ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 2,Mar. 2023,pp. 274－282
DOI:10. 16337/j. 1004-9037. 2023. 02. 004

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

一种轻量级全频带语音增强网络模型

胡沁雯，侯仲舒，乐笑怀，卢　晶

(南京大学声学研究所,近代声学教育部重点实验室,南京 210093)

摘　要：基于深度神经网络的全频带语音增强系统面临着计算资源需求高以及语音在各频段分布不平衡的困难。本文提出了一种轻量级全频带网络模型。该模型在双路径卷积循环网络模型的基础上,利用可学习的频谱压缩映射对高频段频谱信息进行有效压缩,同时利用多头注意力机制对频域的全局信息进行建模。实验结果表明本文模型只需0.89×10^6的参数即可实现有效的全频带语音增强,验证了本文模型的有效性。

关键词：全频带语音增强;深度学习;多头注意力机制

中图分类号：TP301　　**文献标志码**:A

A Light-Weight Full-Band Speech Enhancement Model

HU Qinwen, HOU Zhongshu, LE Xiaohuai, LU Jing

(Key Laboratory of Modern Acoustics, Institute of Acoustics, Nanjing University, Nanjing 210093, China)

Abstract: Deep neural network based full-band speech enhancement systems face challenges of high demand of computational resources and imbalanced frequency distribution. In this paper, a light-weight full-band model is proposed based on dual path convolutional recurrent network with two dedicated strategies, i. e., a learnable spectral compression mapping for more effective high-band spectral information compression, and the utilization of the multi-head attention mechanism for more effective modeling of the global spectral pattern. Experiments validate the efficacy of the proposed strategies and show that the proposed model achieves competitive performance with only 0.89×10^6 parameters.

Key words: full-band speech enhancement; deep learning; multi-head attention mechanism

引　　言

在过去10年中,基于深度神经网络(Deep neural networks, DNN)的数据驱动语音增强方法取得了较大的进展[1],与传统的信号处理方法相比,这类方法在噪声抑制和语音保留上都具有更好的效果。大多数语音增强系统专注于处理宽带(采样率16 kHz)或窄带(采样率8 kHz)语音,而全频段(采样率48 kHz)语音增强方法仍有待探索,以便应用于需要高保真音频的场景。直接将宽带处理网络的频率维度拓宽以应用于全频段语音增强是不可取的,这将导致内存需求和计算负担的显著增加,使得其无法应用于计算资源有限的便携语音交互设备。此外,以统一的方式处理所有频带并非最优方式,因为

基金项目:国家自然科学基金(12274221)。
收稿日期:2022-07-07;**修订日期**:2022-08-12

语音的大部分能量和谐频信息均集中在中低频范围[2]。解决方案之一是用频谱包络来压缩频域特征[3-4],但采用该方案的网络模型受制于频域低分辨率。此外,使用针对不同频段优化的子网络[5]可以专注于低频处理,以显著提高语音增强性能,但它在处理机制上很难减小模型尺寸。

本文基于宽带的双路径卷积循环网络模型(Dual path convolutional recurrent network, DPCRN)[6],提出了一种轻量级全频带语音增强网络模型。DPCRN是一个非常有竞争力的轻量级语音增强模型,在第3届深度噪声抑制挑战赛(Deep Noise Suppression-3,DNS-3)[7]中排名第3位,比前两名的模型具有更少的参数和更低的计算负担:DPCRN需要0.8×10^6参数和3.7×10^9乘加数(Multiply-accumulate operation,MAC),第1名模型需要6.4×10^6参数和6.0×10^9 MAC,第2名模型需要5.2×10^6参数和52.5×10^9 MAC。DPCRN模型分别在两条路径上,即处理频域信息的块内路径和处理时域信息的块间路径上,均使用循环神经网络(Recurrent neural network,RNN)进行建模。为了更好地利用频谱之间的内在联系,本文模型中引入了注意力机制,即用多头注意力(Multi-head attention,MHA)网络[8]取代用于块内处理的RNN,而在块间处理上依然保留RNN。为了更有效地处理语音分布稀疏的高频分量,本文还将一个可学习的频谱压缩映射(Spectral compression mapping,SCM)及其反演变换分别添加到网络预处理和后处理中,可以有效地压缩整个模型的大小。该模型被命名为频谱压缩映射双路径注意力循环网络(Spectral compression mapping-dual path attention recurrent network, SCM-DPARN)。在只有0.89×10^6参数的条件下,SCM-DPARN模型可以获得与其他高性能全频带模型类似的语音增强效果。

