_t + \alpha \cdot loss\_f \tag{7}$$

其中，超参数 α 在实验中设置为 0.2。

4　实验设置

4.1　数据集

仿真实验使用了 VoiceBank-DEMAND 英文语音数据集[26]与 THCHS30[27]中文语音数据集。VoiceBank-DEMAND 数据集中包含有带噪语音，THCHS30 数据集中全部为干净语音。为了便于与基线模型进行对比，采用 VoiceBank-DEMAND 数据集进行增强模型训练与仿真。在 THCHS30 语音数据集上构造电台带噪语音数据集，电台噪声数据来源于实际采集与 freesound 网站[1]以－7.5 dB，－2.5 dB，2.5 dB，7.5 B 与干净语音合成带噪语音，用于模型训练与实测实验。

4.2 网络配置

网络在实现过程中的具体配置如表 1 所列。

表 1　CLU-Net 网络的具体配置

Table 1　Specific configurations of CLU-Net

网络结构	输入	网络配置	输出
	通道数×样点个数	卷积核大小，步长	通道数×样点个数
上采样	1×64000	—	1×256084
编码层 1	1×256084	8,4,48	48×64020
编码层 2	48×64020	8,4,96	96×16004
编码层 3	96×16004	8,4,192	192×4000
编码层 4	192×4000	8,4,384	384×999
通道注意力＋LSTM	384×999	—	384×999
解码层 1	384×999	8,4,192	192×4000
解码层 2	192×4000	8,4,96	96×16004
解码层 3	96×16004	8,4,48	48×64020
解码层 4	48×64020	8,4,1	1×256084
下采样	1×256084	—	1×64000

编解码器各 4 层，中间层使用了一个通道注意力块和一个每层 384 个神经元的双向 LSTM 网络，双向 LSTM 最后使用线性层合并双向的输出。

4.3　网络训练

CLU-Net 的训练轮数设置为 500，训练批次大小为 16；使用了 Adam 优化器，步长为 3×10^{-4}，动量 $\beta_1 = 0.9$，分母动量 $\beta_2 = 0.999$。实验在 Ubuntu20.04 系统平台上进行，该平台带有一块 Xeon Gold 5118(2.3 GHz)的 CPU 与一块 GeForceRTX 2080Ti 的 GPU。

4.4　评价指标

本文采用 5 项客观指标来评价模型性能。语音质量感知评估(PESQ)[28]用于评估语音总体感知质量，评分范围为－0.5～4.5。短时客观可理解性(STOI)[29]用于评估语音可懂度，评分范围为 0～1。3 种基于平均意见得分(MOS)的测量方法[30]分别是测量语音信号失真的平均意见得分 CSIG、测量背景噪声干扰的平均意见得分 CBAK、评估语音整体质量的平均意见得分 COVL，这 3 种平均意见得分(MOS)的评分范围都为 1～5。5 项客观指标的评分值都与语音

[1] https://freesound.org/browse/

综合质量成正相关。同时利用模型参数占用内存来衡量模型的大小。

5 实验及结果分析

5.1 消融实验

在 VoiceBank-DEMAND 公开数据集上，以 Wave-U-Net 模型为基础，实现了包含 4 层一维卷积层的基线模型，对 CLU-Net 中的深度可分离卷积、双向 LSTM 以及注意力模块分别进行消融实验，结果如表 2 所列。其中，标识“√”表示增强模型中采用了相应模块。不同模块组合对应的模型分别命名为 CLU-Net1～CLU-Net4。

表 2 CLU-Net 模块的消融实验

Table 2 Ablation experiments of CLU-Net module

模型	深度可分离卷积	双向 LSTM	通道注意力	PESQ	参数/MB
基线				2.50	23.2
CLU-Net1	√			2.53	15.9
CLU-Net2	√	√		3.08	18.3
CLU-Net3		√	√	3.13	26.6
CLU-Net4	√		√	2.71	17.3
CLU-Net	√	√	√	3.15	19.6

