基于变种注意力的双鉴别器语音增强算法

李正,周斌*
(中南民族大学 计算机科学学院,武汉 430074)

摘 要 日常通信以及说话人识别和语音唤醒等领域的前端任务,都需要干净的语音信号以保证准确的语音可懂度和高效的性能.现有的语音增强算法存在模型参数量大、过度关注评价指标而忽视增强语音真假性等问题.针对上述问题,提出一种基于变种注意力的双鉴别器语音增强算法对含噪语音进行时频域增强.含噪语音信号经过短时傅里叶变换和幂律压缩等一系列操作传入生成器,编码器首先使用稠密卷积模块进行特征提取,后经过维度变化分别利用变种注意力获取时域和频域特征,解码器恢复信号的幅度和复数频谱.最终分别利用评价指标和梅尔频谱训练两个同结构但不同输入的鉴别器.实验结果证明:该模型与SE-Conformer相比,语音质量感知、噪声失真测度和综合质量测度分别提升0.15、0.27和0.25.

关键词 语音增强;变种注意力;双鉴别器
中图分类号 TP399.41 **文献标志码** A **文章编号** 1672-4321(2025)02-0245-08
doi:10.20056/j.cnki.ZNMDZK.20250214

Dual-discriminator speech enhancement algorithm based on variant attention

LI Zheng, ZHOU Bin*
(College of Computer Science, South-Central Minzu University, Wuhan 430074, China)

Abstract In the realms of everyday communication, speaker identification, and voice-activated systems, pristine audio signals are imperative for ensuring clear speech comprehension and optimal performance. The prevailing voice enhancement algorithms are encumbered by substantial model parameters and a disproportionate focus on evaluative metrics, often at the expense of the authenticity of the enhanced speech. To counter these challenges, this study introduces a novel dual-discriminator voice enhancement algorithm that leverages a variant attention mechanism for the enhancement of noisy voice signals in the time-frequency domain. The noisy voice signal undergoes a sequence of transformations, including short-time Fourier transformation and power-law compression, before being processed by the generator. The encoding phase initiates with a dense convolutional module to extract salient features, which are subsequently subjected to dimensionality alterations to harness the variant attention for capturing temporal and spectral characteristics. The decoding phase then reconstructs the signal's amplitude and complex frequency spectrum. The algorithm employs two discriminators of identical architecture but distinct inputs, trained concurrently on evaluative metrics and Mel-spectrogram data. The results indicated that, compared to the SE-Conformer model, the proposed model in this paper achieved significant improvements in terms of PESQ, CBAK, and CVOL, with enhancements of 0.15, 0.27, and 0.25 respectively.

Keywords speech enhancement; variant attention; dual-discriminator

语音增强算法是指从带噪语音信号中恢复出尽可能干净的语音信号,提高噪声条件下语音的质量和可懂度[1].作为说话人识别等领域的前端任务,语音增强算法一直是研究热点.现有的语音增强算法主要分为两类,一类是基于传统信号处理的语音增强算法,另一类则是基于深度学习的语音增强算法.

收稿日期 2024-08-03
* **通信作者** 周斌(1971-),男,教授,博士,研究方向:大数据处理,E-mail:binzhou@mail.scuec.edu.cn
基金项目 湖北省技术创新专项基金资助项目(2019ADC071);中央高校基本科研业务费专项资金资助项目(CZY23006)

基于传统信号处理的语音增强算法,通过对噪声的加性假设和先验估计来恢复清晰的语音,例如维纳滤波[2]以及子空间法[3]等.其具有计算资源消耗低等优点,但在处理复杂多变的噪声环境时效果欠佳.基于深度学习的语音增强算法依据建模方式的不同主要分为两类,一类是基于判别式的语音增强算法[4],一类则是基于生成式的语音增强算法.判别式的语音增强算法直接建立带噪语音与干净语音之间的映射关系,以减少原始语音与增强语音之间的差异.该技术依赖于一对一对应的带噪语音和干净语音数据集,但获取包含各种噪声类型、混响效果以及不同说话者的数据集并不现实.为了解决上述问题,基于生成式的语音增强算法被提出.

