DOI: 10.19650/j.cnki.cjsi.J2412362

双分支复频谱下多特征聚合的轻量化语音增强方法*

张天骐,沈夕文,唐 娟,谭 霜
(重庆邮电大学通信与信息工程学院 重庆 400065)

摘 要:针对目前多种改进的卷积循环网络(CRN)在单掩蔽或单映射的编解码结构下提取特征单一、捕获全局特征不强、参数量较大等问题,提出一种多特征聚合卷积模块与高效 Transformer 融合注意力机制结合的复频谱联合掩蔽和映射的单通道语音增强高效网络。在编解码层设计一种双分支门控协作单元(DGCU),提取复频谱多层次特征后交互、聚合以弥补特征提取单一问题;中间层设计一种通道时频注意力融合模块,聚焦语音的时频、空间局部细节特征。最后在 THCHS30 数据集上进行消融和对比实验,实验结果表明,该网络以最低参数量、较低计算量实现了轻量化,在匹配和不匹配噪声下 PESQ 分别提升了 10.5%~50.6%、16.3%~94.5%,客观、主观指标都优于其他对比的网络模型,表现出较高的降噪性能和网络泛化能力。

关键词: 语音增强;复频谱掩蔽和映射;多特征聚合;高效 Transformer;轻量化
中图分类号: TH701 TN912.35 **文献标识码**: A **国家标准学科分类代码**: 510.40

A lightweight speech enhancenment method based on dual branch complex spectrum with multiple feature aggregation

Zhang Tianqi, Shen Xiwen, Tang Juan, Tan Shuang
(*School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*)

Abstract: To address the issues with current variations of Convolution Recurrent Networks (CRN), which often extract limited features, capture global characteristics poorly, and have large parameter sizes under single masking or mapping encoder-decoder structures, this paper proposes an efficient single-channel speech enhancement network. This network combines a multi-feature aggregation convolution module, leveraging complex spectrum joint masking and mapping, with an efficient Transformer-based attention mechanism. In the encoder-decoder layer, a Dual-branch Gated Cooperative Unit (DGCU) is designed to interact and aggregate multi-level complex spectral features, addressing the problem of singular feature extraction. The intermediate layer incorporates a Channel-Time-Frequency Attention Fusion Module, focusing on spatial and time-frequency local detail features of speech. Ablation and comparative experiments on the THCHS30 dataset demonstrate that this network achieves lightweight efficiency with the lowest parameter count and relatively low computational cost. It improves PESQ by 10.5% ~ 50.6% and 16.3% ~ 94.5% under matched and mismatched noise conditions, respectively. Both objective and subjective metrics outperform other comparative network models, exhibiting superior noise reduction performance and network generalization capability.

Keywords: speech enhancement; complex spectral mapping and masking; multiple feature aggregation; efficient Transformer; lightweight

0 引 言

在现实生活中,人们经常会遭受各种各样的噪声干扰,如环境噪声、电磁噪声等,这些噪声干扰导致语音信号质量下降,降低了人们对语音信息的理解和辨别能力。语音增强技术旨在通过消除混合的噪声,恢复高质量和高可懂度的语音信号。这不仅提升了用户体验感和语音通信可靠性,也推动了语音识别、视频会议、自动驾驶等应用领域的发展。

收稿日期:2024-01-07 Received Date: 2024-01-07
*基金项目:重庆市自然科学基金(cstc2021jcyjmsxmX0836)项目资助

在深度学习领域中,语音增强在处理域的估计方法主要是时域估计[1-2]和时频域估计[3-6]。基于时域估计方法是将含噪语音波形直接输入到模型中训练以学习纯净语音特征。而基于时频域估计方法首先对含噪语音波形进行短时傅里叶变换(short time fourier transform,STFT),得到含噪语音的语谱图,然后输入到模型中训练以估计语音的复频谱[7-11]、幅度谱[12]和相位谱[13-14]等方式来学习纯净语音特征。早期的时频域估计方法是增强幅度谱,通过将含噪语音相位谱和增强后语音的幅度谱相结合来恢复纯净语音,如卷积循环网络(convolution recurrent network,CRN)[12],虽然含噪语音相位和增强语音的不匹配对语音可懂度没有较大影响,但会严重降低语音的听感音质。因此 Tan 等[4]在 CRN 基础上提出复卷积循环网络(complex convolution recurrent network,CCRN),其使用双解码的复频谱映射,估计语音的实虚部特征来增强对相位的隐式估计,增强了模型的性能。随后,又提出门控卷积循环网络(gate convolution recurrent network,GCRN)[5],在 CCRN 基础上将普通卷积替换为门控卷积单元,输出端添加线性层映射实虚部,进而提升了语音的质量。后来,深度复卷积循环网络(deep complex convolution recurrent network,DCCRN)[6]通过使用(deep complex U-network,DCUNET)[15]网络结构中的复卷积运算,再一次提升了 GCRN 模型的性能上限。然而由于上述模型中巨大的卷积核尺寸、数量和复杂的复卷积运算,使得模型的参数量较高。

语音增强方法也分为掩蔽[16]和映射[17-18]方法。映射方法是直接估计纯净语音,掩蔽方法是预测纯净语音的掩蔽值,然后与含噪语音相乘得到增强语音,如理想比率掩蔽(ideal ratio masking, IRM)[19]。

然而上述 CRN、CCRN、GCRN、DCUNET、DCCRN 等网络都是基于映射或掩蔽的方法,存在参数量巨大,对语音序列的长时间依赖性较高,特征提取单一,捕获上下文依赖性不强等缺点。

早期的 Transformer[20-21]模型在计算机视觉领域大放光彩,近年来 Transformer 应用到语音处理领域,对语音长序列上下文建模能力较强。TSTNN[22]使用编解码结构,中间层加入双路径 Transformer 结构和掩蔽模块预测时间序列特征;CAUNET[23]在 UNet 结构基础上在中间层加入双路径 Transformer 结构,对时域语音信号的长范围序列的局部和全局信息建模,性能得到一定的提升,但这两种双路径的 Transformer 只是对于时域上特征的研究,且存在提取局部特征较弱、空间维度信息提取能力不强,计算量大等缺点。

为解决上述问题,本文提出双分支复频谱联合掩蔽与映射的编解码网络模型,使用通道注意力、改进的时频注意力融合 Transformer 对语音长序列特征、多维度信息进行建模,通过估计语音的实虚部分量来恢复语音信号,以更低的参数量和相当的计算量实现更优的性能。在模型编码层、解码层网络设计了一种双分支门控协作单元(dual-branch gated collaboration unit,DGCU),通过两个支路的不同卷积核尺寸提取双支路多特征信息,双支路协作后能更好学习潜在的细节特征。中间层网络则首先由通道注意力机制获取通道维度的空间上下文特征,再输入到改进的时/频注意力融合 Transformer(improved time/frequency attention fusion transformer,IT/FAT)中,对语音的时间和频率信息进行建模。最后,在由清华大学创建的 THCHS30 语料库与 118 种噪声混合下的数据集上,验证本文提出的网络以更加轻量化的特点实现最高的语音增强效果,表明本文提出的网络的更具鲁棒性和有效性。

1 语音增强流程

设混入了加性噪声的含噪语音信号模型为:

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

其中,$\boldsymbol{y}$、$\boldsymbol{s}$、$\boldsymbol{n} \in \mathrm{R}^{1\times L}$ 分别是含噪语音、纯净语音和噪声的时域信号表示,L 为时域波形的长度。式(1)中时域波形经过短时傅里叶变换(STFT)得到复频域信号:

$$\boldsymbol{Y} = \boldsymbol{Y}_{\mathrm{r}} + \mathrm{j}\boldsymbol{Y}_{\mathrm{i}} = (\boldsymbol{S}_{\mathrm{r}} + \boldsymbol{N}_{\mathrm{r}}) + \mathrm{j}(\boldsymbol{S}_{\mathrm{i}} + \boldsymbol{N}_{\mathrm{i}}) \tag{2}$$