1 SCM-DPARN模型

1.1 问题描述和模型结构

时频(Time-frequency,T-F)域含噪语音可以描述为

$$X(t,f)=S(t,f)+N(t,f) \tag{1}$$

式中:t表示时帧序数;f表示频点序数;$S(t,f)$、$N(t,f)$分别表示该时频点的纯净语音和噪声成分,且$S(t,f)$是可能带有混响的信号。下文在指代整个频谱图时省略t、f序数。

本文的处理目标不考虑去混响,因此模型的目标是对带混响语音S进行估计,其估计值为$\tilde{S}$。具体的方法是将输入的噪声语音X直接映射到$\tilde{S}$上,而非间接地进行掩膜估计。模型的训练由最小化在$\tilde{S}$和S上计算的损失函数L来实现,表达式为

$$\tilde{S}=\mathcal{F}(X) \tag{2}$$

$$L=\mathrm{Loss}\left(\tilde{S},S\right) \tag{3}$$

式中$\mathcal{F}$为作用于复频谱的网络函数。使用基于映射的方法直接估计目标频谱的实部和虚部,其好处在于即使输入信号经过低通滤波,网络也可以恢复原始语音的部分高频段频谱。

与DPCRN[6]类似,本文模型由1个编码器、1个双路径处理模块和2个解码器组成,如图1所示。编码器连接在1.2节描述的频谱压缩映射层之后,包含多个二维卷积层;而每个解

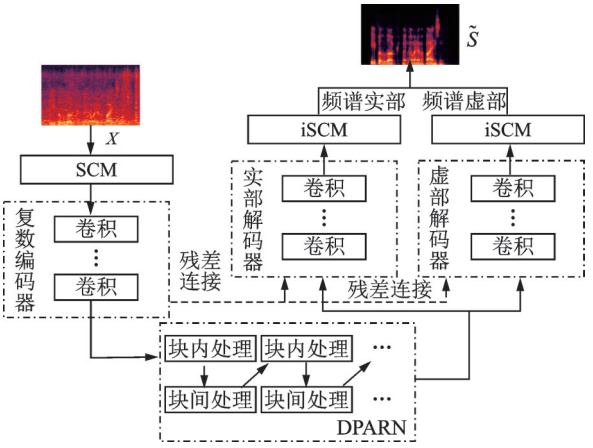

图1 本文提出的SCM-DPARN模型整体结构图

Fig.1 Network architecture of the proposed SCM-DPARN model

码器包含多个转置二维卷积层，其后分别连接一个逆频谱压缩映射（inverse Spectral compression mapping，iSCM）层，用以重建频谱的实部或虚部。在编码器和解码器中的相应层之间使用残差连接。编解码器的结构示意图如图2所示。

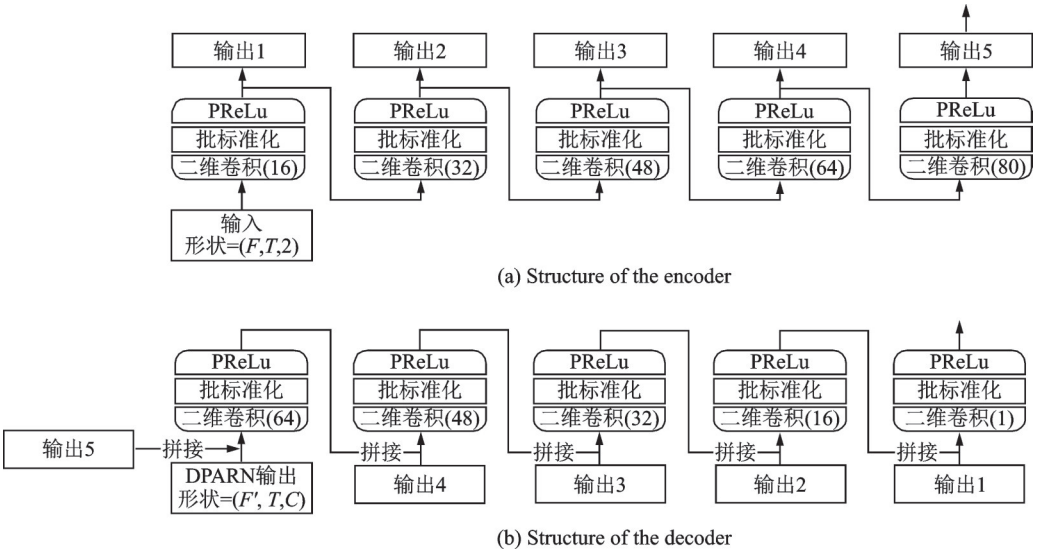

图2 编码器结构和解码器结构示意图

Fig.2 Structures of the encoder and the decoder

1.2 频谱压缩映射

在全频带语音增强任务上，如果保持频谱的分辨率不变，将宽带网络直接扩展成全频带模型，模型计算量将至少提高到原来的3倍，同时它会将三分之二的计算资源分配给信息量较少的高频段（8～24 kHz），从而显著增加计算负担和学习困难。因此，有必要引入频谱信息映射策略来有效压缩高频段。SCM层的设计遵循梅尔尺度滤波器组的类似形式。梅尔尺度滤波器通过对数函数将频率转换为梅尔尺度[9-10]。为了进一步保留关键的中低频段信息，模型完整保留低于5 kHz的频段信息，只对高于5 kHz的频段以对数形式进行拉伸映射。映射曲线定义为

