计算机科学
COMPUTER SCIENCE

基于时频域生成对抗网络的语音增强算法

尹文兵，高戈，曾邦，王霄，陈怡

引用本文

尹文兵，高戈，曾邦，王霄，陈怡. 基于时频域生成对抗网络的语音增强算法[J]. 计算机科学, 2022, 49(6): 187-192.

YIN Wen-bing, GAO Ge, ZENG Bang, WANG Xiao, CHEN Yi. Speech Enhancement Based on Time-Frequency Domain GAN[J]. Computer Science, 2022, 49(6): 187-192.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210500114

基于时频域生成对抗网络的语音增强算法

尹文兵[1] 高 戈[1] 曾 邦[1] 王 霄[1] 陈 怡[2]
1 武汉大学国家多媒体软件工程技术研究中心 武汉 430072
2 华中师范大学计算机学院 武汉 430077
(912228963@qq.com)

摘 要 传统基于生成对抗网络的语音增强算法(Speech Enhancement Algorithm Based on Generative Adversarial Networks,SEGAN)在时域上对语音进行增强处理,完全忽略了语音样本在频域上的分布情况。在低信噪比条件下,语音信号会淹没在噪声中,带噪语音的时域分布信息很难捕获,因此,SEGAN的增强性能会急剧下降,其增强语音的语音质量和语音可懂度很低。针对该问题,提出了基于时频域生成对抗网络的语音增强算法(Time-Frequency Domain SEGAN,TFSEGAN)。TFSEGAN采用了时频域双判别器的模型结构和时频域L1损失函数,时域判别器的输入为语音样本的时域特征,频域判别器的输入为语音样本的频域特征。在训练过程中,时域判别器将语音样本的时域分布信息作为判别标准,而频域判别器将语音样本的频域分布信息作为判别标准。在两个判别器的作用下,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。实验证明,在低信噪比条件下,与SEGAN相比,TFSEGAN的语音质量与可懂度分别提升了约17.45%和11.75%。

关键词:语音增强;生成对抗网络;时频域;低信噪比;语音质量;语音可懂度
中图法分类号 TN912.35

Speech Enhancement Based on Time-Frequency Domain GAN

YIN Wen-bing[1],GAO Ge[1],ZENG Bang[1],WANG Xiao[1] and CHEN Yi[2]
1 National Engineering Research Center for Multimedia Software,Wuhan University,Wuhan 430072,China
2 School of Computer Science,Central China Normal University,Wuhan 430077,China

Abstract The traditional speech enhancement algorithm based on generative adversarial networks (SEGAN) enhances speech in the time domain,and completely ignores the distribution of speech samples in frequency domain. Under the condition of low signal-to-noise ratio,the speech signal will be submerged in noise,and the time-domain distribution information of noisy speech is difficult to capture. Therefore,the enhancement performance of SEGAN will drop sharply,and the speech quality and speech intelligibility of its enhanced speech are very low. To solve this problem,this paper proposes a speech enhancement algorithm (time-frequency domain SEGAN,TFSEGAN) based on time-frequency domain generation confrontation network. TFSEGAN adopts the model structure of the time-frequency domain dual discriminator,and a time-frequency L1 loss function. The input of time domain discriminator is time domain feature of the speech sample,and the input of frequency domain discriminator is frequency domain feature of the speech sample. In the training process,time-domain discriminator uses the time-domain distribution information of speech sample as the criterion,and frequency-domain discriminator uses the frequency-domain distribution information of the speech sample as the criterion. Under the action of two discriminators,the generator of TFSEGAN could simultaneously learn the distribution rules and information of speech samples in time domain and frequency domain. Experiments prove that,compared with SEGAN,the speech quality and intelligibility of TFSEGAN improve by about 17.45% and 11.75% respectively at low signal-to-noise ratio.

