融合多尺度特征与上下文信息的语音增强方法

更藏措毛[1,2]，黄鹤鸣[1,2]，杨毅杰[1,2]

（1.青海师范大学计算机学院，青海 西宁 810008；2.藏语智能信息处理及应用国家重点实验室，青海 西宁 810000）

摘　要：在语音增强中，常用自编码器结构自动提取特征，但这样得到的特征单一或者冗余且不能较好地捕获语音信号的上下文依赖关系。因此，提出一种融合多尺度特征和上下文信息的语音增强方法 MSF-CI。首先，利用多尺度卷积块提取语音信号的多尺度特征，解决特征单一问题；其次，利用注意力机制关注所提取特征的空间与通道关键信息，解决特征冗余问题；最后，使用门控卷积循环神经网络学习语音信号中跨度较长的上下文依赖关系，并通过门控线性单元提高该网络的非线性学习能力，从而提高模型的泛化性。实验结果表明，MSF-CI 在低信噪比和不同噪声环境下增强语音信号的语音感知质量、短时客观可懂度等多个指标上均优于 GRN、DPT-FSNet、U-Net 等同类的单通道语音增强模型。在信噪比为 0 dB 时，该方法的平均语音感知质量和平均语音客观可懂度达到 1.49 和 0.761。在构建的安多藏语语料库上验证模型的泛化性，平均语音感知质量和平均语音客观可懂度相对于噪声提高了 20.7%和 11.3%，MSF-CI 模型不仅可以提升语音的质量与可理解度，而且具有较优的泛化性。

关键词：语音增强；多尺度特征；注意力机制；门控卷积循环神经网络；对数能量谱

源代码链接：https://gitcode.net/gzcm931205/msf-ci.git

中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0067970

Speech Enhancement Method Incorporating Multi-Scale Features and Contextual Information

Gengzangcuomao[1,2], HUANG Heming[1,2], YANG Yijie[1,2]

(1. School of Computer Science and Technology, Qinghai Normal University, Xining 810008, Qinghai, China;
2. State Key Laboratory of Tibetan Intelligent Information Processing and Application, Xining 810000, Qinghai, China)

【Abstract】 In speech enhancement, Auto-Encoder (AE) structures are typically used to extract features automatically. However, the features obtained in this manner are singular, redundant, and cannot adequately capture the contextual dependencies of speech signals. Therefore, a speech-enhancement method, MSF-CI, that incorporates multi-scale features and contextual information is proposed. First, a multi-scale convolutional block is used to extract multi-scale features of speech signals to solve the issue of single features. Second, the attention mechanism is applied to focus on the spatial and channel key information of the extracted features to eliminate feature redundancy. Finally, a Gated Convolutional Recurrent Neural(GCRN) network is used to learn the long-span context-dependent relations of the speech signal, whereas gated linear units are employed to improve the nonlinear learning ability and thus improve the generalization of the network. Experimental results show that the proposed MSF-CI method outperforms similar single-channel speech-enhancement models such as GRN, DPT-FSNet, and U-Net in terms of speech-perception quality and the short-term objective intelligibility of enhanced speech signals at low Signal-to-Noise Ratios(SNR) and in different noise environments. Under a SNR is 0 dB, the average speech-perception quality and average speech objective intelligibility of the proposed method are 1.49 and 0.761, respectively. The generalizability of the proposed method is verified on the Ando Tibetan corpus. Additionally, its average speech-perception quality and average speech objective intelligibility improved by 20.7% and 11.3%, respectively, with respect to noise. Therefore, the MSF-CI model not only enhances speech quality and intelligibility but also provides better generalization.

【Key words】 speech enhancement; multi-scale feature; attention mechanism; Gated Convolutional Recurrent Neural(GCRN) network; Logarithmic Power Spectrum(LPS)

收稿日期：2023-06-29　　**修回日期：**2023-09-26

基金项目：青海省基础研究计划项目(2022-ZJ-925)；国家自然科学基金(62066039)；省部共建藏语智能信息处理及应用国家重点实验室自主课题(2022-SKL-002，2022-SKL-007)；2021 年青海师范大学自然科学中青年项目科研基金(KJQN2021001)。

通信作者 E-mail：1021489068@qq.com

0 引言

语音增强通过消除背景噪声的干扰，提高语音的质量和可理解度，从而提高自动语音识别、电话会议、助听器等语音相关应用产品的性能。

从通道数量来看，语音增强分为单通道语音增强和多通道语音增强 2 种。由于空间信息的缺乏，使得单通道语音增强依旧面临挑战，因此本文重点关注单通道语音增强。传统的单通道语音增强方法包括谱减法[1]、基于子空间的方法[2-3]、基于统计的方法[4-5]及非负低秩稀疏矩阵分解[6]等。但这些方法消除非平稳噪声的效果不明显。

近年来，一些研究者采用深度神经网络消除非平稳噪声信号，取得了明显的效果。基于深度神经网络的语音增强主要分为时域方法[7-8]和频域方法[9-10]。时域方法直接将语音波形提供给神经网络，并以完整的端到端方式学习去噪模型；频域方法先利用短时傅里叶变换(STFT)得到原始语音的语谱图，再通过深度神经网络降低增强语音与干净语音之间的均方误差。

语音增强方法也可以分为映射方法[11-13]和掩蔽方法[14-16]。映射方法直接估计干净目标，其中频域方法估计干净频谱，时域方法估计干净语音波形。掩蔽方法则预测目标语音的掩蔽，即频域掩蔽[17-18]和时域掩蔽[19]。

目前，基于映射的单通道语音增强方法主要存在以下问题：1)提取的特征单一，表征能力不足；2)提取的特征冗余，影响模型性能；3)不能较好地捕获语音信号的上下文依赖关系。

为解决这些问题，本文提出一种融合多尺度特征与上下文信息的语音增强方法(MSF-CI)，主要贡献如下：1)利用多尺度特征(MSF)提取模块，提取多尺度特征，解决特征单一问题，提高数据的表征能力；2)利用注意力模块关注所提取特征的空间与通道关键信息，消除特征冗余；3)利用门控卷积循环神经(GCRN)网络进行语音增强，能够学习跨度较长的上下文依赖关系，并且增强了模型的非线性学习能力。

1 相关工作

特征对语音增强系统的性能有很大影响，但常用的语音增强方法不能很好地提取不同尺度的语音特征，限制了系统的性能。为解决这一问题，文献[20]提出多尺度与注意力机制相结合的卷积神经网络模型，通过提取不同类型的时域特征，实现了特征间的互补。文献[21]提出结合自注意力机制和密集连通性的嵌套 U-Net 模型，利用多尺度聚合块探索来自不同尺度的上下文信息。文献[22]利用多尺度阶梯时频 Conformer 模块分别提取时域和频域的全局和局部特征。这些方法通过提取多尺度语音特征，提高了模型对语音的重构能力，但忽略了冗余特征对系统性能的影响。

