DOI:10.19651/j.cnki.emt.2107441

基于 WDGAN-div 的语音增强方法*

韩鑫怡[1] 张洪德[1] 柳 林[2] 柳 扬[1]
(1.陆军工程大学 通信士官学校 重庆 400035; 2.合肥讯飞数码科技有限公司 合肥 230088)

摘 要: 针对在低信噪比环境下传统语音增强方法适应性差和增强效果不理想的问题,提出一种基于 Wasserstein 散度的深度生成对抗网络的语音增强方法。该方法以 5 个生成器和 1 个判别器为基础组成深度生成对抗网络,利用 5 个生成器进行 5 次增强处理,有效提高对抗网络在低信噪比条件下的增强效果,使用 Wasserstein 散度优化网络训练,改善传统 GAN 网络训练过程中存在的训练不稳定等问题,提高深度生成对抗网络训练的稳定性。在低信噪比环境下该方法相比于传统语音增强方法噪声适应性和增强效果都有明显提升。实验结果表明,与原始带噪语音相比,增强语音的分段信噪比平均提高 6.1 dB,语音质量感知评估测度和短时客观可懂度分别平均提升 28.9%和 10.6%。

关键词: 语音增强;生成对抗网络;深度学习;卷积神经网络;Wasserstein 散度
中图分类号: TP391.4 **文献标识码:** A **国家标准学科分类代码:** 510.40

Speech enhancement method based on WDGAN-div

Han Xinyi[1] Zhang Hongde[1] Liu Lin[2] Liu Yang[1]
(1. Communication Sergeants College, PLA Army Engineering University, Chongqing 400035, China;
2. Hefei Iflytek Digital Technology Limited Company, Hefei 230088, China)

Abstract: Aiming at the problem of poor adaptability and unsatisfactory enhancement effects of traditional speech enhancement methods in low signal-to-noise ratio environments, this paper proposes a speech enhancement method based on Wasserstein divergence deep generative adversarial networks. The DGAN is based on five generators and one discriminator. Five generators are used to enhance the noisy speech signal five times, which effectively improves the enhancement effect of the DGAN in low signal-to-noise ratio environments. At the same time, Wasserstein divergence is used to optimize the network training which can solve the problems in the traditional GAN training process and improve the stability of the DGAN training process. Comparing with traditional speech enhancement methods, the noise adaptability and enhancement effect of this method are significantly improved in low signal-to-noise ratio environments. The experimental results show that, compared with the original noisy speech, SegSNR of the enhanced speech is improved by an average of 6.1 dB. PESQ is increased by an average of 28.9% and STOI is increased by an average of 10.6%.

Keywords: speech enhancement; generative adversarial networks; deep learning; convolutional neural networks; Wasserstein divergence

0 引 言

随着人工智能技术的发展,语音信号处理技术的应用越来越广泛,语音增强作为其中的关键环节,也是研究者关注的热点。目前语音增强可以分为两大类:无监督增强和有监督增强,其中无监督增强是指无需使用大数据预先训练的一类方法,所需资源较少,有着计算量小,实时性高等优点,如比较经典的谱减法(spectral subtraction, SS)。但此类方法仅在平稳噪声环境下有较好的效果,同时由于需在语音和噪声之间进行某些不合理假设,导致增强效果进一步受到影响。有监督增强是以深度学习[1]和神经网络[2]为基础,通过建立关于信号自身特性和模型参数的优化函数,预先学习后再进行增强的一类方法,此类方法相对于无监督增强能够更好的处理非平稳噪声,且有较好的噪声泛化能力,如文献[2]提出的基于深层神经网络的语音增强方法(deep neural net, DNN)。但无论是无监督或有监督语

收稿日期:2021-07-29
*基金项目:军内科研项目(LJ20191C070659)资助

音增强方法，往往忽略语音相位信息的变化，导致增强效果达不到最佳[3]。2014 年，Goodfellow 等[4]以零和博弈思想为基础提出生成对抗网络（generative adversarial net，GAN）在图像处理、计算机视觉等领域取得较好的效果。2017 年文献[5]首次将 GAN 应用于语音增强领域，提出基于 GAN 的语音增强方法（speech enhancement generative adversarial network，SEGAN），由于 SEGAN 方法在时域直接处理信号，因此有效地改善了因相位变化信息的缺失导致增强效果不佳的缺陷，但该方法也存在 GAN 中常见的训练不稳定、模式崩溃、梯度消失等问题，由于这些问题的影响，导致基于 GAN 的语音增强模型的增强效果受到影响[6-7]。同年，文献[8]针对传统 GAN 存在的训练不稳定等问题提出一种基于 Wasserstein 距离的生成对抗网络（Wasserstein generative adversarial networks，WGAN），使用 Wasserstein 距离替代传统 GAN 中的 JS 散度，以改进 JS 散度在两分布差异较大时无法准确衡量真实数据分布和生成数据之间的距离的缺陷，提高模型训练的稳定性，但 WGAN 受的 1-Lipschitz 条件的约束，导致其网络性能受到一定影响。