Keywords Speech enhancement,Generative adversarial network,Time-frequency domain,Low signal-to-noise ratio,Speech quality,Speech intelligibility

1 引言

语音增强是通过一定方法来抑制和降低语音中的噪声的技术,其主要目的是提高语音的质量和改善语音的可懂度。语音增强技术常被应用于现代语音通信系统,以提高通信的质量和人的主观听觉舒适度。近年来,语音增强技术发展迅速,先后出现了许多优秀的语音增强算法。

谱减法[1]、维纳滤波算法[2]、基于统计模型的语音增强算法[3]和信号子空间算法[4]是常见的几种经典语音增强算法。传统语音增强算法在线性平稳噪声的环境中表现良好,但是对非平稳噪声的处理能力较弱。Wang[5]提出了基于理想二值掩模的语音增强算法。该算法的原理是对语音信号中的

到稿日期:2021-05-17 返修日期:2021-09-04
通信作者:高戈(gaoge@whu.edu.cn)

噪声信号进行时频掩蔽处理，从而得到较为干净的语音信号。Srinivasan 等[6]提出了基于理想比率掩模的语音增强方法，该算法的基本原理和基于理想二值掩模的语音增强算法的原理基本相同。相比基于理想二值掩模的语音增强算法，基于理想比率掩模的语音增强方法的增强效果更好。广泛地使用频谱或倒频谱作为映射谱，会丢失语音中大量有价值的信息，如相位信息。针对该问题，Oord[7]等将 WaveNet 网络结构用于语音增强。WaveNet 可以在时域中将带噪语音直接映射到干净语音，保留了完整的相位信息，增强效果更佳[8-9]。随着生成对抗网络的提出和发展，Pascual 等[10]提出了一种基于生成对抗网络(Generative Adversarial Network，GAN)的语音增强方法 SEGAN。该方法将原始带噪语音直接映射到干净语音，保留了大量原始语音的底层信息，从而防止语音失真。基于生成对抗网络的语音增强算法大多使用单个生成器来处理带噪语音，对语音进行单步的映射，无法进一步细化语音和噪声的差别。为了解决该问题，Phan 等[11]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和基于深度生成对抗网络的语音增强算法。两个改进算法都是通过增加生成器的数量来对语音进行多重映射，从而达到进一步细化语音和噪声差别的目的。除了上述基于生成对抗网络的语音增强算法外，还有许多其他基于生成对抗网络的语音增强算法被陆续提出[12-17]。

由于噪声具有不稳定性和随机性，在实际应用中，传统的语音增强算法很难取得很好的增强效果。虽然基于深度学习的语音增强算法能克服噪声的不稳定性和随机性等问题，并且其性能优于传统语音增强算法，但这些算法大多需要假设语音样本的数据分布符合某类特殊的数据分布，如高斯分布。而语音样本的真实分布情况往往不符合这种假设，导致基于深度学习的语音增强算法在实际应用中表现不佳。

SEGAN 在时域上对语音进行增强处理，即它是一种基于时域的语音增强算法，其生成器和判别器的输入均为语音的时域特征。因此，SEGAN 的判别器会将语音样本在时域中的分布信息作为唯一的判别标准，完全忽略了语音样本在频域上的分布情况。在低信噪比条件下，语音信号会淹没在噪声中，带噪语音的时域分布信息很难捕获，SEGAN 的生成器无法通过对比带噪语音和干净语音的分布信息来更新参数。在该条件下，SEGAN 的增强性能急剧下降，其增强语音的语音质量和语音可懂度很低。

近两年，许多基于将时频域特征信息融合的语音增强方法被提出，较好地解决了低信噪比下增强性能不足的问题[18-20]。同样，为了解决 SEGAN 存在的问题，本文提出了基于时频域生成对抗网络的语音增强算法 TFSEGAN。TFSEGAN 采用了时频域双判别器的模型结构和时频域 L1 损失函数，时域判别器的输入为语音样本的时域特征，频域判别器的输入为语音样本的频域特征，并且采用时频域 L1 损失函数作为 TFSEGAN 的损失函数。

2 基于生成对抗网络的语音增强算法

SEGAN 的模型框架与条件生成对抗网络类似，如图 1 所示。带噪语音作为额外的辅助信息输入生成器和判别器，以提升网络的性能和复杂样本数据的处理能力。在训练过程中，带噪语音和随机噪声输入生成器，经过生成器作用后输出增强语音；增强语音和干净语音轮流输入判别器，判别器进行判别打分；判别器输出的判别分数将通过损失函数反馈给生成器，以此促进生成器学习语音的样本分布信息。

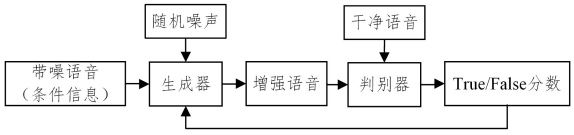


图 1 SEGAN 模型框架

Fig. 1 Framework of SEGAN model

SEGAN 生成器的主要作用是对带噪语音进行增强处理，是整个系统的语音增强模块。SEGAN 判别器的主要作用是监督生成器进行训练，以提高生成器的学习能力和增强效果。判别器只工作于模型的训练阶段，并不参与模型测试等。SEGAN 的生成器采用端到端(End to End)的模型结构，其输入为带噪语音的波形级时域特征(特征点的值为时域波形的幅度值)，输出为增强语音的波形级时域特征。生成器的网络结构与深度自动编码器(Deep Auto Encoder)的网络结构类似，可以分为编码器和解码器两个部分，如图 2 所示。

