ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 3,May 2026,pp. 814－824
DOI:10. 16337/j. 1004-9037. 2026. 03. 014

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于门控混合膨胀卷积的轻量级语音增强

孙林慧，魏鹏滨，王春艳，叶　蕾，邵　曦

（南京邮电大学通信与信息工程学院，南京 210003）

摘　要：针对主流语音增强模型存在的参数量膨胀以及计算复杂度剧增的问题，本文提出了一种基于门控混合膨胀卷积的轻量级语音增强网络。首先，设计了一种门控混合膨胀卷积模块，该模块结合门控线性单元与混合膨胀卷积，实现对语音信号的多尺度特征提取以及对噪声敏感区域的精准抑制，从而在有效保留语音长短时特征的同时，增强模型的鲁棒性；其次，设计了一种层级通道注意力模块，通过层级式特征融合，在低参数量条件下提升对通道维度中语音特征相关性的捕捉能力。在VoiceBank＋DEMAND数据集上进行的实验结果表明，本文模型以仅0.41 M的参数量，在语音质量感知评价（Perceptual evaluation of speech quality，PESQ）、短时客观可懂度（Short-time objective intelligibility，STOI）、倒谱信噪比（Cepstral signal-to-noise ratio，CSIG）、倒谱背景噪声（Cepstral background noise，CBAK）、倒谱总体响度（Cepstral overall loudness，COVL）五项指标上表现良好，实现了模型轻量化与良好性能的有机结合。

关键词：语音增强；门控混合膨胀卷积；多尺度特征融合；通道注意力；低参数量

中图分类号：TN912.3　　**文献标志码：**A

引用格式：孙林慧，魏鹏滨，王春艳，等. 基于门控混合膨胀卷积的轻量级语音增强[J]. 数据采集与处理，2026，41(3)：814-824. SUN Linhui，WEI Pengbin，WANG Chunyan，et al. Lightweight speech enhancement based on gated hybrid dilated convolution[J]. Journal of Data Acquisition and Processing，2026，41(3)：814-824.

引　　言

语音增强是语音信号处理领域的一项关键技术，其核心目标在于提高被各种噪声源干扰的语音信号的质量和可懂性[1]。作为语音处理系统的关键预处理环节，它在自动语音识别[2]、语音编码[3]和人机交互[4]等多个领域被广泛使用，发挥着至关重要的作用。随着移动设备和物联网的普及，人们对噪声环境下的语音质量提出了更高要求。然而，现实场景中的复杂声学环境可能包括非平稳噪声、混响和多人说话等干扰，这使得传统语音增强方法面临巨大挑战。近年来，深度学习技术大幅提升了语音增强的性能，然而现有方法在计算效率、模型轻量化和噪声适应性方面仍存在许多不足。特别是在资源受限的边缘设备上，如何在保证实时性的同时维持高语音质量，成为当前语音增强研究的核心问题之一。

针对上述所涉及到的问题，本文提出一种轻量级语音增强网络，其核心模块包括门控混合膨胀卷积模块、幻影卷积模块、层级通道注意力模块和频率变换块。该网络以仅0.41M的模型参数量，取得了良好的语音增强效果，为语音增强领域提供了一种高效、轻量化的设计方案。本文的主要工作如下：

（1）设计了门控混合膨胀卷积（Gated hybrid dilated convolution，GHDC）模块，该模块将门控线性单元（Gated linear unit，GLU）与混合膨胀卷积（Hybrid dilated convolution，HDC）相结合，实现多尺度特征提取的同时对噪声敏感区域进行动态抑制。该模块能够有效保留语音长短时特征，从而增强模型鲁棒性。

基金项目：江苏省科技重大专项(BG2024027)；国家自然科学基金(61901227)。
收稿日期：2025-05-21；**修订日期：**2025-08-30

(2) 设计了一种层级通道注意力(Hierarchical channel attention,HCA)模块,通过层级式特征融合,结合平均池化与最大池化特征,动态强化时频域中语音的关键分布,从而提升模型对通道维度中语音特征相关性的捕捉能力。

(3) 引入幻影卷积(Ghost convolution,GC)模块,以低成本线性变换替代冗余的卷积操作,从而降低模型的计算复杂度。

(4) 引入频域变换块(Frequency transformation block,FTB),将频率全连接子模块与时频注意力(Time-frequency attention,T-F attention)子模块相结合,隐式学习谐波相关性,从而优化全局频域特征表示。

1 相关工作

传统的语音增强方法包括谱减法[5]、维纳滤波法[6]和基于统计的方法[7]等。其中谱减法和维纳滤波法都基于平稳噪声假设,因而对非平稳噪声的抑制效果较为有限。基于统计模型的方法通过概率框架部分缓解了上述问题,但对噪声分布的强假设仍限制了其在非平稳噪声环境下的泛化性能。

