ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 6,Nov. 2017,pp. 1069-1081
DOI:10.16337/j.1004-9037.2017.06.001

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:+86-025-84892742

语音去混响技术的研究进展与展望

张雄伟[1]　李轶南[1]　郑昌艳[1]　曹铁勇[1]　孙　蒙[1]　闵　刚[2]

(1.中国人民解放军陆军工程大学，南京,210007；2.国防科技大学信息通信学院,武汉,430010)

摘　要：语音交互技术在实际的话音驱动应用中得到日益普及。然而,当声源距离传声器较远时,由于实际环境中混响现象的影响,语音交互的性能还远不能使人满意。针对混响问题,数十年来学者们不断地进行大量的研究,并提出了很多实用的方法。特别是近期兴起的在很大程度上改变语音处理格局的深度学习技术,在单通道去混响方面也取得了很多令人瞩目的效果。然而,目前系统性总结分析基于深度学习的去混响方法与经典算法之间联系的工作仍然比较匮乏。因此,本文对单通道语音去混响技术的发展脉络进行系统的梳理和总结,并讨论了有待进一步研究的开放问题。

关键词：语音去混响;语音交互;深度学习

中图分类号：TN912.3　　**文献标志码**：A

Speech Dereverberation: Review of State-of-the-Arts and Prospects

Zhang Xiongwei[1], Li Yinan[1], Zheng Changyan[1], Cao Tieyong[1], Sun Meng[1], Min Gang[2]

(1. Army Engineering University of PLA, Nanjing, 210007, China; 2. College of Information and Communication, National University of Defense Technology, Wuhan, 430010, China)

Abstract: Speech interaction technology is becoming increasingly popular in practical voice-driven applications. However, due to the inferences caused by reverberation in real-world environments, the performances of speech interaction in the distant-talking condition are far from being satisfactory. Decades of efforts are devoted to solving the reverberation problem and spawning a vast variety of practical methods. Recently, the deep learning technique, which is developing rapidly and has greatly reshaped the speech processing community, also acquires remarkable performance in speech dereverberation. However, a systematic analysis and summary of the inherent relationship between the recent deep learning based methods and the previous classical methods is rarely seen. As such, we give a comprehensive overview of the current and past development of single channel speech dereverberation. Then, the main challenges are discussed. Finally, we share some views of its future development.

Key words: speech dereverberation; speech interaction; deep learning

引　　言

语音信号处理在过去的数十年间取得了长足的进步,很多话音驱动的系统已经逐步融入到人们的

基金项目:国家自然科学基金(61471394,61402519,61701535)资助项目;江苏省自然科学基金(BK20140071,BK20140074)资助项目。
收稿日期:2017-09-19;**修订日期**:2017-10-09

日常生活之中。然而,如今大多数应用仍需要将传声器尽可能地靠近说话人,这种手持式或者穿戴式的使用方式在一定程度上会令使用者感到不适[1]。开发出能够允许说话人在位于传声器较远距离时,仍能正常工作的系统是非常必要的。此类系统能够帮助提升诸如:助听器、免提模式下工作的电子产品、电视电话会议语音识别系统、远程会议中的自动翻译系统和互动式电视等实际产品应用的性能。

当说话人距离传声器较远时,采集到的声音往往会被混响和背景噪声所污染,使得声音的清晰度和可懂度大幅下降。为了解决此问题,很多算法被提出。处理加性背景噪声的算法常常将背景噪声的谱结构限定于一帧或者少量数帧短时信号之中[2,3]。与之形成对照的混响效应则与加性背景噪声不同,其持续影响常常能够跨越很多语音帧,因此需要设计专门的算法对其进行处理。

数十年来,针对混响语音处理的研究取得了很多进展,特别是将混响问题的实际需求与房间声学、计算听觉场景分析、听觉心理学、机器学习、语音建模和最优滤波等技术相结合,诞生出了很多解决混响问题的新思路[4]。随着诸如盲反卷积、非负矩阵分解(Non-negative matrix factorization, NMF)以及深度学习等方法在不同时期和不同阶段取得不断进步,语音去混响的处理水平也在不断提高。

早期的去混响算法通常基于一些简化的统计模型,给定先验假设,实现对于语音以及混响声学特征的描述和逼近,这些算法往往简便易行,需要的内存和计算开销也相对较小,可以很容易地在硬件平台上加以实现。然而,模型的简化往往意味着描述能力的不足,所添加的先验假设在实际环境中,特别是低信噪比、长混响时间条件下,极难得到满足,导致处理效果很难令人满意。