针对目前低信噪比条件下现有增强方法增强效果不理想、传统 GAN 存在的训练不稳定、WGAN 受约束条件限制性能受影响等问题，本文提出一种基于 Wasserstein 散度的深度生成对抗网络（Wasserstein divergence deep generative adversarial networks，WDGAN-div）的语音增强方法。为提高低信噪比环境下的增强效果，在传统 GAN 仅使用单个生成器的基础上，WDGAN-div 使用多个生成器组成深度生成对抗网络，并利用 Wasserstein 散度即能够保留 Wasserstein 距离的良好性质，又无需服从 1-Lipschitz 的约束这一特点，使网络模型能够在更宽松的条件下更稳定的训练。实验结果表明，本文方法的网络模型训练稳定，低信噪比环境下的增强效果有明显提升。

1 基于 Wasserstein 散度的生成对抗网络

由于传统 GAN 目标函数使用 JS 散度无法有效地衡量真实数据分布和生成数据之间的距离[7]，导致模型往往难以训练，容易出现模式崩溃和梯度消失等问题。因此文献[8]提出使用 Wasserstein 距离替代 JS 散度，Wasserstein 距离的定义如下：

$$W(p_r,p_z)=\inf_{\gamma\sim\Pi(p_r,p_z)}E_{(x,y)\sim\gamma}[\|x-y\|] \tag{1}$$

式中：p_r 和 p_z 分别表示真实数据分布和生成数据分布；$\Pi(p_r,p_z)$ 表示 p_r 和 p_z 的联合分布。

Wasserstein 距离即为在联合分布下能够取到所有 x 和 y 距离期望的下界值。因此基于 Wasserstein 距离的 WGAN 的目标函数为：

$$\begin{cases}L(D)=\min E_{x\sim p_r}[D(x)]-E_{z\sim p_z}[D(G(z))]\\ L(G)=\min -E_{z\sim p_z}[D(G(z))]\end{cases} \tag{2}$$

式中：z 为随机噪声分布。

虽然 WGAN 相比传统 GAN 有着诸多优点，但其必须添加限制以服从 1-Lipschitz 约束[8]，文献[8]通过权重裁剪（Weight Clipping）将模型参数限制在 $[-C,C]$ 的范围来满足 1-Lipschitz 约束。但实际训练中“强制裁剪”容易导致网络参数的“极端化”，使网络在训练过程中出现梯度消失或爆炸等问题。

为改进上述问题，文献[9]使用梯度惩罚（Gradient Penalty）代替权重裁剪，通过梯度惩罚确保梯度小于 1，以更好的满足 1-Lipschitz 约束，有效避免因参数的“极端化”导致的梯度消失或爆炸等问题，使网络模型训练稳定性增强，生成数据更接近真实数据。于是基于 Gradient Penalty 的 WGAN（WGAN-GP）的目标函数可以表示为：

$$\begin{cases}L(D)=\min E_{z\sim p_z}[D(G(z))]-E_{x\sim p_r}[D(x)]+\\ \quad\lambda E_{X\sim p_{rz}}[(\|\nabla_X D(X)\|-1)^2]\\ L(G)=\min -E_{z\sim p_z}[D(G(z))]\end{cases} \tag{3}$$

式中：$\|\nabla_X D(X)\|$ 表示梯度，λ 为梯度惩罚因子，p_{rz} 是由 p_r 和 p_z 之间随机插值所得到的分布。

虽然 WGAN-GP 能够有效改善 WGAN 中存在的问题，但其依旧必须满足 1-Lipschitz 约束，且梯度惩罚只针对生成分布与真实分布之间的分布空间，缺少对整个分布空间的数据进行惩罚。因此针对 Wasserstein 距离存在的缺陷，文献[6]提出 Wasserstein 散度进行改进，Wasserstein 散度的具体定义如下：

$$W_{k,p}(p_r,p_z)=\max_{D\sim C_1}E_{x\sim p_r}[D(x)]-E_{x\sim p_z}[D(G(z))]-kE_{X\sim p_{radon}}[(\|\nabla_X D(X)\|)^p] \tag{4}$$

