DOI: 10.13382/j.jemi.B2306619

联合频谱映射与掩蔽估计的协作式语音增强方法*

罗庆予 张天骐 方 蓉 张慧芝

(重庆邮电大学通信与信息工程学院 重庆 400065)

摘 要:为提高目前基于掩蔽与基于频谱映射的语音增强方法性能上界以及复杂环境下的泛化能力,提出了一种在联合复频谱与复掩蔽学习框架下的协作式单通道语音增强方法。该方法采用编码器-双分支解码器结构,在编解码部分设计了一种交互协作学习单元(ICU)来监督交互语音信息流,并提供有效的潜在特征空间;中间层则是设计出一种多尺度融合 Transformer,以少量参数在空间-通道维度上多尺度地提取细节信息后融合输出,同时对语音子频带与全频带信息建模。在大、小数据集与 115 种噪声环境下进行实验,结果表明该方法仅以 0.57 M 的参数量,取得比大部分先进且相关方法更优的主、客观指标,具有良好的鲁棒性与有效性。

关键词: 语音增强;复频谱映射;复掩蔽;多尺度融合 Transformer;轻量型网络

中图分类号: TN912.35 **文献标识码**: A **国家标准学科分类代码**: 510.4040

Collaborative speech enhancement method combining spectral mapping and masking estimation

Luo Qingyu Zhang Tianqi Fang Rong Zang Huizhi

(School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: In order to improve the performance upper bound and generalization ability of current speech enhancement methods based on masking and spectrum mapping, a collaborative monaural speech enhancement method based on the learning framework of combined complex spectrum and masking is proposed. An interactive cooperative learning unit (ICU) is designed in the codec part to monitor the interactive speech information flow and provide an effective potential feature space. In the middle layer, a multi-scale fusion Transformer is designed to extract multi-scale details in the spatial-channel dimension with a small number of parameters for fusion output, at the meanwhile, modeling the voice sub-band and full band information. Experiments on large and small data sets and 115 noise environments show that the proposed method only uses 0.57 M parameters to obtain better subjective and objective indicators than most advanced and related methods, which has good robustness and effectiveness.

Keywords: speech enhancement; complex spectral mapping; complex masking; multi scale fusion Transformer; lightweight network

0 引 言

单通道语音增强旨在解决噪声抑制问题的同时提升语音的感知质量和可懂度。在当今社会上的常见的语音识别[1]、视频会议等的前端处理任务均需要用到语音增强技术。

目前在深度学习领域中,单通道语音增强主要分为基于时域估计[2]与基于时频域估计[3]的两种方法。早期研究主要是通过利用含噪语音的相位谱与估计语音的幅度谱相结合来重建增强语音,但含噪语音的相位谱会影响语音质量,并且相位会因在网络训练过程中产生偏移

收稿日期: 2023-06-09 Received Date: 2023-06-09

*基金项目:国家自然科学基金(61671095,61771085)、重庆市自然科学基金(cstc2021jcyj-msxmX0836)、重庆市教育委员会科研项目(KJ1600427, KJ1600429)资助

而难以估计。虽然时域语音估计方法不用考虑相位问题,但仍达不到在时频域上语音和噪声分离的精细程度,因此在复频域上处理语音成为一种主流趋势。为了联合估计幅度谱与相位谱,继而提出了基于复掩蔽、基于复频谱映射[4-6]的估计方法,通过估计语音实部、虚部分量来学习语音的相位信息。研究学者们在模型尺寸小且计算快的卷积神经网络(convolution neural network,CNN)基础上,将其搬移至复频域进行语音去噪,提出了一种复卷积递归网络(complex convolution recurrent network,CCRN)[4],通过利用复频谱映射的方式来估计语音实、虚部分量,取得了一定效果。文献[7]提出了一种连续取值的理想比率掩蔽(ideal ratio masking,IRM),能有效抑制语音中的噪声。Williamson 等[8]提出采用复理想比率掩蔽(complex ideal ratio masking,CIRM)的方式来有效估计语音幅度谱与相位信息。目前基于复掩蔽估计与复频谱映射的方法都有着各自的优势,可由于不同的训练目标对增强结果有很大影响,且上述研究均仅限于单个训练目标,并未考虑到对不同训练目标对模型性能边界的影响。

近年,Transformer 模型在解决长时性建模问题上表现优异[9],但由于计算成本高且参数量大而不适用于低资源应用。最近,TSTNN[10]、CAUNet[11]等研究模型采用双路径 Transformer 模块对输入语音序列的全局与局部特征进行建模,并取得了一定效果,但这些研究均只局限于简单的时域特征。同时,所提出的相关 Transformer 仅关注语音的空间信息,而忽略了特征在通道维度上的潜在价值,极大的限制了模型的语音增强性能。

针对上述问题,本文搭建了一种联合复掩蔽估计与复频谱映射的神经网络学习框架,并在此框架下提出一种融合多尺度高效 Transformer 的协作式单通道语音增强方法,该方法采用编码器-中间网络-双分支解码器结构。在模型编解码部分,通过设计一种交互协作单元(interactive collaboration unit,ICU)来提升主干网络提取特征的能力并通过双信息流实现交互学习。同时,解码器两分支分别估计输出针对语音实部与虚部分量的掩蔽与直接映射的复频谱,利用可学习参数将两目标的优势结合后输出以进一步提高模型的性能上界。中间层网络则由提出改进的多尺度融合 Transformer 分别对语音子频带和全频带信息建模,其内部设计了一种细节感知多头注意(detail aware multi-head self-attention,DA-MHSA)机制来将语音空间特征与通道维度上的细节特征相融合,为提高解码过程的鲁棒性,在此基础上提取多尺度信息后再传入解码器。最后,分别在由 VoiceBank-DEMAND 英文数据集、THCHS30 中文数据集与 115 种常见噪声下,验证本文网络的增强性能与泛化能力。

1　系统结构

1.1　语音增强模型

单通道语音增强模型的输入描述为:

$$\boldsymbol{y}=\boldsymbol{s}+\boldsymbol{n} \tag{1}$$

其中,$\boldsymbol{y}$、$\boldsymbol{s}$、$\boldsymbol{n}\in\boldsymbol{R}^{1\times L}$ 分别表示含噪语音、纯净语音和噪声的时域波形,L 为波形长度。通过对时域波形进行短时傅里叶变换(short time Fourier transform,STFT)[12]将其转换为时频域得到:

$$\boldsymbol{Y}(t,f)=\boldsymbol{S}(t,f)+\boldsymbol{N}(t,f) \tag{2}$$

其中,$\boldsymbol{Y}$、$\boldsymbol{S}$、$\boldsymbol{N}\in\boldsymbol{R}^{2\times T\times F}$ 分别代表了含噪语音、纯净语音、噪声在时频域上的复值,F 为频率维度,T 为时间帧数。本文提出的神经网络模型如图 1 所示,直接将含噪语音复频谱 $\boldsymbol{Y}$ 作为模型中编码器的输入,中间层网络作为传输过渡层,解码器则采用双分支结构。

