DOI:10.19651/j.cnki.emt.2417250

融合双通道卷积和改进型 Conformer 的两阶段语音增强算法*

徐佳瑜[1,2] 郑展恒[1,2,3] 曾庆宁[1,2] 王 健[1,2]

(1.桂林电子科技大学信息与通信学院 桂林 541004; 2.桂林电子科技大学认知无线电与信息处理教育部重点实验室 桂林 541004; 3.南宁桂电电子科技研究院有限公司 南宁 530000)

摘 要:针对语音关键特征提取不充分、模型结构单一的问题,提出一种两阶段下融合多尺度特征和改进型门控 Conformer 的语音增强方法。首先,针对关键特征提取不充分的问题,提出双通道卷积融合模块,采用不同感受野的二维卷积多尺度提取语音关键信息,并结合门控机制增强网络的短期与长期序列相关性,从而提升模型在复杂环境下的语音增强效果;提出改进型 Conformer,采用时间注意和频率注意分别在时域和频域上进行建模,并结合膨胀卷积模块高效提取局部与全局上下文信息,从而增强网络在语音序列建模中的表现能力。其次,针对模型结构单一的问题,采用两阶段处理结构,将复杂问题分步处理。在第一阶段首先接收噪声频谱的幅值,初步估计出干净语音的幅值,并与噪声相位进行重构,得到粗糙的复频谱。第二阶段在第一阶段得到粗谱的基础上进一步提取更精细的特征,增强语音信号的细节表现能力。最后,在 VoiceBank+DEMAND 数据集上进行测试,实验结果表明,所提算法相比带噪语音的语音感知质量和短时客观可懂度分别提升 50.25%、3.26%,表明该网络能够更有效地提高语音的可懂度,同时改善语音信号的整体质量,具有较强的降噪能力。

关键词:深度学习;语音增强;Conformer;多尺度特征提取;两阶段

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

Two-stage speech enhancement algorithm incorporating dual-channel convolution and improved Conformer

Xu Jiayu[1,2] Zheng Zhanheng[1,2,3] Zeng Qingning[1,2] Wang Jian[1,2]

(1. School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China; 2. Key Laboratory of Cognitive Radio and Information Processing, Ministry of Education, Guilin University of Electronic Technology, Guilin 541004, China; 3. GUET-Nanning E-Tech Research Institute Co., Ltd., Nanning 530000, China)

Abstract: In order to solve the problems of insufficient extraction of key speech features and single model structure, a double-stage speech enhancement method incorporating multi-scale features and improved gated Conformer was proposed to solve the problems of insufficient extraction of key features of speech and single model structure. Firstly, in order to solve the problem of insufficient extraction of key features, a two-channel convolutional fusion module was proposed, which used two-dimensional convolutional multi-scale extraction of speech key information with different receptive fields, and combined with the gating mechanism to enhance the short-term and long-term sequence correlation of the network, so as to improve the speech enhancement effect of the model in complex environments. An improved Conformer is proposed, which uses time attention and frequency attention to model in the time and frequency domains respectively, and combines the dilated convolution module to efficiently extract local and global context information, so as to enhance the performance ability of the network in speech sequence modeling. Secondly, for the problem with a single model structure, a two-stage processing structure is adopted to deal with the complex problem step by step. In the first stage, the amplitude of the noise spectrum is received, the amplitude of the clean speech is preliminarily estimated, and the noise phase is reconstructed to obtain the rough complex spectrum. In the second stage, on the basis of the rough spectrum obtained in the first stage, more refined features were further extracted to enhance the detailed expression ability of the speech signal. Finally, the experimental results are carried out on the VoiceBank+DEMAND dataset, and the experimental results show that the objective evaluation index and short-term intelligibility of this model are increased by 50.25% and 3.26%, respectively, compared with the noisy voice, indicating that the proposed algorithm can improve the intelligibility of speech more effectively, and at the same time improve the overall quality of speech signals, and has strong noise reduction ability.

Keywords: deep learning; speech enhancement; Conformer; multi-scale feature extraction; two phases

0 引 言

语音增强旨在从噪声环境中尽可能高效地提取出纯净语音,该技术在语音通信、安全监控、交通导航等多种领域运用广泛。根据采集语音信号的麦克风数量,语音增强可分为单通道语音增强和多通道语音增强。多通道语音增强

收稿日期:2024-10-31

*基金项目:认知无线电与信息处理教育部重点实验室项目(CRKL230103)资助

使用多个麦克风同时捕捉不同方向的声音，尽管能显著提高增强效果，但成本和系统复杂度较高。相比之下，单通道语音增强仅使用一个麦克风，具备系统简单、模型轻量的优势，更具实用性。然而，由于缺乏空间信息，单通道增强面临更大的技术挑战，研究价值也更为突出。

目前现有的语音增强算法包括两类：一类是传统的语音增强算法，另一类是基于深度学习的语音增强算法。常见的传统单通道算法包括谱减法[1]、滤波法[2]、基于统计模型[3]等，然而这些传统算法必须假设噪声平稳，对非平稳噪声的处理效果较差。而基于深度学习的单通道语音增强算法并不需要基于这一假设，并且对非平稳噪声的处理效果比较好。

