计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(7):179−185 [doi: 10.15888/j.cnki.csa.008587]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

多阶段生成器与时频鉴别器的 GAN 语音增强算法①

陈　宇[1], 尹文兵[2], 高　戈[2], 王　霄[2], 曾　邦[2], 陈　怡[3]

[1](公安部第一研究所, 北京 100048)
[2](武汉大学 国家多媒体软件工程技术研究中心, 武汉 430072)
[3](华中师范大学 计算机学院, 武汉 430077)
通信作者: 陈　宇, E-mail: beijingchy@sina.com

摘　要: 传统生成对抗网络的语音增强算法 (SEGAN) 将时域语音波形作为映射目标, 在低信噪比条件下, 语音时域波形会淹没在噪声中, 导致 SEGAN 的增强性能会急剧下降, 语音失真现象较为严重. 针对该问题, 提出了一种多阶段的时频域生成对抗网络的语音增强算法 (multi-stage-time-frequency SEGAN, MS-TFSEGAN). MS-TFSEGAN 采用了多阶段生成器与时频域双鉴别器的模型结构, 不断对映射结果进行完善, 同时捕获时域与频域信息. 另外, 为了进一步提升模型对频域细节信息的学习能力, MS-TFSEGAN 在生成器损失函数中引入了频域 L1 损失. 实验证明, 在低信噪比条件下, MS-TFSEGAN 的语音质量和可懂度与 SEGAN 相比分别提升了约 13.32% 和 8.97%, 作为语音识别前端时在 CER 上实现了 7.3% 的相对提升.

关键词: 语音增强; 生成对抗网络; 低信噪比; 语音质量; 语音可懂度; 语音识别; 多阶段模型; 深度学习

引用格式: 陈宇,尹文兵,高戈,王霄,曾邦,陈怡.多阶段生成器与时频鉴别器的 GAN 语音增强算法.计算机系统应用,2022,31(7):179–185. http://www.c-s-a.org.cn/1003-3254/8587.html

GAN Speech Enhancement Algorithm with Multi-stage Generator and Time-frequency Discriminator

CHEN Yu[1], YIN Wen-Bing[2], GAO Ge[2], WANG Xiao[2], ZENG Bang[2], CHEN Yi[3]

[1](Frist Research Institute of the Ministry of Public Security of PRC, Beijing 100048, China)
[2](National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan 430072, China)
[3](School of Computer Science, Central China Normal University, Wuhan 430077, China)

Abstract: The traditional speech enhancement generative adversarial network (SEGAN) takes the waveform of time-domain speech as the mapping target. When it comes to a low signal-to-noise ratio, the waveform of time-domain speech is drowned in the noise, resulting in a dramatic degradation of the enhancement performance of SEGAN and more serious speech distortion. In response, a multi-stage-time-frequency SEGAN (MS-TFSEGAN) is proposed for speech enhancement. MS-TFSEGAN employs multi-stage generators with dual time-frequency discriminators to continuously refine the mapping results. It captures both time- and frequency-domain information at the same time. In addition, for the further enhancement of learning ability in the frequency domain, MS-TFSEGAN introduces L1 loss in the generator loss function. Experimental results show that the speech quality and intelligibility of MS-TFSEGAN are improved by about 13.32% and 8.97%, respectively, compared with SEGAN under low SNR. A relative improvement of 7.3% in CER is achieved when MS-TFSEGAN is used as the front-end of speech recognition.

Key words: speech enhancement; generative adversarial network; low signal-to-noise ratio; speech quality; speech intelligibility; speech recognition; multi-stage model; deep learning

① 收稿时间: 2021-10-14; 修改时间: 2021-11-12; 采用时间: 2021-11-30; csa 在线出版时间: 2022-05-31

1 引言

在现实环境中录制的语音信号常常会受到背景噪声的干扰. 语音增强可以通过去除带噪语音中的噪声获得干净的语音信号, 从而提高语音质量与可懂度. 它在听觉辅助设备、语音通信、语音识别前端等应用中发挥着重要作用.

传统语音增强算法以谱减法[1]、二值掩码[2]、维纳滤波法[3] 以及最小均方差法[4] 为主. 这些方法假设语音信号是平稳的, 只能在高信噪比环境下对稳态的加性噪声发挥作用. 然而现实中大部分带噪语音的信噪比较低, 并且带有混响与非平稳的噪声, 传统语音增强方法无法处理该类低信噪比的带噪语音.