$$q_{\mathrm{c}}(m)=\begin{cases} q(m) & 0\ \mathrm{kHz}\leqslant q(m)\leqslant 5\ \mathrm{kHz} \\ 2\,500\left[\ln\left(\dfrac{q(m)-2\,500}{2\,500}\right)+2\right] & 5\ \mathrm{kHz}<q(m)\leqslant 24\ \mathrm{kHz} \end{cases} \tag{4}$$

式中：m为频点的索引；$q(m)$为原始频谱中第m个频点的物理频率；$q_{\mathrm{c}}(m)$为该频点在变换域对应的数值。考虑将一个F维谱压缩成F_{c}维频谱，在变换域中重新采样F_{c}个均匀分布的点$q_{\mathrm{c}}(k)(k=1,2,\cdots,F_{\mathrm{c}})$。上述的对数映射和重新采样的变换过程可以描述为

$$\boldsymbol{x}_{\mathrm{comp}}=\boldsymbol{W}_{\mathrm{SCM}}\boldsymbol{x}\qquad \boldsymbol{x}\in\mathbf{R}^{F},\ \boldsymbol{x}_{\mathrm{comp}}\in\mathbf{R}^{F_{\mathrm{c}}} \tag{5}$$

$$\boldsymbol{W}_{\mathrm{SCM}}=\begin{bmatrix}\boldsymbol{I}_{K\times K} & \boldsymbol{0}_{K\times(F-K)} \\ & \boldsymbol{G}\end{bmatrix}\in\mathbf{R}^{F_{\mathrm{c}}\times F} \tag{6}$$

$$\boldsymbol{G}=\left[\boldsymbol{g}_{K+1},\boldsymbol{g}_{K+2},\cdots,\boldsymbol{g}_{k},\cdots,\boldsymbol{g}_{F_{\mathrm{c}}}\right]^{\mathrm{T}}\in\mathbf{R}^{(F_{\mathrm{c}}-K)\times F} \tag{7}$$

$$\boldsymbol{g}_k(m)=\begin{cases}0 & q(m)<q(q_c(k-1))\\ \dfrac{q(m)-q(q_c(k-1))}{q(q_c(k))-q(q_c(k-1))} & q(q_c(k-1))<q(m)\leqslant q(q_c(k))\\ \dfrac{q(q_c(k+1))-q(m)}{q(q_c(k+1))-q(q_c(k))} & q(q_c(k))<q(m)\leqslant q(q_c(k+1))\\ 0 & q(m)>q(q_c(k+1))\end{cases} \tag{8}$$

式中:$\boldsymbol{x}$为原始频谱;$\boldsymbol{x}_{\text{comp}}$为变换后的频谱;$K$为对应于5 kHz阈值的频点的索引;$\boldsymbol{g}_k$为第$k$个三角滤波器;$q(q_c(k))$为式(4)中对数函数的逆映射,即$q(q_c(k))=2\,500\left(\mathrm{e}^{\frac{q_c(k)}{2500}-2}+1\right)$;$m=1,2,\cdots,F$;$k=K+1,K+2,\cdots,F_c$。

这种变换依然具有局限性。式(4)的对数映射能够更合理地在不同频段分配计算资源,但它并不能有效匹配高频范围内语音频谱的稀疏分布。因此,直接采用这种压缩模式将导致高频部分不能得到有效处理,在高频段有较大的噪声残留。

为了更有效地适应高频范围内语音的稀疏分布,模型中使用一个部分可学习的压缩矩阵$\tilde{\boldsymbol{W}}_{\text{SCM}}$来实现频谱压缩映射,并且该压缩矩阵由$\boldsymbol{W}_{\text{SCM}}$来进行初始化。其中低频带映射$\tilde{\boldsymbol{W}}_{\text{SCM}}^{\text{low}}=[\boldsymbol{I}\ \boldsymbol{0}]\in\mathbf{R}^{K\times F}$是固定的,高频带部分$\tilde{\boldsymbol{W}}_{\text{SCM}}^{\text{high}}\in\mathbf{R}^{(F_c-K)\times F}$由网络进行学习,并且由式(7)中的$\boldsymbol{G}$来进行初始化。相应地,逆频谱压缩映射iSCM也通过可学习矩阵$\tilde{\boldsymbol{W}}_{\text{iSCM}}\in\mathbf{R}^{F\times F_c}$实现,且$\tilde{\boldsymbol{W}}_{\text{iSCM}}$采用随机初始化。