生成式语音增强算法的核心在于利用无监督学习来掌握干净语音的特征分布,进而利用特征作为消噪的指导原则.目前其代表是基于生成对抗网络的语音增强算法.生成对抗网络结构(Generative Adversarial Networks,GAN)[5]近年来逐渐被应用于语音增强领域.Pascual等提出一种基于生成对抗网络的时域语音端到端模型[6],其生成器直接在波形级别上处理含噪语音信号,鉴别器判别真假.后续语音增强算法[7-10]大都在此基础上进行提升,但由于在早期研究中目标函数主要聚焦于增强谱图与目标谱图之间的LP-norm距离,导致目标函数与评价指标之间缺乏直接联系,优化损失函数后评价指标分数也未得到提升.MetricGAN[11]直接通过鉴别器来训练语音增强的客观评价指标,并反馈给生成器进行训练优化,从而提升语音质量.但由于其鉴别器容易发生灾难性遗忘,Fu等对MetricGAN加以改进提出MetricGAN+模型,实现了更优的语音增强效果[12].但其过多关注于评价指标,却忽略了增强语音的真实性以及自然性,仍存在提升空间.

Transformer[13]作为近期注意力结构的代表被广泛应用于语音增强领域[14-17],其可以有效获取全局特征但却难于抽取局部特征.卷积可以对局部特征进行很好的建模,但如果是全局特征则需要堆砌非常深的卷积层,在一定程度上削弱了表达全局信息的能力.现有方法尝试将Transformer和卷积相结合,但其大多直接对序列层进行平均取值,将导致全局角度上的建模失真.Conformer[18]用于来捕捉语音波形或频谱图中的长距离依赖信息,它的提出有效解决了上述问题.受此启发,SE-Conformer[19]首次将Conformer应用在基于生成对抗网络的语音增强中,但由于其只研究基于时域的波形语音信号,导致其效果并不明显且模型参数量过大.

综上所述,为了更充分地捕捉特征信息,减少模型参数量,并在利用评价指标训练鉴别器的同时考虑到增强语音的真实性以及自然性,本文提出一种基于变种注意力的双鉴别器语音增强算法(Dual-Discriminator Speech Enhancement Algorithm Based on Variant Attention,DDSE-VA).DDSE-VA的输入为带噪语音的实部、虚部以及幅度频谱值,可以更好地获取时域和频域特征;注意力采用串联变种注意力结构,通过对数据维度的改变分别考虑时频域特征,这将更好地捕捉数据特征、减少模型参数量以及提升模型收敛速度;采用双鉴别器结构,一个利用复杂评价指标进行训练,另一个直接传入语音梅尔频谱图进行训练,由此能更加全面地考虑增强语音的特性.

1 语音增强算法

1.1 模型框架

DDSE-VA采用生成对抗网络结构,生成器负责生成无限接近原始干净语音的增强语音信号,鉴别器负责判断模型输入的是原始干净语音还是增强语音,双方互相博弈训练.它主要由一个生成器、两个同结构但不同输入的鉴别器以及串行变种注意力结构组成,如图1所示.假设输入一个含噪语音信号,经过数据预处理操作将结果作为输入传入生成器.生成器由一个编码器以及两个解码器构成.输入信号经过编码器的稠密卷积模块进行语音信号

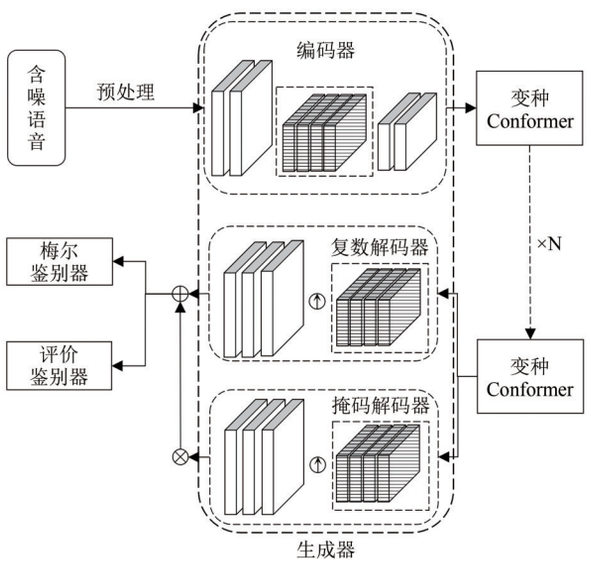

图1 DDSE-VA总框架图

Fig. 1 DDSE-VA framework diagram

特征提取，后利用输入信号的维度转变，通过串行的变种注意力结构分别获取语音信号的时频域特征，再由掩码解码器和复数解码器，分别获得输入信号的幅度频谱和复数频谱，并利用逆短时傅里叶变换获得增强后的语音信号．获取增强语音信号后，将其幅度频谱以及原始干净语音信号的幅度频谱传入一个鉴别器以训练评价指标鉴别器，将两个语音信号的梅尔频谱传入另一个鉴别器以训练真假鉴别器．