在提取语音特征的模型中引入合适的注意力机制，使模型关注到特征中的关键信息。文献[23]利用自适应频谱时间注意力模块和自适应层次注意力模块聚合中间层次的上下文信息。文献[24]利用多头注意力的 Transformer 模型对重要序列信息进行建模。文献[25]利用具有自注意力机制的膨胀卷积神经网络进行语音增强，膨胀模块提取多尺度特征，注意力模块聚合重要信息。文献[26]使用并行的时间注意力分支和频率通道注意力分支，显式地利用位置信息生成二维注意力图，描述显著的时频语音分布。这些方法有效避免了冗余特征的不利影响，提高了语音增强模型的性能。

语音具有长时依赖关系，捕获这些长时依赖关系有利于提升语音增强的效果。卷积循环神经网络(CRNN)既有卷积神经网络(CNN)的空间特征提取能力又有循环神经网络(RNN)的时间建模能力，因而能够同时有效捕捉语音信号的空间上下文依赖关系和时序上下文依赖关系。文献[27]将原始波形输入到 CRNN 模型进行语音增强，能够很好地提取时域上下文信息。文献[28]利用复数谱映射(CSM)卷积循环神经网络，从带噪语音中分别估计干净语音的实部上下文特征谱图和虚部上下文特征谱图，同时增强带噪语音的幅度和相位，实现了与噪声和说话人无关的语音增强因果系统。

门控线性单元在语音识别中也取得了不错的效果。文献[9]在复数卷积循环神经网络的基础上增加门控线性单元，在解决梯度消失问题的同时保留了非线性能力。FAN 等[29]使用门控循环融合方法，保留模型的非线性能力。TAO 等[30]利用门控层鉴别有用信息，降低带噪信息对识别模型性能的影响。

2 MSF-CI 方法

本节详细介绍 MSF-CI 方法的结构和数据处理过程。该方法主要由注意力机制多尺度特征(AMSF)提取模块和门控卷积循环神经网络增强模块组成。MSF-CI 方法结构如图 1 所示。

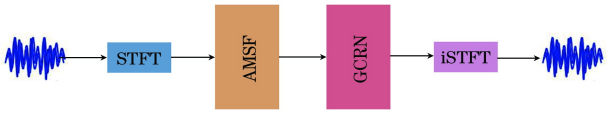

图 1 MSF-CI 方法结构

Fig. 1 Structure of MSF-CI method

首先，通过短时傅里叶变换和对数运算得到带噪语音的对数能量谱(LPS)；其次，利用 AMSF 提取模块从 LPS 中提取多尺度特征并输入到门控卷积循环神经网络进行增强；最后，通过傅里叶逆变换得到增强后的原始语音。

2.1 基于注意力机制的多尺度特征提取模块

AMSF 提取模块由多尺度特征提取模块和空间通道并行压缩与激励(SC_SE)模块构成，如图 2 所示。

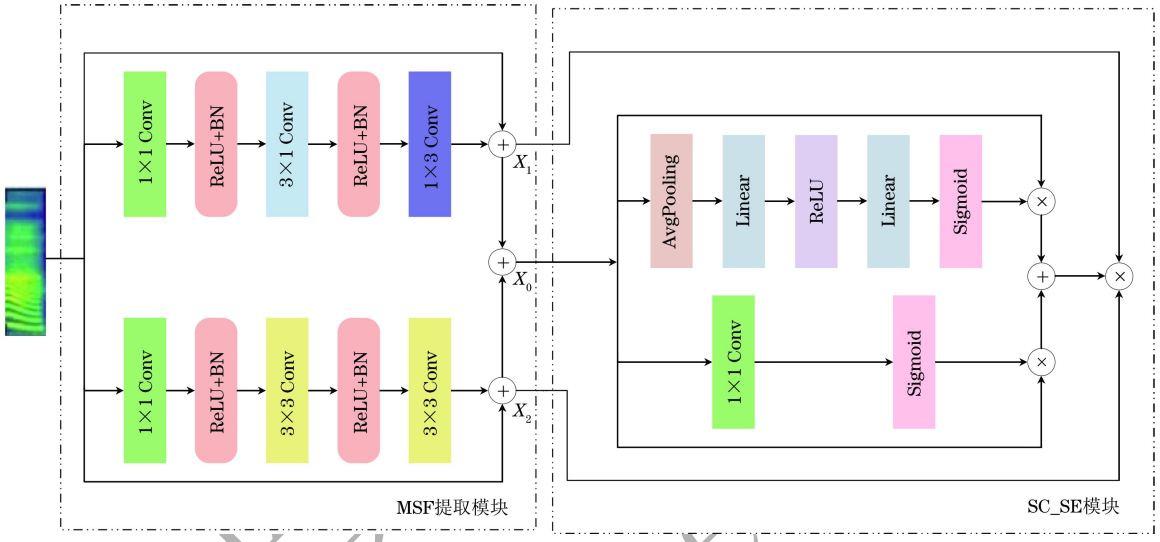

图 2 基于注意力机制的多尺度特征提取模块结构

Fig. 2 Structure of attention mechanism-based multi-scale feature extraction block

2.1.1 多尺度特征提取模块

单纯增加卷积核的个数会使卷积神经网络提取到更多冗余的局部特征。为了解决此问题，本文提出多尺度特征提取方法，通过双分支卷积核来改变感受野，从而更加充分地提取时频域的局部与全局特征。MSF 提取模块由上、下 2 个分支组成。

在上分支中，通过 1×1 的卷积实现数据降维，同时增强模型的非线性能力，将 3×3 的卷积核分解成 1 个 1×3 和 1 个 3×1 的卷积核，分别捕获时域局部特征和频域局部特征。最后，在卷积操作间加入 ReLU 激活函数与 Batch Normalization 加速网络收敛。

下分支与上分支基本相同，不同之处在于使用 2 个级联的 3×3 卷积核来代替 1 个 5×5 的卷积核，在提取全局特征的同时降低模型参数量。

为防止梯度消失或梯度爆炸，本文将上、下分支的输出分别和原始输入 LPS 进行残差连接。在 MSF 提取模块中，数据的处理过程如下：首先，输入的 LPS 分别通过异构的 2 个并行分支提取特征，丰富特征的多样性；其次，每个分支提取到的特征分别与 LPS 通过加操作实现残差连接，分别得到输出 X_1 和 X_2；最后，将 X_1 和 X_2 相加得到多尺度特征 X_0。