1.3 双路径注意力循环网络

原始的DPCRN在两个不同的路径上使用了RNN,即块内RNN和块间RNN。这里的“块”指代单帧的频谱。块内RNN作用于频域,用于对单帧中各频率之间的相关性进行建模;块间RNN作用于时域,用于对时间依赖关系进行建模。考虑到全频带语音具有明显更宽的频率跨度,用MHA替换块内RNN,因为它可以更有效地模拟长序列的全局频谱模式。另一方面,时间轴上的全局信息对于语音增强来说不是必需的,因此保留了块间RNN。

图3给出了DPARN处理模块的详细结构以及MHA的具体结构。在SCM层之后,编码器进一步对频域进行压缩,并且在各时频点上提取维度为C的局部特征向量。在块内MHA的输入上添加三角位置编码(Positional encoding,PE)[8]。在块内MHA中,时域方向上进行并行处理,注意力层的输入为序列长度为F'的C维向量,其中F'表示SCM和卷积编码器压缩后的频率维度,$F'<F_c<F$。查询向量$\boldsymbol{Q}\in\mathbf{R}^{F'\times C}$、键向量$\boldsymbol{K}\in\mathbf{R}^{F'\times C}$和值向量$\boldsymbol{V}\in\mathbf{R}^{F'\times C}$都是与注意力层输入相同的向量。注意力机制由$H$个平行的注意力头来实现。在每个注意力头中,$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$分别被线性地投影到$d_q$、$d_k$($d_k=d_q$)和$d_v$维度上,即

$$\boldsymbol{Q}_h=\boldsymbol{Q}\boldsymbol{W}_h^Q,\ \boldsymbol{K}_h=\boldsymbol{K}\boldsymbol{W}_h^K,\ \boldsymbol{V}_h=\boldsymbol{V}\boldsymbol{W}_h^V\quad \boldsymbol{W}_h^Q\in\mathbf{R}^{C\times d_q},\ \boldsymbol{W}_h^K\in\mathbf{R}^{C\times d_k},\ \boldsymbol{W}_h^V\in\mathbf{R}^{C\times d_v} \tag{9}$$

式中h为注意力头的索引。之后将缩放的点积注意力计算应用于$\boldsymbol{Q}_h$、$\boldsymbol{K}_h$和$\boldsymbol{V}_h$,即

$$\text{Attention}(\boldsymbol{Q}_h,\boldsymbol{K}_h,\boldsymbol{V}_h)=\text{softmax}\left(\frac{\boldsymbol{Q}_h\boldsymbol{K}_h^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_h \tag{10}$$

式(10)计算每2个频率点在投影空间中的相似性,并相应地为值向量的投影值$\boldsymbol{V}_h$分配权重,再加权求和作为每个注意力头对应的输出。不同注意力头的输出连接起来,并被线性投影回一系列C维向量,即

$$\begin{aligned}&\text{MHA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(\text{head}_1,\text{head}_2,\cdots,\text{head}_H)\boldsymbol{W}^O\\ &\text{head}_h=\text{Attention}(\boldsymbol{Q}_h,\boldsymbol{K}_h,\boldsymbol{V}_h)\quad \boldsymbol{W}^O\in\mathbf{R}^{(H\times d_v)\times C}\end{aligned} \tag{11}$$