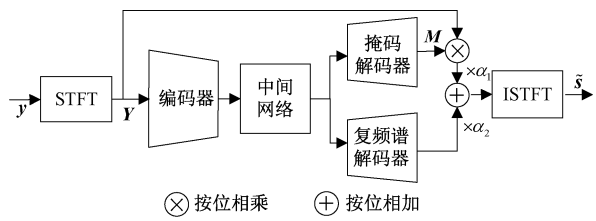

图 1　语音增强流程

Fig. 1　Flowchart of speech enhancement

解码器的上分支用于直接输出估计的增强语音实、虚部分量的掩蔽 $\boldsymbol{M}\in\boldsymbol{R}^{2\times T\times F}$,通过与模型输入的复频谱按位相乘以恢复增强语音复频谱 $\tilde{\boldsymbol{S}}_{\text{mask}}\in\boldsymbol{R}^{2\times T\times F}$,如式(3)所示。解码器下分支则是直接利用神经网络映射估计出增强语音的复频谱 $\tilde{\boldsymbol{S}}_{\text{spectral}}$。通过利用可学习参数 α_1 与 α_2 对两输出进行加权求和,得到本文模型最终输出的增强语音复频谱 $\tilde{\boldsymbol{S}}$,可以表示为:

$$\tilde{\boldsymbol{S}}_{\text{mask}}=\boldsymbol{M}\otimes\boldsymbol{Y} \tag{3}$$

$$\tilde{\boldsymbol{S}}=\alpha_1\tilde{\boldsymbol{S}}_{\text{mask}}\oplus\alpha_2\tilde{\boldsymbol{S}}_{\text{spectral}} \tag{4}$$

其中,⊗、⊕分别表示按位相乘,按位相加。最后,进行短时傅里叶逆变换(inverse short time Fourier transform,ISTFT)将输出恢复到时域空间,重构出增强语音时域波形 $\tilde{\boldsymbol{s}}$:

$$\tilde{\boldsymbol{s}}=\text{ISTFT}(\tilde{\boldsymbol{S}}(t,f)) \tag{5}$$

本文通过神经网络模型来将复掩蔽估计与复频谱映射作为联合训练目标,提高网络对相位的感知能力,避免了直接估计幅度谱与含噪语音相位相结合输出带来的影

响。同时,将两目标的优势结合到一个联合学习框架中,进一步改善语音增强模型的性能边界,网络模型中各部分的功能模块将在后续小节具体介绍。

1.2 交互协作单元

在实际多样化的语音环境下,本文根据混合专家神经网络 MoE[13] 的思想,将结构复杂的深度神经网络简化为具有不同感受野的卷积,提出一种交互协作单元(ICU)来提取并恢复网络编解码部分中的语音特征。ICU 中不仅包含了左右两"专家"的卷积分支,还利用门控机制进一步指导与控制特征信息的生成。

如图 2(a)所示为编码器中的交互协作单元(encoder interactive collaboration unit, En_ICU)。令输入特征为 $\boldsymbol{X} \in \boldsymbol{R}^{B\times C\times T\times F}$,其中 B 代表批量数、C 代表通道数。首先,通过卷积核大小 s 为 1×1 的二维卷积层,并将 C 减半以减轻训练参数负担。接下来,特征输入到两个并行卷积层,作为两个"专家"分支,其卷积核大小分别为 2×3、2×5,通过不同感受野以捕获各种不同语音信息。其中步长 s 设为 1×2,特征尺寸在频域维度上减半,从而获得更鲁棒的低维特征。每个分支采用门控的方式与另一"专家"分支进行信息交互:即通过 1×1 卷积与 Sigmoid 激活函数的输出与另一分支输出相乘,实现两信息流的相互补偿的目的。最后,将相加后的信息矩阵传递给卷积核大小为 1×1 的卷积层,以恢复特征图的通道数。输出特征图的维度为 $B\times C\times T\times F/2$,该输出层卷积依次经过批量归一化(batch normalization, BN)与参数修正线性单元(parametric rectified linear unit, PReLU)激活函数操作。

图 2(b)所示为解码器中的交互协作单元(decoder interactive collaboration unit, De_ICU),其中采用二维反卷积以逐步恢复频率维度上的大小。另外,在掩码解码器分支最后一个 De_ICU 的末端输出层采用 Tanh 激活函数来输出估计的掩蔽值。ICU 不仅能高效替代编解码器中的普通卷积层以及密集卷积层,还能提升主干网络对特征处理能力,促使该网络在交互协作学习模式下能加以区分并学习语音的潜在信息空间。

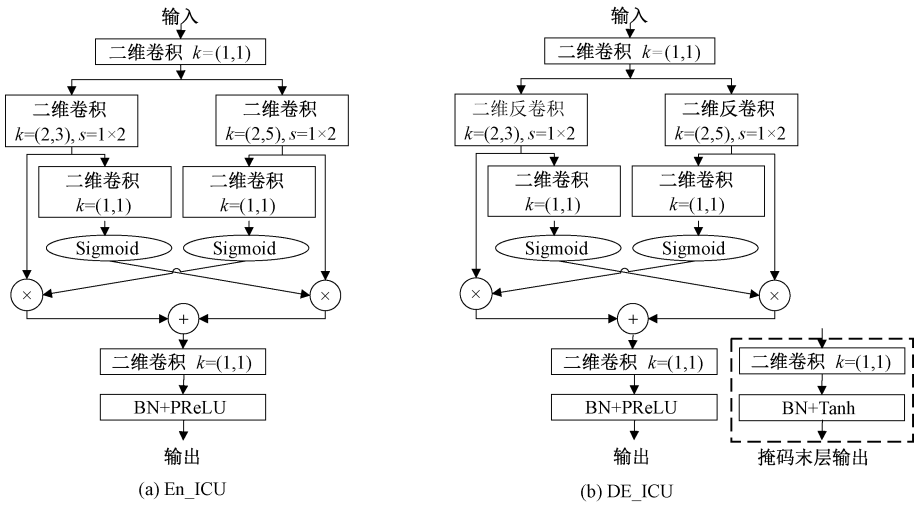

图 2 编解码器中的交互协作单元

Fig. 2 Interactive collaboration units in encoder and decoder

1.3 细节感知多头注意力层

Transformer 起初广泛应用于计算机视觉任务[14],因其中的多头注意力机制能对长序列依赖性高效建模而受到近年语音领域的关注。卷积神经网络的大部分特征信息都包含在信道中[15],而相关研究往往是利用多头注意力机制对语音的时频空间信息建模,这忽视了在通道维度上的潜在有用信息。针对此问题,本文提出一种细节感知多头注意力(DA-MHSA)机制,如图 3 所示。在原有多头注意力机制上引入通道注意分支,两者相互作用来提取语音复频谱中的有效细节特征,同时增强信息的多维度传递。