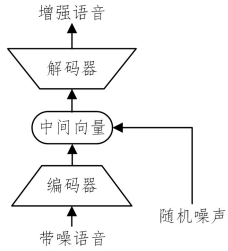


图 2 SEGAN 的生成器

Fig. 2 Generator of SEGAN

图 2 中，编码器部分的输入为带噪语音的时域特征，输出为高维度的中间向量；解码器部分的输入为编码器输出的中间向量和随机噪声，经过解码器的解码后，输出增强语音的时域特征。

SEGAN 的生成器中还使用了跳跃连接将编码器部分与解码器部分连接起来。SEGAN 将生成器编码器部分与解码器部分对应的反卷积层相连，在端到端网络模型中，该连接被称为跳跃连接。与残差连接类似，网络较深时，跳跃连接可以较好地解决模型的梯度消失等问题，并且能加快训练过程。

SEGAN 的损失函数采用了 LSGAN 的损失函数的形式，并且与条件生成对抗网络(CGAN)一样增加了额外的条件信息。SEGAN 的判别器的损失函数如式(1)所示：

$$J(D)=\frac{1}{2}E_{x,x_c\sim p_{\text{data}}(x,x_c)}[(D(x,x_c)-1)^2]+\frac{1}{2}E_{z\sim p_z(z),x_c\sim p_{\text{data}}(x_c)}[D(G(z,x_c),x_c)^2] \tag{1}$$

其中，D 表示判别器，G 表示生成器，x 表示干净语音样本，x^c 表示带噪语音样本，z 表示随机噪声。在训练过程中，带噪语音将作为条件信息与干净语音一起输入判别器，与随机噪声一起输入生成器。

SEGAN 的生成器损失函数如式(2)所示：

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

其中，λ 为控制 L1 损失项的超参数，一般设置为 100。

3 基于时频域生成对抗网络的语音增强算法

为了解决 SEGAN 在低信噪比条件下性能不佳的问题，本节提出基于时频域生成对抗网络的语音增强算法 TFSEGAN。TFSEGAN 采取了时频域双判别器的模型结构，如图 3 所示。其中，第一个判别器被称为时域判别器，其结构与 SEGAN 的判别器相同，输入为干净语音和增强语音的时域特征；另一个判别器被称为频域判别器，其输入为干净语音和增强语音的频域特征(傅里叶幅度谱特征)。

在 TFSEGAN 模型的训练过程中，带噪语音经过分帧、采样后，将作为条件信息与随机噪声一起输入生成器，生成器的输出为增强语音的时域特征。一方面，增强语音的时域特征将直接输入时域判别器；另一方面，对增强语音的时域特征做傅里叶变换(这里采用 FFT)得到其傅里叶幅度频谱，然后将增强语音的幅度谱特征输入频域判别器。两个判别器将分别进行判别打分，打分结果将通过生成器的损失函数反馈给生成器，达到监督和促进生成器训练的目的。

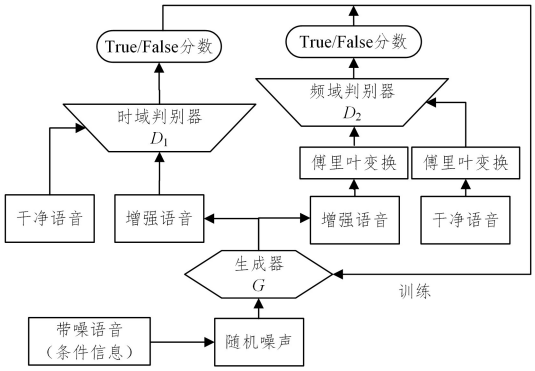


图 3 TFSEGAN 模型框架

Fig. 3 Framework of TFSEGAN model

TFSEGAN 的生成器的网络结构与 SEGAN 的生成器的网络结构相同，可以分为编码器部分和解码器部分。TFSEGAN 生成器的编码器部分由 11 层一维跨步卷积层堆叠构成，其网络结构图如图 4 所示。在编码器中，通过 11 层卷积层，原始经过分帧采样得到的16 384维语音时域特征将被逐层降维，最后降低为 8 维度的中间向量。

