文章编号：2096-1618(2025)02-0137-06

基于注意力机制和掩码学习的GAN语音增强算法

李彤岩， 裴浩延， 裴 燕， 陈 旭， 王 涛
(成都信息工程大学通信工程学院，四川 成都 610225)

摘要：语音增强是自动语音识别的重要组成之一，近年来，生成对抗网络及其变体模型在语音增强中的建模能力逐渐增强，但仍有泛化能力弱、无法适应低信噪比环境等问题。对此，提出一种结合注意力机制双向长短期记忆网络及掩码学习的GAN语音增强模型。该框架创新了语音增强机制，利用双向长短期记忆网络及注意力层作为生成对抗网络的生成器，并引入掩码学习进行频谱重构，将滤波后的信号与原始信号进行叠加得到增强信号，输入判别器后，两个网络相互博弈达到语音增强的目的。采用TIMIT数据集，通过对比语音质量客观评估和短时客观可懂度等语音评价指标，在不同信噪比环境下对该模型进行评估。实验结果表明，该模型的语音增强效果相比基准生成对抗网络等模型平均提升了11.8%，在噪声干扰大的环境下仍有较强的声学建模能力。

关 键 词：语音增强；生成式对抗网络；注意力-BLSTM；掩码重构；低信噪比
中图分类号：TP391.1 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2025.02.003

0 引言

语音增强在语音通信、语音识别和语音合成等领域广泛应用，可有效提高语音信号的质量和整体表现效果。然而，传统的语音增强技术往往只能通过增加滤波器和降噪处理等方法来减少周围噪声和改善语音信号的质量，而不能对语音信号进行有效的重构和增强。因此，近年来，学术界和工业界对于开发更加高效、精确的语音增强技术的需求越来越高。生成对抗网络(generative adversarial networks，GAN)技术的出现为语音增强提供了新的思路和方法。GAN可以通过构建合适的神经网络模型来生成更加逼真的语音信号，并且可以根据语音信号本身进行增强，具有很强的自适应性和适用性。因此，GAN语音增强算法的研究具有非常重要的意义和价值。

目前，国内外已经有一些关于GAN语音增强算法的相关研究，主要包括以下几个方面的内容：

发展初期学者提出了基于卷积神经网络(convolutional neural networks，CNN)的GAN语音增强算法的研究。GAN网络在计算机视觉领域获得很好的效果，SEGAN首次将它应用在语音增强领域[1]，这种算法主要利用CNN在特征提取和分类方面的优势，同时结合GAN在生成器和判别器的共同作用，可以有效提高语音信号的质量和主要特征。基于注意力机制[2-3]的GAN语音增强算法的研究被提出，这种算法主要考虑语音信号中不同特征之间的相关性和关联性，利用注意力机制进行特征提取和重构，可以更有效地提高语音信号的质量和表现效果。

随后，Wang等[4]在Transformer模型提出一种基于自注意力的架构，称为两阶段自注意力机制神经网络(two-stage transformer based neural network for speech enhancement in the time domain，TSTNN)，用于时域中的端到端语音去噪。通过屏蔽模块创建一个掩码，该掩码将与编码器输出相乘。最后，解码器使用屏蔽编码器功能重建增强语音。为更好地捕获音频信号中的局部特征和全局依赖，Kong等[5]提出结合CNN和Transformer的模型Conformer，在语音识别领域取得了显著的成果。同时，Cao等[6]提出一种基于Conformer的MetricGAN-CMGAN，用于在振幅谱和复数域频谱上进行语音增强。该生成器包含了两层基于时频的Conformer模块，能够捕获时间域和频率域的长距离依赖和局部特征。指标判别器的引入在提高生成语音质量的同时，不会对其他指标产生不利影响。

在这些研究中，结合不同的神经网络结构和机制，可以得到不同领域的优秀语音增强算法。但上述模型对于语音信号等长序列节点的分析能力有所欠缺，且计算资源消耗较大。

为改善低信噪比下的声学建模问题，本文提出一种结合注意力机制双向长短期记忆网络及掩码学习的GAN语音增强算法。该算法使用基于注意力机制的BLSTM(bi-directional long short-term memory)作为生成器，两个互为反向拷贝的LSTM捕捉特定时间步骤的过去和未来输入特征，同时采用自注意层将整个输

收稿日期：2023-09-25
基金项目：四川省科技厅资助项目(2023YFS0422)
通信作者：李彤岩. E-mail：lty@cuit.edu.cn.