具体来说,首先将输入特征 $\boldsymbol{X}$ 调整为三维并通过空间多头注意分支,通过 K 个平行注意力层,其中第 i 层的线性变换参数矩阵为 $\boldsymbol{W}_i^{\mathrm{Q}}$、$\boldsymbol{W}_i^{\mathrm{K}}$、$\boldsymbol{W}_i^{\mathrm{V}}$,分别对输入 $\boldsymbol{X}$ 映射输出得到查询 $\boldsymbol{Q}_i$、键 $\boldsymbol{K}_i$、值 $\boldsymbol{V}_i$ 参数矩阵,该操作可表示为:

$$[\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i] = \boldsymbol{X}[\boldsymbol{W}_i^{\mathrm{Q}}, \boldsymbol{W}_i^{\mathrm{K}}, \boldsymbol{W}_i^{\mathrm{V}}] \tag{6}$$

$$\boldsymbol{A}_i^{\mathrm{spatial}} = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_i \boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d}}\right) \tag{7}$$

其中,$i \in [1,2,\cdots,K]$,d 为比例因子。将 $\boldsymbol{Q}_i$ 与 $\boldsymbol{K}_i$ 的矩阵乘积进行尺度归一化,并经过 Softmax 激活函数得到的空间注意力权重矩阵 $\boldsymbol{A}_i^{\mathrm{spatial}}$,如式(7)所示。在通道注意分支上,首先将 $\boldsymbol{Q}_i$ 的转置矩阵与 $\boldsymbol{K}_i$ 的乘积进行尺度归一化,然后分别利用最大池化层 Maxpool(·) 与平均池化层 Avgpool(·) 操作来对通道维度上的特征进行压缩

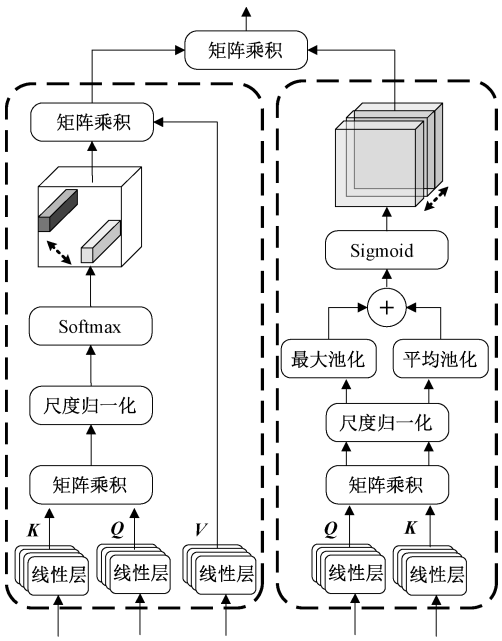

图 3　细节感知多头注意力层

Fig. 3　Detail aware multi-head self-attention layer

激活。最后,对输出求和并通过用 σ 表示的 Sigmoid 激活函数来得到通道注意权重矩阵 $\boldsymbol{A}_i^{\text{channel}}$,具体过程如式(8)所示:

$$\boldsymbol{A}_i^{\text{channel}} = \sigma\left(\text{Maxpool}\left(\frac{\boldsymbol{Q}_i^{\text{T}}\boldsymbol{K}_i}{\sqrt{d}}\right) + \text{Avgpool}\left(\frac{\boldsymbol{Q}_i^{\text{T}}\boldsymbol{K}_i}{\sqrt{d}}\right)\right) \tag{8}$$

在表征子空间中,所求出的空间与通道注意的权重矩阵能更高精度的区分纯净语音信号和噪声信号,将 $\boldsymbol{A}_i^{\text{spatial}}$ 与 $\boldsymbol{A}_i^{\text{channel}}$ 以及 $\boldsymbol{V}_i$ 进行矩阵乘积,得到了融合空间与通道语音细节信息的第 i 层输出 $\boldsymbol{H}_i$,最终把每一层输出进行级联 Concat(·) 得到细节感知多头注意力的输出 $\boldsymbol{H}$,如式(9)、(10)所示:

$$\boldsymbol{H}_i = \boldsymbol{A}_i^{\text{spatial}}\boldsymbol{V}_i\boldsymbol{A}_i^{\text{channel}} \tag{9}$$

$$\boldsymbol{H} = \text{Concat}(\boldsymbol{H}_1, \boldsymbol{H}_2, \cdots, \boldsymbol{H}_K) \tag{10}$$

细节感知多头注意力机制能利用空间注意特征和通道注意特征的相互作用,并采用不同权重的参数矩阵来增强该模块对细节信息的感知,从而提高网络的去噪能力,本文将用 DA-MHSA 代替普通 MHSA 集成到 Transformer 模块中。

1.4　多尺度融合 Transformer

Transformer 由编码器和解码器组成,但在语音处理任务中一般主要采用 Transformer 编码器[8],其由多头注意力层与前馈网络构成。本文提出的细节感知多头注意力层(DA-MHSA)虽能模拟长序列信息,但仅对序列中的元素进行单独通道变换而缺少对局部信息的关注,故本文采用膨胀卷积与普通卷积的组合来代替全连接层作为前馈网络,以提供对局部特征的关注。本文将细节感知多头注意力层与膨胀卷积前馈网络相结合,提出了一种多尺度融合 Transformer,如图 4 所示。在多尺度融合 Transformer 中,首先将 DA-MHSA 的三维输入特征 $\boldsymbol{X}$ 与输出 $\boldsymbol{H}$ 残差连接后再逐点相加,以防止梯度消失,再经过层归一化(layer normalization,LN)得到输出 $\boldsymbol{O}_{\text{DA-MHSA}}$,如式(11)所示:

$$\boldsymbol{O}_{\text{DA-MHSA}} = \text{LN}(\boldsymbol{X} + \boldsymbol{H}) \tag{11}$$