1.2　编码器

将噪声视为加性噪声问题，即含噪语音信号由干净语音和噪声相加产生．现假设输入一个含噪波形语音信号 X，经过短时傅里叶变换将该波形语音信号转换为复数频谱 $X_0 \in R^{T\cdot F\cdot 2}$，$T$ 和 F 分别代表语音信号的时间和频率维度．对 X_0 进行幂律压缩即可得到压缩后的幅度频谱 X_m，后经欧拉公式以及实虚部分离等操作可得公式(1)和(2)．将得到的幅度频谱以及分离后的实部和虚部 $X_{in}=\{X_m, X_r, X_i\}$ 作为输入传入编码器．

$$e^{j\theta}=\cos(\theta)+j\sin(\theta), \tag{1}$$

$$X=|X_o|^c e^{jX_p}=X_m\left(\cos X_p+j\sin X_p\right)=X_r+jX_i, \tag{2}$$

编码器主要由两个卷积块以及一个扩张型稠密卷积模块组成，结构如图 1 所示．两个卷积模块结构相似，首先都是进行卷积操作，随后进行实例归一化处理，最终通过 PReLU 激活函数进行非线性转换．第一个卷积模块负责将原始输入信号的 3 个特征进行维度扩展，最终生成具有 C 个通道的特征映射．中间衔接的扩张型稠密卷积模块由 4 组卷积模块组成，每组分别由稠密卷积、实例归一化处理以及 PReLU 激活函数组成，扩张系数依次为 $\{1,2,4,8\}$．扩张卷积在不增加卷积核数量和层级深度的前提下，有效扩展了模型的感知范围．第二个卷积模块负责将输入信号的频率维度减半至 F'，以降低模型的复杂性．

1.3　变种注意力

Conformer 通过融合 Transformer 和卷积网络的优点，被用来捕捉语音波形或频谱图中的长距离依赖信息．由于模块的堆叠导致参数量增加并且对计算资源要求过高，由此提出一种串联的变种 Conformer 结构．通过顺序部署 N 组变种 Conformer 块(两个变种 Conformer 为一组)，分别获取语音信号的时间和频域特征．假设给定一个原始特征映射为 $X_{\text{in}} \in R^{B\cdot T\cdot F\cdot 3}$，其中 B 表示批次大小．通过编码器后可获得 $X'_{\text{in}} \in R^{B\cdot T\cdot F'\cdot C}$，首先将其变形为 $X''_{\text{in}} \in R^{B\cdot F'\cdot T\cdot C}$ 并通过第一个变种 Conformer 块获取其时域特征，输入结果与 X'_{in} 相加以保留最完整的原始特征．再将其变形为 $X'''_{\text{in}} \in R^{B\cdot T\cdot F'\cdot C}$ 传入第二个变种 Conformer 块获取其频域特征并再次与初始特征相加．以上为一组变种 Conformer 流程，通过 N 组结构可以更好地获取输入特征的时频域特征．

变种 Conformer 与原始结构类似，主要由 4 个子模块堆叠在一起，即前馈模块、多头注意力模块、卷积模块以及第二个前馈模块，具体结构如图 2 所示．两个前馈层都由一层用于调整和缩放每个输入特征的归一化层以及两组线性层和随机丢失层组成．多头注意力模块相对 Transformer 结构，采用相对正弦位置编码使得自注意力模块更好地泛化到不同长度的输入信号中．对于卷积模块而言，将其替换为变种卷积模块(主要由深度卷积和逐点卷积组成)．深度卷积允许每个输入通道通过各自独立的滤波器进行卷积运算，若有 D 个输入通道就有 D 组卷积核．该结构的优点是在不增加参数量的情况下，允许模型在每个通道内学习空间特征．如图 2 所示的变种卷积模块，从归一化层开始，随后紧跟的就是逐点卷积和门控线性单元，门控线性循环单元的输出流向由逐点卷积和点对点卷积组成的深度可分离卷积模块，并应用 Swish 激活函数，之后通过另一个逐点卷积层并在最后阶段采用 Dropout 技术进行正则化处理．在变种卷积模块的第一层逐点卷积和最后一层逐点卷积之间加入残差连接，并在每个变种 Conformer 之后添加一个门控线性单元．

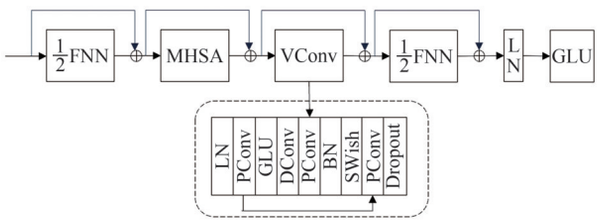

图2　变种注意力框架图

Fig. 2　Variant attention framework diagram

1.4　解码器

解码器部分由掩码解码器和复数解码器组成，它们共同从变种 Conformer 输出中提取信息．掩码解码器负责生成一个特定的掩码与输入信号的幅度值进行逐元素相乘．复数解码器负责直接计算输入信号的实部和虚部．两个解码器结构都采用稠密卷积结构，并通过卷积技术对频率维度进行上采样，具体结构如图 1 所示．在掩码解码器中，首先通过一个卷积层将通道数减少为单一通道，随后通过

PRelu激活函数以及相应卷积层来确定最终掩码值.复数解码器结构与掩码解码器一致,最终输出时不采用任何激活函数.

