计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190306
57(5):928-953,2020

单声道语音降噪与去混响研究综述

蓝 天 彭 川 李 森 叶文政 李 萌 惠国强 吕忆蓝 钱宇欣 刘 峤
(电子科技大学信息与软件工程学院 成都 610054)
(lantian1029@uestc.edu.cn)

An Overview of Monaural Speech Denoising and Dereverberation Research

Lan Tian, Peng Chuan, Li Sen, Ye Wenzheng, Li Meng, Hui Guoqiang, Lü Yilan, Qian Yuxin, and Liu Qiao
(*School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu* 610054)

Abstract Speech enhancement refers to the use of audio signal processing techniques and various algorithms to improve the intelligibility and quality of the distorted speech signals. It has great research value and a wide range of applications including speech recognition, VoIP, tele-conference and hearing aids. Most early work utilized unsupervised digital signal analysis methods to decompose the speech signal to obtain the characteristics of the clean speech and the noise. With the development of machine learning, some supervised methods which aim to learn the relationship between noisy and clean speech signals were proposed. In particular, the introduction of deep learning has greatly improved the performance. In order to help beginners and related researchers to understand the current research status of this topic, this paper conducts a comprehensive survey of the development process of the monaural speech enhancement, and systematically summarizes from the aspect of model methods, datasets, features, evaluation metrics, etc. First, we divide speech enhancement into noise reduction and de-reverberation, then respectively sort out the existing work of traditional and machine-learning-based methods in these two directions. Moreover, we briefly introduce the main ideas of typical solutions, and compare the performance of different methods. Then, commonly used datasets, features, learning objectives and evaluation metrics in experiments are enumerated and illustrated. Finally, four major challenges and corresponding issues in this area are summarized.

Key words speech enhancement; speech denoising; speech dereverberation; machine learning; deep neural network

摘 要 语音增强是提高语音质量与可懂度的关键技术,在语音识别、语音通话、电话会议和听力辅助等领域具有广泛应用前景与重要研究价值.从模型方法、数据集、特征、评估指标等方面,对单声道语音增

收稿日期:2019-05-27;**修回日期**:2019-11-18
基金项目:国家自然科学基金项目(U19B2028,61772117);提升政府治理能力大数据应用技术国家工程实验室开放基金项目(10-2018039);四川省科技服务业示范项目(2018GFW0150);中央高校基本科研业务费专项资金(ZYGX2019J077)
This work was supported by the National Natural Science Foundation of China (U19B2028, 61772117); the Big Data Application on Improving Government Governance Capabilities National Engineering Laboratory Open Fund Project (10-2018039), the Sichuan Hi-Tech Industrialization Program (2018GFW0150), and the Fundamental Research Funds for the Central Universities (ZYGX2019J077).
通信作者:刘峤(qliu@uestc.edu.cn)

强研究工作的发展现状进行了全面调研和深入分析.1)对传统的与基于机器学习的单声道语音降噪以及语音去混响的已有研究工作进行了梳理分类,简要介绍了典型方法的研究思路,并对不同方法的实验结果进行了综合比较;2)对在实验与结果评估过程中所涉及到的常用数据集、常见特征、学习目标与评估指标等进行了整理与介绍;3)对目前单声道语音增强仍然面临的主要问题与挑战进行了总结.

关键词　语音增强;语音降噪;语音去混响;机器学习;深度神经网络

中图法分类号　TP391.4; TN912.3

语音增强是指利用音频信号处理技术及各种算法提高失真语音信号的可懂度或整体感知质量,从而进一步在语音识别、语音通话、电话会议、场景录音、军事窃听和听力辅助等场景中改善应用效果.语音增强属于语音分离的一项内容,而后者还包括说话人分离等.狭义的语音增强单指语音降噪,而广义的语音增强还包括语音去混响[1],因为语音去混响也是提高语音质量的重要手段.根据接收端麦克风数目的不同,可以将语音增强分为单声道(单个麦克风)与多声道(多个麦克风)2类.单声道语音增强算法只需单个麦克风,实现的成本较低,在实际生活中得到了广泛的应用[1].由于单声道增强算法获取的音频信息量较少,且无法利用声音传播的空间信息,它的实现更具挑战[2-3].本文着重关注广义的单声道语音增强(为简化叙述,后文如无特别说明则省略"单声道"限定语),对语音降噪与语音去混响两方面的研究工作都进行了调研分析.

早期的语音降噪或去混响主要通过数字信号分析方法,如谱减法、滤波法等,从时域、频域或时频结合的方式对语音信号进行分解,找到纯净语音或噪声的特征,从而将二者分离,属于无监督的方法.随着机器学习技术的演进,有监督的方法不断地被提出,学者们开始尝试通过各种机器学习模型去自动发现带噪(带混响)语音与纯净语音信号之间的关系,近年来最有代表性的莫过于深度学习在本领域的应用,它极大提升了语音降噪、去混响的效果.

本文对单声道语音增强的现有研究工作进行了梳理分类,简要介绍了典型方法的研究思路,并对具备可比性的实验结果进行了综合比较,有助于本领域研究人员进一步分析这些方法之间的联系与区别;对在实验与评估过程中所涉及到的相关基本概念进行了整理与简介,并提供出处来源,有利于初学者查阅所需预备知识;在全面分析相关研究工作现状的基础上,探讨了目前单声道语音增强仍然面临的主要问题与挑战,可供本领域研究人员参考归纳未来的研究方向.

1　传统的语音降噪方法

语音降噪是语音处理领域的一个基本问题,旨在从受噪声干扰的信号中有效地分离出目标信号.噪声干扰对语音活动检测和语音识别等任务的准确率具有很大的影响,因而研究解决噪声对后续语音处理任务的影响一直受到学术界的广泛关注[4].传统的语音降噪方法主要是基于数字信号处理等算法,主要包括谱减法、维纳滤波、基于统计模型以及子空间的方法等.

1.1　谱减法

谱减法是最早期提出的降噪算法之一,它基于一个简单假设:噪声是加性噪声.通过从带噪语音谱中减去对噪声谱的估计来得到降噪后的语音谱,其基本做法如图1所示,做出这一假设是基于噪声的平稳性或者是一种慢变的过程[5].由于实际噪声的非平稳特性,在使用过程中,这种方法很容易由于谱减过程中减去谱成分的过大或过小造成语音失真,即产生令人困扰的音乐噪声.为减轻由谱减过程引入的语音失真,最常用的一种方式就是采用过减因子来控制失真程度,众多学者提出了不同的准则来计算过减因子[6-8],例如对差分谱做半波整流(half-wave rectification, HWR)和基于心理声学掩蔽阈值的方法.随着小波技术的发展,Zhong等人[9]根据硬阈值和软阈值改进了基于小波降噪的阈值函数算

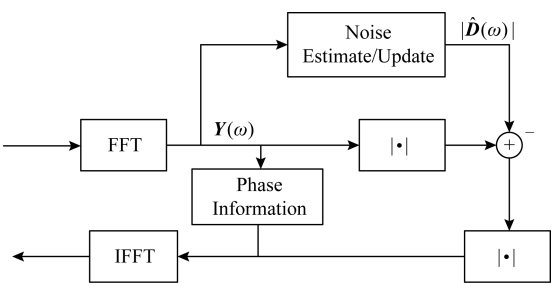


Fig. 1　Spectral subtraction based speech enhancement method

图1　基于谱减法的语音增强方法

法,该方法有效地减少了降噪后信号中的毛刺现象.但是受到假设条件的限制,谱减法始终不能有效地解决音乐噪声的问题.

1.2 滤波法

不同于基于简单假设的谱减法,维纳滤波器的提出是基于最小均方误差意义的最优解,通过求解最优化均方误差计算得到增强信号[10],基本流程如图 2 所示,但是它的推导仍然是基于所分析信号具有平稳性这一假设,不能有效地处理非平稳信号的情况.在后续改进中,通过使用卡尔曼(Kalman)滤波器,滤波法成功地被推广到处理非平稳信号和噪声的场景下[11-12].Wang 等人[13]提出了一种使用卡尔曼滤波器进行调制域语音增强的算法,利用高斯环统计模型将语音和噪声频谱幅度进行结合,通过高斯混合来模拟复数傅里叶域中语音和噪声的先验分布;Andersen 等人[14]将多声道技术,即基于语音失真加权的帧间维纳滤波器(speech-distortion weighted inter-frame Wiener filter)应用于单声道,进一步利用二次高分辨率滤波器组(secondary higher resolution filter bank)改进了对帧间相关性(inter-frame correlation, IFC)的估计,更好地在语音降噪和失真之间找到一个平衡参数,减轻了增强语音失真;Peng 等人[15]在线性预测残差域中结合人类听觉系统的掩蔽特性,进一步抑制了残留噪声.

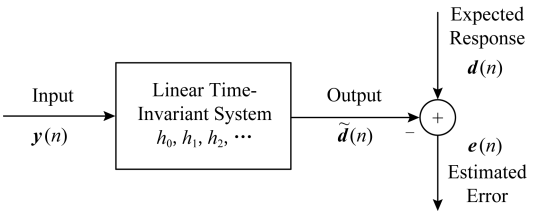


Fig. 2 Wiener filtering based speech enhancement method

图 2 基于维纳滤波法的语音增强方法

1.3 基于统计模型的方法

最小均方误差(minimum mean-square error, MMSE)估计是一种常用的基于统计模型的语音降噪方法,与维纳滤波的区别在于,基于 MMSE 的语音降噪方法可以得到对降噪语音谱的非线性估计[16-18].该方法对短时频谱幅度(short time spectral amplitude, STSA)进行最优估计,即得到关于估计幅度与实际幅度均方误差的最小优化估计器:

$$e=E\{\hat{X}_k-X_k\}, \tag{1}$$

其中,$\hat{X}_k$ 是估计幅度,X_k 是纯净语音的幅度.Faraji 等人[19]推导出加性高斯噪声中语音增强的最小均方误差和最大后验(maximum a posteriori, MAP)估计的闭合形式解,减少了增强信号的失真;考虑到实际噪声环境中语音出现的随机性,Jia 等人[20]采用二元假设模型提出了一种改进的最小均方误差对数谱振幅(minimum mean squared error log spectral amplitude, MMSE-LSA)估计方法,在给定信号模型的基础上,利用音高同步分割对信号进行优化分析,通过频率和类相关的先验分布来解决谐波检索问题;Stahl 等人[21]提出了一种联合检测-估计框架,先对语音中的随机和确定部分做出假设,然后通过制定最大后验标准来解决谐波恢复问题,最终利用所有信息来估计语音的频谱图.因为基于信号处理的方法对信号具有严格的假设限制,在不满足假设条件的真实环境下,特别是在低信噪比条件下,这些方法通常会失效[22].

1.4 子空间方法

子空间方法是一种基于线性代数理论的语音降噪方法,这类算法假设纯净信号可以被视为带噪信号在 Euclidean 空间中的一个子空间,通过将带噪信号向量空间分解为纯净信号主导和噪声信号主导的 2 个子空间,从而可以简单地通过去除落在“噪声空间”中的带噪向量分量来估计纯净信号[22].带噪信号分解为 2 个子空间常用的正交矩阵方法有奇异值分解(singular value decomposition, SVD)[23-24]和特征值分解(eigenvalue decomposition, EVD).Ephraim 等人[25]提出了利用协方差矩阵的特征值分解,通过利用 Karhunen-Loéve 变换(Karhunen-Loéve transform, KLT)进行信号分解,在满足残余噪声低于预设阈值约束的同时实现了语音失真最小化.

我们统计并比较了传统的语音降噪方法在不同噪声环境以及不同信噪比下的主观语音质量评估(perceptual evaluation of speech quality, PESQ)和短时客观可懂度(short-time objective intelligibility, STOI)指标,如表 1 和表 2 所示.其中 PESQ 取值范围为 −0.5～4.5,STOI 取值范围为 0～1,两者的数值越高表示降噪效果越好,详见 4.3 节所述.

2 基于机器学习的语音降噪方法

语音降噪问题可以视为一个监督性学习问题,很多学者考虑使用机器学习的方法来解决语音降噪的问题.由于计算机硬件的限制,早期的有监督模型一般都是在浅层模型以及小数据集上实现的;在

Table 1　Comparison of PESQ Scores in Traditional Speech Denoising Methods

表 1　传统语音降噪方法的 PESQ 指标对比

Method	Babble				White				Factory				Experimental Configuration	
	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB	Speech Corpus	Test-set Size (Number of Utterance)
GSVD[15]	1.39	1.79	2.16	2.52	1.4	1.81	2.21	2.6	1.37	1.76	2.18	2.56	TIMIT	200
PCGSVD[15]	1.43	1.84	2.22	2.62	1.46	1.87	2.28	2.67	1.43	1.83	2.26	2.68	TIMIT	200
CMMSE[15]	1.52	1.94	2.33	2.69	1.78	2.2	2.59	2.93	1.5	1.94	2.35	2.72	TIMIT	200
PCMMSE[15]	1.7	2.09	2.45	2.79	1.78	2.23	2.6	2.94	1.68	2.08	2.46	2.83	TIMIT	200
NLM+VMD[26]		1.92	2.14	2.41		2.51	2.85	3.01		2.29	2.46	2.71	TIMIT	10
SS+PLL[27]		1.53	1.79										NOISEUS	
MMSE-STSA+PLL[27]		1.42	1.73										NOISEUS	
NMF+PLL[27]		1.72	1.92										NOISEUS	
LSA+PI[28]	1.21	1.72	2.19	2.58	1.58	1.91	2.24	2.55	1.51	1.99	2.35	2.65	GRID Corpus	50
LSA+PQI[28]	1.23	1.71	2.17	2.56	1.54	1.82	2.14	2.46	1.51	1.93	2.29	2.62	GRID Corpus	50
LSA+Bi-Phase[28]	1.20	1.71	2.17	2.55	1.55	1.86	2.19	2.51	1.5	1.95	2.33	2.63	GRID Corpus	50

Table 2　Comparison of STOI Scores in Traditional Speech Denoising Methods

表 2　传统语音降噪方法的 STOI 指标对比

Method	Babble				White				Factory				Experimental Configuration	
	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB	Speech Corpus	Test-set Size (Number of Utterance)
LSA+PI[28]	0.44	0.58	0.72	0.81	0.57	0.65	0.72	0.78	0.50	0.63	0.74	0.82	GRID Corpus	50
LSA+PQI[29]	0.45	0.60	0.74	0.84	0.58	0.67	0.74	0.81	0.51	0.65	0.76	0.84	GRID Corpus	50
LSA+Bi-Phase[29]	0.45	0.59	0.73	0.83	0.59	0.67	0.73	0.79	0.51	0.64	0.75	0.83	GRID Corpus	50
SMPost(DD)[29]		0.76*	0.88*	0.96*		0.68*	0.78*	0.84*					NOISEUS	30
SMPost(MDD)[29]		0.78*	0.90*	0.97*		0.72*	0.79*	0.88*					NOISEUS	30
SMPrio(DD)[29]		0.72*	0.81*	0.92*		0.62*	0.72*	0.82*					NOISEUS	30
SMPrio(MDD)[29]		0.78*	0.88*	0.94*		0.68*	0.79*	0.88*					NOISEUS	30

Note: The symbol * comes from our estimation of the graph in the reference paper.