其中,LN(·) 表示层归一化操作。在输入膨胀卷积前馈网络前需要将三维特征还原调整(Reshape)为四维,该网络由一个 1×1 的二维卷积与 3 个卷积核大小为 3×3 且膨胀率 d 分别为 1、6、12 的膨胀卷积组成,利用不同感受野的膨胀卷积来充分提取网络中间层被忽略的局部特征,并且该网络前后同样采用残差连接与 LN 操作。

另外,受 TSTNN[10] 对分帧的时域语音信号的局部、全局特征建模的启发,本文针对时频域语音信号,利用多尺度融合 Transformer 分别对语音子频带的局部频谱与全频带的全局依赖性建模。具体来说,中间网络中共有 $2N$ 个多尺度融合 Transformer 模块,每两个模块为一组,且每一个多尺度融合 Transformer 模块输出前后均进行残差连接以防止网络梯度消失。令第 l 组的三维输入为 $\boldsymbol{X}_l \in \boldsymbol{R}^{C\times(B\times T)\times F}$ 且 $l \in [1,2,\cdots,N]$,利用第 1 个多尺度融合 Transformer 模块对输入中的每个子频带上的所有时间步长进行建模。接着,将该模块的输出调整(Reshape)为 $\boldsymbol{X'}_l \in \boldsymbol{R}^{C\times T\times(B\times F)}$ 后输入第 2 个多尺度融合 Transformer 模块,利用该模块对所有子频带进行整合并对全频带信息建模,具体处理步骤为:

$$\boldsymbol{X'}_l = f(\boldsymbol{X}_l[:,:,i], i = 1,2,\cdots,F) \tag{12}$$

$$\boldsymbol{X}_{l+1} = f(\boldsymbol{X'}_l[:,j,:], j = 1,2,\cdots,T) \tag{13}$$

其中,$f(\cdot)$ 表示多尺度融合 Transformer 的映射函数,且 $\boldsymbol{X}_l[:,:,i]$ 表示第 i 子频带在所有时间步长下的序列,$\boldsymbol{X'}_l[:,j,:]$ 表示第 j 时间步长下包含所有子频带信息的序列即全频带,$\boldsymbol{X}_{l+1}$ 为第 $l+1$ 组的输入。

多尺度融合 Transformer 中的 DA-MHSA 不仅能对语音信号的复频域上特征的子频带与全频带建模,且同时关注到了该特征通道维度上的信息,融合输出的特征进一步经过膨胀卷积前馈网络的多尺度地捕获局部信息进行补偿,增强网络感知语音细节信息的能力,使得模型更好地抑制背景噪声。

1.5　网络模型结构

网络整体结构由编码器、中间层网络、双分支解码器组成,如图 5 所示。

在网络的输出端,首先通过 1 × 1 的二维卷积(Conv2d)来捕获初始化语音特征,并将特征通道数提升为 64。编码器采用 4 个交互协作单元(En_ICU)提取出高精度的复特征,并逐步压缩特征图尺寸以减小后续网

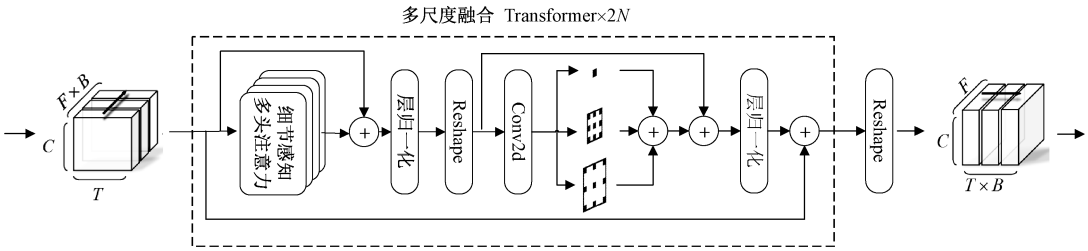

图 4 多尺度融合 Transformer 模块

Fig. 4 Multi-scale fusion Transformer module

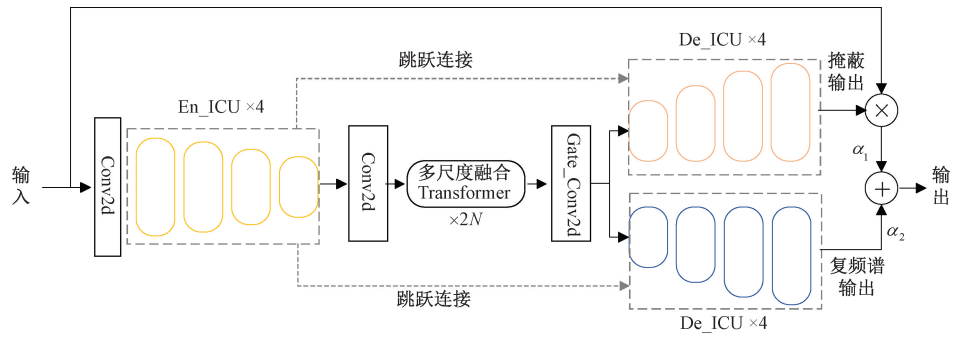

图 5 网络模型结构

Fig. 5 Network model structure diagram

络参数量;作为传输层的中间网络则首先利用 1 × 1 的二维卷积(Conv2d)将通道数减半为 32,接着通过 2N 个多尺度融合 Transformer 模块学习并传递特征,最后利用门控卷积层(Gate_Conv2d)来平滑输出值并恢复通道数。解码器的上下分支则是由 4 个交互协作单元(De_ICU)组成,恢复出压缩特征的位置信息并对特征信息进行补偿,并且输出层末端均将通道数还原为 2。谱映射解码器直接输出估计复频谱,掩码分支输出的估计掩蔽与网络模型输入特征逐点相乘得到此分支估计的复频谱,最后将两分支输出进行加权求和,得到最终估计的纯净语音复频谱。

除此之外,在编码层与对应层级的解码层间采用跳跃连接,以通道维度将深、浅层特征拼接并输入下一解码层,使解码过程更具鲁棒性从而提高模型的精度。

表 1 给出了网络模型参数的详细描述,k、s、d 分别为卷积核大小、步长和膨胀率,每一层模块的输入输出维度按照 $C \times T \times F$ 的格式,其中 T 由输入语音时长决定,且双分支解码部分包含相同的两模块,表中所示的 Conv2d 后均有 LN 与 PReLU 激活函数操作。

