DOI:10.19652/j.cnki.femt.2204536

基于生成对抗网络的车载语音增强应用

石 瑞 杨立东 郭 勇 牛大伟 张丹丹
(内蒙古科技大学信息工程学院 包头 014010)

摘 要:语音增强对智能车载系统和未来汽车工业的发展具有重要意义,为了解决汽车行驶过程中驾驶员语音被噪声污染的问题,提出一种基于高效通道注意力机制的最小二乘生成对抗网络模型。首先在生成网络模型中引入注意力机制,自适应选择一维卷积核大小生成通道权重,在降低模型复杂度的同时带来了明显的性能增益;然后利用最小二乘损失函数来代替 Sigmoid 交叉熵损失函数,使收敛速度更快,避免出现梯度消失的问题;最后经过生成对抗网络对抗博弈不断优化训练,从而实现语音增强。实验表明,该方法相较基线方法在语音质量和清晰度方面都有良好的提升,语音质量感知评估(PESQ)指标平均提升了 3.79%,短时客观可懂度(STOI)指标平均提升了 4.76%,因此更适合实际应用。

关键词:生成对抗网络;语音增强;注意力机制;车载语音系统
中图分类号: TN912 **文献标识码**:A **国家标准学科分类代码**: 510.40

Vehicle voice enhancement application based on generative adversarial network

Shi Rui Yang Lidong Guo Yong Niu Dawei Zhang Dandan
(School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, China)

Abstract: Voice enhancement is of great significance to the development of intelligent on-board system and the future automobile industry. In order to solve the problem of driver voice noise pollution in the process of car driving, a least squares generation adversarial network model based on the efficient channel attention mechanism is proposed. Firstly, the attention mechanism is introduced in the generative network model to automatically select the one-dimensional convolution kernel size to generate the channel weight, which brings obvious performance gain, and then uses the least squares loss function to replace the Sigmoid cross-entropy loss function to make the convergence rate faster and avoid the problem of gradient disappearance. Finally, the speech enhancement is realized. Experiments show that the proposed method has a good improvement in both quality and clarity over the baseline method, The PESQ index average increased by 3.79%, the STOI index average increased by 4.76%, so it is more suitable for practical applications.

Keywords: generative adversarial network; voice enhancement; attention mechanism; vehicle voice system

0 引 言

随着汽车工业的飞跃发展,车载语音系统已成为汽车设备必不可少的组成部分[1],通过语音控制系统为驾驶人员提供一个全新的人机交互体验,解放驾驶员双手,提高驾车专注度,从而有助于安全驾驶,减少交通事故的发生[2],例如通过语音控制命令收拨电话、调节空调温度、播放音乐、设置导航等。在理想安静的环境下,车载语音系统能够准确地识别驾驶员语音信号,然而在现实情况下,汽车行驶中环境复杂多变,各种噪声会影响识别准确率,当噪声污染非常严重时,语音控制系统甚至可能失去正常工作的能力。针对这一问题,语音增强具有重要的实际意义。

语音增强的定义是当语音信号被各种噪声干扰时,从噪声背景中获取有用的语音信号,并抑制、降低噪声干扰。传统的语音增强技术有谱减法[3]、维纳滤波[4]、子空间算法[5]等,这些方法对于平稳噪声的降噪效果比较显著,但是对于非平稳噪声,往往还是不能得到很好的降噪效果。目前在汽车领域大多采用改进的传统去噪方法,但在实际应用中仍有待提高。近年来,随着人工智能的迅速发展,

收稿日期:2022-11-25

深度学习的方法在语音增强中被广泛应用，如卷积神经网络(convolutional neural network，CNN)[6]、深度神经网络(deep neural network，DNN)[7]、循环神经网络(recurrent neural network，RNN)[8]等，这些方法对于非平稳噪声取得了不错的去噪效果。Pascual 等[9]提出了基于生成对抗网络的语音增强(speech enhancement generative adversarial network，SEGAN)模型，这属于端到端的语音增强方法，利用编码器－解码器的全卷积结构，能够快速地进行去噪处理，Yang 等[10]通过正则化改进了基于生成对抗网络的语音增强，使其在低信噪比复杂噪声环境下降噪效果更加明显，但是模型训练缓慢且不稳定。