信号重建首先通过解码器以解耦的方式从串联变种注意力结构中提取输出.解码包含两条路径,掩码解码器利用掩码与输入信号的幅度值逐元素相乘,复数解码器直接生成增强语音的实部与虚部.后将增强后的掩码幅度与含噪语音信号的相位值相结合,形成增强后的复数频谱,再将该复数频谱与复数解码器的输出逐元素相加得到最终的复数频谱.最后,逆顺序执行数据预处理操作,完成整个信号重建过程.

1.5 双鉴别器

早期语音增强算法研究仅关注增强语音和原始干净语音之间的差异距离,导致目标损失函数与评价指标之间缺乏直接联系.复杂的评价指标很难直接集成到损失函数中,所以MetricGAN直接通过评价指标训练鉴别器.该方法虽然提高了模型的评价指标得分,却忽略了增强语音的真假性,导致生成的语音不够自然.为解决上述问题,提出一种结构相同但输入不同的双鉴别器结构.

一个鉴别器将增强语音和干净语音的频谱图作为输入,通过模型训练可以准确预测增强语音信号的评价指标分数;另一个鉴别器将增强语音和干净语音的梅尔频谱作为输入,可有效提升降噪处理之后语音的自然性并突出语音信号的关键特征.双鉴别器的具体结构如图3所示,两个鉴别器首先都由4个卷积单元组成,每个卷积单元包含一个二维卷积层、实例归一化处理层以及PReLU激活函数.在第1个二维卷积和第4个二维卷积之间引入残差连接,该设计有助于提升模型对复杂特征的学习能力并保持模型训练的稳定性.在卷积单元之后附设一个最大池化层和两个全连接层,以缓解梯度消失和过拟合问题.

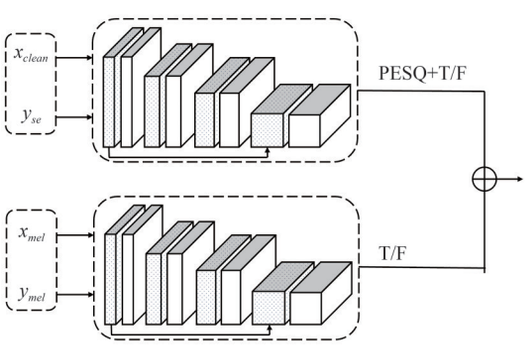

图3 鉴别器框架图

Fig. 3 Discriminator framework diagram

1.6 损失函数

损失函数主要由生成器和鉴别器两部分组成.生成器的损失函数考虑了时域和频域两方面,如下公式所示:

$$\begin{cases} L_{\text{time}} = E_{x,y}\|x - y\|_1 \\ L_{\text{mag}} = E_{x_{\text{clean}},y_{\text{se}}}\|D(x_{\text{clean}},y_{\text{se}}) - 1\|^2 \\ L_{\text{tf}} = E_{x_{\text{clean}},y_{\text{se}}}\|x_{\text{clean}} - y_{\text{se}}\|^2 + E_{x^r_{\text{clean}},y^r_{\text{se}}}\|x^r_{\text{clean}} - y^r_{\text{se}}\|^2 + \\ \quad E_{x^i_{\text{clean}},y^i_{\text{se}}}\|x^i_{\text{clean}} - y^i_{\text{se}}\|^2 \end{cases}, \tag{3}$$

其中时域部分损失函数由x和y计算得出,x和y分别代表原始语音和增强语音的时域波形图.频域损失函数计算不仅使用原始语音和增强的幅度值$\{x_{\text{clean}},y_{\text{se}}\}$,还使用两种语音信号复数频谱的实部和虚部$\{x^r_{\text{clean}},y^r_{\text{se}}\}$和$\{x^i_{\text{clean}},y^i_{\text{se}}\}$.