表 1 模型参数设置

Table 1 Parameters setting of network model

模块	参数设置	输入维度	输出维度
Conv2d	$k = 1 \times 1, s = 1 \times 1$	$2 \times T \times 512$	$64 \times T \times 512$
En_ICU_1	$s = 1 \times 2$	$64 \times T \times 512$	$64 \times T \times 256$
En_ICU_2	$s = 1 \times 2$	$64 \times T \times 256$	$64 \times T \times 128$
En_ICU_3	$s = 1 \times 2$	$64 \times T \times 128$	$64 \times T \times 64$
En_ICU_4	$s = 1 \times 2$	$64 \times T \times 64$	$64 \times T \times 32$
Conv2d	$k = 1 \times 1, s = 1 \times 1$	$64 \times T \times 32$	$32 \times T \times 32$
多尺度融合 Transformer(× 2N)	$d = 1, 6, 12$	$32 \times T \times 32$	$32 \times T \times 32$
Gate_Conv2d	$k = 1 \times 1, s = 1 \times 1$	$32 \times T \times 32$	$64 \times T \times 32$
De_ICU_4(×2)	$s = 1 \times 2$	$(64 + 64) \times T \times 32$	$64 \times T \times 64$
De_ICU_3(×2)	$s = 1 \times 2$	$(64 + 64) \times T \times 64$	$64 \times T \times 128$
De_ICU_2(×2)	$s = 1 \times 2$	$(64 + 64) \times T \times 128$	$64 \times T \times 256$
De_ICU_1(×2)	$s = 1 \times 2$	$(64 + 64) \times T \times 256$	$2 \times T \times 512$

1.6　损失函数

该网络的学习目标是恢复出纯净语音的复频谱,本文的损失函数采用复损失、幅度损失与时域损失函数的线性组合,并且通过功率压缩[16]来缩小频谱幅值范围并保持语音相位信息,在极坐标下压缩后的复频谱 $\boldsymbol{S}^c$ 为:

$$\boldsymbol{S}^c = |\boldsymbol{S}|^p e^{i\theta} \tag{14}$$

其中,压缩参数 p 设置为 0.3。幅度损失为幅度谱上计算出的纯净语音压缩后幅值与估计语音幅值的最小均方误差损失(mean square error,MSE);复损失为在时频域上分别计算出压缩后纯净语音与估计语音复频谱对应实、虚部分量的 MSE。另外,为进一步改善恢复纯净语音的质量,本文引入时域波形的最小绝对值误差损失(mean absolute error,MAE)。将幅度损失 L_{mag}、复损失 L_{RI} 以及时域损失 L_{time} 定义为:

$$\begin{cases} L_{\text{mag}} = \dfrac{1}{N}\sum_{j=1}^{N}\left(|\tilde{\boldsymbol{S}}|^p - |\boldsymbol{S}|^p\right)^2 \\ L_{\text{RI}} = \dfrac{1}{N}\sum_{j=1}^{N}\left[\left(\tilde{\boldsymbol{S}}_{\text{r}}^{\text{C}} - \boldsymbol{S}_{\text{r}}^{\text{C}}\right)^2 + \left(\tilde{\boldsymbol{S}}_{\text{i}}^{\text{C}} - \boldsymbol{S}_{\text{i}}^{\text{C}}\right)^2\right] \\ L_{\text{time}} = \dfrac{1}{M}\sum_{i=1}^{M}\left(|\tilde{\boldsymbol{s}} - \boldsymbol{s}|\right) \end{cases} \tag{15}$$

其中,N 表示训练样本数量,$\tilde{\boldsymbol{S}}_{\text{r}}^{\text{C}}$、$\tilde{\boldsymbol{S}}_{\text{i}}^{\text{C}}$、$|\tilde{\boldsymbol{S}}|^p$ 分别表示功率压缩后增强语音复频谱的实部、虚部以及幅值,$\boldsymbol{S}_{\text{r}}^{\text{C}}$、$\boldsymbol{S}_{\text{i}}^{\text{C}}$、$|\boldsymbol{S}|^p$ 分别表示纯净语音复频谱的实部、虚部以及幅值。利用平衡因子 λ_1 与 λ_2 将复损失与时域损失同幅度损失相结合,其中 λ_1、λ_2 根据经验分别设置为 0.1 与 0.2,网络训练总损失函数 L 定义为:

$$L = L_{\text{mag}} + \lambda_1 L_{\text{RI}} + \lambda_2 L_{\text{time}} \tag{16}$$

2　实验与结果分析

2.1　数据集设置

为多角度地验证本文语音增强效果,并考虑到信噪比范围、语音数量、噪声种类以及在实际应用场景下的泛化能力等因素,本文将设置大、小两规模的中、英语音数据集来分别对本文方法进行训练与测试,数据集具体构成如下:

数据集 1 为权威公开的 VoiceBank-DEMAND 数据集,该数据集包含了纯净语音和对应的含噪语音,训练、验证集由 14 名男性和 14 名女性的语音对组成,测试集由 1 名男性和 1 名女性的语音对组成,其中噪声为来自 DEMAND 数据集中常见的 15 种生活环境噪声。该训练集在信噪比(signal-to-noise ratio,SNR)为 2.5、7.5、12.5、17.5 dB 的条件下生成了 11 572 对语音数据,测试集在信噪比为 0、5、10、15 dB 的条件下生成了 824 对语音数据。在训练集中随机抽取了 1 157 对纯净-含噪语音对作为验证集,剩下的 10 415 对语音数据作为训练集。

数据集 2 选取了清华大学提供的中文 THCHS30 语料库中的纯净语音,并通过随机抽取的方式分别得到 2 000 条、400 条和 150 条纯净语音用作训练、验证和测试。在训练和验证过程中共混合了 115 种噪声,其中包括来自文献[17]的非语音噪声、NOISEX92 中的工业噪声和 Aurous 数据库中的常见生活噪声如汽车、街道、餐厅和展览会等。同时,所有噪声拼接成一个长矢量,随机切割成与输入纯净语音等长度的噪声,并在信噪比为[-5 dB,10 dB]的范围中以 1 dB 为间隔进行随机选择来混合。为了测试网络的泛化能力,测试集噪声从 NOISEX92 中选择了另外 4 种不匹配噪声如 Babble、M109、Fatory2、White,并分别在信噪比[-5 dB,10 dB]的范围内以 5 dB 为间隔进行混合以生成测试数据集。最终,该数据集约包含了 54 h 的训练时长、6 h 的验证时长和 1 h 的测试时长。