现阶段,随着计算能力的快速发展、存储设备价格的日益降低以及海量数据的不断涌现,设计更精确、更复杂的描述模型成为可能。以深度学习为代表的机器学习技术的革命,使得语音处理相关研究的各个领域都取得了令人瞩目的成果[5-7]。相应地,语音去混响算法也得以使用更多的先验信息,使新方法的处理性能提升到一个新的层次。纵观技术发展的总体历程,其发展的总趋势是使用更多的先验数据,开发出表示能力更强的深层结构,逐步使对于混响/纯净语音信号更有效的刻画成为可能。然而,目前针对去混响技术的发展脉络的系统性梳理工作还较为少见,尽管有一些综述性的工作,但是往往局限在一个特定的方面,比如 Mosayyebpour 侧重于介绍和分析基于模型假设、无训练样本模式下的去混响情况[8];Yoshioka 围绕自动语音识别(Automatic speech recognition, ASR)各个不同层次分别论述了处理混响的相应手段[4],而与 ASR 联系不紧密的混响相关的研究则很少涉及;Kinoshita 探讨了参加 REVERB 挑战赛[9]的各种算法,所论述的算法主要集中在 2013 年以前较为流行的去混响算法[10],一些新近提出的,特别是基于深度学习的去混响算法论述相对较少。国内近几年关于语音去混响的综述尚未看到。针对此问题,本文试图从语音去混响技术的整个发展历程的视角来回顾现有的算法,找出当前各类算法之间的区别与联系,理清整个技术体系的发展脉络,并对该领域的研究和发展提供一个详实的参考。

1 混响语音信号的特征

1.1 混响声学模型

当连续语音信号 $s(t)$ 从房间某个位置传播开来,由于多径效应,较远处的传声器捕获到的信号 $y(t)$ 将会是直接传播的语音信号与经由墙壁、地板、天花板以及其他物体一次乃至数次发射所得语音信号的叠加。在数学上,混响的语音信号可以由纯净语音和房间冲击响应(Room impulse response, RIR) $h(t)$ 通过线性卷积来得到,即

$$y(t)=\sum_{\tau=0}^{T}h(\tau)s(t-\tau)+n(t)=h(t)*s(t)+n(t) \tag{1}$$

式中:符号"$*$"为卷积操作;T 为房间冲击响应的长度;$s(t)$,$h(t)$,$n(t)$,$y(t)$ 分别为纯净语音信号、房间冲击响应、加性背景噪声和混响语音信号。为了聚焦于去混响算法的开发,有时也忽略加性背景噪声

$n(t)$。其中房间冲击响应,由房间的声学特性,说话人和传声器的位置等因素决定,当说话人和传声器的位置发生改变时,房间冲击响应也会随之发生改变。

通常将房间冲击响应划分为 3 个组成部分,如图 1 所示,即:直达声音(Direct sound)、早期反射(Early reflection)和晚期混响(Late reverberation)。直达声音是不经反射,直接从声源发出被传声器收集到的声音,紧随直达声音的,被称作早期反射。早期反射由几个比较强的反射所构成,通常由反射次数较少的强反射构成。最后的一部分是由一系列难以区分的反射构成,称作晚期混响[1]。通常,紧随直达声音 50 ms 时间范围内的反射信号被称为早期反射,早期反射与说话人和传声器所在的位置高度相关,因此对于说话人和传声器的位置运动较为敏感;而在直达声音 50 ms 以后的则是晚期混响,大致呈现出指数衰减的模式并与位置相独立。晚期混响对于说话人和传声器位置不敏感的特性,可以用来开发对于说话人运动鲁棒的去混响算法。因此,常常将混响分为

$$h_1(t)=\begin{cases}h(t) & t<\Delta\\ 0 & t\geqslant\Delta\end{cases},\ h_2(t)=\begin{cases}h(t+\Delta) & t\geqslant\Delta\\ 0 & t<\Delta\end{cases} \tag{2}$$

