计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(4): 852−862, 2025

结合卷积增强窗口注意力的双分支语音增强神经网络

张晨辉　原之安　钱宇华
（山西大学大数据科学与产业研究院　太原　030006）
（山西省机器视觉与数据挖掘工程研究中心（山西大学）　太原　030006）
（15383661801@163.com）

Dual-Branch Speech Enhancement Neural Network with Convolutional Enhancement Window Attention

Zhang Chenhui, Yuan Zhi'an, and Qian Yuhua
（*Institute of Big Data Science and Industry*, *Shanxi university*, *Taiyuan* 030006）
（*Shanxi Machine Vision and data Mining Engineering Research Center* (*Shanxi university*), *Taiyuan* 030006）

Abstract　In complex environments and under sudden background noise conditions, speech enhancement tasks are extremely challenging due to the limited capturing of spectrogram features by existing methods, especially in capturing local information of the spectrogram. Previous work on Transformer models primarily focused on global information of the audio while neglecting the importance of local information. Many models only utilized the magnitude information and ignored the phase information after the audio underwent short-time Fourier transform (STFT), resulting in suboptimal capturing of spectrogram features and unsatisfactory speech enhancement results. Based on this, we propose a dual-branch speech enhancement neural network with convolutional enhancement window attention. The model adopts a U-NET architecture and simultaneously models the magnitude and phase information of the audio through the dual-branch structure. A complex computation module is introduced for information interaction between the two branches. The convolutional enhancement window attention module is employed in the skip-connection part between the encoder and decoder layers. It performs self-attention based on non-overlapping windows, significantly reducing the computational complexity of the speech enhancement model while capturing local contextual information. The proposed model is evaluated on the publicly available Voicebank-Demand dataset. Compared with the baseline models DCUNET 16 and DCUNET20, the proposed model achieves improvements of 0.51 and 0.47, respectively, in PESQ (perceptual evaluation of speech quality) metric. Other evaluation metrics also show significant enhancements. Compared with various existing speech enhancement models, the proposed model outperforms them in various metrics, particularly demonstrating remarkable improvements in PESQ scores.

Key words　speech enhancement; dual-branch network; spectrogram features; convolutional enhancement window attention; global information; local information

摘　要　在复杂环境以及突发背景噪音条件下,语音增强任务具有极大的困难和挑战.主要原因是现有的语音增强方法未能有效捕获语谱图特征,尤其是局部信息.在过去的研究中,Transformer 模型更专注于音

收稿日期: 2023-09-20; **修回日期**: 2024-11-19
基金项目: 国家自然科学基金重点项目(62136005); 新一代人工智能国家科技重大专项(2021ZD0112400); 山西省科技重大专项计划“揭榜挂帅”项目(202201020101006)
This work was supported by the Key Program of National Natural Science Foundation of China (62136005), the National Science and Technology Major Project (2021ZD0112400), and the Shanxi Provincial Science and Technology Major Special Plan "Unveiled" Project (202201020101006).
通信作者: 钱宇华(jinchengqyh@sxu.edu.cn)

频的全局信息，而忽略了局部信息的重要性. 在音频经过短时傅里叶变换（STFT）处理后，多数模型仅使用幅值信息，而忽略了相位信息，导致它们未能有效捕获语谱图特征，从而影响了语音增强的效果. 基于此设计出一个带有卷积增强窗口注意力的双分支语音增强神经网络. 该模型采用 U-NET 架构，通过双分支结构对音频的幅值和相位信息同时建模；在 2 个分支之间引入复值计算模块以实现信息交互；在编码器层和解码器层之间的跳跃连接部分采用卷积增强窗口注意力模块，该模块执行基于非重叠窗口的自注意力操作，在捕获局部上下文信息的同时显著降低了语音增强模型的计算复杂度. 该模型在公开的 Voicebank-Demand 数据集上进行测试，与基线模型 DCUNET 16 和 DCUNET20 相比，在客观语音质量评估指标 PESQ（perceptual evaluation of speech quality）分别提高了 0.51 和 0.47. 除了 PESQ 指标外，其他指标也都有显著的提升. 相较于现有的各类语音增强模型，该模型在各项指标上均处于领先水平，尤其是在 PESQ 得分方面的提升更为显著.

关键词 语音增强；双分支网络；语谱图特征；卷积增强窗口注意力；全局信息；局部信息

中图法分类号 TP391.4；TN912.3

DOI: 10.7544/issn1000-1239.202330751 **CSTR:** 32373.14.issn1000-1239.202330751

语音增强任务旨从含噪的音频信号中尽可能地去除背景噪音，提取出干净的音频信号，从而提高语音的客观感知质量和可懂度. 语音增强的应用前景十分广泛，包括了语音识别、声纹识别、听力修复、语音通信、智能家居设备等. 在过去几十年中，传统的语音增强技术，如维纳滤波法[1]、谱减法[2]、基于子空间[3]的方法等备受研究者们的青睐. 然而，这些方法依赖于语音和噪声的统计特性，在处理复杂环境下的语音信号时其表现效果不佳.

近年来，随着深度学习以及高性能计算机的发展，深度神经网络[4-6]（deep neural network, DNN）的优势已经推动了语音增强技术的快速发展，尤其是卷积神经网络[7-8]（convolutional neural network, CNN）和循环神经网络[9]（recurrent neural network, RNN）. 尽管研究人员一直以来都在尝试直接在时域语音信号上构建深度学习模型，然而由于人耳对频率感知更为敏感，往往通过短时傅里叶变换（short-time Fourier transform, STFT）方式处理音频进而构建语音增强模型[10-11]会比时域信号模型[12-13]有更优的表现. 早期的这些研究[14-18]主要集中在增强幅值信息和利用含噪语音的相位信息重构音频信号. 然而，这样的做法过于关注幅值信息，而忽略了相位信息的重要性. 近几年的研究表明[19-21]，在音频感知质量和可理解性方面，幅值信息和相位信息都具有重要的作用. 因此，已有工作开始同时考虑幅值和相位的估计. 早期的相位敏感掩码[19]（phase-sensitive mask, PSM）就将相位信息纳入掩码估计，在此基础上提出了复值比率掩码[20]（complex-valued ratio mask, CRM）来估计真实分量和复杂分量. 该算法通过同时增强干净语音和带噪语音，将语谱图分割成实部和虚部，从而能够很好地重构语音. 近几年，深度复杂网络[21]被开发出来，复值模型实现了先进的语音增强性能. 然而，之前的工作存在 2 个问题：1）建立在 U 型网络（U-NET）[22]结构或者卷积循环网络上，过度地依赖卷积层的表示能力. 2）同时对幅值信号和相位信号进行处理，但两者之间缺乏参数的交互.

Conformer[23]是在 Transformer[24]基础上改进的原始构造模型. 它引入了 2 个重要模块：自注意模块和卷积模块，自注意模块融合了空间信息和通道信息，以提高卷积层的表示能力. 自注意模块和卷积模块分别对全局信息和局部信息进行建模. 这些模块也在 Conformer 中证明了它们的有效性，尤其是在语音分离任务中展现了强大的时序建模能力. 然而，由于 Conformer 中的自注意力机制需要计算全部信息的长期依赖，导致其会造成更大的计算开销. 图像中的 Swin Transformer[25]采用了非重叠窗口的 Transformer 机制，它基于计算窗口注意力机制从而显著减小了计算复杂度. 然而，以往的工作中，窗口注意力[26−28]通常在语音信号的数据帧上计算，虽然窗口注意力能够通过滑动窗口的机制对特征图的不同区域进行特征提取，降低了计算复杂度，但是语音信号中的局部特征并不是均匀分布的，而是集中在特定时间段内. 在这种情况下，窗口注意力机制可能无法有效地捕获到关键的局部特征，使得在语音信号局部数据帧上信息交互的能力不尽如人意.