随着深度学习技术的突破,基于深度神经网络的语音增强方法凭借强大的特征学习能力迅速成为研究主流。常用的深度神经网络有循环神经网络(Recurrent neural network, RNN)[8]和卷积神经网络(Convolutional neural network, CNN)[9]。RNN通过隐状态传递捕获语音的长时依赖性;CNN用局部感受野与权重共享机制高效提取时频特征,但单一尺度的卷积核难以覆盖语音信号的全局特性;卷积循环网络(Convolutional recurrent network, CRN)[10]结合了CNN与RNN的优势,通过编码器-解码器架构实现多分辨率特征融合,但在深层网络中会面临梯度消失与计算效率低下的问题。为了更有效地进行复数域上的运算,Hu等[11]在2020年将卷积循环网络与长短时记忆网络(Long short term memory, LSTM)相结合,同时引入卷积编解码结,提出了深度复杂卷积递归网络(Deep complex convolution recurrent network,DCCRN)。Xiang等[12]提出FullSubNet网络,该网络融合了全频带模型与子频带模型,既可以捕获全局上下文信息,又保留了对信号平稳性进行建模和关注局部频谱模式的能力。在双路径卷积循环网络模型(Dual path convolutional recurrent network, DPCRN)的基础上,胡沁雯等[13]提出了一种轻量级的全频带语音增强网络模型,该模型在只有0.89M参数的条件下,获得了与其他高性能全频带模型类似的语音增强效果。

为了解决传统卷积网络长时依赖性建模能力不足的问题,注意力机制开始被广泛应用到语音增强领域。张玥等[14]将时频注意力机制与U-Net相结合,引导模型关注语音的低频部分特征,并且重构高频成分。近年来,通道注意力机制由于对关键通道特征具有出色的表达能力,逐渐受到研究者们的青睐。姚瑶等[15]提出了一种多维注意力机制,通过将通道注意力和全局、局部时间注意力进行级联,充分挖掘深度神经网络各通道间语音特征的长短时相关性。洪依等[16]在全卷积时域音频分离网络的基础上,提出了一种基于超轻量通道注意力的端对端语音增强网络。尽管这些研究同样考虑到了模型的轻量化设计,但是为了获取相对理想的语音增强效果,在引入复杂结构和通道注意力机制后,这些模型依然面临着参数量膨胀以及计算复杂度剧增的问题。

2 基于门控混合膨胀卷积的轻量级语音增强

2.1 整体网络结构

本文所提出的整体网络结构如图1所示。该网络首先对输入的含噪语音进行短时傅里叶变换处理,将经过幂律变换后的幅度谱特征作为网络的输入特征;接着,输入特征通过GHDC模块处理,该模块采用双分支结构,分别提取语音的短时瞬态特征与长时全局特征,并结合残差连接来融合多尺度信

息;随后,将提取到的多尺度特征输入到幻影卷积层中,通过线性扩展优化特征冗余,从而降低计算复杂度;紧接着将优化后的特征输入到HCA模块中,利用双池化特征交互动态强化关键通道权重,抑制噪声干扰;再经过门控线性单元层增加网络的非线性学习能力,并通过一个卷积块进行降维,随后输入到FTB中,频域变换块可以隐式地学习谐波之间的相关性,捕获沿频率轴的全局相关性;最后将经过频域变换块后的幅度谱特征与原始信号的相位谱逐位相乘,再经由逆幂律变换和短时傅里叶逆变换得到增强后的语音信号。

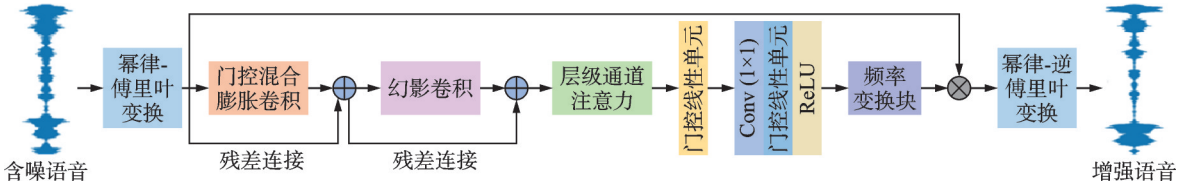

图1 整体网络结构

Fig.1 Overall network structure

整个网络在均方误差(Mean squared error,MSE)损失函数的约束下进行训练,其计算方式为

$$\mathcal{L}_{\mathrm{MSE}}=\frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{x}_i)^2 \tag{1}$$

式中:N为样本的总数量,x_i表示第i个样本的真实值,$\hat{x}_i$表示第i个样本的预测值。

2.2 门控混合膨胀卷积模块

HDC可解决单一膨胀卷积所引起的网格化效应,该效应是指当连续使用相同或具有公约数的膨胀率时,卷积核覆盖区域会呈现离散间隔,导致相邻像素间的信息在膨胀卷积过程中丢失的现象。HDC通过并行或串联使用无公约数的膨胀率序列,扩大网络感受野,达到近似覆盖输入特征图的效果,从而避免信息丢失。