其中,$\boldsymbol{Y}_{\mathrm{r}}$、$\boldsymbol{Y}_{\mathrm{i}}$、$\boldsymbol{S}_{\mathrm{r}}$、$\boldsymbol{S}_{\mathrm{i}}$、$\boldsymbol{N}_{\mathrm{r}}$、$\boldsymbol{N}_{\mathrm{i}} \in R^{1\times T\times F}$ 分别表示含噪语音、纯净语音、噪声的复频谱的实部和虚部,T 是时间帧,F 是频率帧。如图 1 所示,本文提出的网络模型采用编解码结构,将含噪语音 $\boldsymbol{Y}$ 的实部 $\boldsymbol{Y}_{\mathrm{r}}$ 和虚部 $\boldsymbol{Y}_{\mathrm{i}}$ 输入到编码层,中间层为信息传输层,解码层为复频谱掩蔽 + 映射的双分支结构。上分支通过基于复频谱掩蔽恢复出的增强语音实虚部掩蔽为 $\boldsymbol{M} = \mathrm{cat}(\tilde{\boldsymbol{S}}_{\mathrm{r}}, \tilde{\boldsymbol{S}}_{\mathrm{i}}) \in R^{2\times T\times F}$,其与含噪语音的复频谱按位相乘得到上分支增强语音 $\tilde{\boldsymbol{S}}_{\mathrm{up}}$,如式(3)所示。下分支通过基于复频谱映射恢复出的增强语音为 $\tilde{\boldsymbol{S}}_{\mathrm{down}} = \mathrm{cat}(\tilde{\boldsymbol{S}}_{\mathrm{r}}, \tilde{\boldsymbol{S}}_{\mathrm{i}}) \in R^{2\times T\times F}$,下分支增强语音在训练过程中隐式地保留了相位信息,能更好重建语音信号。其中 cat(·) 表示将实部、虚部在通道维度的拼接,上分支及模型最终输出的增强语音如下:

$$\tilde{\boldsymbol{S}}_{\mathrm{up}} = \boldsymbol{M} \otimes \boldsymbol{Y} \tag{3}$$

$$\tilde{\boldsymbol{S}} = \alpha\tilde{\boldsymbol{S}}_{\mathrm{up}} \oplus \beta\tilde{\boldsymbol{S}}_{\mathrm{down}} \tag{4}$$

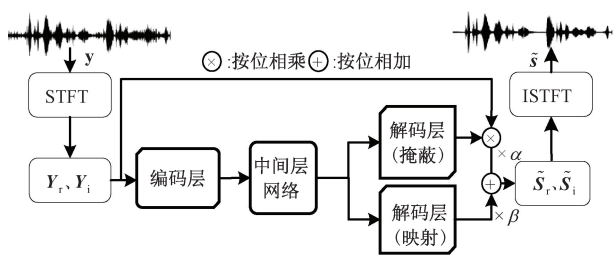


图 1 语音增强流程图

Fig. 1 Flowchart of speech enhancement

其中,$\boldsymbol{M}$、$\boldsymbol{Y}$ 分表表示解码层上分支掩蔽值和含噪语音,α、β 为可学习参数,初始值均设置为 0.5。⊕、⊗表示按位相加和按位相乘。最后将输出的增强复频谱信号进行短时傅里叶逆变换(ISTFT)得到增强后的时域语音波形 $\tilde{\boldsymbol{s}}$:

$$\tilde{\boldsymbol{s}} = \mathrm{ISTFT}(\tilde{\boldsymbol{S}}) \tag{5}$$

2　系统结构

2.1　双分支门控协作单元

由于在单支路上堆叠卷积会提取过多的局部冗余特征,受门控线性单元[24]和 Yin 等[13]提出双向信息交互的 DNN 结构的启发,本文在编、解码层提出了 DGCU 来替代传统编解码网络的卷积层和密集卷积层,DGCU 通过不同感受野的卷积核、双分支信息协作学习和门控机制,可以更好地关注双支路的局部特征中隐藏的规律,交互学习不同特征并控制信息的产生。

编码层的双分支门控协作单元(double branch gate control unit of encoding layer,DGCU_En)如图 2(a)所示,输入特征为 $\boldsymbol{Y} \in \mathrm{R}^{B\times C\times T\times F}$,其中 B 为批次数,C 为通道数,T 为帧数,F 为帧大小。首先通过通道数为 $C/2$,步长为 1 × 2 的卷积减半通道数和帧大小,减少参数量和训练成本;然后输入到双支路中,两个支路分别使用卷积核尺寸为 2 × 3、2 × 5 的卷积以在不同尺度上提取多个角度的语音局部细微特征,从而创建更加丰富和多样化的特征表示;再将两个支路信息在通道维度拼接聚合,随后使用 1×1 的卷积核还原通道数并捕获潜在特征,输出端通过 sigmoid 激活函数后分别与两个不同卷积核尺寸的卷积输出相乘,以进行多层次特征的信息交互;最后将两分支的特征逐点相加以进行信息补偿,相加后的信息输入到 1×1 的二维卷积中来还原通道数,并经过层归一化和参数修正线性单元(parametric rectified linear unit,PRelu)激活函数来加速网络收敛,得到输出特征 $Y' \in R^{B\times C\times T\times F/2}$。

解码层的双分支门控协作单元(double branch gate control unit of decoding layer,DGCU_De)结构与编码层类似,如图 2(b)所示。不同的是,为保持编解码层对特征维度处理的对称性,每层解码层的输出端均采用反卷积以逐层恢复原始输入信号频率维度数。值得注意的是,解码层的上分支末层(掩蔽端)输出使用 Tanh 激活函数估计纯净语音的掩膜。DGCU 在双分支交互协作学习下,可以更容易学习深层次空间信息和辨别潜在特征,提升编解码层网络对复频谱多层次特征的捕捉能力,从而增加网络的表达能力。

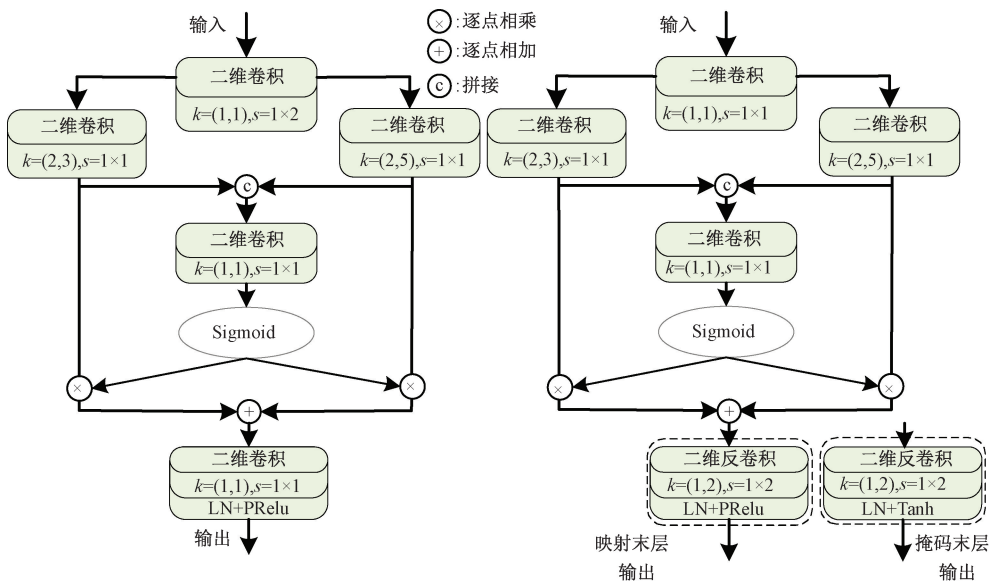


图 2　编码层和解码层中的双分支门控协作单元

Fig. 2　Dual-branch gated collaboration in encoder and decoder

2.2　改进的时/频注意力融合 Transformer

Vaswani 等[25]提出 Transformer 模型中的编码器由位置编码、多头注意力、前馈网络构成。由于位置编码不适合声学序列,TSTNN[21]在模型设计中放弃了位置编码的使用,并将前馈网络中的第 1 个线性层替换为门控循环单元(gated recurrent unit,GRU),以实现对忽略的位置信息进行学习。因此本文提出的改进的 IT/FAT 只保留了 Transformer 模型中的多头注意力和前馈网络,并在多头注意力后增加了时/频特征增强模块,且将前馈网络中的第 1 个线性层改为深度前馈顺序记忆网络(deep feed-

forward sequential memory networks, DFSMN), 以更好学习声学序列的位置信息和捕获长期序列信息, DFSMN 详细介绍见 2.3 节。