基于此，本文提出了一种结合卷积增强窗口注意力的双分支语音增强网络（dual-branch speech enhancement neural network with convolutional enhancement

window attention, DCEWA), 旨在利用 U-NET[22] 层次结构的多尺度特征融合能力以及 Transformer 的注意力机制, 更多地去除细小的噪音. DCEWA 模型是基于深度复值 U 型网络[21](deep complex U-NET, DCUNET), 延续了其编码器和解码器之间的架构, 在跳跃连接处采用了卷积增强窗口注意力(convolution enhancement window attention, CEWIN)模块. CEWIN 是一个高效且有效的模块, 它采用计算非重叠窗口注意代替全局自注意, 显著降低了高分辨率特征映射的计算复杂度. 由于本文构建了分层特征映射并保持窗口大小不变, 基于窗口的低分辨率自注意力能够更好地捕获更多的全局依赖. 另一方面, 局部上下文对于语音增强至关重要. 因此, 本文进一步在 Transformer 块的前馈网络中引入卷积增强模块, 以更有效地捕获局部上下文. 本文进行了消融实验以证明提出的卷积增强前向传播模块相对于传统的 Transformer 的前向传播模块的性能优势. 除了提出卷积增强窗口注意力机制外, 本文受到文献 [29] 的启发, 考虑到幅值信息和相位信息的关联, 提出了双分支结构, 分别对幅值信息和相位信息建模. 在损失函数的部分, 采用联合损失函数的计算方法, 对幅值信息和相位信息分别进行损失计算, 以及对音频进行计算尺度不变信噪比[29](scale-invariant source-to-noise ratio, *SI-SNR*)损失函数, 共同优化时频域和时域的 CRM 掩码估计. 消融实验结果表明, 这样的操作具有更佳的性能表现. 本文的主要贡献在于:

1)提出的卷积增强窗口注意力模块充分利用语谱图特征, 解决了传统 Transformer 在捕获局部上下文信息上的能力不足以及 Transformer 在计算全局注意力时计算复杂度高的问题.

2)提出双分支深度 U-NET 神经网络同时建模幅值信息和相位信息, 并且基于双分支网络设计出联合损失函数.

3)在消融实验证明所提出双分支网络、卷积增强模块与联合损失的优越性, 并且在公共 VCTK[30] 语料库上与先进的算法进行比较, 结果证明本文所提方法的优越性.

1 预备知识

将干净音频信号 $\boldsymbol{X}(n)$ 分割为长度为 $\boldsymbol{M}$、位移为 H 的重叠帧, 计算其短时傅里叶变换, 对每一帧应用窗函数 $W(n)$, 然后对其进行离散傅里叶变换(discrete Fourier transform, DFT), 将其转换到频域. 假设采用单边 M 点 DFT, 得到复数谱 $\boldsymbol{x}\in\mathbb{C}^{K\times L}$, $\boldsymbol{x}$ 定义为

$$\boldsymbol{X}_{k,l}=\sum_{n=0}^{M-1}\boldsymbol{x}(lh+n)\boldsymbol{w}(n)\mathrm{e}^{-\mathrm{j}2\pi\frac{kn}{M}}, \tag{1}$$

其中 k 是频率索引, l 是帧索引, k=m+2/l 是频点的数量, l 是时间帧的数量.

复数谱在极坐标下可以表示为幅度谱 $|\boldsymbol{X}|$ 和相位谱 $\phi\boldsymbol{X}$:

$$\boldsymbol{X}=|\boldsymbol{X}|\mathrm{e}^{\mathrm{i}\phi X}. \tag{2}$$

在语音增强的背景下, 考虑一个加性噪声模型:

$$\boldsymbol{X}=\boldsymbol{S}+\boldsymbol{V}, \tag{3}$$

其中 $\boldsymbol{S}$ 和 $\boldsymbol{V}$ 分别是纯净语音信号和加性噪声分量. 给定噪声信号 $\boldsymbol{X}$, 任务是计算估计干净信号 $\boldsymbol{S}=|\boldsymbol{S}|\mathrm{e}^{\mathrm{i}\hat{\phi}}$, 将噪声信号 $\boldsymbol{X}$ 输入到语音增强模型中, 经过神经网络的优化, 得到乘性掩码 $\boldsymbol{M}$, 将此掩码 $\boldsymbol{M}$ 与噪声 $\boldsymbol{X}$ 相乘并将结果转换回时域, 得到估计的干净信号 $\hat{\boldsymbol{S}}$.

2 本文所提方法

由于语音频谱图上的信息较为丰富, 充分捕获和利用语谱图特征是语音增强的有效方法. 在本节中针对语谱图特征的提取, 首先描述了 DCEWA 整体设计以及模型的层次结构; 其次, 对 DCEWA 模型各个部分进行了详细的描述.

2.1 整体设计

图 1 展示了本文设计的 DCEWA 的整体架构示意图. 它是一个带有跳跃连接的 U 型网络, DCEWA 的输入是带噪音频通过短时傅里叶变换得到的频域的幅值谱以及相位谱, 其可表示为 $\boldsymbol{X}$=[T, F, W], 其中 T 代表帧数, F 代表频点数. DCEWA 包含 2 个分支: 一个分支是用来预测音频的幅值部分, 另一个是预测相位部分. 这 2 个分支具有相同的网络架构. 具体来讲, 对于给定的带噪音频, DCEWA 首先通过短时傅里叶变化将音频分为幅值和相位 2 部分, 得到 ***Mag*** = [B, C, H, W], ***Phase*** = [B, C, H, W], 其中 ***Mag*** 代表幅值信息, ***Phase*** 代表相位信息. 接下来, 按照 U 型结构, 特征 ***Mag*** 和 ***Phase*** 分别经过 K 个编码器以及 K 个解码器. 在 2 个分支的编码器之间, 通过一个复值计算模块实现了幅值信息和相位信息的交互, 在编码模块和解码模块之间采用结合卷积增强非重叠窗口注意力模块的跳跃连接方式进行连接, 在得到每个分支的输出之后, 通过 CRM 进行计算掩码操作. 最后, 通过逆短时傅里叶变化得到增强后的音频.