本文将HDC与GLU结合设计了一种GHDC。在GHDC模块中,HDC的作用是在不增加参数量的情况下增大网络的感受野,实现多尺度特征提取,全面地捕获语音信号的短时瞬态特征与长时稳态特征。GLU则可以对HDC各膨胀率分支输出特征进行跨通道权重分配,并动态抑制噪声敏感区域,从而增强模块的非线性表达能力。HDC模块具体结构如图2所示,该模块主要由两个并行的膨胀卷积分支组成,每个分支各包括3个卷积核大小为5×5的膨胀卷积层。上侧分支采用膨胀率递减设计,每个膨胀卷积层的膨胀率以5、2、1的顺序依次减小,通过逐步缩小感受野聚焦于语音短时特征;而下侧分支则采用膨胀率递增设计,每个膨胀卷积层的膨胀率以1、2、5的顺序依次增大,通过逐步扩大感受野以建模语音长时依赖性。每个膨胀卷积层输出经Sigmoid激活后,再与两侧同层特征图通过逐位相乘强化共性区域,达到精确抑制噪声敏感区域的效果。然后将两侧分支的特征在通道维度上进行拼接,从而

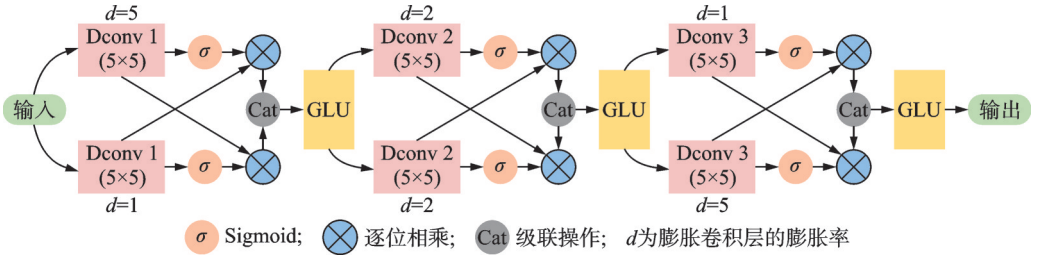

图2 门控混合膨胀卷积模块

Fig.2 Gated hybrid dilated convolution module

保留原始输入的多尺度信息,提供更全面的互补性特征表示,最后送入GLU后作为两侧分支中下一层膨胀卷积层的输入。

2.3 幻影卷积模块

幻影卷积技术由Han等[17]提出,其核心思想是通过高效的线性变换对原始特征图进行扩展,生成一系列能够充分挖掘特征内在信息的"幻影"特征图,从而减少冗余特征,显著降低模型复杂度。

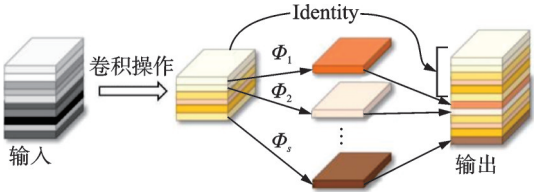

图3 幻影卷积模块

Fig.3 Ghost convolution module

如图3所示,幻影卷积模块一共有3个步骤:首先对原始特征图进行少量的常规卷积操作,生成M个内在特征图;然后对这些内在特征图施加一系列低成本的线性变换,生成S个额外的幻影特征图;最后将得到的幻影特征图与原始特征图进行拼接,从而生成完整的输出特征图。该模块的具体运算过程为

$$Y = X \cdot f \tag{2}$$

$$y_{ij} = \Phi_{ij}(y_i) \quad \forall i = 1, 2, \cdots, M;\ j = 1, 2, \cdots, S \tag{3}$$

$$X_o = [Y, y_{11}, y_{12}, \cdots, y_{MS}] \tag{4}$$

式中:X为输入的原始特征图,X_o为输出特征图,f为卷积过程中所使用的滤波器,Y为原始特征图通过卷积操作所生成的内在特征图,y_i为Y中的第i个内在特征图,y_{ij}为y_i的第j个幻影特征图,Φ_{ij}为用于生成幻影特征图y_{ij}的线性运算。

2.4 层级通道注意力模块

语音频谱蕴含丰富的时频特征信息,但其能量分布在不同频域呈现显著的非均匀特性:低频区域通常能量集中,而高频区域则相对分散。为了更有效地利用这些信息,需要对不同的频率区域给予不同的关注度。通道注意力(Channel attention,CA)机制为解决这一问题提供了一种有效的途径,其能够识别并强化时频域中语音特征的关键分布,同时减少输入特征图中不相关区域的干扰。

传统CA方法通常采用双分支并行架构,分别对输入特征进行平均池化和最大池化处理,这种方式虽然有效,但往往会损失一部分输入的特征信息。为了解决该问题,本文对传统的CA方法进行了改进,设计了一种HCA。