过去几年里, 深度神经网络 (deep neural networks, DNNs) 逐渐应用到语音增强任务中来[5]. 凭借对复杂映射的强大建模能力, DNNs 可以从数据中学习到语音或噪声的深层特征, 例如使用卷积神经网络 (convolutional neural network, CNN)[6] 或循环神经网络 (recurrent neural network, RNN)[7] 学习带噪语音到干净语音的频谱映射过程, 从而达到去噪的目的. 为了引入语音的上下文信息, 长短时记忆网络 (long short-term memory, LSTM)[8] 也被应用到语音增强中. 这些方法通常使用短时傅里叶变换 (short-time Fourier transform, STFT) 所得的频域幅度谱作为网络映射目标, 同时使用带噪相位进行语音重构, 这会导致幅度谱与相位谱不匹配的情况出现. 为了解决该问题, 时域语音增强方法[9−11] 逐渐受到人们重视. 该类方法通过直接增强语音的时域波形, 避免了逆短时傅里叶变换 (ISTFT) 过程, 使其性能不依赖相位估计的准确性[12].

生成对抗网络 (generative adversarial network, GAN)[13] 通过学习底层数据分布来生成类似于真实数据的样本. GAN 作为最先进的深度生成模型被迅速应用到语音相关的任务中来, 如语音转换[14]、语音合成[15] 等. Pascual 等人提出了一种基于生成对抗网络的语音增强方法 (speech enhancement generative adversarial network, SEGAN)[16], 该方法使用生成器将带噪语音的时域波形直接映射生成干净语音波形, 保留了大量原始语音的底层信息. 同时使用鉴别器区分干净语音信号与增强语音信号, 将鉴别结果反馈给生成器, 指导生成器学习类似于真实干净语音的信号分布. 尽管已有实验证明 GAN 在语音增强任务上的应用是成功的, 但增强语音失真与缺乏对各种语音特征的考虑[12] 等问题依然存在. 在图像处理任务中, 许多人通过修改损失函数[17] 或改进生成器和鉴别器结构[18,19], 以改善 GAN 的效果. 但在语音增强任务中该问题还未得到广泛研究, SEGAN 仍存在语音失真与低信噪比条件下表现不佳的问题.

为了解决该问题, Phan 等人[20] 提出了 SEGAN 的改进算法, 即基于迭代生成对抗网络的语音增强算法 (ISEGAN) 和基于深度生成对抗网络的语音增强算法 (DSEGAN), 通过对增强语音进行多重映射, 达到进一步细化语音和噪声差别的目的. 同时, 文献 [21] 提出了一种基于时频域生成对抗网络的语音增强算法 (time-frequency domain SEGAN, TFSEGAN), 该方法采用了时频双鉴别器的模型结构和时频域 L1 损失函数, 提升了低信噪比下 SEGAN 增强语音的语音质量和语音可懂度, 但该方法仍存在增强语音失真的现象.

受文献 [20,21] 启发, 本文提出一种新的生成对抗网络语音增强框架, 该框架包含多个生成器与多个鉴别器. 其中串联的多生成器可以对语音信号进行多阶段映射, 通过不断对增强语音进行优化, 取得更优的生成结果. 并联的多鉴别器分别将增强语音的时域特征与频域特征作为输入, 指导生成器间接地学习语音频域特征分布. 相应的本文使用时频域联合损失函数, 通过在生成器的损失函数中引入频域损失, 提高生成器对频域细节信息的捕获能力. 实验结果表明, 本文所提出的方法在 PESQ[22] 与 STOI[23] 评价指标上比 SEGAN 基线表现更优, 并且在低信噪比条件下相较 DSEGAN[20] 与 TFSEGAN[21] 方法具有更好的去噪效果与更少的语音失真.

2 基于生成对抗网络的语音增强算法

SEGAN 与 CGAN[24] 的结构类似, 由一个生成器 G 与一个鉴别器 D 组成. 在训练过程中, 生成器将随机采样的带噪语音 $\tilde{x}$ 作为输入, 通过映射的方式产生相应的增强语音信号 $\hat{x}$. 干净语音 x 与增强语音 $\hat{x}$ 分别与带噪语言 $\tilde{x}$ 成对送入鉴别器. 鉴别器接收到输入信号, 需要鉴别该信号是否为真实的干净语音, 并将结果反馈给生成器. 为了生成一个真实的样本, 生成器被训练来欺骗鉴别器, 而鉴别器被训练来区分真实样本 x 和 $\hat{x}$. SEGAN 通过最小最大化博弈的对抗性训练来模拟真实数据的复杂分布, 以此促进模型学习语音的样本分布信息, 结构如图 1 所示, 其中⊕为特征拼接 (contact) 操作.