受 CAUNET[23] 提出的时域双路径全局和局部 Transformer 的启发, 本文使用 IT/FAT 对时间和频率维度的细粒特征进行建模。具体来说, 中间层网络有 N 个 IT/FAT 模块, 每个 IT/FAT 模块由改进的时间注意力融合 Transformer (improved time attention fusion transformer, ITAT) 和改进的频率注意力融合 Transformer (improved frequency attention fusion transformer, IFAT) 构成, 模块的输入和输出添加了残差连接以避免梯度消失, ITFA 与 TFTA 的结构相同只是处理的维度不同, 如图 3 所示。

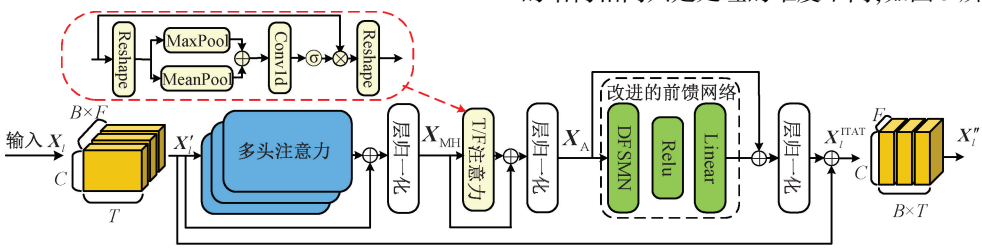


图 3 改进的时/频注意力融合 Transformer

Fig. 3 Improved time/frequency attention fusion Transformer

设中间层中第 l 层 ITAT 的输入特征为 $\boldsymbol{X}_l \in R^{B\times C\times T\times F}$ 且 $l \in [1,2,\cdots,N]$, 首先变形 (Reshape) 为 $\boldsymbol{X'}_l \in R^{(B\times F)\times C\times T}$ 输入到 ITAT 中对时间信息建模, 将得到的输出值变形为 $\boldsymbol{X}''_l \in R^{(B\times T)\times C\times F}$ 再输入到 IFAT 中对频率信息建模, 对时间和频率信息建模的具体处理步骤为:

$$\boldsymbol{X}_l^{\mathrm{ITAT}} = f^{\mathrm{ITAT}}(\boldsymbol{X}'_l[:,:,i], i = 1,2,\cdots,T) \tag{6}$$

$$\boldsymbol{X}''_l = \mathrm{Reshape}(\boldsymbol{X}_l^{\mathrm{ITAT}}) \tag{7}$$

$$\boldsymbol{X}_l^{\mathrm{IFAT}} = f^{\mathrm{IFAT}}(\boldsymbol{X}''_l[:,:,j], j = 1,2,\cdots,F) \tag{8}$$

其中, $f^{\mathrm{ITAT}}(\cdot)$ 、$f^{\mathrm{IFAT}}(\cdot)$ 分别表示 ITAT 和 IFAT 模块的映射函数, $\boldsymbol{X}'_l[:,:,i]$ 表示第 i 时间步长下的频率特征序列, $\boldsymbol{X}''_l[:,:,j]$ 表示第 j 频率特征在所有时间步长下的序列。

式(9)的具体操作步骤是首先将 $\boldsymbol{X}'_l$ 输入到改进的时间注意力 Transformer(ITAT) 中, 对于多头注意力和层归一化处理操作, 表达式如下:

$$[\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i] = [\boldsymbol{W}_i^{\mathrm{Q}},\boldsymbol{W}_i^{\mathrm{K}},\boldsymbol{W}_i^{\mathrm{V}}]\boldsymbol{X}'_l \tag{9}$$

$$\boldsymbol{head}_i = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{\mathrm{d}}}\right)\boldsymbol{V}_i \tag{10}$$

$$\boldsymbol{MH} = \mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Concat}(\boldsymbol{head}_1,\cdots,\boldsymbol{head}_{\mathrm{k}})\boldsymbol{W}^{\mathrm{o}} \tag{11}$$

$$\boldsymbol{X}_{\mathrm{MH}} = \mathrm{LN}(\boldsymbol{MH} + \boldsymbol{X}'_l) \tag{12}$$

其中, $i \in [1,2,3\cdots,k]$, $\boldsymbol{W}_i^{\mathrm{Q}}$、$\boldsymbol{W}_i^{\mathrm{K}}$、$\boldsymbol{W}_i^{\mathrm{V}}$ 分别为 $\boldsymbol{Q}_i$、$\boldsymbol{K}_i$、$\boldsymbol{V}_i$ 的第 i 层参数矩阵, 且维度为 $B\times C\times C$, Concat(·) 表示对每层输出 $\boldsymbol{head}_i$ 级联操作。然后输入到 T/F 注意力中, 对 $\boldsymbol{X}_{\mathrm{MH}}$ 的时间维度进行最大池化(max pooling, MP) 和平均池化(average pooling, AP) 提取关键特征, 在通道维度拼接后经过卷积和 sigmoid 激活函数操作, 并与多头注意力的层归一化后的输出 $\boldsymbol{X}_{\mathrm{MH}}$ 进行逐点相乘、层归一化后得到输出 $\boldsymbol{X}_{\mathrm{A}}$:

$$\boldsymbol{X}'_{\mathrm{A}} = \sigma\{f^{1\times1}\{\mathrm{Cat}[\mathrm{MP}(\boldsymbol{X}_{\mathrm{MH}}),\mathrm{AP}(\boldsymbol{X}_{\mathrm{MH}})]\}\} \tag{13}$$

$$\boldsymbol{X}_{\mathrm{A}} = \mathrm{LN}(\boldsymbol{X'}_{\mathrm{A}} \otimes \boldsymbol{X}_{\mathrm{MH}} + \boldsymbol{X}_{\mathrm{MH}}) \tag{14}$$

其中, σ 表示 sigmoid 激活函数, $f^{1\times1}(\cdot)$ 表示卷积核尺寸为 1 的一维卷积, Cat(·) 表示拼接, $\otimes$ 表示逐点相乘。最后经过改进的前馈网络、层归一化和残差连接, 改进的前馈网络包括 DFSMN、ReLU 激活函数和线性层, 得到 ITAT 的最终输出:

$$\boldsymbol{X}_l^{\mathrm{ITAT}} = \mathrm{LN}[\mathrm{IFFN}(\boldsymbol{X}_{\mathrm{A}}) + \boldsymbol{X}_{\mathrm{A}}] + \boldsymbol{X'}_l \tag{15}$$

其中, IFFN(·) 表示改进的前馈网络处理操作。将 ITAT 输出值 $\boldsymbol{X}_l^{\mathrm{ITAT}}$ 的维度变为 $(B\times \mathrm{T})\times \mathrm{C}\times \mathrm{F}$ 后再输入到 IFAT 对频率维度进行处理。

2.3 深度前馈顺序记忆网络

深度前馈顺序记忆网络(deep feedforward sequential memory networks, DFSMN)[26] 是对 LSTM 的改进, 高效提升性能的同时还减少近 3/4 参数量。DFSMN 在隐藏层加入记忆单元, 记忆单元可以对隐藏层中的前后单元进行编码以更好捕捉上下文序列信息, 学习时间和频率的长序列相关性。其流程图如图 4 所示。

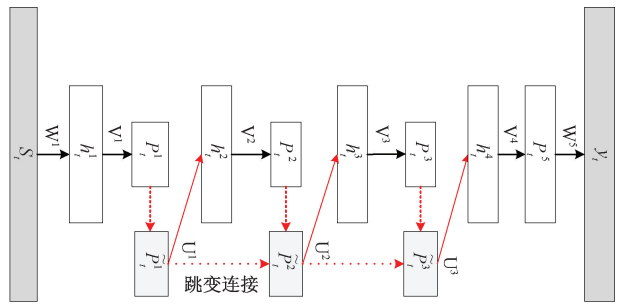


图 4 深度前馈顺序记忆网络

Fig. 4 Deep Feedforward Sequential Memory Networks

在本文中, 令输入到 DFSMN 的特征为 $\boldsymbol{Y} \in R^{T\times(B\times F)\times C}$, 首先变化为 $\boldsymbol{s}_t \in R^{(B\times F)\times C\times T}$, 对于每个 t 时