从表 2 中可见，在基线模型中分别加入深度可分离卷积、双向 LSTM、通道注意力机制后，CLU-Net1，CLU-Net2，CLU-Net3，CLU-Net4 的 PESQ 相比基线模型，分别能够提升 0.03，0.58，0.63，0.21。从实验结果可以得到，CLU-Net1 引入了深度可分离卷积，相比普通卷积而言，网络参数减少了，语音增强性能保持相当的水平。CLU-Net2 和 CLU-Net4 在 CLU-Net1 的基础上分别加入了双向 LSTM 和通道注意力模块，对比性能可以知道引入双向 LSTM 改善语音增强效果更为显著。而 CLU-Net3 同时采用双向 LSTM 和通道注意力，取得了比 CLU-Net2 和 CLU-Net4 更好的效果，这是因为双向 LSTM 和通道注意力对语音深层提取所起的作用有区别，但是参数量却较大。CLU-Net 同时采用了 3 种处理模块，相比 CLU-Net3 参数量减少了约 7MB，而 PESQ 达到最优的 3.15，相比基线模型提升了 0.65，语音增强性能得到显著提高。

5.2 对比实验

仿真实验在 VoiceBank-DEMAND 数据集与构建的电台带噪数据集上进行，在公开数据集 VoiceBank-DEMAND 上的对比如表 3 所列。

表 3 公开数据集 VoiceBank-DEMAND 上的模型性能对比

Table 3 Model performance comparison on VoiceBank-DEMAND

模型	处理域	PESQ	STOI	CSIG	CBAK	COVL	参数/MB
SEGAN[3]	T	2.16	0.93	3.48	2.94	2.80	43.2
Wave-U-Net[31]	T	2.40	—	3.52	3.24	2.96	38.1
MetricGAN[32]	F	2.86	—	3.99	3.18	3.42	—
PHASEN[33]	F	2.99	—	4.21	3.55	3.62	—
DeepMMSE[34]	F	2.95	0.94	4.28	3.46	3.64	—
DEMUCS[21]	T	3.07	0.95	4.31	3.40	3.63	130.5
TSTNN[35]	T	2.96	0.95	4.33	3.53	3.67	3.51
CleanUNet[36]	T	2.90	0.95	4.33	3.42	3.64	46.07
CLU-Net	T	3.15	0.95	4.37	3.58	3.69	19.6

注：“—”表示原文未提供。

针对 VoiceBank-DEMAND 数据集中的各类背景噪声，CLU-Net 能够在保持较低参数量的条件下，PESQ，STOI，CSIG，CBAK，COVL 分别达到 3.15，0.95，4.37，3.58，3.69，其中 CLU-Net 的 PESQ 值相比表中的对比模型 DEMUCS，TSTNN，CleanUNet 分别提高了 0.08，0.19，0.25；参数占用内存相比采用普通卷积网络的 DEMUCS 与 CleanUNet 分别减少了 110.9MB 与 26.47MB。虽然 TSTNN 采用了膨胀卷积与亚像素卷积，参数量较少，但由于其感受野有限，对上下文联系与长时相关性建模不足，因此，联合通道注意力与长短时记忆网络的深度可分离 U 形网络 CLU-Net 具有一定优势，具备较好的增强性能。