2.1.2 SC_SE 模块

MSF 提取模块在提取高级特征的同时会导致特征冗余。SC_SE 模块能够关注频谱中的关键空间信息和通道信息，获取鉴别性强的特征。数据处理过程如下：

1)压缩操作。在空间维度进行特征压缩，将每个二维的特征通道变成实数。这个实数具有某种程度的全局感受野，并且输出的维度和输入的特征通道数相匹配。具体操作是对原特征图 $U \in \mathbb{R}^{C\times T\times F}$ 进行全局平均池化，得到 1 个具有全局感受野的特征图 $U \in \mathbb{R}^{1\times 1\times C}$。

2)激励操作。类似于循环神经网络的门控机制，通过参数 ω 为每个特征通道生成权重，对特征通道间的相关性进行显式建模。

3)重新标定。通过重加权操作，激励输出权重代表的是经过特征选择后每个特征通道的重要性，通过乘法逐通道加权到先前特征上，从通道维度上完成对原始特征的重新标定。计算过程如式(1)～式(4)所示：

$$z_k = \frac{1}{T\times F}\sum_{i=1}^{T}\sum_{j=1}^{F}U_k(i,j) \tag{1}$$

$$\hat{U}_{sSE} = F_{sSE}(U) = [\sigma(q_{1,1})u^{1,1},\cdots,\sigma(q_{T,F})u^{T,F}] \tag{2}$$

$$\hat{U}_{cSE}=F_{cSE}(U)=$$
$$[\sigma(\hat{z}_1)u_1,\sigma(\hat{z}_2)u_2,\cdots,\sigma(\hat{z}_c)u_c] \tag{3}$$

$$\hat{U}_{scSE}=\hat{U}_{cSE}+\hat{U}_{sSE} \tag{4}$$

输入特征图 $U=[u_1,u_2,\cdots,u_c]$ 由 C 个通道 $u_i\in\mathbb{R}^{T\times F}$ 构成；空间压缩由全局平均池化层执行，表示第 i 个通道重要性的 $\sigma(\hat{z}_i)$ 被重新缩放，能够自适应地调整网络，使它强调重要通道而忽略非重要通道，同理，得到通道压缩空间激励块。

将 SC_SE 模块的输出 X 与 MSF 提取模块的输出 X_1 和 X_2 相乘，能够更好地聚合频域上下文信息，获得对噪声的动态表征。

2.2 基于门控卷积循环神经网络的语音增强

本文在利用 GCRN 网络进行语音增强时，CNN 提取输入信号的特征，而 RNN 实现对信号时间进行动态建模。因此，GCRN 网络能够同时捕获输入信号的频域和时域信息。

门控卷积循环神经网络结构如图 3 所示。GCRN 网络分别由 5 层门控卷积块、注意力机制、长短时记忆(LSTM)网络以及 5 层门控反卷积块构成。

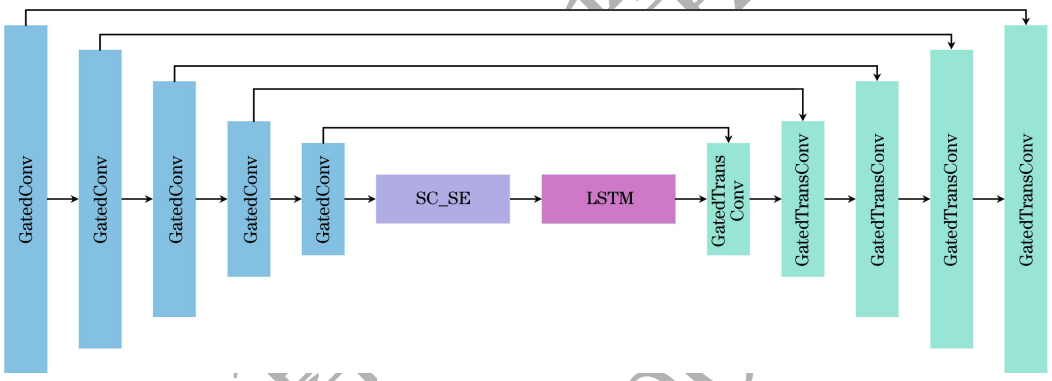

图 3 门控卷积循环神经网络结构

Fig.3 Structure of gated convolutional recurrent neural network

门控卷积块由上下分支构成：上分支由 1 个卷积层和 Sigmoid 函数组成；下分支由 1 个卷积层和 ReLU 激活函数组成。2 个分支的输出相乘后用 Batch Normalization 进行归一化。计算过程如式(5)所示：

$$Y=\mathrm{ReLU}(x\times w_1+b_1)+(x\times w_2+b_2) \tag{5}$$

门控反卷积块具有类似的结构。门控卷积块和门控反卷积块的结构如图 4 所示。

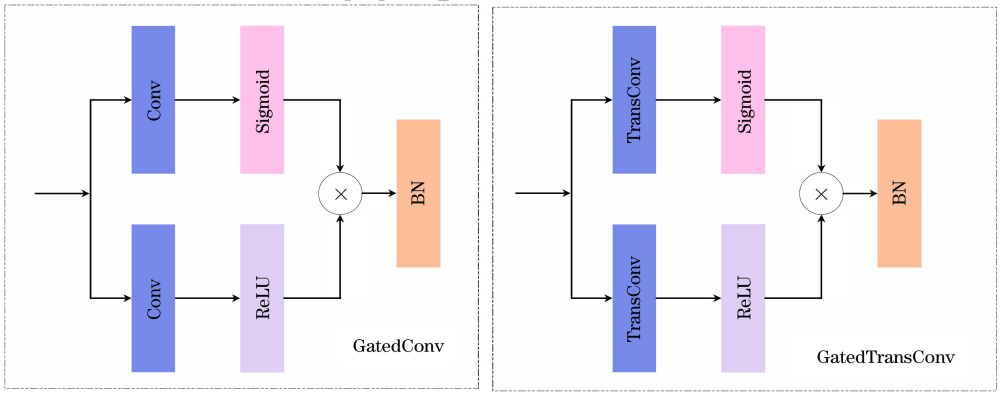

图 4 门控卷积块与门控反卷积块结构

Fig.4 Structure of gated convolutional block and gated deconvolutional block

GCRN 网络对数据的处理过程：首先，对经过 5 层门控卷积块后的输入信息使用注意力机制关注更多时域信息；其次，使用 LSTM 提取语音的时间信息；最后，通过 5 层的门控反卷积块还原原始语音。

3 实验

本文实验环境为 Windows 10 操作系统。GPU 显卡 NVIDIA GeForce RTX 3060，显存为 12 GB 以及 CUDA 11.3、PyTorch 1.11 和 Python 3.8 的软件平台。

3.1 数据集

在清华大学录制的开放式中文数据集 THCHS30 上验证 MSF-CI 方法的性能。THCHS30 包括 1 万多条语音，时长为 40 多个小时，内容主要以诗句为主并且全部都是女声。噪声集使用 Noisex-92。

