第 35 卷 第 12 期 信 号 处 理 Vol.35 No.12
2019 年 12 月 Journal of Signal Processing Dec. 2019
文章编号: 1003-0530(2019)12-1931-11

基于深度神经网络的单通道语音增强方法回顾

鲍长春 项 扬
(北京工业大学信息学部语音与音频信号处理研究室,北京 100124)

摘 要: 语音增强是一种试图从噪声中分离出语音的技术,目的是提高语音的质量和可懂度。在过去的几十年里,人们提出了多种类型的语音增强方法,但这些方法在非平稳噪声环境中的表现还未达到最佳程度,因为他们没有充分利用语音和噪声的先验信息。近年来,随着深度学习的发展,深度神经网络已成为当下实现语音增强的主流方法,在改善语音质量和提升可懂度方面发挥了积极作用。本文从深度神经网络的结构出发,回顾了基于深度学习的单通道语音增强方法。首先,介绍了语音增强的背景;其次,详细描述了四种不同类型神经网络实现语音增强的方法;最后,给出了未来语音增强方法的建议和本文的结论。

关键词: 语音增强;深度神经网络;深度学习;先验信息;监督学习
中图分类号: TN912.35 **文献标识码:** A **DOI:** 10.16798/j.issn.1003-0530.2019.12.001

引用格式: 鲍长春,项扬. 基于深度神经网络的单通道语音增强方法回顾[J]. 信号处理,2019,35(12):1931-1941. DOI: 10.16798/j.issn.1003-0530.2019.12.001.
Reference format: Bao Changchun, Xiang Yang. Review of Monaural Speech Enhancement Based on Deep Neural Networks [J]. Journal of Signal Processing, 2019, 35(12): 1931-1941. DOI: 10.16798/j.issn.1003-0530.2019.12.001.

Review of Monaural Speech Enhancement Based on Deep Neural Networks

Bao Changchun Xiang Yang
(Speech and Audio Signal Processing Lab, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: Speech enhancement tries to separate speech from noise and aims to improve the quality and intelligibility of speech. In the past several decades, many types of speech enhancement methods have been proposed. However, these methods cannot always achieve the best performance for non-stationary noise because they do not make best use of prior information of speech and noise. In recent years, with the advance of deep learning, the deep neural network (DNN) has become a mainstream strategy to conduct speech enhancement, and is playing important role in improving speech quality and increasing intelligibility. Based on the structure the DNN, the DNN-based monaural speech enhancement methods are reviewed in this paper. First, the background of speech enhancement is introduced. Next, four different types of the DNN used for conducting speech enhancement are carefully described. Finally, some comments of speech enhancement for future work and conclusions are given.

Key words: speech enhancement; deep neural network; deep learning; prior information; supervised learning

1 引言

在移动电话和微信等日常语音通信中,环境噪声和其他干扰不可避免地影响了通话质量。如何有效地消除环境噪声和干扰一直是语音信号处理领域的一项挑战性课题。语音增强的目的就是消除噪声和干扰,最大可能地提高语音听觉质量和可懂度。目前很多的语音增强算法仅仅改善了语音质量,在可懂度方面存在很大不足,即在低信噪比条件下,噪声得到了减少,但引入了较大的语音失

收稿日期: 2019-10-11; 修回日期: 2019-12-02
基金项目: 国家自然科学基金资助项目(61831019, 61471014)

真。因此,语音增强的主要挑战是如何设计一个高效的算法,在不引入语音失真的前提下,有效抑制噪声和干扰。

语音增强的具体解决方案与应用场景、噪声或者干扰信号的特性、噪声与语音的关系、麦克风的数量密切相关。干扰信号可能是类似噪声的信号(如风扇噪声),也有可能是类似语音的信号(如多人说话时其他说话人的声音);噪声对于信号而言可能是加性的,也有可能是乘性的(如房间混响情况)。再者,噪声与纯净的语音信号之间可能是统计相关或者无关的;麦克风可以使用单个,也可以使用多个。麦克风阵列可有效排除噪声干扰,达到减少语音失真的目的,但其缺点是需要准确的声源定位信息。

由上可以看出,语音增强是一项较为复杂和繁琐的科研课题,要同时照顾到主、客观的听觉感受,这不仅涉及到信号处理方面的知识,同时还要基于人耳听觉感知和语音学的相关原理。

由于语音增强在实际应用中的重要性,自上世纪 50 年代以来,语音增强已经吸引了国内外众多学者的关注[1-2]。语音增强大体上可以分为无监督学习和有监督学习两类。无监督学习方法不依赖于先验语音信息,所需计算量较少,对硬件要求低。因此,这类方法在实际应用时相对容易实现。但这类方法往往基于很多假设,而这些假设在实际中经常是不准确或是错误的,因此,其实际性能受到了限制。近年来,随着计算机处理能力的提升,基于深度学习的有监督学习方法已经成为当今语音增强的主流方法。这类方法在实现语音增强时,一般分为线下和线上两个阶段。在线下,运用有监督学习的训练方式,得到观测信号和纯净语音特征的映射关系。之后,在线上运用该映射关系进行语音增强。尽管这类方法在线下训练时,需要大量数据,花费很长的时间,且对计算机硬件的要求也很高,但是与传统方法相比,它具有一定的噪声泛化能力,因此,能够获得更好的语音质量,且更容易能满足实际场景的需要。

由于单麦克风语音增强不需要声源定位,同时对混响不敏感,因此,以谱减法、统计模型法和子空间法为代表的无监督学习方法,在过去的几十年里得到了深入研究,并衍生出众多的改良方法,在消除噪声方面发挥了重要作用。谱减法[3-4]假设噪声是加性的,它能从观测信号谱中减去估计的噪声谱,得到重构的语音谱。统计模型法[5-8]则是将语音增强作为统计估计框架中的一个问题提出,即在给定观测信号的测量参数(如傅里叶变换系数)后,希望对目标参数(语音信号的傅里叶变换系数)进行线性或非线性估计。子空间算法[9-11]则以线性代数理论为基础,即将观测信号的向量空间分解为两个子空间,其中一个子空间主要包括纯净信号,另一个子空间主要包括噪声信号,这样就可以简单的通过清除观测信号向量空间中“噪声子空间”的部分内容,达到估计纯净信号的目的。