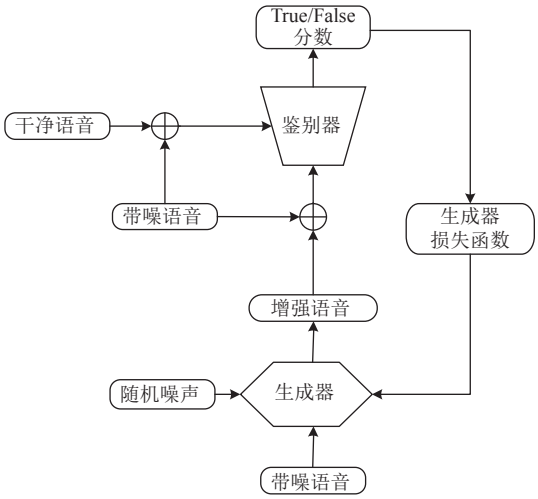

图 1 SEGAN 模型框架

SEGAN 生成器是一个编码器-解码器结构, 负责实现语音增强的功能, 其结构如图 2 所示. 编码器的输入为带噪语音原始波形, 经过多层卷积与 PReLU 压缩得到高维的中间向量c. 该中间向量c与随机噪声 z 一起送往解码器. 解码器结构与编码器网络结构镜像对称, 通过反卷积与 PReLU 激活函数将编码器结果还原为语音时域特征, 即增强后的语音波形$\hat{x}$. 由于使用的语音时域波形特征是直接对语音进行分帧、采样得到的, 没有其他特征提取操作, 模型的输出无需进行语音波形重构等操作来得到增强语音, 避免了传统谱映射方法中估计幅度谱与带噪相位谱不匹配的问题.

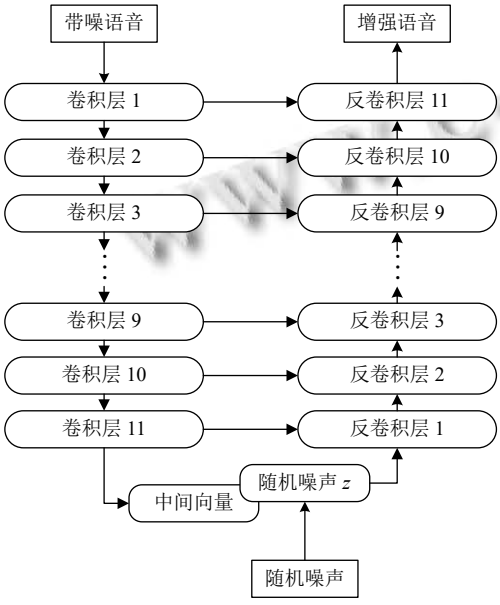

图 2 SEGAN 的生成器

SEGAN 鉴别器的主要作用是监督生成器进行训练, 只工作于模型的训练阶段, 并不参与模型测试等阶段. 鉴别器由 12 个 1 维卷积层与 1 个全连接层组成, 其结构如图 3 所示. 鉴别器对输入信号进行分类, 判断其真假, 并将判别打分反馈给生成器. 在训练过程中, 鉴别器学会将这对信号$(x,\tilde{x})$分类为真, $(x,\tilde{x})$分类为假, 而生成器试图欺骗鉴别器, 使鉴别器将$(x,\tilde{x})$分类为真.

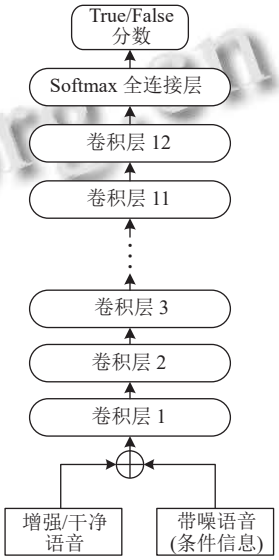

图 3 SEGAN 的鉴别器

受最小二乘法 GAN[25] 的启发, SEGAN 采用最小二乘法损失来提高鉴别器的稳定性. 鉴别器 D 和生成器 G 的最小二乘目标函数被明确写为:

$$J(D)=\frac{1}{2}E_{x,\tilde{x}\sim p_{data}(x,\tilde{x})}\left[(D(x,\tilde{x})-1)^2\right]+\frac{1}{2}E_{z\sim p_z(z),\tilde{x}\sim p_{data}(\tilde{x})}\left[D(G(z,\tilde{x}),\tilde{x})^2\right] \quad (1)$$

$$J(G)=\frac{1}{2}E_{x,\tilde{x}\sim p_{data}(x,\tilde{x})}\left[(D(G(z,\tilde{x}),\tilde{x})-1)^2\right]+\lambda\|G(z,\tilde{x})-x\|_1 \quad (2)$$

其中, D 表示鉴别器, G 表示生成器, x 表示干净语音, $\tilde{x}$表示带噪语音, z 表示随机噪声, $\hat{x}\equiv G(z,\tilde{x})$为生成器输出的增强语音.