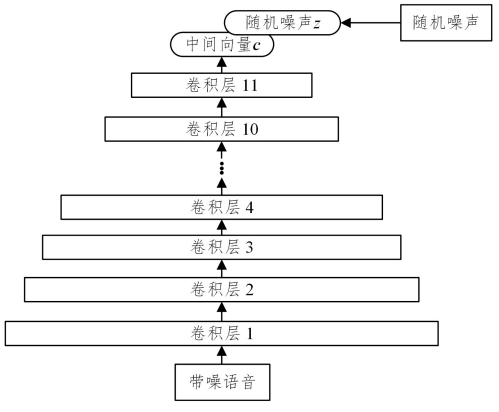


图 4 TFSEGAN 生成器的编码器部分

Fig. 4 Encoder part of TFSEGAN generator

TFSEGAN 生成器的解码器部分由 11 层一维反卷积层组成，其网络结构如图 5 所示。与 SEGAN 的生成器结构一致，为了将编码器部分输出的中间向量还原为 16 384 维的语音时域特征，解码器部分需要与编码器部分对称。通过 11 层反卷积层，8 维的中间向量被逐层还原为 16 384 维的语音时域特征。

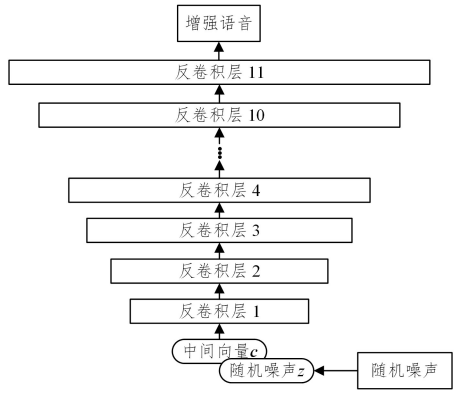


图 5 TFSEGAN 生成器的解码器部分

Fig. 5 Decoder part of TFSEGAN generator

反卷积层，也称转置卷积层，与卷积层相对应。卷积层的作用是将输入的特征转换为更高阶的特征，而反卷积层的作用则是将输入的高阶特征还原为更低阶的特征。实际上，反卷积层的操作与卷积层的操作类似，不同的是反卷积操作需要先对输入特征平面进行补零，然后再进行卷积操作。

TFSEGAN 的判别器的网络结构与 SEGAN 的判别器的网络结构不同，SEGAN 的判别器由 12 层卷积层和 1 层全连接层组成，而 TFSEGAN 的两个判别器都由 7 层卷积层和 1 层全连接层组成。

相比 SEGAN，TFSEGAN 减少了判别器的网络层数，主要原因是，在生成对抗网络的训练过程中，如果判别器的性能太好，即能将绝大部分生成器生成的样本判断为“假”样本，会导致生成器产生梯度消失的问题。TFSEGAN 含有两个判别器，在训练过程中同时与生成器做对抗训练，更容易产生以上问题。通常，网络越深，网络模型的性能越好。因此，为了降低判别器的性能，需要适当减少其网络层数。

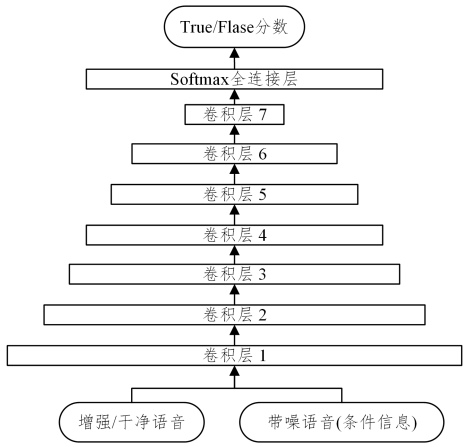


图 6 TFSEGAN 的时域判别器

Fig. 6 Time domain discriminator of TFSEGAN

TFSEGAN 的时频域判别器的网络结构都由 7 层卷积层和 1 层全连接层组成。时域判别器和频域判别器的网络结构

分别如图 6 和图 7 所示。为使生成器能够同时学习语音相关时域和频域的特征信息，且学习过程中没有偏向，频域判别器和时域判别器需要在网络结构上具有一致性。频域判别器与时域判别器唯一的区别是输入的特征不同，时域判别器的输入为语音的时域特征，而频域判别器的输入为语音的频域特征(傅里叶幅度谱特征)。