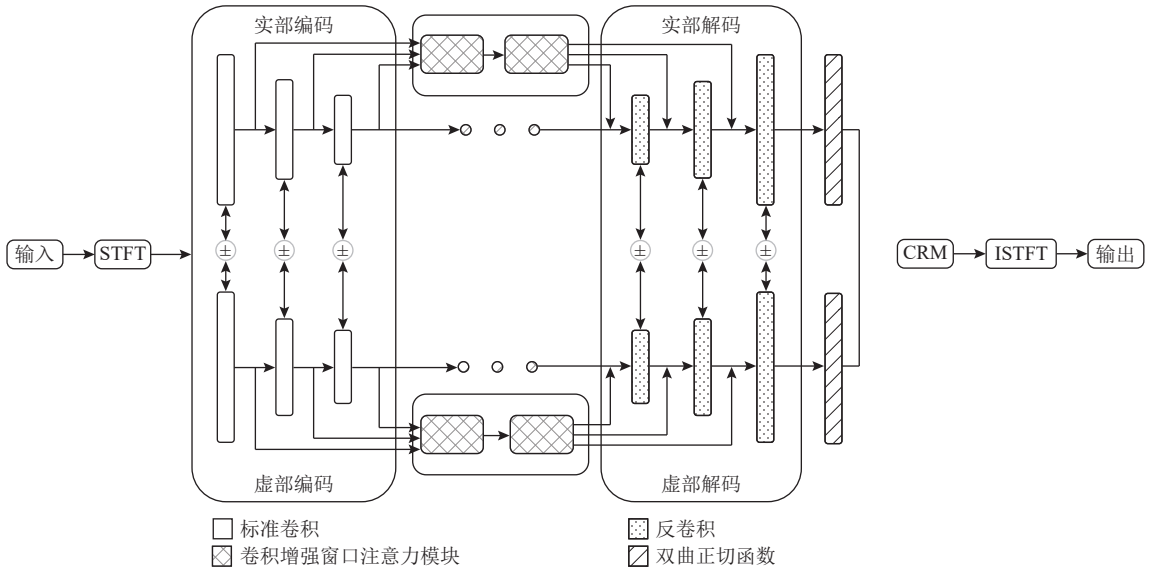


Fig. 1 The network structure of DCEWA

图 1 DCEWA 网络结构

2.2 编码器-解码器

本文的编码器模块得益于 DCUNET 的启发，两者不同之处在于，前者将幅值信息与相位信息进行分离，分别对其进行建模. 具体操作为将幅值信息和相位信息分别输入编码器后，首先通过 2D 卷积层实现下采样操作，在分别完成卷积操作后，通过复值计算模块与另外一个分支建立连接，其可表示为

$$\begin{aligned} \boldsymbol{Mag} &= Conv2d(\boldsymbol{Mag}) - Conv2d(\boldsymbol{Phase}), \\ \boldsymbol{Phase} &= Conv2d(\boldsymbol{Mag}) + Conv2d(\boldsymbol{Phase}), \end{aligned} \tag{4}$$

其中 $\boldsymbol{Mag}$ 代表幅值信息，$\boldsymbol{Phase}$ 代表相位信息，$Conv2d$ 代表 2D 卷积层.

实现了实部和虚部信息的交互，然后将 2 个特征信息分别通过归一化（batch normalization, BN）层和修正线性单元（rectified linear unit, ReLU）层. 解码器模块参照编码器模块，依旧采用幅值和相位分离的策略进行建模. 音频特征进入 2D 反卷积层后同样通过交互模块实现信息交互，紧接着进入 BN 层和 ReLU 层. 不同之处在于，在解码模块的最后一层，进入一个 2D 反卷积层后，会通过双曲正切函数（hyperbolic tangent function, TANH）约束估计的实部和虚部，以便在 CRM 掩码模块更好地估计掩码 $\boldsymbol{M}$.

在设计编码器-解码器模块时，本文参照了 DCUNET 中编码器-解码器操作. 具体设计了 7 层编码器和 10 层解码器，具体的网络参数相较于 DCUNET 又有些改进，整体的设计跟 DCUNET 16 和和 DCUNET 20 接近，但是在通道数上有所改进. 其中图 2 表示的是 DCEWA 16 所用卷积层的层数与卷积的参数，图 3 代表的是 DCEWA 20 所用到卷积层的层数与卷积的参数，每个卷积操作之后是 BN 层和 ReLU 层.

2.3 卷积增强窗口注意力模块

卷积增强窗口注意力模块旨在更好地提取语音频谱图上的信息. 由于噪声类型的多样性以及干净音频和带噪音频之间的差异，这非常具有挑战. 在之前的工作中，Tranformer 捕获局部上下文信息的能力十分有限，且进行全局注意力计算又会造成过多的计算量. 已有实验证明[31]，当帧长度较短时，相位谱变得更加重要，而幅度谱的重要性降低. 当同时估计幅度和相位时，较短的帧可以显著提高具有显式相

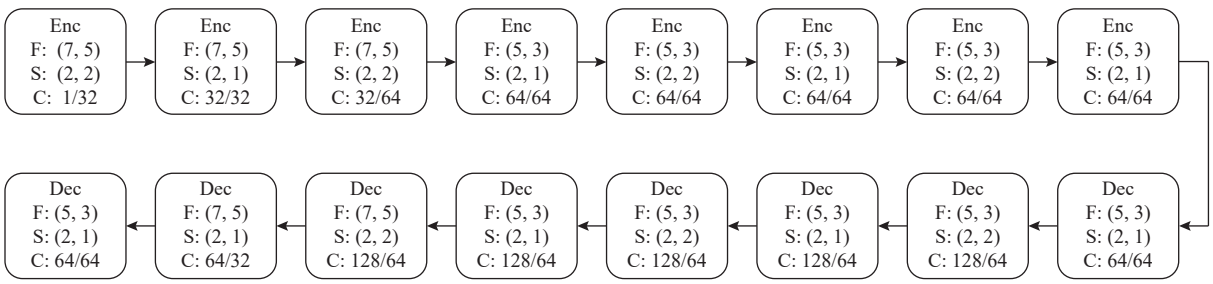


Fig. 2 DCEWA 16 (16 layers): a model with 16 convolutional layers

图 2 DCEWA 16（16 层）：16 层卷积层的模型

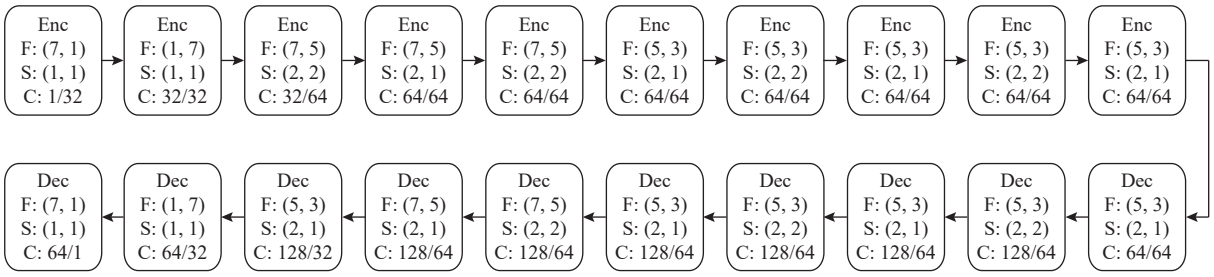


Fig. 3 DCEWA 20 (20 layers): a model with 20 convolutional layers

图 3 DCEWA 20（20 层）：20 层卷积层的模型

位估计的 DNN 的性能. 音频局部信号相比音频全局信号在更短的帧上处理音频能够有效提升语音增强性能. 为了解决这个问题, 得益于在 Uformer[32] 网络里面 LeWin 对图像去噪的应用, 以及 Conformer 在语音识别领域的优异性, 本文设计出了语音增强的 CEWIN 模块, 将其应用于 U-NET 里面的跳跃连接处, CEWIN 采用 Transformer 的自注意力捕获长距离依赖, 同时 Transformer 联合卷积增强模块, 以更好地捕获有用的局部上下文信息. CEWIN 模块包含非重叠窗口注意力 W-MSA(window-based multi-head self-attention)模块、卷积增强(convolution enhancement, CE)模块、卷积增强前向传播(convolution enhancement forward propagation, CEFP)模块. 如图 4(a)所示.

图 4(a)中 W-MSA 模块采用非重叠窗口的自注意力机制, 通过该方法, 在非重叠窗口内计算自注意力机制, 有效降低了计算量.