式中:直达声音和早期混响共同被记作 $h_1(t)$,晚期混响被记作 $h_2(t)$,Δ 表示的是二者的时间边界,通常取直达声音抵达后的 50 ms。

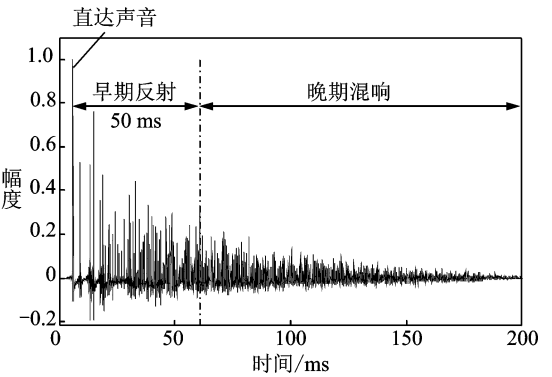


图 1 房间冲击响应的构成

Fig. 1 The composition of a room impulse response

1.2 衡量混响的常见参数

有很多参数可用于对混响进行衡量,这些参数对于预测自动语音识别的正确率非常重要。混响时间 T_{60} 是最常见的用以描述房间声学特性的参数。T_{60} 是晚期混响相对于直达声音衰减 60 dB 所需要的时间[4]。典型的办公室和家庭环境,混响时间 T_{60} 大约是 200 ms 到 1 000 ms。除混响时间外,还有一些其他的常见指标,比如直达混响比(Direct-to-reverberant ratio, DRR),D_{50},清晰指数 C_{50} 以及中心时间 T_S[11]。其中,DRR 的计算公式为

$$\mathrm{DRR}=10\log_{10}\left(\frac{E_d}{\left(\sum_{m=0}^{M-1}h^2(m)-E_d\right)}\right) \tag{3}$$

式中:E_d 为直达能量,通过将 sinc 函数与房间冲击响应直达通路附近的样本进行卷积得到

$$E_d=\max_{\sigma}\sum_{m=-\eta}^{\eta}\left(\mathrm{sinc}\left(\pi(m+\sigma)\right)h(m+n_d)\right)^2 \tag{4}$$

式中:$\eta=8$ 为 sinc 函数的旁瓣数,并且 $\sigma=[-1:1]$是寻找最大能量值时所考虑的偏移。

类似地,C_{50} 和 D_{50} 为

$$C_\tau=10\log_{10}\left(\frac{\sum_{m=0}^{N_\tau}h^2(m)}{\sum_{m=N_\tau+1}^{M-1}h^2(m)}\right) \tag{5}$$

$$D_\tau=10\log_{10}\left(\frac{\sum_{m=0}^{N_\tau}h^2(m)}{\sum_{m=0}^{M-1}h^2(m)}\right) \tag{6}$$

式中:N_τ 为房间冲击响应 $h(m)$ 从直达声音开始,在 τ ms 时间内的样点总个数,如上所述,50 ms 常常

用于区分早期反射和晚期混响的时间边界，因此，常常也以 50 ms 作为边界计算混响参数。

此外，中心时间 T_S 测量的是均方房间冲击响应的质心，其计算公式为

$$T_S = \frac{\sum_{m=0}^{M-1} \left(\frac{m}{f_s}\right) h^2(m)}{\sum_{m=0}^{M-1} h^2(m)} \tag{7}$$

式中：f_s 为采样频率。

Kuttruff 指出 D_{50} 可以指示语音的可懂度，也可以预测自动语音识别中词的准确率[11]，Tsilfidis 指出 C_{50} 是在房间声学参数中（包括 T_{60}，C_{50}，D_{50} 等）与自动语音识别最为相关的参数[12]。Pablo 通过皮尔逊相关系数（Pearson correlation coefficient）以及变量之间的互信息（Mutual information）验证了 Tsilfidis 的观点[13]。

1.3 自然混响的统计特征

为了研究人类听觉系统是否能够从混响环境的统计特性中推断出声源以及所处空间的信息。近期针对实际环境中混响的统计特性，文献[14]收集并调研了自然界中 271 种不同环境中（包括位于不同城市的体育馆、超市、森林、餐馆和公寓商店等）的混响冲击响应（Reverberant impulse responses），系统全面地分析了这些信号的统计特性，并在与人类听觉高度相关的耳蜗谱（Cochleagram）上进行研究。研究结果表明，混响的冲击响应具有一定的普遍规律，这些规律总结如下：