此外，图 7 给出了以 Wave-U-Net 为基础的基线模型与 CLU-Net 增强后的结果。

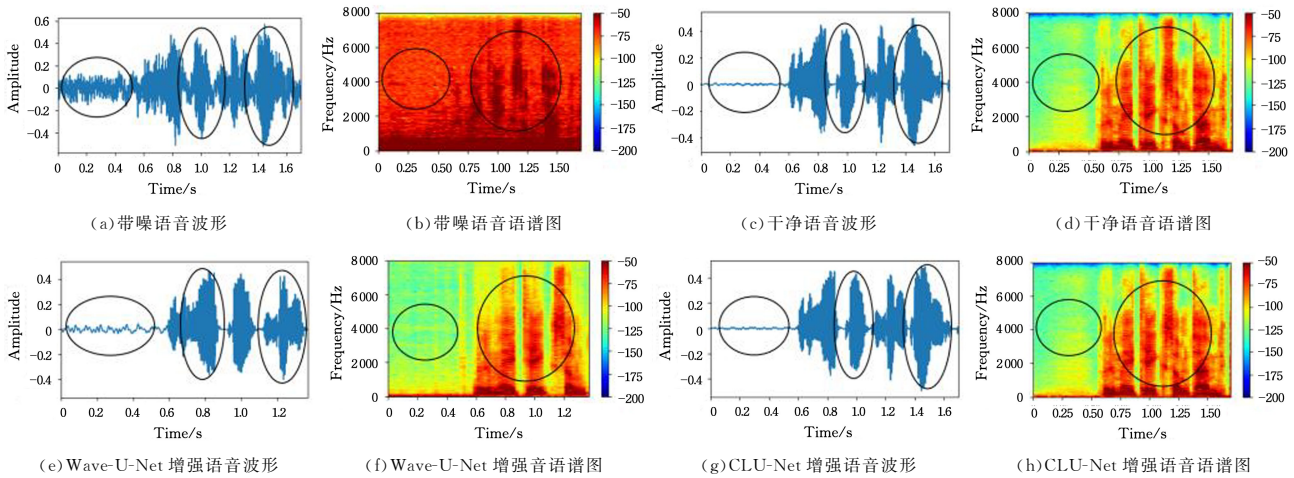


图 7 模型增强效果对比

Fig. 7 Comparison of model enhancement results

其中图 7(a)与图 7(b)分别为带噪语音波形与语谱图，图 7(c)与图 7(d)分别为干净语音波形与语谱图，图 7(e)与图 7(f)分别为 Wave-U-Net 增强语音波形与语谱图，图 7(g)与图 7(h)分别为 CLU-Net 增强语音波形与语谱图。

从图 7(e)和图 7(g)所示的椭圆中可以看出，CLU-Net 相比 Wave-U-Net 网络增强的语音波形，在无声段处的残留

噪声较小，在有声段的增强波形更接近干净语音波形。从图 7(f)和图 7(h)所示的语音图可以看出，Wave-U-Net 增强语音在无声段出现了干净语音频谱中没有的谐波结构，在有声段的高频频谱始终保持较大能量，与真实语音高频能量分布并不一致。而 CLU-Net 的频谱分布与真实语音的频谱分布更为相近，失真较小。

在构建的中文电台带噪语音测试集上，增强语音的 PESQ，STOI，CSIG，CBAK 以及 COVL 可以分别达到 2.9，0.93，4.25，3.47，3.53，对电台带噪数据的增强结果如表 4 所列。

表 4　中文电台带噪数据集上模型的性能对比

Table 4　Model performance comparison on Chinese radio speech

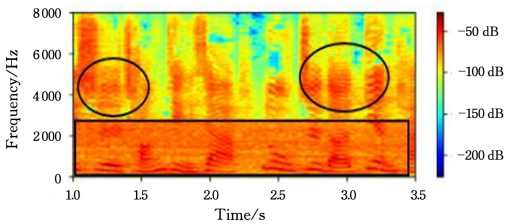
模型	PESQ	STOI	CSIG	CBAK	COVL	参数/MB	训练时间/s
Wave-U-Net	2.29	0.92	3.37	3.16	2.85	38.1	475
DEMUCS	2.81	0.93	4.17	3.31	3.42	130.5	513
CLU-Net	2.90	0.93	4.25	3.47	3.53	19.6	377

如表 4 所列，在中文电台带噪语音数据集上，将 CLU-Net 与同样以 U-Net 为基础的 Wave-U-Net 和 DEMUCS 模型进行对比，可以明显看出，CLU-Net 相比 DEMUCS 的 PESQ 能够提高 0.09，CSIG，CBAK，COVL 也分别提高了 0.08，0.16，0.11，性能得到显著提高。从结果可以看出，CLU-Net 对于信道噪声的抑制效果能达到较好的水平，在电台通信场景下具备一定的应用潜力。在计算时间与存储空间占用方面进行进一步对比，CLU-Net 相比 Wave-U-Net 减少了 18.5MB 的参数占用空间，同时 CLU-Net 相比 Wave-U-Net 训练模型一轮的时间缩短了 98s，模型的参数存储空间与计算时间都得到了优化。

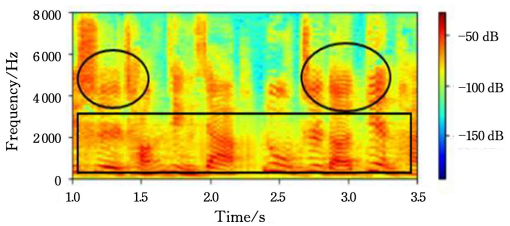
5.3　实测实验

本文分别在消声室(室内)和树林中(室外)采集了不同的电台带噪语音，这些场景中的环境噪声分别代表了室内无混响理想环境的噪声和室外噪声。利用 CLU-Net 增强框架对这些带噪语音进行了增强，不同场景以及对应的带噪语音、增强语音频谱如图 8、图 9 所示。

(a)实测场景 1：消音室(室内)