W-MSA 模块的实现如图 5 所示. 给定音频信号 $\boldsymbol{X}=[B, C, H, W]$, 将音频特征信息 H, W 根据窗口大小 M, 通过补零填充的方式将音频特征对齐为窗口 M 的倍数. 然后, 将音频信号 $\boldsymbol{X}$ 拆分成窗口大小为 $M\times M$ 的非重叠窗口, 从每个窗口中得到平移和转置特征 $\boldsymbol{X}^i\in\mathbb{R}^{M^2\times C}$, 在每个窗口内执行自注意力机制. 设多头注意力机制的头是 k, 第 k 个头的非重叠窗口计算过程可以描述为

$$\begin{aligned}&\boldsymbol{X}=\{\boldsymbol{X}^1, \boldsymbol{X}^2, \cdots, \boldsymbol{X}^N\},\ N=HW/\boldsymbol{M}^2,\\&\boldsymbol{Y}_k^i=Attention(\boldsymbol{X}^i\boldsymbol{W}_k^{\boldsymbol{Q}}, \boldsymbol{X}^i\boldsymbol{W}_k^{\boldsymbol{K}}, \boldsymbol{X}^i\boldsymbol{W}_k^{\boldsymbol{V}}),\ i=1, 2, \cdots, N,\\&\hat{\boldsymbol{X}}_k=\{\boldsymbol{Y}_k^1, \boldsymbol{Y}_k^2, \cdots, \boldsymbol{Y}_k^N\},\end{aligned}\tag{5}$$

其中 $\boldsymbol{W}_k^{\boldsymbol{Q}}, \boldsymbol{W}_k^{\boldsymbol{K}}, \boldsymbol{W}_k^{\boldsymbol{V}}\in\mathbb{R}^{C\times d_k}$ 分别表示第 k 个头的查询 $\boldsymbol{Q}$(query)、键 $\boldsymbol{K}$(key)、值 $\boldsymbol{V}$(value), $\hat{\boldsymbol{X}}_k$ 表示的是第 k 个头的输出. 所有的头的输出将被连接起来, 接下来经过线性变换得到最终的结果.

标准的 Transformer 模块在捕获局部上下文信息方面受到限制. 由于近邻音频对于语音增强来说至关重要, 因此本文提出 CE 模块来捕获局部上下文信息, 如图 4(b)所示. 该结构是一个非对称的结构, 其中 Liner Layer 代表全连接层, GELU 代表激活函数, Conv 是一个 2D 卷积层. 通过消融实验证明, CE 模块应用到前向传播模块里面可以明显提升模型性能.

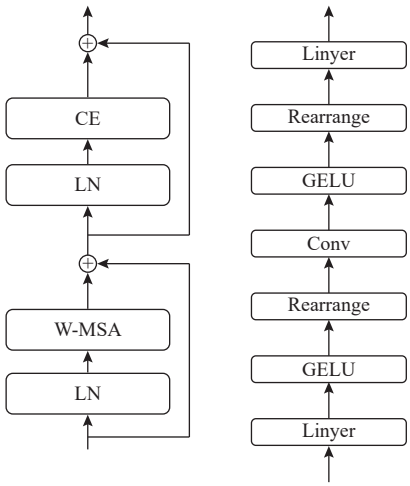


Fig. 4 Illustration of CEWIN block and CE block

图 4 CEWIN 模块和卷积增强模块示意图

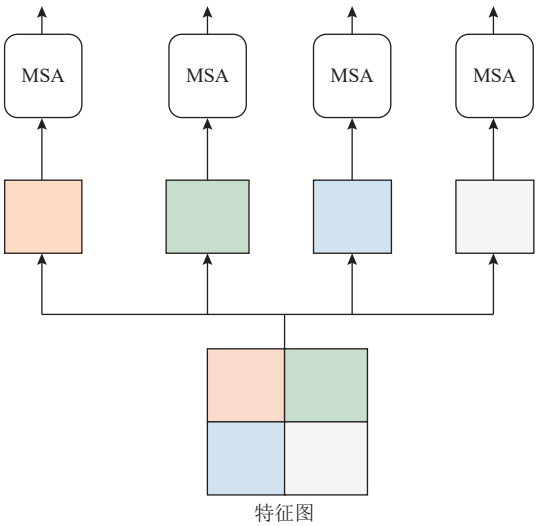


Fig. 5 Illustration of W-MSA block

图 5 W-MSA 模块示意图

为了更好地捕获语谱图上的信息, 本文在之前

的工作基础上重新设计了前向传播网络，提出了CEFP. 将以往的全连接层、激活层等变成上述的CE模块，并在CE模块前又加入线性归一化（linear normalization, LN）层实现归一化，保证数据特征分布的稳定性. 具体操作可见图4（b）. 前向传播模块可以表示为：

$$\begin{aligned} \boldsymbol{X}_l' &= \boldsymbol{W} - MSA(LN(\boldsymbol{X}_{l-1})) + \boldsymbol{X}_{l-1}, \\ \boldsymbol{X}_l &= CE(LN(\boldsymbol{X}_l')) + \boldsymbol{X}_l', \end{aligned} \tag{6}$$

其中$\boldsymbol{X}_l'$和$\boldsymbol{X}_l$分别表示W-MAS模块和CE模块的输出，LN代表归一化.

2.4 CRM 掩码模块

通过编码器和解码器模块后，将通过CRM掩码模块得到掩码$\boldsymbol{M}$. 给定一个带噪音频$\boldsymbol{x}(n)=\boldsymbol{y}(n)+\boldsymbol{z}(n)$，其中$\boldsymbol{x}(n)$表示带噪音频，$\boldsymbol{y}(n)$表示干净音频，$\boldsymbol{z}(n)$表示噪声音频，语音增强的任务是从$\boldsymbol{x}(n)$里面估计出$\boldsymbol{y}(n)$，将$\boldsymbol{x}$和$\boldsymbol{y}$经过短时傅里叶变换后得到$\boldsymbol{X}=\boldsymbol{X}_{\mathrm{r}}+\mathrm{j}\boldsymbol{X}_i, \boldsymbol{Y}=\boldsymbol{Y}_{\mathrm{r}}+\mathrm{j}\boldsymbol{Y}_i$，CRM可以定义为

$$\boldsymbol{M}=\frac{\boldsymbol{X}_{\mathrm{r}}\times\boldsymbol{Y}_{\mathrm{r}}+\boldsymbol{X}_i\times\boldsymbol{Y}_i}{\boldsymbol{X}_{\mathrm{r}}^2+\boldsymbol{X}_i^2}+\mathrm{j}\frac{\boldsymbol{X}_{\mathrm{r}}\times\boldsymbol{Y}_i-\boldsymbol{X}_i\times\boldsymbol{Y}_{\mathrm{r}}}{\boldsymbol{X}_{\mathrm{r}}^2+\boldsymbol{X}_i^2}. \tag{7}$$

本文采用信号近似（SA）的方法，直接将干净语音的幅值或复杂语谱图与使用掩码的噪声幅值或复谱图之间的差异最小化. 损失函数表示为$Loss(\hat{\boldsymbol{m}}\times\boldsymbol{Y},\boldsymbol{S})$，具体的公式表示为

$$\hat{\boldsymbol{Y}}=(\boldsymbol{X}_{\mathrm{r}}\times\hat{\boldsymbol{M}}_{\mathrm{r}}-\boldsymbol{X}_i\times\hat{\boldsymbol{M}}_i)+\mathrm{j}(\boldsymbol{X}_{\mathrm{r}}\times\hat{\boldsymbol{M}}_j-\boldsymbol{X}_i\times\hat{\boldsymbol{M}}_{\mathrm{r}}). \tag{8}$$

2.5 损失函数

为了探究幅值信息和相位信息对语音增强性能的影响，本文提出分别对幅值信息和相位信息设计损失函数，并联合时域信息建立联合损失函数. 在对时域信息建立损失函数时，采用了与基准模型DCUNET相同的加权信噪比（weighted-signal-to-distortion ratio, *WSDR*）损失函数和DPRNN[33]中的*SI-SNR*损失函数进行对比. 实验结果表明采用*SI-SNR*损失函数的性能会更优越，对幅值信息和相位信息分别设计损失函数可以显著提高模型性能.