本文将数据集分为训练集和测试集。训练集包

括 1 500 条语音，每条语音和 Babble、Factory 及 Pink 3 种噪声分别在－5 dB、0 dB 和 5 dB 3 种信噪比(SNR)下依次混合，生成 13 500 条含噪语音。测试集包括 45 条语句，每条语音和已知噪声 Babble 和 Pink 分别在－5 dB、0 dB、5 dB、10 dB 和 15 dB 5 种信噪比下依次混合。同时，每条语音和未知噪声 White 和 Buccaneer1 分别在－6 dB、－3 dB、0 dB、3 dB 和 15 dB 5 种信噪比下依次混合，最终生成 900 条语音。

在西北民族大学构建的安多藏语数据集 XBMU-AMDO31 上验证 MSF-CI 方法的泛化能力。该数据集由 66 名以安多藏语为母语的说话人进行录制。实验测试 200 条语句，取其平均值表征 MSF-CI 方法的泛化性能。

3.2 实验设定与基线实验

本文将所有语音均下采样至 16 kHz。受 GPU 显存的限制，将每条语音切分成 4 s 的片段；使用 512 点 STFT；汉宁窗口的长度设置为 25 ms，相邻 2 帧之间有 75%的重叠。本文方法的学习率为 0.02，优化器为 Adam，在每 5 个 epoch 之后学习率会按照 0.1 的比例进行衰减，以 16 个小批量训练 150 个 epoch，输入的振幅谱大小为(513，257)。

基线方法为基于频域映射的卷积循环神经网络，搭建了 5 层的卷积和 5 层的反卷积，将 LPS 输入到卷积循环神经网络中，完成实验验证。

本文提出模型的参数设置如表 1 所示。输入与输出的尺寸分别表示通道数×时域特征×频域特征。超参数分别表示卷积核大小、步长和通道的大小。SC_SE 的输入尺寸为[256，503，3]，超参数为 1×1，1×1，16，输出尺寸为[256，503，3]。LSTM 输入尺寸为[503，768]，超参数为 768，输出尺寸为[503，768]。

表 1 本文所提网络模型的参数设置

Table 1 Parameter settings of the network model proposed in this paper

模块	层名	输入尺寸	超参数	输出尺寸
多尺度特征提取	ConvBNReLU_0	[1，513，257]	1×1，1×1，1	[1，513，257]
	ConvBNReLU_1	[1，513，257]	1×3，1×1，16	[16，513，257]
	ConyBNReLU_2	[16，513，257]	3×1，1×1，16	[16，513，257]
	ConvBNReLU_3	[1，513，257]	1×1，1×1，16	[16，513，257]
	ConvBNReLU_4	[16，513，257]	3×3，1×1，16	[16，513，257]
	ConvBNReLU_5	[16，513，257]	3×3，1×1，16	[16，513，257]
编码器	GatedConvBlock_1	[16，513，257]	3×9，1×2，48	[48，511，126]
	GatedConvBlock_2	[48，511，126]	3×9，1×2，64	[64，509，60]
	GatedConvBlock_3	[64，509，60]	3×9，1×2，128	[128，507，27]
	GatedConvBlock_4	[128，507，27]	3×9，1×2，256	[256，505，11]
	GatedConvBlock_5	[256，505，11]	3×9，1×2，256	[256，503，3]
解码器	GatedTransConvBlock_1	[256，503，3]	3×9，1×2，256	[16，256，505，11]
	GatedTransConvBlock_2	[16，256，505，11]	3×9，1×2，128	[16，128，507，27]
	GatedTransConvBlock_3	[16，128，507，27]	3×9，1×2，64	[16，64，509，60]
	GatedTransConvBlock_4	[16，64，509，60]	3×9，1×2，48	[16，48，511，126]
	GatedTransConvBlock_5	[16，48，511，126]	3×9，1×2，1	[16，1，513，257]

3.3 评价标准

本文利用 5 种评价指标评估方法的性能：语音质量感知评估(PESQ)、短时目标可理解度(STOI)、语音信号失真的平均意见得分(CSIG)、CBAK 和 COVL。这些指标的值越高表明方法的性能越优。

1)PESQ 的取值范围为－0.5～4.5。

2)STOI 取值范围为 0～1。

3)CSIG 取值范围为 1～5。

4)CBAK 表示背景噪声干扰性的 MOS 预测，取值范围为 1～5。

5)COVL 表示整体处理后语音质量的 MOS 预测，取值范围为 1～5。

3.4 实验结果

3.4.1 消融实验

为验证每个模块的有效性以及模型设计的合理性，本文分别设计平稳噪声环境下和非平稳噪声环境下的消融实验，如表 2 和表 3 所示，加粗表示最优数据(下同)，S 表示 SNR(下同)。MSF+GCRN 表示

在 GCRN 之前加入多尺度特征提取模块；AMSF＋GCRN 表示在 MSF＋GCRN 后加入注意力机制。从表 2 和表 3 可以看出，可以得到如下结论：

1）GCRN 在已知噪声与未知噪声下均有效地提升了 STOI 与 PESQ 的值，说明利用 GCRN 进行语音增强是有效的。例如，在 Pink 噪声环境下，PESQ 的平均值从带噪语音 1.206 提升至 GCRN 1.682，STOI 的平均值从 73.82% 提升至 81.02%。

2）MSF＋GCRN 模型的 PESQ 值在已知噪声与未知噪声的任何 1 个 SNR 上均高于模型 GCRN 的 PESQ 值，说明 MSF 提取了更有效的特征，提高了语音质量。除了部分 Pink 噪声和在 SNR 为 15 dB 时的 STOI 值略低于 GCRN 以外，MSF＋GCRN 模型的 STOI 值在其他 SNR 情况下均高于 GCRN 的值。例如，在 SNR 为－5 dB 时 White 噪声和 SNR 为－5 dB 噪声下，MSF＋GCRN 模型的性能均优于 GCRN 的性能。

3）相较于仅有 MSF 模块，AMSF＋GCRN 模型随着已知平稳噪声 Pink 下 SNR 的增加，PESQ 的值均有所提升，在不同 SNR 的非平稳噪声下，取得了较好或相似的性能，而 STOI 值在平稳噪声及非平稳噪声环境下，得到了较好或相似的性能结果。例如，在 SNR 为 0 dB 下，PESQ 平均值从 MSF＋GCRN 模型 1.57 提升至 AMSF＋GCRN 模型的 1.61，Pink 噪声的平均值从 1.712 提升至 1.744。