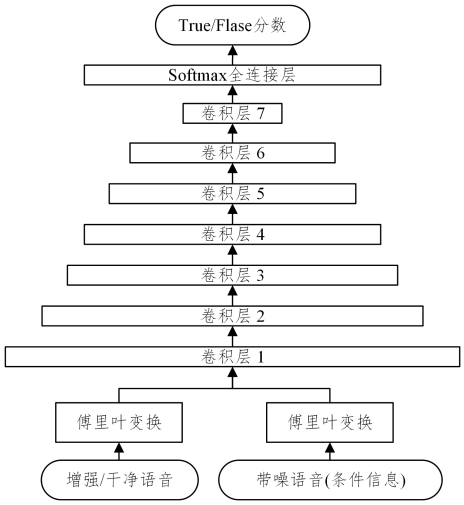


图 7 TFSEGAN 的频域判别器

Fig. 7 Frequency domain discriminator of TFSEGAN

TFSEGAN 在训练时，两个判别器输出的判别分数都将反馈给生成器，以帮助生成器同时捕捉语音样本的时域特征分布信息和频域特征分布信息。因此，TFSEGAN 的生成器的损失函数不仅含有两个判别器的判别分数损失项，还含有时频域 L1 损失项。TFSEGAN 的生成器的损失函数如式(3)所示：

$$J(G)=\frac{1}{2}E_{x,x_c\sim p_{\text{data}}(x,x_c)}[(D_1(x,x_c)-1)^2]+\frac{1}{2}E_{z\sim p_z(z),x_c\sim p_{\text{data}}(x,x_c)}[(D_2(FFT[G(z,x_c)],FFT[x_c])^2]+\lambda\|G(z,x_c)-x\|_1+\mu\|FFT[G(z,x_c)]-FFT[x]\|_1 \tag{3}$$

其中，G 表示生成器；D_1 表示时域判别器；D_2 表示频域判别器；FFT 表示快速傅里叶变换；λ 为控制 L1 损失项的超参数，与 SEGAN 相同，设置为 100。

TFSEGAN 的时域判别器和频域判别器的损失函数如式(4)和式(5)所示：

$$J(D_1)=\frac{1}{2}E_{x,x_c\sim p_{\text{data}}(x,x_c)}[(D_1(x,x_c)-1)^2]+\frac{1}{2}E_{z\sim p_z(z),x_c\sim p_{\text{data}}(x,x_c)}[D_1(G(z,x_c),x_c)^2] \tag{4}$$

$$J(D_2)=\frac{1}{2}E_{x,x_c\sim p_{\text{data}}(x,x_c)}[(D_2(FFT[x],FFT[x_c])-1)^2]+\frac{1}{2}E_{z\sim p_z(z),x_c\sim p_{\text{data}}(x,x_c)(x_c)}[D_2(FFT[G(z,x_c)],FFT[x_c])^2] \tag{5}$$

4 实验与分析

4.1 实验设置

增强模块网络的带噪语音训练集由包含 340 个说话人、共 150 h 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 数据集中加上 −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。增强模块网络的测试集由包含 20 个说话人、共 10 h 的 Aishell-1 干净语音数据集和噪声数据集 MUSAN 仿真而成。通过 SOX 工具给 AIshell-1 数据集中加上 −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测试数据集。实验结果将由 PESQ，STOI 和 SSNR 这 3 种参数进行评估。

在本实验中，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 *batch-size* 设为 50，初始化学习率为 0.002，优化方式采用 RMSProp 优化器。

TFSEGAN 生成器编码器部分的网络参数如表 1 所列，解码器部分的网络参数如表 2 所列，判别器(包括时域判别器和频域判别器)的网络参数如表 3 所列。

表 1 TFSEGAN 生成器编码器的网络参数

Table 1 Network parameters of TFSEGAN generator's encoder

网络层	卷积核大小	步长	输入大小	输出大小	激活函数
卷积层 1	31	2	16 384×1	8 192×16	PReLU
卷积层 2	31	2	8 192×16	4 096×32	PReLU
卷积层 3	31	2	4 096×32	2 048×32	PReLU
卷积层 4	31	2	2 048×32	1 024×64	PReLU
卷积层 5	31	2	1 024×64	512×64	PReLU
卷积层 6	31	2	512×64	256×128	PReLU
卷积层 7	31	2	256×128	128×128	PReLU
卷积层 8	31	2	128×128	64×256	PReLU
卷积层 9	31	2	64×256	32×256	PReLU
卷积层 10	31	2	32×256	16×512	PReLU
卷积层 11	31	2	16×512	8×1 024	PReLU