为更好地均衡生成器各部分损失函数的作用,增设3个损失函数μ_1、μ_2、μ_3,具体公式如下:

$$L_{\text{gan}} = \mu_1 L_{\text{time}} + \mu_2 L_{\text{mag}} + \mu_3 L_{\text{tf}}. \tag{4}$$

鉴别器损失函数融合PESQ评价标准,同时为提升增强语音的自然性将干净语音和增强语音的梅尔频谱$\{x_{\text{mel}},y_{\text{mel}}\}$作为输入传入鉴别器,具体损失函数如下:

$$L_d = E_{x_{\text{clean}}}\|D(x_{\text{clean}},x_{\text{clean}})\|^2 + E_{x_{\text{clean}},y_{\text{se}}}\|D(x_{\text{clean}},y_{se}) - Q_{\text{PESQ}}\|^2 + E_{x_{\text{mel}},y_{\text{mel}}}\|D(x_{\text{mel}},y_{\text{mel}}) - 1\|^2. \tag{5}$$

2 实验与结果分析

2.1 数据集

实验采用公开数据集VoiceBank[20]+DEMAND[21],该数据集由训练集和测试集组成.训练集通过28个不同的说话人录制不同声音,并向其添加10种不同类型的噪声(包含2种人造噪声以及2种来自DEMAND噪声)以构建出11572条语音数据.训练集混合噪声的分贝范围为0 dB~15 dB,5 dB一增加.测试集通过按信噪比{2.5 dB,7.5 dB,12.5 dB,17.5 dB}添加5种未在训练集出现的DEMAND噪声,以构建出824条测试语音数据.数据集里噪声多样,包含办公室等公共环境噪声、公共汽车等交通噪声以及风声等自然环境噪声等.

2.2 实验配置及环境

本文首先将数据集的采样率从48 K统一下采样至16 K.训练集被等分为2 s一段且50%重叠率的等长语音数据,测试集则保留其原始的完整音频长度.采取512的汉明窗函数、长度为256的位移以及50%的重叠,通过短时傅里叶变换生成257个频率

段.该设置提供了一个较好的平衡点，有效避免了较短或较长窗口函数的缺点，有效平衡频谱和时间分辨率，减少频谱泄漏和边界效应，从而提升了信号分析的精度和连续性.模型训练采用AdamW参数优化器，采用0.001的学习率并根据训练批次将学习率减半更新.训练一共进行75个批次，采用大小为1的Batchsize，各个损失函数权重依次为0.3、0.7、1、0.01.本文实验环境使用Linux操作系统，通过RTX 4070 12 GB显卡训练模型，具体配置如表1所示.

表1　实验环境配置表

Tab. 1　Experimental environment configuration table

配置名称	型号
CPU	Intel i5 13490F
GPU	NVIDIA GeForce RTX 4070
操作系统	Ubuntu 22.04
深度学习框架	Pytorch 1.10.0
加速环境	CUDA 11.3
语言	Python 3.8.0

2.3　评价指标

为评估模型增强后的语音质量，采用多种指标，包括主观和客观评价.主观指标涵盖失真测度（CSIG）、噪声失真测度（CBAK）和综合质量测度（COVL），它们的评分范围均为1到5.客观指标则包含语音质量感知（PESQ）、短时可懂度（STOI）和分段信噪比（SSNR）.PESQ的评分范围在-0.5到4.5之间，STOI的评分介于0到1之间.由于信噪比（SSNR）与主观质量的相关性较低，因此选用基于帧的SSNR来衡量语音质量，其评分范围为0~30.以上指标的数值越高，表示语音质量越佳.

2.4　对比实验

为验证模型的有效性，特与经典模型以及近期模型进行对比.选取的基于深度学习的语音增强算法包含SEGAN、SE-Conformer、MetricGAN、MANNER[22]、FRCRN[23]、TSTNN[24]、PHASEN[25]以及DCCARN[26].T代表只考虑时域特征，F代表只考虑频域特征，TF代表同时考虑时频域特征.SEGAN模型直接将时域波形图传入模型并第一次将生成对抗网络应用到语音增强方向.SE-Conformer第一次将Conformer注意力结构引入到基于生成对抗网络的语音增强模型，提升了增强模型获取全局和局部特征的能力.MetricGAN首次利用评价指标训练鉴别器，有效提升了增强模型的评价指标得分.PHASEN和FRCRN都是以频域特征作为输入的增强算法，二者不同之处在于PHASEN利用卷积神经网络作为注意力结构，而FRCRN则利用一种变体长短时记忆网络作为注意力结构，以更少的参数获取更高的性能.TSTNN是一种以Transformer网络结构为注意力结构的语音增强模型，其提出双阶段Transformer结构用于提取语音信号的全局和局部上下文信息.具体实验结果如表2所示.