2006 年 Hinton 等人[30]提出了一种基于受限玻尔兹曼机的逐层学习方案,并将其应用于深层神经网络(deep neural network, DNN)的网络训练中,解决了 DNN 训练中的局部最优问题,显示出监督性学习的建模优势.此后,得益于 DNN 的层次化非线性处理能力,深度学习的概念被广泛应用于语音[31-32]、图像[33]及自然语言处理[34]任务中,迅速发展成为机器学习领域的一个重要分支,越来越多的学者开始探索深度学习在语音降噪方面的应用.

早期的经典 DNN 模型通常由一个输入层,若干非线性隐含层以及一个输出层组成,层与层之间相互堆叠,前一层的输出传递到后一层,形成一个深层网络.相比于浅层网络,深层模型更擅长从原始数据中学习对目标有用的特征表示,比较典型的神经网络有卷积神经网络(convolutional neural network, CNN)[35-36]、循环神经网络(recurrent neural network, RNN)[35]以及 2014 年提出的生成对抗网络(generative adversarial network, GAN)[37]等.在基于深度学习的语音降噪任务中,根据神经网络是否对语音时域波形直接处理可以分为非端到端和端到端的语音降噪;在非端到端的语音降噪任务中,根据网络的学习目标的不同,可以把降噪方法分为:基于时频掩蔽(time-frequency mask)的语音降噪算法、基于频谱映射的语音降噪算法和基于信号近似的语音降噪算法;一些学者也提出了基于端到端的算法以及深度学习与传统方法结合的算法.本节将介绍传统机器学习、非端到端方法以及端到端的方法在语音降噪领域的应用.

图 3 给出了非端到端的语音降噪算法结构图，在训练阶段首先通过时频分解、特征提取将原始的时域波形处理为时频表示，随后将时频表示的特征送入到神经网络中进行训练，将估计出的目标作用于带噪语音得到降噪后的语音；经过多轮迭代调整网络参数，使其更好地学习带噪语音与纯净语音之间的复杂映射关系.在测试阶段，提取特征后的带噪语音被输入到训练好的降噪模型中，降噪后的语音时频表示与带噪语音的相位结合便可得到时域的波形信号.与图 3 类似，图 4 给出了端到端的语音降噪模型，通过直接学习时域波形层级的映射关系，在保留更多原始波形信息的同时，简化了处理流程.

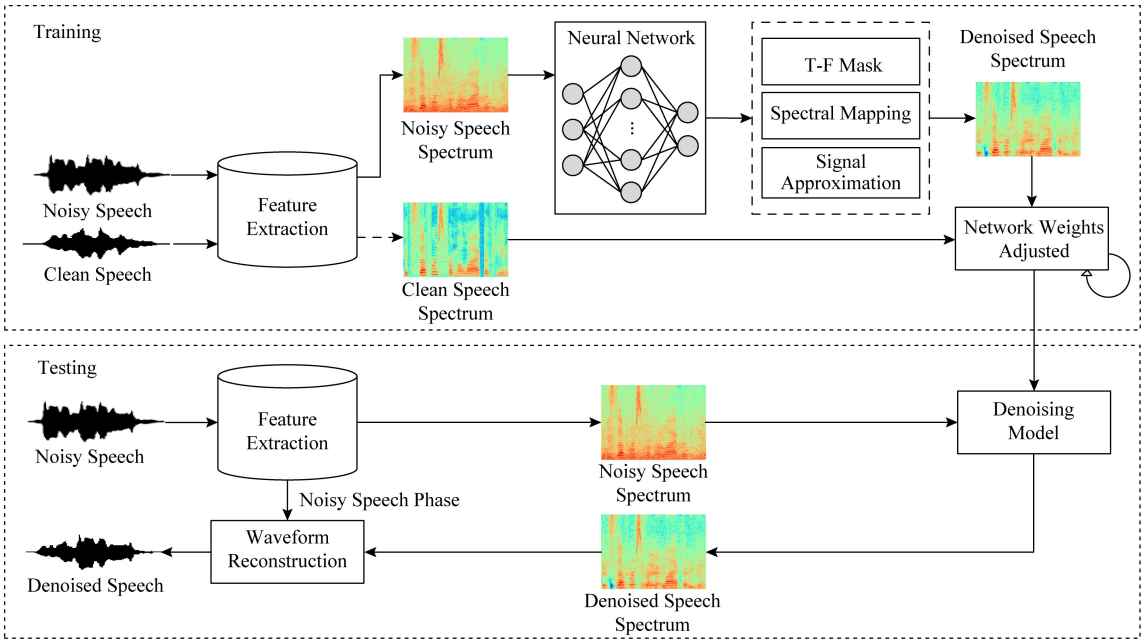


Fig. 3 A block diagram of non-end-to-end speech denoising system based on deep learning

图 3 基于深度学习的非端到端语音降噪系统结构框图

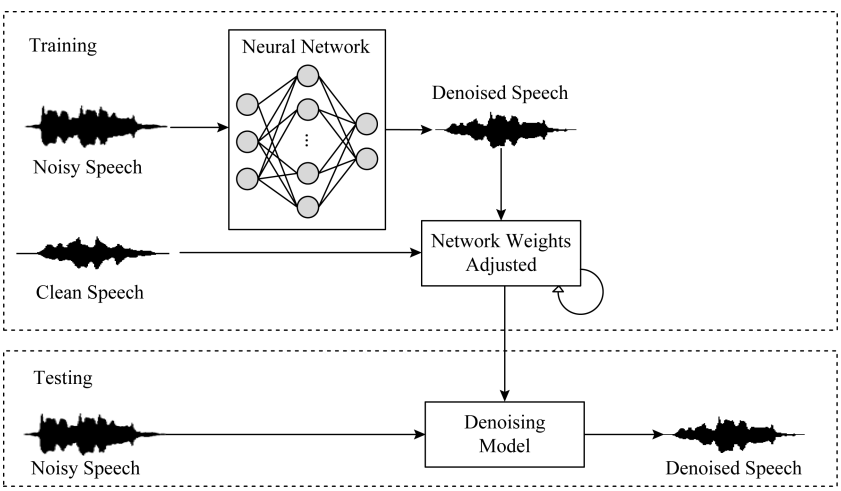


Fig. 4 A block diagram of end-to-end speech denoising system based on deep learning

图 4 基于深度学习的端到端语音降噪系统结构框图

2.1 基于传统机器学习模型的方法

早期语音降噪系统模型主要是一些浅层模型，经典的方法包含高斯混合模型(Gaussian mixture model, GMM)、支持向量机(support vector machine, SVM)和非负矩阵分解(nonnegative matrix factorization, NMF).

高斯混合模型通过多个高斯分布函数的线性组合，来模拟复杂的分布.Kim 等人[38]利用 GMM 对

时频单元进行建模,通过输入给定的频带特征,输出语音主导和噪声主导的概率.利用估计的二值掩蔽和混合语音的 Gammatone 滤波输出合成语音的时域波形.但由于该模型是单独对每一个频带进行建模,忽略了频带间的相关性,不具有较强的实用性.

支持向量机通过在高维特征空间中寻找最优分类面对数据进行分割.Han 等人[39]利用 SVM 对每个频带的时频单元进行建模,学习被目标语音主导的时频单元和被噪音主导的时频单元最优区分面,通过计算到分类面的距离实现时频单元的分类.相比于 GMM,SVM 具有更好的分类准确性和泛化性能.

非负矩阵分解是最常用的有监督语音降噪方法[40-41].NMF 算法对纯净语音和噪声单独训练,分别得到对语音和噪声的信号基表示,从而在带噪语音中分离出纯净语音.为了减少具有与语音信号类似的特征的残余噪声分量,Chung 等人[42]提出了基于 NMF 的类条件基矢量的训练和补偿算法.但是当遇到在训练阶段没有出现过的语音或者噪音时,算法性能会出现下降.

2.2　基于深度学习模型的方法

2.2.1　基于时频掩蔽的方法

基于时频掩蔽的语音降噪方法将描述纯净语音与噪声之间相互关系的时频掩蔽作为学习目标.研究表明,基于时频掩蔽的方法可以有效地提高复杂环境下的语音可懂度[38,43],但该方法需要假设纯净语音与噪声之间有一定的独立性.理想二值掩蔽(ideal binary mask, IBM)[44]是最早用于语音降噪的时频掩蔽之一,它实际上是一个定义在二维空间(时间和频率)上的一个二值(0 或 1)矩阵,其中每个元素:

$$f_{\mathrm{IBM}}(t,f)=\begin{cases}1, & f_{\mathrm{SNR}}(t,f)>\rho_{\mathrm{LC}},\\ 0, & \text{otherwise},\end{cases} \tag{2}$$

其中,t 和 f 分别表示时刻和频率,$f_{\mathrm{SNR}}(t,f)$ 表示在时刻 t、频率 f 处时频单元的局部信噪比.当局部信噪比大于局部阈值(local criterion, LC) ρ_{LC} 时,IBM 在此处赋值为 1,否则赋值为 0,这代表 IBM 将每个时频单元判定为以语音为主或以噪声为主.除此之外,也有一些基于比值的掩蔽或复数域掩蔽相继被提出,例如理想比值掩蔽(ideal ratio mask, IRM)[45]、最优比值掩蔽(optimal ratio time-frequency mask, ORM)[46]、频谱幅度掩蔽(spectral magnitude mask, SMM)[47]、相位敏感掩蔽(phase-sensitive mask, PSM)[48]以及复数域理想比值掩蔽(complex ideal ratio mask, cIRM)[49]等.这些掩蔽根据语音及噪声的幅度谱或功率谱计算得到,随后通过将逆变换技术作用于估计的时频掩蔽上,从而合成目标语音的时域波形.

Wang 等人[50-51]将 DNN 引入语音分离与降噪领域,并对该工作进行扩展.他们将受限玻尔兹曼机(restricted Boltzmann machine, RBM)预训练的前馈 DNN 作为二元分类器来估计 IBM,并考虑了语音的时间动态特性,引入结构化感知机和条件随机场来改进模型.实验证明:相比于传统方法,基于 DNN 的方法在匹配和不匹配的噪声情况下均取得了很好的降噪效果.在扩展工作中,Wang 等人[52]对通过 Gammatone 滤波器组的子带信号使用 DNN 来学习输入信号的特征,他们将训练网络中最后一个隐藏层的输出与输入特征串联起来送入 SVM 中估计 IBM,经过实验评估作者取得了高的语音可懂度,但是语音质量损失较为严重;Healy 等人[43]将该算法扩展为 2 阶段训练方式,利用数据的上下文信息,显著提高了分类精度.作者在专业测验中测试了该算法,结果表明,对于听力正常和听力受损的听众,语音可懂度均显著提高.

Narayanan 和汪德亮[53-54]将理想比率掩蔽 IRM 作为目标,在梅尔谱域估计 IRM,并在一定程度上提高了语音识别的鲁棒性;Madhu 等人[55]也发现连续性学习目标相比于二值目标可以取得更好的性能;Nie 和 Zhang 等人[56-57]提出了一种用于 IBM 估计的深度叠加网络,并使用掩码进行基音估计,提高了掩码估计和基音估计的精度;Williamson 等人[49]提出复数理想比例掩蔽 cIRM 并使用 DNN 同时估计 cIRM 的实部和虚部,极大提高了语音可懂度;Hui 等人[58]使用卷积网络,通过 Maxout 和 Dropout 方法分别解决了训练的饱和问题以及泛化问题,并在客观可懂度和语音质量方面均超过了基于 DNN 的方法;Wang 等人[47]在语音分离任务中分析对比了一系列时频掩蔽的训练目标,从论文结果中可以看出,以 IRM 为训练目标的方法可以得到更好的语音质量与可懂度.

2.2.2　基于特征映射的方法

基于特征映射的语音降噪方法利用带噪语音特征与纯净语音特征之间的复杂关系,学习两者间的映射.网络的输入与输出通常是同种类型的声学特征,并且在实现过程中,几乎没有对语音和噪声信号做任何假设.常见的特征映射包括目标幅度谱(target magnitude spectrum, TMS)、Gammatone

域目标功率谱(Gammatone frequency target power spectrum, GF-TPS)以及短时傅里叶变换幅度谱(short-time Fourier transform spectrum, SFTS)等.其中,TMS[59-62]从带噪语音中估计纯净语音幅度谱、功率谱或梅尔谱等,然后将得到的幅度与带噪语音相位结合,得到估计语音波形;GF-TPS[47]是基于Gammatone滤波器的听觉谱(cochleagram),通过听觉谱转换,可以很容易地将GT-TPS的估计结果转换为降噪的语音波形;SFTS是语音的时域信号经过分帧、加窗以及短时傅里叶变换得到的时频表示.若不考虑相位不匹配的影响,则可直接估计目标语音的短时傅里叶变换(short-time Fourier transform, STFT)幅度谱,结合带噪语音相位信息后,通过短时傅里叶逆变换(inverse short-time Fourier transform, ISTFT)可估计得到目标语音的时域波形.

自动编码器是基于特征映射的语音降噪算法中的一类典型结构,Vincent等人[63]在2008年首次提出降噪自动编码器(denoising autoencoder, DA),并将其用于提取鲁棒性的特征;在此基础上,Maas等人[64]提出了循环降噪自动编码器(recurrent denoising autoencoder, RDA),并将该方法应用到语音识别的前端降噪任务上,降低了语音识别的错误率;Xia等人[65]利用降噪自动编码器估计纯净语音的频谱,然后用最小控制迭代平均的方法估计噪声,进而计算出先验信噪比,最后用维纳滤波的方法得到纯净语音的频谱估计;Lu等人[59]提出用堆叠式自动编码进行语音降噪,将多个训练好的自动编码器(autoencoder, AE)叠加成一个深层自动编码器(deep autoencoder, DAE),然后使用反向传播算法对其进行监督微调.通过DAE学习一个梅尔域带噪语音到纯净语音的功率谱映射,并在匹配噪声的情况下取得了一定的降噪效果.