入序列不同位置的信息联系起来,通过掩码学习增强生成器定位的准确性。将卷积神经网络作为判别器,输出二元分类结果。研究采用对抗损失、均方误差(mean squared error,MSE)和掩码损失的加权和作为生成器的损失函数,采用交叉熵损失作为鉴别器的损失函数,以此达到较低信噪比下的语音增强的作用。

1 基于生成对抗网络的语音增强

生成对抗网络[7]由生成器和判别器两部分组成,这两部分可以为卷积神经网络或循环神经网络等。在网络训练过程中,生成器和判别器相互博弈,最终达到一种纳什平衡。

GAN 生成器的输入为含有噪声的语音信号,输出为消除噪声后的语音信号,其结构通常采用循环神经网络或卷积神经网络。判别器的输入为含有噪声的语音信号或生成器产生的假样本,输出为真或假的二元分类结果,其结构也通常采用卷积神经网络。研究表明,在与语音技术有关的应用中,利用 GAN 旨在学习一个合适的映射函数,并准确地重建增强的语音,同时保留语音质量和可懂度。生成器的最小二乘函数如下:

$$L_G=\frac{1}{2}E_{z\sim p_z(z),x\sim p_{\text{data}}(x)}[(D(\boldsymbol{G}(z,x),x)-1)^2]+\lambda E_{z\sim p_z(z),\hat{x},x\sim p_{\text{data}}}\|\boldsymbol{G}(z,x)-\hat{x}\|_1 \quad (1)$$

判别器的最小二乘函数如下:

$$L_D=\frac{1}{2}E_{z\sim p_z(z),x\sim p_{\text{data}}}[D(\boldsymbol{G}(z,x),x)^2]+\frac{1}{2}E_{x,\hat{x}\sim p_{\text{data}}(x,\hat{x})}[(D(\hat{x},x)-1)^2] \quad (2)$$

其中,$\hat{x}$ 为纯净语音信号,z 为噪声信号,带噪语音信号 $x=\hat{x}+z$,通过训练生成器,当损失函数 L_G 最小化时,将 x 映射为其对应的 $\hat{x}$,即可得到生成的语音信号。鉴别器则通过最小化损失函数 L_D 来区分真实语音信号和虚假语音信号。

2 结合语音掩码的 LAGAN 模型

本文提出一种基于注意力 BLSTM 和掩码学习的 GAN 语音增强算法框架(Mask-LAGAN)。在该框架中,使基于注意力机制的 BLSTM 网络作为生成器,不仅生成去噪语音信号,还生成掩码信号,用于对原始语音信号进行加权处理。为实现这一目标,BLSTM 训练了两个 LSTMs 来处理输入序列。第二个 LSTM 是第一个 LSTM 的反向拷贝,旨在捕捉特定时间步骤的过去和未来输入特征,在处理长序列节点时表现出良好的性能。另外,注意力层用于将整个输入序列的不同位置的信息联系起来,利用缩放的点积注意力计算注意力分布。最后,通过掩码学习,生成器能够更准确地定位噪声部分和语音部分,从而更好地去除噪声。

为评估生成器的性能,采用卷积神经网络作为判别器,用于输出真或假的二元分类结果。在生成器的损失函数中,采用对抗损失、均方误差和掩码损失的加权和。对抗损失用于促使生成器生成逼真的增强语音,MSE 损失用于衡量生成的增强语音与目标语音之间的差异,掩码损失通过对比生成的增强语音和目标语音之间的差异,指导生成器准确地识别噪声部分和语音部分,从而更好地抑制噪声干扰。鉴别器的损失函数采用交叉熵损失。