上述经典的无监督学习的语音增强方法能在平稳噪声条件下取得令人满意的增强效果,但对非平稳噪声而言,这类方法的性能不佳,原因是他们不能适应噪声随时随地的变化。为此,基于深度神经网络[12]的有监督学习的语音增强算法应运而生,这类方法充分利用了语音的先验信息,克服了无监督学习方法所存在的很多假设和噪声估计不准确问题,大大提升了增强性能。下面我们将介绍当前热衷研究的四类基于神经网络的语音增强方法,他们依次是基于多层感知器(multilayer perceptions, MLPs)的语音增强方法、基于卷积神经网络(convolutional neural networks, CNNs)的语音增强方法、基于循环神经网络(recurrent neural networks, RNNs)的语音增强方法和基于生成对抗神经网络(generative adversarial networks, GANs)的语音增强方法。

2 基于 MLP 的语音增强方法

通过一系列组合仿射运算和非线性的层,MLP 能有效解决预测和分类的监督学习问题,其输入层与输出层之间采用全连接方式,输入层经过隐含层到输出层的计算是以反馈的方式进行的,即使用反向传播(Backpropagation algorithm, BP)算法[13]训练网络。图 1 给出了 MLP 的结构示意图。通常,一个 MLP 是由 $L+1$ 层神经元组成。其中,第一层为输入层,最后一层为输出层,其他层均为隐含层。图 1 中的输入向量为 $\boldsymbol{X}=(x_1,x_2,\cdots,x_m)$,输出向量为 $\hat{\boldsymbol{Y}}=(y_1,y_2,\cdots,y_s)$,第 l 个隐含层的输出向量 $\boldsymbol{H}^{(l)}=(h_1^{(l)},h_2^{(l)},\cdots,h_{s_l}^{(l)})$。通过使得输出矢量 $\hat{\boldsymbol{Y}}$ 和目标矢量 $\boldsymbol{Y}$ 之间的损失函数 $L(\hat{\boldsymbol{Y}},\boldsymbol{Y})$ 最小,可最终得到网络的偏差和权重参数。

基于 MLP 网络结构,2013 年 Bingyin Xia 和 Changchun Bao 提出了加权去噪自动编码器(Weighted Denoising Auto-encoder, WDA)[14-15]模

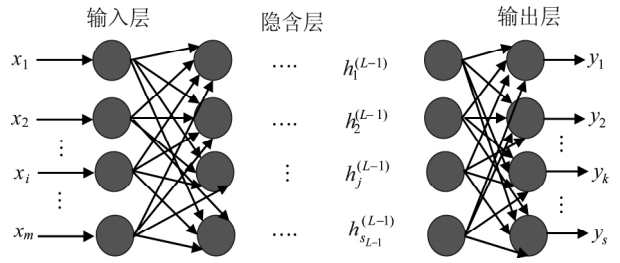

图 1　MLP 的结构示意图

Fig.1　Sketch of the MLP structure

型,用于描述纯净语音和含噪语音功率谱的仿射关系,较好地估计了纯净语音的功率谱,提高的语音增强的性能。

在 2014 年, Yong Xu 等人提出将含噪语音的对数功率(log-power spectral, LPS) 作为 MLP 的输入,利用 MLP 非线性函数匹配的能力,在 MLP 的输出端预测得到纯净语音的 LPS [16] 。图 2 给出了该方法利用深度神经网络实现语音增强的原理框图。它实现语音增强分为训练和增强两个阶段,在训练阶段,需要提取含噪语音和语音的特征对 MLP 进行训练; 在增强阶段,首先对带噪语音 $y(n)$ 进行特征提取,得到带噪语音的幅度谱 $Y_l(k,m)$ 及其相位角 θ,然后利用之前已经训练好的 DNN,将带噪语音的特征 $Y_l(k,m)$ 作为输入,即可获得所需的训练目标 $\hat{X}_l(k,m)$ 。最后通波形重构,就可获得增强语音 $\hat{x}(n)$ 。图 2 下面的虚线框内给出了语音特征提取及波形重构的原理。值得注意的是,该方法在进行语音增强时,没有对语音和噪声之间的关系做任何假设,因此该方法能够适用于广泛的噪声环境中且能取得更好的增强表现。为了能够提升该算法的性能, Yong Xu 等人在 2015 年 [17] 提出了三种策略对原有网络进行了改进。这三种策略分别是: (1) 解决神经网络过匹配的全局方差(global variance,GV) 均衡策略; (2) 提升神经网络泛化能力的退出(Dropout) 策略 [18] ; (3) 提高网络泛化能力的噪声意识训练(noise aware training,NAT) [19] 。从文献 [17] 给出的针对未知的非平稳噪声的 PESQ(perceptual evaluation of speech quality) 测试结果看,对于展览馆内噪声、驱逐舰发动机噪声和高频信道噪声,在-5 dB,0 dB,5 dB,10 dB,15 dB 和 20 dB 信噪比条件下,组合三种策略的 PESQ 都远远高于基线 DNN 和传统的对数最小均方误差方法。对于展览馆内噪声,基线 DNN 的平均 PESQ 为 2.54, 对数最小均方误差方法的平均 PESQ 为 2.30,组合三种策略的平均 PESQ 为 2.76; 对于驱逐舰发动机噪声,基线 DNN 的平均 PESQ 为 2.80, 对数最小均方误差方法的平均 PESQ 为 2.62,组合三种策略的平均 PESQ 为 3.03; 对于高频信道噪声,基线 DNN 的平均 PESQ 为 2.45, 对数最小均方误差方法的平均 PESQ 为 2.45,组合三种策略的平均 PESQ 为 2.69。

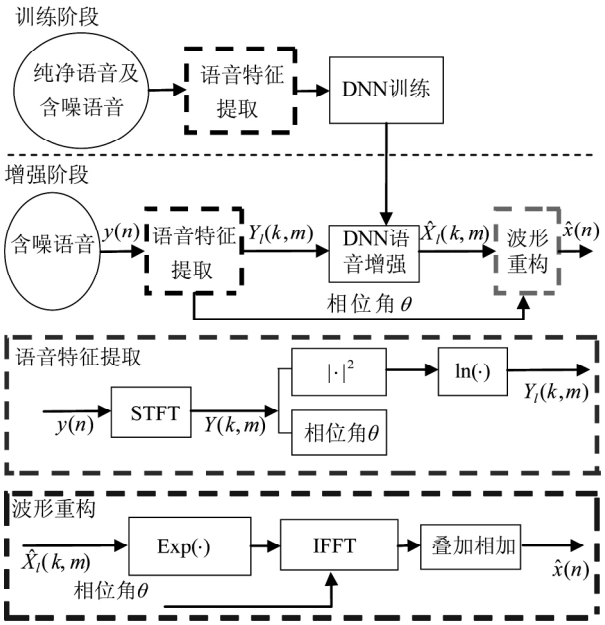

图 2　深度神经网络实现语音增强的原理框图 [16]