Xu等人[61,66]提出把深层神经网络视为一个回归模型,作者使用带RBM预训练的DNN将带噪语音的对数功率谱映射到纯净语音的对数功率谱上,然后使用混合语音的相位,通过ISTFT得到目标语音的时域波形信号;作者使用了多种噪声来构建训练数据集,降噪后的PESQ比带噪语音高0.4～0.5,明显高于传统语音降噪方法,并且具有较好的泛化性能.Han等人[62]使用DNN来学习带混响和噪声的语音到纯净语音的映射关系,提高了语音可懂度与信噪比;Tu等人[67]在DNN非连续层之间添加了跳连接,间接地迫使神经网络学习IRM,另外,作者将网络结构堆叠起来,取得了更好的评估结果;Wang等人[68]发现直接使用标准的前馈神经网络把带噪信号映射到纯净信号的效果不理想,所以他们将傅里叶逆变换融合到神经网络中.Karjol等人[69]考虑到单个DNN可能无法更好地挖掘语音信号的时空结构信息,所以他们使用了添加门控网络的多DNN策略来训练数据,并取得了优于单个DNN的降噪效果.也有一些基于CNN的方法被用于频谱映射,通常CNN模型由输入层、卷积层、池化层、全连接层和输出层组成,通过卷积层与池化层的级联挖掘特征信息,另外CNN中的权重共享可以减少训练参数的数量.Park等人[70]提出冗余卷积编码解码网络(redundant convolutional encoder-decoder, R-CED),通过删去池化层、加入跳跃连接的方式优化训练过程.Fu等人[71]提出了一种SNR-Aware(signal to noise ratio aware)的CNN语音降噪模型,并在实际应用中验证了该方法的泛化性.Gao等人[72]采用长短期记忆网络[73](long short-term memory, LSTM)显式学习特定信噪比的中间目标,引入密集连接的渐进学习,将输入以及中间目标的估计拼接起来,再一起学习下一个目标.这种方式缓解了信息丢失的问题,语音可懂度在各种实验噪声下均有提高.

一些学者将GAN应用到了语音降噪领域,GAN中的对抗机制来源于二人博弈的思想,它同时训练2部分模型:生成模型和判别模型,分别用M_G和M_D表示.M_G的目标是生成更加"真实"的样本以欺骗M_D,M_D的目标是更准确地分辨真实样本与M_G生成的样本之间的差异;通过迭代训练,在持续的竞争中共同推动2种模型提高性能,直到M_D无法区分M_G生成的样本与真实样本为止.Michelsanti等人[74]借鉴图像领域的Pix2Pix[75]框架,通过M_G对带噪语音频谱图降噪,M_D用来将M_G生成的降噪频谱与纯净语音频谱区分开,作者取得了与DNN相当的降噪效果.Donahue等人[76]探索了GAN在语音鲁棒性识别中的应用,在频域上应用GAN,提出了FSEGAN(frequency-domain speech enhancement GAN)并在语音鲁棒性识别中相比于传统多风格训练(multi-style training, MTR)有7%的性能提升.

2.2.3 基于信号近似的方法

基于信号近似(signal approximation, SA)的方法是利用神经网络估计掩蔽,并将其作用于带噪语音幅度谱上,得到估计语音的幅度谱。该掩蔽能最小化纯净语音幅度谱与估计语音幅度谱之间的差异:

$$f_{SA}(t,f)=[\hat{f}_{SMM}(t,f)|\boldsymbol{Y}(t,f)|-|\boldsymbol{S}(t,f)|]^2, \tag{3}$$

其中,$\hat{f}_{\mathrm{SMM}}(t,f)$表示对 SMM 的估计,因此 SA 可以看作一个结合了比值掩蔽和频谱映射的目标,其目的是取得最大的信噪比.在此方法中,掩蔽预测值实际上是一个要学习的"隐目标",而掩蔽值作用于带噪语音频谱特征后得到的估计语音是要学习的"真目标",通过这种方式学习到的掩蔽与实际目标是更为相关的.

Huang 等人使用 DNN 与 DRNN(deep RNN)对说话语音进行降噪与分离[77-78],DRNN 是多层 RNN 的堆叠,与 RNN 类似,是一类具有短期记忆能力的神经网络,其神经元既可以接受其他神经元的信息,也可以接受自身的信息,形成具有环路的网络结构,比较适合对语音信号这种序列化数据建模;通过 DRNN 估计出目标语音和干扰语音的掩蔽值,由区分性训练的方式将掩蔽值引入到损失函数中,最小化混合语音重构误差,实验结果相比于 NMF 方法有很大提升.然而,在 RNN 中很容易出现梯度消失和梯度爆炸的问题[79],为缓解这一问题引入了 LSTM,通过门控机制将上下文信息保持在记忆单元中,Weninger 等人[80]使用 LSTM 模型实现信号近似来预测掩蔽值,在时频域内估计误差,在随后的工作中加入了相位信息并应用到了鲁棒性语音识别的任务中[81].

2.2.4　基于端到端的方法

大部分监督性语音降噪是在时频域进行的,近年来,一些学者开始将注意力转移到端到端的解决方式上,即对原始时域波形信号直接进行处理.由于不依赖于频域表示,端到端的方法避免了相位信息丢失以及重构降噪语音时使用带噪语音相位而可能引发的降噪效果不佳的问题;端到端的处理方式可以减少语音信号的处理工序,避免了信号在时频域的来回切换,使得流程更加简化.

Qian 等人[82]提出贝叶斯 WaveNet[83]框架 BaWN(Bayesian WaveNet)用于语音降噪,利用 WaveNet 对原始波形的强大建模能力,将输出正则化到语音空间,显示出贝叶斯框架中语音先验分布的有效性,并取得了较好的泛化性能;随后,Rethage 等人[84]也在 WaveNet 的基础上进行语音降噪,利用非因果扩张卷积来预测一系列目标,而不是单一目标.实验结果表明,该方法优于基于幅度谱的 Wiener 滤波方法;Fu 等人[85-86]提出了全卷积神经网络(fully convolutional neural network, FCN)来对语音进行降噪,他们发现全连接层不易同时映射语音信号的高频分量与低频分量,所以删除了卷积网络的全连接层.作者将神经网络应用于整句语音波形信号,并改进了损失函数,使得语音降噪效果得到改善.Venkataramani 等人[87]提出了一种基于卷积自动编码器的前端变换,用来替代 STFT.该编码器可以自动从数据的原始波形发现数据特定的频域表示,该方法相比于基于 STFT 的方法,取得了更好的性能,可以用于端到端的语音降噪任务中.Pascual 等人[88]提出了基于 GAN 的端到端语音降噪模型,其 M_{G} 是一个全卷积网络,用于对语音进行降噪处理,鉴别器 M_{D} 与 M_{G} 有着同样的结构,它对 M_{G} 生成的波形以及纯净原始信号波形进行判别,并将判别结果反馈给 M_{G}.通过作者的实验,GAN 可以在一定程度上对语音进行降噪,但是在评估指标 PESQ 上略低于 Wiener 滤波.

2.3　结合传统方法与深度学习的方法

并非所有的语音增强方法都是单纯基于神经网络的,一些学者将深度学习的方法与传统方法相结合.Vu 等人[89]将 DNN 与稀疏非负矩阵分解(sparse non-negative matrix factorization, SNMF)结合应用到噪声环境下的自动语音识别(automatic speech recognition, ASR)任务中.如图 5 所示,作者在已标记数据上对语音和噪声基向量进行无监督 SNMF 学习,并进行有监督的 SNMF 特征提取,通过构建神经网络来学习 SNMF 激活系数之间的非线性映射,使降噪信号的对数谱与目标语音的对数谱之间的均方误差最小.Roux 等人[90]将 NMF 扩展为深层结构,并在各种噪声和混响条件下进行测试,取得了较大的性能提升.

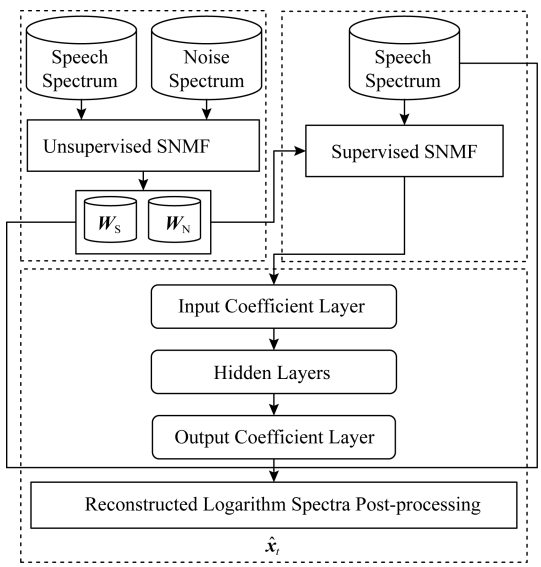


Fig. 5　Combination of DNN and NMF method

图 5　DNN 与 NMF 结合的方法[89]

Yang 等人[91]提出了一种利用 DNN 估计自回归模型(autoregressive model, AR)参数的新方法,训练神经网络学习纯净语音与噪声 AR 模型的参数,利用学习到的 AR 模型参数构造 AR-Wiener 滤波器;采用语音存在概率对 AR-Wiener 滤波器进行了改进,消除了谐波间的残余噪声.Bando 等人[92]最近提出了一种半监督语音降噪方法 VAE-NMF (variational autoencoder NMF),该方法采用了基于变分自编码器(variational autoencoder, VAE)的语音概率生成模型和基于 NMF 的噪声概率生成模型,并在未知噪声下取得了比传统 DNN 监督学习更好的性能.我们对不同信噪比和不同噪声条件下的深度学习算法进行了对比,并比较了他们的 PESQ 和 STOI 性能,如表 3 和表 4 所示:

Table 3 Comparison of PESQ Scores in Deep Learning Based Speech Denoising Methods

表 3 深度学习语音降噪方法的 PESQ 指标对比

Method	Babble			Factory			Experimental Configuration				
	−5 dB	0 dB	5 dB	−5 dB	0 dB	5 dB	Speech Corpus	Training-set Size (Number of Utterance)	Noise Corpus	Noise Types	SNR/dB
AECNN[93]	2.2	2.65	2.95	2.2	2.65	2.95	TIMIT	2 000	NOISEX-92	5	−5,0
CASADNN-pRM[94]	2.26	2.58	2.89	1.81	2.22	2.79	TIMIT	1 500	NOISEX-92	4	−5,0,5
GAN[95]	1.72	2.18	2.62	1.78	2.27	2.68	TIMIT	2 000	NOISEX-92	5	−5,0
GRN[96]	2.16	2.63	2.97				WSJ0 SI-84	6 622	Sound Ideas		−5,−4,−3,−2,−1,0
CRN[97]	2.17	2.44#					WSJ0 SI-84	6 622	Sound Ideas		−5,−4,−3,−2,−1,0
DNN-ORM[98]	1.93#	2.3	2.60#	2.63#	2.8	2.95#	IEEE Corpus	600	NOISEX-92	4	−3,0,3
RSD-CDM[99]	1.88#	2.12	2.39#				IEEE Corpus		Aurora	4	−8,−4,0,4,8
BLSTM-SE[100]	1.62#*	2.10*	2.65#*	1.75#*	2.30*	2.70#*	LibriSpeech	80h (≈ 20 000)	Sound Ideas	138	−6,−3,0,3,6,9
CED[100]	1.55#*	2.12*	2.45#*	1.68#*	2.18*	2.60#*	LibriSpeech	80h (≈ 20 000)	Sound Ideas	138	−6,−3,0,3,6,9
NG-Pix2Pix[74]		1.2	1.49				RSR2015		FUB, Self-synthesized	5	10,20
DAE[59]					2.82	3.19		350		2	0,5,10

Note: The symbol * comes from our estimation of the graph in the reference paper, and the symbol # comes from the rounding of the data in the reference paper.

Table 4 Comparison of STOI Scores in Deep Learning Based Speech Denoising Methods

表 4 深度学习语音降噪方法的 STOI 指标对比

Method	Babble			Factory			Experimental Configuration				
	−5 dB	0 dB	5 dB	−5 dB	0 dB	5 dB	Speech Corpus	Training-set Size (Number of Utterance)	Noise Corpus	Noise Types	SNR/dB
AECNN[93]	2.2	2.65	2.95	2.2	2.65	2.95	TIMIT	2 000	NOISEX-92	5	−5,0
CASADNN-pRM[94]	2.26	2.58	2.89	1.81	2.22	2.79	TIMIT	1 500	NOISEX-92	4	−5,0,5
GAN[95]	1.72	2.18	2.62	1.78	2.27	2.68	TIMIT	2 000	NOISEX-92	5	−5,0
IFFT-DNN[68]	0.61	0.75	0.85	0.65	0.78	0.86	TIMIT	2 000	NOISEX-92	4	−5,0
GRN[96]	0.80	0.89	0.93				WSJ0 SI-84	6 622	Sound Ideas		−5,−4,−3,−2,−1,0
CRN[97]	0.79	0.85#					WSJ0 SI-84	6 622	Sound Ideas		−5,−4,−3,−2,−1,0
DNN-STOI[101]		0.82					WSJ0	20 000	CHiME3, Self-synthesized	6	[−5,10]
DNN-ELC[102]	0.66	0.82	0.90				WSJ0	20 000	CHiME3, Self-synthesized	6	[−5,10]

Continued (Table 4)

Method	Babble			Factory			Experimental Configuration				
	−5 dB	0 dB	5 dB	−5 dB	0 dB	5 dB	Speech Corpus	Training-set Size (Number of Utterance)	Noise Corpus	Noise Types	SNR/dB
DNN-EMSE[102]	0.51	0.59	0.77				WSJ0	20 000	CHiME3, Self-synthesized	6	[−5,10]
DNN-ORM[98]	0.71#	0.80	0.87#	0.86#	0.90	0.92#	IEEE Corpus	600	NOISEX-92	4	−3,0,3
BLSTM-SE[100]	0.58#*	0.76*	0.89#*	0.66#*	0.80*	0.90#*	LibriSpeech	80 h (≈20 000)	Sound Ideas	138	−6,−3,0,3,6,9
NG-Pix2Pix[74]		0.46	0.60				RSR2015		FUB, Self-synthesized	5	10,20
DNN-Two targets[103]	0.64*	0.77*	0.88*				NB Tale	1 932	Aurora, NOISEX-92, Guoning Hu's Collection		−5,0,5,10,15,20

Note: The symbol * comes from our estimation of the graph in the reference paper, and the symbol # comes from the rounding of the data in the reference paper.