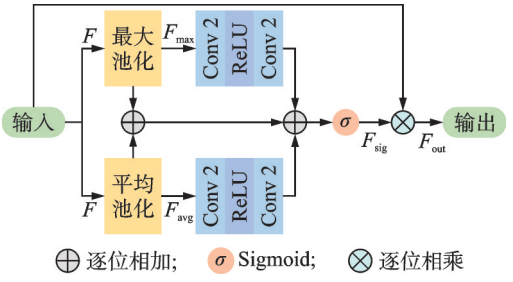

图4 层级通道注意力模块

Fig.4 HCA module

HCA的具体结构如图4所示。在原有的通道注意力机制的基础上,HCA将平均池化特征和最大池化特征逐位相加,获得初步组合特征;然后将该组合特征与两个独立池化分支的输出进行逐位相加,从而得到融合特征,输入到Sigmoid激活函数层,对每个通道的重要性进行建模。最后HCA将通过Sigmoid层后的融合特征与一开始输入的原始特征进行逐位相乘,最终得到整个模块的输出特征。这样的层级式特征融合方式不仅可以保留原始输入的全部信息,还可通过多级特征交互增强表征的完备性和准确性,从而提高网络的表征能力。HCA的具体运算过程为

$$F_{mix} = F_{avg} + F_{max} \tag{5}$$

$$F_{sig} = \sigma(\mathrm{conv}\, F_{avg} + \mathrm{conv}\, F_{max} + F_{mix}) \tag{6}$$

$$F_{out} = F \otimes F_{sig} \tag{7}$$

式中：F_{avg}和F_{max}分别表示平均池化特征和最大池化特征，σ表示Sigmoid激活函数运算，conv表示经过卷积块的运算，F_{out}表示HCA模块的输出特征，“⊗”表示逐位相乘。

2.5 频域变换块

本文基线模型FTB来自文献[18]，原模型采用双流网络架构，分别通过幅值流和相位流进行对应的频谱分量预测，本文所引入的FTB取自原模型中的幅值流分支。FTB的结构如图5所示，图中变量含义参见文献[18]。

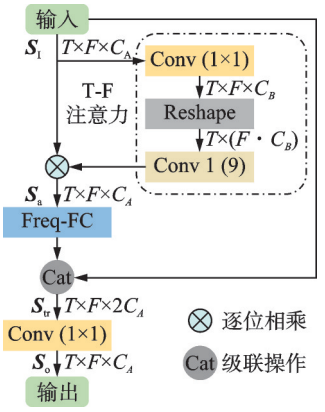

图5 频域变换块

Fig.5 Frequency transformation block

图5包含两个关键子模块：T-F注意力机制模块和频率全连接(Frequency fully connected，Freq-FC)模块，具体运算过程为

$$\boldsymbol{S}_a=\boldsymbol{f}_{attn}(\boldsymbol{S}_I) \tag{8}$$

$$\boldsymbol{S}_{tr}=\boldsymbol{f}_{req}(\boldsymbol{S}_a) \tag{9}$$

$$\boldsymbol{S}_o=\mathrm{conv}(\mathrm{concat}(\boldsymbol{S}_{tr}+\boldsymbol{S}_I)) \tag{10}$$

$$\boldsymbol{S}_{tr}(t_0)=\boldsymbol{X}_{tr}\cdot\boldsymbol{S}_a(t_0) \tag{11}$$

式中：$\boldsymbol{S}_I$和$\boldsymbol{S}_o$分别表示FTB的输入特征和输出特征；$\boldsymbol{f}_{attn}(\cdot)$表示T-F注意力机制模块对应的注意力映射函数；$\boldsymbol{S}_a$表示T-F注意力机制模块的输出特征；$\boldsymbol{S}_{tr}$表示Freq-FC模块的输出特征；$\boldsymbol{f}_{req}(\cdot)$表示频率全连接变换函数；concat(·)表示特征拼接操作；$\boldsymbol{S}_{tr}(t_0)$表示时间步长t_0时转换后的特征切片，与$\boldsymbol{S}_a(t_0)$具有相同的维度；$\boldsymbol{X}_{tr}\in\mathbf{R}^{F\times F}$表示可训练的频率转换矩阵；$\boldsymbol{S}_a(t_0)\in\mathbf{R}^{F\times C_A}(t_0\in\{0,1,\cdots,T-1\})$表示每个时间步长的特征切片。

3 实验结果与分析

3.1 数据集与实验设置

实验采用的数据集是VoiceBank+DEMAND[19]，该数据集语言为英语，其在语音识别和语音合成研究等任务中有广泛使用。该数据集的训练集有男女各14位说话人，包含11 572条干净语音数据。训练集添加了10种噪声，混合噪声的分贝等级分别为0、5、10和15 dB，其中8种取自Demand数据集，另外2种是人造噪声。测试集有男女各1位说话人，包含824条干净语音数据。测试集添加了Demand数据集中的5种噪声，均未在训练集噪声中出现，混合噪声的分贝等级分别为2.5、7.5、12.5和17.5 dB。

本实验采用的处理器为P100 GPU，内存为16 GB。编码语言为Python，软件环境为Python 3.10，深度学习框架主要使用Torch 2.1.2，语音处理库为Librosa 0.10.1。实验所用数据采样率为16 kHz，通过400点短时傅里叶变换(Short-time Fourier transform，STFT)将语音信号转换到频域，采用窗长为400的汉宁窗，帧移为100，具有25%的重叠。整个训练的批次是50个，批量大小是8。优化器采用Adam优化器，学习率设置为0.000 5。本文采用语音质量感知评价(Perceptual evaluation of speech quality，PESQ)、短时客观可懂度(Short-time objective intelligibility，STOI)、倒谱信噪比(Cepstral signal-to-noise ratio，CSIG)、倒谱背景噪声(Cepstral Background noise，CBAK)和倒谱总体响度(Cepstral overall loudness，COVL)共5种语音评价指标对网络的性能进行评估。