刻的序列 $\boldsymbol{s}_t \in R^{(B\times F)\times C}$，经过一系列过程如下：

$$\boldsymbol{p}_t^l = \boldsymbol{V}^l(\boldsymbol{W}^l \boldsymbol{s}_t^{\,l} + \boldsymbol{b}^l) + \boldsymbol{b}^l \tag{16}$$

$$\tilde{\boldsymbol{p}}_t^l = \boldsymbol{H}(\tilde{\boldsymbol{p}}_t^{l-1}) + \boldsymbol{p}_t^l + \sum_{i=0}^{N_L} \boldsymbol{a}_i^l \odot \boldsymbol{p}_{t-i}^l + \sum_{j=1}^{N_R} \boldsymbol{c}_j^l \odot \boldsymbol{p}_{t+j}^l \tag{17}$$

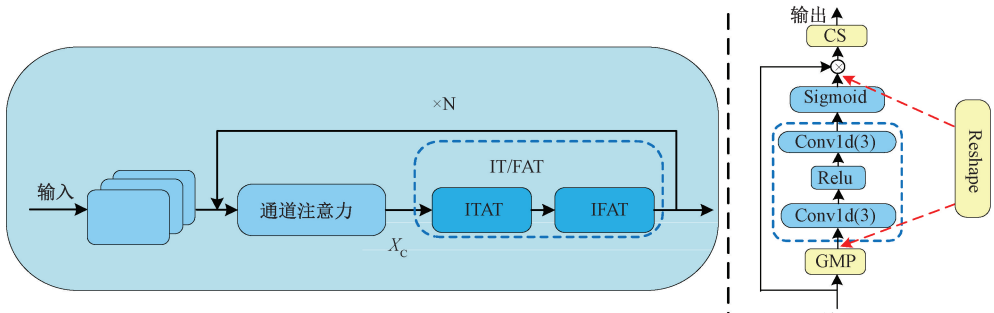
其中，$\boldsymbol{p}_t^l$ 表示第 l 层的线性映射，H(·) 表示跳变连接符号，$\tilde{\boldsymbol{p}}_t^{l-1}$ 表示第 $l-1$ 层计算后得到的值，N_R、N_L 分别表示前向、后向编码的单元数目，⊙ 表示逐元素点乘。

然后通过下式计算下一层隐藏单元的激活值：

$$\boldsymbol{h}_t^{l+1} = f(\boldsymbol{U}^l \tilde{\boldsymbol{p}}_t^{\,l} + \boldsymbol{b}^{l+1}) \tag{18}$$

2.4　通道时频注意力融合模块

由于多层卷积后的特征在通道维度上包含了大部分的特征信息，而本文提出的 IT/FAT 只捕获了语音时频空间特征却忽略了通道维度的潜在特征。因此在 IT/FAT 前引入了通道注意力，提出的通道时频注意力融合模块如图 5(a) 所示。通道注意力如图 5(b)，首先通过全局最大池化，将输入维度 $B\times C\times T\times F$ 压缩为 $B\times C\times 1\times 1$ 并重塑(Reshape)维度为 $B\times 1\times C$，再通过卷积核尺寸为 3 的一维卷积、Relu 激活函数、一维卷积、sigmoid 激活函数和 Reshape 操作后与输入信号相乘以聚合通道中的关键特征，最后进行通道混洗(channel shuffle，CS)操作使通道间信息进行交互学习和避免多层卷积扰乱通道的相关性。



(a) 通道时频注意力融合模块
(a) Channel time frequency attention fusion block

(b) 通道注意力
(b) Channel attention

图 5　通道时频注意力融合模块和通道注意力

Fig. 5　Channel time frequency attention fusion block and channel attention

通道注意力输出的特征 $\boldsymbol{X}_C$ 依次输入到串行的 ITAT 和 IFAT 模块中，得到输出 $\boldsymbol{O}$：

$$\boldsymbol{O} = \boldsymbol{O}_F[\boldsymbol{O}_T(\boldsymbol{X}_C)] \tag{19}$$

其中，$\boldsymbol{O}_T(\cdot)$、$\boldsymbol{O}_F(\cdot)$ 分别为 ITAT 和 IFAT 模块处理。

2.5　总体网络结构

总体网络结构如图 6 所示，由编码层、中间层、双分支(掩蔽+映射)解码层组成。

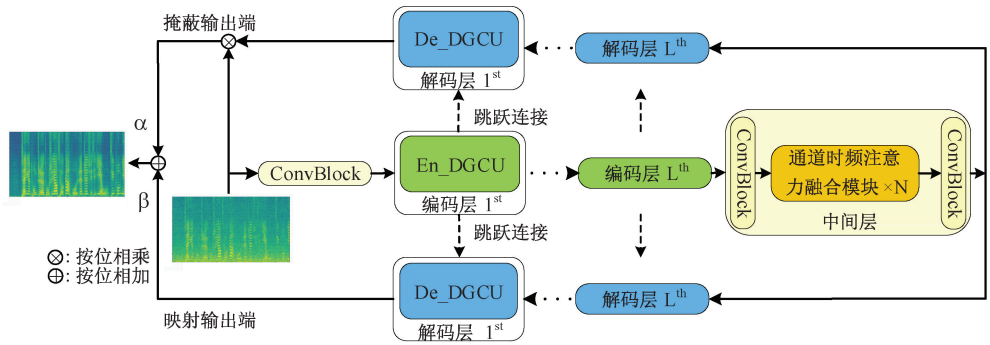


图 6　总体网络结构图

Fig. 6　Overall network structure

语音特征首先在网络输入端由卷积模块(ConvBlock)将通道数由 2 扩展到 64。编码层使用 4 层双分支门控协作单元提取语音的复频谱特征，通过逐层减半频率特征以降低参数量。中间层首先通过 ConvBlock 使通道数减半至 32，增强网络的非线性能力并聚合特征，然后输入到通道时频注意力融合模块中学习并传递语音关键特征，最后经过 ConvBlock 还原通道数至 64。所有的 ConvBlock 模块包含卷积核尺寸为

1 × 1 的二维卷积、层归一化和 PRelu 激活函数,以加速网络收敛。中间层中的 DFSMN 设置为 2 层,每层包含 64 个隐藏单元。解码层的上下分支分别使用 4 个双分支门控协作单元,每一层接收来自对应编码层的跳变信息,将高分辨率信息与低分辨率信息融合,重建编码层压缩所丢失的信息,使得网络具有更高的鲁棒性和特征精度。上下分支末层的输出端均将通道数还原为 2,两分支的加权求和,得到增强后的复频谱纯净语音信号。

表 1 为总体网络模型参数的详细说明,c、s、k 分别为卷积核的输出通道数、步长和尺寸,每层结构的输入、输出维度的格式为 $C \times T \times F$。

表 1 总体网络模型的参数设置

Table 1 Paramters setting of overall network model

网络总结构	模块	参数设置	输入维度	输出维度
编码层	ConvBlock	$k=1\times1, c=64$	$2\times T\times512$	$64\times T\times512$
	DGCU_En_1	$s=1\times2, c=64$	$64\times T\times512$	$64\times T\times256$
	DGCU_En_2	$s=1\times2, c=64$	$64\times T\times256$	$64\times T\times128$
	DGCU_En_3	$s=1\times2, c=64$	$64\times T\times128$	$64\times T\times64$
	DGCU_En_4	$s=1\times2, c=64$	$64\times T\times64$	$64\times T\times32$
中间层	ConvBlock	$k=1\times1, s=1\times1, c=32$	$64\times T\times32$	$32\times T\times32$
	通道时频注意力融合模块	-	$32\times T\times32$	$32\times T\times32$
	ConvBlock	$k=1\times1, s=1\times1, c=64$	$32\times T\times32$	$64\times T\times32$
解码层(映射+掩蔽)	DGCU_De_4(×2)	$s=1\times2, c=64$	$(64+64)\times T\times32$	$64\times T\times64$
	DGCU_De_3(×2)	$s=1\times2, c=64$	$(64+64)\times T\times64$	$64\times T\times128$
	DGCU_De_2(×2)	$s=1\times2, c=64$	$(64+64)\times T\times128$	$64\times T\times256$
	DGCU_De_1(×2)	$s=1\times2, c=2$	$(64+64)\times T\times256$	$2\times T\times512$