3　语音去混响

语音去混响的目标是将混响语音转化为无混响语音,是一项具有挑战的任务.混响是声信号从声源通过多条路径传播到人耳或麦克风(接收器)的过程.接收器接收到的信号中,包括未经过任何障碍物反射而直接到达的语音成分,以及随后到达的混响成分.一般从直达语音到达后算起,50 ms 内到达的混响,称为早期混响,超过 50 ms 到达的称为晚期混响[104-105].相比于晚期混响,早期混响反射次数较少,信号强度较高,与说话人和接收器的位置高度相关;晚期混响在经过多次反射后,强度大致呈指数衰减,与位置无关,并且会改变语音的时间包络,对语音质量的影响较大[106-107].

语音去混响技术可概括为 3 类:1)假设带混响语音由线性系统产生,首先估计声学系统的参数,再得到无混响信号的估计,称作混响消除方法;2)假设带混响语音由加性过程产生,且混响与语音无关,称作混响抑制方法;3)对混响声学系统未知,直接从带混响语音映射到无混响语音,这一类的典型代表是基于深度学习的语音去混响方法[104,106].

3.1　混响消除方法

混响消除方法利用卷积失真模型对信号建模,将纯净语音信号 $s(n)$ 与线性系统冲激响应 $a(n)$ 卷积,再加上噪声 $u(n)$ 形成带混响和噪声的语音 $x(n)$,在时域可表示为

$$x(n)=\sum_{\tau=0}^{L-1}a(\tau)s(n-\tau)+u(n). \tag{4}$$

在不考虑噪声干扰情况下,式(4)在经过傅里叶变换并取幅度值后,可表示成矩阵形式:

$$\boldsymbol{X}=\boldsymbol{A}\boldsymbol{S}, \tag{5}$$

其中,$\boldsymbol{S}$,$\boldsymbol{X}$ 分别表示纯净语音与带混响语音的时频域幅值矩阵,矩阵 $\boldsymbol{A}$ 由冲激响应 $\boldsymbol{a}$ 转换.

早期混响消除法的一个基本思路是对冲激响应求逆,通过混响的逆过程将语音还原.Neely 等人[108]最先开展这方面研究,针对冲激响应恰好是最小相位的情况,设计了一个逆滤波器,在一定程度上消除了冲激响应对语音信号的影响,但在多数情况下冲激响应是非最小相位的,因此该方法有一定的局限性;Wu 等人[109]利用逆滤波器解决早期混响的非平坦频率响应使语音频谱失真的问题,但发现不能去除晚期混响,于是采用谱减法进一步处理,实验表明逆滤波器和谱减法都改善了语音质量;Dong 等人[110]研究如何提升室内公共广播系统的语音清晰度,提出将 Taal 等人[111]的感知失真测量语音增强方法(perceptual distortion measure based speech enhancement)方法与 Kirkeby 等人[112]的快速逆滤波法(fast inverse filtering, FIF)结合,新设计了一种基于 Gammatone 滤波器的 FIF 方法,比原 FIF 方法能进一步减少传输信道的失真,如图 6 所示.有的工作根据式(3)构建 NMF 模型以消除混响.Liang 等人[113]使用 NMF 对纯净无混响语音建模,并推导出一种有效的闭式变分期望最大化算法来估计混响和噪声参数.Mohammadiha 等人[114]提出的方法使用卷积传递函数的非负近似(non-negative approximation of the convolutive transfer function, N-CTF)来同时估计语音信号和 RIR(room impulse

responses)的幅度谱.在 N-CTF 模型中,假设幅度谱中每个频点的 STFT 系数大小是由纯净语音信号的幅度与 RIR 的卷积决定,其优势在于无需对 RIR 相位建模.同时为了利用语音的频谱结构,应用 NMF 对语音的频谱建模;Zhang 等人[115]考虑到在真实环境中,RIR 可能较长而导致对其幅度谱的估计难以收敛,于是在结合 N-CTF 和 NMF 的基础上,分 2 阶段分别处理混响和噪声,缩短了处理时间并提升了性能;Mohanan 等人[116]提出构建非卷积的 NMF 模型,这样将更容易在时域或频域中引入新的约束,以及扩展到有加性噪声的场景.

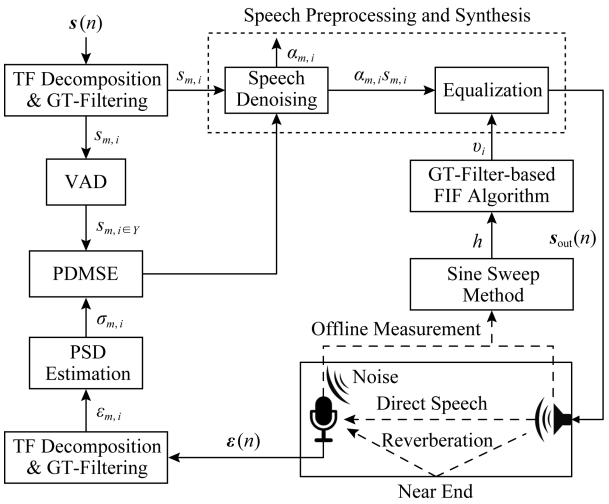


Fig. 6 A method for reverberation elimination

图 6 一种混响消除方法[110]

3.2 混响抑制方法

混响抑制方法利用加性失真模型对信号建模,纯净语音信号 $s(n)$、混响 $r(n)$与噪声 $u(n)$相加形成带混响和噪声的语音 $x(n)$,在时域可表示为

$$s(n)=x(n)+r(n)+u(n). \tag{6}$$

在混响抑制方法中,早期混响因为混响时间极短且对语音质量有益,一般将它假设为纯净语音的一部分.而晚期混响因为失真且混响时间较长,假设其与纯净语音、早期混响无关,是需要被抑制的部分[104,117].

基于以上晚期混响与语音的加性假设和无关假设,语音降噪方法可以应用于去混响.例如 Dionelis 等人[118]提出将调制域的自适应卡尔曼滤波用于单声道语音降噪和去混响,如图 7 所示,该算法需要估计语音对数幅度谱的后验分布,滤波器的更新步骤对语音、噪声和混响之间的非线性关系进行建模,实验证明了算法的有效性.Peng 等人[15]对噪声与混响同时加以抑制,使用广义奇异值分解(generalized singular value decomposition, GSVD)的方法,提出了一种基于约束最小均方误差(constrained minimum mean square error, CMMSE)的线性预测残差估计(linear prediction residual estimator, LPRE)算法,称作 CMMSE-GSVD-LPRE.在含有混响和噪声的实验中,该算法优于谱减法,但仍有混响成分残留,于是他们在线性预测残差域利用了人的听觉掩蔽特性,进一步提升性能.

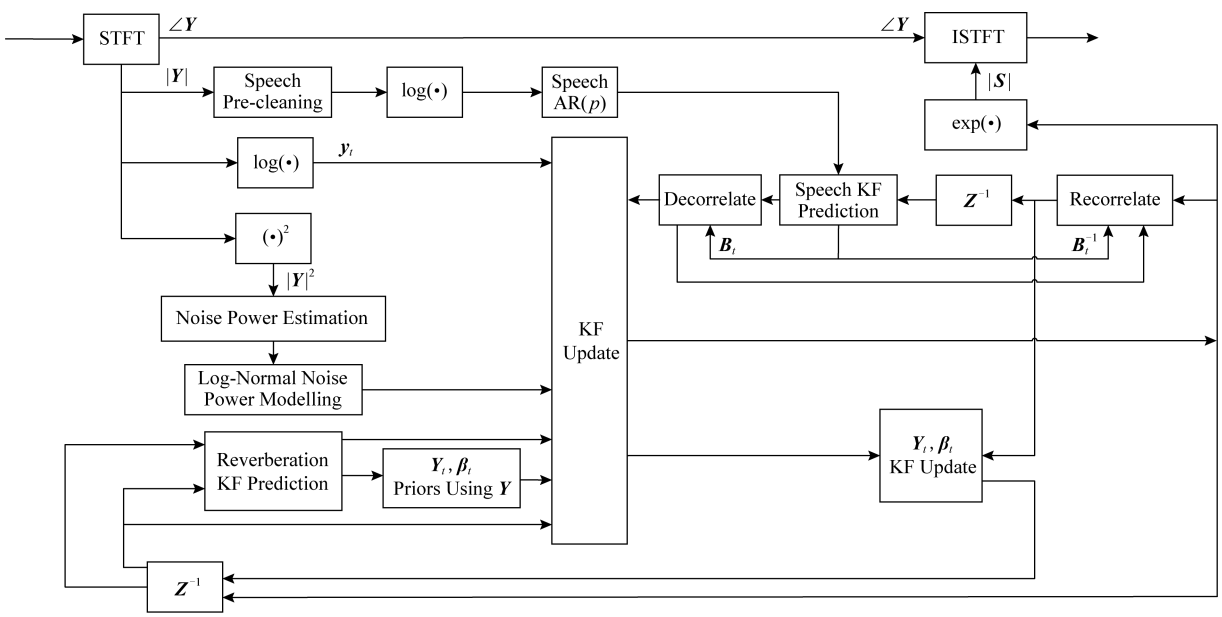


Fig. 7 A method for reverberation suppression

图 7 一种混响抑制方法[118]

带有权重的预测线性误差(weighted linear prediction error, WPE)方法早在2008年被提出[119],是目前应用广泛的混响抑制方法,有不少研究是基于此方法[120-121].虽然其数学模型是基于多声道的,但也能有效地应用到单声道.WPE的基本思路是构造滤波器,使用从倒数第$K+\Delta$帧开始的共K帧语音,估计出当前语音帧的混响,再用当前语音减去混响估计,得到对纯净语音的估计,WPE去混响可表示成:

$$\hat{y}_{t,f}=x_{t,f}-\sum_{\tau=\Delta}^{K+\Delta-1}g_{\tau,f}x_{t-\tau,f}, \tag{7}$$

其中,$\hat{y}_{t,f}$与$x_{t,f}$分别是当前时频点的去混响结果与混响语音,$g_{\tau,f}$是滤波器的权重.WPE假设当前时频点的无混响语音$x_{t,f}$服从复高斯分布,其均值为0且方差为$\lambda_{t,f}$,通过多次迭代过程估计λ与g.为了保证估计的准确性,WPE需要较长的语音,而DNN有处理短语音片段的能力,Kinoshita等人[121]提出用DNN估计未知参数λ,再计算出g,替代了多次迭代估计过程,提升了WPE算法的实时性.Drude等人[120]对比了WPE与基于神经网络的波束成形方法,并将两者结合,取得了比两者独立去混响更好的结果.

3.3　基于深度学习的语音去混响方法

混响消除和抑制方法都对产生混响的信号模型做出假设,估计模型的参数,恢复出纯净语音.还有一类方法不估计信号模型的参数,直接将带混响的语音转换成纯净语音.近年来,这类方法的主要研究方向是用深度学习模型,通过大量数据训练,建立混响语音到纯净语音的非线性映射.目前为止,涌现出的相关研究已经应用了多种神经网络,并根据语音混响特点,结合其他机器学习方法做出创新.

基于深度学习的语音去混响方法在探索初期主要采用DNN.Han等人[122]提出了基于DNN的去混响算法,首先从混响语音中提取出频谱,采用MLP估计纯净语音的耳蜗谱,最后重构语音信号,取得了比非深度网络的方法更好的结果;随后,Wu等人[123]提出混响时间感知模型,将混响时间作为一个控制参数,引入到特征抽取和模型训练阶段,以适当地选择输入的语音帧长和帧移;Zhao等人[124]针对噪声和混响同时存在的场景,分2个阶段建模,第1阶段用DNN估计掩码的方式去除噪声,第2阶段用另一个DNN直接估计频谱的方式去除混响,第1阶段的输出经过特征提取输入到第2阶段的DNN,在训练过程中,2个DNN是分别单独训练,然后再联合训练的;在重构语音阶段,这项工作没有直接使用带噪带混响语音的相位,而是使用Griffin等人[125]提出的时域信号重构技术;实验结果表明,该方法明显优于单阶段方法.

除了DNN以外,也有研究工作使用CNN[126-127],RNN或LSTM[128-132]等深度学习模型.Guzewich等人[127]提出了一个基于CNN的去混响模型,参考了VGG模型[133]基本思路,用大量小卷积核提升神经网络的能力,包含9个卷积层、4个池化层和最后2个全连接层;实验表明该模型比参考的基线模型更好,并且优于Wu等人[123]提出的DNN模型,该模型在说话人识别任务中有效降低了错误率;考虑到早期混响对语音的可懂度有益,而晚期混响则会降低可懂度[134],Zhao等人[132]提出用LSTM神经网络对混响语音中的长期依赖信息建模,估计出晚期混响成分并从混响语音中减去,而非直接估计出无混响语音;Yu等人[135]提出一个隐含层有CNN和LSTM结构的神经网络模型,用于语音关键词检测的前端去噪和去混响;在Zhao等人[136]提出的神经网络模型中,使用卷积层学习时频域中的局部模式,再用双向循环连接层对相邻语音帧间的动态相关性建模,最后用全连接层估计纯净语音的频谱;Santos等人[129]采用了相似的建模思路,使用了卷积层和循环连接层构建神经网络,还在输入层、隐含层及输出层间加入了残差连接.

值得注意的是,近年有工作开始使用GAN的对抗策略训练去混响模型.Ernst[126]借鉴了全卷积网络在图像处理领域的成功经验,用频谱图表示混响语音信号,使用U-Net[137]学习混响语音频谱到无混响语音频谱的映射.他们利用了CGAN(conditional GAN)[74]训练U-Net,这是CGAN首次应用于去混响.Li等人[138]使用了对抗训练策略,其中语音增强模型是一个包含卷积层、双向LSTM层和全连接层的神经网络,与之对抗的判别器模型同样包含卷积层、双向LSTM层和全连接层.