(b)消音室(室内)带噪语音语谱图



(c)增强语音语谱图

图 8　室内增强效果对比

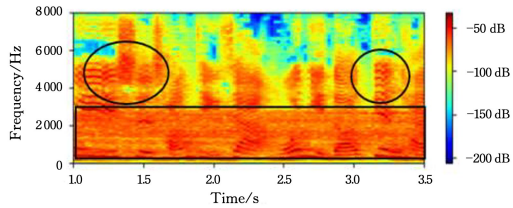
Fig. 8　Comparison of enhancement effect in indoor scenes

1)消音室(室内)场景

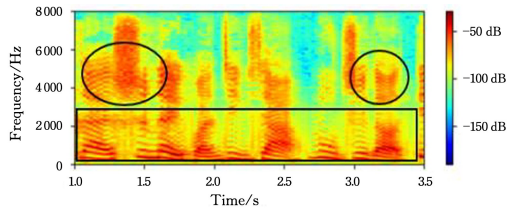
实测音频是在图 8(a)所示的消音室环境中采集的，背景噪声主要为信道噪声。从图 8(b)的矩形框中可以看出，信道噪声集中分布于 0～2000 Hz 之间。从图 8(c)增强的语谱图可以看出，语音增强能够明显去除电台背景噪声。从椭圆框位置中可以看出，增强后的语谱图中的谐振峰结构更加清晰。

2)树林(室外)场景

室外树林环境中，除了信道噪声依然明显外，同时还存在低频风声、高频树叶沙沙声以及其他说话人的微弱语音等背景噪声。与图 9(b)中的带噪语音相比，图 9(a)中带噪语音的谱结构更加不清晰。图 9(b)为增强后的语谱图，从椭圆框中可以看出，室外场景带噪语音中的背景噪声、低频段的较强信道噪声均得到一定抑制，增强语音中呈现清晰谐振峰结构。



(a)树林(室外)带噪语音语谱图



(b)增强语音语谱图

图 9　室外增强效果对比

Fig. 9　Comparison of enhancement effect in outdoor scenes

结束语　为了提高电台语音通信的质量，本文提出了一种基于联合通道注意力与 LSTM 的深度可分离 U 形网络 CLU-Net，采用深度可分离卷积、通道注意力与长短时记忆网络，实现了低参数量条件下的高效语音建模。在不同数据集上的测试结果均表明，所提的 CLU-Net 增强框架能够有效抑制不同类别环境噪声与信道噪声，在低信噪比条件下的增强性能优于其他同类型的增强网络。今后将进一步研究语音增强的轻量化处理，实现增强模型在实际电台设备中的移植，并开展相应的测试和模型优化。

参 考 文 献

[1] WANG Y P, WEI G H, PAN X D, et al. Prediction model and experiment of out-of-band dual-band interference of communication station[J]. Acta Electronica Sinica, 2019, 47(4): 826-831.

[2] LI S, CAO F. Research on end-to-end framework model analysis and trend of intelligent speech technology [J]. Computer Science, 2022, 49(S1): 331-336.

[3] PASCUAL S, BONAFONTE A, SERRA J. SEGAN: Speech Enhancement Generative Adversarial Network[C]// Conference of the International Speech Communication Association. 2017: 3642-3646.

[4] PANDEY A, WANG D. TCNN: Temporal Convolutional Neural Network for Real-time Speech Enhancement in the Time Domain[C]// 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019). Brighton, UK, 2019: 6875-6879.

[5] PANDEY A, WANG D L. Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain[C]// 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020). IEEE, 2020: 6629-6633.

[6] FAN J Y, YANG J B, ZHANG X W, et al. Single-channel speech enhancement based on multi-head attention mechanism in U-net network[J]. Acta Acoustica Sinica, 2022, 47(6): 703-716.

[7] LI L, ZHU Y, ZHU Z. Automatic Modulation Classification Using ResNeXt-GRU With Deep Feature Fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-10.

[8] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1251-1258.

[9] BENGIO Y, SIMARD P, FRASCONI P, Learning long-term dependencies with gradient descent is difficult[J]. IEEE Transactions on Instrumentation and Measurement, 1994, 5(2): 157-166.

[10] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[11] BANG J Y, SUN M, ZHANG X W, et al. Lightweight Model for Bone-Conducted Speech Enhancement Based on Convolution Network and Residual Long Short-Time Memory Network[J]. Journal of Data Acquisition & Processing, 2021, 36(5): 921-931.