式中：$k>0$；$p>1$；C_1 为一阶连续函数族；p_{radon} 表示 Radon 测度。

相比 Wasserstein 距离，Wasserstein 散度仅需满足条件 C_1，解除必须服从 1-Lipschitz 约束的限制。同时得益于通用逼近定理[6]（universal approximation theory）和现代神经网络架构，条件 C_1 能够轻松的通过神经网络实现参数化。因此 Wasserstein 散度是在保留 Wasserstein 距离良好性质的前提下，解除 1-Lipschitz 约束，使网络模型能够在更宽松的条件下更稳定的训练，生成数据也更加接近真实数据。于是基于 Wasserstein 散度的 GAN（WGAN-div）的目标函数为：

$$\begin{cases}L(D)=\max E_{x\sim p_r}[D(x)]-E_{z\sim p_z}[D(G(z))]-\\ \quad kE_{X\sim p_{r(x)}}[(\|\nabla_X D(X)\|)^p]\\ L(G)=\min -E_{z\sim p_z}[D(G(z))]\end{cases} \tag{5}$$

式中：$p_{r(x)}$ 分布在 WGAN-div 中限制较为宽松，可以有多种设置方法[6]，为方便后续仿真实验对比，本文将其设置为与式(3)中 p_{rz} 相同的分布。

2　基于 WDGAN-div 的语音增强

2.1　基于 GAN 的语音增强

文献[5]提出的基于生成对抗网络的语音增强方法(SEGAN),其基本原理是将带噪语音信号通过生成器,生成接近纯净语音信号分布的信号,以实现语音增强,具体模型如图 1 所示。

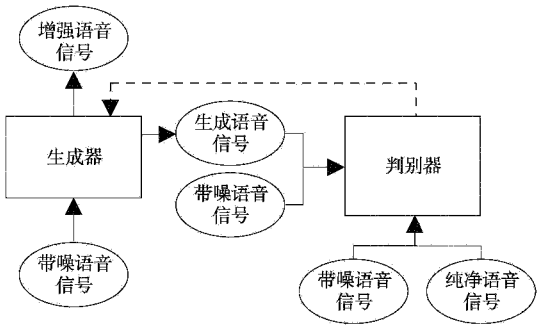

图 1　基于 GAN 的语音增强模型

训练阶段将生成语音信号与纯净语音信号分别输入到判别器中进行判别,同时结合条件生成对抗网络的思想[10],在判别器输入端同时加入带噪语音信号,降低网络参数更新随机性,增强网络训练稳定性,提高降噪性能,而后将判别结果反馈回生成器中,训练不断优化生成器的参数。

增强阶段则直接将带噪语音信号输入生成器中生成增强语音信号。

2.2　基于 WDGAN-div 的语音增强

为提高低信噪比环境下的语音增强效果,本文提出一种基于 Wasserstein 散度的深度生成对抗网络语音增强方法。

在语音增强领域,噪声干扰情况下的生成数据分布与真实数据分布之间可能差异较大,差异衡量的准确度直接影响语音增强的效果,使用 Wasserstein 散度能够更准确的衡量数据间的距离,提升网络模型的增强效果,并且 Wasserstein 散度在保留 Wasserstein 距离良好性质的前提下,解除 1-Lipschitz 约束,使网络模型能够在更宽松的条件下更稳定的训练,有效改善传统 GAN 存在的训练不稳定失等问题。同时针对单个生成器在强噪声干扰条件下增强效果不理想的问题,本文提出使用多生成器组成深度生成对抗网络,利用多个生成器多次生成信号以获得更好的增强效果。

通过多次实验测试,并综合考虑现有设备运算能力和训练时长等多方因素,最终将生成器数量设置为 5 个,既能够保证低信噪比条件下的增强效果,也不会因模型复杂,参数过多而导致过拟合的问题。具体模型如图 2 所示。