4）对于未知的平稳噪声和非平稳噪声，注意力机制模块能够有效提升 PESQ 的值。例如，在 Buccaneer1 噪声下，MSF-CI 模型的 PESQ 平均值从 1.368 提升至 1.394。对已知噪声，相较于 AMSF＋GCRN 模型，MSF-CI 模型具有相近的 PESQ 值，但其 STOI 的值更高，证明注意力机制模块可提升语音的可理解度。

表 2　在平稳噪声下 STOI 及 PESQ 结果

Table 2　Results of STOI and PESQ in stationary noise

模型	STOI/%										PESQ									
	已知噪声 Pink					未知噪声 White					已知噪声 Pink					未知噪声 White				
	$S=$ −5 dB	$S=$ 0 dB	$S=$ 5 dB	$S=$ 10 dB	$S=$ 15 dB	$S=$ −6 dB	$S=$ −3 dB	$S=$ 0 dB	$S=$ 3 dB	$S=$ 15 dB	$S=$ −5 dB	$S=$ 0 dB	$S=$ 5 dB	$S=$ 10 dB	$S=$ 15 dB	$S=$ −6 dB	$S=$ −3 dB	$S=$ 0 dB	$S=$ 3 dB	$S=$ 15 dB
带噪语音	56.3	66.1	75.4	83.0	**88.3**	59.1	64.8	70.3	75.3	**89.7**	1.03	1.04	1.10	1.27	1.59	1.04	1.04	1.05	1.07	1.47
GCRN	71.9	78.6	**82.7**	**85.1**	86.8	**71.6**	76.1	77.9	81.2	86.7	1.31	1.54	1.72	1.87	1.97	1.33	1.46	1.58	1.68	1.97
MSF＋GCRN	72.1	78.6	82.6	84.8	86.2	**71.6**	**76.8**	**78.8**	81.2	86.1	1.35	1.57	1.76	1.90	1.98	**1.37**	1.51	1.62	1.73	1.98
AMSF＋GCRN	71.1	77.9	82.0	84.5	86.1	70.1	74.8	77.9	80.5	86.0	**1.36**	**1.61**	**1.80**	**1.93**	**2.02**	**1.37**	1.51	**1.64**	**1.76**	2.01
MSF-CI	**71.9**	**78.7**	**82.7**	**85.1**	86.5	71.1	76.6	78.7	**81.3**	86.3	1.34	1.58	1.77	**1.93**	**2.02**	1.36	**1.52**	**1.64**	1.74	**2.03**

表 3　在非平稳噪声下 STOI 及 PESQ 的结果

Table 3　Results of STOI and PESQ in non-stationary noise

模型	STOI/%										PESQ									
	已知噪声 Babble					未知噪声 Buccaneer1					已知噪声 Babble					未知噪声 Buccaneer1				
	$S=$ −5 dB	$S=$ 0 dB	$S=$ 5 dB	$S=$ 10 dB	$S=$ 15 dB	$S=$ −6 dB	$S=$ −3 dB	$S=$ 0 dB	$S=$ 3 dB	$S=$ 15 dB	$S=$ −5 dB	$S=$ 0 dB	$S=$ 5 dB	$S=$ 10 dB	$S=$ 15 dB	$S=$ −6 dB	$S=$ −3 dB	$S=$ 0 dB	$S=$ 3 dB	$S=$ 15 dB
带噪语音	54.4	64.4	71.3	81.5	**87.1**	51.3	57.0	62.9	68.9	**86.7**	1.05	1.07	1.15	1.36	1.73	1.03	1.04	1.05	1.07	1.53
GCRN	67.3	76.2	81.7	84.7	86.8	53.8	61.6	67.7	72.9	84.5	1.21	1.42	1.66	1.83	1.97	1.11	1.17	1.26	1.37	1.87
MSF＋GCRN	68.7	77.0	82.0	84.8	86.5	54.0	62.1	68.4	74.0	84.6	**1.26**	**1.49**	**1.72**	1.88	1.98	**1.12**	1.18	1.28	1.42	1.90
AMSF＋GCRN	66.7	75.7	81.0	84.1	86.0	52.7	60.5	67.0	72.6	84.1	**1.26**	**1.49**	**1.72**	**1.90**	**2.02**	**1.12**	1.17	1.26	1.38	1.91
MSF-CI	**69.5**	**77.3**	**82.3**	**85.0**	86.8	**55.7**	**63.3**	**69.7**	**74.8**	85.2	**1.26**	**1.49**	1.71	1.89	**2.02**	**1.12**	**1.19**	**1.29**	**1.43**	**1.94**

3.4.2　与其他模型的对比

为了评估模型增强性能，表 4 与表 5 分别给出了在平稳噪声和非平稳噪声下不同模型的性能对比。CED 表示卷积编码器-解码器模型；CRN 表示卷积循环神经网络；GCRN 表示具有门控单元的卷积循环神经网络；U-Net 表示基于时频掩蔽的 U 型网络；DPT-FSNet 和 TSTNN 分别表示双路的 Transformer 网络和基于时域特征的 Transformer 网络。从表 4 和表 5 可以看出，相较于现有的深度神经网络模型，MSF-CI 模型的 PESQ 值均取得了

较优的结果。例如，对比经典的 CRN 模型，White 噪声在 SNR 为−6 dB、−3 dB、0 dB、3 dB 和 15 dB 下 PESQ 值分别提高了 0.21、0.29、0.36、0.42 和 0.68。在 STOI 指标上性能较差的原因是在高 SNR 条件下，语音出现了过平滑现象，导致语音部分谐波信息丢失。

表 4 在平稳噪声环境下不同模型的实验结果对比

Table 4 Comparison of experimental results of different models in a stationary noise environment