2.2　实验参数设置及评估标准

实验环境如表 2 所示,所有语音信号均采样到 16 kHz,采用窗长约为 63 ms,窗移为 16 ms 的汉明窗进行短时傅里叶变换,且由于频谱具有共轭对称性,故仅采用 1/2 频率维度进行训练。可学习的加权因子 α_1、α_2 均初始化设置为 0.5。实验中批处理大小为 2,Epoch 设置为 120,并采用 Adam 优化器优化网络模型参数,前 30 轮的初始学习率为 0.000 5 并保持不变,30 轮后若验证集损失连续 1 个训练轮次不减少,则学习率衰减 0.5 倍,若连续 5 个训练轮次不减少,则停止训练。

表 2　实验环境设置

Table 2　Experimental environment setup

类别	环境条件
显卡	NVIDIA GeForce RTX 3090
深度学习框架	Pytorch1.9
CUDA 版本	11.1
CUDNN 版本	11.1
脚本语言	Python3.8

本文实验采用的客观指标为:语音感知质量测评(perceptual evaluation of speech quality,PESQ),其取值范围为[-0.5,4.5],值越大表示语音听觉感受越好[18];短时客观可懂度(short-time objective intelligibility, STOI),其取值范围为[0,1],值越大则语音可懂度越高[19],本文均采用百分比数值来表示;对数谱距离(log spectral distance,LSD)评估恢复语音的失真程度,值越小代表失真程度越少。主观评价指标的为 MOS,由于 MOS 评估成本较高,多用信号失真测度(CSIG)、噪声失真测度(CBAK)和综合质量测度(COVL)等客观计算方法来拟合,其取值范围均为[1,5],值越大则语音质量越好。

2.3 实验结果及对比分析

1) 消融实验

首先在数据集 1 上考察中间层多尺度融合 Transformer 模块组数对网络性能的影响。如表 3 为增加 N 的数量对网络性能评价指标 PESQ 和 STOI 的影响,可以观察出在 N 取值为 3 或 4 时,两个指标得分均较高且相近。当 N 取值为 5,两个指标得分均有下降的趋势,说明该网络参数量已冗余并对性能产生了影响。尽管多尺度融合 Transformer 组数 $N=3$ 时,在 STOT 上得分相比于 $N=4$ 时降低了 0.2,但模型总参量却减少 15%,故本文从指标得分与参数量上综合考量,将 $N=3$ 设置为网络固定参数。

表 3 不同模型参数对语音增强效果的比较

Table 3 Comparison of different model parameters on speech enhancement effect

多尺度融合 Transformer 组数 N	PESQ	STOI/%	参数量/M
1	2.81	93.05	0.38
2	2.96	94.96	0.47
3	**3.19**	95.24	**0.57**
4	3.14	**94.44**	0.67
5	3.09	94.93	0.77

表 4 为探究网络输出方式对语音增强性能的影响,在本文网络模型基础上分别考察了以下情况:解码器仅采用单分支输出复掩蔽、单分支输出复频谱、采用双分支联合输出、采用双分支联合并加权输出。同样在数据集 1 上进行验证。可以看出采用联合两训练目标的方法时,网络性能相对于单目标情况均有大幅提升,这说明联合估计能有效提升单个掩码估计与复谱映射在语音增强上的性能上界。另外,本文在此基础上进一步对双分支输出进行加权求和,使得 PESQ 与 STOI 指标分别提升了 0.08 与 0.06。

表 4 不同网络模块的消融实验结果比较

Table 4 Comparison of ablation results of different network modules

模型	PESQ	STOI/%	CSIG	CBAK	COVL
-复掩码	3.04	94.92	4.34	3.59	3.79
-复谱映射	3.02	94.71	4.28	3.53	3.74
-联合相加	3.11	95.18	4.45	3.60	3.84
-联合加权相加	**3.19**	**95.24**	**4.58**	**3.70**	**3.83**

2) 对比实验

表 5 横向对比了多种相关的语音增强网络的性能,其中除了经典的语音模型还包括了近年来含有 Transformer 的语音增强模型如处理时域信号的 TSTNN、CAUNet、CPTNN 以及处理复频域信号的 T-GSA,和同样处理复频域语音信号的 DCCRN、PHASEN 模型等。从表 5 中可以看出,与同样引入 Transformer 的方法相比,本文在 PESQ 上的提升幅度为 0.12~0.23,说明本文改造的多尺度融合 Transformer 相比普通双路径 Transformer 明显利于网络性能的提升。其次,相比于复频谱处理域上的语音增强模型,本文方法在 PESQ 得分指标上显著提升了 0.29~0.51,进一步证明了加权联合训练目标的有效性。综上所述,所提的网络模型能在各项主、客观指标上取得提升,另外相比于 CPTNN,网络参数量仍能减少 25%,说明本文方法能更好地抑制噪声,同时实现了模型精度与复杂度间的良好平衡。

表 5 在 VoiceBank-DEMAND 数据集上的语音增强方法性能对比

Table 5 Performance comparison of speech enhancement methods on VoiceBank-DEMAND dataset

	处理域	PESQ	STOI/%	CSIG	CBAK	COVL	参数量/M
含噪语音	—	1.97	91	3.34	2.44	2.63	—
Wave U-Net, 2018[20]	时域	2.40	—	3.52	3.24	2.96	10
MetricGAN, 2019[21]	幅值域	2.86	—	3.99	3.18	3.42	—
T-GSA, 2020[9]	复频域	3.06	—	4.18	3.59	3.62	63
DEMUCS, 2020[22]	时域	3.07	**95**	4.31	3.40	3.63	33.5
DCCRN, 2020[6]	复频域	2.68	94	3.88	3.18	3.27	3.7
PHASEN, 2020[12]	复频域	2.99	—	4.21	3.55	3.62	5.05
TSTNN, 2021[10]	时域	2.96	**95**	4.33	3.53	3.67	0.92
CAUNet, 2021[11]	时域	2.96	**95**	4.22	3.53	3.60	1.04
CPTNN, 2022[23]	时域	3.07	**95**	4.40	3.59	3.76	0.76
Propose	复频域	**3.19**	**95**	**4.58**	**3.70**	**3.83**	**0.57**

为考察本文方法在真实应用场景中的泛化抗噪性能,采用数据集 2 进行验证。如表 6 所示,选取了 3 中在复频谱处理域上的先进语音增强模型进行比较,包括了 CCRN[4]、用门控线性模块替代 CCRN 中普通卷积层的门控卷积递归网络(gate convolution recurrent network, GCRN)[5]、以及网络整体采用复数运算的深度复卷积递归网络(deep convolution recurrent network, DCCRN)[6]。从表 6 中可以看出,随着信噪比增加,各方法能恢复出更

