基于生成对抗网络合成噪声的语音增强方法研究

夏 鼎，徐文涛
（南京航空航天大学 理学院，江苏 南京 211106）

摘 要：在语音增强领域，深度神经网络通过对大量含有不同噪声的语音以监督学习方式进行训练建模，从而提升网络的语音增强能力。然而不同类型噪声的获取成本较大，噪声类型难以全面采集，影响了模型的泛化能力。针对这个问题，提出一种基于生成对抗网络(Generative Adversarial Networks，GAN)的噪声数据样本增强方法，该方法对真实噪声数据进行学习，根据数据特征合成虚拟噪声，以此扩充训练集中噪声数据的数量和类型。通过实验验证，所采用的噪声合成方法能够有效扩展训练集中噪声来源，增强模型的泛化能力，有效提高语音信号去噪处理后的信噪比和可理解性。

关键词：语音增强；生成对抗网络；数据增强
中图分类号：TN912.3 **文献标识码**：A DOI：10.16157/j.issn.0258-7998.200327

中文引用格式：夏鼎，徐文涛. 基于生成对抗网络合成噪声的语音增强方法研究[J].电子技术应用，2020，46(11)：56-59，64.
英文引用格式：Xia Ding，Xu Wentao. Research on speech enhancement method based on generating noise using GAN[J]. Application of Electronic Technique，2020，46(11)：56-59，64.

Research on speech enhancement method based on generating noise using GAN

Xia Ding，Xu Wentao
(School of Science，Nanjing University of Aeronautics and Astronautics，Nanjing 211106，China)

Abstract：In the field of speech enhancement，deep neural network can improve the enhancement ability of the model by training and modeling a large number of data with different noises in the supervised learning way. However，the acquisition cost of different types of noise is large and the noise types are difficult to be comprehensive，which affects the generalization ability of the model. Aiming at this problem，this paper proposes a noise data augmentation method based on generative adversarial network(GAN)，which learns from the real noise data and synthesizes virtual noises according to the data features，so as to expand the number and type of the noise data in the training set. Experimental results show that the method of noise synthesis adopted in this article can effectively expand the source of noise in the training set，enhance the generalization ability of the model，and effectively improve the signal-to-noise ratio and intelligibility of speech signal after denoising.

Key words：speech enhancement；generative adversarial network；data augmentation

0 引言

在语音信号处理的过程中，背景噪声和环境干扰严重影响了信号处理的可靠性，需要通过语音增强处理方法去除信号中的噪声干扰，改善含噪语音的质量。因此，语音增强技术在语音识别、听力辅助和语音通信等领域中具有非常重要的作用。

传统的语音增强方法有谱减法[1]、维纳滤波[2-3]以及之后出现的基于统计模型的处理方法[4]等，这些方法都是基于已知噪声的统计特性来进行建模，得到噪声的功率谱信息，对含噪语音信号进行降噪处理，以估计纯净语音信号。这些传统方法的准确性严重依赖数据特征工程处理方法和数据类型，对于未知的噪声干扰，其适应能力较差[5]。随着人工智能的发展，深度神经网络被应用于语音增强领域[6]。利用深层神经网络的特征学习，可以将含噪语音映射为纯净语音，达到去除噪声的目的。为了提高深度神经网络进行语音增强方法的泛化能力，最直接的手段是进行数据增强，包括增加数据的多样性、扩大数据集等。实验表明，在深度神经网络训练的过程中采用更多种类的噪声数据，语音信噪比质量可以显著提高[7-8]。但是，真实的噪声数据获取难度较大，成本较高，这限制了网络去噪能力的适用性。针对这一问题，本文基于生成对抗网络 GAN 设计了一种训练数据集增强方法，通过生成虚拟噪声，扩充训练集中噪声数据的类型和数量，提高模型的泛化能力。

1 生成对抗网络

1.1 GAN 的简单介绍

GAN[9]可对样本数据的分布规律进行特征学习，并以此生成具有相似规律的数据，作为一个生成模型，最

初被应用于生成图像。经典的 GAN 由两部分组成：一个生成器 G 和一个判别器 D。其均由神经网络构成，经过训练，生成器能够对样本的分布规律进行建模，并根据建立出来的新分布生成新的数据，判别器能够判别样本是来自真实数据样本的分布还是来自生成器。