基于车载环境噪声复杂、时变、不可预测的特点[11]，本文在生成对抗网络(generative adversarial network，GAN)基础上进行了改进，采用注意力机制与最小二乘生成对抗网络(least squares generative adversarial networks，LSGAN)相结合的方法，首先在生成器模型中引入高效通道注意力机制(FCA)，自适应选择一维卷积核大小生成通道权重，再利用最小二乘损失函数代替 Sigmoid 交叉熵损失函数，经过生成对抗网络对抗训练，从而实现语音信号在质量和清晰度的提升。

1 相关工作

1.1 GAN

GAN[12]是通过生成网络(generator，G)和判别网络(discriminator，D)对抗训练来实现生成的样本与真实数据分布极其接近[13-14]。其中 D 的任务是尽可能准确地判断样本是真实数据还是 G 产生的数据；G 的任务是尽可能生成 D 无法区分的样本。这两个网络经过不断地交替训练，最后收敛时，如果 D 无法判断出一个样本的来源，那就认为 G 可以生成与真实数据分布相符的样本[15]。训练过程如图 1 所示。

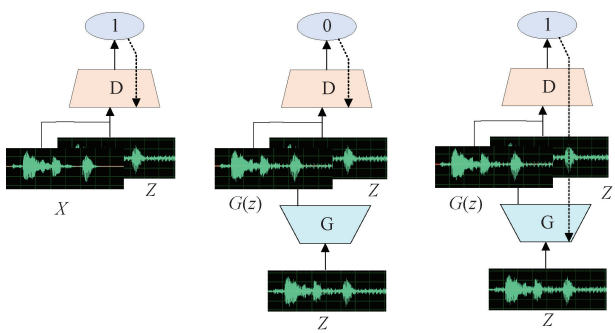

图 1 训练过程

图 1 中 X 表示纯净语音，Z 表示含噪语音，$G(z)$表示生成网络生成的语音，训练过程中 GAN 的目标函数如下：

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z))] \tag{1}$$

对于 G 网络来说，目标是最小化 $V(D,G)$，使生成增强语音 $G(z)$更接近纯净语音 X，对于 D 网络来说，目标是最大化 $V(D,G)$，$\log D(x)$越大，$\log(1-D(G(z))$就越小，使得 D 分辨率越高。通过最大值和最小值的相互对抗，不断优化训练，最终达到纳什平衡。

1.2 LSGAN

GAN 模型采用交叉熵作为模型损失函数，在训练过程中，当 G 生成的数据分布和真实数据分布存在较大差异时，会出现梯度消失问题，从而导致生成语音质量较差，为了解决这个问题，Mao 等[16]提出了 LSGAN，采用最小二乘损失函数来代替 Sigmoid 交叉熵损失函数，不仅提高了生成语音质量，而且使得模型收敛的更快，同时显著的增加了稳定性。为了使生成数据与真实数据更加接近，本文在生成网络的损失函数中增加 L_1 范数，L_1 范数作用是衡量生成数据和真实数据的差距[17]。LSGAN 的目标损失函数如下：

$$\min_D V_{\mathrm{LSGAN}}(D) = \frac{1}{2}E_{x,\tilde{x}\sim p_{data}(x,\tilde{x})}[(D(x,\tilde{x})-1)^2] + \frac{1}{2}E_{z\sim p_z(z),\tilde{x}\sim p_{data}(\tilde{x})}[D(G(z,\tilde{x}),\tilde{x})^2] \tag{2}$$

$$\min_G V_{\mathrm{LSGAN}}(G) = \frac{1}{2}E_{z\sim p_z(z),\tilde{x}\sim p_{data}(\tilde{x})}[(D(G(z,\tilde{x}),\tilde{x})-1)^2] + \lambda \| \boldsymbol{G}(z,\tilde{x}) - \boldsymbol{x} \| \tag{3}$$