整个 WDGAN-div 网络结构由 5 个生成器和 1 个判别器构成,工作流程分为生成和判别两个阶段。生成阶段

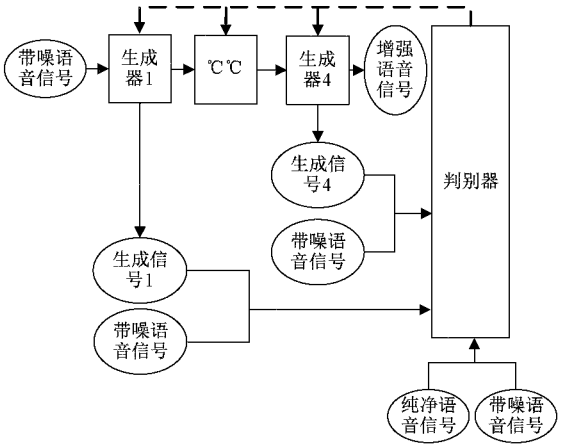

图 2　基于 WDGAN-div 的语音增强模型

5 个生成器"串联"工作,带噪语音信号依次经过 5 个生成器进行处理。判别阶段判别器分别对每个生成器的生成信号单独判别以保证每个生成器之间参数相互独立。

结合图 2 的语音增强模型和 WGAN-div 的目标函数,WDGAN-div 的目标函数可以表示为:

$$\begin{cases} L(D)=\max E_{x\sim p_r}[D(x,z_0)]- \\ \quad \sum_{n=1}^{N}\frac{1}{N}E_{z_{n-1}\sim p_{z_{n-1}}}[D(G_n(z_{n-1}),z_0)]- \\ \quad \sum_{n=1}^{N}\frac{k}{N}E_{X_n\sim p_{rz_{n-1}}}[(\|\nabla_{X_n}D(X_n)\|)^p] \\ L(G)=\min-\sum_{n=1}^{N}\frac{1}{N}E_{z_{n-1}\sim p_{z_{n-1}}}[D(G_n(z_{n-1}),z_0)]+ \\ \quad \lambda\sum_{n=1}^{N}\omega_n\|G_n(z_{n-1})-x\|_1 \end{cases} \tag{6}$$

式中:N 为生成器个数,本文中取值为 5,z_{n-1} 为生成器 n 的输入;z_0 为原始带噪语音信号;$p_{z_{n-1}}$ 表示服从生成器 n 输入数据的分布;$\|\nabla_{X_n}D(X_n)\|$ 表示对生成器 n 的判别梯度。

同时在生成器目标函数中加入纯净语音信号与增强语音信号差值的 L1 范数作为正则项,防止训练过拟合,提高网络性能,其中 ω_n、λ 分别为正则项权重系数和影响因子。

2.3　WDGAN-div 的模型结构

WDGAN-div 的生成器结构如图 3 所示,为类似 U-NET 网络[11]的全卷积网络,分为编码和解码两个部分。为保留更多语音特征信息,提升模型增强效果,加入跳跃链接(Skip Connection)[12]。编码部分由卷积层和卷积池化层构成,为了保留更多信号的细节信息,使用卷积池化层替代传统池化层;解码部分则是与编码部分相对应的反卷积和反卷积池化层;为增强模型鲁棒性,将感知向量 $\boldsymbol{C}$ 添加随机噪声 $\boldsymbol{Z}$ 后输入解码部分。生成器各层卷积核个数分别为 16,32,32,64,128,128,256,512,512,1 024,512,512,256,128,128,64,32,32,16,1,激活函数除最后一层使用 Tanh 函数外,其他各层均使用 PReLU 函数。

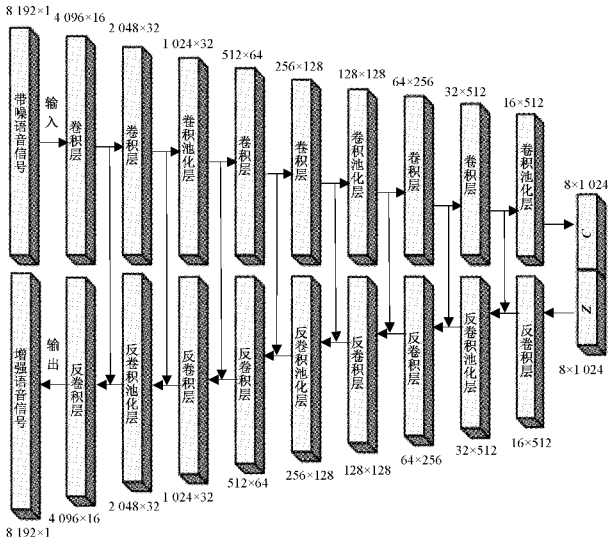

图 3　生成器结构

WDGAN-div 的判别器结构如图 4 所示，类似生成器的编码部分，除最后输出层为全连接层以外，其他各层均为卷积层或卷积池化层。由于常用的标准批量化(batch normalization，BN)与 Wasserstein 散度所需满足的条件 C_1 不兼容，因此本文使用层批量化(layer normalization，LN)替代 BN[9]，以加快网络模型收敛。判别器各层卷积核个数分别为 16，32，32，64，128，128，256，512，512，1024，1，激活函数均使用 Leaky ReLU 函数。同时为防止训练出现过拟合，在输入端添加高斯白噪声，在输出端设置一个 dropout 层。