表2　对比模型实验结果

Tab. 2　Comparative model experimental results

模型	参数量	特征	PESQ	STOI	CSIG	CBAK	COVL	SSNR
SEGAN	97.47 M	T	2.16	0.92	3.48	2.94	2.80	7.73
MetricGAN		F	2.86		3.99	3.18	3.42	
PHASEN	8.8 M	F	2.99		4.21	3.55	3.62	10.18
TSTNN	0.92 M	T	2.96	0.95	4.10	3.77	3.52	9.70
SE-Conformer		T	3.13	0.95	4.45	3.55	3.82	
MANNER	24.7 M	T	3.21	0.95	4.53	3.65	3.91	9.56
FRCRN	6.9 M	TF	3.21		4.23	3.64	3.73	2.31
DCCARN		TF	2.83		3.91	3.60	3.43	
DDSE-VA	1.5 M	TF	3.28	0.96	4.51	3.82	4.07	10.70

由表2可知，DDSE-VA实验结果除MANNER外在各评价指标上均全部超越其余对比模型，分别与对比模型在PESQ、STOI、CBAK、COVL和SSNR的最高得分超出0.07、0.01、0.05、0.16和0.52.这得益于双鉴别器结构的设计，同时考虑增强语音的频谱图和梅尔频谱真假性，使生成器生成的语音更加自然且可懂度更高.利用评价指标训练鉴别器可有效提升提高语音的语音质量感知得分.将原始Conformer的卷积模块替换为深度可分离卷积模块，其中深度卷积允许每个通道通过各自的滤波器进行卷积操作，深度卷积之后紧跟一层逐点卷积将每个输出通道与其他通道相结合以学习跨通道特征.该操作在参数量减少的情况下仍能够捕捉到输入数据的重要特征，使DDSE-VA虽然在CSIG评价指标上略低于MANNER模型，但其模型参数大大降低且远低于MANNER模型参数，其余指标均超越MANNER模型.由特征列可知，传统时频域语音增强算法直接将语音的幅度值和相位值传入模型，导致模型抽取特征不充分并容易造成伪影问题.DDSE-VA模型以语音的频谱以及分离后的实部和虚部作为输入并采用串行变种注意力结构分别获取输入的时域和频域特征，以充分获取模型输入的局部和全局特征.

为更好地展示各个模型之间的性能差异，选取对比试验中的3个经典模型SEGAN、MetricGAN、PHASEN以及DDSE-VA结果进行可视化处理，分别展示模型生成增强语音的时域波形图和频谱图.可视化图（图4）共分为两列，第1列为语音的时域波形

图,第2列为语音的频谱图.由左至右由上至下依次为原始干净语音、含噪语音、SEGAN、MetricGAN、PHASEN以及DDSE-VA四个模型增强语音的可视化结果图.为更好展现语音的频域特征,频域图以分贝为坐标.以第1行原始干净语音的波形图和频谱图做基线,可看出第1行的含噪语音含有大量不规则噪声.由第2行可视化图可知SEGAN和MetricGAN模型在时域部分消噪效果不错,但由于其模型输入只考虑时域特征,导致频域部分效果不佳.由第3行可视化图可知,PHASEN虽考虑频域特征输入,但DDSE-VA得益于串行变种注意力结构,效果明显优于PHASEN.由以上数据可知,基于变种注意力结构的双鉴别器模型在进行语音降噪时,处理噪声细节部分更加得力,在时频域特征的各个频段都可以有效地进行语音消噪,在语音增强领域具有较好的性能表现.