通过以上的网络结构和损失函数设计,本文提出的模型可以在较低信噪比下实现有效的语音增强,图 1展示了模型框架。

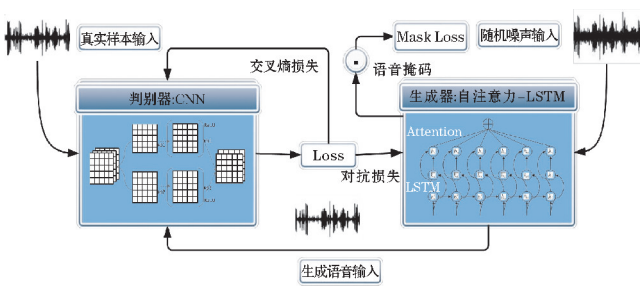


图 1 Mask-LAGAN 语音掩码模型

2.1 生成器:注意力机制 BLSTM 结合掩码学习

采用结合注意力机制的 BLSTM 作为 GAN 的生成器,并通过语音掩码算法进行语音增强,增强声学建模能力。生成器架构如图 2 所示。

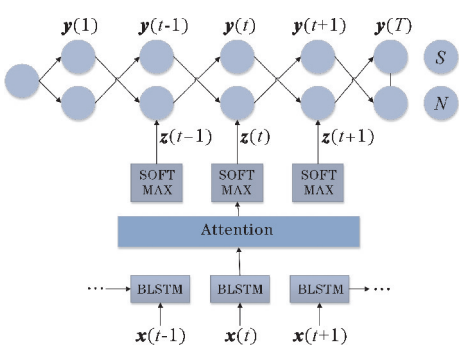


图 2 GAN 生成器网络结构

长短时记忆(long short-term memory,LSTM)网络已被证明能有效地执行长序列数据分析任务[8-9]。不同于前馈神经网络,它只能根据当前的输入特征来预测输出标签,而 LSTM 可以保存从历史上获得的重要相关信息。一个 LSTM 单元通过控制 3 个门来执行记

忆和遗忘的功能,其执行步骤如下。

输入带噪语音序列 $\boldsymbol{x}=\{x_1,x_2,x_3,\cdots,x_{n-1},x_n\}$ 至网络结构中,各步骤公式如下:

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{i_x}\cdot[\boldsymbol{h}_{t-1},\boldsymbol{c}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i) \tag{3}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{f_x}\cdot[\boldsymbol{h}_{t-1},\boldsymbol{c}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f) \tag{4}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{o_x}\boldsymbol{x}_t+\boldsymbol{W}_{o_h}\boldsymbol{h}_{t-1}+\boldsymbol{W}_{o_c}\boldsymbol{c}_{t-1}+\boldsymbol{b}_o) \tag{5}$$

$$\boldsymbol{c}_t=i_t\varphi(\boldsymbol{W}_{c_x}\cdot[\boldsymbol{h}_{t-1},\boldsymbol{c}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_c) \tag{6}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\cdot\tanh(\boldsymbol{c}_t) \tag{7}$$

其中,$\boldsymbol{y}_t$ 为输出纯净语音序列,$\boldsymbol{i}_t$ 为输入门,$\boldsymbol{f}_t$ 为遗忘门,$\boldsymbol{o}_t$ 为记忆单元,通过 3 个门控单元得到 t 时刻状态 $\boldsymbol{c}_t$ 和当前节点输出 $\boldsymbol{h}_t$,$\boldsymbol{W}_{i_x}$,$\boldsymbol{W}_{o_h}$,$\boldsymbol{W}_{o_c}$ 等为权重矩阵,$\boldsymbol{b}$ 为偏置向量,σ 为激活函数 Sigmoid。