Fig.2　Block diagram of the DNN-based speech enhancement

值得注意的是噪声意识训练作为一种简单有效的提升神经网络性能的方法,吸引了大量学者对其进行研究,期望获得更好的增强语音。在文献 [17] 中,运用的是静态的噪声意识训练方法,该方法不能很好地追踪噪声的变化,为此 Yong Xu 等人提出了一种动态的噪声意识训练方法 [20] 。该方法能对噪声进行更加准确的估计,因此其也能获得质量更好的增强语音。另外, 在 2016 年微软的 Dinei Florencio 等人也提出了一种运用动态噪声意识训练方法的深度神经网络来进行语音增强 [21] 。值得注意的是,在该篇论文中 [21] ,作者还将心理声学模型 [22] 融入到了神经网络的训练中,并将该模型作为训练网络过程中的误差函数的一部分。因此,通过该神经网络进行增强的语音能够更好的符合人耳心理听觉特性。此外,该网络也有较好的泛化性能,能够在多种混合噪声环境中实现语音增强。除了运用噪声意识训练来提升 MLP 的性能外,多目标学习也能有效地提高 MLP 的性能。 Yong Xu 在

2015 年提出了运用一个 MLP 同时学习多个纯净语音特征的方法[23]，提高了语音增强性能。该项工作使用了 MLP 去学习新的用于语音增强的特征，不仅提升了原来所学习特征的表现，还将新学习的特征用在了语音增强的后处理中。通过实验，作者发现将噪声的 LPS 作为网络输入，去同时匹配纯净语音的 LPS、梅尔倒谱系数以及理想二值掩膜(ideal binary mask，IBM)能实现更好的增强效果。另外，Qing Wang 等人在 2017 年也提出了一种基于多目标学习的方法[24]，该方法将纯净语音的 LPS 及理想比值掩膜(ideal ratio mask，IRM)作为了学习目标，并利用语音的子带特征来进行噪声感知训练，提高了噪声意识训练的有效性，语音质量及可懂度都有了较大的提升。在利用 MLP 实现语音增强的研究中，D.L Wang 的研究也是十分值得关注的。D.L Wang 的研究主要是利用语音的频域特征或根据计算听觉场景分析(computational auditory scene analysis，CASA)所得到的特征作为 MLP 的输入，去预测 IRM(ideal ratio mask)或者 IBM[25-26]。这种方法虽然假定了语音和噪声是相互独立的，但通过实验结果发现，这类算法的性能并不差于利用 MLP 去直接估计 LPS 的算法。此外，利用 MLP 计算一些具有相位特性的 IRM[27-28]，可以进一步提升算法性能。除了上述方法，一些与传统方法结合的神经网络语音增强方法[29-30]，也实现了较好的语音增强效果。

3 基于 CNN 的语音增强方法

CNN[31] 也是一种能够有效实现语音增强的神经网络。近几年来，CNN 在语音处理的相关领域得到广泛应用。

与 MLP 相比，CNN 能够更加准确地获取输入语音信号的局部特征，因此，将其运用在语音增强中时，它能够更好地恢复出语音信号的高频成分，提高增强语音的质量及可懂度。在 2017 年，Tomas 等人将 CNN 有效地应用到了语音增强任务中[32]。在该项工作中，作者用一维 CNN 实现了语音增强，一维 CNN 结构如图 3 所示。该网络从输入层到输出层的连接依次是：卷积层，最大值池化层，卷积层及两个全连接层。作者将含噪语音的 LPS 作为网络输入，分别预测 IRM 及纯净语音的 LPS。其与 MLP 类似，利用了随机梯度下降法对 CNN 进行训练。在输入信噪比分别为 -3 dB，0 dB，2 dB，5 dB，7 dB，10 dB条件下，通过与基于 MLP 的掩蔽增强的方法比较，文中提出的基于映射的卷积和全连接的降噪自动编码器[32] 明显提高了 PESQ 分，如在 -3 dB 时，基于 MLP 的掩蔽增强 DNN 的 PESQ 为 2.15，基于 CNN 的掩蔽增强的 PESQ 为 2.19，基于 MLP 的映射增强 DNN 的 PESQ 为 2.22，基于 CNN 的映射增强的 PESQ 为 2.36，这一方面证明了具有拓扑结构的 CNN 能比 MLP 结构的 DNN 更有效提高语音质量，另一方面也证明了映射式增强要比掩蔽式增强效果好。但在该 CNN 的结构中，由于仍然含有全连接层，因此网络所含参数较多，不太适应于移动设备。为此，高通的 Park 和卡内基梅隆大学的 Jinwon 给出了一套解决方案[33]，即运用全卷积层的卷积神经网络减少 CNN 对硬件内存占用。在该 CNN 结构中，全部连接层均是卷积层，无任何全连接层。因此该网络所含参数的数目可以达到原 MLP 的千分之一。此外，在该网络中，卷积层之间还加入了残差连接[34]，这大大提高了该网络的性能。值得注意的是，虽然该网络的参数数目减少了，但其性能并没有比原先 MLP 的差。受到这项工作的启发，Fu，Szu-Wei 等人提出了一种在给定含噪语音时域波形的条件下，运用全卷积神经网络直接匹配得到纯净语音时域波形的语音增强方法[35]。直接匹配纯净语音时域波形的增强方法有效避免了在以往方法中要进行相位估计的问题，这为今后语音增强的研

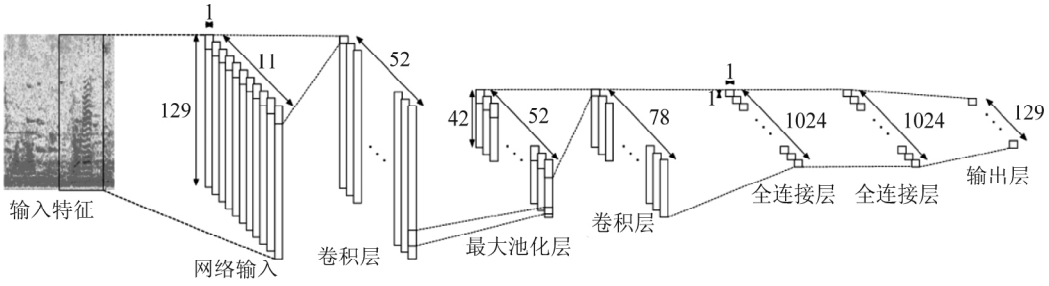

图 3 实现语音增强的一维 CNN 结构图[32]