纯净的语音。本文方法除了在-5 dB 下的 STOI 指标略低于 DCCRN,在其余信噪比下的 PESQ 与 STOI 均高于其他网络。这说明本文方法明在低信噪比下的抗噪能力还需加强,但总体的 PESQ 平均提升了 0.25~0.36,STOI 平均提升了 0.32~1.44,尤其是在高信噪比下的两指标提升明显。除此之外,本文在参数量上依然保持着优势,只占表 6 中其余模型参数量的 3%~15%。综上所述,本文方法在不匹配噪声的情况下,具有更好的泛化能力与鲁棒性且模型相对轻量。

表 6　在大数据下且不同信噪比下各方法的泛化性能对比

Table 6　Comparison of generalization performance of different networks under unmatched noise

SNR	PESQ				STOI/%				参数量/M
	-5	0	5	10	-5	0	5	10	—
含噪语音	1.04	1.08	1.20	1.50	67.05	77.32	85.18	90.42	—
CCRN	1.76	2.22	2.61	2.96	83.97	89.88	94.13	95.83	17.44
GCRN	1.86	2.29	2.63	2.94	84.47	90.23	94.34	96.46	9.77
DCCRN	1.91	2.35	2.70	3.06	**85.03**	91.68	94.84	96.73	3.74
Propose	**2.15**	**2.57**	**2.97**	**3.31**	85.01	**92.445**	**94.97**	**97.13**	**0.57**

图 6 展示了在短时突发背景噪声下,各方法恢复出的纯净语音波形,其中突发背景噪声为随机选择出的汽车鸣笛声。从图 6 中的椭圆框可以观察出,面对突发噪声,DCCRN 与本文方法能恢复出基本的波形,而 CCRN、GCRN 虽去掉了背景噪声但波形幅值失真较为严重。另外,可以观察出本文相比于 DCCRN 在该片段前后恢复出的语音包络还原度更高,说明在面对生活中常见的突发短时噪声的情况下,本文方法能更有效地保留原始语音清晰度,对突发噪声的抑制能力较强,增强语音的整体听觉效果更好。

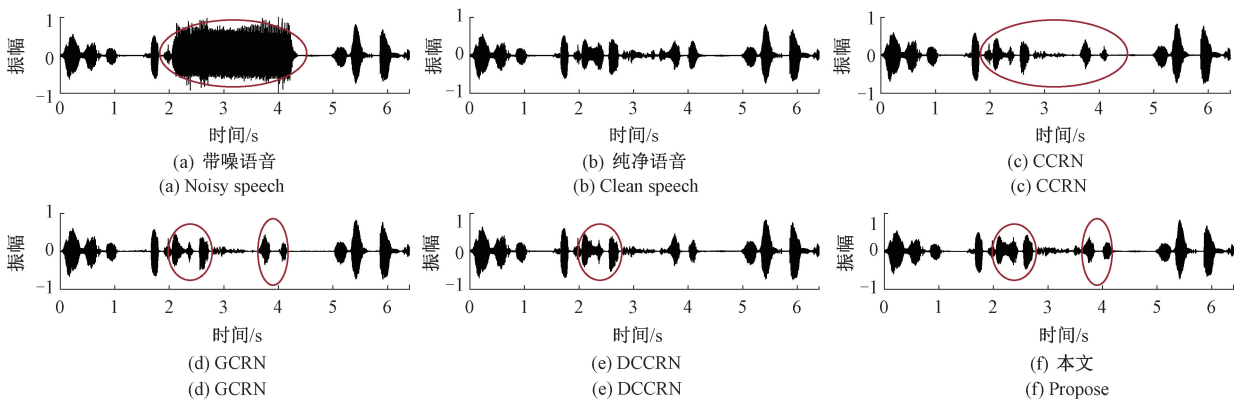

图 6　短时突发噪声下不同方法的增强语音波形图

Fig. 6　Enhanced speech waveforms using different methods under short time burst noise

在数据集 2 下,图 7 讨论了各方法在不同噪声的不同信噪比下的平均 LSD 值。对比曲线可以观察到,所有方法对于 White 噪声均有更好的抑制能力,而本文方法除了在 M109 噪声上与 DCCRN 取得相近 LSD 外,在其余噪声下,均取得更低 LSD,这说明下在大部分噪声下,本文模型对噪声学习并抑制的能力更强,恢复出语音波性的失真程度更小,能更准确的对语音序列建模。

图 8 直观比较了各方法估计的增强语音的语谱图,其中是以一条-5 dB 下随机选取的含 M109 噪声的女声语音为例,图 8(a)、与(b)分别为含噪语音、纯净语音的语谱图。从各方法得到语谱图可看出,CCRN 与 GCRN 虽基本恢复出语谱形状,但在低频段的局部部分出现失真情况,故整体增强效果一般。DCCRN 虽能恢复出大致细节信息,但在 0~2 s 处的轻声语音被当作噪声抑制掉了。从图 8(f)可以看出,本文方法对低频语音上的噪声仍抑制更充分,并且能良好恢复谐波部分,语谱图还原度整体相对较高。

3　结　论

本文提出了一种基于多尺度融合 Transformer 且联合复掩码估计与复频谱映射两训练目标的协作式单通道语音增强方法。在主干网络上提出交互协作单元来促使不同的信息流进行补偿与监督;中间网络则是提出多尺度融合 Transformer 来充分捕获语音细节信息;本文复频域上联合掩蔽与频谱映射,进一步加强该方法抑制噪声的