*WSDR*损失函数是DCUNET中考虑到音频信号原始波形包含固有的相位信息的一种联合时域评估策略的损失函数，它是在信号失真比（signal-to-distortion ratio, *SDR*）的基础上改进而来的. 具体可以表示为

$$\begin{aligned} &\max_{\hat{\boldsymbol{y}}} SDR(\boldsymbol{y},\hat{\boldsymbol{y}}) := \max_{\hat{\boldsymbol{y}}} \frac{\langle\boldsymbol{y},\boldsymbol{y}\rangle}{\|\boldsymbol{y}\|^2\|\boldsymbol{y}\|^2-\langle\boldsymbol{y},\boldsymbol{y}\rangle^2} \propto \min_{\hat{\boldsymbol{y}}} \frac{\|\boldsymbol{y}\|^2}{\langle\boldsymbol{y},\boldsymbol{y}\rangle^2}, \\ &Loss_{\mathrm{ven}}(\boldsymbol{y},\hat{\boldsymbol{y}}) := -\frac{\langle\boldsymbol{y},\hat{\boldsymbol{y}}\rangle^2}{\|\hat{\boldsymbol{y}}\|^2}, \end{aligned} \tag{9}$$

其中$\boldsymbol{y}$是干净音频，$\hat{\boldsymbol{y}}$是增强的音频，$\langle\cdot,\cdot\rangle$表示2个向量之间的点积.

*SI_SNR*是被普遍用来代替均方误差（mean square error, MSE）的评价指标. *SI_SNR*可以表示为

$$\begin{cases} S_{\mathrm{target}} := (\langle\tilde{\boldsymbol{s}},\boldsymbol{s}\rangle\times\boldsymbol{s})/\|\boldsymbol{s}\|_2^2, \\ e := \tilde{\boldsymbol{s}}-\boldsymbol{s}_{\mathrm{target}}, \\ SI_SNR := 10\lg\left(\dfrac{\|\boldsymbol{s}_{\mathrm{target}}\|_2^2}{\|\boldsymbol{s}_{\mathrm{target}}\|_2^2}\right), \end{cases} \tag{10}$$

其中$\boldsymbol{s}$和$\hat{\boldsymbol{s}}$分别表示纯净的和估计的时域波形，$\|\cdot\|_2$代表L2范式.

在双分支网络的推动下，幅值信息和相位信息实现分离. 为了充分利用语谱图上的音频信息，并让损失函数与CRM掩码估计相关，本文设计提出了损失函数*Loss_D_R*和*Loss_D_I*，用于计算幅值损失和相位损失. *Loss_D_R*的具体操作为将增强后的幅值信息和干净音频的幅值信息经过函数*D_Real*，且*D_Real*包含5层卷积层和2层注意力层，除最后一层卷积层外，所有的卷积层后面都跟有Leaky ReLU激活函数. *D_Real*函数具体排列方式如图6所示. 然后，再经过式（11）表示得到*Loss_D_R*. 对相位分支进行的*D_Imag*函数和*Loss_D_I*损失函数的操作相同.

$$\begin{aligned} &Loss_D_R(\boldsymbol{mag_enh},\boldsymbol{mag_cle}) = \\ &((D_real(\boldsymbol{mag_enh})-\boldsymbol{s})^2+(D_real(\boldsymbol{mag_cle}))^2)/2, \end{aligned} \tag{11}$$

其中***mag_enh***表示增强音频的幅值信息，***mag_cle***表示干净音频的幅值信息，$\boldsymbol{s}$表示全1矩阵.

基于此，本文对比分析了4种损失函数，来验证不同损失函数对模型性能的影响. 具体为

$$\begin{aligned} L1 &= WSDR, \\ L2 &= WSDR + Loss_{DR} + Loss_{DI}, \\ L3 &= SI_SNR, \\ L4 &= SI_SNR + Loss_{DR} + Loss_{DI}. \end{aligned} \tag{12}$$

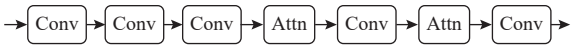


Fig. 6 Illustration of D_Real block

图6 D_Real模块示意图

3 实验设置

3.1 数据集

为了检验模型性能，本文选择了对比模型常用的VCTK语料库进行相关的实验. 为了方便跟其他模型进行对比，本文选取了其中的带有28个说话人的训练集（包含14个男性说话人和14个女性说话

人), 总共有 11 572 个混合音频, 测试集包含 2 个说话人, 共有 824 个混合音频. 训练集采用 Demand[34] 数据集, 从中选取出了 8 种噪声, 人工合成 2 种噪声, 采用 4 种信噪比, 分别为 15 dB, 10 dB, 5 dB, 0 dB 来合成共 40 种噪声信号, 测试集在 Demand 中选取出 2 种噪声, 采用 4 种信噪比, 分别为 17.5 dB, 12.5 dB, 7.5 dB, 2.5 dB 来合成共 20 种噪声信号. 原始音频信号的采样率为 48 kHz, 包括了清晰的语音信号以及各种类型和水平的噪声信号. 这些噪声信号被添加到语音信号中, 以模拟真实世界中的语音信号干扰情况. 除了语音信号和噪声信号, 该数据集还提供了相应的信噪比值和噪声类型的注释信息, 以方便模型的训练和评估. 为了与其他实验进行对比, 本文将所有的音频信号重采样为 16 kHz. 本文用 0 填充方法将音频补齐到 10 s.

3.2 实验设置以及评价指标

本文在数据预处理过程中, 采用了具有 40 ms 大小的汉宁(Hanning)的窗口和 10 ms 的帧长. 对于所有的模型在 VCTK 数据集上训练 50 轮, 本文使用 PyTorch 平台和 Adam 优化, 在 DCEWA 16 的模型上设置 *BatchSize* 为 4, 在 DCEWA 20 上设置 *BatchSize* 为 2, 设置学习率为 0.001 和 Decay 为 0.1.

本文使用多个客观测量来评估所提出的语音增强模型, 包括: 语音质量的感知评估[35](PESQ), 其值范围为−0.5~4.5; 短时客观可懂度[36](STOI), 得分在 0~1. 本文还采用主观平均意见得分(MOS), 包括用于信号失真的 CSIG(1~5)、用于噪声失真评估的 CBAK(1~5)和用于总体质量评估的 COVL(1~5).

4 对比实验及结果分析

实验部分主要包括 4 部分: 1)将本文设计的模型与基准模型以及现有的先进模型进行对比. 2)进行对比实验, 探索不同损失函数对本文提出模型的语音增强性能的影响. 3)进行消融实验 1, 探索设计的卷积增强前向传播网络和传统的前向传播网络对性能的影响. 4)进行消融实验 2, 探索双分支模型和卷积增强窗口注意力对性能的影响.

4.1 不同模型性能对比

在本节, 本文提出的模型将与基准模型以及包括时域方法和时频域(T-F)在内的现有 SOTA(state-of-the-art)方法进行比较. 其中时域方法包括 Wave-U-Net[27], SEGAN[37], WaveCRN[38], TSTNN[39], MSANet[40]; 时频域方法, 包括 DCUNET 16 和 DCUNET 20 以及 MetricGAN+[41], FTDDN[42], DB-AIAT[43], DBT-Net[44]. 所有模型都使用 VCTK 语料库进行增强实验.