（1）在直达声音抵达后约 50 ms 的时间范围内，混响的冲击响应时间序列呈现出类似高斯噪声的局部统计特性，即晚期混响可以近似为被幅度包络调制的高斯噪声。

（2）在不同频带上，混响的能量大致以指数形式衰减。

（3）不同频带呈现出不同的衰减速率，并且相较于低频和高频，中频部分（200～2 000 Hz）的衰减相对更加缓慢。

（4）伴随着混响总能量的增加，不同频率的衰减速率的变化更加显著。

充分有效地利用混响的这些规律能够帮助开发出性能更加优异的去混响算法。

1.4 语音特征表示

语音信号特征表示有很多种不同的计算方式[15]，最直接的方法是使用语音时域波形（Waveform signal，WAV）。

此外，还有诸如伽马通频率特征（Gammatone frequency feature，GF）[16]、伽马通频率倒谱系数（Gammatone frequency cepstral coefficients，GFCC）[17]、多分辨率的耳蜗谱（Multiresolution cochleagram，MRCG）[18]、伽马通频率调制系数（Gammatone frequency modulation coefficients，GFMC）[19]、基于基音的特征（Pitch-based feature，PITCH）[20]、对数幅度谱特征（Log-magnitude spectral feature，LOG-MAG）、感知线性预测特征（Perceptual Linear prediction feature，PLP）[21]、相对谱变换 PLP 特征（Relative spectral transform PLP feature，RASTA-PLP）[22]、幅度调制谱（Amplitude modulation spectrogram，AMS）[23]、Gabor 滤波器组特征（Gabor filterbank，GFB）[24]、梅尔频率倒谱系数（Mel-frequency cepstral coefficients，MFCC）、对数梅尔滤波器组特征（Log-Mel filterbank feature，LOG-Mel）、相对自相关序列 MFCC（Relative autocorrelation sequence MFCC，RAS-MFCC）[25]、相位自相关 MFCC（Phase autocorrelation MFCC，PAC-MFCC）[26]、自相关 MFCC（Autocorrelation MFCC，AC-MFCC）[27]、能量归一化倒谱系数（Power-normalized cepstral coefficients，PNCC）[28]。

表 1 对上述语音特征进行了简要说明,文献[15]全面比较了混响和噪声条件下,应用不同语音特征对于去混响的影响与效果。

表 1 语音特征表示及简要描述

Tab. 1 Speech features and their brief descriptions

语音特征	简 要 描 述
WAV	直接使用信号时域波形
GF	与听觉感知器高度相关
GFCC	为自动说话人辨识设计的一种参数
MRCG	能够同时体现局部和上下文信息
GFMC	为缓解噪声和混响影响的特征
PITCH	利用听觉场景分析重要的基音作为特征
LOG-MAG	对于语音幅度谱取对数运算
PLP	抑制说话人相关细节的特征
RASTA-PLP	添加 RASTA 滤波器的 PLP 特征
AMS	调制特征
GFB	通过一系列 Gabor 滤波器组得到的特征
MFCC	语音信号处处理常用特征
LOG-MEL	在语音识别和分离中广泛使用的特征
RAS-MFCC	噪声鲁棒的特征
PAC-MFCC	基于信号相位轨迹的特征
AC-MFCC	用 MFCC 处理较高滞后性的自相关序列
PNCC	更加噪声和混响鲁棒的改进版 MFCC 特征

1.5 评价语音去混响算法的主要指标

语音去混响算法的性能指标大致可以分两个角度来评价:(1)从语音增强的角度来评价去混响的性能;(2)从自动语音识别的角度来评价[29]。当从语音增强角度来对增强后的语音质量评价时,评价方式又可以进一步分为客观评价和主观评价,常见的客观评价指标包括倒谱距离(Cepstrum distance, CD)[30]、对数自然率(Log likelihood ratio, LLR)[30]、频率权重的分段信噪比(Frequency weighted segmental SNR, FWSegSNR)[30]、语音混响调制能量比(Speech to reverberation modulation energy ratio, SRMR)[31],语音质量的感知评价(Perceptual evaluation of speech quality, PESQ)[32],客观的语言清晰度度量(Short-time objective intelligibility, STOI)[33]。表 2 给出了这些客观评价指标的简要说明。主观