Fig.3 Speech enhancement using one-dimension CNN

究提供了新思路。此外,Emad 提出了基于卷积神经网络,并运用降噪自编码器的单声道声源分离算法[36]。与上述运用 CNN 进行语音增强方法不同的是,该算法用的是二维卷积神经网络而非一维的,这种方法在声源分离任务上也取得了出色的表现。此外,Yang and Bao 提出了一种基于堆叠式的卷积自编码器语音增强方法[64],将语音对数谱作为网络的输入,自回归参数作为训练目标,并结合码书谐波恢复算法实现了语音增强。该方法对带噪语音的谐波恢复取得了较好的效果。另外,Cui 和 Bao 提出了一种基于 MLP 的自编码器[65],将带噪语音对数谱作为输入特征,预测了一系列自回归参数,并结合维纳滤波器实现了语音增强,其也取得了较好的语音谐波恢复效果。

4　基于 RNN 的语音增强方法

在以往利用 MLP 或者 CNN 进行语音增强时,很难将某一帧语音和其周围帧的关系都考虑到,这限制了 MLP 和 CNN 在进行语音增强时的性能。然而 RNN[37] 的出现或许可以缓解该问题。与之前的 MLP 和 CNN 相比,RNN 能够利用之前的信息对当前内容进行预测和判断。因此,其更适用于处理和时间相关的信息。图 4 给出了一个 RNN 的简单示意图。由该图可以发现,RNN 在预测当前时刻信息时,不仅考虑了当前时刻的输入,同时也用到了在它之前时刻所输出的信息。在图 4 中,$X_i(i=0,1,\cdots,t)$ 代表网络的输入,A 代表 RNN 网络,$h_i(i=0,1,\cdots,t)$ 代表网络的输出,i 代表其时间索引。从图 4 可以看出,前一时刻的输出将作为后一时刻网络的输入,因此,它用到了之前时刻的信息。此外,它训练 RNN 所用的方法是时序反向传播算法(Back Propagation Through Time, BPTT)。Andrew 等人提出运用循环降噪自编码器[38]来对含噪语音进行降噪从而提高自动语音识别系统的表现。在这项工作中,作者将含噪语音特征作为降噪自编码器网络的输入,直接预测得到纯净语音的梅尔倒谱系数。通过实验,作者指出这种方法能有效的提高语音识别的准确率。Po-Sen 等人[39]提出使用 RNN 进行唱歌信号的语音分离方法,其目的是从当前唱歌信号中分别分离出人声信号及纯净音乐信号。在该项工作中,作者首先使用 RNN 同时预测人声信号和纯净音乐信号的幅度谱,然后用该幅度谱去估计人声和音乐信号的掩膜值,最后结合观测信号的相位信息得到所期望的信号。与以往用 MLP 进行分离的结果相比,用 RNN 所分离的信号有更高的质量。基于该项研究,Po-Sen 等人又提出使用 RNN 进行三种不同的音源分离任务(语音分离,唱歌信号分离,语音增强)[40]。在这项研究工作中,作者针对三种不同类型的音源分离任务,提出了与其任务相对应的最合适的 RNN 结构,以及误差准则。图 5 展示了在该项研究中用到的不同的 RNN 的结构。图 5 中左侧的图展示了最简单的 RNN 结构,其包含一个输入层,一个 RNN 层以及一个输出层。图 5 中间的图是对左侧图的改进,其在输入层与 RNN 层以及 RNN 层与输出层之间加入了多个全连接层。而在图 5 的右侧图中,在输入层和输出层之间使用了多个 RNN 层且无全连接层。

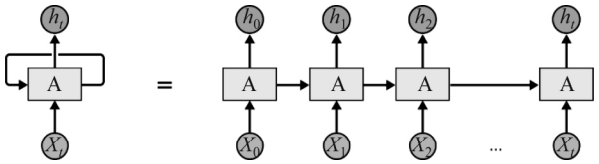

图 4　RNN 的简明示意图[37]

Fig.4　Brief sketch of the RNN

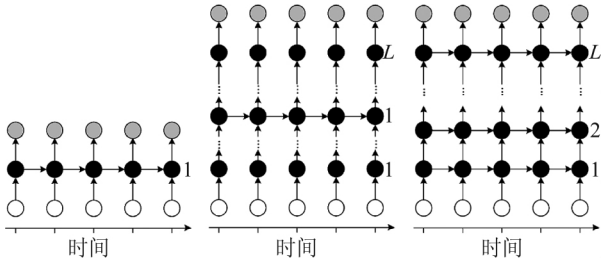

图 5　用于不同类型音源分离任务的 RNN 结构[40]

Fig.5　The structures of the RNN for different source separation task

虽然 RNN 在音源分离上有着出色的表现,但是它存在着梯度消失或者梯度爆炸的问题,这严重影响了 RNN 的性能。为此,长短时记忆(LSTM, long short-term memory)循环神经网络[45]被用于了音源分离任务中[41-42,46]。与 RNN 相比,在 LSTM 中引入了一个记忆细胞模块,该模块能够有效地解决之前 RNN 所存在的问题,并有更加广泛的应用。图 6 展示了一个 LSTM 的结构框图。从该图中我们可以发现,对于一个 LSTM 结构,其包含一个记忆细胞和三个门限。其中,忘记门限负责控制从胞腔中应该丢弃多少之前的信息,而输入门限负责控制有多少新的信息应该被加入到胞腔中,输出门限负责输出当前的信息。图 6 中符号 C_t 表示记忆细胞,f_t 表示忘

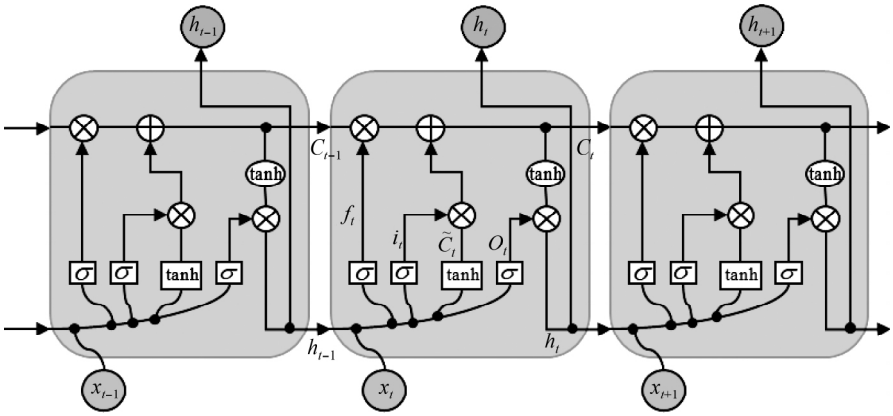

图 6 一个 LSTM 的示意图[45]