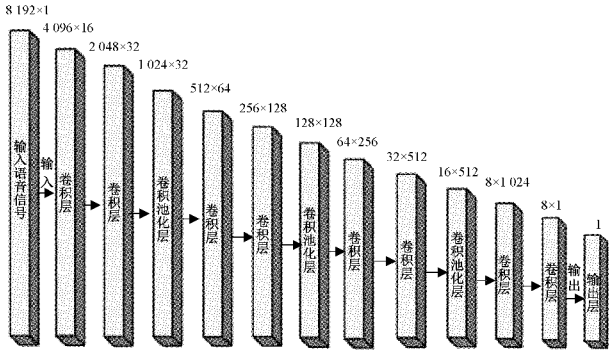

图 4　判别器结构

3　实验仿真

3.1　实验设置

本实验基于 TensorFlow 深度学习框架进行，具体实验环境如表 1 所示。

实验训练集的纯净语音使用 Valentini2016 的训练的纯净语音[13]，包含不同性别的 28 位说话人共计 11 572 条语音。噪声数据使用中科大噪声集[14]的 100 种非言语噪声外加实际录制的 10 种噪声，共计 110 种。以 5 种不同信噪比(−10、−5、0、5、10 dB)构建训练集。测试集的纯净语音使用 Valentini2016 测试集的纯净语音，包含 2 位说话人 824 条语音。测试噪声为 NOISEX-92 数据集中 15 种不同类型噪声。同样以 5 种不同信噪比(−10、−5、0、5、10 dB)构建测试集。训练集和测试集中所有语音都以 16 kHz 采样率进行重采样，并进行分帧处理，帧长设置为 8 192，在训练阶段帧移设置为 50%，测试阶段则为 100%。为增强信号高频分量，在输入端进行预加重，并在输出端进行相应的去加重，预加重和去加重系数均为 0.95。

表 1　实验环境

实验环境	具体配置
操作系统	Windows10 64 位
CPU	AMD Ryzen5 5600X
GPU	NVIDIA RTX3060
内存	16 G
显存	12 G

经过多次实验测试，本算法的最佳参数设置如下：使用的卷积核宽度均设置为 13，步长设置为 2。激活函数 Leaky ReLU 的斜率为 0.3，dropout 层保留率为 0.5。根据文献[6]将 Wasserstein 散度系数分别设置为 $k=2$，$p=6$，正则项权重系数和影响因子分别设置为 $\omega_n=2^{n-N}$，$\lambda=100$。随机噪声 $\boldsymbol{Z}$ 的均值为 0 方差为 1，高斯白噪声方差为 0.5。其余参数均使用方差为 0.02，均值为 0 的截断正态分布初始化。

网络模型的训练采用分批次训练，批量大小为 50，总计训练 200 轮。由于整个网络深且复杂，为缩短训练时间，本文采用文献[15]提出的两种不同更新率(two time-scale update rule，TTUR)的方法进行训练，分别将生成器和判别器的学习率设置为 0.000 1 和 0.000 5，生成器与判别器更新比为 1，即训练 1 次生成器，判别器也训练 1 次。使用参数为 $\beta_1=0$，$\beta_2=0.9$ 的 Adam 优化器进行网络参数更新。

3.2　实验评价标准

为验证本文提出的语音增强方法的实际增强性能，需对增强效果进行语音质量评价。本文分别从语音时域波形、语谱图、分段信噪比 SegSNR)、短时客观可懂度(STOI)以及语音质量感知评估测度(PESQ)5 个方面进行综合验证。

其中，SegSNR 是判断信号中残留噪声的多少，取值越高表示残留噪声越少。PESQ 是针对语音质量进行评价，其取值范围为[−0.5，4.5]之间，分数越高表示语音质量越好。而 STOI 针对语音可懂度进行评价，取值范围为[0，1]，分数越高越优表示语音可懂度越高。