在式 (2) 中, 干净语音x和增强语音$\hat{x}$时域幅值之间的 L1 距离被计算在内, 以鼓励生成器 G 生成更精细和真实的结果[16]. L1-norm 项的影响由超参数λ调节, λ在文献 [16] 中被设定为 100.

3 MS-TFSEGAN

为了提升 SEGAN 在低信噪比条件下的增强性能,

同时减少增强语音失真对语音识别后端的影响, 本文提出了一种多阶段生成器与时频鉴别器的生成对抗网络 (MS-TFSEGAN), 其结构如图 4 所示.

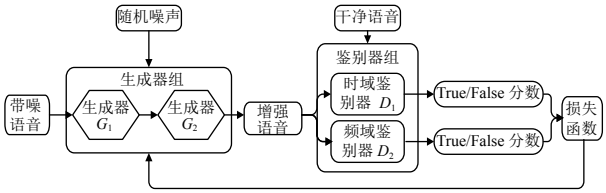

图 4 MS-TFSEGAN 模型框架

3.1 多阶段生成器

文献 [26] 在 GAN 的基础上使用了额外的生成器组成串联生成器的结构, 取得了更好的图像重构性能. 以此为鉴, Phan 等人在文献 [20] 中探索了 SEGAN 多重映射的方法, 证实了该方法在语音增强中同样有效. 本文采用文献 [20] 中的 DSEGAN 的生成器框架, 采用两个生成器进行串联, 实现多阶段的增强, 取代 SEGAN 中单生成器单阶段增强的方法, 其结构如图 5.

生成器 G_1 与 G_2 与 SEGAN 的生成器结构相同, 但参数相互独立. G_1 的输入为带噪语音信号$\tilde{x}$与随机噪声z_1, 生成增强语音信号$\hat{x}_1$. 生成器 G_2 接收到前一级 G_1 的输出$\hat{x}_1$以及随机噪声z_2, 再次进行增强映射, 输出一个更好的增强信号$\hat{x}_2$, 即:

$$\hat{x}_1 = G_1(z_1, \tilde{x}) \tag{3}$$

$$\hat{x}_2 = G_2(z_2, \hat{x}_1) \tag{4}$$

生成器输出$(\hat{x}_1, \tilde{x})$, $(\hat{x}_2, \tilde{x})$都会送往鉴别器, 鉴别器需要将这两个数据对都判断为假, 将干净语音与带噪语音组成的数据对$(x, \tilde{x})$判断为真.

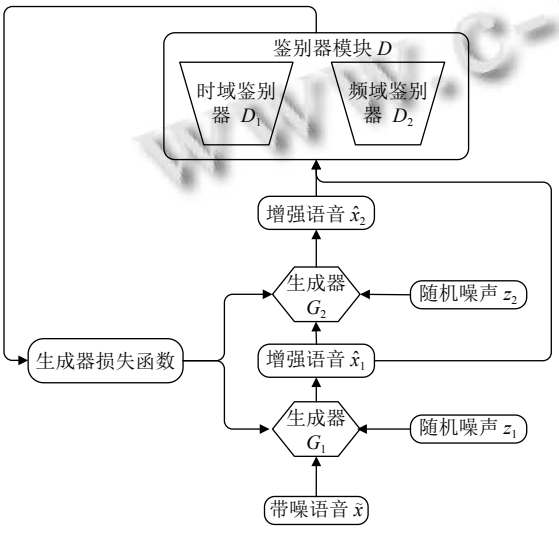

图 5 MS-TFSEGAN 的生成器

3.2 时频鉴别器

传统 SEGAN 的生成器与鉴别器输入均为语音时域特征, 完全忽略了语音信号在频域上的特征分布情况. 而低信噪比条件下, 语音信号会淹没在噪声中, 使带噪语音的时域分布信息难以捕获. 鉴于此, 本文采用双路鉴别器结构, 使生成器能够同时学习语音的时域与频域中的特征分布, 如图 6 所示, 其中鉴别器 D_1 为时域鉴别器, 鉴别器 D_2 为频域鉴别器.