式中：λ 为控制 L_1 正则化的超参数。其中，$L_1 = \lambda \| \boldsymbol{G}(z,\tilde{x}) - \boldsymbol{x} \|$

1.3 基于 ECA 的 LSGAN

近几年，注意力机制在深度卷积神经网络中应用频繁并具有巨大潜力[18]，目前的方法大多是集中于开发更加复杂的注意力模块，用来实现性能方面的提升，但是反而增加了模型的复杂度，为了解决性能和复杂度的矛盾问题，本文采用一种 ECA 模块[19]，该模块只增加了少量的参数，却能获得明显的性能增益。

ECA 模块是一种超轻量级的注意力模块，利用不降维局部交叉通道交互的策略，避免降维给通道注意预测带来副作用，而适当的跨通道交互会使保持性能的同时降低模型的复杂性。ECA 模块使用不降维的 GAP 聚合卷积特征后，首先自适应地确定核大小 k，然后进行一维卷积，最后使用 Sigmoid 函数来学习信道注意。ECA 原理如图 2 所示。

本文将这种方法利用到车载语音系统去噪中。在 LSGAN 架构上做了改进，G 的编码层中添加了 ECA 模块，ECA 自适应选择一维卷积核大小生成通道权重，在生成高质量语音的同时提高了训练效率。网络结构如图 3 所示。

G 网络和 D 网络都是由一维卷积层构成，其中 G 是最重要的一部分，由 22 层对称 U 型卷积和反卷积层构成[20]，激活函数使用 PReLU。其结构相当于一个自动编

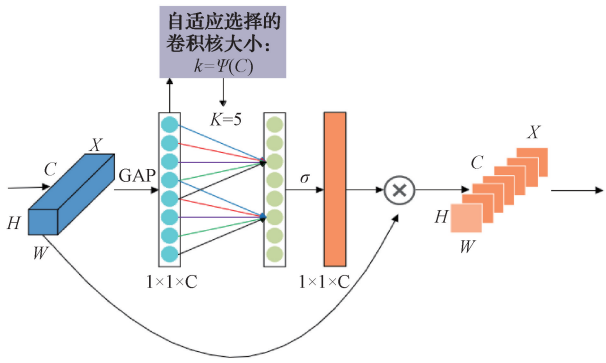

图 2　ECA 原理

码器，在编码阶段，首先输入语音信号被卷积层投影和压缩后，送入 ECA 模块，再通过线性整流层，然而解码阶段和编码阶段恰好相反，最后输出增强语音，并且编码层和解码层跳跃式连接，以便使输入和输出底层结构相同。判别网络相当于一个二分类器[21]，激活函数使用 Leaky ReLU，为了更好的训练模型，使用 VBN 稳定训练过程，最后的卷积层后使用全连接层输出判别结果。

2　实验及分析

本文是基于 ECA 的 LSGAN 模型对车载语音进行增强，实验平台为 Ubuntu18 64 位操作系统，GPU 使用 NVIDIA GTX 1080Ti，同时采用 Pytorch 深度学习框架进