而 BLSTM[10-11] 训练两个 LSTMs 对输入序列进行处理。第二个 LSTM 是第一个 LSTM 的反向拷贝,旨在捕捉一个特定时间步骤的过去和未来输入特征。在带噪语音序列输入网络后,前向 LSTM 和后向 LSTM 的状态输出为:

$$\overrightarrow{\boldsymbol{h}}_t=\overrightarrow{\mathrm{LSTM}}(\boldsymbol{x}_t,\overrightarrow{\boldsymbol{h}}_{t-1}) \tag{8}$$

$$\overleftarrow{\boldsymbol{h}}_t=\overleftarrow{\mathrm{LSTM}}(\boldsymbol{x}_t,\overleftarrow{\boldsymbol{h}}_{t+1}) \tag{9}$$

模型利用 n 个隐状态线性加权求和将不同长度的语音序列编码,在这之前,利用注意力层进行权重分配计算,注意力机制[10] 使神经网络检查输入的特定区域的语音信号,以降低任务复杂性,并排除不相关的信息。其将权重 w_j 分配给每一帧特征 $\boldsymbol{u}_j$,最后通过加权求和函数生成隐藏的声学特征向量 $\boldsymbol{r}$:

$$\boldsymbol{p}_j=\tanh(\boldsymbol{W}_u\boldsymbol{u}_j+\boldsymbol{b}_u),\boldsymbol{p}_j\in[-1,1] \tag{10}$$

$$w_j=\frac{e^{\boldsymbol{u}_j}}{\sum_{t=1}^{N}e^{\boldsymbol{u}_t}},\sum_{j=1}^{N}w_j=1 \tag{11}$$

$$\boldsymbol{r}=\sum_{t=1}^{N}w_j\boldsymbol{u}_j,\boldsymbol{r}\in R^{2L} \tag{12}$$

其中,w_j 和 $\boldsymbol{b}_u$ 为注意力层的权重、偏差。

最后通过一种信号近似方法训练一个语音掩码估计器[12],使干净的语音和估计的语音的频谱幅度之间的差异最小。语音掩码损失为:

$$L_m=E_{\boldsymbol{x}\sim p_{\mathrm{data}}}\|\mathrm{IRM}\otimes\boldsymbol{X}-\hat{\boldsymbol{X}}\|_2 \tag{13}$$

$$\mathrm{IRM}=\sqrt{\frac{\hat{\boldsymbol{X}}(t,f)^2}{\hat{\mathrm{X}}(t,f)^2+\hat{\boldsymbol{V}}(t,f)^2}} \tag{14}$$

其中,IRM 为理想语音掩码,$\hat{\boldsymbol{X}}(t,f)$ 为语音频谱,$\hat{\boldsymbol{V}}(t,f)$ 为噪声频谱。

研究采用均方误差 MSE 作为生成器重构损失,结合对抗损失和掩码损失,列出生成器损失函数如下:

$$L_G=\frac{1}{2}E_{z\sim p_z(z),\boldsymbol{x}\sim p_{\mathrm{data}}(\boldsymbol{x})}[(D(\boldsymbol{G}(\boldsymbol{z},\boldsymbol{x}),\boldsymbol{x})-1)^2]+\lambda E_{z\sim p_z(z),\hat{\boldsymbol{x}},\boldsymbol{x}\sim p_{\mathrm{data}}}\|\mathrm{MSE}(\boldsymbol{G}(\boldsymbol{z},\boldsymbol{x}))-\hat{\boldsymbol{x}}\|_1+\alpha L_m \tag{15}$$

其中,α 为控制语音掩码损失的系数,在文献[13]中被规定为 30 时训练效果较好。

2.2　判别器

在语音增强模型中,目标函数与评价指标往往不会直接关联[14-16],因此在最优化目标函数后,也会存在评价指标未达标的情况。使用语音质量感知评价 PESQ 指标作为标签,采用 3 个卷积块作为判别器,其间包含 PReLU 激活,之后添加 Sigmoid 激活,以训练判别器来估计最大 PESQ 指标,网络结构如图 3 所示。