近年来，基于深度学习的语音增强算法发展迅速，许多神经网络被提出，包括：卷积神经网络(convolutional neural networks，CNN)[4]、长短期记忆网络(long short-term memory，LSTM)[5]、生成对抗网络(speech enhancement generative adversarial network，SEGAN)[6]等。其中，LSTM 凭借其能够有效建模语音信号中长时依赖的能力，被广泛应用于语音增强任务中。Tan 等[7]提出的门控循环神经网络(gated convolution recurrent network，GCRN)利用 LSTM 建模时间依赖性，Hu 等[8]的深度复数卷积递归网络(deep complex convolution recurrent network，DCCRN)则进一步通过复数运算结合复数 LSTM 处理长时依赖性，增强了相位信息的建模能力。然而 LSTM 在捕捉长距离依赖性时存在计算复杂度高、并行效率低的问题，且难以充分建模全局依赖关系。为此，Transformer[9]模型通过自注意力机制有效捕捉全局依赖，显著提高了计算效率。随后，Conformer[10]模型在保留自注意力机制捕捉全局依赖性的同时，融合卷积操作以增强局部特征提取能力。然而，该模型在特征提取上仍然存在一定的局限性。原始 Conformer 仅针对单一尺度进行建模，未能充分捕捉语音信号中多尺度的时频依赖关系。

近年来，在许多任务中，多阶段学习已被证实比单阶段网络更有效，如图像领域[11-12]和语音领域[13-14]。基于此，提出一种两阶段算法，旨在将原始增强问题分成多个更容易解决的子问题。

综上所述，提出两阶段下融合多尺度特征和改进型门控 Conformer 的语音增强方法(a double-stage speech enhancement method incorporating multi-scale features and improved gated conformer，DGC_Net)，采用双通道融合卷积和改进型 Conformer 模块对语音进行多尺度建模，两阶段网络解决单阶段网络性能的局限性。在编解码层，通过双通道卷积融合模块，将 2 个不同膨胀率的二维卷积结合起来，多尺度提取音频的时频信息，更好地学习潜在细节特征。在中间层，设计了双分支改进型 Conformer 模块(two-branch improved conformer，TBI-Conformer)，替换普通卷积为膨胀卷积以此建立全局时间依赖，并采用复值时间注意和复值频率注意建模局部依赖，增强网络在语音序列的建模能力。采用两阶段网络，其中每个子网络均采用 U-Net 结构，在第一阶段粗谱的基础上进一步提取更精细的特征。最后，在公开数据集 VoiceBank+DEMAND 上进行多次实验，结果显示，所提算法实现了较高的语音增强效果，进一步表明所提模型降噪能力更强，更具有效性。

1　模　　型

所提两阶段算法 DGC_Net 如图 1(a)所示。它包含两个处理阶段，通过先粗略估计再精细优化的策略，能有效提升语音信号的清晰度和准确性。在第 1 阶段，融合改进型门控 Conformer 的幅值网络(a magnitude speech enhancement network based on improved gated Conformer，MGC_Net)接收噪声频谱的幅值，初步估计出干净语音的幅值，并与噪声相位相结合，转化为实部和虚部，得到粗糙的复频谱。第 2 阶段融合改进型门控 Conformer 的复值网络(a complex network based on improved gated Conformer，CGC_Net)进一步增强该复频谱，同时通过残差网络恢复第一阶段可能丢失的信息，最终得到增强以后的完整频谱。两阶段网络计算过程为：

$$|\hat{S}_1| = \Im_1(|X|;\phi_1) \tag{1}$$

$$\hat{S}_{r1} = \psi(|S_1| e^{j\theta_X}); \hat{S}_{i1} = \xi(|S_1| e^{j\theta_X}) \tag{2}$$

$$(\hat{S}_{r2}, \hat{S}_{i2}) = \Im_2(\hat{S}_{r1}, \hat{S}_{i1}, X_r, X_i;\phi_2) \tag{3}$$

其中，$\Im_1$ 和 $\Im_2$ 分别表示 MGC_Net 和 CGC_Net 的映射函数，参数集分别为 ϕ_1 和 ϕ_2，ψ 和 ξ 分别表示实操作和虚操作，X 表示原始语音输入信号，$\hat{S}_1$ 和 $\hat{S}_2$ 分别表示 MGC_Net 和 CGC_Net 的估计输出，r 和 i 分别表示复数的实部和虚部。

各个网络的细节如图 1(b)～(c)所示。MGC_Net 采用编码器-解码器结构，其中编码器与解码器各由 6 个双通道卷积融合模块组成，且中间增强阶段引入多个 TBI-Conformer 模块。在编码器中，通过膨胀卷积对输入特征进行下采样，有效提取语音信号的全局和局部特征，能够同时捕捉短时和长时依赖，从而提高噪声抑制能力。中间增强阶段采用多个 TBI-Conformer 结构，对特征信息进行充分提取。解码器则结合不同膨胀率的卷积，从多尺度重建语音信号，确保在大尺度上重建语音信号的同时，保持局部细节的准确性。为防止 MGC_Net 网络中编码器压缩过程中的信息丢失，采用跳跃连接将每个解码器块与对应的编码器块相连接。CGC_Net 网络与 MGC_Net 网络类似，每个编码器和解码器均采用 6 个双通道卷积融合模块，其中 2 个解码器结构分别用于精确估计实部和虚部频谱图。