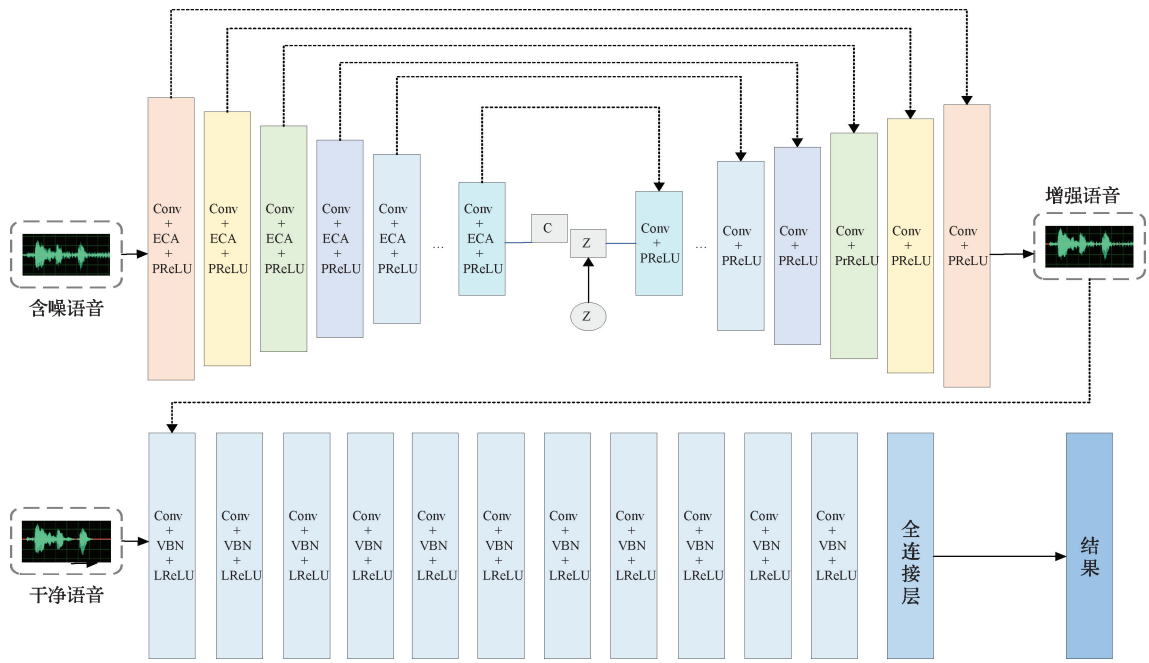

图 3　ECA-LSGAN 网络结构

行实验。并选取 3 种基线方法进行对比，方法为谱减法、维纳滤波语音、SEGAN。

2.1　数据集

纯净语音选取希尔贝壳 AISHELL-2 中文语音数据库中的 24 名说话者录音，说话者是具有不同口音的男性和女性，每人 104 条语音。语音内容为汽车行驶中的不同控制指令，如播放某首歌曲、设置空调具体温度、导航等，此数据集采样频率为 16 kHz。噪声来源于 Noise-92、tosound 数据库，在训练阶段共有 12 种类型噪声，分别为白噪声、汽车噪声等，通过语音合成为不同信噪比的含噪语音(−5、0、5、10、15 dB)，进行混合构成 24 000 条含噪语音进行训练。

基于汽车行驶中环境噪声复杂多变的特性，测试阶段使用汽车环境中收集的实际语音数据，设置 4 个实际场景进行真实模拟，噪声选取 tosound 数据库汽车行驶中车内录制的噪声，场景如下：1)汽车正常行驶过程中发动机高速运转的场景；2)汽车行驶中雨水拍打玻璃的场景；3)汽车行驶中播放广播的场景；4)汽车行驶中车内聊天的场景。其中每个场景内都包含车体本身噪声(如轮胎摩擦噪声、车体振动产生的噪声等)，纯净语音选取 AISHELL-2 中文语音数据库中的 1 名说话者录音，采样频率为 16 kHz。通过语音合成为不同信噪比的含噪语音(−10、−5、0、5、10 dB)，共 1 000 条含噪语音，评估不同信噪比(SNR)下模型的性能。

2.2　模型训练

模型使用 RMSprop 优化器进行训练，生成器和判别器的学习率都设置为 0.000 2，训练次数 epoch 设置为 60，批量处理大小 Batch Size 设置为 100。

2.3　评价指标

为了更准确评估方法的性能，分别采用了主观评价指

标和客观评价指标对结果进行评定，其中客观评价指标选取语音质量感知评估（perceptual evaluation of speech quality，PESQ）和短时客观可懂度（short-time objective intelligibility，STOI）。PESQ得分位于0～5，值越高代表语音质量越好；STOI得分位于0～1，值越大代表语音可懂度越高、也越清晰，同时STOI对于车载语音系统更为重要，STOI越高，车载语音系统更能准确识别语音信号。主观评价采用平均意见得分（mean opinion score，MOS），采取5分制对被测语音质量进行评价，分值越高代表试听者更加认可。