4　性能分析

为全面评估本文基于 WDGAN-div 的语音增强方法(SEWDGAN)的实际增强性能，分别使用无监督与有监督增强方法进行对比实验，并通过单样本具体性能分析和全样本基于统计学总体分析两个角度对实验结果进行分析研究。

无监督增强方法为经典的谱减法(SESS)，有监督增强方法为文献[5]基于 GAN 的语音增强方法(SEGAN)和文献[16]基于 WGAN-GP 的语音增强方法(SEWGAN)，同时为保证实验条件统一，SEGAN 和 SEWGAN 均使用与本文相同的训练集和实验设置进行训练。

关于增强模型的训练，本文通过多次实验测试发现，即使在不同的参数配置条件下，本文使用的 Wasserstein 散度均能够稳定的训练由 5 个生成器组成的深度生成对抗网络，网络模型均能够稳定收敛，且增强效果稳定。相比之下，传统 GAN 语音增强模型，在某些参数设置情况下，会出现网络模型不收敛或出现模式崩溃等问题，导致模型的增强效果受到影响。综上所述，本文提出的方法在网络训练稳定性方面相对于传统 GAN 有明显提升。

4.1　单样本性能分析

选取一段叠加－5 dB 的 Factory 噪声的语音信号作为样本对 4 种方法增强方法进行具体性能分析。从图 5～8 的时域波形图和语谱图对比分析可以发现，4 种增强方法均大幅度的减少噪声的干扰。但从图 6(a)的时域波形图可以发现，即使 SESS 首尾的噪声残留要少于其他 3 种方法，但整体语音波形失真最严重。其主要原因为 SESS 对噪声进行了强制谱减，虽然能够较大幅度的去除噪声，但同时也减去部分语音成分，导致语音失真严重。同时从图 8(a)的语谱图中可以发现，SESS 增强语音在频率 7 000 Hz 左右的频段中产生大量高频的“音乐噪声”。表 2 中 SESS 的 PESQ 分数和 STOI 分数均远低于其他 3 种方法，甚至低于初始值，更是从数据直接验证上述结论。

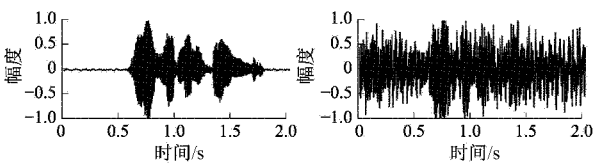

图 5　纯净语音与带噪语音的时域波形图

表 2　4 种增强方法对－5 dB 的 Factory 噪声增强结果

增强方法	SegSNR/dB	PESQ	STOI
NOISE	－6.195	1.098	0.779
SESS	－1.198	1.059	0.735
SEGAN	－3.089	1.343	0.806
SEWGAN	－0.486	1.341	0.805
SEWDGAN	0.193	1.534	0.895