[12] ZHANG Q, SONG Q, NI Z, et al. Time-frequency attention for monaural speech enhancement [C]// 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022). IEEE, 2022: 7852-7856.

[13] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3-19.

[14] TOLOOSHAMS B, GIRI R, SONG A H, et al. Channel-attention dense u-net for multichannel speech enhancement [C]// 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020). Barcelona, Spain. IEEE, 2020: 836-840.

[15] ZHU X, LI J, LIU Y, et al. A Survey on Model Compression for Large Language Models[J]. arXiv: 2308.07633, 2023.

[16] ANDREW G H, MENGLONG Z, BO C, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv: 1704.04861, 2017.

[17] ZHANG X, ZHOU X, LIN M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6848-6856.

[18] ZENG Y, LI Y, ZHOU Z, et al. Domestic activities classification from audio recordings using multi-scale dilated depthwise separable convolutional network[C]// 2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2021: 1-5.

[19] TAN K, WANG D L. A convolutional recurrent neural network for real-time speech enhancement[C]// Interspeech 2018. 2018: 3229-3233.

[20] LE X, CHEN H, CHEN K, et al. DPCRN: Dual-path convolution recurrent network for single channel speech enhancement[C]// Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association. Brno, Czechia, 2021: 2811-2815.

[21] DEFOSSEZ A, SYNNAEVE G, ADI Y. Real time speech enhancement in the waveform domain[C]// Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event. 2020: 3291-3295.

[22] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.

[23] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3146-3154.

[24] PARK H J, KANG B H, SHIN W, et al. Manner: Multi-view attention network for noise erasure[C]// 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022). Singapore, IEEE, 2022: 7842-7846.

[25] LI Y, WANG W, CHEN H, et al. Few-shot speaker identification using depthwise separable convolutional network with channel attention[J]. arXiv: 2204.11180, 2022.

[26] VALENTINI-BOTINHAO C, WANG X, TAKAKI S, et al. Investigating RNN-based speech enhancement methods for noise-robust text-to-speech[C]// SSW. 2016: 146-152.

[27] WANG D, ZHANG X. Thchs-30: A free chinese speech corpus [J]. arXiv: 1512.01882, 2015.

[28] RIX A W, BEERENDS J G, HOLLIER M P, et al. Perceptual evaluation of speech quality (PESQ) — a new method for speecn quality assessment of telephone networks and codecs[C]// Proceedings of the 26th International Conference on Acoustics, Speech, and Signal Processing. Utah: IEEE, 2001: 749-752.

[29] TAAL C H，HENDRIKS R C，HEUSDENS R，et al. An algorithm for intelligibility prediction of time-frequency weighted noisy speech[J]. IEEE Transactions on Audio，Speech，and Language Processing，2011，19(7)：2125-2136.

[30] HU Y，LOIZOU P C. Evaluation of objective quality measures for speech enhancement[J]. IEEE Transactions on Audio，Speech，and Language Processing，2007，16(1)：229-238.

[31] MACARTNEY C，WEYDE T. Improved speech enhancement with the Wave-U-Net[J]. arXiv：1811. 11307，2018.

[32] FU S W，LIAO C F，TSAO Y，et al. Metricgan：Generative adversarial networks based black-box metric scores optimization for speech enhancement[C]// International Conference on Machine Learning. PMLR，2019：2031-2041.

[33] YIN D，LUO C，XIONG Z，et al. Phasen：A phase-and-harmonics-aware speech enhancement network[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020，34(5)：9458-9465.

[34] ZHANG Q Q，AARON M N，WANG M J，et al. Deepmmse：A deep learning approach to mmse-based noise power spectral density estimation[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，IEEE，2020，28(1)：1404-1415.

[35] WANG K，HE B，ZHU W P. TSTNN：Two-Stage Transformer Based Neural Network for Speech Enhancement in the Time Domain[C]// 2021 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP 2021). Toronto，ON，Canada，2021：7098-7102.

[36] KONG Z，PING W，DANTREY A，et al. Speech denoising in the waveform domain with self-attention[C]// 2022 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP 2022). IEEE，2022：7867-7871.

YAO Yao，born in 1998，postgraduate. Her main research interests is intelligent speech processing.

YANG Jibing，born in 1978，Ph.D，associate professor. His main research interests include speech and acoustic signal processing.

（责任编辑：喻藜）