在不同场景和不同信噪比条件下，不同方法的客观评价结果如表1所示。

表1 PESQ与STOI评测结果

场景	方法	PESQ						STOI					
		−10 dB	−5 dB	0 dB	5 dB	10 dB	均值	−10 dB	−5 dB	0 dB	5 dB	10 dB	均值
Engine	Noisy	1.78	1.94	2.02	2.16	2.26	2.03	0.53	0.60	0.67	0.72	0.79	0.66
	谱减法	1.79	1.96	2.04	2.18	2.28	2.05	0.52	0.59	0.66	0.71	0.77	0.65
	Wiener	2.40	2.58	2.70	2.86	3.02	2.71	0.77	0.82	0.86	0.89	0.93	0.85
	SEGAN	2.07	2.16	2.34	2.50	2.63	2.34	0.72	0.72	0.76	0.81	0.84	0.77
	本文方法	2.50	2.68	2.82	2.94	3.01	2.79	0.85	0.89	0.90	0.91	0.93	0.90
Rain	Noisy	1.91	1.96	2.14	2.26	2.44	2.14	0.64	0.70	0.77	0.81	0.85	0.75
	谱减法	1.98	1.98	2.17	2.30	2.46	2.18	0.63	0.69	0.76	0.80	0.84	0.74
	Wiener	2.40	2.48	2.68	2.84	3.10	2.70	0.77	0.81	0.87	0.90	0.94	0.86
	SEGAN	2.31	0.41	2.63	2.81	3.02	2.24	0.74	0.78	0.83	0.85	0.88	0.82
	本文方法	2.52	2.66	2.76	2.96	3.18	2.82	0.85	0.88	0.91	0.92	0.94	0.90
Broadcast	Noisy	1.88	1.98	2.16	2.25	2.36	2.13	0.60	0.66	0.71	0.76	0.80	0.71
	谱减法	1.97	2.00	2.21	2.28	2.40	2.17	0.67	0.63	0.69	0.73	0.78	0.70
	Wiener	2.34	2.41	2.57	2.70	2.87	2.58	0.77	0.80	0.84	0.88	0.91	0.84
	SEGAN	1.99	2.28	2.43	2.57	2.80	2.41	0.71	0.76	0.80	0.84	0.86	0.80
	本文方法	2.31	2.49	2.61	2.72	2.86	2.60	0.76	0.82	0.84	0.87	0.90	0.84
Speeking	Noisy	1.77	1.89	2.02	2.23	2.38	2.06	0.55	0.60	0.67	0.74	0.79	0.66
	谱减法	1.79	1.95	2.08	2.25	2.40	2.09	0.53	0.59	0.65	0.72	0.77	0.65
	Wiener	2.28	2.43	2.59	2.72	2.90	2.58	0.74	0.79	0.84	0.87	0.92	0.83
	SEGAN	1.94	2.17	2.45	2.72	2.82	2.42	0.66	0.70	0.77	0.82	0.85	0.76
	本文方法	2.23	2.39	2.59	2.81	2.92	2.59	0.76	0.79	0.84	0.86	0.9	0.83

由表1可知，本文方法的PESQ和STOI两项指标的平均值都高于其他基线方法。在场景1（Engine）和场景2（Rain）中，相较于维纳滤波法，PESQ指标平均提升4.7%，STOI指标平均提升5.9%，而在场景3（Broadcast）和场景4（Speeking）中，本文方法的PESQ和STOI两项指标在部分信噪比下略低于维纳滤波方法，特别在场景4中，该类噪声对驾驶员语音控制声音形成干扰，在时域和频域上有一定程度的重叠，降低了模型的性能，低信噪比下最为严重，导致增强后的语音质量和可懂度受到影响。