2.6 损失函数

时域损失函数能够更好平衡 PESQ 和其他指标的性能,为了更精确地恢复纯净语音复频谱的实部和虚部,本文在时间域和频率域联合了时域损失函数、复频域损失和幅度损失函数,并且使用 Li 等[27]提出的幂律压缩来压缩幅值和和保留隐藏的相位信息。压缩后的复频谱为:

$$\boldsymbol{S}^{c} = |\boldsymbol{S}|^{p} e^{i\theta} \tag{20}$$

式(20)中压缩系数 p 为 0.3,使用幂律压缩后的估计语音复频谱来计算训练损失。复数域损失 L_{RI} 是分别对压缩后的纯净语音与增强语音的实部和虚部分量使用最小均方误差损失函数(mean square error,MSE);幅度损失 L_{mag} 是对压缩后纯净语音与增强语音的幅值使用 MSE。时域损失 L_{time} 是对纯净语音和增强语音的时域波形使用最小绝对值误差损失函数(mean absolute error,MAE);3 种损失及总损失 L 的定义如下:

$$\begin{cases} L_{\mathrm{time}} = \mathrm{MAE}(|\tilde{\boldsymbol{s}} - \boldsymbol{s}|) \\ L_{\mathrm{RI}} = \mathrm{MSE}(\tilde{\boldsymbol{S}}_{\mathrm{r}}^{c} - \boldsymbol{S}_{\mathrm{r}}^{c}) + MSE(\tilde{\boldsymbol{S}}_{\mathrm{i}}^{c} - \boldsymbol{S}_{\mathrm{i}}^{c}) \\ L_{\mathrm{mag}} = \mathrm{MSE}(|\tilde{\boldsymbol{S}}|^{p} - |\boldsymbol{S}|^{p}) \\ L = \gamma_1 L_{\mathrm{time}} + \gamma_2 L_{\mathrm{RI}} + L_{\mathrm{mag}} \end{cases} \tag{21}$$

其中,$|\tilde{\boldsymbol{S}}|^{p}$、$\tilde{\boldsymbol{S}}_{r}^{c}$、$\tilde{\boldsymbol{S}}_{\mathrm{i}}^{c}$、$|\boldsymbol{S}|^{p}$、$\boldsymbol{S}_{\mathrm{r}}^{c}$、$\boldsymbol{S}_{\mathrm{i}}^{c}$ 分别表示幂律压缩后增强语音和纯净语音复频谱的幅值、实部和虚部,γ_1、γ_2 为平衡时间损失与复值损失的平衡因子,根据经验分别设置 0.1、0.2。

3 实验设置及结果对比分析

3.1 数据集设置

为多方面的验证本文网络的语音增强效果,考虑到不同信噪比、不同噪声种类以及具体应用场景下的泛化能力等影响因素,本文选取不同噪声数据集的多种日常生活中常见噪声,用于训练和测试阶段,具体数据集的构成如下:

THCHS30 数据集:随机抽取中文 THCHS30 语料库中的 2 500 条纯净语音用于训练和验证、400 条语音用于测试。数据集中混合的日常生活中常见噪声包括 100 种 Nonspeech 的非语言环境噪声、Noise92 中的 Babble、factory 等噪声、Demand 中 Bus、Office 等噪声,共 28 种噪声。在训练和验证过程中随机选取其中的 21 种噪声,随机截取噪声并与纯净语音等长度,将截取后的噪声与纯净语音在信噪比为-5~10 dB 范围内,以 1 dB 为步长随机混合以生成训练和验证数据集;在测试过程中选取了训练集中的 8 种噪声作为匹配噪声来评估模型的增强效果,余下未训练过的 7 种噪声作为不匹配噪

声来测试模型的泛化能力,并分别在信噪比为-5、0、5、10 dB 条件下混合以生成测试数据集,测试集中两种类型噪声均模拟了日常生活中的居家、休闲活动、出行、办公等不断变化的环境噪声,如洗衣机滚筒噪声、公园大自然噪声、汽车行驶噪声、飞机起降噪声、公交站车流声、办公区过道噪声、工厂机器轰鸣声等。表 2 为噪声数据集的详细划分,最终训练集、验证集共产生 40 000 对含噪-纯净语音,约 45 小时,其中训练时长约 41 小时、验证时长约 4 小时,测试集产生 3 200 对含噪-纯净语音,时长约 4 小时。

表 2　噪声类型

Table 2　Noise types

类型	噪声
训练噪声	Babble、White、Buccaneer1、F16、M109、Factory1、Bus、Kitchen、Livingroom、Meeting、Park、River、Office、Traffic、Volvo、Animal、Wind、Laugh、Bell、Machine、Street
匹配噪声	Buccaneer1、White、Factory1、M109、Meeting、River、Traffic、Park
不匹配噪声	Car、Factory2、Hfchannel、Pink、Station、washing、Hallway

3.2　参数设置及评价指标

所有语音信号均采样到 16 kHz,帧长为 63.937 5 ms,帧移为 16 ms。解码层上下分支输出端的可学习因子 α、β 均初始化为 0.5。实验中的批次数(batch_size)设置为 2,轮次数(Epoch)设置为 100,网络模型采用 Adam 优化器进行训练,前 30 轮的初始学习率设置为 0.000 5 保持不变,30 轮后若验证集损失连续 2 个训练轮次不减少,学习率减半,若连续 5 个 Epoch 不降低,则停止训练。本文实验及对比的所有实验在 RTX3090 的 GPU 上,基于 Pytorch1.9、CUDA11.1、python3.8 等开发环境搭建的模型下训练。

本文实验在客观指标和主观评价指标上评估语音质量,客观评估指标:语音质量的感知评估(perceptual evaluation of speech quality,PESQ),指标范围从-0.5~4.5,值越大语音听觉感受越好;短时客观可懂度(short-time objective intelligibility,STOI)[28],指标范围从 0 到 1,分数越高语音越易理解;对数谱距离(log spectral distance,LSD)评估增强后语音的失真程度,值越低语音的失真程度越小。主观评估指标(subjective mean opinion scores,SMOS)[29]:CSIG 用于信号失真评估、CBAK 用于噪声失真评估、COVL 用于总体质量评估,各指标得分越高语音质量越好,所有的 SMOS 得分在 1~5 范围内。

3.3　消融实验

首先在 THCHS30 数据集上纵向对比实验,对本文提出网络模型进行消融实验,表 3 为探究不同编解码层、不同中间层个数对 PESQ 和 STOI(%)的影响(PESQ 和 STOI(%)是在 7 种不匹配噪声与 4 种信噪比条件下的均值)。由表 3 的对比结果可知,当编解码层数为 4,中间层数为 4 时网络的 PESQ 最高,虽然 STOI 稍差于中间层为 5 的网络,但参数量降低了约 10%,故编码层数为 4 层,中间层网络为 4 层被设置为本网络的最优配置,后续对比实验均使用此配置。

表 3　不匹配噪声下不同网络配置对语音增强效果对比

Table 3　Comparison of different network configuration on speech enhancement effect under unmatched noise

编/解码层数	中间层数	PESQ	STOI/%	参数量/M
3	4	2.51	87.5	0.32
3	5	2.48	87.7	0.36
4	3	2.50	88.3	0.33
4	4	**2.62**	88.9	**0.37**
4	5	2.53	**90.0**	0.41

表 4 为本网络模型中不同模块组合的消融实验对比,其中 GConv 表示使用 GCRN[7] 中的门控卷积单元,ITransformer 表示 IT/FAT 中无 T/F 注意力,IT/FAT 和 DGCU 分别是本文提出的改进的时频注意力融合 transformer 和双分支门控卷积单元。id1 表示编解码层使用门控卷积单元(GConv)和中间层使用 Itransformer;id2 是 id1 基础上在 Transformer 中加入 T/F 注意力;id3 是 id1 基础上将 GConv 替换为 DGCU;id4 是 id3 基础上在 Transformer 中加入 T/F 注意力;id5 是使用 DGCU、IT/FAT 和通道注意力的网络。从实验结果可以看出本文提出的 DGCU、T/F 注意力、IT/FAT 等模型都提升了网络的性能,此外,加入通道注意力的模型(Propose)能够进一步提升网络的性能上限,PESQ、STOI(%)分别提升 0.06 和 0.5%。