在对 GAN 模型的参数训练过程中，生成器从潜在空间中随机取样作为输入，其输出需要尽量模仿训练集中的真实样本，去尽可能欺骗判别器。判别器的输入为真实样本或生成器的输出，其目的是尽力分辨生成器的输出和真实样本。两个网络相互对抗、不断调整参数，最终使得判别器无法判断出生成器的输出结果是否真实。生成器产生的数据和真实样本便具有相似的特征。

生成对抗网络用以下的目标函数进行训练：

$$\min_G \max_D V(D,G)=E[\log(D(\boldsymbol{y}))]+E[\log(1-D(G(z)))] \tag{1}$$

其中，$\boldsymbol{y}$ 是真实数据的分布上一个真实样本，z 是随机分布上一个随机噪声向量，G 为生成器，D 为判别器。

经典 GAN 本质上学习的是随机噪声向量 z 到目标真实样本 $\boldsymbol{y}$ 的一个映射，即 $G(z)\rightarrow\boldsymbol{y}$。在语音信号的去噪处理中，要保证语音信号的不变性，因此需要对网络设置条件，带条件的生成对抗网络[10](cGANs)学习的是在已知样本 x 下，随机噪声向量 z 到目标真实样本 $\boldsymbol{y}$ 的一个映射，即 $G(z,x)\rightarrow\boldsymbol{y}$。cGAN 用以下的目标函数进行训练：

$$\min_G \max_D V(D,G)=$$
$$E[\log(D(\boldsymbol{x},\boldsymbol{y}))]+E[\log(1-D(\boldsymbol{x},G(\boldsymbol{x},z)))] \tag{2}$$

1.2 噪声生成网络 WGAN

经典 GAN 作为一个生成模型，理论上已经能够进行噪声数据的生成，但是实验上却容易出现问题，在文献[11]中指出经典的生成对抗网络容易出现模式崩坏，即生成的声音虽然与真实数据相似，但是生成的数据之间缺乏多样性，这与本文数据增强的目标相违背。

借鉴 Wasserstein GAN[11](WGAN)这个改进后的神经网络，相对于经典 GAN，其做了以下 3 点改动：

(1)D 最后一层去掉激活层。

(2)G 和 D 的损失函数(loss)不取对数。

(3)每次更新 D 的参数之后，将其绝对值截断到不超过一个固定常数，即梯度截断(gradient clipping)；或者使用梯度惩罚(gradient penalty)。

WGAN 所做的 3 点改动解决了 GAN 训练困难和不稳定、模式坍塌等问题，而且 G 的损失函数越小，对应生成的图片质量就越高，WGAN 虽然需要更长的训练时间，但收敛更加稳定。

基于 WGAN 如上的优点，设计了如下的网络结构。

1.2.1 输入和输出

为了减少特征化工作量，对原始的一维波形声音(即时域信号)直接进行训练，这一点在 SEGAN[12]中证明是有效的。为了增强训练数据，对于采样到 16 kHz 的时域信号数据进行分帧，每 16 384 个(约 1 s)数据为一个训练样本。

本文的目的是让训练好的生成器 G 去生成不同于已有真实噪声的人造噪声数据，其输出也是一维的时域信号。这样一种端到端的训练网络在减少特征化工程量的同时，提高了网络的优化效果。

1.2.2 网络结构

参考 SEGAN 针对语音信号处理的网络结构，设计出针对于语音生成的 WGAN。其中，依据网络的输入、输出的特征维度，即形状为 1×16 384 的一维时域信号，生成器 G 的输出和判别器 D 的输入均设置为 1×16 384。

生成器 G 的输入为 100 维的随机噪声，其分布服从于标准正态分布，经过全连接层和维数变换层(Reshape)后，得到 1 024×8 的特征信号，再经过 11 次一维反卷积(deconv)和 10 次带参数的非线性激活(PReLU)操作后得到形状为 1×16 384 的输出，其中反卷积使用的滤波器大小均为 32，卷积步长为 2，填充(padding)为 15。最后一层使用双曲正切激活(Tanh)将数据缩放到 −1~+1 之间。图 1 简要展示了网络的结构。