为了更直观看出语音增强的效果，本文对比了场景2汽车正常驾驶，−5 dB信噪比下通过不同方法处理后的语谱图（Adobe Audition音频编码软件截图），如图4所示。

从图4可以看出，基线方法和本文方法都能有效的去除背景噪声，但是与本文方法相比，维纳滤波与谱减法增强后的语音仍残留较多的噪声，同时语音本质也是随机噪声，而SEGAN去除噪声太彻底，将一些有用语音信号丢失，因此本文方法更值得采用。

上述客观评价中维纳滤波是基线方法中效果最好的，所以主观评价主要测试维纳滤波与本文方法在不同场景下−5、0、5 dB信噪比的主观性能。本文邀请15位试听者进行主观评测，其中10位从事语音信号处理，5位未接触过语音处理领域，为了避免试听者的听觉疲劳，测试时间不宜过长。每个信噪比选取30个语音数据进行评测，随机的分给试听者，为每个试听者提供3种类型，分别为含噪语音、维纳滤波处理过的语音、本文方法处理过的语音。评估期间，试听者在相同环境下对被测语音进行主观打分，最后取平均值。表2为主观评测结果。

由表2可以看出，这两种方法在语音增强方面都能有良好的效果，而这两种方法相比较，本文方法的语音质量评价略高于维纳滤波，受到试听者的认可，同时主观评价与客观评价相一致，因此可以应用到实际环境中。

3 结 论

本文提出了一种基于ECA的LSGAN在车载语音增强的方法，针对车载环境中噪声复杂、多变的问题，利用

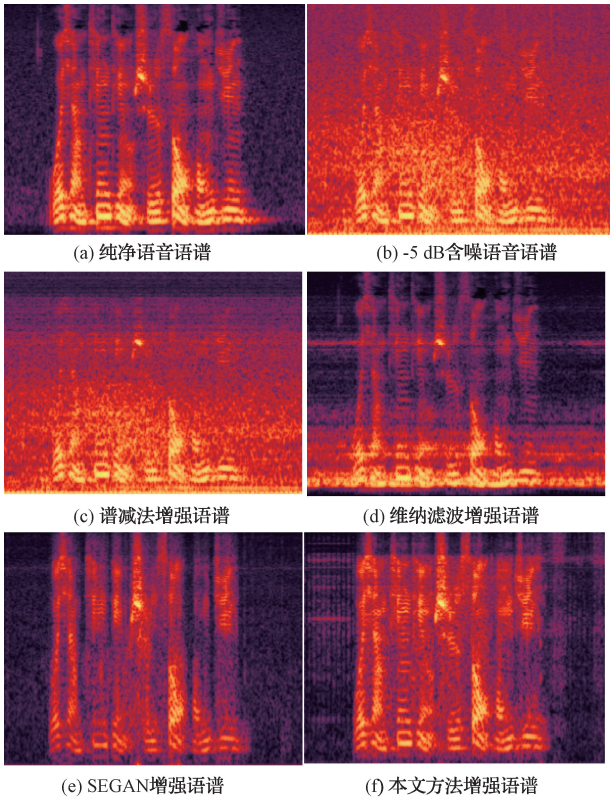

图 4　信号语谱

表 2　主观评价

度量标准	含噪语音	维纳滤波	本文方法
MOS	2.38	3.73	3.79

ECA 自适应选择一维卷积核大小生成通道权重的特点，与 LSGAN 结合进行训练，最终生成高质量语音，不仅在性能上有明显提升同时大大提高了训练效率。讨论了 ECA-LSGAN 在不同场景不同信噪比下的性能，实验结果表明，对于非平稳场景噪声（类似于雨噪声、发动机噪声）相比于其他方法，本文方法去噪效果更好，主观质量和客观质量都有了明显提升。但是类似于广播、谈话等噪声场景下效果不理想，原因是此类噪声属于语音干扰噪声，与驾驶员语音控制声音在时域和频域上有很大程度的重叠，加大了去噪难度，影响模型的性能，下一步将继续对车载语音中这类重叠噪声进行研究。

参 考 文 献

[1] MA J, DING Y N. The impact of in-vehicle voice interaction system on driving safety[J]. Journal of Physics: Conference Series, 2021, 1802(4): 042083.