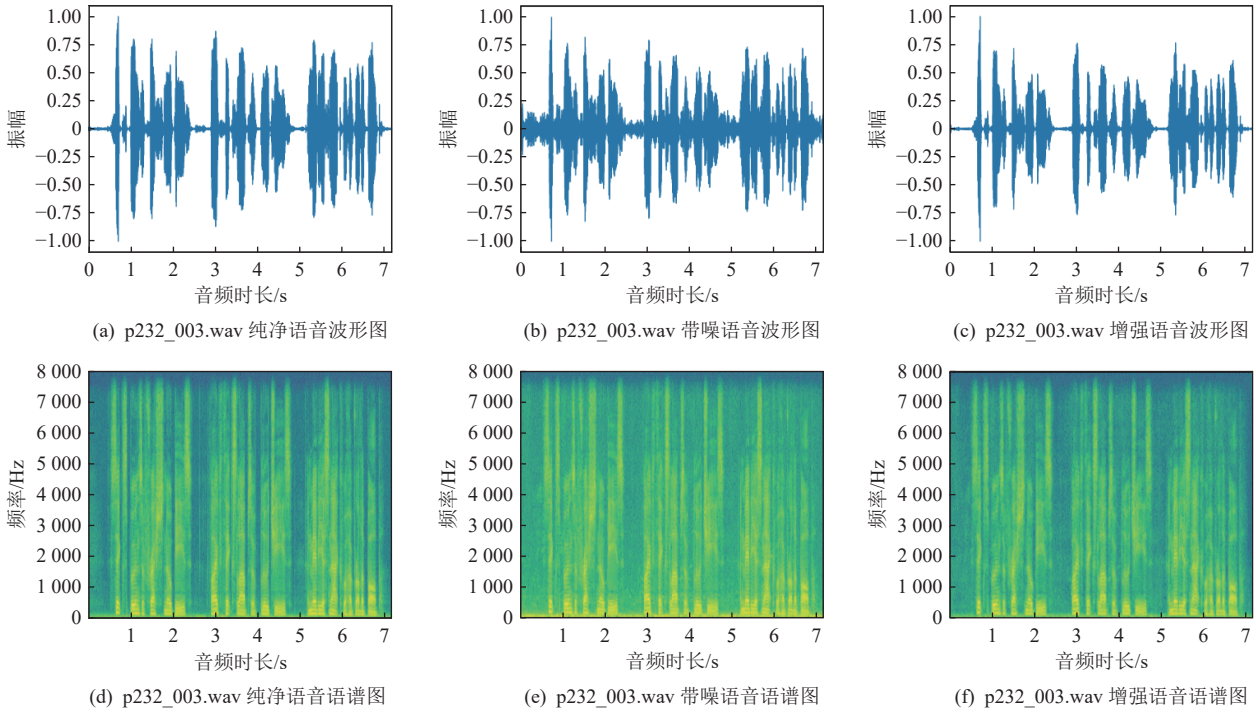
表 1 列出了对比结果. 相较于基线模型 DCUNET 16, DCEWA 16 在 PESQ, CSIG, CBAK, COVL 等评价指标有了 0.51, 0.12, 0.19, 0.34 的提升, DCEWA 20 相比于 DCUNET 20, 在 PESQ, CSIG, CBAK, COVL 得分上有了 0.47, 0.15, 0.06, 0.30 的提升. 本文提出的 DCEWA 20 模型在绝大多数指标上都优于其他对比模型, 尤其是在 PESQ 指标上的提升更为明显, 远远优于其他模型. 特别地, 可以发现 MSANet, DB-AIAT, DBT-Net 的 CSIG 分数略高于本文所提模型, 这 3 种模型都是基于 Transformer 的基线方法. CSIG 指标用于衡量信号失真度, 相比于注意力机制依靠全局依赖处理音频, 本文使用的窗口注意力在音频信号处理过程中将音频频谱图分割成不同区域进行处理, 原始音频会在信号处理过程中发生略微的改变或扭曲, 使得信号失真度略高, 从而造成了略低的 CSIG 得分. 具体的可视化结果如图 7 和图 8 所示, 其中图 7 和图 8 给出了纯净音频、带噪音频、DCEWA 20 模型增强后的波形图和语谱图. 对比波形图得出, DCEWA 20 模型增强后的波形图与纯净音频几乎一致. 在语谱图的表现上可以看出, DCEWA 20 模型处理低信噪比时对背景噪声的抑制更加充分, 对增强信号的谐波成分保持得更好, 增强带来的失真较小. 可以看出

Table 1 Evaluation Scores of DCEWA Baseline Models Using VCTK Corpus

表 1 使用 VCTK 语料库对 DCEWA 基线模型的评测分数

方法	评价指标				
	PESQ	STOI	CSIG	CBAK	COVL
Noisy	1.97	0.92	3.34	2.44	2.63
SEGAN	2.16		3.48	2.94	2.80
Wave-U-Net	2.40		3.52	3.24	2.96
WaveCRN	2.64		3.94	3.37	3.29
DCUNET 16	3.01		4.10	3.77	3.52
DCUNET20	3.13		4.24	4.00	3.69
TSTNN	2.96	0.95	4.10	3.77	3.52
MSANet	3.11	0.95	4.44	3.63	3.81
MetricGAN+	3.15	0.95	4.14	3.16	3.64
FTDDN	3.02	0.95	4.25	3.49	3.63
DB-AIAT	3.31	0.95	**4.61**	3.75	3.96
DBT-Net	3.30	0.95	4.59	3.75	3.92
DCEWA 16 (本文)	3.52	0.94	4.22	3.96	3.86
DCEWA 20 (本文)	**3.60**	**0.95**	4.39	**4.06**	**3.99**

注: 黑体数值表示最优值.

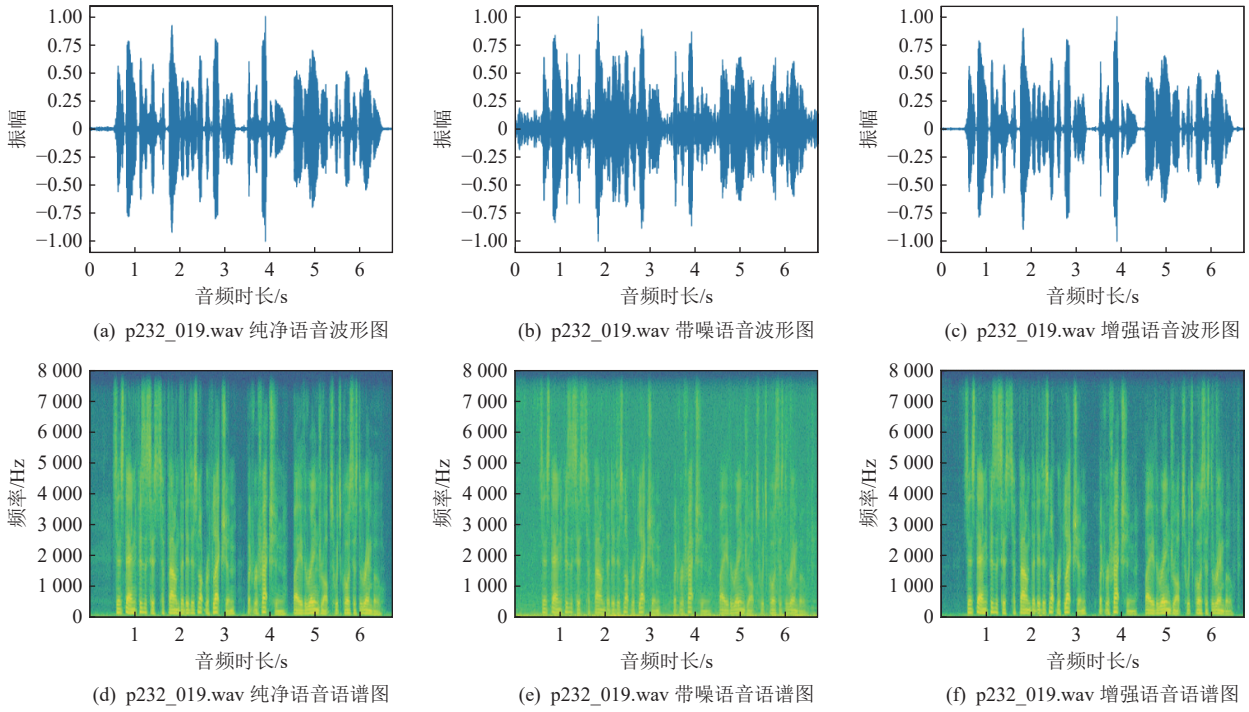


(a) p232_003.wav 纯净语音波形图 (b) p232_003.wav 带噪语音波形图 (c) p232_003.wav 增强语音波形图

(d) p232_003.wav 纯净语音语谱图 (e) p232_003.wav 带噪语音语谱图 (f) p232_003.wav 增强语音语谱图

Fig. 7 Time-frequency spectrograms of clean speech, noisy speech, and enhanced DCEWA 20 model