表 4　不匹配噪声下不同网络模型消融实验对比

Table 4　Comparison of ablation experiments using different network models under unmatched noise

模型	id	PESQ	STOI(%)	CSIG	CBAK	COVL
GConv+ITransformer	1	2.39	87.4	3.61	2.77	2.94
GConv+IT/FAT	2	2.43	87.7	3.66	2.83	3.00
DGCU+ITransformer	3	2.48	88.2	3.75	2.99	3.11
DGCU+IT/FAT	4	2.55	88.4	3.84	3.05	3.20
Propose	5	**2.62**	**88.9**	**3.88**	**3.09**	**3.25**

表 5 为对比不同输出方法对 PESQ、STOI(%)、CSIG、CBAK、COVL 等指标(7 种不匹配噪声与 4 种信噪比条件下平均值)的影响,对比的输出方法有:单支路解码层基于掩蔽(单掩蔽)、单支路解码层基于映射(单映射)、双支路解码层基于映射的可学习参数相加(双映射)、双支路解码层基于映射+掩蔽的常规相加(平均相加)和双支路解码层基于映射+掩蔽的可学习参数相加(联合加权)。在 THCHS30 数据集上测试的结果表明,联合映射+掩蔽的方法可以有效提升单掩蔽或单映射性能,与此同时,联合掩蔽+映射也可以提升双映射的性能上限,虽然 CSIG 得分略低于双映射方法,但总体上各性能优于双映射。另外,通过使用可学习参数联合加权求和,再次将基于映射+掩蔽的常规相加模型的 PESQ 和 STOI 分别提升了 0.07 和 0.4%。

表 5 不匹配噪声下不同语音增强输出方法的实验性能对比

Table 5 Performance comparison of different speech enhancement output methods under unmatched noise

模型	PESQ	STOI(%)	CSIG	CBAK	COVL
单掩蔽	2.54	88.1	3.83	2.98	3.09
单映射	2.52	88.2	3.86	2.96	3.11
双映射	2.59	88.6	**3.91**	3.06	3.22
平均相加	2.54	88.5	3.76	3.02	3.15
联合加权(本文)	**2.62**	**88.9**	3.88	**3.09**	**3.25**

3.4 对比实验

为验证本文网络性能的优越性和泛化能力,再进行横向对比实验,对比的网络模型为 6 种在时频域处理的先进语音增强模型,分别为 DCUNET[15]、RCED[30]、CCRN、GCRN、DCCRN、CAUNET[23],其中 Propose-L 是表 3 中编解码层数为 3,中间层数为 4 的模型配置。DCUNET 是使用复卷积运算的编解码结构网络,RCED 是一种经典的复频域处理的编解码网络,CAUNET 是基于时域处理且使用双路径 Transformer 的编解码结构网络。所有对比网络的参数和结构与原论文设置一致,在训练时的学习率、损失函数等模型参数与本文方法保持一致,且所有网络的最优模型均达到收敛状态。

表 6、表 7 分别为各网络在匹配噪声和不匹配噪声下的不同信噪比的平均 PESQ 和 STOI(%)得分,其中 Mask、Mapping 分别表示基于掩蔽和基于映射的方法。由表 6 可知,匹配噪声下只使用卷积处理的 RCED 网络降噪能力有限;加入 LSTM 的 CCRN 使得全卷积网络的增强效果得到了提升;使用复卷积运算的 DCUNET 进一步提升了网络性能;GCRN 在 CCRN 基础上使用门控卷积单元,控制网络的信息传输,提取到的特征更加丰富,可懂度与语音质量更高;DCCRN 相比 GCRN 取得更高的性能,原因可能在于使用了复卷积运算,增加了复频谱实虚部之间信息的交流。CAUNET 使用双路径 Transfomer,在语音长序列建模的原因下使得语音的恢复度较好,PESQ 和 STOI 指标更接近本文网络的指标得分。对比上述所有基于复频谱和幅度谱的网络,本文网络在所有信噪比下的 PESQ、STOI 均取得了最高的分数,平均得分分别提升了 0.09~0.39、1.2%~4.9%,说明本网络对噪声的抑制效果最好,随着信噪比增加,恢复出的语音更加纯净。

表 6 匹配噪声下的不同信噪比的各网络泛化性能对比

Table 6 Comparison of generalization performance of under different signal-to-noise ratios under matched noise

SNR	特征类型	PESQ					STOI(%)					参数量/M
	-	-5	0	5	10	均值	-5	0	5	10	均值	-
含噪语音	-	1.23	1.41	1.66	1.99	1.57	69.5	76.7	83.5	89.1	79.7	-
RCED	幅度谱(Map)	1.82	2.13	2.51	2.89	2.34	76.6	84.8	90.2	93.2	86.2	4.13
DCUNET	复频谱(Mask)	1.89	2.21	2.55	2.92	2.39	78.0	85.4	90.5	93.9	87.0	3.60
CCRN	复频谱(Map)	1.89	2.22	2.59	2.97	2.42	77.2	86.0	91.2	94.3	87.2	9.06
GCRN	复频谱(Map)	1.98	2.31	2.70	3.03	2.51	78.9	87.7	91.7	94.4	88.2	9.77
DCCRN	复频谱(Map)	2.01	2.39	2.78	3.08	2.57	79.6	87.9	92.0	94.6	88.5	3.70
CAUNET	时域波形(Map)	2.09	2.42	2.84	3.11	2.62	80.4	88.1	92.3	94.9	88.9	1.04
Propose-L	复频谱(Map+Mask)	2.13	2.46	2.87	3.18	2.66	80.3	87.9	92.2	94.9	88.8	0.32
Propose	复频谱(Map+Mask)	**2.17**	**2.53**	**2.92**	**3.29**	**2.73**	**82.7**	**89.2**	**93.1**	**95.5**	**90.1**	**0.37**

为验证本文方法在真实应用场景的网络泛化抗噪声能力,在测试集使用不匹配噪声进行验证。如表 7 所示,不匹配噪声下 DCUNET 的各指标比 CCRN 有较大提升,但在匹配噪声下的 DCUNET 与 CCRN 有相似的性能,原因可能是复卷积运算能够更好学习未知的噪声特征。另外,不匹配噪声下的所有网络性能均有一定的下降,但本

表 7　不匹配噪声下的不同信噪比的各网络泛化性能对比

Table 7　Comparison of generalization performance of under different signal-to-noise ratios under unmatched noise

SNR	特征类型	PESQ					STOI(%)					参数量/M
	—	-5	0	5	10	均值	-5	0	5	10	均值	-
-	-5	0	5	10	均值	-5	0	5	10	均值	-	
含噪语音	-	1.30	1.44	1.62	1.85	1.55	66.8	74.9	82.3	88.3	78.1	-
RCED	幅度谱(Map)	1.68	1.88	2.20	2.64	2.10	71.2	80.2	88.1	92.7	83.1	4.13
DCUNET	复频谱(Mask)	1.78	2.08	2.45	2.84	2.29	74.4	83.8	89.8	93.5	85.4	3.60
CCRN	复频谱(Map)	1.77	2.02	2.35	2.76	2.22	71.7	81.3	88.6	93.2	83.7	9.06
GCRN	复频谱(Map)	1.81	2.14	2.46	2.85	2.32	72.9	83.1	89.4	93.5	84.7	9.77
DCCRN	复频谱(Map)	1.84	2.17	2.51	2.92	2.36	73.7	83.7	90.2	93.8	85.4	3.70
CAUNET	时域波形(Map)	1.92	2.28	2.67	3.02	2.47	78.4	87.0	91.5	94.4	87.8	1.04
Propose-L	复频谱(Map+Mask)	1.95	2.34	2.70	3.06	2.51	78.1	86.6	91.1	94.3	87.5	0.32
Propose	复频谱(Map+Mask)	**2.04**	**2.42**	**2.81**	**3.19**	**2.62**	**80.1**	**88.0**	**92.5**	**95.1**	**88.9**	**0.37**

文网络指标下降最少且所有信噪比下的 PESQ、STOI 指标得分均最高,平均得分分别提升了 0.15~0.52 和 1.1%~6.9%,说明本文具有较高的鲁棒性、更强的网络泛化能力和抗噪性。