有的工作将深度学习与其他机器学习方法结合.Lee等人[139]提出的去混响模型包含多个DAE,根据集成学习的思想,每个DAE处理特定声学环境中的语音,用融合函数将各DAE处理结果整合得到去混响语音;刘斌等人[128]提出用LSTM神经网络去混响,发现LSTM估计的纯净语音过于平滑而降低了语音信号的感知质量,于是采用NMF对

LSTM 的输出做后处理,有效抑制了过平滑问题;Chien 等人[140]设计了一种由矩阵分解方法构建的神经网络层,称作 STNF(spectro-temporal neural factorization)层,用于提取语音中的时频域特征,STNF 的前向计算和反向传播都可视作矩阵分解过程,实验表明 STNF 层相比于全连接层的去混响效果更好;Raikar 等人[141]提出用最大后验估计建模方法,将独立的去混响和降噪过程结合到一起,其中降噪部分使用的是 SEGAN 模型[88],其输入是去混响的结果,其去噪的结果又会提升混响卷积矩阵的估计准确率.表 5 中是不同去混响方法在不同的混响时间(T_{60})下的 PESQ,STOI 以及语音混响调制能量比(speech-to-reverberation modulation energy ratio, SRMR)指标统计.SRMR 是一种非侵入式无需纯净语音进行计算的指标,用于评估语音质量与可懂度,它的值越高表示去混响的效果越好.

Table 5 Comparison of Scores in Speech Dereverberation Methods

表 5 语音去混响方法的指标对比

Method	T_{60}/s									Experimental Configuration			
	0.3			0.6			0.9						
	PESQ	STOI	SRMR	PESQ	STOI	SRMR	PESQ	STOI	SRMR	Speech Corpus	Training-set Size (Number of Utterance)	Training T_{60}/s	RIR
Cross-Corpora CDNN[127]	2.710*	0.930*		2.620*	0.920*		2.480*	0.900*		TIMIT	46 200	0.1 – 1	10
RTA-DNN[123]	3.160			2.820			2.590			TIMIT	46 200	0.3 – 0.9	3
PCMMSE-GSVD-LPRE[15]			3.600#			4.420#			3.410#	TIMIT	9 000	0.1 – 1	10
N-CTF+NMF[114]	3.150#			2.640#						TIMIT	600	0.3 – 0.9	3
ASSE+LSTM-IRM[130]	3.410										75 000	0.3	48
TDSE[135]	3.630									self-recorded	45 000	0 – 0.6	15 000
Derogate[138]	2.760		5.420							WSJ0	38 328	0 – 0.7	3
CARNN[129]	2.790*	0.870*	3.200*	2.570*	0.840*	2.700*	2.420*	0.820*	2.600*	IEEE Corpus	35 150	0.2 – 2.0	740
Two-stages TDR[124]	2.573	0.876		2.433	0.863		2.223	0.831		IEEE Corpus	15 000	0.3 – 0.9	10
IDEA[139]	2.470			2.306			2.147			MHINT	2 250	0.3 – 1.0	3

Note: The symbol * comes from our estimation of the graph in the reference paper and the symbol # comes from the rounding of the data in the reference paper.

4 实验与评估

本节介绍了在语音增强实验及评估中的一些必要内容,主要包括数据集、特征和评估指标.语音增强实验都需要根据实验目的准备特定的数据集,并使用数据集对算法的有效性及性能进行检验.对大多数学习算法而言,进行学习前需要先从数据中提取更易于学习的特征,因为直接学习原始数据往往是比较困难的.此外,对实验结果进行评估也是必要的,一方面可以从评估分数判断实验结果的好坏,另一方面不同算法的实验结果很难直接进行比较,在进行评估后就可以方便地对比每个算法的性能.

4.1 数据集

数据集是语音增强实验的关键部分,作用于模型训练、验证、测试的整个过程.通常,数据集的大小和数据的多样性对模型的性能及泛化能力有很大影响.在语音增强中,数据的多样性包括语料的多样性、噪声的多样性、信噪比的多样性、说话人的多样性.经实验证明[61,142],在一定范围内,随着数据集数据量的增加和数据多样性的提高,语音增强模型的噪声、信噪比、说话人甚至是语言的泛化能力都有所提高.

在语音增强中音频数据集一般可以分为纯净语音数据集、噪声数据集以及带噪语音数据集,实验大多会使用公开数据集,但此外一些有特殊需求的研究者会自行构建数据集.当实验需要用到带噪语音时,可以使用已有的带噪语音数据集,也可以使用语音噪声混合工具,如滤波与噪声添加工具(filtering and noise adding tool, FaNT)[143]将纯净语音和噪

声混合,通过调整参数得到特定信噪比的带噪语音.在进行去混响实验时,主要通过将语音信号与不同混响时间的房间脉冲响应 RIR 进行卷积得到混响语音信号.语音增强中常见的音频数据集如表 6 所示:

Table 6　Common Voice and Noise Datasets

表 6　常见语音和噪声数据集

Dataset	Language	Type	Dataset Size		
			Total Duration Time/h	Number of Sentences	Number of Noise Types
TIMIT[144]	English	Speech	5.5	6 300	
WSJ0[145]	English	Speech	42		
IEEE Corpus[146]	English	Speech		720	
Voice Bank[147]	English	Speech	300		
Librispeech[148]	English	Speech	1 000		
TSP	English	Speech		1 444	
WSJ1	English	Speech	162		
Aishell	Chinese	Speech	178		
THCHS-30	Chinese	Speech	30		
ATR[149]	Japanese	Speech			
NB Tale	Norwegian	Speech			
NOIZEUS	English	Noisy		5 760	
Aurora[150]	English	Noisy		86 950	
CHiME-2[151]	English	Noisy			
CHiME-3[152]	English	Noisy		1 600	
ChiME-4	English	Noisy			
MHINT[153]	Chinese	Noisy			
NOISEX-92[154]		Noise			15
DEMAND[155]		Noise			18
RSG-10[156]		Noise			38
Guoning Hu's Collection		Noise			20
100-Nonspeech		Noise			20
SPIB		Noise			15

4.2　特　征

语音信号是一种非平稳、时变的随机过程,很难直接对其学习,因此往往需要进行特征提取,而提取不同的特征会对增强性能有很大的影响.数十年来,为提高语音质量及可懂度,学者们提出了多种语音特征,这些特征都有各自的优势和不足.在单声道语音增强的早期研究中,主要使用基于基音的特征[157]和幅度调制谱(amplitude modulation spectrum, AMS)[158],这些特征提取过程相对简单,但表示能力不足.接着逐步提出了更多单声道特征,包括梅尔倒谱系数(mel-frequency cepstral coefficient, MFCC)[159]、感知线性预测(perceptual linear prediction, PLP)[160]、相对频谱表示(representations relative spectra, RASTA-PLP)[161],这些特征虽然在一定程度上提高了语音增强性能,但单个特征还是难以取得很好的效果.针对这一问题,Wang 等人[159]使用 Group Lasso 特征选择器,得到了 1 组互补的特征组合,包括 AMS,RASTA-PLP,MFCC,这个特征组合在多种条件下相对单个特征显著地提高了增强性能,在很多研究中得到了应用.同时,短时傅里叶变换幅度谱和短时傅里叶变换对数幅度谱也常用于语音增强,且由于高频部分幅度较小,故对数幅度相对幅度更能凸显高频成分.然而有研究[162]发现,短时傅里叶变换幅度谱的性能比短时傅里叶变换对数幅度谱略好.此外,学者们还在 Gammatone 滤波的基础上提出了 Gammatone 特征(Gammatone feature,

GF)、Gammatone 倒谱系数(Gammatone frequency cepstral coefficient, GFCC)[163]、Gammatone 调制频谱(Gammatone frequency modulation spectral based cepstral, GFMC)[164].随后,又有学者对已有的特征进行研究与改进,在 MFCC 的基础上提出了 Delta 倒谱系数(delta spectral cepstral coefficients, DSCC)[165]、相对自相关序列 MFCC(relative auto-correlation sequence MFCC, RAS-MFCC)[166]、自相关序列 MFCC(auto-correlation sequence MFCC, AC-MFCC)[167]、相位自相关 MFCC(phase auto-correlation MFCC, PAC-MFCC)[168].陈纪同等人[169]提出了多分辨率听觉谱(multi-resolution cochleagram, MRCG)特征,它同时计算出 4 种不同分辨率的倒谱,从而可以同时提取到局部性信息和整体性信息,现已成为最常用的特征之一.

下面对一些常见的特征进行介绍:

1) MRCG

MRCG 由 4 种不同分辨率的倒谱组成,高分辨率倒谱捕捉局部信息,3 个低分辨率倒谱捕捉不同尺度的上下文信息.为得到 MRCG,首先将信号进行 64 通道的 Gammatone 滤波得到一个听觉谱,称作 CG1,并在每个时频单元进行取对数操作;类似地,可用 200 ms 的帧长和 10 ms 的帧移计算得到第 2 个听觉谱,称作 CG2;其次使用一个长为 11 帧和宽为 11 频带的方形窗对 CG1 进行平滑,得到第 3 个听觉谱,称作 CG3;和 CG3 的计算相似,使用 23×23 的方形窗对 CG1 进行平滑,得到第 4 个听觉谱,称作 CG4;串联 CG1, CG2, CG3, CG4 得到一个 64×4 的向量,即为 MRCG.

2) MFCC

MFCC 即梅尔倒谱系数,首先对输入信号作分帧操作,经验上取 10～30 ms 帧长,5～15 ms 帧移;其次对每一帧进行加窗处理,一般使用汉明(Hamming)窗;然后进行 FFT 计算得到对应的频谱,再将频谱通过 Mel 滤波器组转换为梅尔域,最后在 Mel 频谱上进行倒谱分析,得到 MFCC.

3) GF

该特征由 Gammatone 听觉滤波得到,首先用 Gammatone 滤波器组对信号进行处理,然后对每个滤波输出以 100Hz 的频率进行采样,最后采样结果通过立方根操作进行幅度压缩得到 GF.

4) GFCC

GF 特征一般由 64 个频率成分组成,但在实际系统中由于 GF 特征矢量的维度比较大,计算量也较大.此外,由于相邻的滤波器通道有重叠的部分,导致 GF 特征矢量相互之间存在相关性.因此为减小 GF 特征矢量的维度及相关性,对每一个 GF 特征矢量进行离散余弦变换(discrete cosine transform, DCT)得到 GFCC.实验表明,前若干维及最后若干维的 GFCCs 系数对语音的区分性能较大,因此一般取前 26 维的 GFCC 系数作为特征.

5) PLP

PLP 即感知线性预测系数,它能够最大限度地消除说话人不同带来的影响,同时可以留下关键的共振峰结构,由于该特征与语音内容比较相关,因此常用于语音识别.

4.3 评估指标

评估实验结果需要设定评估指标,不同的指标从不同角度对实验结果进行评分.语音增强任务有多种评估指标,这些指标按评估方法可以分为主观方法和客观方法.主观方法的评估主体为人,以人耳感受为判别标准,带有一定的主观因素;客观方法是指计算机直接以一定的计算方法来为语音评分,在实验中多采用客观方法.从评估目标级别的角度可分为信号级别和感知级别,信号级别的指标目的是量化信号增强或干扰降低的程度,如信噪比(signal to noise ratio, SNR);而感知级别的指标更关注语音增强对于语音的可懂度和感知质量的提高,如 PESQ, STOI.表 7～9 中分别列举了语音增强中的客观指标、主观指标以及语音去混响的指标:

Table 7 Speech Enhancement Objective Evaluation Index

表 7 语音增强客观评估指标

Metric Abbreviation	Full Name
PESQ	Perceptual Evaluation of Speech Quality
STOI	Short-Time Objective Intelligibility
ESTOI	Extended Short-Time Objective Intelligibility
segSNR	Segmental Signal to Noise Rate
SDR	Signal to Distortion Rate
SIR	Signal to Interference Rate
SAR	Signal to Artifacts Rate
LLR	Logarithmic Likelihood Ratio
WSSD	Weighted Spectral Slope Distance
SII	Speech Intelligibility Index
exSII	Extended Speech Intelligibility Index
SRT	Speech Recognition Threshold
POLQA	Perceptual Objective Listening Quality Assessment
PSNR	Phase Signal to Noise Rate

Table 8　Speech Enhancement Subjective Evaluation Index
表 8　语音增强主观评估指标

Metric Abbreviation	Full Name	Value Range
MOS	Mean Opinion Score	[1,5]
MUSHRA	Multi Stimulus Test with Hidden Reference and Anchor	[0,100]

Table 9　Speech Dereverberation Evaluation Index
表 9　语音去混响评估指标

Metric Abbreviation	Full Name
SRMR	Speech Reverberation Modulation Energy Ratio
CD	Cepstrum Distance
SRR	Speech Reverberation Ratio

对 4 个常用的评估指标进行详细介绍:

1) 平均主观意见分(mean opinion score, MOS)

MOS[170]常用于衡量通信系统语音质量,由人对语音质量的真实反映得出,但其受测试条件的限制和测试人员主观因素的影响,且不满足实时性要求.由不同人分别对原始语料和经过系统处理后失真的语料进行主观感觉对比,最后求平均得到 MOS 值,MOS 值取值范围为 1～5 分.

2) PESQ

PESQ 指标[171]的设计目的是评估电话网络和编解码的语音质量,与 MOS 高度相关,侧重于评估语音的清晰度.它是感知分析测量系统(perceptual analysis measurement system, PAMS)和感知语音质量增强版 PSQM99(perceptual speech quality measure 99)集成的结果,应用范围广泛,包括模拟连接、编解码器、报文丢失、可变延迟.同时它是国际电信联盟电信标准化部门(ITU-T) P.862 建议书提供的客观 MOS 评估方法.PESQ 值介于－0.5～4.5 之间,但是对于正常的主观测试材料,该值介于 1.0(差)和 4.5(无失真)之间.在极高的失真度下 PESQ 值可能会低于 1.0,但这种情况非常少见.