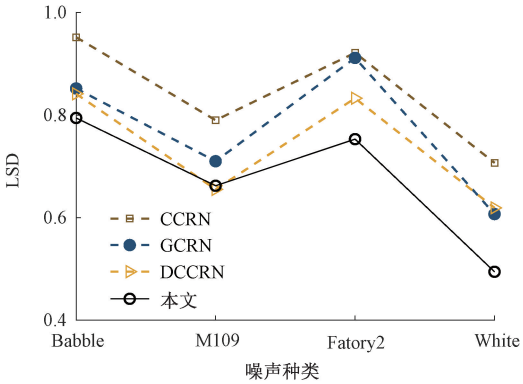

图 7 不同噪声下各方法的平均 LSD 得分

Fig. 7 Average LSD scores of each method under different noises

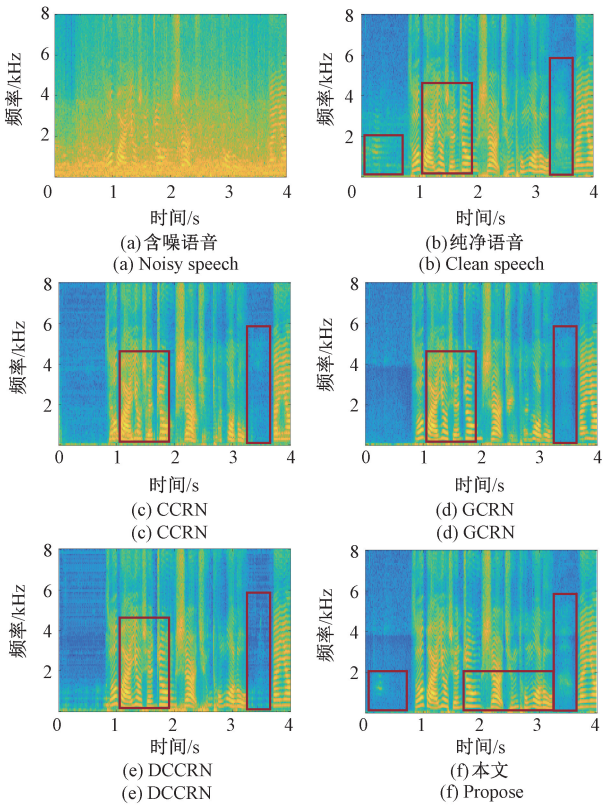

图 8 语谱图

Fig. 8 Spectrogram

性能。本文方法在大、小中英数据集上得到充分验证,证明了该方法在语音增强上存在一定优越性,并且能实现模型参数与精度间的良好平衡。后续研究还需进一步优该方法在复频域上优化估计过程,考虑更复杂的噪声环境。

参考文献

[1] TRINH V A, BRAUN S. Unsupervised speech enhancement with speech recognition embedding and disentanglement losses[C]. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: 391-395.

[2] 韩鑫怡, 张洪德, 柳林, 等. 基于 WDGAN-div 的语音增强方法[J]. 电子测量技术, 2021, 44(21): 64-70.
HAN X Y, ZHANG H D, LIU L, et al. Speech enhancement method based on WDGAN-div [J]. Electronic Measurement Technology, 2021, 44(21): 64-70.

[3] 李吉祥, 倪旭昇, 颜上取, 等. 基于 A-DResUnet 的语音增强方法[J]. 电子测量与仪器学报, 2022, 36(10): 131-137.
LI J X, NI X SH, YAN SH Q, et al. Speech enhancement method based on A-DResUnet[J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(10): 131-137.

[4] TAN K, WANG D L. Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement [C]. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore (ICASSP), 2019: 6865-6869.

[5] TAN K, WANG D L. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement [J]. IEEE/ACM Transaction on Audio, Speech, and Language Processing, 2020, 28: 380-390.

[6] HU Y X, LIU Y, LV S B, et al. DCCRN: Deep complex convolution recurrent network for phase-aware speech Enhancement[C]. Proceeding of the Interspeech 2020, 2020: 2472-2476.

[7] NARAYANAN A, WANG D. Ideal ratio mask estimation using deep neural networks for robust speech recognition[C]. IEEE International Conference on Acoustics, Speech and Signal Processing, 2013: 7092-7096.

[8] WILLIAMSON D S, WANG Y, WANG D. Complex ratio masking for monaural speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 24(3): 483-492.

[9] KIM J, EI-KHAMY M, LEE J. T-GSA: Transformer with Gaussian-weighted self-attention for speech enhancement [C]. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 6649-6653.

[10] WANG K, HE B B, ZHU W P. TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain[C]. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021: 7098-7102.

[11] WANG K, HE B B, ZHU W P. CAUNet: Context-aware U-Net for speech enhancement in time domain[C]. 2021 IEEE International Symposium on Circuits and Systems (ISCAS), 2021: 1-5.

[12] YIN D L, LUO C, XIONG Z W, 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.

[13] CHAZAN S E, GOLDBERGER J, GANNOT S. Speech enhancement using a deep mixture of experts[J]. 2017, DOI:10.48550/arXiv.1703.09302.

[14] 黄星华, 吴天舒, 杨玉龙, 等. 一种面向旋转机械的基于 Transformer 特征提取的域自适应故障诊断[J]. 仪器仪表学报, 2022, 43(11): 210-218.
HHUANG X H, WU T SH, YANG Y L, et al. Domain adaptive fault diagnosis based on Transformer feature extraction for rotating machinery[J]. Chinese Journal of Scientific Instrument, 2022, 43(11): 210-218.

[15] XIANG X X, ZHANG X J, CHEN H Z. A nested U-Net with self-attention and dense connectivity for monaural speech enhancement [J]. IEEE Signal Processing Letters, 2022, 29: 105-109.

[16] LI A D, ZHENG C S, PENG R H, et al. On the importance of power compression and phase estimation in monaural speech dereverberation [J]. JASA Express Letters, 2021, 1(1): 014802.

[17] HU G N, WANG D L. A tandem algorithm for pitch estimation and voiced speech segregation [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(8): 2067-2079.

[18] ITU T P. 826 perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs [S]. International Telecommunications Union (ITU-T) Recommendation, 2001: 862.

[19] 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.

[20] RETHAGE D, PONS J, SERRA X. A wavenet for speech denoising [C]. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018: 5069-5073.

[21] FU S-W, LIAO C F, TSAO Y, et al. MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement [C]. Proceedings of the 36th International Conference on Machine Learning (PMLR), 2019: 2031-2041.

[22] DEFOSSEZ A, SYNNAEVE G, ADI Y. Real time speech enhancement in the waveform domain[J]. 2020, DOI:10.48550/arXiv.2006.12847.

[23] WANG K, HE B B, ZHU W P. CPTNN: Cross-parallel transformer neural network for time-domain speech enhancement [C]. 2022 International Workshop on Acoustic Signal Enhancement (IWAENC), 2022: 1-5.

作者简介

罗庆予(通信作者),2021 年于山东工商学院获得学士学位,现为重庆邮电大学在读硕士研究生,主要研究方向为语音增强与语音分离。

E-mail: luoqingyu_ban@163.com

Luo Qingyu (Corresponding author) received her B. Sc. degree from Shandong Technology and Business University in 2021. Now she is a M. Sc. candidate in Chongqing University of Posts and Telecommunications. Her main research interests include speech enhancement and speech separation.

张天骐,2003 年于电子科技大学获得博士学位,现为重庆邮电大学教授、博士生导师,主要研究方向为通信信号的调制解调、盲处理、图像与语音信号处理。

E-mail: zhangtg@cqupt.edu.cn

Zhang Tianqi received his Ph. D. degree from University of Electronic Science and Technology of China in 2003. Now he is a professor and doctoral supervisor in Chongqing University of Posts and Telecommunications. His main research interests include modulation and demodulation of communication signals, blind processing, and image and speech signal processing.