图 7 纯净语音、含噪语音以及增强后的 DCEWA 20 模型语音时频谱



(a) p232_019.wav 纯净语音波形图 (b) p232_019.wav 带噪语音波形图 (c) p232_019.wav 增强语音波形图

(d) p232_019.wav 纯净语音语谱图 (e) p232_019.wav 带噪语音语谱图 (f) p232_019.wav 增强语音语谱图

Fig. 8 Time-frequency spectrograms of clean speech, noisy speech, and enhanced DCEWA 20 model

图 8 纯净语音、带噪语音以及增强后的 DCEWA 20 模型语音时频谱

DCEWA 20 模型充分地利用了 U-NET 层次结构的多尺度特征融合能力以及 Transofrmer 的注意力机制，对语音信号的建模更为准确.

4.2 不同损失函数对模型性能的影响

对本文提出的 4 种损失函数进行对比，分别对 DCEWA 16 以及 DCEWA 20 进行了实验对比. 通过表 2 可以看出，*SI-SNR* 损失函数相比于 *WSDR* 在绝大多数指标上具有更优的性能. 本文设计提出的 *Loss_D_R* 和 *Loss_D_I* 在与 *SI_SNR* 损失函数联合时降噪效果有着明显的改善，然而在与 *WSDR* 进行联合时效果

Table 2 Evaluation Scores of Different Loss Functions Using VCTK Corpus

表 2 使用 VCTK 语料库时的不同损失函数的性能评测分数

方法	评价指标				
	PESQ	STOI	CSIG	CBAK	COVL
DCEWA 16-L1	3.41	0.94	4.04	3.90	3.72
DCEWA 16-L2	3.37	0.94	3.97	3.95	3.66
DCEWA 16-L3	3.45	0.94	4.14	4.00	3.66
DCEWA 16-L4	3.52	0.94	4.22	3.96	3.86
DCEWA 20-L1	3.48	0.94	4.21	3.97	3.83
DCEWA 20-L2	3.45	0.94	4.20	4.01	3.82
DCEWA 20-L3	3.55	0.94	4.29	4.03	3.91
DCEWA 20-L4	**3.60**	**0.95**	**4.39**	**4.06**	**3.99**

注：黑体数值表示最优值.

有些许下降. 消融实验证明本文所提出的联合损失函数的优越性.

4.3 消融实验 1

在本节, 为了探索所提出的带卷积增强的前向传播网络(CEFP)与传统前向传播(forward propagation, FP)性能的对比, 本文分别采用 DCEWA-CEFP 以及 DCEWA-FP. 根据表 3 可以看出, 本文所提出的卷积增强前向传播网络的各方面性能都优于传统前向传播网络的性能, 其中在 PESQ 指标上的提升更为显著.

Table 3 Performance Evaluation Scores of Different Forward Propagation Using VCTK Corpus in This Paper

表 3 使用 VCTK 语料库时的本文不同前向传播性能评测分数

方法	评价指标				
	PESQ	STOI	CSIG	CBAK	COVL
DCEWA 16-FP	3.49	0.94	4.21	3.94	3.86
DCEWA 16-CEFP	3.52	0.94	4.22	3.96	3.86
DCEWA 20-FP	3.56	0.95	4.20	4.05	3.94
DCEWA 20-CEFP	**3.60**	**0.95**	**4.39**	**4.06**	**3.99**

注：黑体数值表示最优值.

4.4 消融实验 2

通过消融实验分别检验双分支模块和卷积增强窗口注意力模块在 DCEWA 网络中的有效性. 引入 2 种网络模型分别在 DCEWA 16 和 DCEWA 20 上进行对比, 具体实验结果如表 4 所示. 2 种模型包括了单分支 U 形网络(single branch U-NET, SUNET), 该网络仅对幅值信息进行处理, 双分支 U 形网络(dual branch U-NET, DUNET), 该网络在幅值和相位上对信号分别处理, 2 个模型均采用普通的跳跃连接方式. 根据表 4 可以看出, 本文提出的双分支模型能捕获更多的时频信息, 有效提升增强效果, 同时在跳跃连接部分采用卷积增强窗口注意力模块, 相较于普通的跳跃连接可以更好地捕获局部上下文信息, 提升增强性能.

Table 4 Evaluation Scores of Our Proposed Different Network Models Using VCTK Corpus in Ablation Experiment

表 4 使用 VCTK 语料库时的本文所提不同的网络模型的消融实验评测分数

方法	评价指标				
	PESQ	STOI	CSIG	CBAK	COVL
SUNET16	3.16	0.93	3.74	3.65	3.43
DUNET16	3.46	0.94	4.21	3.99	3.85
DCEWA 16	3.52	0.94	4.22	3.96	3.86
SUNET20	3.19	0.93	4.02	3.59	3.59
DUNET20	3.52	0.94	4.29	4.04	3.90
DCEWA 20	**3.60**	**0.95**	**4.39**	**4.06**	**3.99**

5 结 论

本文提出了一种带有卷积增强窗口的双分支语音增强神经网络, 进一步地对深度复杂网络模型进行了改进. 在这一结构中, 采用双分支网络, 同时对幅值和相位进行建模. 在编码器和解码器之间引入卷积增强窗口模块, 以加强音频信号特征的提取, 通过此模块高效地提取了局部和全局上下文信息. 实验结果表明, 本文的模型在多数评价指标上都优于目前的大多数方法. 此外, 本文提出了 4 种损失函数进行对比, 结果表明本文联合损失函数具有更佳的性能表现. 相较原始模型, 本文提出的模型具有更好的性能, 通过实验研究证明了所提出的方法在相同的数据集上的优越性能. 尽管如此, 本文直接在频域进行语音增强和采用复值比率掩码存在计算复杂度高的问题. 下一步的工作是考虑采用频带结构代替频谱, 聚焦语音的频谱包络与周期性, 保证增强性能的同时进一步降低计算复杂度, 以达到在带噪语音上进行高质量实时增强的目的.

作者贡献声明: 张晨辉提出研究问题和研究方法, 收集和分析数据, 完成实验和撰写论文; 原之安设计实验并完善研究方法, 修改引言部分; 钱宇华协

调成员之间合作，并对论文整体结构和逻辑进行了把控和修改；张晨辉和原之安对本文具有同等贡献.