3) STOI

STOI 指标由 Taal 等人[172]于 2011 年提出,它是基于纯净语音与带噪语音的时间包络相关系数计算得到,在实验中表现出与语音可懂度的高度相关性.计算 STOI 包括 3 个步骤:首先去除静音帧(silent frames),即删除能量少于 50 dB 的帧,因为静音对语音可懂度没有影响;其次,对信号进行基于 DFT 的 1/3 倍频带分解,汉明窗的长度为 25 ms,256 个频率覆盖,频率范围为 0～5 kHz;最后通过相关过程计算输出 STOI.STOI 取值范围为[0,1],且与主观语音可懂度正相关,即值越大表示语音可懂度越好.

4) 分段信噪比(segmental SNR,segSNR)

segSNR 指标主要用于语音增强、语音编码后的测试.由于语音信号是非平稳信号,有很多低能量和高能量区域,并且这些区域与语音的理解密切相关.segSNR 不计算整段语音的信噪比,而是计算短期(15～20 ms)SNR 的平均值,因此能够反映语音的局部失真水平.与 SNR 相比,segSNR 与 MOS 的相关度更高.

5　问题与挑战

在研究者们的努力下,传统方法或深度学习方法的语音增强算法性能都得到了一定提高.但语音增强领域仍存在着一些问题和挑战,包括低信噪比环境下的语音增强问题、增强算法的泛化问题、相位失真问题、测度不匹配问题等.

5.1　低信噪比环境下的语音增强问题

在低信噪比环境中实现有效且稳定的语音增强仍然面临着挑战.在－5 dB 环境下,语音功率不及噪声功率的 1/3,语音幅度常常只有噪声幅度的一半.短时傅里叶变换后,幅度谱以噪声为主导,使得一些基于掩蔽的模型失去了优势,常用的 IBM 会把噪声与语音混合的部分划分为噪声而全部过滤,这种情况下基于掩蔽的模型的效果往往不如基于映射的模型.

面对低信噪比条件下的复杂环境,PL(progressive learning)模型及其与多任务学习和集成学习结合的方法进入了研究者的视野[72,173].PL 模型与普通模型的差别是它把一个学习目标拆分为多个子目标,每个子目标相较前一个目标更加接近最终目标.如图 8 所示,处理 SNR 为 0 dB 的信号的过程可以拆分为先达到 10 dB、再到 20 dB、最后获得目标纯净信号 3 个阶段.实验证明,PL 模型比一般模型

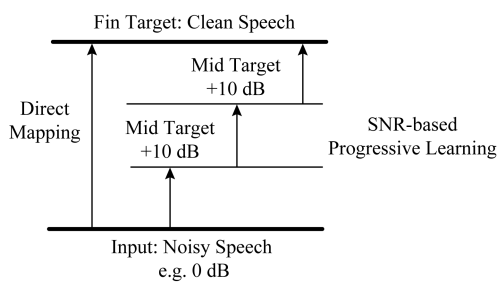


Fig. 8　The PL model for speech enhancement[174]
图 8　语音增强的 PL 模型[174]

更加适合训练海量数据或复杂特征.一种解释是一般模型训练海量数据时,随着训练数据的增加,模型发生了灾难性的遗忘,丢失了之前学到的部分信息.复杂的低信噪比环境下,一般模型也更容易受到影响.而 PL 模型的结构可以使之按阶段保留过去学习到的信息,最后把每个阶段的信息集成到对最终目标的训练中去.因此,在低信噪比或多信噪比环境下,PL 模型可以学习并保留更多特征,泛化性更强.然而,如何选择中间阶段的训练目标是 PL 模型要解决的问题,简单地把训练目标指定为一个固定 SNR 的语音,可能无法发挥模型真正的效能.而在结合多任务学习的 PL 模型中,如何选择训练目标也是一个问题.研究者可以探索一种产生对信噪比环境自适应的阶段目标算法,也可以选择其他的评估指标.

在结合多任务的模型中,模型使用了不同滤波方法提取的声音特征,MFCC 和 GFCC 是 2 种提取声音特征的方式[175],提取后的特征会存在相似或者不同的地方,研究者可能需要选择具有互补特征的训练目标.Fu 等人[71]将 *SNR* 感知结构和语音增强模型相结合,提出了 2 个基于 CNN 的模型,它们在低信噪比条件下取得了更好的效果.前一个模型学习环境中的 *SNR* 级别,在目标函数中加入环境真实的 *SNR* 值,形成一个多任务学习模型.模型在降噪的同时还会判断环境的 *SNR*,以此适应不同的环境;后一个模型先评测环境的 *SNR*.然后根据不同的 *SNR*,选择不同的降噪模型.实验表明,这 2 个模型性能都优于简单 CNN 模型,这说明对于不同的 *SNR* 环境,可以通过加入 *SNR* 评测的方法来提高模型能力.而且实验中还发现后一个模型在 12 dB 和 −12 dB 的 *SNR* 环境下测得的一些指标优于前一个模型,这意味着对应不同 *SNR* 环境使用不同的语音增强模型可能得到更好的结果.

5.2 模型算法泛化问题

基于深度学习的语音增强模型在面对未知环境时,性能会明显恶化.模型的泛化能力不良一直是个难题.语音增强算法的泛化能力可以分为 3 个方面:对未知种类噪声的泛化能力、对未知信噪比环境的泛化能力和对未知说话人的泛化能力.一种简单有效提高模型泛化能力的方法是在大量不同的噪声数据集上训练模型,而且使用 RNN 模型比 DNN 模型更加有优势[96].近几年,Park 等人[176]提出了基于 CNN 编码的语音增强模型,在未知噪声和未知信噪比环境下表现较好.同时,利用编码 CNN 或扩张 CNN 模型也能提高对未知说话人语音增强的能力[96-97].

ASAM[177]提供了另外一种提升增强模型对噪声的泛化能力的思路.ASAM 是一个利用注意力机制和长期记忆的语音分离模型,它利用双向 LSTM 对混合语音和纯净语音的幅度谱作高维映射.再将纯净语音幅度谱的映射融合为一个向量,表示为纯净语音的特征,存入长期记忆中.然后利用该段记忆来关注混合语音中属于同一说话人的映射的向量.长期记忆结构中存在一个存储空间来临时保存未知说话人语音的记忆.这是一个语音分离模型,但可以把要移除的语音替换作噪声.在测试阶段,把捕获的不含语音的未知噪声看作未知语音输入模型,将其特征存入模型的长期记忆中.这类似一种实时获取噪声特征的方法.此后可以利用不同噪声的特征结合语音特征来增强语音.

5.3 相位失真问题

目前常用的基于深度学习的语音增强过程是先对带噪语音计算短时傅里叶变换得到幅度谱和相位谱,再对幅度谱进行处理,最后将产生的幅度谱与原始带噪信号的相位信息合成纯净语音.但是近些年,研究者开始注意到相位信息在语音增强中的重要性.

除了利用相位信息的掩蔽层的模型[48],研究者探索更好的方法去使用带噪信号的相位重构纯净语音信号的相位.在频域的无监督语音增强的相位重构方法中,有 2 类方法:基于振幅的方法和基于模型的方法.基频法是一种基于模型的方法,最近研究者提出利用基频的方法[178-185].短时傅里叶变换相位改良法[182]是一种先进的相位重构方法,但该方法会引入额外的蜂鸣声.而 Wakabayashi 等人[185]利用了相位失真特征,抑制了额外的声音,在 PESQ 上表现超过短时傅里叶变换相位改良法[181],但在 STOI 指标上没有有效地提高.

一些研究者直接在时域上利用 CNN 处理带噪语音[84,93],这样避免了原始带噪信号的相位的使用,提升了一定的模型性能.但是这种做法只将时域上的信息输入神经网络,未利用神经网络处理频域信息,或忽略了信号在频域上的信息,这样可能丢失了一部分必要的纯净语音信息.将模型结合多任务学习的方法可能会有进一步提高.

5.4 语音增强算法测度不匹配问题

语音增强的一个目标是增加语音的可懂度,把错字率(word error rate, WER)看作评估语音增强算法能力的指标可能更为直接.但这种做法要结合

语音识别系统的测试或人工识别测试,评估难度较大.简单地计算增强语音的 SDR,SIR,SAR 指标可以避免语音识别中繁杂的流程,但同时这些指标存在与语音可懂度的相关度不够的问题.于是后来出现了一些匹配人类听觉感知方法的指标,如 STOI.

同时,不匹配的问题也存在于深度学习增强算法所常用的损失函数 MSE(mean-square error).一个好的损失函数可以提高模型的性能.MSE 简单地计算预测语音和正确语音波形或幅度谱的欧氏距离,有时不能完全反映增强语音的质量.因此,出现了新的基于不同的语音评估指标的损失函数.STOI 是目前评估增强语音可懂度的重要指标,它接近人类评估语音方式.但一般使用的损失函数 MSE 与这种方式不匹配,在优化模型时不一定能改善 STOI[186].如何改良损失函数以匹配 STOI 的运算方式是最近的一个研究点.有研究者以提高语音有限的 SNR 为目标来训练模型,却取得了更好的效果,由此发现人类对语音质量的评估与损失函数 MSE 存在不匹配问题[103].Zhao 等人[186]提出了以 STOI 指标为训练目标的损失函数:

$$\mathcal{L}(m)=(1-f(\boldsymbol{X}_m^{24},\boldsymbol{Y}_m^{24}))^2+\lambda\|\boldsymbol{X}_m^{24}-\boldsymbol{Y}_m^{24}\|_{\mathrm{F}}/24,\quad(8)$$

其中,$f(\cdot)$表示 STOI 的计算函数,取值范围在 0 和 1 之间.$\boldsymbol{X}$ 代表带噪信号的语谱图,$\boldsymbol{Y}$ 代表纯净信号的语谱图.$\|\cdot\|_{\mathrm{F}}$ 代表 Frobenius 范数.它对原 STOI 算法作了调整,使之匹配经 STFT 提取的信息.损失函数以最大化 STOI 指标为目标,同时也用语谱图的差值作微调.使用以 STOI 为目标的损失函数有效地提高了模型在 STOI 指标上的性能.然而,模型在其他指标上的性能却没有明显提高.如表 10 所示,PESQ 在 MSE 上的性能仍高于基于 STOI 的损失函数.而且,由于 PESQ 指标的计算较为复杂,将 PESQ 加入损失函数较为困难,如何有效地优化 PESQ 仍是一个可以探索的问题.此外,Venkataramani 等人[87]使用了一种以优化 SDR 为目标的损失函数.对照传统的 MSE,研究者用基于 SDR 的损失函数在 SDR,SIR,SAR 指标上做了实验.从实验结果中可以发现,基于 SDR 的损失函数的指标实验结果较 MSE 有一些提高,同时指标结果的方差更小,这意味着它们更加稳定[87].通过引入指标来优化算法损失函数的一个问题是通常只加入对某一单个指标的运算,如计算 STOI 或 SDR,而没有同时顾及多个指标.基于 STOI 的损失函数可以显著提升 STOI 指标,但没有改善 PESQ 指标.研究者可以设计一种以多指标为目标的更加符合人类听觉的优化方法.

Table 10　Performance Comparison of Models Using STOI and MSE Loss Functions

表 10　使用 STOI 和 MSE 损失函数的模型性能比较[86]

Metric	Loss Function	*SNR*					Average
		−12 dB	−6 dB	0 dB	6 dB	12 dB	
STOI	STOI	**0.562**	**0.699**	**0.814**	**0.888**	**0.931**	**0.779**
	MSE	0.496	0.647	0.780	0.864	0.909	0.739
PESQ	STOI	1.434	1.608	1.877	2.205	2.587	1.942
	MSE	**1.536**	**1.754**	**2.078**	**2.481**	**2.909**	**2.152**

Note: The bold indicates better performance under the same metric and SNR.

6　结束语

语音识别被认为是人工智能未来发展的重要方向之一,而语音增强是其中一项核心关键技术,此外它也能应用于语音通话、电话会议、场景录音、军事窃听和听力辅助等场景,因此具有重要的理论研究与实际应用价值.本文从方法、数据集、特征、评估指标等方面,对单声道语音增强(包括降噪与去混响)研究工作的发展现状进行了全面调研和深入分析,并对该工作面临的重要挑战和关键问题进行了总结.尽管国内外研究人员已经提出了多种单声道语音增强方法,深度学习的引入也为该领域研究带来了新的突破,但已有工作还存在泛化性差、相位失真、测度差异等问题,特别是在低信噪比环境下的应用效果还很不理想,所以这仍是一个充满挑战、值得研究的领域.