表 2 客观语音去混响的评价指标

Tab. 2 Objective measurements for dereverberated speech

评价指标	简要描述
CD	处理所得的语音与纯净语音之间的倒谱距离。数值越小,表明效果越好。
LLR	基于 LPC 测量的指标,描述处理所得语音与纯净语音平滑谱之间的差异程度。数值越小,表明效果越好。
FWSegSNR	根据人类听觉感知系统的敏感程度分配权重。越高的打分表明越好的语音质量。
SRMR	无需纯净语音便可计算的非侵入式指标。越高的打分表明越好的语音质量。
PESQ	ITU-T 推荐的评价指标。越高的打分表明越好的语音质量
STOI	与带噪语音的可懂度高度相关

评价可以使用 ITU-R 推荐的(Multi stimulus test with hidden reference and anchor, MUSHRA)测试来进行[34]。从自动语音识别的角度来评价时,语音识别的错词率(Word error rate, WER)常常被用作评价指标。

2 典型的语音去混响方法

根据利用先验知识从少到多的顺序,本节介绍 3 种典型的去混响方法。

2.1 盲反卷积方法

盲反卷积方法利用最少的先验信息,既不需要语音信号样本,也不需要 RIR 信号样本,便可以对混响语音信号进行处理。使用逆滤波估计就是最直接的处理混响语音的思路之一[35]。然而,其所经常采用的最小相位假设在实际应用中几乎从未得到满足。文献[36]通过最大化线性预测(Linear prediction, LP)残差的峰度来估计固定长度的 RIR 逆滤波器以抑制早期混响。然而,逆滤波只在从 200～400 ms 这样很短的混响时间区间之内才有效。当混响时间较长时,基于峰度的自适应逆滤波的目标函数出现很多的鞍点,导致估计的不准确[37]。

另一类广泛采用的盲反卷积算法基于简化的统计模型对 RIR 进行建模[10]。在这种方法中,将 RIR $h(t)$ 视作被指数衰减包络所调制的白噪声,衰减速率由混响时间 T_{60} 所决定,为

$$h(t)=\begin{cases}a(t)\exp\left(-\dfrac{3\ln(10)}{T_{60}}t\right) & t>0\\ 0 & \text{其他}\end{cases} \tag{8}$$

式中:$a(t)$ 为以 0 为均值,以 σ_a^2 为方差的白噪声,并且 T_{60} 为混响时间。

假设观测到的混响信号是由纯净语音信号和如上所述的简化 RIR 通过时域卷积所得到,那么就可以将混响功率谱 $|\hat{R}_n[f]|^2$ 看作是由过去帧的功率谱 $|Y_{n-K}[f]|^2$(n 表示的是语音帧的序号),通过添加一定的权重所得到,则

$$|\hat{R}_n[f]|^2=|Y_{n-K}[f]|^2\exp\left(-\frac{6\ln(10)}{T_{60}}T_d\right) \tag{9}$$

式中:$K=\left|\dfrac{T_d f_s}{D}\right|$ 并且符号 $|\cdot|$ 表示向下取整,T_d 通常设置为 50 ms,f_s 为采样频率,D 为计算短时傅里叶变换(Short-Time Fourier transformation, STFT)时的帧移样点数。

此时,纯净语音信号可以通过谱减法得到,即通过从观测功率谱 $Y_n[f]$ 中减去估计出的混响功率谱 $|\hat{R}_n[f]|^2$ 得到。然而谱减法可能会带来令人不悦的音乐噪声[8]。

利用线性时不变滤波器(Linear time-invariant filter, LTI filter)是进行盲反卷积的另一个思路。比较有代表性的方法被称作带有权重的预测线性误差 (Weighted linear prediction error, WPE) 方法[37]。WPE 在 STFT 的每个频点上进行如下的长时线性预测,则

$$\boldsymbol{y}_n[f]=\sum_{\tau=D}^{D+O}\boldsymbol{G}_\tau[f]^{\mathrm{H}}\boldsymbol{y}_{n-\tau}[f]+\boldsymbol{s}_n[f] \tag{10}$$

式中:n 为时间帧的序号,$\boldsymbol{y}_n[f]$ 为不同频率 f 上的 STFT 的系数构成,$\boldsymbol{s}_n$ 为预测误差向量,矩阵 $\boldsymbol{G}_\tau$ 是被称作预测矩阵(Prediction matrix)的复数值方阵,符号 $(\cdot)^{\mathrm{H}}$ 表示共轭转置。D 为预测步长,O 为预测滤波器阶数。