1.1　双通道卷积融合模块

门控机制最早是为 LSTM 开发的，它在整个网络中能够起到动态调节和管理信息流动的功能。通过多个门控单元，LSTM 能够选择性地从输入信息中决定保留、遗忘或传

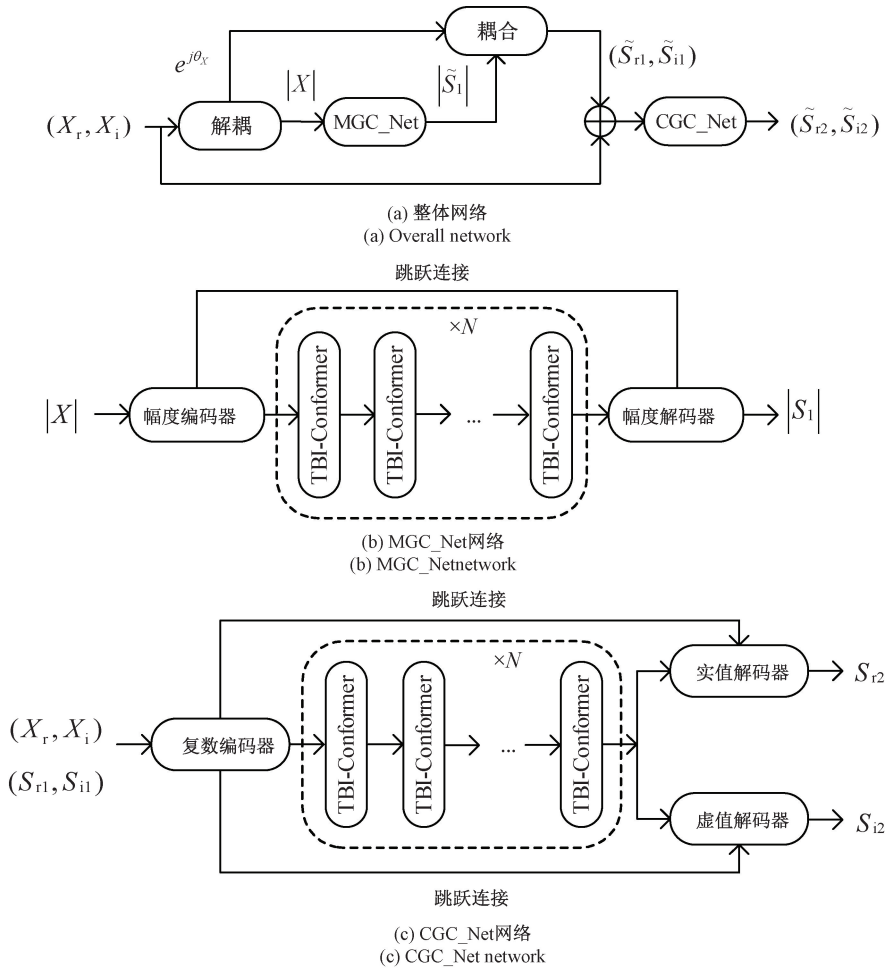


图 1 网络结构图

Fig. 1 Network structure diagram

递哪些信息到下一层,从而有效增强模型处理长序列信息的能力。然而,LSTM 存在结构复杂、计算开销大的缺点,为了解决这些问题,门控线性单元(gated linear unit, GLU)被提出,其不仅继承了 LSTM 灵活控制信息流的优势,而且同时减少了模型的参数量,提升了计算效率。GLU 的表达式为:

$$\boldsymbol{X}_{L+1}=(\boldsymbol{W}_L*\boldsymbol{X}_L+\boldsymbol{B}_L)\otimes\sigma(\boldsymbol{V}_L*\boldsymbol{X}_L+\boldsymbol{C}_L) \tag{4}$$

其中,$\boldsymbol{X}_{L+1}$ 和 $\boldsymbol{X}_L$ 表示第 $L+1$ 层和第 L 层网络的输出,$\boldsymbol{W}_L$、$\boldsymbol{V}_L$ 和 $\boldsymbol{B}_L$、$\boldsymbol{C}_L$ 表示第 L 层的卷积核和偏置,$\otimes$ 和 $*$ 分别表示逐点相乘和卷积,σ 表示 sigmoid 函数,其表达式为:

$$\sigma(x)=\frac{1}{1+e^{x}} \tag{5}$$

膨胀卷积是一种特殊的卷积操作,在保持参数量不变的情况下,通过在卷积核的元素之间插入间隔来控制卷积核的稀疏度,在获得更大范围信息的同时不会有损分辨率。

本文采用二维膨胀卷积,同时在时间和频率两个维度上进行膨胀,旨在扩展时频两个尺度上的感受野,增强模型对时频域中更多上下文信息的捕捉能力。设二维膨胀卷积的表达式为:

$$\boldsymbol{Y}=\boldsymbol{X}*\boldsymbol{K}_d \tag{6}$$

其中,$\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 分别是输入特征和输出特征,$*$ 表示卷积运算,d 表示膨胀率,$\boldsymbol{K}_d$ 表示膨胀卷积的卷积核。

设原始二维普通卷积的卷积核为 $k\times k$,经过膨胀后,卷积核实际大小 $\boldsymbol{K}_d$ 为 $k_d\times k_d$,其中单层膨胀卷积核 k_d 的大小为:

$$k_d=k+(k-1)\times(d-1) \tag{7}$$

基于门控机制以及膨胀卷积的优点,本文提出双通道卷积融合模块,如图 2 所示。在该模块中,为了获取更多丰富的时频信息,使用二维膨胀卷积扩大感受野,并使用门控机制来控制整个网络的信息流。左侧卷积膨胀率 $d=1$,右侧卷积膨胀率 $d=3$,使网络既能获取局部特征,又能获取全局特征。当输入的特征图信息每通过一次膨胀卷积操作时,再通过 sigmoid 函数将特征压缩到(0,1),与另一分支的膨胀卷积相乘,旨在建立网络的短时序