模型	STOI/%										PESQ									
	已知噪声 Pink					未知噪声 White					已知噪声 Pink					未知噪声 White				
	S= −5 dB	*S*= 0 dB	*S*= 5 dB	*S*= 10 dB	*S*= 15 dB	*S*= −6 dB	*S*= −3 dB	*S*= 0 dB	*S*= 3 dB	*S*= 15 dB	*S*= −5 dB	*S*= 0 dB	*S*= 5 dB	*S*= 10 dB	*S*= 15 dB	*S*= −6 dB	*S*= −3 dB	*S*= 0 dB	*S*= 3 dB	*S*= 15 dB
带噪语音	56.3	66.1	75.4	83.0	88.3	59.1	64.8	70.3	75.3	89.7	1.03	1.04	1.10	1.27	1.59	1.04	1.04	1.05	1.07	1.47
CED	58.3	65.7	70.4	72.5	73.0	60.0	63.5	65.8	67.8	70.8	1.09	1.22	1.37	1.39	1.36	1.17	1.24	1.31	1.35	1.34
CRN	61.9	69.7	74.3	76.6	77.2	61.0	65.9	69.3	72.0	76.0	1.16	1.29	1.41	1.45	1.43	1.15	1.21	1.27	1.32	1.35
GCRN	71.9	78.6	**82.7**	**85.2**	86.8	**71.7**	76.1	**79.0**	**81.4**	86.7	1.31	1.54	1.72	1.87	1.97	1.33	1.46	1.57	1.68	1.97
TSTNN[19]	64.6	73.3	78.8	82.1	84.0	63.7	70.0	73.8	77.0	83.3	1.25	1.45	1.66	1.80	1.89	1.28	1.37	1.48	1.59	1.89
DPT-FSNet[24]	59.9	69.8	78.5	85.0	**89.4**	64.4	71.3	76.5	80.5	**91.2**	1.05	1.11	1.26	1.52	1.89	1.11	1.16	1.24	1.34	1.95
U-Net[15]	56.3	73.6	75.4	93.0	88.3	61.7	67.8	72.2	75.7	81.9	1.20	1.30	1.39	1.45	1.47	1.15	1.22	1.29	1.36	1.48
MSF-CI	**71.9**	**78.7**	**82.7**	85.1	86.5	71.1	**76.6**	78.7	81.3	86.3	**1.34**	**1.58**	**1.77**	**1.92**	**2.02**	**1.36**	**1.50**	**1.63**	**1.74**	**2.03**
增益值	15.6	12.6	7.3	2.1	−1.8	12.0	11.8	8.4	6.0	−3.4	0.31	0.54	0.67	0.65	0.43	0.32	0.46	0.58	0.67	0.56

表 5 在非平稳噪声环境下不同模型的实验结果对比

Table 5 Comparison of experimental results of different models in a non-stationary noise environment

模型	STOI/%										PESQ									
	已知噪声 Babble					未知噪声 Buccaneer1					已知噪声 Babble					未知噪声 Buccaneer1				
	S= −5 dB	*S*= 0 dB	*S*= 5 dB	*S*= 10 dB	*S*= 15 dB	*S*= −6 dB	*S*= −3 dB	*S*= 0 dB	*S*= 3 dB	*S*= 15 dB	*S*= −5 dB	*S*= 0 dB	*S*= 5 dB	*S*= 10 dB	*S*= 15 dB	*S*= −6 dB	*S*= −3 dB	*S*= 0 dB	*S*= 3 dB	*S*= 15 dB
带噪语音	54.4	64.4	71.3	81.5	87.1	51.3	57.0	62.9	68.9	86.7	1.05	1.07	1.15	1.36	1.73	1.03	1.04	1.05	1.07	1.53
CED	53.9	62.8	68.9	72.0	72.8	54.2	59.0	63.2	67.0	72.3	1.10	1.24	1.38	1.40	1.38	1.70	1.11	1.18	1.26	1.35
CRN	56.8	65.8	72.1	75.5	76.8	51.6	57.1	65.9	66.9	75.9	1.11	1.23	1.38	1.46	1.45	1.11	1.15	1.21	1.27	1.41
GCRN	67.3	76.1	81.7	84.7	86.8	53.8	61.6	67.7	72.9	84.5	1.21	1.42	1.66	1.84	1.97	1.11	1.16	1.26	1.37	1.86
TSTNN[19]	57.5	68.3	76.1	80.7	83.3	54.9	61.5	67.2	72.1	82.2	1.13	1.30	1.52	1.71	1.84	**1.13**	1.20	1.28	1.39	1.79
DPT-FSNet[24]	54.4	65.9	75.1	82.4	**87.7**	50.5	59.0	66.0	71.6	**87.3**	1.06	1.11	1.23	1.47	1.86	1.06	1.08	1.11	1.15	1.71
U-Net[15]	63.0	72.2	77.6	80.2	81.8	54.0	61.2	66.8	71.2	80.6	1.16	1.26	1.35	1.41	1.45	1.09	1.13	1.17	1.23	1.44
MSF-CI	**69.5**	**77.3**	**82.3**	**85.0**	86.8	**55.7**	**63.3**	**69.7**	**74.8**	85.2	**1.26**	**1.49**	**1.71**	**1.89**	**2.02**	1.12	**1.19**	**1.29**	**1.43**	**1.94**
增益值	15.1	12.9	11.0	3.5	−0.3	4.4	6.3	6.8	5.9	−1.5	0.21	0.42	0.56	0.53	0.29	0.09	0.15	0.24	0.36	0.41

图 5 所示为本文模型 MSF-CI 和其他深度神经网络模型在 4 种信噪比下 5 种不同噪声类型的平均 PESQ 增益值与 STOI 增益值(彩色效果见《计算机工程》官网 HTML 版，下同)。其中，4 种信噪比分别为−5 dB、0 dB、5 dB 和 10 dB，5 种噪声分别指 Babble、White、Buccaneer1、Pink 和 Factory1，可以得到如下结论：

1)MSF-CI 可更好地提高语音的质量。相较于其他模型，MSF-CI 方法获得了最高的 PESQ 增益值，如图 5(a)所示。

2)在低 SNR 下，MSF-CI 模型的鲁棒性更强，故 STOI 增益值明显。从表 4 和表 5 可以看出，在高 SNR 下本文模型的结果在 STOI 指标上已接近最大值，因此其增益值较小，如图 5(b)所示。

图 6 所示为对比的深度神经网络和本文方法 MSF-CI 在−5 dB、0 dB、5 dB 和 10 dB 信噪比下，5 种不同噪声类型的平均 CSIG、CBAK 和 COVL 值。从图 6 可以看出：1)所有方法均有效地增强了语音，在 3 种指标下，所有模型的结果值均高于带噪语音值；2)相较于其他模型，本文方法的每种指标都高，说明性能更优；3)在 3 种评价指标下，MSF-CI 方法性能效果最佳，验证了本文方法的有效性。

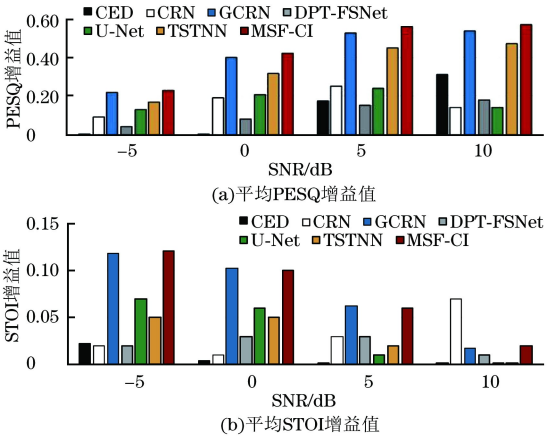

图 5 不同信噪比下 5 种噪声的平均 PESQ 增益值与平均 STOI 增益值

Fig. 5 Average PESQ gain value and average STOI gain value of 5 types of noise under different signal-to-noise ratios