整个网络的参数设置如表1所示，其中所有模块的输入维度和输出维度都按照$C\times T\times F$的格式，这3个参数分别表示特征图通道数、时间帧和频率维度。表中超参数k、s、p和d分别表示卷积核大小、步长、输入特征图的填充量以及膨胀卷积层的膨胀率，其中p和d在表中按照$m\times n$的格式，m和n分别表示GHDC模块的上下两条支路分别对应的输入特征图的填充量和膨胀卷积层的膨胀率。

表1 模型网络参数

Table 1 Network parameters of model

模块	输入维度	输出维度	超参数
GHDC	[1,361,201]	[16,361,201]	Dcov 1(×2):k=5,p=10×2,d=5×1
	[16,361,201]	[64,361,201]	Dcov 2(×2):k=5,p=4×4,d=2×2
	[64,361,201]	[64,361,201]	Dcov 3(×2):k=5,p=2×10,d=1×5
Ghost-conv	[64,361,201]	[32,361,201]	Conv 2:k=1,s=1
	[32,361,201]	[32,361,201]	Dw-conv 2:k=3,p=1,s=1
	[32,361,201]	[64,361,201]	
HCA	[64,361,201]	[64,1,1]	卷积块:k=1
	[64,1,1]	[64,1,1]	
	[64,1,1]	[64,361,201]	
GLU	[64,361,201]	[32,361,201]	—
卷积块	[32,361,201]	[1,361,201]	Conv 2:k=1
FTB	[1,361,201]	[5,361,201]	—

3.2 混合膨胀卷积层通道数对网络性能的影响

混合膨胀卷积层的通道数是网络结构的一个重要超参数。对于混合膨胀卷积层而言,更多的通道数意味着网络能够提取更多不同尺度的信息特征,从而增强网络的特征表示能力。然而,通道数的增加会导致模型参数量的增加和计算复杂度的提升,因此,在设计网络结构时,需要在性能提升和计算效率之间做出权衡,设置合适的通道数。

为了获取最佳的通道数,本文针对GHDC模块设置了4组不同的通道数进行比较,具体实验结果如表2所示。表中MAC(Multiply-accumulate operation)表示乘加运算次数。在前3组实验中,当Dcov 3输出通道数由16增至32,再增至64时,语音评价的各个指标均逐步提升,说明在这3组实验中,适当增加通道数强化了增强网络的特征表示能力,但当Dcov 3输出通道数为从64增加至128时,所得数据结果的各个指标反而出现下降,这可能是由于通道数过多后,会在训练过程中学到大量冗余特征,导致泛化能力下降。

表2 GHDC设置不同通道数的网络增强性能对比

Table 2 Comparison of speech enhancement performance for networks with different channel numbers in GHDC

序号	通道数	PESQ	STOI	CSIG	CBAK	COVL	MACs/G	Number of parameter/10^6
1	Dcov 1(×2):1,4; Dcov 2(×2):4,16; Dcov 3(×2):16,16	2.57	0.91	3.94	3.04	3.27	**0.23**	0.16
2	Dcov 1(×2):1,8; Dcov 2(×2):8,32; Dcov 3(×2):32,32	2.61	0.92	4.00	3.12	3.31	**0.76**	0.21
3	Dcov 1(×2):1,16; Dcov 2(×2):16,64; Dcov 3(×2):64,64	**2.68**	**0.93**	**4.10**	**3.21**	**3.41**	**2.97**	0.41
4	Dcov 1(×2):1,32; Dcov 2(×2):32,128; Dcov 3(×2):128,128	2.34	0.88	1.25	2.36	1.35	**11.84**	1.20

根据实验结果，当GHDC模块设置为：Dcov 1输入通道数为1，输出通道数为16；Dcov 2输入通道数为16，输出通道数为64；Dcov 3输入通道数为64，输出通道数为64时，网络取得最佳性能，且模型参数量不大，符合轻量化要求。

3.3 消融实验

为了对各个部分的有效性进行探究，本文以FTB作为基线模型设计了消融实验，结果如表3所示，其中，标识“√”表示在网络中使用了该模块。

表3 消融实验结果

Table 3 Results of ablation experiment

FTB	GHDC	HCA	Ghosts	PESQ	STOI	CSIG	CBAK	COVL
√				2.29	0.90	3.59	2.74	2.92
√	√			2.58	0.91	4.01	3.07	3.30
√	√	√		2.62	0.91	4.03	2.98	3.32
√	√	√	√	**2.68**	**0.93**	**4.10**	**3.21**	**3.41**