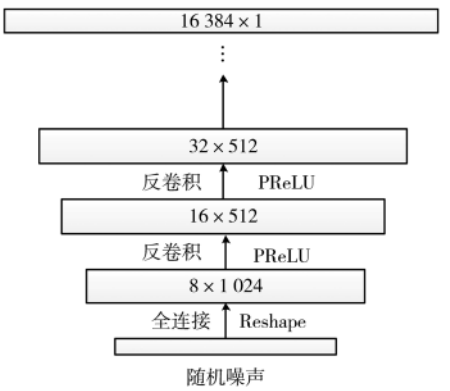

图 1 WGAN 的生成器

判别器 D 的输入为 1×16 384 的时域信号，经过 10 次卷积(conv)、批归一化(Batch Normalization，BN)和非线性激活(LeakyReLU)的处理操作，得到 2 048×16 的特征信号，其中卷积使用的滤波器大小为 31，卷积步长为 2，填充(padding)为 15；然后使用卷积核为 1×1 的卷积操作进行降维，并减少训练的参数，将 2 048×16 的特征信号转化为 1×16 的特征信号，最后通过一个全连接层得到一个有关真假判断的数字，注意这里不使用激活函数。在第 3 层、第 6 层、第 8 层之后使用正则化技术(dropout)减少过拟合。图 2 简要展示了网络的结构。

下面展示了网络的具体参数：

(1)生成器 G

8×1 024，16×512，32×512，64×256，128×128，256×64，512×32，1 024×32，2 048×16，4 096×16，8 192×16，16 384×1

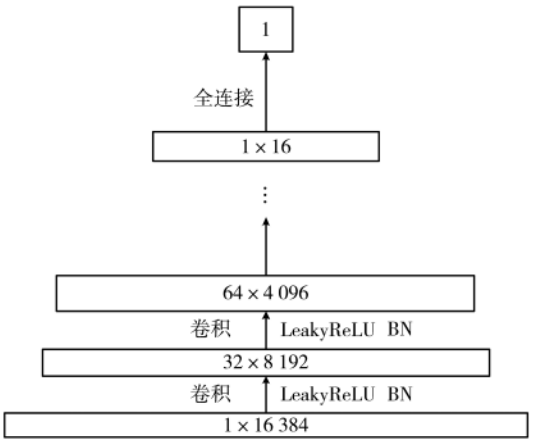

图 2 WGAN 的判别器

层与层之间用带参数的非线性激活(PReLU),最后一层用双曲正切函数激活(Tanh)。

(2)判别器 D

16 384×1,8 192×32,4 096×64,2 048×64,1 024×128,512×128,256×256,128×256,64×512,32×512,16×2 048,16×1,1

层与层之间用批归一化(Batch Normalization,BN)和非线性激活(LeakyReLU),最后一层不使用非线性激活(Sigmoid)。

1.3 语音增强网络 SEGAN

SEGAN[12]类似于 VAE/GAN[13]的网络结构,其生成器 G 则类似于自动编码器[14],可以进行语音增强。在编码阶段,输入信号被投影,通过多个跨步卷积(strided conv)和非线性激活(PReLU)操作,直到得到一个低维向量 c 的压缩表示,它与噪声向量 z(一般服从于标准正态分布)相连接。在解码阶段,编码过程通过反卷积(deconv)和带参数的非线性激活(PReLU)操作,将低维向量 c 转化为一维的语音数据。

如图 3 所示,生成器 G 具有跳过连接的特点,每个编码层直接连接到其相应的解码层,绕过了编码的阶段,将原始数据的特性保存在了解码过程中,这种结构使得语音增强的效果更加稳定。

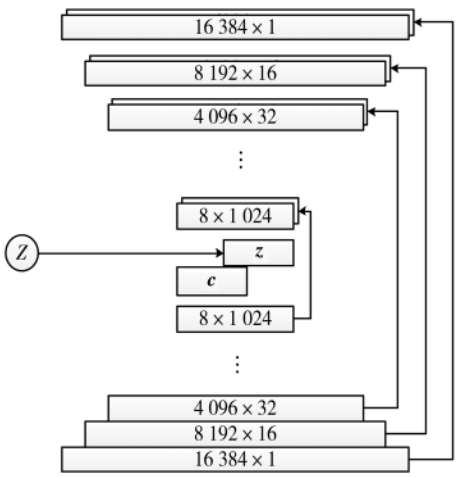

图 3 SEGAN 的生成器

具体的网络参数如下:

(1)编码器

16 384×1,8 192×16,4 096×32,2 048×32,1 024×64,512×64,256×128,128×128,64×256,32×256,16×512,8×1 024

(2)解码器

解码器与编码器相同,但是还有服从正态分布上的 z 向量,所以解码器的结构为:

8×2 048,16×1 024,32×512,64×512,128×512,256×256,512×128,1 024×128,2 048×64,4 096×64,8 192×32,16 384×2

(3)判别器

16 384×1,4 096×64,2 048×64,1 024×128,512×128,256×256,128×256,64×512,32×512,16×1 024,8×2 048,8×1,1

2 实验

2.1 实验配置

WGAN 在来自 Perception and Neurodynamics 实验室的 100 类噪声 NonSpeech100[15]上进行训练,以此训练出具有生成噪声的能力的 WGAN,来进行数据增强,具体生成 NonSpeech100 50%规模的不同的噪声(采用分帧进行训练,设置 1 s 为一个数据,统计结果,NonSpeech100 共有 412 个数据,故需要生成 204 段噪声)。

为了量化测试泛化能力,设计了一组对比实验。设置了两个包含不同数量噪声种类的训练集,训练集中所采用的不含噪的语音是 TIMIT[16]语料库的训练集,所用的噪声数据为 NonSpeech100 以及由 WGAN 得到的生成噪声。

两个训练集采用不同的噪声数据集合成。具体合成方法如下。

训练集 1 由 Nonspeech100 真实噪声以及 TIMIT 语音合成,训练集 2 由 Nonspeech100 和生成噪声共同与 TIMIT 语音合成。两个训练集均包含分窗后的 8 576 段(每段 1 s)含噪语音,每个语音样本采用以下的方法合成:从 8 576 个样本中无放回地随机选取 1 个样本,同时从噪声集合中(训练集 1 有 412 个样本,训练集 2 有 618 个样本)有放回地随机选取 1 种噪声,将该噪声按照 0 dB、-5 dB、5 dB 中的随机一种的全局信噪比和纯净语音样本合成。

由于噪声选取是有放回的,为了噪声选取的极端情况出现,每个样本被选取的次数限制为 20 次,超过 20次便不再放回。

训练集所使用的纯净语音为 TIMIT 语料库的测试集,噪声数据使用的是噪声库 Noisex92[17]。纯净语音集合共 3 264 个样本,噪声集合共 4 004 个样本。按照 3 个不同的全局信噪比合成 3 个测试集,分别为:-5 dB 测试

集、0 dB 测试集、5 dB 测试集。

WGAN 训练时，每 5 个 epoch 设置一个断点，训练50 个 epochs，batch size 设置为 8，在每个断点生成一定的噪声，最后选取最佳的生成噪声的断点进行噪声生成。

SEGAN 训练时，每 5 个 epoch 设置一个断点，训练 100 个 epochs，batch size 设置为 64，之后评价语音质量在各个 epoch 上进行测试，选取最佳的数值。

2.2 实验结果

2.2.1 生成噪音数据样本

为了证实 WGAN 生成不同类型噪声的能力，随机选取两个经 50 个 epoch 训练后的生成噪声，画出两段虚拟噪声的语谱图，如图 4 和图 5 所示，可以看出两段噪声在频域的能力分布特征相差很大，以此证明了 WGAN 生成多类型的噪声的能力。