图 7 所示为 MSF-CI 和其他方法增强后的语谱图的对比结果。从图 7 可以看出：本文的语谱图与干净语音相似，证明 MSF-CI 能够有效去除语音噪声。

表 6 所示为 MSF-CI 在安多藏语语料库上的性能。从表 6 可以看出：相较于带噪语音，MSF-CI 在 5 种评价指标上均有所提升。例如，在 0 dB 下，PESQ、STOI、CSIG、CBAK 和 COVL 的性能分别提升了 0.25、8、0.87、0.22、0.56；相较于 GCRN，性能分别提升了 0.1、3.9、0.51、0.16 和 0.32。因此，

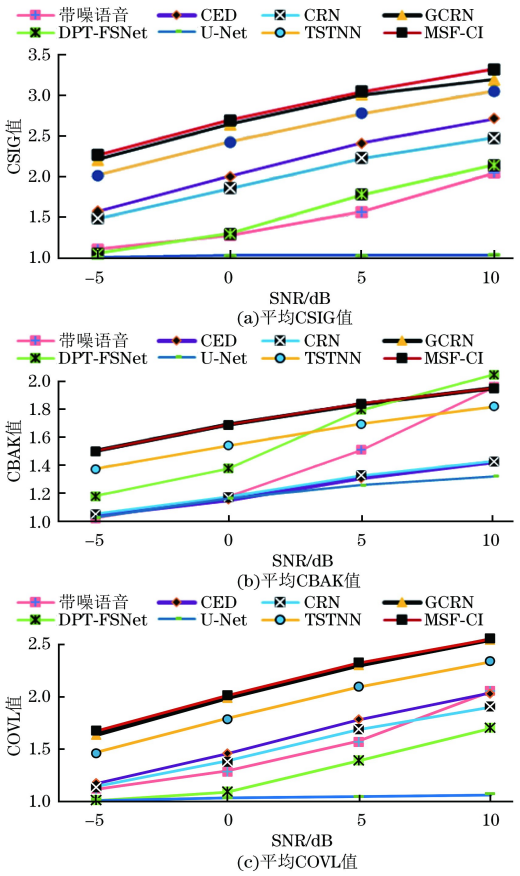

图 6 不同信噪比下 5 种噪声的平均 CSIG 值、CBAK 值及 COVL 值

Fig. 6 Average CSIG, CBAK and COVL values for 5 noise under different SNR

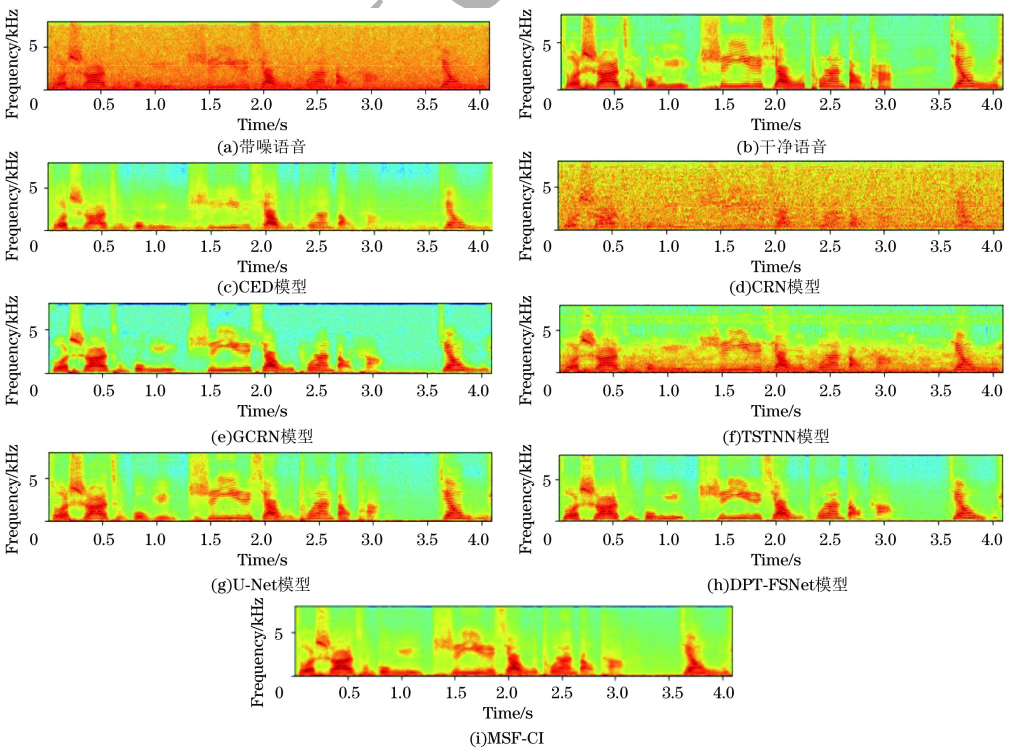

图 7 不同模型的语谱图对比

Fig. 7 Comparison of spectrum maps of different models

MSF-CI 在藏语语料库上也能有较好的效果，说明该方法有较好的泛化能力。

表 6 MSF-CI 在安多藏语语料库上的性能

Table 6 The performance of MSF-CI on the Ando Tibetan language corpus

模型	PESQ			STOI/%			CSIG			CBAK			COVL		
	$S=$ −5 dB	$S=$ 0 dB	$S=$ 5 dB	$S=$ −5 dB	$S=$ 0 dB	$S=$ 5 dB	$S=$ −5 dB	$S=$ 0 dB	$S=$ 5 dB	$S=$ −5 dB	$S=$ 0 dB	$S=$ 5 dB	$S=$ −5 dB	$S=$ 0 dB	$S=$ 5 dB
带噪语音	1.07	1.10	1.17	52.8	64.4	75.2	1.34	1.53	1.80	1.23	1.47	1.76	1.13	1.23	1.40
CED	1.11	1.25	1.33	59.0	68.5	74.0	1.68	1.89	2.07	1.44	1.53	1.60	1.32	1.47	1.61
CRN	1.14	1.22	1.31	55.9	63.5	69.2	1.83	2.05	2.23	1.35	1.43	1.50	1.37	1.53	1.68
GCRN	1.16	1.25	1.33	59.9	68.5	74.0	1.68	1.89	2.08	1.44	1.53	1.61	1.32	1.47	1.61
TSTNN[19]	1.21	1.22	1.31	62.3	66.1	73.0	1.95	1.94	2.13	**1.59**	1.61	1.69	1.48	1.49	1.65
DPT-FSNet[24]	1.12	1.13	1.15	**64.1**	66.1	67.5	1.46	1.45	1.46	1.47	1.54	1.61	1.18	1.18	1.20
U-Net[15]	**1.24**	**1.38**	1.47	60.9	71.2	77.4	1.00	1.06	1.16	1.06	1.31	1.45	1.01	1.06	1.16
MSF-CI	1.21	1.35	**1.48**	62.7	**72.4**	**79.0**	**2.12**	**2.40**	**2.64**	1.58	**1.69**	**1.78**	**1.57**	**1.79**	**2.00**