在MHA层之后,用一个前馈网络进一步处理每个频点的信息,该前馈网络包括2个全连接层和

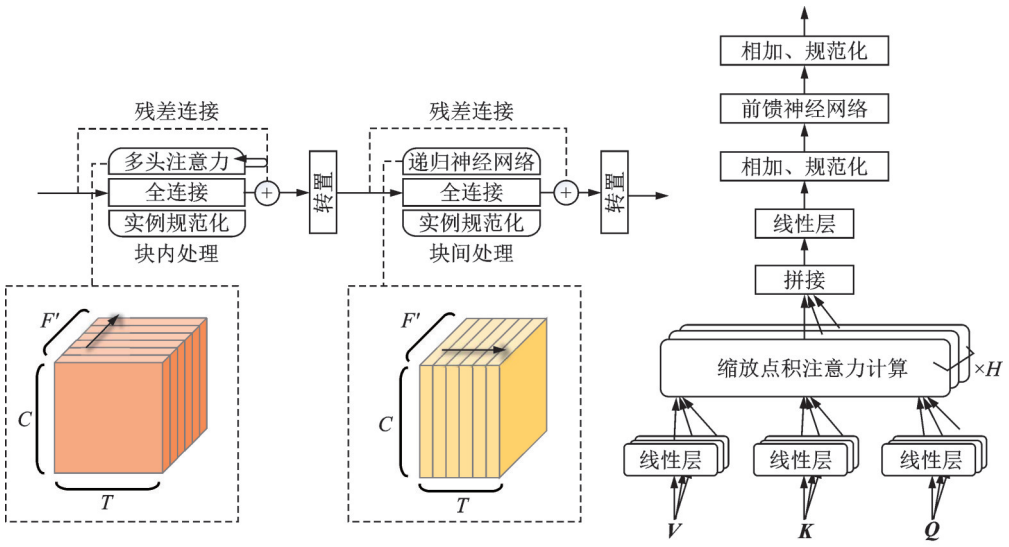

图3 DPARN模块和MHA模块的结构示意图

Fig.3 Structures of DPARN module and MHA module

1个ReLU激活函数层，即

$$\mathrm{FFN}(\boldsymbol{z})=\max(\boldsymbol{0},\boldsymbol{z}\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2 \tag{12}$$

式中：$\boldsymbol{z}$为前馈网络输入；$\boldsymbol{W}_1\in\mathbf{R}^{C\times(4C)}$；$\boldsymbol{W}_2\in\mathbf{R}^{(4C)\times C}$；$\boldsymbol{b}_1\in\mathbf{R}^{4C}$；$\boldsymbol{b}_2\in\mathbf{R}^{C}$。

将1个MHA层和1个前馈网络视为1个MHA模块。B个重复的MHA模块之后为1个全连接层和1个实例标准化层(Instance normalization，IN)。进一步将输出张量转置之后传入块间RNN中。在块间RNN中，使用长短时记忆网络(Long-short term memory，LSTM)处理时序关系[6]，后面依然接有1个全连接层和1个IN层。在每个MHA模块内，以及块间处理部分与块内处理部分之间应用残差连接[6]。

1.4 训练目标

网络的学习目标是纯净语音频谱的实部和虚部$S=S_{\mathrm{real}}+\mathrm{i}S_{\mathrm{imag}}$，其中$S_{\mathrm{real}}$表示实部，$S_{\mathrm{imag}}$表示虚部。使用功率压缩损失函数[11]作为训练目标，可以更好地处理低功率时频点中的信息。复数谱在极坐标下可以写为$S=|S|\mathrm{e}^{\mathrm{i}\theta_S}$，其幂压缩谱可以描述为$S^{\mathrm{C}}=|S|^{\gamma}\mathrm{e}^{\mathrm{i}\theta_S}$，其中$\gamma$为压缩系数，上标C表示功率压缩模式。因此，实部和虚部分别表示为

$$S_{\mathrm{real}}^{\mathrm{C}}=|S|^{\gamma}\cos\theta_S,\ \ S_{\mathrm{imag}}^{\mathrm{C}}=|S|^{\gamma}\sin\theta_S \tag{13}$$

网络估计得到的纯净语音的$\tilde{S}_{\mathrm{real}}^{\mathrm{C}}$和$\tilde{S}_{\mathrm{imag}}^{\mathrm{C}}$遵循相同的定义。用于恢复复数谱和幅度谱的损失函数$L_{\mathrm{RI}}\left(\tilde{S},S\right)$和$L_{\mathrm{Mag}}\left(\tilde{S},S\right)$分别定义为

$$L_{\mathrm{RI}}\left(\tilde{S},S\right)=\left\|S_{\mathrm{real}}^{\mathrm{C}}-\tilde{S}_{\mathrm{real}}^{\mathrm{C}}\right\|_{\mathrm{F}}^{2}+\left\|S_{\mathrm{imag}}^{\mathrm{C}}-\tilde{S}_{\mathrm{imag}}^{\mathrm{C}}\right\|_{\mathrm{F}}^{2} \tag{14}$$

$$L_{\mathrm{Mag}}\left(\tilde{S},S\right)=\left\|\,|S|^{\gamma}-\left|\tilde{S}\right|^{\gamma}\right\|_{\mathrm{F}}^{2} \tag{15}$$

式中$\|\cdot\|_{\mathrm{F}}$为矩阵的Frobenius范数，最终用于训练SCM-DPARN的损失函数为$L_{\mathrm{RI}}\left(\tilde{S},S\right)$和$L_{\mathrm{Mag}}\left(\tilde{S},S\right)$之和，表达式为

$$L=L_{\mathrm{RI}}\left(\tilde{S},S\right)+L_{\mathrm{Mag}}\left(\tilde{S},S\right) \tag{16}$$