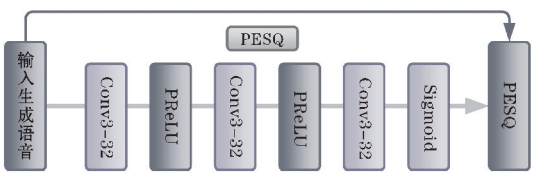


图 3　判别器网络结构

判别器损失函数如下:

$$L_D=E_{z\sim p_z(z),\boldsymbol{x}\sim p_{\mathrm{data}}}[(D(\boldsymbol{G}(\boldsymbol{z},\boldsymbol{x}),\boldsymbol{x})-1)^2]+E_{\boldsymbol{x},\hat{\boldsymbol{x}}\sim p_{\mathrm{data}}(\boldsymbol{x},\hat{\boldsymbol{x}})}[(D(\hat{\boldsymbol{x}},\boldsymbol{x})-Q_{\mathrm{PESQ}})^2] \tag{16}$$

其中,D 指的是判别器,QPESQ 指的是正则化的 PESQ 得分,将其归一化为范围[0,1]。

3　实验设置及结果分析

3.1　实验设置

本实验环境为 64 位 Windows 操作系统,采用 Kaldi、Pytorch 等开发工具,以德州仪器公司、麻省理工学院和 SRI 国际公司合作建立的语料库 TIMIT 为数据集,并验证本文所提出的模型的可靠性。数据库中收集了 438 名男性和 192 名女性(70% 的演讲者为男性,大部分为白人),语料库中的训练集和测试集为不同的发音人。在训练集中,将语料库 60% 干净的语料与背景噪声混合,添加信噪比为0 dB、3 dB、6 dB的随机噪声;研究采取 TIMIT 中 20% 音频数据作为验证集,以调整超参数;其余 20% 作为测试集,并添加信噪比为0 dB、3 dB、6 dB的随机噪声,设置语音采样频率为 16 kHz[17]。

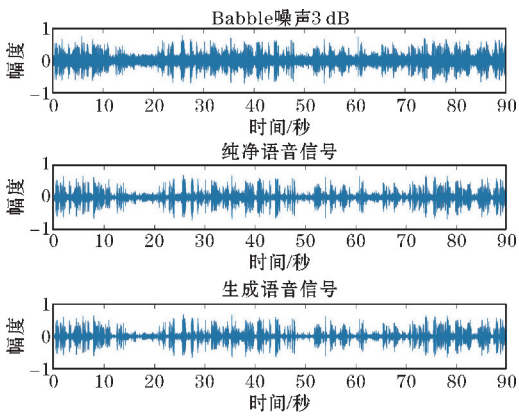
生成器中,输入 40 维 MFCC 特征参数,每个 BLSTM 层之后,一个有 320 个节点的线性层被用来结合前向和后向 LSTM 输出。注意力层包括 10 个宽度为 100 的中心卷积滤波器。

激活函数为 tanh 来解决梯度消失问题,输出层使用 Softmax 网络进行分类处理。

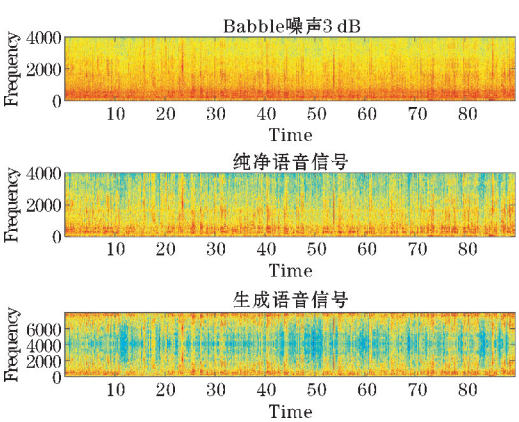
判决器网络设置参考文献[18-19]。本文通过语音增强领域普遍认可的语音质量客观评估(perceptual evaluation of speech quality,PESQ)、短时客观可懂度(short-time objective intelligibility,STOI)和综合语音可懂度指标(composite speech intelligibility gain,CSIG)等评价指标对模型进行评估[20]。PESQ 是一种用于评估语音质量的客观指标,可以对单个语音信号的质量进行评估,并提供一个0~4.5的得分,表示语音质量的好坏;STOI 用于评估经过失真或干扰的语音信号的可懂度,其取值范围从0~1,表示从完全不可懂到完全可懂的语音信号;CSIG 基于语音信号和背景噪声的能量分布,并通过计算信号与干扰加噪声的比例,确定语音信号的清晰度,从而得出 CSIG 得分。