Fig.6 Diagram of an LSTM block

记门限，i_t表示输入门限，o_t表示输出门限，$\tilde{C}_t$表示新的待选值，$\otimes$ 表示矢量内积，$\oplus$ 表示矢量相加。

对于一个 LSTM 结构，定义其在第 t 时刻的输入和输出分别为x_t和h_t，则其在 t 时刻的忘记门限为:

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \tag{1}$$

忘记门限决定了 LSTM 要丢弃的信息。接下来 LSTM 要决定保存在胞腔中的新信息，即要确定如下的输入门限值:

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i) \tag{2}$$

接着生成一个如下的待选胞腔:

$$\tilde{C}_t=\tanh(W_C\cdot[h_{t-1},x_t]+b_C) \tag{3}$$

然后如下更新原有胞腔:

$$C_t=f_t*C_{t-1}+i_t*\tilde{C}_t \tag{4}$$

最终的输出为:

$$O_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o) \tag{5}$$

$$h_t=O_t*\tanh(C_t) \tag{6}$$

在式(1)到式(6)中的 W_x，b_x分别代表相应模块的权值和偏置，其激活函数 σ 和 tanh 可以写做:

$$\sigma(s)=\frac{1}{1+e^{-s}} \tag{7}$$

$$\tanh(s)=\frac{e^{s}-e^{-s}}{e^{s}+e^{-s}} \tag{8}$$

从式(1)到式(6)可以看出，输入门限i_t与忘记门限f_t依赖于当前输入 x_t和之前的输出 h_{t-1}，这让记忆细胞对于背景信息变得更敏感，从而去进行自身的更新，使其能够更好地去处理复杂的时序信息。由于 LSTM 能够充分利用时序信息，因此，与之前的 MLP 和 CNN 相比，将其应用于语音增强时，可以更好地获得噪声、带噪语音以及纯净语音内部间的时序相关信息，提高增强语音的质量。

一个成功的例子是 Lei 等人运用 LSTM-RNN 实现的语音增强。在该项工作中，他们用 LSTM-RNN 分别预测纯净语音的对数谱和 IRM 来实现语音增强[43]。此外，他们也进行了多目标学习，用一个 LSTM-RNN 同时预测了纯净语音对数谱和 IRM。他们的实验结果表明，和用基于 MLP 的 DNN 比，LSTM-RNN 能有效提升语音质量和可懂度。各种信噪比下的平均 PESQ 分从 DNN 的 2.339 提高到了 LSTM-RNN 的 2.663，可懂度从 DNN 的 80.2% 提高到了 84.8%。此外，文献[66]和[67]也验证了 LSTM 在多说话者环境中也能对语音实现有效的增强，提高其可懂度。

尽管 MLP，CNN，RNN，LSTM 在实现语音增强上有各自的优势，但也有各自的不足。因此，将这三种网络结构结合到一起[44]，或许可以实现最佳的语音增强效果。

5 基于 GAN 的语音增强方法

GAN[47]是 2014 年提出的一种强大的神经网络结构。GAN 可以在给定随机噪声输入的情况下，生成出与真实信号相一致的东西(如生成真实的图片或者语音)。换句话说，GAN 能够从一个随机分布的样本空间 z 中，生成真实的样本数据 x。对于一个 GAN，可以将其分为两个部分，一个生成器(G)，和一个判别器(D)。生成器的目的是利用一些随机样本去生成模仿真实分布的数据，进而去欺骗判别器。而判别器的任务是，去区分给它的数据是由生成器生成的还是确实就是真实的数据。因此，在生

成器和判别器之间就形成了一个对抗学习的过程。该过程可由如下的目标函数表示[47]:

$$\min_G \max_D V(D,G) = \mathrm{E}_{x\sim p_{\mathrm{data}}(x)}[\log D(x)] + \mathrm{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (9)$$

其中 x 是在分布 P_{date} 下的真实样本，z 是在分布 P_z 下的随机噪声输入。在式(9)中，$D(x)$ 代表了数据 x 来自于数据样本的概率。通过训练 D 来最大化其分配正确标签给训练样本和生成样本的概率。与此同时，通过训练 G 来最小化 $\log(1-D(G(z)))$。

由于由式(9)所生成的真实数据样本存在一定的不确定性，一种叫做条件生成对抗网络(conditional generative adversarial network，cGAN)的结构被提了出来[48]。

cGAN 能在给定一些特定的样本输入的情况下，获得更多的匹配信息，从而得到所期望的输出。相对于之前的 GAN，在 cGAN 中，它的生成器和判别器的输入多加入了一个观测得到的矢量 y。因此，生成器能够在 y 的控制之下合成所期望的数据。此外，判决器能够学会去区分一对真实的数据(x，y)或者是假的(生成的)数据($G(y,z)$，y)。其训练的目标函数可以表示为:

$$\min_G \max_D V_{\mathrm{cGANs}}(G,D) = \mathrm{E}_{x,y\sim p_{\mathrm{data}}(x,y)}[\log D(x,y)] + \mathrm{E}_{z\sim p_z(z),y\sim p_{\mathrm{data}}(y)}[\log(1-D(G(y,z),y))] \quad (10)$$

此外，对于生成器 G，还可以用 L1 重构误差对其进行优化[48]，因此用于训练 G 的误差函数可以表示为:

$$\min_G V_{\mathrm{cGANs}}(G) = \mathrm{E}_{z\sim p_z(z),y\sim p_{\mathrm{data}}(y)}[\log(1-D(G(y,z),y))] + 1\cdot\mathrm{E}_{x,y\sim p_{\mathrm{data}}(x,y),z\sim p_z(z)}[\|x-G(y,z)\|_1] \quad (11)$$