2 仿真实验与结果分析

2.1 消融实验

2.1.1 数据集

首先进行消融实验,以评估SCM层的作用以及DPARN与其他双路径模型相比的优势。所有模型均在一个小数据集上进行训练,其中包括来自英语数据集VCTK[12]和法语数据集SIWIS[13]的纯净语音,以及来自DEMAND[14]和QUT-NOISE[15]的噪声数据。所有音频都以48 kHz采样。纯净语音数据集的大小总共约为45 h。将音频随机分成10 s长的片段,总共生成16 000个纯净语音片段。其中14 500个片段用于训练,其余用于验证。首先将10% 的纯净数据与从openSLR26和openSLR28[16]中随机选取的房间脉冲响应进行卷积,然后以−5~10 dB(间隔1 dB)的信噪比(Signal-to-noise ratio,SNR)将语音片段与噪声片段随机混合,从而生成带噪语音片段。对于带有混响的带噪语音,模型的训练目标为带混响的纯净语音,即在本文的实验中仅以去除加性噪声为目标,不考虑去混响。

对于测试数据集,使用来自DAPS[17]的纯净语音和来自Saki[18]的噪声。测试数据的模拟生成方式和训练数据相同。测试的SNR级别为{−5 dB, 0 dB, 5 dB, 10 dB}。

2.1.2 参数设置

短时傅里叶变换(Short-time Fourier transform,STFT)的窗口长度为25 ms,帧移为12.5 ms。离散傅里叶变换长度为1 200个点,即输入网络的频率特征的维度为601,执行STFT时使用汉宁窗。

在SCM层,$F_c=256$,$K=125$。编码器包含5个二维卷积层。输出通道维度为{16, 32, 48, 64, 80},卷积核大小为{(5,2),(3,2),(3,2),(3,2),(2,1)},步长为{(2,1),(1,1),(1,1),(1,1),(1,1)},其中第1个数字指代频率轴上的配置,第2个数字指代时间轴上的配置。解码器中的转置卷积层是反向排序的。每个(转置)卷积层后皆为1个批标准化层(Batch normalization,BN)和1个PReLU函数层。残差连接通过在通道维度上进行拼接来实现。在本实验中,DPARN处理块仅由1个块内处理模块和1个块间处理模块组成。在块内MHA中,$B=2$, $H=8$, $d_k=d_q=d_v=C/H$,其中$C=80$;在块间RNN中,LSTM的隐藏层大小为127。

采用预热训练法[8]来训练SCM-DPARN。预热法调节学习率从一个很小的值开始,先上升再下降,它能够防止网络权重在一开始波动太大,使其找到一个合适的收敛方向。学习率α随训练步长ψ而变化:$\alpha=\frac{1}{\sqrt{C}}\times\min\left(\frac{1}{\sqrt{\psi}},\frac{\psi}{\sqrt{\Psi^3}}\right)$,其中预热步数$\Psi$为40 000。实验中使用Adam优化器来进行学习,具体参数$\beta_1=0.9$, $\beta_2=0.98$, $\epsilon=10^{-9}$;压缩参数$\gamma=\frac{2}{3}$。

2.1.3 基线模型和评价指标

基线模型包括DPCRN[6]与4个变体:带SCM层的DPCRN(SCM-DPCRN)、使用块内RNN与块间MHA的模型(SCM-DPRAN)、使用块内MHA和块间MHA的模型(SCM-DPAAN)。无SCM层的DPCRN模型使用完整的601维频率特征进行训练,不进行频谱压缩。块间MHA中MHA模块的数量$B=2$。评估指标包括语音质量感知评估(Perceptual evaluation of speech quality,PESQ)、短时客观可懂度(Short-time objective intelligibility,STOI)和尺度不变信号失真比(Scale-invariant signal-to-distortion ratios,SI-SDR)。

2.1.4 实验结果及分析

消融实验中各模型在不同信噪比测试集上的PESQ、STOI、SI-SDR分数分别如表1~3所示。其中PESQ、STOI分数是将信号全部从48 kHz采样率降采样至16 kHz后测得的。SI-SDR分数是在原始的48 kHz采样率信号上计算得到的。从各个指标来看,不进行任何频谱压缩的DPCRN模型性能比带有

SCM层的DPCRN模型性能相对更弱。由此可见,对于全频带语音增强模型,保持完整的频率信息不仅在计算中是多余的,而且会恶化模型的性能。在这种设置下,模型分配了大量的计算资源用于学习语音能量分布稀疏的高频部分,既无法在高频上提取有效的信息,又影响了模型对低频成分的建模。本文提出的SCM-DPARN在各指标、各信噪比数据上性能皆为最佳,并且其优势在低信噪比下更加明显。得益于SCM模块,SCM-DPARN能够有效地提取中低频的谐波信息;同时由于多头自注意力网络的全局信息处理优势,它对频带较宽的辅音成分能达到较好的保留效果。SCM-DPRAN和SCM-DPAAN两个模型有性能下降,甚至劣于SCM-DPCRN。说明至少对于轻量级网络而言,在MHA、RNN层数较少的情况下,仅沿频率轴应用多头注意力机制确实是最佳选择。