[2] 景亚鹏，苏海涛，王绍，等. 汽车内驾驶员语音增强评价研究[J]. 声学技术，2021，40(6)：832-838.

[3] NA S, LI W X, LIU Y. An improved spectral subtraction algorithm for speech enhancement system[C]. Proceedings of 2016 6th International Conference on Information Engineering for Mechanics and Materials(ICIMM 2016), 2016: 329-334.

[4] ZHAO Y L, OU S F, GAO Y. Improved wiener filter algorithm for speech enhancement [J]. Automation, Control and Intelligent Systems, 2019, 7(3): 92-98.

[5] SUN C L, MU J S. An eigenvalue filtering based subspace approach for speech enhancement[J]. Noise Control Engineering Journal, 2015, 63(1): 36-48.

[6] PANDEY A, WANG D L. A new framework for CNN-based speech enhancement in the time domain[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(7): 1179-1188.

[7] CHENG J M, LIANG R Y, ZHAO L. DNN-based speech enhancement with self-attention on feature dimension[J]. Multimedia Tools and Applications, 2020, 79(43-44): 32449-32470.

[8] SALEEM N, KHATTAK M I, et al. Multi-objective long-short term memory recurrent neural networks for speech enhancement [J]. Journal of Ambient Intelligence and Humanized Computing, 2021, 12(10): 9037-9052.

[9] PASCUAL S, BONAFONTE A, SERRA J. SEGAN: Speech enhancement generative adversarial network[J]. Neural Information Processing Systems (NI-PS), 2017, 3(12): 1703-1710.

[10] YANG F, WANG Z T, LI J F, et al. Improving generative adversarial networks for speech enhancement through regularization of latent representations[J]. Speech Communication, 2020, 118: 1-9.

[11] JING Y P, SU H T, WANG S, et al. Research on speech enhancement evaluation of driver in vehicle[J]. Technical Acoustics, 2021, 40(6): 832-838.

[12] 陈亮，吴攀，刘韵婷，等. 生成对抗网络 GAN 的发展与最新应用[J]. 电子测量与仪器学报，2020，34(6)：70-78.

[13] 杨鸿杰，陈丽，张君毅. 基于生成对抗网络的数字信号生成技术研究[J]. 电子测量技术，2020，43(20)：127-132.

[14] 吴培良，刘瑞军，李瑶，等. 一种基于生成对抗网络与模型泛化的机器人推抓技能学习方法[J]. 仪器仪表学报，2022，43(5)：244-253.

[15] 王延年，李文婷，任劼. 基于生成对抗网络的单帧图像超分辨算法[J]. 国外电子测量技术，2020，39(1)：26-32.

[16] 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.

[17] 张睿琦,李迟生,陈颖.一种改进生成对抗网络的短波信号增强方法[J].现代电子技术,2021,44(17):56-60.

[18] ZHAO Z P, BAO Z T, ZHANG Z X, et al. Self-attention transfer networks for speech emotion recognition[J]. Virtual Reality & Intelligent Hardware, 2021, 3(1): 43-54.

[19] WANG Q L, WU B G, ZHU P F, et al. ECA-Net: Effificient channel attention for deep convolutional neural networks[C]. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11534-11542.

[20] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2015: 234-241.

[21] CAO Z H, HUANG Z H, GE W P, et al. Influence of attention mechanism on generative adversarial network speech enhancement transfer learning model[J]. Technical Acoustics, 2021, 40(1): 77-81.

作 者 简 介

石瑞,硕士,主要研究方向为语音信号处理。
E-mail:784962012@qq.com

杨立东,博士,教授,硕士生导师,主要研究方向为音频信号处理与模式识别。
E-mail:yld_nkd@imust.edu.cn