表 8 为各网络模型在不匹配噪声下的 3 种主观评价指标(CSIG、CBAK、COVL)得分,实验结果表明本文网络的 3 种主观指标得分均高于其他网络,可知本网络对背景噪声的抑制能力较强,能有效恢复语音成分,听觉效果和整体质量较好。

表 8　不匹配噪声下各网络的平均 CSIG、CBAK、COVL 得分

Table 8　CSIG、CBAK、COVL average scores of each network under unmatched noise

对比方法	含噪语音	RCED	DCUNET	CCRN	GCRN	DCCRN	CAUNET	Popose-L	Popose
CSIG	2.43	3.31	3.37	3.33	3.42	3.51	3.72	3.79	**3.88**
CBAK	1.77	2.68	2.77	2.69	2.71	2.78	2.92	3.01	**3.09**
COVL	1.92	2.71	2.80	2.77	2.89	2.96	3.07	3.13	**3.25**

图 7 为各网络模型的参数量和每个时间帧的计算量对比图。RCED 由于其卷积层数较多、卷积核的数量多、尺寸大及跳变连接后的维度拼接导致参数量和计算量庞大;DCUNET 由于其复杂的复卷积运算和较大的卷积核尺寸与卷积核数量,模型的复杂度较高;CCRN、GCRN、DCCRN 模型中巨大的卷积核数量和多层 LSTM 结构大大增加了参数量和计算量,其中 DCCRN 的复杂卷积运算和复数 LSTM 运算再次增加了计算量,而由于其使用较小的卷积核使得参数量较低;CAUNET 中使用多层密集卷积层(Dense-Block),且其中的卷积核的数量和尺寸较大,使得模型计算量和参数量均较高。本文在编解码处使用的卷积核的数量及尺寸均很小,且对卷积核的数量进行了优化,参数量较低,但由于编解码模块中较多的卷积运算,增加了计算量。

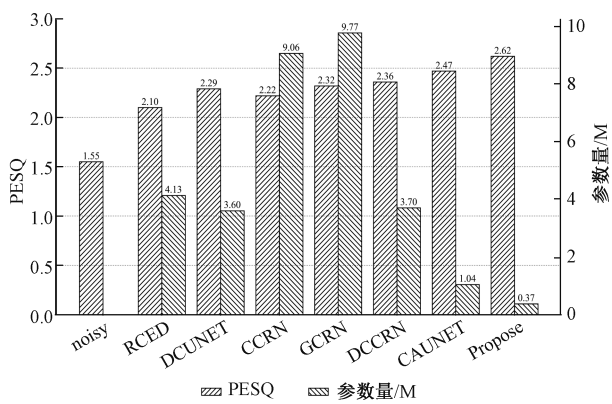


图 7　不匹配噪声下各网络的参数量和计算量对比

Fig. 7　Comparison in terms of parameters and computations of each network under unmatched noise

从图 7、表 7 可以看得出,本文提出模型的低配置 Propose-L 的计算量与图 7 中 CCRN 计算量相当,但 PESQ、STOI 分别提升 43.3%、67.9%;本文提出的模型

Propose 参数量与 GCRN 参数量相当，但 PESQ、STOI 分别提升 39.0%、63.6%。通过对比分析上述模型的参数量、计算量和性能表现可以得出，Propose 模型的计算量虽略微高于部分模型，但参数量最低且性能最优，Propose-L 计算量与最低计算量的模型相当，但参数量和性能远优于该模型。最后，可以得出本文提出的模型最高效，语音质量和可懂度得分最高、去噪声效果最好。

图 8 为在 Babble、White、M109、Factory2、Pink、Hfchannel、Car、Station 等噪声下不同网络模型的 LSD 得分(4 种信噪比下 LSD 得分的均值)，由图中对比结果可知本文提出的网络在 Hfchannel 噪声下取得了与 CAUNET 网络相近的 LSD 得分，但其余噪声均取得了最小的 LSD 得分，说明本文方法对多数噪声特征的学习捕获能力较强，且语音失真更少。

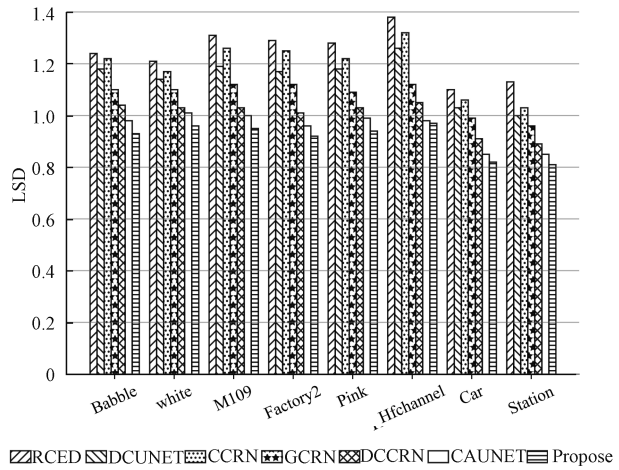


图 8 不同噪声下各网络的平均 LSD 得分

Fig. 8 Average LSD scores of each network under different noises

为了更加直观比较各网络的降噪能力，将含噪语音的增强效果可视化，如图 9、10 所示，以含 Station 噪声且

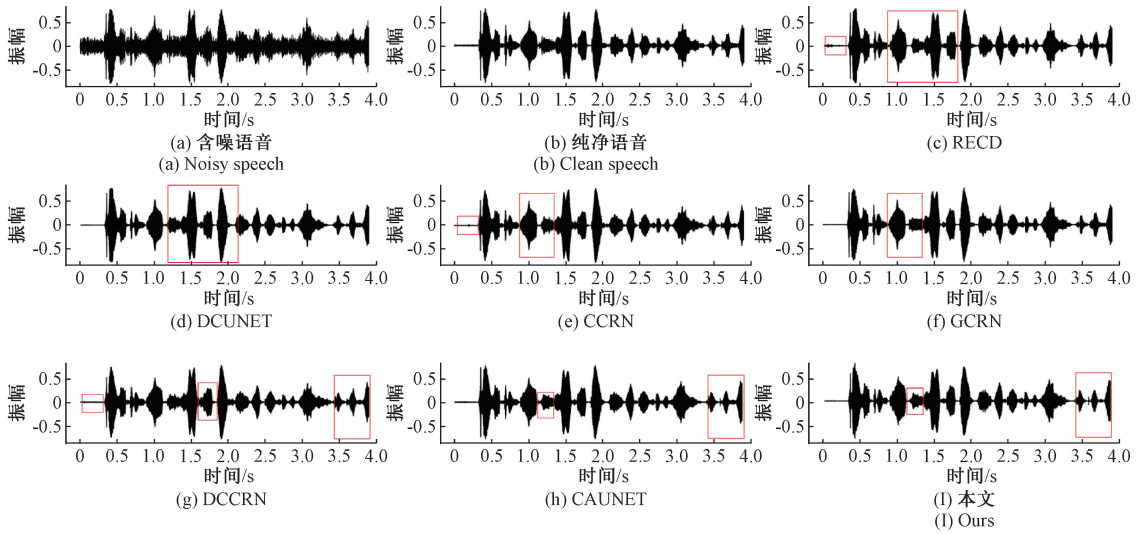
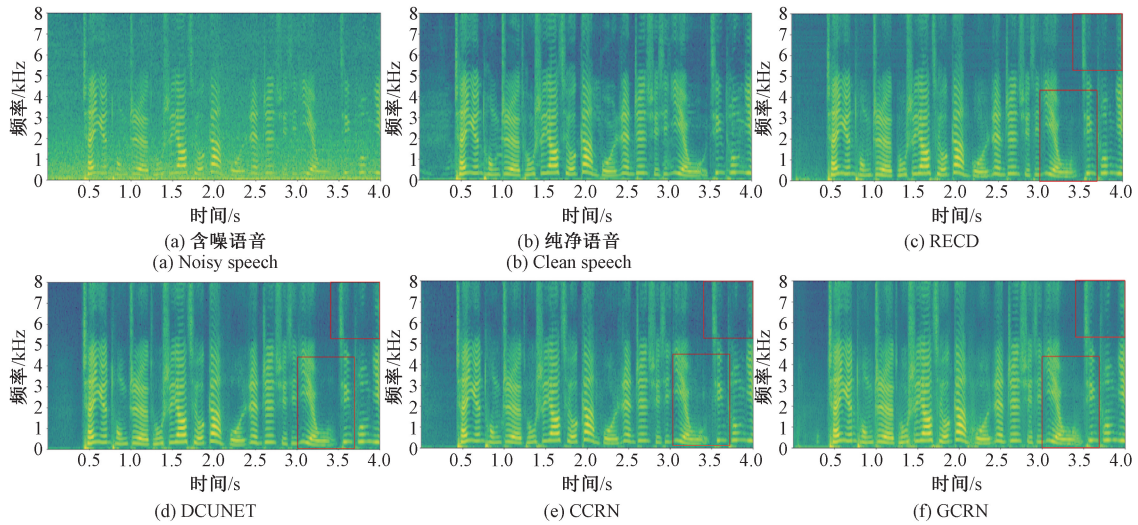