表1 不同信噪比测试集上的PESQ评分

Table 1 PESQ scores on test datasets with different SNRs

模型	SNR/dB			
	−5	0	5	10
原带噪语音	1.15	1.23	1.37	1.64
DPCRN	1.45	1.94	2.07	2.36
SCM-DPCRN	1.69	2.33	2.53	2.99
SCM-DPAAN	1.47	1.84	2.15	2.58
SCM-DPRAN	1.54	2.04	2.32	2.78
SCM-DPARN	**1.84**	**2.42**	**2.61**	**3.03**

表2 不同信噪比测试集上的STOI评分

Table 2 STOI scores on test datasets with different SNRs %

模型	SNR/dB			
	−5	0	5	10
原带噪语音	77.6	86.3	91.1	95.5
DPCRN	76.8	89.0	90.7	91.6
SCM-DPCRN	83.2	91.8	94.2	96.9
SCM-DPAAN	82.0	89.6	92.5	95.6
SCM-DPRAN	81.1	90.5	93.5	96.5
SCM-DPARN	**85.3**	**92.9**	**94.7**	**97.0**

表3 不同信噪比测试集上的SI-SDR评分

Table 3 SI-SDR scores on test datasets with different SNRs

模型	SNR/dB			
	−5	0	5	10
原带噪语音	−4.99	−0.01	5.00	10.00
DPCRN	3.68	8.64	9.79	11.65
SCM-DPCRN	6.05	10.66	12.20	15.06
SCM-DPAAN	4.57	9.08	10.92	13.79
SCM-DPRAN	4.30	9.25	11.53	14.14
SCM-DPARN	**7.07**	**11.51**	**12.82**	**15.41**

2.2 增强实验

2.2.1 数据集和参数设置

为了将SCM-DPARN与近年的其他全频带、超宽带(采样率32 kHz)语音增强论文模型进行比较,在公开的VCTK-DEMAND数据集[19]上进一步训练和测试提出的模型。该数据集仅提供纯净语音与

带噪语音,不提供原始的噪声数据。纯净语音来自VCTK[12]数据集,其中28个说话人的数据用于训练,2个说话人的数据用于测试。噪声数据包括用于训练的2种人工生成的噪声类型(语音形噪声和咿呀声)和来自DEMAND[14]的8种真实噪声,以及用于测试的其他5种噪声。训练的SNR级别为{0 dB,5 dB,10 dB,15 dB},测试的SNR级别为{2.5 dB,7.5 dB,12.5 dB,17.5 dB}。训练数据总共大约10 h。STFT配置、网络参数设置和训练策略与消融实验相同,预热步长为5 000。

2.2.2 基线模型和评价指标

基线模型包括:RNNoise[20]、PerceptNet[3]、DeepFilterNet[4]和S-DCCRN[5]。基线模型的分数摘取于相关论文中[3-5,20]。在这些模型中,只有S-DCCRN[5]是在VCTK-DEMAND训练集上训练的,且在文献[5]中应用于超宽带语音增强,对应的处理难度低于全频带语音增强。

2.2.3 实验结果及分析

在公开的VCTK-DEMAND测试集上的结果如表4所示。虽然RNNoise[20]只有0.06×10^6的参数,但它在性能上相较其他模型有较大劣势。在训练数据集较小的情况下,本文提出的SCM-DPARN模型在PESQ、STOI、SI-SDR三个指标上都取得了最好的分数,且模型大小只有0.89×10^6参数。结果表明,利用SCM对全频带信息进行压缩、使用多头自注意力对频谱结构进行建模对于全频带语音增强模型是有效的改进策略。

表4 VCTK-DEMAND数据集上各模型性能比较

Table 4 Performance comparison of different models on VCTK-DEMAND dataset

模型	参数量/10^6	采样率/kHz	PESQ	STOI/%	SI-SDR
原带噪语音	—	48	1.97	92.1	8.41
RNNoise[20]	**0.06**	48	2.29	—	—
PerceptNet[3]	8.00	48	2.73	—	—
DeepFilterNet[4]	1.80	48	2.81	—	16.63
S-DCCRN[5]	2.34	32	2.84	94.0	—
SCM-DPARN	0.89	48	**2.92**	**94.2**	**18.28**

3 结束语

本文提出了一种轻量级的全频带语音增强模型SCM-DPARN。该模型利用可学习的频谱压缩映射来更有效地压缩信息量较少的高频段频谱,用多头注意力网络取代循环神经网络对全频段频谱的全局结构进行建模。本文通过消融实验验证了频谱压缩的有效性,并进一步确认了DPARN相较于其他双路径语音增强模型的优势。在VCTK-DEMAND数据集上的实验显示,与几种全频带语音增强模型相比,本文提出的SCM-DPARN仅使用0.89×10^6参数就实现了较好的语音增强效果。

参考文献:

[1] WANG Deliang, CHEN Jitong. Supervised speech separation based on deep learning: An overview[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(10): 1702-1726.