参　考　文　献

[1] Benesty J, Makino S, Chen Jingdong. Speech Enhancement [M]. Berlin: Springer Science & Business Media, 2005

[2] Tu Jingxian, Xia Youshen, Zhang Songchuan. A complex-valued multichannel speech enhancement learning algorithm for optimal tradeoff between noise reduction and speech distortion [J]. Neurocomputing, 2017, 267: 333-343

[3] Araki S, Hayashi T, Delcroix M, et al. Exploring multi-channel features for denoising-autoencoder-based speech enhancement [C] //Proc of the 40th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2015: 116-120

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

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

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

[7] Gustafsson H, Nordholm S E, Claesson I. Spectral subtraction using reduced delay convolution and adaptive averaging [J]. IEEE Transactions on Speech and Audio Processing, 2001, 9(8): 799-807

[8] Hu Yi, Bhatnagar M, Loizou P C. A cross-correlation technique for enhancing speech corrupted with correlated noise [C] //Proc of the 26th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2001: 673-676

[9] Zhong Xinmei, Dai Yunzhong, Dai Yong, et al. Study on processing of wavelet speech denoising in speech recognition system [J]. International Journal of Speech Technology, 2018, 21(3): 563-569

[10] Chen Jingdong, Benesty J, Huang Yiteng, et al. New insights into the noise reduction Wiener filter [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(4): 1218-1234

[11] Paliwal K K, Basu A. A speech enhancement method based on Kalman filtering [C] //Proc of the 12th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 1987: 177-180

[12] Gabrea M, Grivel E, Najun M. A single microphone Kalman filter-based noise canceller [J]. IEEE Signal Processing Letters, 1999, 6(3): 55-57

[13] Wang Yu, Brookes M. Model-based speech enhancement in the modulation domain [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(3): 580-594

[14] Andersen K T, Moonen M. Robust speech-distortion weighted interframe Wiener filters for single-channel noise reduction [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(1): 97-107

[15] Peng Renhua, Tan Zhenghua, Li Xiaodong, et al. A perceptually motivated LP residual estimator in noisy and reverberant environments [J]. Speech Communication, 2018, 96: 129-141

[16] Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator [J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1984, 32(6): 1109-1121

[17] 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, 1985, 33(2): 443-445

[18] Martin R. Speech enhancement using mmse short time spectral estimation with Gamma distributed speech priors [C] //Proc of the 25th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2002: I-253-I-256

[19] Faraji N, Kohansal A. Mmse and maximum a posteriori estimators for speech enhancement in additive noise assuming a t-location-scale clean speech prior [J]. IET Signal Processing, 2018, 12(4): 532-543

[20] Jia Hairong, Wang Weimei, Wang Dong, et al. Speech enhancement using modified MMSE-LSA and phase reconstruction in voiced and unvoiced speech [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2019, 33(2): 1958002-1-1958002-16

[21] Stahl J, Mowlaee P. A pitch-synchronous simultaneous detection-estimation framework for speech enhancement [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(2): 436-450

[22] Loizou P C. Speech Enhancement: Theory and Practice [M]. Boca Raton, FL: CRC, 2007

[23] Hu Yi, Loizou P C. A generalized subspace approach for enhancing speech corrupted by colored noise [J]. IEEE Transactions on Speech and Audio Processing, 2003, 11(4): 334-341

[24] Jensen S H, Hansen P C, Hansen S D, et al. Reduction of broad-band noise in speech by truncated qsvd [J]. IEEE Transactions on Speech and Audio Processing, 1995, 3(6): 439-448

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

[26] Srinivas N, Pradhan G, Shahnawazuddin S. Enhancement of noisy speech signal by non-local means estimation of variational mode functions [C] //Proc of the 19th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2018: 1156-1160

[27] Pallavi P, Rao C V R. Phase-locked loop (PLL) based phase estimation in single channel speech enhancement [C] //Proc of the 19th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2018: 1161-1164

[28] Barysenka S Y, Vorobiov V I, Mowlaee P. Single-channel speech enhancement using inter-component phase relations [J]. Speech Communication, 2018, 99: 144-160

[29] Saleem N, Irfan M. Noise reduction based on soft masks by incorporating SNR uncertainty in frequency domain [J]. Circuits Systems and Signal Processing, 2018, 37(6): 2591-2612

[30] Hinton G, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507

[31] Dahl G E, Yu Dong, Deng Li, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 30-42

[32] Hinton G, Deng Li, Yu Dong, et al. Deep neural networks for acoustic modeling in speech recognition [J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97

[33] Cireşan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification [C] //Proc of the 25th Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2012: 3642-3649

[34] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning [C] //Proc of the 25th Int Conf on Machine Learning. New York: ACM, 2008: 160-167

[35] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436

[36] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[37] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C] //Proc of the 28th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2014: 2672-2680

[38] Kim G, Lu Yang, Hu Yi, et al. An algorithm that improves speech intelligibility in noise for normal-hearing listeners [J]. The Journal of the Acoustical Society of America, 2009, 126(3): 1486-1494

[39] Han Kun, Wang Deliang. Towards generalizing classification based speech separation [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(1): 168-177

[40] Virtanen T. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(3): 1066-1074

[41] Mohammadiha N, Smaragdis P, Leijon A. Supervised and unsupervised speech enhancement using nonnegative matrix factorization [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(10): 2140-2151

[42] Chung H, Badeau R, Plourde E, et al. Training and compensation of class-conditioned NMF bases for speech enhancement [J]. Neurocomputing, 2018, 284: 107-118

[43] Healy E W, Yoho S E, Wang Yuxuan, et al. An algorithm to improve speech recognition in noise for hearing-impaired listeners [J]. The Journal of the Acoustical Society of America, 2013, 134(4): 3029-3038

[44] Divenyi P. Speech Separation by Humans and Machines [M]. Berlin: Springer, 2005: 181-197

[45] Naik G R, Wang Wenwu. Blind Source Separation [M]. Berlin: Springer, 2014: 349-368

[46] Liang Shan, Liu Wenju, Jiang Wei, et al. The optimal ratio time-frequency mask for speech separation in terms of the signal-to-noise ratio [J]. The Journal of the Acoustical Society of America, 2013, 134(5): EL452-EL458

[47] Wang Yuxuan, Narayanan A, Wang Deliang. On training targets for supervised speech separation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1849-1858

[48] Erdogan H, Hershey J R, Watanabe S, et al. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks [C] //Proc of the 40th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2015: 708-712

[49] Williamson D S, Wang Yuxuan, Wang Deliang. Complex ratio masking for monaural speech separation [J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2016, 24(3): 483-492

[50] Wang Yuxuan, Wang Deliang. Cocktail party processing via structured prediction [C] //Proc of the 25th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2012: 224-232

[51] Wang Yuxuan, Wang Deliang. Boosting classification based speech separation using temporal dynamics [C] //Proc of the 13th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2012: 1528-1531

[52] Wang Yuxuan, Wang Deliang. Towards scaling up classification-based speech separation [J]. IEEE Transactions on Audio, Speech and Language Processing, 2013, 21(7): 1381-1390

[53] Narayanan A, Wang Deliang. Ideal ratio mask estimation using deep neural networks for robust speech recognition [C] //Proc of the 38th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2013: 7092-7096

[54] Narayanan A, Wang Deliang. Investigation of speech separation as a front-end for noise robust speech recognition [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(4): 826-835

[55] Madhu N, Spriet A, Jansen S, et al. The potential for speech intelligibility improvement using the ideal binary mask and the ideal Wiener filter in single channel noise reduction systems: Application to auditory prostheses [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(1): 63-72

[56] Nie Shuai, Zhang Hui, Zhang Xueliang, et al. Deep stacking networks with time series for speech separation [C] //Proc of the 39th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2014: 6667-6671

[57] Zhang Xueliang, Zhang Hui, Nie Shuai, et al. A pairwise algorithm using the deep stacking network for speech separation and pitch estimation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(6): 1066-1078

[58] Hui Like, Cai Meng, Guo Cong, et al. Convolutional maxout neural networks for speech separation [C] //Proc of the 15th IEEE Int Symp on Signal Processing and Information Technology. Piscataway, NJ: IEEE, 2015: 24-27

[59] Lu Xugang, Tsao Y, Matsuda S, et al. Speech enhancement based on deep denoising autoencoder [C] //Proc of the 14th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2013: 436-440

[60] Han Kun, Wang Deliang. Neural network based pitch tracking in very noisy speech [J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2014, 22(12): 2158-2168

[61] Xu Yong, Du Jun, Dai Lirong, et al. An experimental study on speech enhancement based on deep neural networks [J]. IEEE Signal Processing Letters, 2014, 21(1): 65-68

[62] Han Kun, Wang Yuxuan, Wang Deliang, et al. Learning spectral mapping for speech dereverberation and denoising [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(6): 982-992

[63] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders [C] // Proc of the 25th Int Conf on Machine learning. New York: ACM, 2008: 1096-1103

[64] Maas A L, Le Q V, O'Neil T M, et al. Recurrent neural networks for noise reduction in robust ASR [C] //Proc of the 13th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2012: 22-25

[65] Xia Bingyin, Bao Changchun. Speech enhancement with weighted denoising auto-encoder [C] //Proc of the 14th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2013: 3444-3448

[66] Xu Yong, Du Jun, Dai Lirong, et al. A regression approach to speech enhancement based on deep neural networks [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(1): 7-19

[67] Tu Ming, Zhang Xianxian. Speech enhancement based on deep neural networks with skip connections [C] //Proc of the 42nd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2017: 5565-5569

[68] Wang Yuxuan, Wang Deliang. A deep neural network for time-domain signal reconstruction [C] //Proc of the 40th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2015: 4390-4394

[69] Karjol P, Kumar M A, Ghosh P K. Speech enhancement using multiple deep neural networks [C] //Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 5049-5052

[70] Park S R, Lee J W. A fully convolutional neural network for speech enhancement [C] //Proc of the 18th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2017: 1993-1997

[71] Fu S W, Tsao Y, Lu Xugang. SNR-aware convolutional neural network modeling for speech enhancement [C] //Proc of the 17th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2016: 3768-3772

[72] Gao Tian, Du Jun, Dai Lirong, et al. Densely connected progressive learning for LSTM-based speech enhancement [C] //Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 5054-5058

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

[74] Michelsanti D, Tan Zhenghua. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification [C] //Proc of the 18th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2017: 2008-2012

[75] Isola P, Zhu Junyan, Zhou Tinghui, et al. Image-to-image translation with conditional adversarial networks [C] //Proc of the 28th Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1125-1134

[76] Donahue C, Li Bo, Prabhavalkar R. Exploring speech enhancement with generative adversarial networks for robust speech recognition [C] //Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 5024-5028

[77] Huang Posen, Kim M, Hasegawa-Johnson M, et al. Deep learning for monaural speech separation [C] //Proc of the 39th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2014: 1562-1566

[78] Huang Posen, 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, 2015, 23(12): 2136-2147

[79] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks [C] //Proc of the 30th Int Conf on Machine Learning. Brookline, MA: Microtome Publishing, 2013: 1310-1318

[80] Weninger F, Hershey J R, Roux J L, et al. Discriminatively trained recurrent neural networks for single-channel speech separation [C] //Proc of the 2nd IEEE Global Conf on Signal and Information Processing. Piscataway, NJ: IEEE, 2014: 577-581

[81] Weninger F, Erdogan H, Watanabe S, et al. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR [C] //Proc of the 12th Int Conf on Latent Variable Analysis and Signal Separation. Berlin: Springer, 2015: 91-99

[82] Qian Kaizhi, Zhang Yang, Chang Shiyu, et al. Speech enhancement using Bayesian wavenet [C] //Proc of the 18th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2017: 2013-2017

[83] van den Oord A, Dieleman S, Zen H, et al. WaveNet: A generative model for raw audio [C] //Proc of the 9th ISCA Speech Synthesis Workshop. Grenoble, France: ISCA, 2016: 125-125

[84] Rethage D, Pons J, Serra X. A wavenet for speech denoising [C] //Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 5069-5073

[85] Fu S W, Tsao Y, Lu Xugang, et al. Raw waveform-based speech enhancement by fully convolutional networks [C] // Proc of the 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. Piscataway, NJ: IEEE, 2017: 006-012

[86] Fu S W, Wang Taowei, Tsao Y, et al. End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks [J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2018, 26(9): 1570-1584

[87] Venkataramani S, Casebeer J, Smaragdis P. End-to-end source separation with adaptive front-ends [C] //Proc of the 52nd Asilomar Conf on Signals, Systems, and Computers. Piscataway, NJ: IEEE, 2018: 684-688

[88] Pascual S, Bonafonte A, Serrà J. Segan: Speech enhancement generative adversarial network [C] //Proc of the 18th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2017: 3642-3646

[89] Vu T T, Bigot B, Chng E S. Combining non-negative matrix factorization and deep neural networks for speech enhancement and automatic speech recognition [C] //Proc of the 41st IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2016: 499-503

[90] Roux J L, Hershey J R, Weninger F. Deep NMF for speech separation [C] //Proc of the 40th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2015: 66-70

[91] Yang Yan, Bao Changchun. DNN-based AR-Wiener filtering for speech enhancement [C] //Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 2901-2905

[92] Bando Y, Mimura M, Itoyama K, et al. Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization [C] //Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 716-720

[93] Pandey A, Wang Deliang. A new framework for supervised speech enhancement in the time domain [C] //Proc of the 19th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2018: 1136-1140

[94] Bao Feng, Abdulla W H. A new ratio mask representation for casa-based speech enhancement [J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2019, 27(1): 7-19

[95] Pandey A, Wang Deliang. On adversarial training and loss functions for speech enhancement [C] //Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 5414-5418

[96] Tan Ke, Chen Jitong, Wang Deliang. Gated residual networks with dilated convolutions for monaural speech enhancement [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(1): 189-198

[97] Tan Ke, Wang Deliang. A convolutional recurrent neural network for real-time speech enhancement [C] //Proc of the 19th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2018: 3229-3233

[98] Xia Shasha, Li Hao, Zhang Xueliang. Using optimal ratio mask as training target for supervised speech separation [C] //Proc of the 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. Piscataway, NJ: IEEE, 2017: 163-166

[99] Saleem N, Khattak M I. Regularized sparse decomposition model for speech enhancement via convex distortion measure [J]. Modern Physics Letters B, 2018, 32(22): 1850262-1-1850262-13

[100] Eskimez S E, Soufleris P, Duan Zhiyao, et al. Front-end speech enhancement for commercial speaker verification systems [J]. Speech Communication, 2018, 99: 101-113

[101] Kolbæk M, Tan Zhenghua, Jensen J. On the relationship between short-time objective intelligibility and short-time spectral-amplitude mean-square error for speech enhancement [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(2): 283-295

[102] Kolbæk M, Tan Zhenghua, Jensen J. Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure [C] //Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 5059-5063

[103] Gelderblom F B, Tronstad T V, Viggen E M. Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(3): 583-594

[104] Vincent E, Virtanen T, Gannot S. Audio Source Separation and Speech Enhancement [M]. New York: John Wiley & Sons, 2018

[105] Kuttruff H. Room Acoustics [M]. Boca Raton, FL: CRC, 2016

[106] Zhang Xiongwei, Li Yinan, Zheng Changyan, et al. Speech dereverberation: Review of state-of-arts and prospects [J]. Journal of Data Acquisition and Processing, 2017, 32(6): 1069-1081 (in Chinese)

(张雄伟, 李轶南, 郑昌艳, 等. 语音去混响技术的研究进展与展望[J]. 数据采集与处理, 2017, 32(6): 1069-1081)

[107] Naylor P A, Gaubitch N D. Speech Dereverberation [M]. Berlin: Springer, 2010

[108] Neely S T, Allen J B. Invertibility of a room impulse response [J]. The Journal of the Acoustical Society of America, 1979, 66(1): 165-169

[109] Wu Mingyang, Wang Deliang. A two-stage algorithm for one-microphone reverberant speech enhancement [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(3): 774-784

[110] Dong Huanyu, Lee C M. Speech intelligibility improvement in noisy reverberant environments based on speech enhancement and inverse filtering [J]. Eurasip Journal on Audio, Speech, and Music Processing, 2018, 2018(1): 1-13

[111] Taal C H, Hendriks R C, Heusdens R. Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure [J]. Computer Speech & Language, 2014, 28(4): 858-872

[112] Kirkeby O, Nelson P A, Hamada H, et al. Fast deconvolution of multichannel systems using regularization [J]. IEEE Transactions on Speech and Audio Processing, 1998, 6(2): 189-194