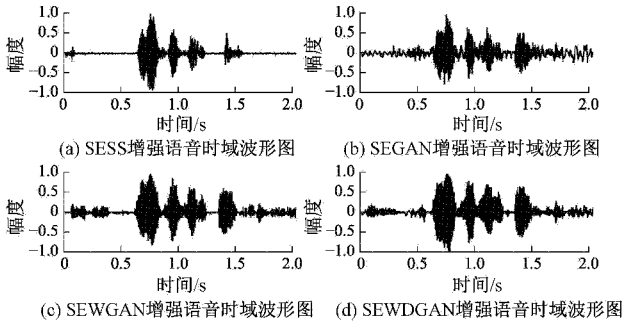

图 6　4 种增强方法增强语音的时域波形图

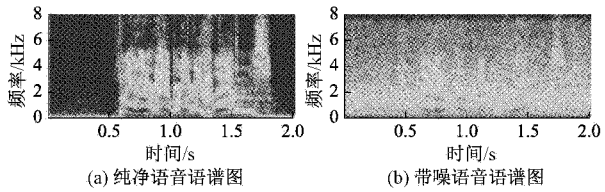

图 7　纯净语音与带噪语音的语谱图

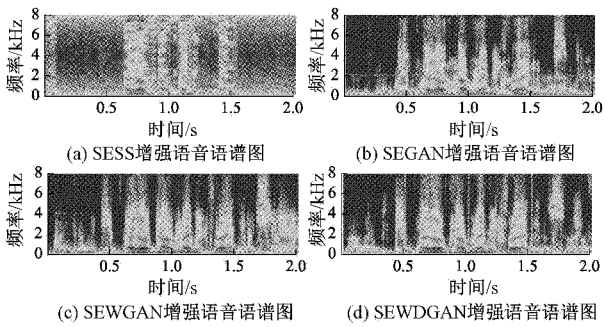

图 8　4 种增强方法增强后的语音的语谱图

SEGAN 增强语音整体存在一定失真，语音部分能量低于其他两种有监督方法，表 2 中 SegSNR 也远低于其他两种有监督方法。SEWGAN 的增强语音整体失真较少，但从图 8(c)的语谱图可以发现，开头前 0.5 s 存在少量低频噪声残余，同时尾部 1.9 s 附近也残留部分噪声。

本文提出的增强方法(SEWDGAN)虽然首尾部分也残留有部分噪声，但语音整体失真最小，尽可能多的保留了原始语音信息。结合表 2 中数据进行分析，本文方法的 STOI 分数、PESQ 分数以及 SegSNR 均为最高，综合增强效果最好。

随机抽取其他噪声样本语音进行单样本实验测试，在其他噪声样本条件下的单样本分析结论均与在－5 dB 的 Factory 噪声条件下的结论相同，说明在其他噪声条件下也具有相同的增强结果。

4.2　全样本性能分析

为更准确更全面地比较 4 种增强方法的增强效果，将 15 种噪声按 5 种信噪比构成的共计 824 条测试语音信号均分别使用 4 种增强方法进行增强处理。表 3～5 为 4 种增强方法 5 种信噪比条件下增强结果的平均值，表 6 为 4 种增强方法增强效果平均的提升程度。

表 3 4 种增强方法不同信噪比条件下的平均 SegSNR

SNR/dB	NOISE	SESS	SEGAN	SEWGAN	SEWDGAN
10	1.647	1.307	5.076	6.845	7.299
5	−1.485	0.671	2.389	4.361	4.902
0	−3.947	−0.257	−0.313	1.501	2.157
−5	−6.391	−1.094	−2.692	−1.091	−0.404
−10	−8.145	−2.100	−4.755	−3.611	−1.789

表 4 4 种增强方法不同信噪比条件下的平均 PESQ 分数

SNR/dB	NOISE	SESS	SEGAN	SEWGAN	SEWDGAN
10	1.404	1.107	1.889	1.870	1.943
5	1.190	1.100	1.515	1.488	1.634
0	1.138	1.068	1.368	1.356	1.486
−5	1.080	1.108	1.203	1.198	1.336
−10	1.055	1.088	1.103	1.113	1.210

表 5 4 种增强方法不同信噪比条件下的平均 STOI 分数

SNR/dB	NOISE	SESS	SEGAN	SEWGAN	SEWDGAN
10	0.876	0.834	0.894	0.891	0.953
5	0.813	0.759	0.843	0.849	0.901
0	0.744	0.701	0.784	0.797	0.851
−5	0.672	0.633	0.698	0.714	0.788
−10	0.590	0.551	0.583	0.614	0.684

表 6 4 种增强方法增强结果的平均提升程度

增强方法	SegSNR/dB	PESQ/%	STOI/%
SESS	3.4	−5.8	−5.9
SEGAN	3.6	19.6	2.8
SEWGAN	5.2	18.8	4.7
SEWDGAN	6.1	28.9	10.6

从表 6 中可以发现，SESS 虽然在不同信噪比环境下 SegSNR 依然能够平均提升 3.4 dB，但强制谱减导致语音失真，同时产生的“音乐噪声”，导致 PESQ 分数和 STOI 分数的提升均为负值，即语音质量和可懂度低于增强前，其原因是语音失真和“音乐噪声”严重影响了语音的质量和可懂度，这也验证了单样本时分析所得出的结论。

SEWGAN 与 SEGAN 虽然在 PESQ 分数即语音质量方面的提升仅相差 0.8%，但 SEWGAN 在 SegSNR 和 SOTI 分数的提升要明显的优于 SEGAN，也就是 SEWGAN 增强后噪声干扰更少且语音可懂度更高。

本文基于 WDGAN-div 的语音增强方法（SEWDGAN）在不同信噪比环境下均能够获得较好的增强效果，无论 SegSNR，PESQ 分数以及 STOI 分数均为最高，其中与与原始带噪语音相比，增强语音的 SegSNR 平均提升 6.1 dB，STOI 分数平均提升 10.6%，而 PESQ 分数的平均提升更是达到 28.9%，远高于其他 3 种方法。说明本文提出的增强方法不仅能够有效地降低噪声干扰，提高语音分段信噪比，同时也能够保证增强语音的质量和可懂度。