列相关性和长期序列相关性。再将两分支输出的特征图结合，并通过二维逐点卷积进行特征融合，旨在通过降低通道数来减少模型复杂度，同时通过卷积操作保留重要声学信息，进一步加强对关键信息的提取。

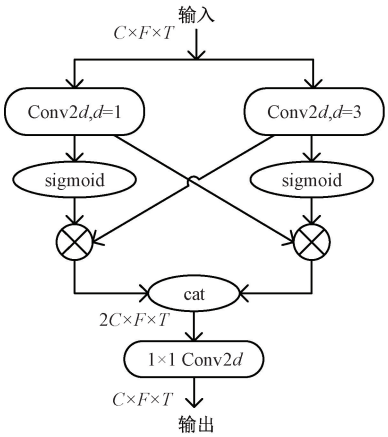


图 2　双通道卷积融合模块

Fig. 2　Dual-channel convolutional fusion module

1.2　TBI-Conformer 结构

在语音识别与语音分离任务中，Conformer 模块展现了其强大的性能和灵活性，它结合了卷积操作与自注意力机制的优点，既能够捕捉局部特征，又能建模全局依赖，使其在处理复杂时序数据时具有更强大的建模能力。然而原始的 Conformer 结构在时间和频率两个维度的关键特征提取方面仍然存在不足，因此，本文对其进行改进，以进一步提升模型对时频特征的捕捉能力，增强模型在处理时频信息上的表现。

TBI-Conformer 结构如图 3 所示，左分支通过 sigmoid 函数使模型专注于对特征的控制，类似于一种选择性过滤机制，从而抑制无关或噪声信息。右分支则直接通过改进型 Conformer 处理后保留了完整的特征信息。左右分支的输出相乘后，最终结果兼具全局依赖和局部依赖的特征表达，既能够保留整体信息，又能灵活捕捉细节，有效提高模型对复杂数据的建模能力。其中，改进型 Conformer 由 2 个前馈层（feed forward，FF）、1 个时间注意（time attention，TA）、1 个频率注意（frequency attention，FA）、1 个膨胀卷积模块（dilated convolution，DC）组成。

1）TA/FA 模块

针对原始自注意力机制在时间、频率维度建模不充分的问题，提出改进型 Conformer 结构，分别从时间维度和频率维度对语音特征进行建模，并且第二阶段 CGC_Net 结构通过堆叠多个复数时间自注意力层和多个复数频率自注意力层，计算序列中每个元素以及其他元素的相似度，捕捉序列内部的依赖关系，从而实现了对时间维度以及频率维度强大的特征提取能力。

在原始的自注意力机制中，将实数输入序列 $\boldsymbol{X}$ 进行线性变换，通过不同的线性层生成 $\boldsymbol{Q}$（query：查询）、$\boldsymbol{K}$（key：

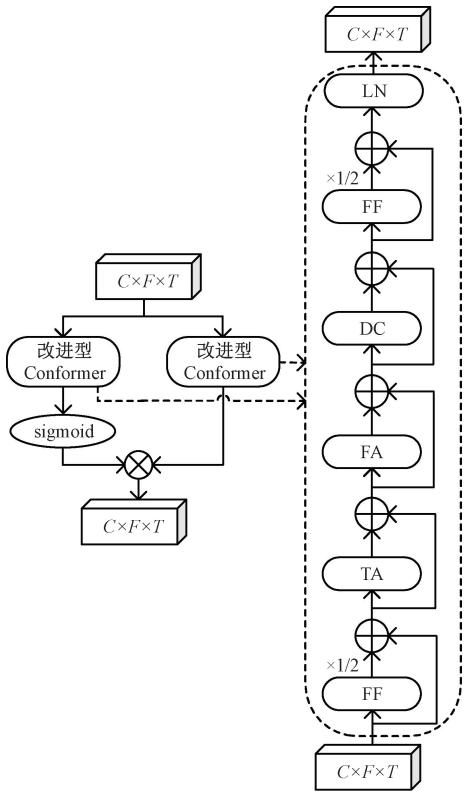


图 3　TBI-Conformer 结构图

Fig. 3　TBI-Conformer structure diagram

键）、$\boldsymbol{V}$（value：值）矩阵：

$$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}^{Q},\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}^{K},\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}^{V} \tag{8}$$

其中，$\boldsymbol{W}^{Q}$、$\boldsymbol{W}^{K}$、$\boldsymbol{W}^{V}$ 是可学习的权重矩阵。

通过 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 的点积计算得到 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 之间的相似度，然后把点积矩阵除以一个缩放因子 $\sqrt{d_k}$，得到注意力得分矩阵，再通过 softmax 函数对注意力得分矩阵进行归一化操作与 V 矩阵相乘，得到加权求和后的输出为：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{9}$$

受到实数自注意力机制的启发，给定复数输入序列为 $\boldsymbol{X}=\boldsymbol{X}_r+\mathrm{j}\boldsymbol{X}_i$，计算复数值 $\boldsymbol{Q}$：

$$\boldsymbol{Q}=\boldsymbol{Q}_r+\mathrm{j}\boldsymbol{Q}_i=(\boldsymbol{X}_r\boldsymbol{W}_r-\boldsymbol{X}_i\boldsymbol{W}_i)+j(\boldsymbol{X}_r\boldsymbol{W}_i+\boldsymbol{X}_r\boldsymbol{W}_i) \tag{10}$$