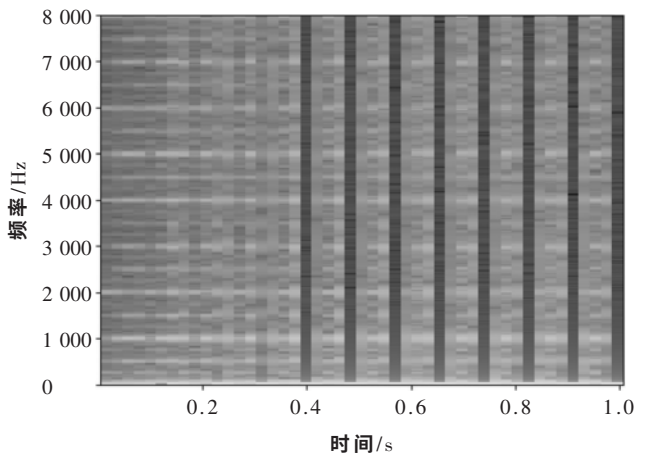

图 4 生成噪声 1 的语谱图

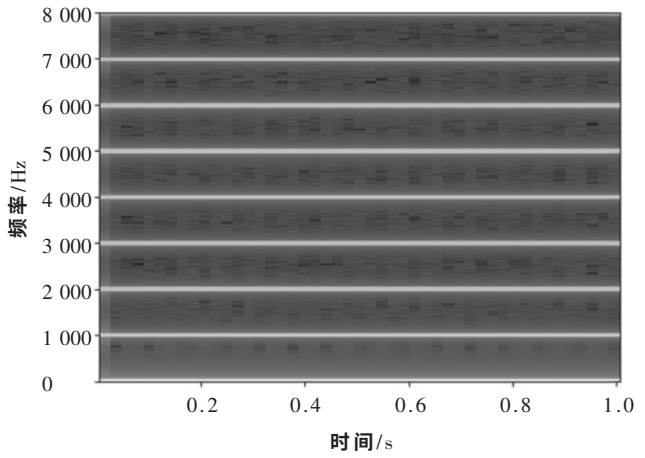

图 5 生成噪声 2 的语谱图

2.2.2 基于 SEGAN 模型的语音增强性能比较

评价指标：

(1)主观语音质量评估(Perceptual Evaluation of Speech Quality，PESQ)：属于客观 MOS 值评价方法的一种，PESQ 得分范围在 1.0~4.5 之间，得分越高表示语音质量越好。

(2)短时客观可懂度(Short Time Objective Intelligibility，STOI)：STOI 得分范围在 0~1 之间，得分越高表示语音可懂度越好。

(3)分段信噪比(dB)：分段计算整个时间轴上的语音信号与噪声信号的平均功率之比。

各评价指标上的增强效果比较如表 1~表 3 所示。

表 1 SEGAN-baseline 下不同训练集的语音增强 PESQ 比较

训练集	测试集信噪比/dB		
	-5	0	5
训练集 1	1.69	1.95	2.23
训练集 2	1.77	2.02	2.35

表 2 SEGAN-baseline 下不同训练集的语音增强 STOI 比较

训练集	测试集信噪比/dB		
	-5	0	5
训练集 1	0.662 7	0.703 7	0.777 4
训练集 2	0.670 2	0.723 1	0.787 2

表 3 SEGAN-baseline 下不同训练集的语音增强分段信噪比比较

训练集	测试集信噪比/dB		
	-5	0	5
训练集 1	-2.28	-0.71	0.42
训练集 2	-2.11	-0.65	0.45

可以看出，3 种评价指标下，训练集 2(即由 Non-speech100 和生成噪声共同与 TIMIT 语音合成的训练集)对应的 SEGAN 模型在 3 种不同信噪比的测试集上都取得了更好的效果。

3 结论

深度神经网络是一种基于大数据的模型，增加数据集合多样性是提高模型泛化能力的简单又有效的重要方法。由于真实噪声收集成本高，本文基于 WGAN 设计了一个真实噪声驱动的，可以生成人造噪声的网络模型，并且通过实验证明了用数据增强后的噪声集训练出的 SEGAN 模型有更好的语音增强效果。进而表明了本文所使用的 WGAN 能够生成有效的人造噪声数据，有效提高了语音增强模型对于不同噪声背景的声音数据的处理能力。

参考文献

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

[2] CHEN J D，BENESTY J，HUANG Y T，et al.New insights into the noise reduction wiener letter[J].IEEE Transactions

(下转第 64 页)