3.2 结果分析

通过对纯净语音信号叠加信噪比分别为3 dB、6 dB、9 dB的 Babble 噪声、Buc 噪声,分别采用 Mask-LAGAN 语音增强算法与 LSTM、SEGAN 语音增强算法等进行对比实现,计算各模型对应的评价指标。图4、5分别为添加3 dB的 Babblle 噪声和 Buc 噪声后,模型生成语音与纯净语音的时域波形图、频谱图对比。

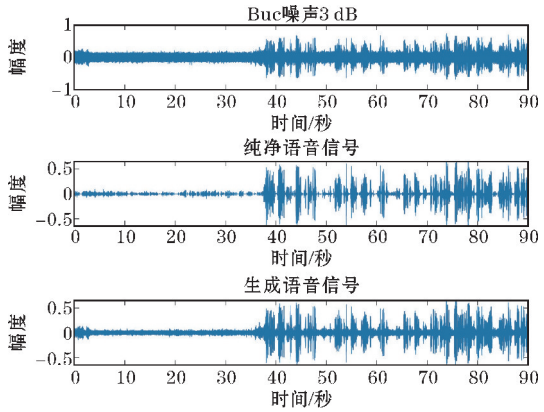


(a) 噪声时域波形图

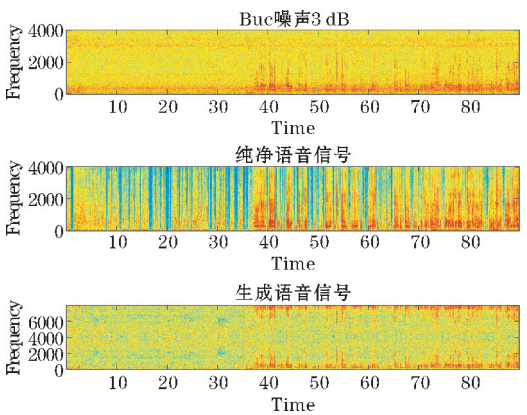


(b) 频谱图

图4 3 dB Babble 噪声波形图



(a) 噪声时域波形图



(b) 频谱图

图5 3 dB Buc 噪声波形图

图4、5展示了模型对3dB Babble 和3dB Buc 的含噪语音增强后的结果,可看出本文模型在去除噪声的同时,又保留了语音的可懂度,相比其他语音增强模型,本模型在对抗迭代阶段着重考虑了语音频率方向的上下文信息,由于语音能量大部分集中于低频段,导致其他模型在增强语音后,部分频段的信息会丢失,而在模型训练过程中加入频谱重构损失则会减少信息丢失,这也验证了频谱重构学习的重要性。

对添加 3 dB、6 dB、9 dB 随机噪声后的语音信号进行评估,并取各信噪比指标的平均得分。从表 1 可得出,相比 LSTM、SEGAN 等语音增强的基准模型,利用模型在训练阶段使用过的数据集评价时,即 seen,模型 Mask-LAGAN 在 PESQ 得分上分别提高了0.393和0.48,在 STOI 得分上分别提高了1.247和0.546,在 CSIG 得分上分别提高了0.3315和0.3941。利用模型在训练阶段未使用过的数据集评价时,即 unseen,模型 Mask-LAGAN 在 PESQ 得分上分别提高了0.254和0.437,在 STOI 得分上分别提高了0.1319和0.0682,在 CSIG 得分上分别提高了0.2313和0.3941。