表 2 TFSEGAN 生成器解码器的网络参数

Table 2 Network parameters of TFSEGAN generator's decoder

网络层	卷积核大小	步长	输入大小	输出大小	激活函数
反卷积层 1	31	2	8×1 024	16×512	PReLU
反卷积层 2	31	2	16×512	32×256	PReLU
反卷积层 3	31	2	32×256	64×256	PReLU
反卷积层 4	31	2	64×256	128×128	PReLU
反卷积层 5	31	2	128×128	256×128	PReLU
反卷积层 6	31	2	256×128	512×64	PReLU
反卷积层 7	31	2	512×64	1 024×64	PReLU
反卷积层 8	31	2	1 024×64	2 048×32	PReLU
反卷积层 9	31	2	2 048×32	4 096×32	PReLU
反卷积层 10	31	2	4 096×32	8 192×16	PReLU
反卷积层 11	31	2	8 192×16	16 384×1	PReLU

表 3 TFSEGAN 判别器的网络参数

Table 3 Network parameters of TFSEGAN discriminator

网络层	卷积核大小	步长	输入大小	输出大小	激活函数
卷积层 1	31	4	16 384×1	4 096×16	LeakyReLU
卷积层 2	31	4	4 096×16	1 024×32	LeakyReLU
卷积层 3	31	4	1 024×32	256×64	LeakyReLU
卷积层 4	31	4	256×64	64×128	LeakyReLU
卷积层 5	31	4	64×128	16×256	LeakyReLU
卷积层 6	31	4	16×256	4×512	LeakyReLU
卷积层 7	1	4	4×512	4×1	LeakyReLU
全连接层	无	无	4×1	1	Softmax

4.2 实验结果与分析

本次实验的基线系统为 SEGAN。ISEGAN(Phan 等于 2020 年提出的基于迭代生成对抗网络的语音增强算法)将作为额外的基线系统，其性能仅供参考。

实验结果将从语音的质量、信噪比和可懂度3个方面进行分析，分别以PESQ，SSNR和STOI作为评估标准。

通过对模型输出的增强语音的质量、信噪比和可懂度进行对比分析来衡量TFSEGAN的性能优劣，并分析其存在的问题和不足。

(1)语音质量结果比较

采用PESQ作为语音质量评估标准，对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4所列。

表4 不同信噪比下TFSEGAN的*PESQ*值

Table 4 *PESQ* values of TFSEGAN with different signal-to-noise ratios

model	10 dB	5 dB	0 dB	−5 dB	−10 dB	−15 dB
Noisy	2.97	2.50	2.18	1.85	1.47	1.23
SEGAN	2.95	2.68	2.43	2.09	1.61	1.12
ISEGAN	3.03	2.74	2.47	2.11	1.64	1.29
TFSEGAN	**3.14**	**2.84**	**2.58**	**2.31**	**1.89**	**1.52**

由表4可知，在提升语音质量方面，与SEGAN和ISEGAN相比，TFSEGAN的性能明显更好。与SEGAN相比，TFSEGAN在各个信噪比条件下的增强性能平均提高了约13.70%；在低信噪比条件下，TFSEGAN的性能平均提升了约17.45%。在信噪比较高的条件下(大于0 dB)，TFSEGAN对语音质量的提升效果与SEGAN相差不大。而在信噪比较低的条件下，TFSEGAN相比SEGAN性能提升十分明显，这说明TFSEGAN在很大程度上解决了SEGAN存在的低信噪比条件下语音质量低下的问题。

当输入语音的信噪比分别为10 dB，5 dB，0 dB，−5 dB，−10 dB和−15 dB时，TFSEAGN的性能相比SEGAN分别提升了约6.44%，5.97%，6.17%，10.53%，17.39%和35.71%。可以看出，输入语音的信噪比越低，TFSEGAN的性能越好。由此说明TFSEGAN更适合处理低信噪比的带噪语音，对低信噪比语音的质量提升效果很好。