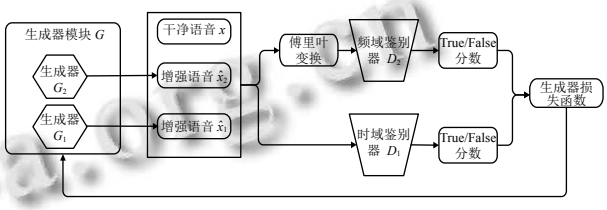

图 6 MS-TFSEGAN 的鉴别器框架

鉴别器 D_1 与 D_2 与 SEGAN 的鉴别器结构相同, 如图 3 所示, D_1 输入为时域波形特征, D_2 输入为语音信号经过 FFT 变换之后的频域特征. 鉴别器 D_1 和 D_2 都是由 12 层卷积层和 1 层 Softmax 全连接层组成, 并在 LeakyReLU 激活前使用虚拟批归一化 (virtual batch-norm)[27], $\alpha = 0.3$. 在训练时, 两个鉴别器都将接收到生成器 G_1 与 G_2 输出的增强语音, 并对两路增强语音进行鉴别打分, 打分结果通过损失函数分别送回生成器 G_1 与 G_2.

3.3 损失函数

MS-TFSEGAN 在训练时, 鉴别器 D_1 与 D_2 将生成器 G_1、G_2 输出的数据对$(\hat{x}_1, \tilde{x})$, $(\hat{x}_2, \tilde{x})$都鉴别为假, 只有$(x, \tilde{x})$鉴别为真. MS-TFSEGAN 时域鉴别器 D_1 的损失函数与 SEGAN 鉴别器的损失函数相似, 但要对生成器 G_1 和 G_2 的输出都进行鉴别打分, 其损失函数如式 (5) 所示:

$$\begin{aligned} J(D_1) = &\frac{1}{2}E_{x,\tilde{x}\sim p_{data}(x,\tilde{x})}[(D_1(x,\tilde{x})-1)^2] \\ &+\frac{1}{4}\sum_{n=1}^{2}E_{z_n\sim p_z(z),\tilde{x}\sim p_{data}(\tilde{x})}[D_1(G_n(z_n,\hat{x}_{n-1}),\tilde{x})^2] \end{aligned} \tag{5}$$

其中, D_1 表示时域鉴别器, G 表示生成器, x表示干净语音, $\tilde{x}$表示带噪语音, z表示随机噪声, $\hat{x}_n \equiv G_n(z_n, \tilde{x})$为第 n个生成器输出的增强语音, 当n为 0 时, $\hat{x}_0 \equiv \tilde{x}$.

MS-TFSEGAN 频域鉴别器 D_2 的输入为傅里叶频谱特征. L1 损失可以直接最小化估计频谱和干净频谱之间的距离[16], 可以带来比 L2 损失更好的语音增强性

能[28], 所以 MS-TFSEGAN 的损失函数采用 L1 范式计算频域损失, 其计算方式如式 (6) 所示:

$$J(D_2)=\frac{1}{2}E_{x,\tilde{x}\sim p_{data}(x,\tilde{x})}[(D_2(FFT[x],FFT[\tilde{x}])-1)^2]+\frac{1}{4}\sum_{n=1}^{2}E_{z_n\sim p_z(z),\tilde{x}\sim p_{data}(\tilde{x})}[D_2(FFT[G_n(z_n,\hat{x}_{n-1})],FFT[\tilde{x}])^2] \tag{6}$$

$$J(G)=\frac{1}{4}\sum_{n=1}^{2}E_{z_n\sim p_z(z),\tilde{x}\sim p_{data}(\tilde{x})}(D_1(G_n(z_n,\hat{x}_{n-1}),\tilde{x})-1)^2+\frac{1}{4}\sum_{n=1}^{2}E_{z_n\sim p_z(z),\tilde{x}\sim p_{data}(\tilde{x})}(D_2(FFT[G_n(z_n,\hat{x}_{n-1})],FFT[\tilde{x}])-1)^2+\sum_{n=1}^{2}\lambda_n\|G_n(z_n,\hat{x}_{n-1})-x\|_1+\sum_{n=1}^{2}\mu_n\|FFT[G_n(z_n,\hat{x}_{n-1})]-FFT[x]\|_1 \tag{7}$$

为了在多个阶段规范损失大小, 式 (7) 中的权值 λ_n和μ_n的值大小分别被设置为$\frac{100}{2^{N-n}}$和$\frac{1}{2^{N-n}}$[20]. 通过上述设置, 一个生成器的增强后输出的 L1 损失项是其前一个生成器的两倍.