表 1　基准模型和 Mask-LAGAN 的评价指标对比

评价指标		LSTM	SEGAN	注意力机制 BLSTM	注意力机制 SEGAN	Mask-LAGAN
PESQ	seen	2.647	2.560	1.877	2.910	3.040
	unseen	2.633	2.450	1.807	2.797	2.887
STOI	seen	0.8153	0.8881	0.8030	0.9241	0.9427
	unseen	0.8102	0.8739	0.7740	0.9203	0.9421
CSIG	seen	3.0773	3.0147	2.4288	3.3138	3.4088
	unseen	3.0654	2.9204	2.3660	3.2180	3.2967

整体来说,LSTM 和 SEGAN 均从评价指标优化的方向出发来进行模型训练,其更注重于整体的损失函数优化,并未聚焦于局部低频段信息,故其 CSIG 指数,即语音可懂度指标分数略低于其余模型。引入注意力机制后,模型可捕捉到时序信息中的局部信息,也可对全局机型建模,但其对语音上下文信息的针对减弱,导致 PESQ 等语音质量指标下降。本文提出的模型虽引入注意力机制,但同时将频谱重构损失加入了整体模型的对抗损失中,使在关联整体网络的情况下对低频段进行着重考虑,故在 3 种语音评价指标上均有提升。

实验表明,本文提出的 Mask-LAGAN 模型各评价指标均远高于基准模型,在环境噪声较大时的声学建模能力更优,并且在低信噪比、失真干扰较大的情况下仍然有较高的鲁棒性。

4　结束语

本文提出一种在低信噪比下仍有优良声学建模能力的语音增强模型 Mask-LAGAN,采用基于注意力机制的 BLSTM 网络作为生成器,并引入语音掩码学习,该方法能够有效解决模型不稳定和泛化能力弱的问题。该模型能够在生成语音时自适应地选择重要的上下文信息、学习语音信号的特征表示,并且针对基准模型消耗计算资源较大的情况,BLSTM 结构具有并行化的潜力,可以加速模型的训练和推理,特别是在失真、干扰影响较大的情况下,仍然能够高质量地增强语音信号。此外,在生成器输出前加入语音掩码学习,通过与判决器进行对抗训练,更准确地恢复语音信号强度。最后,通过 PESQ、STOI、CSIG 等语音质量评价指标对提出的 Mask-LAGAN 模型和基准模型进行对比评估,验证了模型在低信噪比情况下仍具有较高的语音增强能力。

参考文献:

[1] Pascual Santiago, Antonio Bonafonte, Joan Serrà. SEGAN: Speech Enhancement Generative Adversarial Network[C]. Interspeech,2017.

[2] Bahdanau Dzmitry, Jan Chorowski, Dmitriy Serdyuk, et al. End-to-end attention-based large vocabulary speech recognition[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),2015:4945-4949.

[3] Watanabe, Shinji, Takaaki Hori, et al. Hybrid CTC/Attention Architecture for End-to-End Speech Recognition[J]. IEEE Journal of Selected Topics in Signal Processing,2017:1240-1253.

[4] Wang Kai, He Bengbeng, Zhu Weiping. TSTNN: Two-Stage Transformer Based Neural Network for Speech Enhancement in the Time Domain[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021: 7098-7102.

[5] Qiuqiang Kong, Xu Yong, Wang Wenwu, et al. Sound Event Detection of Weakly Labelled Data With CNN-Transformer and Automatic Threshold Optimization[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019(28): 2450-2460.

[6] Cao Ru, Sherif Abdulatif, Bin Yang. CMGAN: Conformer-based Metric GAN for Speech Enhancement [J]. ArXiv,2022.

[7] 张恩琪,顾广华,赵晨,等. 生成对抗网络 GAN 的研究进展[J]. 计算机应用研究,2021,38(4): 968-974.

[8] 庞源焜,张宇山. 句子级状态下 LSTM 对谣言鉴别的研究[J]. 计算机应用研究,2022,39(4): 1064-1070.