(2)语音信噪比结果比较

采用SSNR作为语音信噪比评估标准，对TFSEGAN在不同信噪比条件下的增强语音的信噪比进行评估，评估结果如表5所列。

表5 不同信噪比下TFSEGAN的*SSNR*值

Table 5 *SSNR* values of TFSEGAN with different signal-to-noise ratios

model	10 dB	5 dB	0 dB	−5 dB	−10 dB	−15 dB
SEGAN	10.58	6.20	4.38	2.53	0.90	0.09
ISEGAN	11.08	9.39	**6.70**	4.18	**1.95**	**0.91**
TFSEGAN	**11.97**	**9.54**	6.21	**4.81**	1.56	0.87

由表5可知，在提升语音信噪比方面，TFSEGAN的性能在整体上明显优于SEGAN，平均提高了约189.82%。在低信噪比条件下，TFSEGAN的性能提升了约267.98%。当输入语音的信噪比为10 dB，5 dB，0 dB，−5 dB，−10 dB和−15 dB时，TFSEGAN的性能比SEGAN分别提升了13.14%，53.87%，41.78%，90.12%，73.33%和866.67%。可以看出，随着信噪比的降低，TFSEGAN的性能不断提升，说明TFSEGAN在低信噪比条件下具有很好的增强性能。

(3)语音可懂度结果比较

采用STOI作为语音信噪比评估标准，对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6所列。

表6 不同信噪比下TFSEGAN的*STOI*值

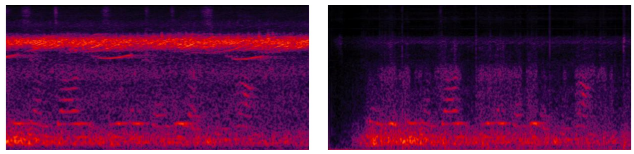
Table 6 *STOI* values of TFSEGAN with different signal-to-noise ratios

model	10 dB	5 dB	0 dB	−5 dB	−10 dB	−15 dB
Noisy	0.897	0.828	0.771	0.707	0.600	0.526
SEGAN	0.869	0.824	0.764	0.677	0.544	0.432
ISEGAN	0.882	0.836	0.775	0.687	0.546	0.423
TFSEGAN	**0.899**	**0.849**	**0.791**	**0.717**	**0.616**	**0.537**

由表6可知，在提升语音可懂度方面，TFSEGAN的性能明显优于SEGAN和ISEGAN。与SEGAN相比，TFSEGAN在提升语音可懂度方面的整体性能平均提高了约8.91%；而在低信噪比条件下，TFSEGAN的性能提升了约11.75%。当输入语音的信噪比为10 dB，5 dB，0 dB，−5 dB，−10 dB和−15 dB时，TFSEGAN的性能比SEGAN分别提升了3.45%，3.03%，3.53%，5.91%，13.25%和24.31%。可以看出，信噪比越低，TFSEGAN的性能越好，这进一步说明了TFSEGAN更适合处理低信噪比语音。

(4)语音频谱比较

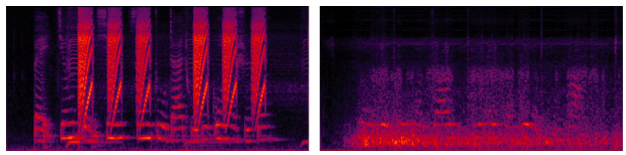
将带噪语音和增强语音的频谱图进行对比，可以直观地看到语音增强的效果和系统的性能。图8−图10分别表示在−15 dB，−10 dB和−5 dB条件下的语音频谱对比图。其中，图8(a)表示带噪语音的频谱图，图8(b)表示增强语音的频谱图。



(a) (b)

图8 −15 dB条件下带噪语音和增强语音的频谱图

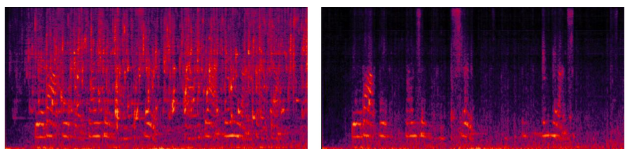
Fig. 8 Spectrum diagram of noisy speech and enhanced speech at −15 dB



(a) (b)

图9 −10 dB条件下带噪语音和增强语音的频谱图

Fig. 9 Spectrum diagram of noisy speech and enhanced speech at −10 dB



(a) (b)