4 实验

4.1 实验设置

本文的实验环境为 64 位操作系统 Ubuntu 16.04, 主要使用的开源工具为 SOX、TensorFlow 和 Kaldi. 本文使用的数据集为开源中文语音数据集 Aishell-1, 噪声数据集则为 MUSAN[29].

训练集设置: 增强模块网络的带噪语音训练集由包含 340 个说话人、共 150 小时的 Aishell-1 干净中文语音数据集和噪声数据集 MUSAN 仿真而成. 通过 SOX 工具给 Aishell-1 数据集加上了−15 dB、−10 dB、−5 dB、0 dB、5 dB 和 10 dB 这 6 组不同信噪比的随机种类噪声, 可以得到不同信噪比的带噪语音训练数据集.

测试集设置: 本次实验的测试集包括 6 个子集, 6 个子集中的带噪语音的信噪比分别为−15 dB、−10 dB、−5 dB、0 dB、5 dB 和 10 dB. 其中, 每个子集包含 1 000 条带噪语音, 子集之间除了信噪比不同, 其他设置均相同. 与训练集相同, 测试集的语音是使用 SOX 工具给 Aishell-1 测试集添加 MUSAN 噪声集. 实验结果将由 PESQ、STOI 和语音识别 CER 这 3 种参数进行评估.

在本实验中, MS-TFSEGAN 生成器与鉴别器 (包括时域鉴别器和频域鉴别器) 的网络参数均与 SEGAN 的网络参数一致. 语音信号采样率为 16 kHz, 帧长为 1 s, 帧移为 500 ms, FFT 采样点数为 16 384. 另外, 模型训练的 batchsize 设为 100, 初始化学习率为 0.000 1, 优化方式采用 RMSProp 优化器.

其中, D_2 表示频域鉴别器, $FFT[\cdot]$指快速傅里叶变换操作.

另外, 文献 [21] 表明, 频域鉴别器虽然能监督生成器捕获部分频域特征, 但该部分特征较为模糊与粗糙, 其细腻程度严重不足并缺乏现实意义. 为了解决该问题, 本文采用文献 [21] 中的时频联合损失函数, 向生成器损失函数中引入频域 L1 损失项. 故整个生成器的损失函数如式 (7) 所示:

4.2 实验结果与分析

本次实验的基线系统为 SEGAN, 选取 DSEGAN 和 TFSEGAN 作为额外的基线系统, 其性能仅作为参考. 实验的结果将从增强后语音的质量、可懂度以及对后端语音识别系统的影响 3 个方面进行分析. 其中, PESQ 作为增强语音质量的评估标准; STOI 作为增强语音可懂度的评估标准; 以 Kaldi 的 Chain model 作为后端语音识别系统, 对增强语音进行识别得到的 CER 作为对后端语音识别系统的影响的评估标准. 本次实验的结果以及分析如下所示.

(1) 语音质量比较

采用 PESQ 作为语音作为增强后语音的评估标准, 对 MS-TFSEGAN 在不同信噪比条件下的增强语音的质量进行评估, 评估结果如表 1 所示.

表 1 4 种模型的 PESQ 对比

信噪比 (dB)	−15	−10	−5	0	5	10
Noisy	1.366	1.628	1.885	2.202	2.513	2.899
SEGAN	1.564	1.901	2.260	2.594	2.862	3.112
DSEGAN	1.649	1.961	2.322	2.627	2.896	3.134
TFSEGAN	1.665	1.991	2.337	2.643	2.917	3.191
MS-TFSEGAN	**1.801**	**2.171**	**2.589**	**2.931**	**3.207**	**3.455**

由表 1 所示, 在提升语音质量方面, 各信噪比下 MS-TFSEGAN 增强效果均比 DSEGAN 与 TFSEGAN 优秀. 与 SEGAN 相比, MS-TFSEGAN 的增强性能平均提升了约 13%, 尤其在−15 dB 条件下, 其性能提升约 15.15%. 这说明 MS-TFSEGAN 很大程度上解决了 SEGAN 在低信噪比条件下增强语音质量较差的问题.

同时如表 1 所示, MS-TFSEGAN 在 DSEGAN 的基础上引入了时频鉴别器, 取得了约 10% 的性能提升.

与 TF-SEGAN 相比, 使用了多阶段的生成映射, 取得了约 9% 的性能提升, 尤其在高信噪比条件下同样取得了约 8.27% 的提升, 解决了 TFSEGAN 在高信噪比下提升效果不明显的问题[21].