其中，r 和 i 分别表示复数的实部和虚部，$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 的计算同理。

因此，复数自注意力的计算可以表示为：

$$\begin{aligned}&\text{ComplexAttention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\\&(\text{Attention}(\boldsymbol{Q}_r,\boldsymbol{K}_r,\boldsymbol{V}_r)-\text{Attention}(\boldsymbol{Q}_r,\boldsymbol{K}_i,\boldsymbol{V}_i)-\\&\text{Attention}(\boldsymbol{Q}_i,\boldsymbol{K}_r,\boldsymbol{V}_i)-\text{Attention}(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_r))+\\&\mathrm{j}(\text{Attention}(\boldsymbol{Q}_r,\boldsymbol{K}_r,\boldsymbol{V}_r)-\text{Attention}(\boldsymbol{Q}_r,\boldsymbol{K}_i,\boldsymbol{V}_r)-\\&\text{Attention}(\boldsymbol{Q}_i,\boldsymbol{K}_r,\boldsymbol{V}_r)-\text{Attention}(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i))\end{aligned} \tag{11}$$

在改进型 Conformer 模块中，输入特征先后经过复数时间注意力层和复数频率注意力层，复数多头自注意力模

块并行堆叠 8 个自注意力层,使模型能从不同角度关注输入序列的不同部分,增强序列的表示能力以及鲁棒性。

2)DC 模块

深度可分离卷积在图像处理领域[15-16]等多种任务中运用广泛,并被证实是有效的。它将一个完整的卷积分为两步,包括逐通道卷积和逐点卷积。

由于原始 Conformer 中的一维卷积在特征提取方面的能力有限,采用二维深度可分离卷积来提升特征提取能力。为了使模型在提取特征时能同时获取全局和局部的关键信息,对原始深度可分离卷积模块进行改进,其中逐通道卷积选用二维膨胀卷积(dilated conv2d,D-Conv2d),并设置膨胀因子,M 个膨胀因子分别为 $1,2,4,\cdots,2^{M-1}$。

DC 模块将逐通道卷积分为左右两个分支,左分支膨胀卷积的扩张率取值依次增长,而右分支膨胀卷积的扩张率取值依次递减,并结合了门控机制,如图 4 所示。DC 模块采用不同扩张率的深度可分离卷积,较大的扩张率能有效获得全局关键信息,从而获取语音的长期依赖关系,而较小的扩张率有助于学习局部的序列相关性,从而提取更加精细的特征信息。此外,门控机制进一步优化了信息流的控制,使整个网络在处理复杂序列数据时更加高效。

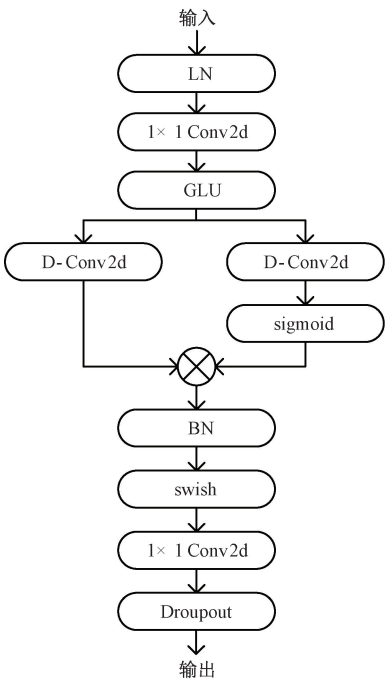


图 4 膨胀卷积模块

Fig. 4 Dilated convolution module

2 实验结果与性能分析

2.1 数据集

在本次实验中,为验证所提模型的有效性,选用公开数据集 VoiceBank+DEMAND 作为实验数据来源。

VoiceBank+DEMAND 数据集主要是由来自 3 个不同国家的 28 名说话者的语音样本组成,采样率为 48 kHz,采样大小为 16 bits。使用的噪声数据集 DEMAND,包含多种真实环境下的噪声录音。训练集是按照 4 种不同信噪比{0,5,10,15} dB 进行合成,一共得到 11 572 组音频文件,测试集是按照信噪比{2.5,7.5,12.5,17.5} dB 进行合成,一共得到 824 组音频文件。在此次实验中,统一将实验所用数据集降采样为 16 kHz。

为提升所提模型在混响条件下的鲁棒性以及泛化能力,并增强模型在不同特征下的学习能力,在训练过程中采用以下数据增强策略:

1)波形丢失:随机将原始波形中的部分片段置零;

2)速度变化:通过速度函数 SOX 来改变原始输入语音的速度;

3)频谱掩蔽:随机遮蔽频谱的某些部分。

2.2 实验设置

本文所用实验环境:处理器为 2.5 GHz 的 Inter Core i5-12400F,GPU 为 GeForce RTX4060Ti,操作系统为 Windows10,Pytorch 版本为 2.8.2,Python 版本为 3.10.13,CUDA 版本为 11.7。

在本次实验中,对于所有模型的设置如下:窗口长度和跳步大小分别为 25 ms 和 6.25 ms,FFT 长度为 512,优化方法选用 Adam,batch_size 设置为 8,采样率固定为 16 kHz,初始学习率为 0.001,学习率的衰减由 ReduceLR0nPlateau 调度器管理,当验证损失在 10 个训练周期内未改善时,学习率将减半,实验还采用了早停策略,以防止模型过拟合,保存效果最佳的模型。