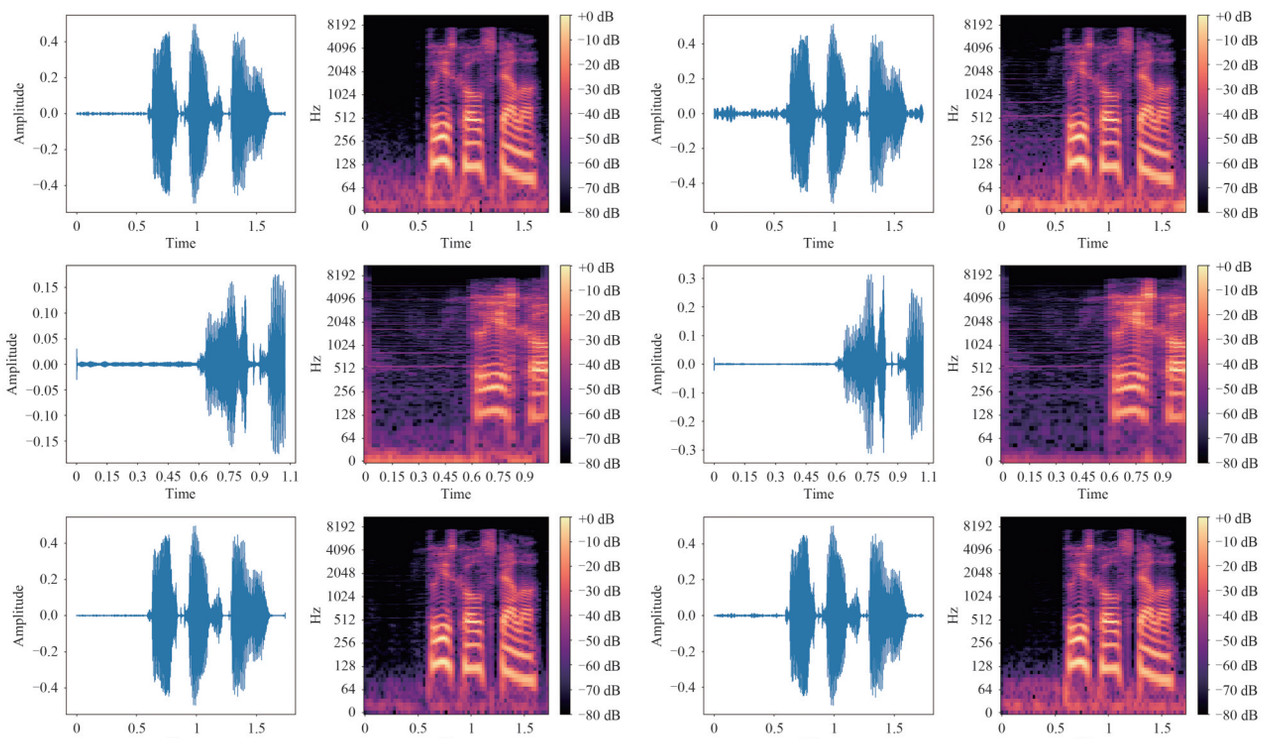

图4 对比实验可视化图

Fig. 4 Comparative experiment visualization chart

2.5 消融实验

为验证本文所设计结构的有效性,特进行下述消融实验.本文消融实验设计由六部分组成:采用并行变种Conformer结构、去掉变种Conformer结构、使用原始Conformer替换变种Conformer结构、采用原始单鉴别器结构、去掉鉴别器损失函数梅尔频谱部分、同时采用单鉴别器并去掉变种注意力结构.通过对上述实验设计的单独实现以验证变种Conformer和双鉴别器结构的作用所在.具体实验结果如表3所示.

表3 消融实验结果

Tab. 3 Ablation experiment results

模型	PESQ	STOI	CSIG	CBAK	COVL	SSNR
Parallel-VCon	3.20	0.96	4.26	3.73	3.89	9.87
No VCon	2.99	0.95	4.16	3.64	3.70	9.65
With Con	3.18	0.96	4.40	3.73	3.89	10.56
One Disc	3.16	0.95	4.41	3.70	3.86	10.60
No Mel	3.23	0.95	4.46	3.78	3.99	10.65
No DVC	2.96	0.95	4.17	3.60	3.71	9.64
DDSE-VA	3.28	0.96	4.51	3.82	4.07	10.70

Parallel-VCon采用并行变种注意力结构其余网络结构不变,通过表3可知采用并行变种注意力结构同时获取输入特征的时域和频域特征再进行特征融合处理的效果不如串行注意力结构.通过串行结构以及维度转变分别获取输入特征的时域特征和频域特征,可充分获取输入的全局以及局部特征.No VCon代表不使用变种Conformer结构,即不采用注意力结构,其余网络结构保持不变,这导致输入特征获取不充分,模型效果下降.With Con代表使用原始Conformer结构,其余结构不变,失去深度可分离卷积模块以及残差连接的加持,模型效果出现下降.One Disc代表使用单鉴别器,训练鉴别器时只考虑频谱特征以及评价指标.No Mel代表鉴别器损失

函数不考虑梅尔频谱部分.No DVC代表使用单鉴别器并去掉变种注意力结构.由上可知,本文消融试验所设计的6组网络结构相对于DDSE-VA网络结构,在除STOI评价指标外其余指标均有不同程度的下降.因此,该消融实验可证明本文在DDSE-VA采用的结构均发挥了不同程度的作用.