图10 −5 dB条件下的带噪语音和增强语音的频谱图

Fig. 10 Spectrum diagram of noisy speech and enhanced speech at −5 dB

结束语 为了解决SEGAN存在的问题，本文提出了基于

时频域生成对抗网络的语音增强算法 TFSEGAN。在模型结构方面，TFSEGAN 采用了时频域双判别器的模型结构，时域判别器的输入为语音样本的时域特征，频域判别器的输入为语音样本的频域特征。在训练过程中，时域判别器将语音样本的时域分布信息作为判别标准，而频域判别器将语音样本的频域分布信息作为判别标准。在两个判别器的作用下，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，TFSEGAN 采用了时频域 L1 损失函数。在 TFSEGAN 的训练过程中，如果生成器仅依靠判别器的监督作用来捕获语音样本在时域和频域上的分布信息，那么其捕获到的分布信息将是十分模糊且不具备现实性的。为了解决该问题，本文提出了时频域 L1 损失函数。

实验表明，在低信噪比条件下，TFSEGAN 的语音质量和可懂度相比 SEGAN 分别提升了约 17.45%和 7.83%。相比 SEGAN，TFSEGAN 的语音可懂度在各信噪比条件下都大于带噪语音的可懂度，说明 TFSEGAN 在很大程度上解决了 SEGAN 语音可懂度太低的问题。另外，相比 SEGAN，TFSEGAN 能够捕获语音样本中更加详细的分布信息，尤其是语音样本在频域中的分布信息；TFSEGAN 的生成器生成的增强语音样本在时域和频域上与干净语音样本更加相似，其细粒度和现实性也更好。

参 考 文 献

[1] BOLL S F. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on Acoustics Speech & Signal Processing, 1979, 27(2): 113-120.

[2] LIM J S, OPPENHEIM A V. Enhancement and bandwidth compression of noisy speech[J]. Proceedings of the IEEE, 2005, 67(12): 1586-1604.

[3] MCAULAY R J, MALPASS M L. Speech enhancement using a soft-decision noise suppression filter[J]. IEEE Trans. Acoust. Speech Signal Process, 1980, 28(2): 137-145.

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

[5] WANG D L. On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis[M]. Springer, US, 2005.

[6] SRINIVASAN S, ROMAN N, WANG D L. Binary and ratio time-frequency masks for robust speech recognition[J]. Speech Communication, 2006, 48(11): 1486-1501.

[7] OORD A, DIELEMAN S, ZEN H, et al. Wavenet: A generative model for raw audio[J]. arXiv: 1609.03499, 2016.

[8] QIAN K, ZHANG Y, CHANG S, et al. Speech Enhancement Using Bayesian Wavenet[C]// Interspeech. 2017: 2013-2017.

[9] RETHAGE D, PONS J, SERRA X. A wavenet for speech denoising[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5069-5073.

[10] PASCUAL S, BONAFONTE A, SERRA J. SEGAN: Speech enhancement generative adversarial network[J]. arXiv: 1703.09452, 2017.

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

[12] ZHANG Z, DENG C, SHEN Y, et al. On loss functions and recurrency training for GAN-based speech enhancement systems[J]. arXiv: 2007.14974, 2020.

[13] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Networks[J]. Advances in Neural Information Processing Systems, 2014, 3: 2672-2680.

[14] MIRZA M, OSINDERO S. Conditional Generative Adversarial Nets[J]. Computer Science, 2014: 2672-2680.

[15] ODENA A. Semi-supervised learning with generative adversarial networks[J]. arXiv: 1606.01583, 2016.

[16] DONAHUE J, KRÄHENBÜHL P, DARRELL T. Adversarial feature learning[J]. arXiv: 1605.09782, 2016.

[17] MAO X, LI Q, XIE H, et al. Least squares generative adversarial networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2794-2802.

[18] YUAN W H, SHI Y L, HU S D, et al. A Speech Enhancement Approach Based on Fusion of Time-Domain and Frequency-Domain Features[J]. Computer Engineering, 2021, 47(10): 75-81.

[19] LIU H, LI Y, YUAN H Q, et al. Speech Signal Separation Based on Generative Adversarial Networks[J]. Computer Engineering, 2020, 46(1): 302-308.

[20] LIU S H, SUN X, LI C B. Emotion Recognition Using EEG Signals Based on Location Information Reconstruction and Time-Frequency Information Fusion[J]. Computer Engineering, 2021, 47(12): 95-102.

YIN Wen-bing, born in 1997, postgraduate. His main research interests include speech enhancement and so on.

GAO Ge, born in 1973, Ph.D, professor, is a member of China Computer Federation. His main research interests include speech processing and computer vision.

（责任编辑：柯颖）