4 结束语

为解决特征单一、特征冗余问题以及为了更好地捕获语音信号的上下文依赖关系，本文提出一种融合多尺度特征和上下文信息的语音增强方法 MSF-CI。该方法使用多尺度特征提取方法提取特征，解决特征单一问题；利用注意力机制解决特征冗余问题；使用 GCRN 进行语音增强，能够更好地捕获语音信号的上下文依赖关系。实验结果表明，MSF-CI 中的每个模块是有效的，用它进行语音增强能够显著提高语音质量和可理解度，并且该方法具有较强的泛化能力。由于利用 MSF-CI 进行语音增强尚未考虑相位带来的影响，也未考虑模型的计算复杂度，因此后续将尝试使用复数域信息弥补相位问题，以降低模型的计算复杂度。

参考文献

[1] BOLL S. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979, 27(2): 113-120.

[2] YANG Y, LIU P P, ZHOU H L, et al. A speech enhancement algorithm combining spectral subtraction and wavelet transform[C]//Proceedings of the 4th International Conference on Automation, Electronics and Electrical Engineering. Washington D. C., USA: IEEE Press, 2021: 268-273.

[3] JABLOUN F, CHAMPAGNE B. A multi-microphone signal subspace approach for speech enhancement[C]//Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Washington D. C., USA: IEEE Press, 2002: 205-208.

[4] CHEN J D, BENESTY J, HUANG Y T, et al. New insights into the noise reduction Wiener filter[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(4): 1218-1234.

[5] GERKMANN T, HENDRIKS R C. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(4): 1383-1393.

[6] ISLAM M S, ZHU Y Y, HOSSAIN M I, et al. Supervised single channel dual domains speech enhancement using sparse non-negative matrix factorization[J]. Digital Signal Processing, 2020, 100: 102697.

[7] 李江和, 王玫. 一种用于因果式语音增强的门控循环神经网络[J]. 计算机工程, 2022, 48(11): 77-82.
LI J H, WANG M. A gated recurrent neural network for causal speech enhancement[J]. Computer Engineering, 2022, 48(11): 77-82. (in Chinese)

[8] 董宏越, 马建芬, 张朝霞. 基于时域波形映射-频域谐波损失的语音增强[J]. 计算机工程与设计, 2021, 42(6): 1677-1683.
DONG H Y, MA J F, ZHANG Z X. Waveform mapping in time domain and harmonic loss in frequency domain based speech enhancement[J]. Computer Engineering and Design, 2021, 42(6): 1677-1683. (in Chinese)

[9] 袁文浩, 时云龙, 胡少东, 等. 一种基于时频域特征融合的语音增强方法[J]. 计算机工程, 2021, 47(10): 75-81.
YUAN W H, SHI Y L, HU S D, et al. A speech enhancement approach based on fusion of time-domain and frequency-domain features[J]. Computer Engineering, 2021, 47(10): 75-81. (in Chinese)

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

[11] 张天骐, 罗庆予, 方蓉, 等. 基于信息提炼与残差特征聚合网络的单通道语音增强[J]. 信号处理, 2023, 39(7): 1285-1298.
ZHANG T Q, LUO Q Y, FANG R, et al. Single-channel speech enhancement method based on hierarchical refinement and residual feature aggregation network[J]. Journal of Signal Processing, 2023, 39(7): 1285-1298. (in Chinese)

[12] ZEZARIO R E, FU S W, CHEN F, et al. Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 31: 54-70.

[13] TESCH K, GERKMANN T. Insights into deep non-linear filters for improved multi-channel speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 563-575.

[14] BORGSTRÖM B J, BRANDSTEIN M S. Speech enhancement via attention masking network (SEAMNET): an end-to-end system for joint suppression of noise and reverberation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 515-526.

[15] AKTER K, MAMUN N, HOSSAIN M A. A T-F masking based monaural speech enhancement using U-Net architecture [C] // Proceedings of the International Conference on Electrical, Computer and Communication Engineering. Washington D. C., USA: IEEE Press, 2023: 1-5.

[16] MARTÍN-DOÑAS J M, JENSEN J, TAN Z H, et al. Online multichannel speech enhancement based on recursive EM and DNN-based speech presence estimation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 3080-3094.

[17] WILLIAMSON D S, WANG Y X, WANG D L. Complex ratio masking for monaural speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(3): 483-492.

[18] ZHAO S K, MA B. D2Former: a fully complex dual-path dual-decoder conformer network using joint complex masking and complex spectral mapping for monaural speech enhancement [C] // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2023: 1-5.

[19] WANG K, HE B B, ZHU W P. TSTNN: two-stage transformer based neural network for speech enhancement in the time domain [C] // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2021: 7098-7102.

[20] XIANG X X, ZHANG X J, CHEN H Z. A convolutional network with multi-scale and attention mechanisms for end-to-end single-channel speech enhancement[J]. IEEE Signal Processing Letters, 2021, 28: 1455-1459.

[21] 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.

[22] 金玉堂,王以松,王丽会,等. 基于多尺度阶梯时频 Conformer GAN 的语音增强算法[J]. 计算机应用, 2023, 43(11): 3607-3615.
JIN Y T, WANG Y S, WANG L H, et al. Speech enhancement algorithm based on multi-scale ladder-type time-frequency Conformer GAN [J]. Journal of Computer Applications, 2023, 43(11): 3607-3615. (in Chinese)

[23] YU G C, LI A D, WANG H, et al. DBT-Net: dual-branch federative magnitude and phase estimation with attention-in-attention transformer for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 2629-2644.

[24] DANG F, CHEN H T, ZHANG P Y. DPT-FSNet: dual-path transformer based full-band and sub-band fusion network for speech enhancement[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2022: 6857-6861.

[25] PANDEY A, WANG D L. Dense CNN with self-attention for time-domain speech enhancement [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1270-1279.

[26] ZHANG Q Q, QIAN X Y, NI Z H, et al. A time-frequency attention module for neural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 462-475.

[27] ASHUTOSH P, WANG D. A new framework for CNN-based speech enhancement in the time domain[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(7): 1179-1188.

[28] TAN K, WANG D L. Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement [C] // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2019: 6865-6869.

[29] FAN C H, YI J Y, TAO J H, et al. Gated recurrent fusion with joint training framework for robust end-to-end speech recognition[EB/OL]. [2023-05-25]. https://arxiv.org/abs/2011.04249.

[30] TAO F, BUSSO C. Gating neural network for large vocabulary audiovisual speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(7): 1290-1302.

编辑 薛晋栋