[9] Zhang Shiqing, Zhao Xiaoming, Tian Qingxi. Spontaneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM[J]. IEEE Transactions on Affective Computing,2022(13):680-688.

[10] 刘继明,孙成,袁野. 基于训练模型改进的语音

问句信息抽取方法[J]. 科学技术与工程, 2021,21(18):7635-7641.

[11] Sultana Sadia M, Zafar Iqbal, Mohammad Reza Selim, et al. Bangla Speech Emotion Recognition and Cross-Lingual Study Using Deep CNN and BLSTM Networks[J]. IEEE Access, 2022(10): 564-578.

[12] Shim, Kyuhong, Jungwook Choi, et al. Understanding the Role of Self Attention for Efficient Speech Recognition[C]. International Conference on Learning Representations, 2022.

[13] Su BoHao, ChiChun Lee. Unsupervised Cross-Corpus Speech Emotion Recognition Using a Multi-Source Cycle-GAN[C]. IEEE Transactions on Affective Computing, 2022.

[14] Wu Bowen, Liu Chaoran, Carlos Toshinori Ishi, et al. Modeling the Conditional Distribution of Co-Speech Upper Body Gesture Jointly Using Conditional-GAN and Unrolled-GAN[C]. Electronics, 2021.

[15] Ju Lin, Niu Sufeng, Adriaan J, et al. Improved Speech Enhancement Using a Time-Domain GAN with Mask Learning[C]. Interspeech, 2020.

[16] 杨海涛, 王华朋, 楚宪腾, 等. 基于卷积循环神经网络的语音逻辑攻击检测[J]. 科学技术与工程, 2022, 22(18): 7937-7944.

[17] Su Jiaqi, Jin Zeyu, Adam Finkelstein. HiFi-GAN-2: Studio-Quality Speech Enhancement via Generative Adversarial Networks Conditioned on Acoustic Features[C]. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2021: 166-170.

[18] Beck, Gustavo Teodoro Döhler, Ulme Wennberg, et al. Wavebender GAN: An Architecture for Phonetically Meaningful Speech Manipulation[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: 6187-6191.

[19] Kim Minsu, Joanna Hong, Yong Man Ro. Lip to Speech Synthesis with Visual Context Attentional GAN[C]. Neural Information Processing Systems, 2022.

[20] Rix A W, Beerends J. G, Hollier M P, et al. Perceptual evaluation ofspeech quality (PESQ)—A new method for speech quality assessmentof telephone networks and codecs[C]. Proc of the 26th IEEE IntConf on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE, 2001: 749-752.

GAN Speech Enhancement Algorithm based on Attention Mechanism and Mask Learning

LI Tongyan, PEI Haoyan, PEI Yan, CHEN Xu, WANG Tao

(College of Communication Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: Speech enhancement is one of the important components in automatic speech recognition (ASR). In recent years, the modeling capability of generative adversarial networks (GANs) and their variants in speech enhancement has been gradually improved. However, they still suffer from weak generalization ability and inability to adapt to low signal-to-noise ratio environments. In response to this issue, a GAN-based speech enhancement model called Mask-LAGAN, which combines attention-based bidirectional LSTM (BLSTM) and mask learning, is proposed. The framework innovatively designs the speech enhancement mechanism by using BLSTM and attention layers as the generator of the GAN, and introduces mask learning for spectrum reconstruction. The enhanced signal is obtained by overlaying the filtered signal with the original signal, followed by input to the discriminator. The two networks engage in a mutual adversarial training to achieve the goal of speech enhancement. The TIMIT dataset is utilized for comparative evaluation under different signal-to-noise ratio conditions, using speech evaluation metrics such as PESQ, STOI, and CSIG. Experimental results demonstrate that the proposed model achieves an average improvement of 11.8% in speech enhancement compared to models like SeGAN.

Keywords: speech enhancement; generative adversarial networks; attention-BLSTM; mask reconstruction; low signal-to-noise ratio