5 结 论

本文提出的基于 WDGAN-div 的语音增强方法利用 Wasserstein 散度即能够保留 Wasserstein 距离良好性质，又无须服从 1-Lipschitz 的约束这一特点，改善传统 GAN 训练不稳定等问题，使由 5 个生成器和 1 个判别器组成的深度生成对抗网络能够在更宽松的条件下更稳定的训练。相比目前使用单个生成器的语音增强方法，本文提出的方法使用 5 个生成器对带噪信号进行 5 次增强处理，在低信噪比条件下的增强效果更好。实验结果表明，本文提出的增强方法能够在低信噪比环境下有效提升语音的信噪比，改善语音质量和可懂度，并且在不同信噪比环境下的增强效果均优于谱减法、基于 GAN 和 WGAN-GP 的语音增强方法。

虽然相比其他单生成器的增强方法有更好的增强效果，但由于本文的增强模型是由 5 个生成器组成的深度生成对抗网络，网络本身较深，在训练过程中需要耗费大量的时间。因此在未来的工作中，还需要继续对网络架构进行优化改进，减少网络复杂程度，提高训练速度。

参考文献

[1] JINKYU L, JAN S, ZAKIZADEH S T, et al. Phase-sensitive joint learning algorithms for deep learning-based speech enhancement[J]. IEEE Signal Processing Letters, 2018, 25:1276-1280.

[2] 徐勇. 基于深层神经网络的语音增强方法研究[D]. 合肥：中国科学技术大学，2015.

[3] PALIWAL K K, WÓJCICKI K K, SHANNON B J. The importance of phase in speech enhancement[J]. Speech Communication, 2011, 53(4):465-494.

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

[5] PASCUAL S, BONAFONTE A, SERRÀ J. SEGAN: Speech enhancement generative adversarial network[C]. Conference of the International Speech Communication Association (ISCA), Stockholm, Sweden: ISCA, 2017:3642-3646.

[6] WU J, HUANG Z, THOMA J, et al. Wasserstein divergence for GANs[J]. Computer Vision-ECCV2018, 2018:673-688.

[7] ARJOVSKY M, BOTTOU L. Towards principled methods for training generative adversarial networks[C]. International Conference on Learning Representations

(ICLR), Los Toulon, France: ICLR, 2017:1-17.

[8] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN[J]. Neural Information Processing Systems Foundation, 2017(12):5768-5778.

[9] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of Wasserstein GANs[C]. Conference of the Neural Information Processing Systems(NIPS), Los Angeles, USA: NIPS, 2017:5769-5779.

[10] MICHELSANTI D, TAN Z H. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification [C]. Conference of the International Speech Communication Association(ISCA), Stockholm, Sweden: ISCA, 2017:2008-2012.

[11] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation [J]. Medical Image Computing and Computer-Assisted Intervention, 2015:234-241.

[12] DROZDZAL M, VORONTSOV E, CHARTRAND G, et al. The importance of skip connections in biomedical image segmentation [J]. Deep Learning and Data Labeling for Medical Applications, 2016:179-187.

[13] VALENTINI-BOTINHAO C, XIN W, TAKAKI S, et al. Investigating RNN-based speech enhancement methods for noise-robust text-to-speech[C]. Proceedings of 9th ISCA Speech Synthesis Workshop (SSW), Sunnyvale, USA: SSW, 2016:159-165.

[14] 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, 2013, 21(1):65-68.

[15] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium[C]. Conference of the Neural Information Processing Systems(NIPS), Los Angeles, USA: NIPS, 2017:6626-6637.

[16] 叶帅帅. 基于 Wasserstein 生成对抗网络的语音增强算法研究[D]. 北京:北京邮电大学, 2019.

作者简介

韩鑫怡,硕士研究生,主要研究方向为战场信息处理。

E-mail: 976455756@qq.com

张洪德,博士,副教授,主要研究方向为电子侦察研究。

E-mail: hdzhang264@126.com

柳林,硕士,工程师,主要研究方向为语音识别、自然语言处理、人工智能。

E-mail: linliu@iflytek.com

柳扬,硕士,副教授,主要研究方向为信号处理。

E-mail: 461687857@qq.com