从表3可见，当加入GHDC模块后，PESQ、STOI、CSIG、CBAK和COVL分别提升了0.29、0.01、0.42、0.33和0.38，说明GHDC模块可以有效提升网络性能。在加入HCA模块后，PESQ、CSIG和COVL分别提升了0.04、0.02和0.02，这说明HCA模块可有效地指导网络更专注于对当前任务有益的关键通道，从而提升网络性能；引入幻影卷积后，PESQ、STOI、CSIG、CBAK和COVL分别提升了0.06、0.02、0.07、0.23和0.09，这证明幻影卷积模块不仅有助于网络的轻量化设计，对语音质量的提升也有一定的贡献。与基线模型相比，完整模型的PESQ提升了0.39，STOI提升了0.03，CSIG、CBAK和COVL分别提升了0.51、0.47和0.49，消融实验结果证明了每个模块的有效性。

为了进一步直观地对比关键模块的语音增强效果，在测试集中随机抽取一条语音，给出该语音在消融实验各阶段的语谱图，具体如图6所示。图6(a)和(b)分别表示干净语音信号和含噪语音信号的语

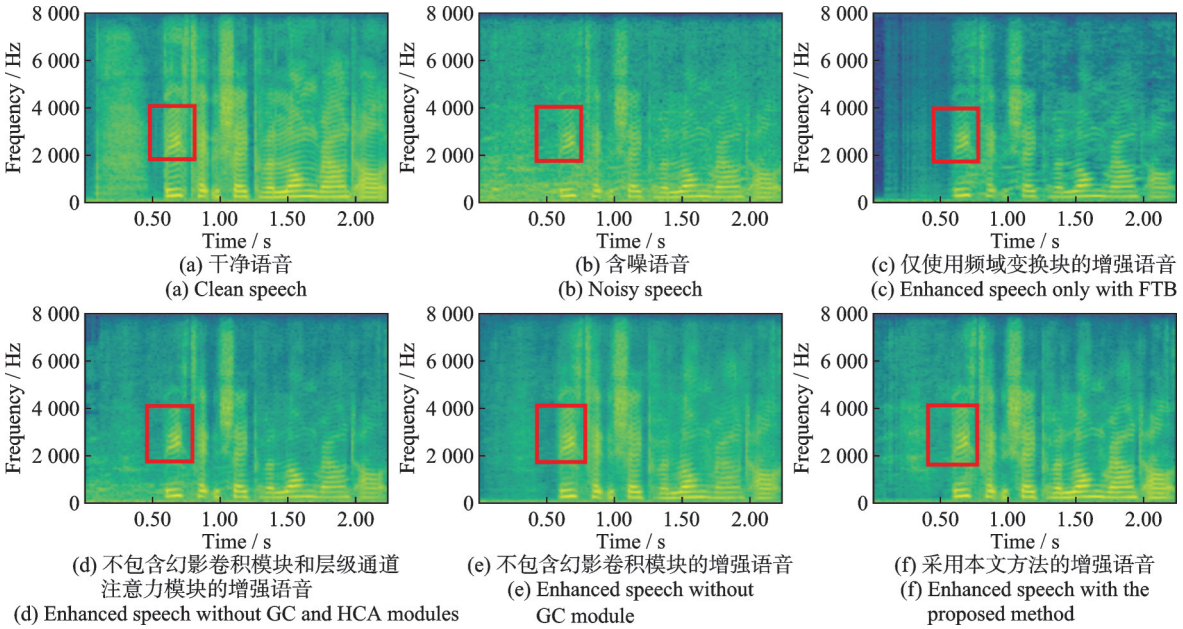

图6 语音信号的语谱图

Fig.6 Spectrograms of speech signal

谱图,图6(c)、(d)、(e)和(f)分别表示经过仅保留FTB模块的模型、去掉幻影卷积模块和HCA模块的模型、去掉幻影卷积模块的模型和完整模型增强后重建的语音信号语谱图。通过对比分析,可以发现所提模型的GHDC模块、HCA模块和幻影卷积模块都有助于提升重建语音信号的质量,采用完整模型进行增强能够消除大部分噪声干扰,此时重建语音效果最好。

3.4 不同网络模型的对比实验

为了展现本文所提出网络模型的优劣,本节将所提出网络模型与其他优秀语音增强网络模型进行对比。选用Wave U-Net[20]、CRN[10]、GCRN[21]、DCCRN[11]、S-DCCRN[22]、Fast FullSubNet[23]共6种模型作为对比模型,具体实验结果如表4所示。

表4 不同网络模型对比结果

Table 4 Comparison among different network models

模型	PESQ	STOI	CSIG	CBAK	COVL	Number of parameter/10^6
Wave U-Net[20]	2.40	—	3.52	3.24	2.96	38.10
CRN[10]	2.59	—	3.78	**3.26**	3.27	6.10
GCRN[21]	2.50	**0.94**	3.66	3.17	3.09	9.80
DCCRN[11]	2.68	0.94	3.88	3.18	3.27	3.70
Fast FullSubNet[23]	2.81	0.94	3.86	3.42	**3.62**	8.67
S-DCCRN[22]	**2.84**	0.94	4.03	3.43	2.97	2.34
本文模型	2.68	0.93	**4.10**	3.21	3.41	**0.41**