图 9 不同方法的增强语音波形图

Fig. 9 Enhanced speech waveforms using different methods



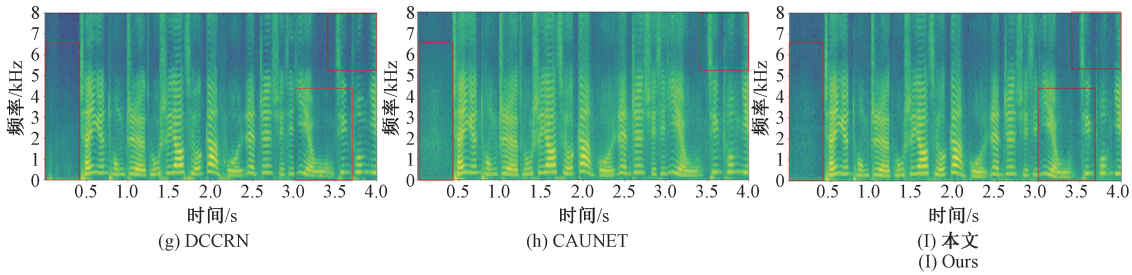


图 10　不同方法的增强语音语谱图

Fig. 10　Enhanced speech Spectrogram using different methods

在-5 dB 下女性说话者为例,对比了纯净语音、含噪语音和不同网络的增强语音的波形图和语谱图。从图 9(c)、(e)、(g)可以看出该 3 种网络模型的语音沉默段仍然含有少量噪声,其他网络模型可以较好消除沉默段的噪声,但部分波形幅值存在失真问题。可以观察到本网络(图 9(I))可以较好消除沉默段和其他片段噪声,对语音波形包络还原度更高,能够更好保留原始语言的清晰度,整体听觉效果更好。从图 10(c)看出 RCED 网络的除噪能力较差,DCUNET、CCRN、GCRN 能恢复部分语音成分,但低频处谐波成分恢复较差,存在一定的背景噪声。虽然 DCCRN、CAUNET 网络进一步降低了残留噪声,但对比本文网络,DCCRN 和 CAUNET 网络的谐波分量还原度较低,高频部分和沉默段仍存在噪声。通过对比各网络模型,本文网络保留更多的谐波特性且除噪能力最好。

4　结　　论

本文提出一种基于复频谱掩蔽和映射的轻量化编解码结构的语音增强网络,在网络的编解码层提出双分支门控协作单元,提取双支路信息流的多层次特征并交互学习;中间层则提出通道时频注意力融合模块来捕获语音的通道、时频多维度细节特征;复频谱掩蔽和映射的方法更利于多特征提取,进一步提升了本模型的降噪能力。在 THCHS30 数据集中测试集的可视化数据结果表明,本网络可以很好降低日常生活中各种多变环境下的噪声以及在语音增强上的性能和复杂度的优越性,各个指标均取得最高的分数,其中在匹配、不匹配噪声下 PESQ 分别提升了 10.5% ~ 50.6%、16.3% ~ 94.5%,且参数量只为对比网络的 3.8% ~ 35.6%,以最小参数量和相当的计算量使得性能提升最大,实现了网络的轻量化。后续的研究会在更大、更复杂的噪声数据集上优化本模型的复频域估计过程和进一步降低模型的计算量。

参考文献

[1] PARK H J, KANG B H, SHIN W, et al. Manner: Multi-view attention network for noise erasure[C]. 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022: 7842-7846.

[2] WANG M, CHEN J Q, ZHANG X L, et al. Multi-modal speech enhancement with bone-conducted speech in time domain[J]. Applied Acoustics, 2022, 200:1-7.

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

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

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

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

[7] MANASWINI B, LAKSHMI V A, CHINTHA L, et al. Gated convolutional recurrent networks with efficient channel attention for monaural speech enhancement[C]. 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), 2023: 217-222.

[8] NUSTEDE E J, ANEMÜLLER J. Single-channel speech enhancement with deep complex U-Networks and probabilistic latent space models[C]. 2023 IEEE

International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023:11331-11335.

[9] LI AN D, LIU W, 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.

[10] ZHAO S, 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]. 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023:13140-13144.

[11] ABDULATIF S, CAO R ZH, YANG B. CMGAN: Conformer-based metric-gan for monaural speech enhancement[J]. ArXiv preprint arXiv: 2209. 11112, 2022.

[12] TAN K, WANG D L. A convolutional recurrent neural network for real-time speech enhancement[C]. Interspeech 2018, 2018: 3229-3233.

[13] YIN D CH, LUO CH, XIONG ZH W, et al. PHASEN: A phase-and-harmonics-aware speech enhancement network[C]. 34th AAAI Conference on Artificial Intelligence, 2020: 9458-9465.

[14] AI Y, LING ZH H. Neural speech phase prediction based on parallel estimation architecture and anti-wrapping losses[C]. 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023: 2010-2014.

[15] 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.

[16] ZHANG G CH, WANG CH L, YU L B, et al. Multi-scale temporal frequency convolutional network with axial attention for multi-channel speech enhancement[C]. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022:9206-9210.

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

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

[19] NARAYANAN A, WANG D L. Ideal ratio mask estimation using deep neural networks for robust speech recogni-tion[C]. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013: 7092-7096.

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

[21] 金宇锋, 陶重犇. 基于 Transformer 的融合信息增强 3D 目标检测算法[J]. 仪器仪表学报, 2023, 44(12): 297-306.
JIN Y F, TAO CH B. Fusion information enhanced method based on Transformer for 3D object detection[J]. Chinese Journal of Scientific Instrument, 2024, 44(12): 297-306.

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

[23] WANG K, HE B B, ZHU W P. CAUNET: Context-aware U-Net for speech enhancement in time domain[C]. 2021 IEEE International Conference on Aconstics, Spoech and Signal Processing, 2021: 7098-7102.

[24] DAUPHIN Y N, FAN A, AULI M, et al. Language modeling with gated convolutional networks[C]. International Conference on Machine Learning, 2017, 70: 933-941.

[25] VASWANI A, SHAZEER N, PARMAR N, et al.

Attention is all you need [C]. Advances in Neural Information Processing Systems, 2017.

[26] BI M X, LU H, ZHANG SH L, et al. Deep feed-forward sequential memory networks for speech synthesis [C]. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018:4794-4798.

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

[28] TAAL C H, HENDRIKS R C, HEUSDENS R, et al. An algorithm for intelligibility prediction of time-frequency weighted noisy speech[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2125-2136.

[29] HU Y, LOIZOU P C. Evaluation of objective quality measures for speech enhancement [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(1): 229-238.

[30] PARK S R, LEE J. A fully convolutional neural network for speech enhancement [C]. Interspeech, 2017: 2017-1465.

作者简介

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

E-mail: zhangtg@ cqupt. edu. cn

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

沈夕文(通信作者),2022 年于阜阳师范大学获得学士学位,现为重庆邮电大学在读硕士研究生,主要研究方向为语音增强与语音分离。

E-mail: 674051078@ qq. com

Shen Xiwen (Corresponding author) received his B. Sc. degree in 2022 from Fuyang Normal University. Now he is a M. Sc. candidate in Chongqing University of Posts and Telecommunications. His main research interests include speech enhancement and speech separation.