虽然 GAN 在图像上已经取得了广泛的应用，且都取得了不错的效果，但是 GAN 在语音上的运用却相对较少。直到 2017 年，Pascual 才将 GAN 运用到了语音增强上[49]。在文献[49]中，作者将含噪语音的时域波形作为 G 的输入，期望 G 能通过训练生成出纯净语音的时域波形，进而达到语音增强的目的。在该项工作中，G 和 D 的网络结构都是一维的 CNN，且加入了残差连接[34]。整个 GAN 的训练误差函数除了原本 GAN 自身的误差函数[47]，还加入了与文献[56]类似的 L1 误差[50]，以提高整个网络的性能。虽然该篇文章最后没有说明 GAN 在语音增强上与以往神经网络相比，谁的性能更好，但这也为今后语音增强的研究提供了新的方向。图 7 和图 8 分别展示了该网络的训练过程及网络结构。图 7 的左侧图展示了将真实特征 x 作为 D 的输入，并结合反向传播算法对其进行训练；图 7 中间的图展示了在将特征 z 作为 G 的输入的情况下，让 G 输出假特征去作为 D 的输入，在这种情况下，利用反向传播对 D 进行训练。图 7 的右侧图与中间的图类似，只不过此时不对 D 进行权值更新，只将 D 的输出作为反馈，利用反向传播算法对 G 进行训练。图 8 展示了实现语音增强 GAN 的网络结构图。由该图我们可以发现，其网络结构就是一个普通的卷积神经网络，并且使用了残差连接。其中 K 和 L 分别表示每个卷积层输出维度的大小及其所用滤波器的维度。值得注意的是，在编码端得到输出 c 之后，其将随机变量 z 加入其中，让 c 和 z 共同作为了解码端的输入，最终在解码端获得了增强语音。

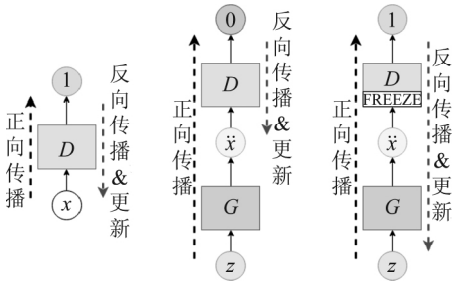

图 7　用于实现语音增强 GAN 的训练过程图[49]

Fig.7　Diagram of GAN's training for speech enhancement

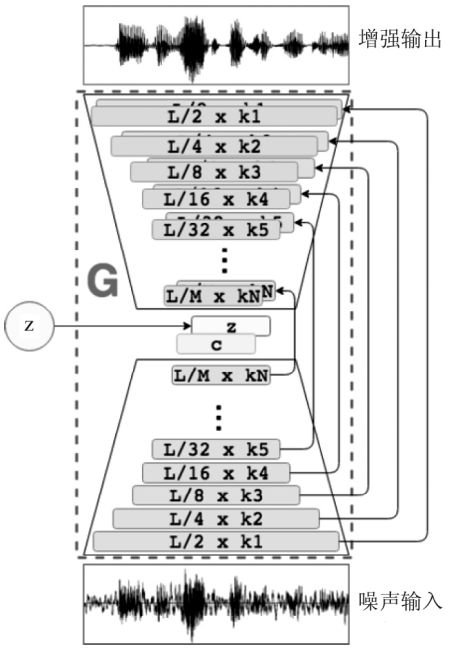

图 8　用于实现语音增强 GAN 的网络结构图[49]

Fig.8　Structure of GAN for speech enhancement

此外，Michelsanti 也对 GAN 在语音增强上的应用进行了研究[51]。其所设计的 GAN 结构与文献[49]的类似，只不过文献[51]所用的特征是语音信号的频域特征，且 G 和 D 的网络结构都是二维的 CNN。受到这文献[49]和[51]两项工作的启发，Fan 等人将 GAN 运用到了音源分离的工作中[52]。在 Fan 的工作中，其 G 和 D 的结构就是最基本 MLP。此外，其实现音源分离的过程与之前 Po-Sen 等人[39]的类似。值得注意的是，在该项工作中[52]，其训练 GAN 的过程与之前的工作[49,51]有所不同。在文献[52]中，生成器 G 首先进行了有监督式的学习(训练的误差函数是均方误差)，之后 G 的训练才加入 D，进行 GAN 的训练。这为今后 GAN 的训练提供了借鉴。此外，文献[53]提出了利用具有泛化性能 GAN 网络，其仍将语音波形作为输入输出特征，并加入了对抗音频回归误差，使其能对已经失真的语音进行增强。

另外，百度研究院的 Anuroop 等人受到 Pascual 工作的启发，将 GAN 用到了语音识别模型的前端，对语音进行降噪[53]。该项工作帮助语音识别系统提高了语音识别的准确率。此外，谷歌的 Bo Li 等人也对 GAN 是否能够提升语音识别系统的鲁棒性进行了研究[54]。在他们的工作中[54]，将 Pascual 所提出的 GAN 的结构[49]直接作为语音识别系统的前端，对含噪语音进行增强(与文献[49]不同的是文献[54]提出了在频域进行语音增强)。通过实验作者指出，GAN 或许不是最合适用于提高语音识别系统的鲁棒性的方法，但其有更广阔的前景。由于传统的 GAN[47]存在训练困难的问题，而 Wasserstein GAN[55]又能很好的解决该问题。因此，UIUC 的 Subakan 等人[56]提出使用 Wasserstein GAN 来完成音源分离的任务，且取得了不错的效果。

随着 GAN 的发展，在之后的研究中，Cycle GAN 网络结构[57]被提了出来。该网络结构最初被用于图像风格迁移任务。值得注意的是，由于该方法在训练阶段的时候不需要匹配的图像对其进行训练，因此其大大降低了对于训练集的要求，使得在现实生活中的大量图片可用于网络的训练。在语音领域，Cycle GAN 最初是被用于语音转换[58-59]上，它解决了在传统语音转换任务上需要平行数据集的问题，且取得了良好的转换效果。在语音增强任务上，Cycle GAN 被用于语音识别系统的前端[60-61]，有效地提高了整个语音识别系统在噪声环境下的鲁棒性，并降低了识别过程中的词错误率。此外，该方法在训练阶段同样也没有使用平行数据集。由此可以看出，Cycle GAN 在语音领域的应用有着很大的潜力。

总的来说，目前对用 GAN 进行语音增强(音源分离)的研究相对较少，并且与以往的 MLP，CNN，RNN 相比是否能取得更好的表现还不确定[62]，但是 GAN 在语音增强上的应用仍有着良好的前景。