由表4中的实验结果可知,与Wave U-Net模型相比,本文模型在PESQ、CSIG、CBAK和COVL这4个性能指标上有很大的提升,且模型参数量大幅度减少。具体而言,Wave U-Net模型参数量是38.10×10^6,而本文模型的参数量仅为0.41×10^6。与CRN相比,本文提出的模型在CBAK指标上稍微逊色,但其他指标都提升较大,且模型轻量化方面也占有明显优势,模型参数量大约为CRN的1/15。与GCRN相比,本文模型只有STOI低了0.01,其他指标均有很大提升,而且参数量比GCRN模型降低了9.49×10^6。虽然本文模型与DCCRN模型在PESQ指标上大小相同,STOI略低0.01,但在CSIG、CBAK和COVL指标上分别高出0.22、0.03和0.14,整体语音增强效果更优,且参数量减少了3.29×10^6。

与Fast FullSubNet和S-DCCRN这两个模型相比,在PESQ、STOI和CBAK指标方面本文模型略差于这两个模型,但CSIG指标方面本文模型优于这两个模型,在COVL指标方面本文模型优于S-DCCRN模型。另外,Fast FullSubNet模型的参数量为8.67×10^6,S-DCCRN模型的参数量为2.34×10^6,远远高过本文所提出模型的参数量。因而在性能相近的情况下,本文所提出模型在参数量上更具优势。

总之,该实验结果表明本文所提出网络的模型参数量小于所对比的模型,且性能优于部分对比模型,说明本文所提模型在轻量化方面和性能方面综合来讲具有一定的客观优势。

4 结束语

针对基于深度学习的语音增强模型计算复杂度和参数量不断增加的问题,本文提出了基于门控混合膨胀卷积的轻量级语音增强网络。首先设计了门控混合膨胀卷积模块进行多尺度特征提取,有效保留语音长短时特征,增强模型鲁棒性;其次引入幻影卷积模块,通过减少特征冗杂,大幅降低了模型的计算复杂度;此外,还设计了一种层级通道注意力模块,能够在低参数量的条件下显著强化模型对通道

维度中语音特征相关性的捕捉能力。实验证明,本文提出的轻量化语音增强网络模型与其他相关模型相比,模型参数量更低,语音增强性能良好,实现了轻量化与优性能的有机结合。然而,本文方法的高度轻量化设计可能导致模型在复杂噪声环境中的抗干扰能力不足。后续需要研究如何改进网络结构来进一步提升网络的性能,以及增强网络的抗干扰能力。

参考文献:

[1] WANG Z, ZHU X, ZHANG Z, et al. SELM: Speech enhancement using discrete tokens and language models[C]// Proceedings of ICASSP2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2024: 11561-11565.

[2] YAO Z, GUO L, YANG X, et al. Zipformer: A faster and better encoder for automatic speech recognition[EB/OL]. (2024-03-24). https://arxiv.org/abs/2310.11230.

[3] 王晶,徐亮,陈晓娇,等. 基于神经网络的低码率语音编码技术研究综述[J]. 信号处理,2024, 40(12): 2261-2280.
WANG Jing, XU Liang, CHEN Xiaojiao, et al. Research review on low bit rate speech coding technology based on neural networks[J]. Journal of Signal Processing, 2024, 40(12): 2261-2280.

[4] CHEN J, SHI Y, LIU W, et al.Gesper: A unified framework for general speech restoration[C]//Proceedings of ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023: 1-2.

[5] UPADHYAY N, KARMAKAR A. Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study[J]. Procedia Computer Science, 2015, 54: 574-584.

[6] ABD EL-FATTAH M A, DESSOUKY M I, ABBAS A M, et al. Speech enhancement with an adaptive Wiener filter[J]. International Journal of Speech Technology, 2014, 17(1): 53-64.

[7] CHOI J H, CHANG J H. On using acoustic environment classification for statistical model-based speech enhancement[J]. Speech Communication, 2012, 54(3): 477-490.

[8] VALIN J M. A hybrid DSP/deep learning approach to real-time full-band speech enhancement[C]//Proceedings of 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP). Vancouver, BC, Canada: IEEE, 2018.

[9] TAN K, WANG D. Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom: IEEE, 2019: 6865-6869.

[10] TAN K, WANG D L. A convolutional recurrent neural network for real-time speech enhancement[C]//Proceedings of Interspeech 2018. Hyderabad, India: International Speech Communication Association, 2018: 3229-3233.

[11] HU Y, LIU Y, LV S, et al. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement[C]// Proceedings of Interspeech 2020. Shanghai, China: International Speech Communication Association, 2020: 2472-2476.

[12] HAO X, SU X, HORAUD R, et al. FullSubNet: A full-band and sub-band fusion model for real-time single-channel speech enhancement[C]//Proceedings of ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2021: 6633-6637.

[13] 胡沁雯,侯仲舒,乐笑怀,等. 一种轻量级全频带语音增强网络模型[J]. 数据采集与处理,2023, 38(2): 274-282.
HU Qinwen, HOU Zhongshu, LE Xiaohuai, et al. A lightweight full-band speech enhancement network model[J]. Journal of Data Acquisition and Processing, 2023, 38(2): 274-282.