(2) 语音可懂度结果比较

采用 STOI 作为语音可懂度评估标准, 对 MS-TFSEGAN 在不同信噪比条件下的增强语音的可懂度进行评估, 评估结果如表 2 所示.

表 2 4 种模型 STOI 的对比

信噪比 (dB)	−15	−10	−5	0	5	10
Noisy	0.562	0.630	0.703	0.772	0.832	0.888
SEGAN	0.500	0.606	0.717	0.799	0.851	0.889
DSEGAN	0.523	0.626	0.734	0.808	0.858	0.893
TFSEGAN	0.524	0.626	0.729	0.805	0.860	0.900
MS-TFSEGAN	**0.577**	**0.682**	**0.785**	**0.854**	**0.896**	**0.927**

由表 2 可知, 在提升语音可懂度方面, MS-TFSEGAN 的性能明显强于 SEGAN、DSEGAN 和 TFSEGAN. MS-TFSEGAN 的 STOI 相较 SEGAN 提升了约 8.97%, 相较 DSEGAN 与 TFSEGAN 也获得了约 6.67% 的提升. 特别是在−10 dB 与−15 dB 时, 由于 SEGAN 存在语音失真的问题, 导致增强后语音的可懂度比原带噪语音更低. 而 MS-TFSEGAN 在−15 dB 与−10 dB 条件下 STOI 分别取得了 15.4% 与 12.54% 的提升, 明显优于其他方法, 进一步说明 MS-TFSEGAN 能够有效缓解语音失真导致的语音可懂度下降.

(3) 语音识别结果比较

采用 CER 作为标准评估 MS-TFSEGAN 对后端语音识别系统准确率的影响. 6 组不同信噪比数据经过各语音增强模型后, 得到的增强语音的识别结果 CER 的值如表 3 所示.

表 3 4 种模型的 CER 对比

信噪比 (dB)	−15	−10	−5	0	5	10
Noisy	61.57	49.72	36.06	22.26	9.11	2.54
SEGAN	71.48	51.36	36.28	21.74	11.39	4.21
DSEGAN	61.76	48.94	34.00	21.54	11.31	3.81
TFSEGAN	61.60	48.91	33.36	19.96	10.18	3.25
MS-TFSEGAN	**56.54**	**41.94**	**26.35**	**12.57**	**5.26**	**1.74**

表 3 显示, 在−15 dB 条件下, SEGAN、DSEGAN 对识别准确率并无提升, 反而由于增强语音存在失真导致识别错误率上升. 相较之下, 由于引入了频域信息, TFSEGAN 与 MS-TFSEGAN 在低信噪比条件下对语音识别准确率有所提升, 尤其是 MS-TFSEGAN 相较 SEGAN 在准确率上有约 10% 的提升. 同时在高信噪比下, 相较于原语音也有 1% 的识别准确率的提升.

在复杂度方面, 由于 MS-TFSEGAN 需要训练两个生成器与两个鉴别器, 导致其训练速度相比于 SEGAN 较慢. 但在增强阶段, 鉴别器组不参与增强过程, MS-TFSEGAN 的计算速度与 SEGAN 基线模型相当.

综合 3 组实验可以得到, MS-TFSEGAN 由于其多阶段映射与时频鉴别器的结构, 无论在语音增强性能上还是提升语音识别准确率的性能方面, 都要优于 SEGAN 等基线模型. 说明 MS-TFSEGAN 更适合处理低信噪比语音以及作为语音识别的前端模块.

5 总结

为了解决 SEGAN 存在的问题和不足, 本文提出了多生成器与时频鉴别器的生成对抗网络语音增强算法 (MS-TFSEGAN). 在模型结构方面, MS-TFSEGAN 采用了串联的生成器结构, 对带噪语音进行多阶段的增强. 同时使用了并联的时频鉴别器结构, 时域鉴别器输入语音样本的时域特征, 频域鉴别器输入语音样本的频域特征. 在两个鉴别器的作用下, MS-TFSEGAN 的生成器能够同时学习语音样本在时域和频域中的分布规律和信息. 在损失函数方面, MS-TFSEGAN 采用了时频域联合损失函数.

实验证明, 在负信噪比条件下, MS-TFSEGAN 的语音质量和可懂度与 SEGAN 相比分别提升了约 14.63% 和 12.47%. 同时语音识别的实验证明, MS-TFSEGAN 相比于 SEGAN 与其他模型能够更好地捕获语音样本中的频域中的分布信息, 在语音识别准确率的提升上更为优秀. MS-TFSEGAN 一定程度上解决了 SEGAN 在低信噪比条件下增强性能不佳的问题, 同时减少了增强语音的失真现象, 在作为语音识别前端模块具有明显优势.