3　总结

本文提出了基于变种注意力结构的双鉴别器语音增强算法,以获取更好的语音降噪效果.为解决幅度值和相位值相加存在误差、易产生伪影的问题,本文采用复数频谱及其实部虚部以获取更好的时频域特征.采用串行变种注意力结构分别获取模型输入的时域和频域特征,以更小的参数量获取更详细的全局和局部特征.采用双鉴别器结构,有效提升了增强语音的可懂性以及自然性.通过对比实验验证了本文模型结构在数据集VoiceBank+DEMAND上取得有效的降噪效果,又经消融实验验证各新增结构的有效性.

尽管本文算法在多变噪声环境下表现出色,但仍有改进空间.未来工作将进一步提高算法对噪声的鲁棒性以提高其在极端噪声环境下的性能,提升算法结构以适应更广泛的应用需求.

参　考　文　献

[1] 范君怡, 杨吉斌, 张雄伟, 等. 基于Transformer的单通道语音增强模型综述[J]. 计算机工程与应用, 2022, 58(12): 25-36.

[2] EXTRAPOLATION I. Smoothing of stationary time series with engineering applications[J]. Policy, 1966, 10: 23.

[3] DENDRINOS M, BAKAMIDIS S, CARAYANNIS G. Speech enhancement from noise: A regenerative approach [J]. Speech Communication, 1991, 10(1): 45-57.

[4] XU Y, DU J, DAI L R, et al. An experimental study on speech enhancement based on deep neural networks [J]. IEEE Signal Processing Letters, 2014, 21(1): 65-68.

[5] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks [J]. Communications of the ACM, 2020, 63(11): 139-144.

[6] PASCUAL S, BONAFONTE A, SERRÀ J. SEGAN: Speech enhancement generative adversarial network [C]// Interspeech 2017.Stockholm: ISCA, 2017: 3642-3646.

[7] PHAN H, MCLOUGHLIN I V, PHAM L, et al. Improving GANs for speech enhancement [J]. IEEE Signal Processing Letters, 2020, 27: 1700-1704.

[8] PANDEY A, WANG D. On adversarial training and loss functions for speech enhancement [C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018: 5414-5418.

[9] ABDULATIF S, ARMANIOUS K, GUIRGUIS K, et al. AeGAN: Time-frequency speech denoising *via* generative adversarial networks [C]//2020 28th European Signal Processing Conference (EUSIPCO). Amsterdam: IEEE, 2021: 451-455.

[10] BABY D, VERHULST S. Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty [C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).Brighton: IEEE, 2019: 106-110.

[11] FU S W, LIAO C F, TSAO Y, et al. MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement [C]//International Conference on Machine Learning. Sanya: PMLR, 2019: 2031-2041.

[12] FU S W, YU C, HSIEH T A, et al. MetricGAN+: An improved version of MetricGAN for speech enhancement [EB/OL]. 2021: 2104.03538. https://arxiv. org/abs/2104.03538v2

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017, 30:5998-6008.

[14] JIANG W, SUN C, CHEN F, et al. Low complexity speech enhancement network based on frame-level swin transformer[J]. Electronics, 2023, 12(6): 1330.

[15] 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: IEEE, 2021: 7098-7102.

[16] YU W, ZHOU J, WANG H, et al. SETransformer: Speech enhancement transformer [J]. Cognitive Computation, 2022, 14(3): 1152-1158.

[17] RAMESH K, XING C, WANG W, et al. Vset: A multimodal transformer for visual speech enhancement [C]//ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021: 6658-6662.

[18] GULATI A, QIN J, CHIU C C, et al. Conformer: Convolution-augmented transformer for speech recognition [EB/OL]. 2020: 2005.08100. https://arxiv. org/abs/2005.08100v1

[19] KIM E, SEO H. SE-conformer: Time-domain speech enhancement using conformer [C]//Interspeech 2021. Brno:ISCA, 2021: 2736-2740.

[20] VEAUX C, YAMAGISHI J, KING S. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database [C]//2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE). Gurgaon: IEEE, 2013: 1-4.

[21] 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. Montreal: ASA, 2013: 1-6.

[22] 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: IEEE, 2022: 7842-7846.

[23] ZHAO S, MA B, WATCHARASUPAT K N, et al. FRCRN: Boosting feature representation using frequency recurrence for monaural speech enhancement [C]//2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022, Singapore: IEEE, 2022: 9281-9285.

[24] WANG K, HE B, ZHU W P. TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain [C]//I2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021: 7098-7102.

[25] YIN D, LUO C, XIONG Z, et al. PHASEN: A phase-and-harmonics-aware speech enhancement network [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 9458-9465.

[26] 余本年, 詹永照, 毛启容, 等. 面向语音增强的双复数卷积注意聚合递归网络[J]. 计算机应用, 2023, 43(10): 3217-3224.

(责编 曹东,校对 雷建云)