2.3 评价指标

在语音增强领域,用于评价音频文件的性能指标主要分为主观和客观两大类。在本次实验中,主要运用 5 个常用的语音评价指标,包括:语音感知质量(perceptual evaluation of speech quality,PESQ)、短时客观可懂度(short-time objective intelligibility,STOI)、语音信号失真的符合度量(mean opinion score prediction of the speech distortion,CSIG)、背景噪声影响的符合度量(mean opinion score prediction of the intrusiveness of background noise,CBAK)、整体语音质量的复合度量(mean opinion score prediction of the overall processed speech quality,COVL)。其中,PESQ 取值范围为[-0.5,4.5],STOI 取值范围为[0,1],CSIG、CBAK、COVL 取值范围为[1,5]。当评价指标越高,表明增强后的音频质量越好,且越接近纯净语音。

2.4 消融实验

为了验证本文语音增强模型所加的各个模块是否有效,在数据集 VoiceBank+DEMAND 上进行消融实验,分别测试了 MGC_Net 和 CGC_Net 两种单级模型的性能,此外,还测试了在 MGC_Net 和 CGC_Net 将常规二维卷积替换为双通道卷积融合模块后的性能以及将 LSTM 替换为

TBI-Conformer 后的性能。实验结果如表 1 所示。其中，Noisy 为原始带噪语音的测试指标。结果表明：1）将普通二维卷积替换为双通道卷积融合模块后，MGC_Net 和 CGC_Net 的 PESQ 指标均有较大提升，表明模型在语音主观质量上的增强效果明显。同时，3 项综合指标（CSIG、CBAK、COVL）也都有大幅度提升，表明模型在减少信号失真和抑制背景噪声方面也取得了明显进步；2）引入 TBI-Conformer 后，MGC_Net 和 CGC_Net 的各项指标进一步提升，3 项综合指标的提升，也表明了模型在减少信号失真和提升整体语音质量方面表现更加突出；3）将 MGC_Net 和 CGC_Net 连接为两阶段模型 DGC_Net 后，第二阶段的 CGC_Net 以第一阶段 MGC_Net 的预增强频谱作为输入，帮助其更好地捕捉和保留语音轮廓。这使得 CGC_Net 可以更精确地增强语音成分，提高整体语音处理效果。

表 1　在 VoiceBank＋DEMAND 数据集上的消融实验

Table 1　Ablation experiments on the VoiceBank＋DEMAND dataset

模型	双通道融合卷积	TBI-Conformer	PESQ	STOI	CSIG	CBAK	COVL
Noisy	—	—	1.97	0.92	3.35	2.44	2.63
MGC_Net	×	×	2.61	0.92	3.59	2.65	3.09
MGC_Net	√	×	2.71	0.93	3.75	2.76	3.22
MGC_Net	√	√	2.79	0.94	3.84	2.81	3.31
CGC_Net	×	×	2.65	0.93	3.73	2.66	3.13
CGC_Net	√	×	2.74	0.94	3.85	2.82	3.29
CGC_Net	√	√	2.85	0.94	4.12	2.95	3.36
DGC_Net	√	√	2.96	0.95	4.31	3.43	3.51

2.5　VoiceBank＋DEMAND 数据集实验对比

将本文提出的模型与近年来比较有代表性的模型在 VoiceBank＋DEMAND 数据集上进行对比，包括 SEGAN、GCRN、DCCRN、CTS-Net、DeepFilterNet[17]、FullSubNet＋[18]、MetricGAN[19]、S-DCCRN[20]。表 2 中，Noisy 为原始带噪语音的测试指标，加粗字体表示在该类指标中最好的测试结果。

表 2　各模型在 VoiceBank＋DEMAND 数据集的对比实验

Table 2　Comparative experiments of each model on the VoiceBank＋DEMAND dataset

模型	PESQ	STOI	CSIG	CBAK	COVL
Noisy	1.97	0.92	3.35	2.44	2.63
SEGAN	2.16	0.92	3.48	2.94	2.80
GCRN	2.51	0.94	3.71	3.24	3.09
DCCRN	2.62	0.93	3.79	2.46	3.27
DeepFilterNet	2.81	—	—	—	—
FullSubNet＋	2.88	0.94	3.86	3.42	3.57
MetricGAN	2.86	—	—	3.18	3.42
CTS-Net	2.92	—	4.25	**3.46**	**3.59**
S-DCCRN	2.84	0.94	4.03	3.43	2.97
DGC_Net	**2.96**	**0.95**	**4.31**	3.43	3.51

可以观察到，DGC_Net 与其他模型相比在 PESQ、STOI、CSIG、COVL 等指标上均表现突出。1）与各个单阶段模型相比，DGC_Net 在各项指标上均有显著提升，表明双阶段模型有效弥补了单阶段模型结构单一的局限，能够更好地捕捉和利用前一阶段的信息，从而提升语音增强效果和整体性能。2）与 GCRN 以及 DCCRN 模型相比，DGC_Net 不仅具备两阶段的优势，还通过在增强阶段引入 TBI-Conformer，在序列建模能力方面优于传统 LSTM，使得模型更有效地捕捉长时依赖性与时序信息，进一步提升语音增强能力。

2.6　TBI-Conformer 的影响

为探究中间层 TBI-Conformer 分别对 MGC_Net 和 CGC_Net 的影响，设计了各个阶段在 LSTM 以及不同个数 TBI-Conformer 下的对比实验。实验结果如表 3 所示。