[2] MONSON B B, LOTTO A J, STORY B H. Analysis of high-frequency energy in long-term average spectra of singing, speech, and voiceless fricatives[J]. The Journal of the Acoustical Society of America, 2012, 132(3): 1754-1764.

[3] VALIN J M, ISIK U, PHANSALKAR N, et al. A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech[C]//Proceedings of Interspeech. [S.l.]: IEEE, 2020: 2482-2486.

[4] SCHRÖTER H, ROSENKRANZ T, MAIER A. DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022: 7407-7411.

[5] LV Shubo, FU Yihui, XING Mengtao, et al. S-DCCRN: Super wide band DCCRN with learnable complex feature for speech

enhancement[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022: 7767-7771.

[6] LE Xiaohuai, CHEN Hongsheng, CHEN Kai, et al. DPCRN: Dual-path convolution recurrent network for single channel speech enhancement[C]//Proceedings of Interspeech. [S.l.]: IEEE, 2021: 2811-2815.

[7] REDDY C K A, DUBEY H, KOISHIDA K, et al. INTERSPEECH 2021 deep noise suppression challenge[C]//Proceedings of Interspeech. Toronto, ON, Canada: IEEE, 2021: 2796-2800.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017. DOI: https://doi.org/10.48550/arXiv.1706.03762.

[9] DAVIS S, MERMELSTEIN P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(4): 357-366.

[10] SKOWRONSKI M D, HARRIS J G. Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition[J]. The Journal of the Acoustical Society of America, 2004, 116(3): 1774-1780.

[11] LI Andong, ZHENG Chengshi, PENG Renhua, et al. On the importance of power compression and phase estimation in monaural speech dereverberation[J]. JASA Express Letters, 2021, 1(1): 014802.

[12] VEAUX C, YAMAGISHI J, MACDONALD K. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit, technical report[R]. Edinburgh: The University of Edinburgh, 2017.

[13] HONNET P E, LAZARIDIS A, GARNER P N, et al. The SIWIS French speech synthesis database—Design and recording of a high quality French database for speech synthesis[R]. Switzerland: IDIAP Research Institute, 2017.

[14] THIEMANN J, ITO N, VINCENT E. The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings[C]//Proceedings of Meetings on Acoustics ICA2013. Montreal Montreal, Canada: [s.n.], 2013: 035081.

[15] DEAN D, SRIDHARAN S, VOGT R, et al. The QUT-NOISE-TIMIT corpus for evaluation of voice activity detection algorithms[C]//Proceedings of the 11th Annual Conference of the International Speech Communication Association. Makuhari, Chiba, Japan: DBLP, 2010: 3110-3113.

[16] KO T, PEDDINTI V, POVEY D, et al. A study on data augmentation of reverberant speech for robust speech recognition [C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, USA: IEEE, 2017: 5220-5224.

[17] MYSORE G J. Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—A dataset, insights, and challenges[J]. IEEE Signal Processing Letters, 2014, 22(8): 1006-1010.

[18] SAKI F, SEHGAL A, PANAHI I, et al. Smartphone-based real-time classification of noise signals using subband features and random forest classifier[C]//Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China: IEEE, 2016: 2204-2208.

[19] VALENTINI-BOTINHAO C, WANG X, TAKAKI S, et al. Investigating RNN-based speech enhancement methods for noise-robust text-to-speech[C]//Proceedings of the 9th ISCA Speech Synthesis Workshop. Sunnyvale, USA: ISCA, 2016: 146-152.

[20] 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: 1-5.

作者简介:

胡沁雯(1999-),女,博士研究生,研究方向:语音增强、语音分离,E-mail: qinwen.hu@smail.nju.edu.cn。

侯仲舒(1998-),男,硕士研究生,研究方向:语音增强、语音分离,E-mail: zhongshu.hou@smail.nju.edu.cn。

乐笑怀(1998-),男,硕士研究生,研究方向:语音增强、语音分离,E-mail: xiaohuaile@smail.nju.edu.cn。

卢晶(1976-),**通信作者**,男,博士,教授,研究方向:声场调控、声信息增强,E-mail: lujing@nju.edu.cn。

(编辑:张黄群)