melanoma detection[J].Open Medicine, 2018, 13: 9-16.
[5] 黄海新, 张东.基于深度学习的人脸活体检测算法[J].电子技术应用, 2019, 45(8): 44-47.
[6] HAMEED N, SHABUT A M, HOSSAIN M A.Multi-class skin diseases classification using deep convolutional neural network and support vector machine[C].2018 12th International Conference on Software, Knowledge, Information Management and Applications, 2018: 1-7.
[7] LOPEZ A R, GIRO-NIETO X, BURDICK J, et al.Skin lesion classification from dermoscopic images using deep learning techniques[C].Proceeding of the 2017 13th IASTED International Conference on Biomedical Engineer.Piscataway, NJ: IEEE, 2017: 49-54.
[8] 郝子煜, 阿里甫·库尔班, 李晓红, 等.基于 CapsNet 的中国手指语识别[J].计算机应用研究, 2019, 36(10): 3157-3159.
[9] LI Y, QIAN M, LIU P, et al.The recognition of rice images by UAV based on capsule network[J].Cluster Computing, 2019, 22: 9515-9524.
[10] AFSHAR P, MOHAMMADI A, PLATANIOTIS K N.Brain tumor type classification via capsule networks[J].arXiv preprint, arXiv: 1802.10200, 2018.
[11] 余成波, 熊递恩.基于胶囊网络的指静脉识别研究[J].电子技术应用, 2018, 44(10): 15-18.
[12] SABOUR S, FROSST N, HINTON G E.Dynamic routing between capsules[C].Advances in Neural Information Processing Systems, 2017: 3856-3866.
[13] 林少丹, 洪朝群, 陈雨雪.结合胶囊网络和卷积神经网络的目标识别模型[J].电讯技术, 2019, 59(9): 987-994.
[14] 何雪英, 韩忠义, 魏本征.基于深度卷积神经网络的色素性皮肤病识别分类[J].计算机应用, 2018, 38(11): 3236-3240.

(收稿日期: 2020-02-22)

作者简介:
李励泽(1995-), 女, 硕士研究生, 主要研究方向: 电子电路与系统设计、智能信息处理。
张晨洁(1983-), 女, 博士研究生, 主要研究方向: 智能信息处理。
杨晓慧(1963-), 通信作者, 女, 教授, 主要研究方向: 电路与系统, E-mail: yangxiaohui1963@163.com。

(上接第 59 页)

on Audio, Speech, and Language Processing, 2006, 14(4): 1218-1234.
[3] LIM L, OPPENHEIM A V.Enhancement and bandwidth compression of noisy speech[J].Proceedings of the IEEE, 1979, 67(12): 1586-1604.
[4] 王晶, 傅丰林, 张运伟.语音增强算法综述[J].声学与电子工程, 2005(1): 22-26.
[5] 何玉文, 鲍长春, 夏丙寅, 等.基于 AR-HMM 在线能量调整的语音增强方法[J].电子学报, 2014, 42(10): 1991-1997.
[6] 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.
[7] WANG Y, CHEN J, WANG D L.Deep neural network based supervised speech segregation generalizes to novel noises through large-scale training[R].Ohio State University Columbus, 2015.
[8] CHEN J, WANG Y, YOHO S E, et al.Largescale training to increase speech intelligibility for hearingimpaired listeners in novel noises[J].The Journal of the Acoustical Society of America, 2016, 139(5): 2604-2612.
[9] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al.Generative adversarial nets[C].Proceedings of Advances in Neural Information Processing Systems.US: NIPS, 2014: 2672-2680.
[10] MEHDI M, SIMON O.Conditional generative adversarial nets[J].arXiv: 1411.1784, 2014.
[11] MARTIN A, SOUMITH C, LÉON B, et al.Wasserstein GAN[J].arXiv: 1701.07875, 2017.
[12] SANTIAGO P, ANTONIO B, JOAN S, et al.SEGAN: speech enhancement generative adversarial network[J].arXiv: 1703.09452v1, 2017.
[13] ANDERS B L L, OLE W.Autoencoding beyond pixels using a learned similarity metric[J].arXiv: 1512.09300v2, 2016.
[14] PIERRE B.Autoencoders, unsupervised learning, and deep architectures[C].JMLR: Workshop and Conference Proceedings, 2012, 27: 37-50.
[15] Hu Guoning.PNL100 nonspeech sounds[OL].[2020-04-22].http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html.
[16] GAROFOLO J S, LAMEL L F, FISHER W M, et al.TIMIT acoustic-phonetic continuous speech corpus[C].Philadelphia: Linguistic Data Consortium, 1993.
[17] VARGA A, STEENEKEN H J M.Assessment for automatic speech recognition: II.NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems[J].Speech Communication, 1993, 12(3): 247-251.

(收稿日期: 2020-04-22)

作者简介:
夏鼎(1999-), 男, 本科, 主要研究方向: 机器学习。
徐文涛(1989-), 男, 博士, 讲师, 主要研究方向: 信号处理、机器学习。

版权声明