可以看出，LSTM 相比 TBI-Conformer 缺少优势，这是因为在处理音频特征时虽然能捕捉长时依赖性，但无法充分关注局部时频特征。MGC_Net 和 CGC_Net 分别在引入 1 个和 2 个 TBI-Conformer 模块时表现最佳，进一步

表 3 在 VoiceBank+DEMAND 数据集下引入不同中间层对模型的影响

Table 3 The impact of introducing different middle layers on the model under the VoiceBank+DEMAND dataset

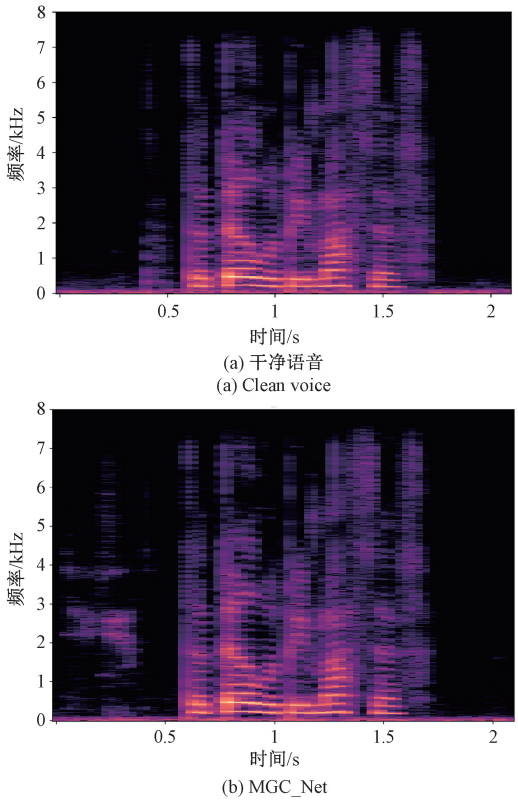
模型	PESQ	STOI	CSIG	CBAK	COVL
MGC_Net+实数 LSTM	2.71	0.93	3.75	2.76	3.22
MGC_Net+实数 TBI-Conformer(1 个)	2.79	0.94	3.84	2.81	3.31
MGC_Net+实数 TBI-Conformer(2 个)	2.76	0.94	3.81	2.74	3.26
CGC_Net+复数 LSTM	2.74	0.94	3.85	2.82	3.29
CGC_Net+复数 TBI-Conformer(1 个)	2.80	0.94	3.94	2.84	3.32
CGC_Net+复数 TBI-Conformer(2 个)	2.85	0.94	4.12	2.95	3.36
CGC_Net+复数 TBI-Conformer(3 个)	2.77	0.94	3.88	2.82	3.32

增加反而会导致性能下降。这是因为随着 TBI-Conformer 数量的增加,模型参数增加、复杂度增加,模型参数和复杂度加剧,导致信息提取的效率降低,模型难以进一步提升,出现性能瓶颈甚至过拟合的现象。

2.7 可视化分析

为直观展示单阶段 MGC_Net、CGC_Net 模型和两阶段 DGC_Net 模型在增强任务中的性能表现,从 VoiceBank+DEMAND 数据集中随机选取一条语音进行语谱图可视化分析,如图 5 所示。图 5(a)~(d)分别对应干净语音、单阶段 MGC_Net、单阶段 CGC_Net 和两阶段 DGC_Net 的语谱图。

从以下 4 个语谱图可以得出,单阶段模型 MGC_Net 和 CGC_Net 相较于干净语音,有较多的噪声干扰。而两阶段模型 DGC_Net 更接近干净语音,保留了一些干净信号的特征,表现最优,能够有效减少噪声并保留更多语音信号。



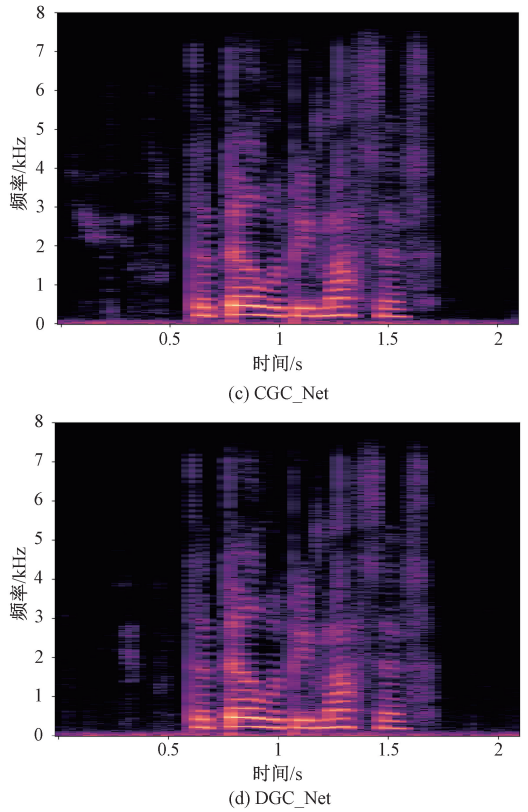


图 5 可视化语谱图

Fig. 5 Visualize the spectral map

3 结 论

本文提出了一种融合双通道卷积和改进型 Conformer 的双阶段语音增强算法 DGC_Net。该网络分为两个阶段对语音信号进行增强,在第 1 阶段 MGC_Net 粗略估计目标语音的幅度,第 2 阶段接收来自第一阶段的粗谱以及第 1 阶段可能丢失的特征信息,通过 CGC_Net 网络抑制残余噪声和干扰语音,得到增强以后的完整频谱图。通过实验证明,两阶段网络较好地弥补了单一网络在性能上的

不足。此外，DGC_Net 通过双通道卷积融合模块和改进型 Conformer 进一步加强了模型对关键特征的提取能力。在未来的工作中，将把模型扩展到其他领域，如语音分离和说话人确认。

参考文献

[1] YANG Y, LIU P P, ZHOU H L, et al. A speech enhancement algorithm combining spectral subtraction and wavelet transform [C]. 2021 IEEE 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE). Piscataway, NJ: IEEE Press, 2021: 268-273.