6 结论与展望

语音增强算法发展至今，已取得了举世瞩目的成就。如今，随着计算机硬件的高速发展，与一些传统语音增强方法相比，基于深度神经网络的语音增强效果更为明显。本文从神经网络的结构(MLP，CNN，RNN 和 GAN)出发，对当下一些主流的基于 DNN 的语音增强算法进行了回顾。这四种网络各有它们的优势及劣势。尽管 RNN 要比 MLP 和 CNN 有更好的语音增强性能，但其计算复杂度相对较高，在一些移动设备上应用会存在一些困难。此外，虽然基于 GAN 的语音增强方法训练过程相对繁琐，但其在增强阶段能更好的拟合纯净语音的能量分布并且其在提高网络的泛化性能上，也有较大的潜力。

在今后的单通道语音增强的研究中，首先我们可以对基于 GAN 的语音增强方法进行更加深入的探索。该类方法由于可以利用非平行数据集对网络进行训练，因此更多现实生活中的语音能被用于网络的训练，进而提升整个网络的泛化性能。此外，一些更加新颖神经网络结构(如胶囊网络等)或训练方式(如增强学习等)，可以引入到语音增强的研究当中，探究其在语音增强上的性能。

参考文献

[1] Benesty J, Makino S, Chen J D. Speech Enhancement [M]. Springer, 2005.

[2] Loizou P C. Speech enhancement: Theory and practice [M]. CRC Press, 2013.

[3] Weiss W R, Aschkenasy E, Parsons T W. Study and the Development of the INTEL Technique for Improving Speech Intelligibility, Technical Report NSC-FR/4023, Nicolet Scientific Corporation, 1975.

[4] Boll S F. Suppression of acoustic noise in speech using spectral subtraction [J]. IEEE Transactions on Acoustics,

Speech, and Signal Processing, April, 1979, 27(2): 113–120.

[5] McAulay R J, Malpass M L. Speech enhancement using a soft-decision noise suppression filter [J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, April, 1980, 28(2): 137–145.

[6] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator [J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, April, 1985, 33(2): 443–445.

[7] Lim J. S, Oppenheim A. V. All-pole modeling of degraded speech [J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, June, 1978, 26(3): 197–210.

[8] Lim J. S, Oppenheim A. V. Enhancement and bandwidth compression of noisy speech [C] //Proceedings of the IEEE, Dec. 1979, 67(12): 1586–1604.

[9] Dendrinos M, Bakamidis S, Carayannis G. Speech enhancement from noise: A regenerative approach [J]. Speech Communication 10.1(1991): 45–57.

[10] Ephraim Y, Van Trees H L. A signal subspace approach for speech enhancement [J]. IEEE Transactions on Speech and Audio Processing, July 1995, 3(4): 251–266.

[11] Jabloun F, Champagne B. Incorporating the human hearing properties in the signal subspace approach for speech enhancement [J]. IEEE Transactions on Speech and Audio Processing, Nov. 2003, 11(6): 700–708.

[12] Bengio Y. Learning deep architectures for AI [M]. Now Foundations and Trends, 2009.

[13] Rumelhart D E, McClelland J L. Learning internal representations by error propagation [M]. MIT Press, 1987: 318–362.

[14] Xia B, Bao C. Speech enhancement with weighted denoising auto-encoder [C] //Interspeech2013, Lyon, France, August 25–29, 2013: 3444–3448.

[15] Xia B, Bao C. Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification [J]. Speech Communication 60(2014): 13–29.

[16] Xu Y, Du J, Dai L, et al. An Experimental Study on Speech Enhancement Based on Deep Neural Networks [J]. IEEE Signal Processing Letters, Jan. 2014, 21(1): 65–68.

[17] Xu Y, Du J, Dai L, et al. A Regression Approach to Speech Enhancement Based on Deep Neural Networks [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Jan. 2015, 23(1): 7–19.

[18] Nitish S, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research 15.1(2014): 1929–1958.

[19] Seltzer M L, Yu D, Wang Y. An investigation of deep neural networks for noise robust speech recognition [C] // 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC Canada, May 26–31, 2013: 7398–7402.

[20] Xu Y, et al. Dynamic noise aware training for speech enhancement based on deep neural networks [C] // Interspeech2014, Singapore, September 14–18, 2014: 2670–2674.

[21] Kumar A, Florencio D. Speech Enhancement in Multiple-Noise Conditions Using Deep Neural Networks [C] // Interspeech 2016, San Francisco, USA, September 8–12, 2016: 3738–3742.

[22] Painter T, Spanias A. A review of algorithms for perceptual coding of digital audio signals [C] //Proceedings of 13th International Conference on Digital Signal Processing, Santorini, Greece, July 2–4, 1997: 179–208.

[23] Xu Y, Du J, Huang Z, et al. Multi-Objective Learning and Mask-Based Post-Processing for Deep Neural Network Based Speech Enhancement [C] //Interspeech2015, Dresden, Germany, September 6–10, 2015: 1508–1512.

[24] Wang Q, Du J, Dai L, et al. Joint noise and mask aware training for DNN-based speech enhancement with SUB-band features [C] //2017 Hands-free Speech Communications and Microphone Arrays(HSCMA), San Francisco, CA, 2017: 101–105.

[25] Wang Y, Narayanan A, Wang D. On Training Targets for Supervised Speech Separation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Dec. 2014, 22(12): 1849–1858.

[26] Narayanan A, Wang D. Ideal ratio mask estimation using deep neural networks for robust speech recognition [C] // 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013: 7092–7096.

[27] Williamson D S, Wang Y, Wang D. Complex Ratio Masking for Monaural Speech Separation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, March 2016, 24(3): 483–492.

[28] Erdogan H, Hershey J R, Watanabe S, et al. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks [C] // 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, QLD, Australia, April 19–24, 2015: 708–712.

[29] Mirsamadi S, Tashev I. Causal Speech Enhancement Combining Data-Driven Learning and Suppression Rule Estimation [C] //Interspeech2016, San Francisco, United states, September 8–12, 2016: 2870–2874.

[30] Phapatanaburi K, et al. Noise robust voice activity detection using joint phase and magnitude based feature enhancement [J]. Journal of Ambient Intelligence and Humanized Computing, 2017: 1-15.

[31] Sainath T N, Mohamed A, Kingsbury B, et al. Deep convolutional neural networks for LVCSR [C] //2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013: 8614-8618.

[32] Kounovsky T, Malek J. Single channel speech enhancement using convolutional neural network [C] // 2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM), Donostia-San Sebastian, Spain, May 24-26, 2017: 1-5.

[33] Park S R, Lee J. A fully convolutional neural network for speech enhancement [C] // Interspeech2017, Stockholm, Sweden, August 20-24, 2017: 1993-1997.