[14] 张玥, 张雄伟, 孙蒙. 基于时频注意力机制与U-Net的骨导语音鲁棒增强方法[J]. 信号处理, 2022, 38(10): 2134-2143.
ZHANG Yue, ZHANG Xiongwei, SUN Meng. Bone-conducted robust speech enhancement based on time-frequency domain attention mechanism and U-Net[J]. Journal of Signal Processing, 2022, 38(10): 2134-2143.

[15] 姚瑶, 杨吉斌, 张雄伟, 等. 基于多维注意力机制的单通道语音增强方法[J]. 南京大学学报(自然科学版), 2023, 59(4): 669-679.
YAO Yao, YANG Jibin, ZHANG Xiongwei, et al. Single-channel speech enhancement based on multi-dimensional attention mechanism[J]. Journal of Nanjing University (Natural Sciences), 2023, 59(4): 669-679.

[16] 洪依, 孙成立, 冷严. 基于超轻量通道注意力的端对端语音增强方法[J]. 智能科学与技术学报, 2021, 3(3): 351-358.
HONG Yi, SUN Chengli, LENG Yan. End-to-end speech enhancement based on ultra-lightweight channel attention[J]. Chinese Journal of Intelligent Science and Technology, 2021, 3(3): 351-358.

[17] HAN K, WANG Y, TIAN Q, et al. GhostNet: More features from cheap operations[C]//Proceedings of the IEEE CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 1580-1589.

[18] 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. New York, USA: AAAI Press, 2020, 34(5): 9458-9465.

[19] VALENTINI-BOTINHAO C. Noisy speech database for training speech enhancement algorithms and TTS models, 2016 [EB/OL]. (2017-08-21)[2026-05-08]. https://datashare.ed.ac.uk/handle/10283/2791;jsessionid=8337E5C3F5D2D8558AD01C715E97A638.

[20] MACARTNEY C, WEYDE T. Improved speech enhancement with the Wave-U-Net[EB/OL]. (2018-11-27)[2026-05-08]. https://arxiv.org/abs/1811.11307.

[21] TAN K, WANG D. 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.

[22] LV S, FU Y, XING M, et al. S-DCCRN: Super wide band DCCRN with learnable complex feature for speech enhancement [C]//Proceedings of ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2022: 7767-7771.

[23] HAO X, LI X. Fast FullSubNet: Accelerate full-band and sub-band fusion model for single-channel speech enhancement [EB/OL]. (2022-12-18). https://arxiv.org/abs/2212.09019.

作者简介:

孙林慧(1979-),女,副教授,博士,硕士生导师,研究方向:语音处理与现代语音通信、深度学习和情感识别,E-mail: sunlh@njupt.edu.cn。

魏鹏滨(2000-),男,硕士研究生,研究方向:深度学习与单通道语音分离及语音增强,E-mail: 1024010335@njupt.edu.cn。

王春艳(1997-),女,硕士研究生,研究方向:深度学习与单通道语音分离及语音增强,E-mail: 2951494232@qq.com。

叶蕾(1978-),女,副教授,博士,研究方向:语音信号处理与图信号处理,E-mail: yel@njupt.edu.cn。

邵曦(1976-),**通信作者**,男,教授,博士生导师,研究方向:多媒体信息系统,E-mail: shaoxi@njupt.edu.cn。

(编辑:刘彦东)

Lightweight Speech Enhancement Based on Gated Hybrid Dilated Convolution

SUN Linhui, WEI Pengbin, WANG Chunyan, YE Lei, SHAO Xi*

(School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: To address the issues of parameter inflation and soaring computational complexity in mainstream speech enhancement models, a lightweight speech enhancement network based on gated hybrid dilated convolution is proposed in this paper. Firstly, a gated hybrid dilated convolution module is designed, which integrates gated linear units with hybrid dilated convolution to achieve multiscale feature extraction of speech signals and precise suppression of noise-sensitive regions, thereby effectively preserving both long-term and short-term speech characteristics while enhancing model robustness. Secondly, a hierarchical channel attention module is proposed to enhance the capture of speech feature correlations in channel dimensions through hierarchical feature fusion, while maintaining low parameter complexity. Experimental results on the VoiceBank+DEMAND dataset demonstrate that the proposed model, with only 0.41 million parameters, achieves competitive performance on the perceptual evaluation of speech quality (PESQ), the short-time objective intelligibility (STOI), cepstral signal-to-noise ratio (CSIG), cepstral background noise (CBAK) and cepstral overall loudness (COVL), thus achieving an organic integration of model lightweighting and high-precision performance.

Highlights:

1. Propose a lightweight speech enhancement network with gated hybrid dilated convolution.
2. Integrate multiscale feature extraction, channel attention, and Ghost convolution for efficient feature modeling.
3. Achieve a good balance between enhancement performance and model complexity on VoiceBank+DEMAND.

Key words: speech enhancement; gated hybrid dilated convolution; multiscale feature fusion; channel attention; low parameter complexity

---

Foundation items: Jiangsu Provincial Major Science and Technology Project (No.BG2024027); National Natural Science Foundation of China (No.61901227).

Received: 2025-05-21; **Revised**: 2025-08-30

***Corresponding author**, **E-mail**: shaoxi@njupt.edu.cn.