[2] 杨诗童，杨飞. 改进 TF-GSC 和改进后置滤波语音增强算法[J]. 电子测量技术，2024，46(17)：118-124.
YANG SH T, YANG F. Improved TF-GSC and improved post filter speech enhancement algorithm[J]. Electronic Measurement Technology, 2024, 46(17): 118-124.

[3] MAHMMOD B M, RAMLI A R, ABDULHUSSIAN S H, et al. Low-distortion MMSE speech enhancement estimator based on Laplacian prior[J]. IEEE Access, 2017, 5: 9866-9881.

[4] KISHORE V, TIWARI N, PARAMASIVAM P. Improved speech enhancement using TCN with multiple encoder-decoder layers [C]. 2020 Interspeech, IEEE Press. 2020: 4531-4535.

[5] NIAN ZH X, TU Y H, DU J, et al. A progressive learning approach to adaptive noise and speech estimation for speech enhancement and noisy speech recognition [C]. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE Press, 2021: 6913-6917.

[6] PASCUAL S, BONAFONTE A, SERRA J. SEGAN: Speech enhancement generative adversarial network[C]. 2017 Interspeech, IEEE Press, 2017: 3642-3646.

[7] 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, 2019, 28: 380-390.

[8] HU Y X, LIU Y, LYU SH B, et al. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement[C]. 2020 Interspeech, IEEE Press, 2020: 2472-2476.

[9] 罗庆予，张天骐，方蓉，等. 联合频谱映射与掩蔽估计的协作式语音增强方法[J]. 电子测量与仪器学报，2023，37(10)：14-23.
LUO Q Y, ZHANG T Q, FANG R, et al. Collaborative speech enhancement method combining spectral mapping and masking estimation[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(10): 14-23.

[10] 高盛祥，莫尚斌，余正涛，等. 基于多维度注意力机制和复数 Conformer 的单通道语音增强方法[J]. 重庆邮电大学学报(自然科学版)，2024，36(2)：393-403.
GAO SH X, MO SH B, YU ZH T, et al. Monaural speech enhancement method based of multi-dimensional attention mechanism and complex Conformer[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2024, 36(2): 393-403.

[11] JIANG R, LIU W. Single image deraining via two stages: bilateral learning and feature decoupling[C]. 2022 4th International Conference on Robotics and Computer Vision (ICRCV). Piscataway, NJ: IEEE Press, 2022: 186-192.

[12] 刘升，古丽巴哈尔·托乎提，补生来，等. 融合改进 YOLOv7 与 UNet 的编码点定位方法[J]. 国外电子测量技术，2024，43(5)：9-17.
LIU SH, GULBAHAR T, BU SH L, et al. Fusion of improved YOLOv7 and UNet codepoint localization methods [J]. Foreign Electronic Measurement Technology, 2024, 43(5): 9-17.

[13] LI A D, LIU W ZH, ZHENG CH SH, et al. Two heads are better than one: a two-stage complex spectral mapping approach for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1829-1843.

[14] 张小恒，李勇明，王品. 双阶段帕金森病语音聚类包络卷积稀疏迁移学习算法[J]. 仪器仪表学报，2022，43(11)：151-161.
ZHANG X H, LI Y M, WANG P, et al. Two-stage PD speech clustering envelope and convolution sparse transfer learning algorithm[J]. Chinese Journal of Scientific Instrument, 2022, 43(11): 151-161.

[15] ZHAO Q F, ZHONG L F, XIAO J H, et al. Efficient multi-organ segmentation from 3D abdominal CT images with lightweight network and knowledge distillation [J]. IEEE Transactions on Medical Imaging, 2023, 42(9): 2513-2523.

[16] LIU Y D, WAN H M. DDC-BlockV2: A more accurate and smaller general purpose lightweight module [C]. 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology(AINIT), IEEE Press, 2024: 1894-1898.

[17] SCHROTER H, ESCALANTE B A N, ROSENKRANZ T, et al. DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering [C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE Press, 2022: 7407-7411.

[18] CHEN J, WANG Z L, TUO D Y, et al. Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement [C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE Press, 2022: 7857-7861.

[19] 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, 2019: 2031-2041.

[20] LYU S B, FU Y H, XING M T, et al. S-dccrn: Super wide band dccrn with learnable complex feature for speech enhancement [C]. IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), IEEE Press, 2022: 7767-7771.

作者简介

徐佳瑜,硕士研究生,主要研究方向为语音增强。

E-mail:1023107894@qq.com

郑展恒(通信作者),高级实验师,硕士生导师,主要研究方向为信号处理。

E-mail:glzzh@guet.edu.cn

曾庆宁,教授,博士,主要研究方向为语音信号处理。

E-mail:qingningzeng@126.com

王健,副教授,硕士生导师,主要研究方向为智能信号处理。

E-mail:wangjian@guet.edu.cn