[34] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Las Vegas, NV, 2016: 770-778.

[35] Fu S, Tsao Y, Lu X, et al. Raw Waveform-based Speech Enhancement by Fully Convolutional Networks [C] //9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Kuala Lumpur, Malaysia, December 12-15, 2017: 6-12.

[36] Grais E M, Plumbley M D. Single Channel Audio Source Separation using Convolutional Denoising Autoencoders [C] //2017 IEEE Global Conference on Signal and Information Processing, Montreal, QC, Canada, November 14-16, 2017: 1265-1269.

[37] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model [C] // Interspeech 2010: 1045-1048.

[38] Maas A L, Le Q V, O'Neil T M, et al. Recurrent neural networks for noise reduction in robust ASR [C] // Interspeech2012, Portland, OR, United states, September 9-13, 2012: 22-25.

[39] Huang P, Kim M, Hasegawa-Johnson M, et al. Singing-Voice Separation from Monaural Recordings using Deep Recurrent Neural Networks [C] // 15th International Society for Music Information Retrieval Conference, Taipei, Taiwan, October 27-31, 2014: 477-482.

[40] Huang P, Kim M, Hasegawa-Johnson M, et al. Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Dec. 2015, 23(12): 2136-2147.

[41] Uhlich S, Porcu M, Giron F, et al. Improving music source separation based on deep neural networks through data augmentation and network blending [C] //2017 IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, United states, March 5-9, 2017: 261-265.

[42] Wöllmer M, Zhang Z, Weninger F, et al. Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise [C] //2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, May 26-31, 2013: 6822-6826.

[43] Sun L, Du J, Dai L, et al. Multiple-target deep learning for LSTM-RNN based speech enhancement [C] // 2017 Hands-free Speech Communications and Microphone Arrays(HSCMA), San Francisco, CA, United states, March 1-3, 2017: 136-140.

[44] Sainath T N, Vinyals O, Senior A, et al. Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks [C] //2015 IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, QLD, Australia, April 19-24, 2015: 4580-4584.

[45] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[46] Mimilakis S I, Drossos K, Santos J F, et al. Monaural Singing Voice Separation with Skip-Filtering Connections and Recurrent Inference of Time-Frequency Mask [C] // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, April 15-20, 2018: 721-725.

[47] Goodfellow I J, Pouget-Abadie J. Mirza M, et al. Generative adversarial nets [M]. Advances in Neural Information Processing Systems, 2014.

[48] Isola P, Zhu J, Zhou T, et al. Image-to-image translation with conditional adversarial networks [C] //30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, United states, July 21-26, 2017: 5967-5976.

[49] Pascual S, Bonafonte A, Serra J. SEGAN: Speech Enhancement Generative Adversarial Network [C] // Interspeech 2017, Stockholm, Sweden, August 20-24, 2017: 3642-3646.

[50] Mao X, Li Q, Xie H, et al. On the effectiveness of least squares generative adversarial networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, Dec.1, 2019, 41(12): 2947-2960.

[51] Michelsanti D, Tan Z. Conditional Generative Adversarial

Networks for Speech Enhancement and Noise-Robust Speaker Verification [C] // Interspeech2017, Stockholm, Sweden, August 20-24, 2017: 2008-2012.

[52] Fan Z, Lai Y, Jang J. SVSGAN: Singing Voice Separation via Generative Adversarial Network [C] // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, April 15-20, 2018: 726-730.

[53] Sriram A, Jun H, Gaur Y, et al. Robust Speech Recognition Using Generative Adversarial Networks [C] // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, April 15-20, 2018: 5639-5643.

[54] Donahue C, Li B, Prabhavalkar R. Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition [C] // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, April 15-20, 2018: 5024-5028.

[55] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. arXiv preprint arXiv: 1701.07875v3, 2017.

[56] Mirza M, Osindero S. "Conditional generative adversarial nets." arXiv preprint arXiv: 1411.1784, 2014.

[57] Zhu J, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C] // 2017 IEEE International Conference on Computer Vision, Venice, Italy, October 22-29, 2017: 2242-2251.

[58] Fang F, Yamagishi J, Echizen I, et al. High-quality nonparallel voice conversion based on cycle-consistent adversarial network [C] // 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing, Calgary, AB, Canada, April 15-20, 2018: 5279-5283.

[59] Kaneko T, Kameoka H. Parallel-data-free voice conversion using cycle-consistent adversarial networks. arXiv preprint arXiv: 1711.11293v2, 2017.

[60] Mimura M, Sakai S, Kawahara T. Cross-domain speech recognition using nonparallel corpora with cycle-consistent adversarial networks [C] // 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, December 16-20, 2017: 134-140.

[61] Meng Z, Li J, Gong Y, et al. Cycle-Consistent Speech Enhancement [C] // Interspeech2018, Hyderabad, India, September 2-6, 2018: 1165-1169.

[62] Wang D, Chen J. Supervised speech separation based on deep learning: an overview [J]. IEEE/ACM Transactions on Audio Speech and Language Processing, October, 2018, 26(10): 1702-1726.

[63] Pascual S, Serrà J, Bonafonte A. Towards Generalized Speech Enhancement with Generative Adversarial Networks [C] // Interspeech2019, Graz, Austria, September 15-19, 2019: 1791-1795.

[64] Yang Y, Bao C. RS-CAE-Based AR-Wiener Filtering and Harmonic Recovery for Speech Enhancement [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(11): 1752-1762.

[65] Cui Z, Bao C. Linear Prediction-based Part-defined Auto-encoder Used for Speech Enhancement [C] // 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing, Brighton, United kingdom, May 12-17, 2019: 6880-6884.

[66] Delfarah M, Wang D. Recurrent Neural Networks for Co-channel Speech Separation in Reverberant Environments [C] // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, April 15-20, 2018: 5404-5408.

[67] Delfarah M, Wang D. Deep Learning for Talker-Dependent Reverberant Speaker Separation: An Empirical Study [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(11): 1839-1848.

作者简介

鲍长春　男，1965 年生，内蒙古赤峰人。现为北京工业大学教授，博士生导师。主要研究方向为语音与音频信号处理。

E-mail: chchbao@bjut.edu.cn

项　扬　男，1994 年生，云南昆明人。2019 年于北京工业大学获硕士学位，目前是丹麦奥尔堡大学(Aalborg University)博士研究生，主要研究方向为语音增强。

E-mail: xiangyang3131777@emails.bjut.edu.cn