[113] Liang Dawen, Hoffman M D, Mysore G J. Speech dereverberation using a learned speech model [C] //Proc of the 40th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2015: 1871-1875

[114] Mohammadiha N, Smaragdis P, Doclo S. Joint acoustic and spectral modeling for speech dereverberation using non-negative representations [C] //Proc of the 40th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ, 2015: IEEE: 4410-4414

[115] Zhang Long, Xu Xu, Chen Huang, et al. Supervised single-channel speech dereverberation and denoising using a two-stage model based sparse representation [J]. Speech Communication, 2018, 97: 1-8

[116] Mohanan N, Velmurugan R, Rao P. A non-convolutive nmf model for speech dereverberation [C] //Proc of the 19th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2018: 1324-1328

[117] Sun Pengfei, Mahdi A, Xu Jianhong, et al. Speech enhancement in spectral envelope and details subspaces [J]. Speech Communication, 2018, 101: 57-69

[118] Dionelis N, Brookes M. Modulation-domain Kalman filtering for monaural blind speech denoising and dereverberation [J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2019, 27(4): 799-814

[119] Nakatani T, Yoshioka T, Kinoshita K, et al. Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation [C] //Proc of the 33rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2008: 85-88

[120] Drude L, Boeddeker C, Heymann J, et al. Integrating neural network based beamforming and weighted prediction error dereverberation [C] //Proc of the 19th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2018: 3043-3047

[121] Kinoshita K, Delcroix M, Kwon H, et al. Neural network-based spectrum estimation for online wpe dereverberation [C] //Proc of the 18th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2017: 384-388

[122] Han Kun, Wang Yuxuan, Wang Deliang. Learning spectral mapping for speech dereverberation [C] //Proc of the 39th IEEE Int Conf on Acoustics, Speech and Signal Processing. IEEE, 2014: 4628-4632

[123] Wu Bo, Li Kehuang, Yang Minglei, et al. A reverberation-time-aware approach to speech dereverberation based on deep neural networks [J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2017, 25(1): 98-107

[124] Zhao Yan, Wang Zhongqiu, Wang Deliang. A two-stage algorithm for noisy and reverberant speech enhancement [C] //Proc of the 42nd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2017: 5580-5584

[125] Griffin D, Lim J. Signal estimation from modified short-time Fourier transform [J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1984, 32(2): 236-243

[126] Ernst O, Chazan S E, Gannot S, et al. Speech dereverberation using fully convolutional networks [C] //Proc of the 26th European Signal Processing Conf. Piscataway, NJ: IEEE, 2018: 390-394

[127] Guzewich P, Zahorian S A, Chen Xiao, et al. Cross-corpora convolutional deep neural network dereverberation preprocessing for speaker verification and speech enhancement [C] //Proc of the 19th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2018: 1329-1333

[128] Liu Bing, Tao Jianhua. A research to speech dereverberation method based on BLSTM recurrent neural networks and non-negative matrix factorization [J]. Journal of Signal Processing, 2017, 33(3): 268-272 (in Chinese)
(刘斌, 陶建华. 联合长短时记忆递归神经网络和非负矩阵分解的语音混响消除方法[J]. 信号处理, 2017, 33(3): 268-272)

[129] Santos J F, Falk T H. Speech dereverberation with context-aware recurrent neural networks [J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2018, 26(7): 1236-1246

[130] Tu Yanhui, Tashev I, Zarar S, et al. A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition [C] //Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 2531-2535

[131] Valentini-Botinhao C, Yamagishi J. Speech enhancement of noisy and reverberant speech for text-to-speech [J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2018, 26(8): 1420-1433

[132] Zhao Yan, Wang Deliang, Xu Buye, et al. Late reverberation suppression using recurrent neural networks with long short-term memory [C] //Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 5434-5438

[133] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [C] //Proc of the 3rd Int Conf on Learning Representations. Brookline, MA: Microtome Publishing, 2015: 1-14

[134] Palomäki K J, Brown G J, Barker J P. Techniques for handling convolutional distortion with ′missing data′ automatic speech recognition [J]. Speech Communication, 2004, 43(1-2): 123-142

[135] Yu Meng, Ji Xuan, Gao Yi, et al. Text-dependent speech enhancement for small-footprint robust keyword detection [C] //Proc of the 19th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2018: 2613-2617

[136] Zhao Han, Zarar S, Tashev I, et al. Convolutional-recurrent neural networks for speech enhancement [C] //Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 2401-2405

[137] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation [C] //Proc of the Int Conf on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234-241

[138] Li Chenxing, Wang Tieqiang, Xu Shuang, et al. Single-channel speech dereverberation via generative adversarial training [C] //Proc of the 19th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2018: 1309-1313

[139] Lee W J, Wang S S, Chen Fei, et al. Speech dereverberation based on integrated deep and ensemble learning algorithm [C] //Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 5454-5458

[140] Chien J T, Kuo K T. Spectro-temporal neural factorization for speech dereverberation [C] //Proc of the 43rd Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 5449-5453

[141] Raikar A, Basu S, Hegde R M. Single channel joint speech dereverberation and denoising using deep priors [C] //Proc of the 6th IEEE Global Conf on Signal and Information Processing. Piscataway, NJ: IEEE, 2018: 216-220

[142] Sun Lei, Du Jun, Dai Lirong, et al. Multiple-target deep Learning for LSTM-RNN based speech enhancement [C] //Proc of the 4th Hands-Free Speech Communications and Microphone Arrays. Piscataway, NJ: IEEE, 2017: 136-140

[143] Hirsch H G, Finster H. The simulation of realistic acoustic input scenarios for speech recognition systems [C] //Proc of the 9th European Conf on Speech Communication and Technology. Grenoble, France: ISCA, 2005: 2697-2700

[144] Garofolo J S, Lamel L F, Fisher W M, et al. TIMIT acoustic phonetic continuous speech corpus [DB/OL]. Philadelphia, PA: Linguistic Data Consortium, 1993 [2019-05-24]. https://catalog.ldc.upenn.edu/LDC93S1

[145] Paul D B, Baker J M. The design for the Wall Street Journal-based CSR corpus [C] //Proc of the Workshop on Speech and Natural Language. Stroudsburg, PA: ACL, 1992: 357-362

[146] Rothauser E, Chapman W, Guttman N, et al. IEEE recommended practice for speech quality measurements [J]. IEEE Transactions on Audio and Electroacoustics, 1969, 17(3): 225-246

[147] Veaux C, Yamagishi J, King S. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database [C] //Proc of the 2013 Int Conf Oriental COCOSDA held jointly with 2013 Conf on Asian Spoken Language Research and Evaluation. Piscataway, NJ: IEEE, 2013: 225-228

[148] Panayotov V, Chen G, Povey D, et al. Librispeech: An asr corpus based on public domain audio books [C] //Proc of the 40th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2015: 5206-5210

[149] Takeda K, Sagisaka Y, Katagiri S. Acoustic-phonetic labels in a Japanese speech database [C] //Proc of the 1st European Conf on Speech Technology. Grenoble, France: ISCA, 1987: 2013-2016

[150] Hirsch H G, Pearce D. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions [C] //Proc of the 1st Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2000: 29-32

[151] Vincent E, Barker J, Watanabe S, et al. The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines [C] //Proc of the 38th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2013: 126-130

[152] Barker J, Marxer R, Vincent E, et al. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines [C] //Proc of the 9th Workshop on Automatic Speech Recognition and Understanding. Piscataway, NJ: IEEE, 2015: 504-511

[153] Wong L L, Soli S D, Liu Sha, et al. Development of the mandarin hearing in noise test (NHINT) [J]. Ear and Hearing, 2007, 28(2): 70S-74S

[154] Varga A, Steeneken H J. 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

[155] Thiemann J, Ito N, Vincent E. The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings [J]. Acoustical Society of America, 2013, 133(5): 3591-3591

[156] Steeneken H J, Geurtsen F W. Description of the RSG-10 noise database [R]. Soesterberg: Institute for Perception, 1988

[157] Hu Guoning, Wang Deliang. A tandem algorithm for pitch estimation and voiced speech segregation [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(8): 2067-2079

[158] Kim G, Lu Yang, Hu Yi, et al. An algorithm that improves speech intelligibility in noise for normal-hearing listeners [J]. Acoustical Society of America, 2009, 126(3): 1486-1494

[159] Wang Yuxuan, Han Kun, Wang Deliang. Exploring monaural features for classification-based speech segregation [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(2): 270-279

[160] Hermansky H. Perceptual linear predictive (PLP) analysis of speech [J]. The Journal of the Acoustical Society of America, 1990, 87(4): 1738-1752

[161] Hermansky H, Morgan N. Rasta processing of speech [J]. IEEE Transactions on Speech and Audio Processing, 1994, 2(4): 578-589

[162] Huang Posen, Kim M, Hasegawa-Johnson M, et al. Singing-voice separation from monaural recordings using deep recurrent neural networks [C/OL] //Proc of the 15th Int Society for Music Information Retrieval Conf. 2014: 477-482 [2019-05-24]. https://zenodo.org/record/1415678

[163] Shao Yang, Wang Deliang. Robust speaker identification using auditory features and computational auditory scene analysis [C] //Proc of the 33rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2008: 1589-1592

[164] Maganti H, Matassoni M. An auditory based modulation spectral feature for reverberant speech recognition [C] //Proc of the 11th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2010: 570-573

[165] Kumar K, Kim C, Stern R M. Delta-spectral cepstral coefficients for robust speech recognition [C] //Proc of the 36th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2011: 4784-4787

[166] Yu Yang, Wang Wenyu, Han Peng. Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks [J]. Eurasip Journal on Audio, Speech, and Music Processing, 2016, 2016(1): 1-18

[167] Shannon B J, Paliwal K K. Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition [J]. Speech Communication, 2006, 48(11): 1458-1485

[168] Ikbal S, Misra H, Bourlard H. Phase autocorrelation (PAC) derived robust speech features [C] //Proc of the 28th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2003: Ⅱ-133-Ⅱ-136

[169] Chen Jitong, Wang Yuxuan, Wang Deliang. A feature study for classification-based speech separation at low signal-to-noise ratios [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1993-2002

[170] Telecommunication Standardization Sector of ITU. Subjective Test Methodology for Evaluating Speech Communication Systems That Include Noise Suppression Algorithm [S]. Geneva, Switzerland: ITU-T, 2003

[171] Rix A W, Beerends J G, Hollier M P, et al. Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs [C] //Proc of the 26th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2001: 749-752

[172] Taal C H, Hendriks R C, Heusdens R, et al. An algorithm for intelligibility prediction of time - frequency weighted noisy speech [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2125-2136

[173] Shu Xiaofeng, Zhou Yi, Cao Yin. A progressive enhancement method for noisy and reverberant speech [C] //Proc of the 23rd IEEE Int Conf on Digital Signal Processing. Piscataway, NJ: IEEE, 2018: 1030-1034

[174] Gao Tian, Du Jun, Dai Lirong, et al. SNR-based progressive learning of deep neural network for speech enhancement [C] //Proc of the 17th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2016: 3713-3717

[175] Wang Zhongqiu, Wang Deliang. All-neural multi-channel speech enhancement [C] //Proc of the 19th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2018: 3234-3238

[176] Park S R, Lee J W. A fully convolutional neural network for speech enhancement [C] //Proc of the 18th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2017: 1993-1997

[177] Xu Jiaming, Shi Jiaming, Liu Guangcan, et al. Modeling attention and memory for auditory selection in a cocktail party environment [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 2564-2571

[178] Krawczyk M, Gerkmann T. STFT phase improvement for single channel speech enhancement [C] //Proc of the 13th Int Workshop on Acoustic Signal Enhancement. Berlin: VDE, 2012

[179] Gerkmann T, Krawczyk M. MMSE-optimal spectral amplitude estimation given the STFT-phase [J]. IEEE Signal Processing Letters, 2013, 20(2): 129-132

[180] Krawczyk M, Rehr R, Gerkmann T. Phase-sensitive real-time capable speech enhancement under voiced-unvoiced uncertainty [C] //Proc of the 21st European Signal Processing Conf. Piscataway, NJ: IEEE, 2013: 1-5

[181] Krawczyk M, Gerkmann T. STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1931-1940

[182] Kulmer J, Mowlaee P, Watanabe M K. A probabilistic approach for phase estimation in single-channel speech enhancement using von mises phase priors [C] //Proc of the 24th IEEE Int Workshop on Machine Learning for Signal Processing. Piscataway, NJ: IEEE, 2014: 61-66

[183] Mowlaee P, Kulmer J. Phase estimation in single-channel speech enhancement: Limits-potential [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(8): 1283-1294

[184] Krawczyk-Becker M, Gerkmann T. On MMSE-based estimation of amplitude and complex speech spectral cofficients under phase-uncertainly [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(12): 2251-2262

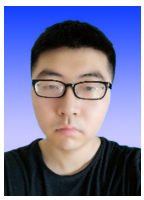
[185] Wakabayashi Y, Fukumori T, Nakayama M, et al. Single-channel speech enhancement with phase reconstruction based on phase distortion averaging [J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2018, 26(9): 1559-1569

[186] Zhao Yan, Xu Buye, Giri R, et al. Perceptually guided speech enhancement using deep neural networks [C] //Proc of the 43rd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 5074-5078

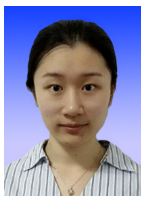
Lan Tian, born in 1977. PhD, associate professor. His main research interests include speech recognition, natural language processing and medical image processing.

Peng Chuan, born in 1994. Master candidate. His main research interests include natural language processing, speech enhancement and speech recognition.

Li Sen, born in 1996. Master candidate. His main research interests include natural language processing and speech recognition.



Ye Wenzheng, born in 1995. Master candidate. His main research interests include speech enhancement and speech recognition.



Li Meng, born in 1995. Master candidate. Her main research interests include speech enhancement and speech separation.

Hui Guoqiang, born in 1996. Master candidate. His main research interests include speech recognition and speech enhancement.

Lü Yilan, born in 1997. Master candidate. Her main research interests include speech enhancement and speech separation.

Qian Yuxin, born in 1997. Master candidate. His main research interests include speech enhancement and speech separation.

Liu Qiao, born in 1974. PhD, professor, PhD supervisor. His main research interests include natural language processing, machine learning and data mining.