参考文献

1 Boll S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979, 27(2): 113–120. [doi: 10.1109/TASSP.1979.1163209]

2 Li N, Loizou PC. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. The Journal of the Acoustical Society of America, 2008, 123(3): 1673–1682. [doi: 10.1121/1.2832617]

3 Lim J, Oppenheim A. All-pole modeling of degraded speech. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, 26(3): 197–210. [doi: 10.1109/TASSP.1978.1163086]

4 Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1984, 32(6): 1109–1121. [doi: 10.1109/TASSP.1984.1164453]
5 Xu Y, Du J, Dai LR, *et al*. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(1): 7–19. [doi: 10.1109/TASLP.2014.2364452]
6 Mamun N, Khorram S, Hansen JHL. Convolutional neural network-based speech enhancement for cochlear implant recipients. Proceedings of the INTERSPEECH 2019, 20th Annual Conference of the International Speech Communication Association. Graz: ISCA, 2019. 4265–4269.
7 Zhao H, Zarar S, Tashev I, *et al*. Convolutional-recurrent neural networks for speech enhancement. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 2401–2405.
8 Weninger F, Erdogan H, Watanabe S, *et al*. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. Proceedings of the 12th International Conference on Latent Variable Analysis and Signal Separation. Liberec: Springer, 2015. 91–99.
9 Pandey A, Wang D. A new framework for supervised speech enhancement in the time domain. Proceedings of the INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association. Hyderabad: ISCA, 2018. 1136–1140.
10 Stoller D, Ewert S, Dixon S. Wave-U-Net: A multi-scale neural network for end-to-end audio source separation. Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR). Paris: ISBN, 2018. 334–340.
11 Rethage D, Pons J, Serra X. A wavenet for speech denoising. Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 5069–5073.
12 Kim HY, Yoon JW, Cheon SJ, *et al*. A multi-resolution approach to GAN-based speech enhancement. Applied Sciences, 2021, 11(2): 721. [doi: 10.3390/app11020721]
13 Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680.
14 Hsu CC, Hwang HT, Wu YC, *et al*. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. arXiv: 1704.00849, 2017.
15 Saito Y, Takamichi S, Saruwatari H. Statistical parametric speech synthesis incorporating generative adversarial networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(1): 84–96. [doi: 10.1109/TASLP.2017.2761547]
16 Pascual S, Bonafonte A, Serrà J. SEGAN: Speech enhancement generative adversarial network. Proceedings of the INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association. Stockholm: ISCA, 2017. 3642–3646.
17 Gulrajani I, Ahmed F, Arjovsky M, *et al*. Improved training of Wasserstein GANs. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 5769–5779.
18 Karras T, Aila T, Laine S, *et al*. Progressive growing of GANs for improved quality, stability, and variation. arXiv: 1710.10196, 2018.
19 Karras T, Laine S, Aittala M, *et al*. Analyzing and improving the image quality of styleGAN. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 8107–8116.
20 Phan H, Mcloughlin IV, Pham L, *et al*. Improving GANs for speech enhancement. IEEE Signal Processing Letters, 2020, 27: 1700–1704. [doi: 10.1109/LSP.2020.3025020]
21 尹文兵. 基于生成对抗网络的语音增强技术研究 [博士学位论文]. 武汉: 武汉大学, 2021.
22 ITU-T. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs: ITU-T P. 862. (2001-02-23).
23 Jensen J, Taal CH. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(11): 2009–2022. [doi: 10.1109/TASLP.2016.2585878]
24 Mirza M, Osindero S. Conditional generative adversarial nets. arXiv: 1411.1784, 2014.
25 Mao XD, Li Q, Xie HR, *et al*. Least squares generative adversarial networks. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 2813–2821.
26 Quan TM, Nguyen-Duc T, Jeong WK. Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss. IEEE Transactions on Medical Imaging, 2018, 37(6): 1488–1497. [doi: 10.1109/TMI.2018.2820120]
27 Isola P, Zhu JY, Zhou TH, *et al*. Image-to-image translation with conditional adversarial networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 5967–5976.
28 Pandey A, Wang DL. On adversarial training and loss functions for speech enhancement. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 5414–5418.
29 Snyder D, Chen GG, Povey D. MUSAN: A music, speech, and noise corpus. arXiv: 1510.08484, 2015.

(校对责编: 孙君艳)