参 考 文 献

[1] Lim J, Oppenheim A. All-pole modeling of degraded speech[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1978, 26(3): 197−210

[2] 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

[3] Ephraim Y, Van Trees H. A signal subspace approach for speechenhancement[J]. IEEE Transactions on Speech and Audio Processing, 1995, 3(4): 251−266

[4] Shi Wenhua, Ni Yongjing, Zhang Xiongwei, et al. Deep neural network based monaural speech enhancement with sparse nonnegative matrix factorization[J]. Journal of Computer Research and Development, 2018, 55(11): 2430−2438 (in Chinese)
(时文华, 倪永婧, 张雄伟, 等. 联合稀疏非负矩阵分解和神经网络的语音增强[J]. 计算机研究与发展, 2018, 55(11): 2430−2438)

[5] Ali M N, Brutti A, Falavigna D. Speech enhancement using dilated Wave-U-Net: An experimental analysis[C]//Proc of the 27th Conf of Open Innovations Association (FRUCT). Piscataway, NJ: IEEE, 2020: 3−9

[6] Zhang Q, Nicolson A, Wang M, et al. DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1404−1415

[7] Park S R, Lee J. A fully convolutional neural network for speech enhancement[J]. arXiv preprint, arXiv: 1609.07132, 2016

[8] Pandey A, Wang Deliang. Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain[C]//Proc of the 45th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2020: 6629−6633

[9] Pandey A, Wang Deliang. Dual-path self-attention RNN for real-time speech enhancement[J]. arXiv preprint, arXiv: 2010.12713, 2020

[10] Ye Moujia, Wan Hongjie. Improved transformer-based dual-path network with amplitude and complex domain feature fusion for speech enhancement[J]. Entropy, 2023, 25(2): 228

[11] Yu Guochen, Li Andong, Zheng Chenshi, et al. Dual-branch attention-in-attention transformer for single-channel speech enhancement[C]//Proc of the 47th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2022: 7847−7851

[12] Lee J, Kang H G. Real-time neural speech enhancement based on temporal refinement network and channel-wise gating methods[J]. Digital Signal Processing, 2023, 133: 103879

[13] Kong Z, Ping W, Dantrey A, et al. Speech denoising in the waveform domain with self-attention[C]//Proc of the 47th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2022: 7867−7871

[14] Baby D, Verhulst S. Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty[C]//Proc of the 44th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2019: 106−110

[15] Hao Xiang, Su Xiangdong, Horaud R, et al. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement[C]//Proc of the 46th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2021: 6633−6637

[16] Tan K, Wang Deliang. A convolutional recurrent neural network for real-time speech enhancement[C]//Proc of INTERSPEECH 2018. Grenoble, France: ISCA, 2018: 3229−3233

[17] Kim J, El-Khamy M, Lee J. T-GSA: Transformer with Gaussian-weighted self-attention for speech enhancement[C]//Proc of the 45th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2020: 6649−6653

[18] Pandey A, Wang Deliang. Dense CNN with self-attention for time-domain speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1270-1279

[19] Erdogan H, Hershey J R, Watanabe S, et al. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks[C]//Proc of the 45th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2015: 708−712

[20] Williamson D S, Wang Yuxuan, Wang Deliang. Complex ratio masking for monaural speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 24(3): 483−492

[21] Choi H S, Kim J H, Huh J, et al. Phase-aware speech enhancement with deep complex U-NET[J]. arXiv preprint, arXiv: 1903.03107, 2019

[22] Macartney C, Weyde T. Improved speech enhancement with the wave-U-NET[J]. arXiv preprint, arXiv: 1811.11307, 2018

[23] Anmol G, Qin J, Chiu C C, et al. Conformer: Convolution-augmented transformer for speech recognition[J]. arXiv preprint, arXiv: 2005.08100, 2020

[24] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proc of the 31st Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2017: 5998−6008

[25] Liu Ze, Lin Yutong, Cao Yue, et al. Swin Transformer: Hierarchical vision transformer using shifted windows[C]//Proc of the 39th IEEE/CVF Int Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 10012−10022

[26] Koizumi Y, Harada N, Haneda Y. Trainable adaptive window switching for speech enhancement[C]//Proc of the 49th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2019: 616−620

[27] Parvathala V, Andhavarapu S, Pamisetty G, et al. Neural comb filtering using sliding window attention network for speech enhancement[J]. Circuits, Systems, and Signal Processing, 2023, 42(1): 322−343

[28] Liang Xinyan, Qian Yuhua, Guo Qian, et al. AF: An association-based fusion method for multi-modal classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(12): 9236−9254

[29] Hu Yanxin, Liu Yun, Lv Shubo, et al. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement[J]. arXiv preprint, arXiv: 2008.00264, 2020

[30] Valentini-Botinhao C, Wang Xin, Takaki S, et al. Investigating RNN-based speech enhancement methods for noise-robust text-to-speech[C]//Proc of the 9th ISCA Int Conf on Speech Synthesis Workshop. Grenoble, France: ISCA, 2016: 146−152

[31] Peer T, Gerkmann T. Phase-aware deep speech enhancement: It's all about the frame length[J]. JASA Express Letters, 2022, 2(10), 104802

[32] Wang Zhedong, Cun Xiaodong, Bao Jianming, et al. Uformer: A general U-shaped transformer for image restoration[C]//Proc of the 41st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 17683−17693

[33] Luo Yi, Chen Zhuo, Yoshioka T. Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation[C]//Proc of the 50th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2020: 46−50

[34] Thiemann J, Ito N, Vincent E. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings[J] Journal of the Acoustical Society of America, 2013, 19(1): 035081

[35] Loizou P C. Speech Enhancement: Theory and Practice[M]. Boca Raton: CRC Press, 2013

[36] Hu Yi, Loizou P C. Evaluation of objective quality measures for speech enhancement[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 16(1): 229−238

[37] Pascual S, Bonafonte A, Serra J. SEGAN: Speech enhancement generative adversarial network[J]. arXiv preprint, arXiv: 1703.09452, 2017

[38] Tsun-An Hsieh, Wang Hsin-Min, Lu Xuguang, et al. WaveCRN: An efficient convolutional recurrent neural network for end-to-end speech enhancement[J]. IEEE Signal Processing Letters, 2020, 27: 2149−2153

[39] Wang Kai, He Bengbeng, Zhu Weiping. TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain[C]//Proc of the 51st IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2021: 7098−7102

[40] Wang Ning, Ma Sihan, Li Jingyuan, et al. Multistage attention network for image inpainting[J]. Pattern Recognition, 2020, 106: 107448

[41] Fu S W, Yu Cheng, Hsieh T A, et al. MetricGAN+: An improved version of metricGAN for speech enhancement[J]. arXiv preprint, arXiv: 2104.03538, 2021

[42] Huang Xiangdong, Chen Honghong, Gan lin. Speech enhancement method based on frequency-time dilated dense network[J]. Journal of Computer Research and Development, 2023, 60(5): 1628−1638 (in Chinese)
(黄翔东, 陈红红, 甘霖. 基于频率-时间扩张密集网络的语音增强方法[J]. 计算机研究与发展, 2023, 60(5): 1628−1638)

[43] Yu Guochen, Li Andong, Zheng Chengshi, et al. Dual-branch attention-in-attention transformer for single-channel speech enhancement[C]//Proc of the 52nd IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2022: 7847−7851

[44] Yu Guochen, Li Andong, Wang Hui, 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



Zhang Chenhui, born in 1999. Master candidate. Student member of CCF. His main research interest includes speech enhancement.
张晨辉, 1999 年生. 硕士研究生. CCF 学生会员. 主要研究方向为语音增强.



Yuan Zhi'an, born in 1998. PhD candidate. Student member of CCF. His main research interests include signal enhancement and machine learning.
原之安, 1998 年生. 博士研究生. CCF 学生会员. 主要研究方向为信号增强、机器学习.



Qian Yuhua, born in 1976. PhD, professor, PhD supervisor. Member of CCF. His main research interests include artificial intelligence, big data, machine learning and data mining.
钱宇华, 1976 年生. 博士. 教授. 博士生导师, CCF 会员. 主要研究方向为人工智能、大数据、机器学习、数据挖掘.