需要指出的是,一方面,由于语音信号具有时变特性,因此纯净语音信号与其过去的帧信号相关性也是时变的;另一方面,晚期混响部分本质上是先前的语音帧,通过一定时间的时延反射所构成,因此晚期混响与纯净语音部分和早期混响部分不相关。换言之,纯净语音与早期混响将会以预测残差的形式残留下来,也就是说,可以将式(10)重写成如下的方式来对纯净语音信号进行估计,则

$$\hat{\boldsymbol{s}}_n[f]=\boldsymbol{y}_n[f]-\sum_{\tau=D}^{D+O}\boldsymbol{G}_\tau[f]^{\mathrm{H}}\boldsymbol{y}_{n+\tau}[f] \tag{11}$$

通常,线性预测方法将预测步长 D 设定为 1。然而,WPE 算法通常将预测步长 D 设定为 2 或者是 3,以减轻线性预测的过度去关联效应。通过将预测步长设定为大于 1 的整数,过度去关联效应会得到改善,其代价是引入某种程度的染色效应[37]。此外,WPE 只对晚期混响部分进行抑制,而对于直达声音和早期混响则加以保留。

2.2 基于非负矩阵分解的去混响方法

基于盲反卷积的方法常常需要为语音或者混响添加一些先验假设。为了便于计算和建模,这些假设往往对于语音信号和 RIR 信号进行了不同程度的简化和近似,使得进一步提升去混响算法遭遇瓶颈。近年来,基于非负矩阵分解(Non-negative matrix factorization, NMF)的算法在语音处理方面获得了成功的应用[38]。其内涵的组合模型能够很好地对语音信号进行建模[39]。最早的基于 NMF 模型的去混响算法由 Kameoka 提出[40],在非负的约束下使用逆滤波的调制变换函数(Modulation transfer function, MTF),并使用求解 NMF 问题经典的乘法迭代来对各个参数进行估计,其基本模型为

$$y(k,t)=\sum_{\tau}h(k,\tau)s(k,t-\tau) \tag{12}$$

式中:y,s 和 h 分别为时频域的混响语音信号、纯净语音信号和 RIR 信号模型。k 为不同频率的子带,t 为语音帧的时序。模型假设语音信号是稀疏的,令 $\sum_t h(k,t)=1$ 以消除模型的尺度模糊(Scale ambiguity)问题。

上述方法仍然是盲反卷积的方法,在此基础上,Mohammadiha[41]将语音字典引入纯净语音的表示,通过对语音信号的更精确刻画以进一步提升去混响算法的效果,则

$$s(k,t)\approx\sum_{r}w(k,r)x(r,t) \tag{13}$$

式中:w 为每个基函数都被归一化的语音字典,k 为不同频率,r 为基函数的序号。式(12)可以改写为

$$y(k,t)=\sum_{\tau}h(k,\tau)\sum_{r}w(k,r)x(r,t-\tau) \tag{14}$$

式中:纯净语音由语音字典进行约束;对 RIR 添加两个约束:(1)上述的避免尺度模糊的约束,(2)对于所有的 τ,满足 $h(k,\tau)<h(k,\tau-1)$ 的约束条件。为了更精确地对语音信号进行建模,具有一定时间连续性描述能力的语音模型被用于对语音信号进行建模[42]。此外,同时考虑背景噪声和混响影响的工作[43]。Mohanan 尝试对 RIR 模型添加更多的约束以提升去除效果,分别讨论了对 RIR 部分添加稀疏约束、给定 RIR 频率子带幅度包络和保留早期反射部分对于去混响的影响[44]。Liang 基于广义逆高斯分布(Generalized inverse-Gaussian, GIG)的先验假设,使用变分推断方法对各个参数进行推断[45]。然而当前常常使用的约束对于精确建模 RIR 来说并不充足,添加的约束往往只能反映出 RIR 的大致趋势,这造成了所估计出的 RIR 与实际的 RIR 信号之间仍然存在着一定的差距,这在一定程度上影响着去混响算法的性能。如何添加合适的约束,使得估计所得结果趋于合理,是基于非负矩阵分解模型的去混响方法需要解决的问题。

2.3 基于深度学习的语音去混响方法

近年来,基于深度学习(Deep learning)的语音处理技术取得了令人振奋的效果[6,7,11,46,47]。特别是在处理回归问题时,基于非线性映射的深层人工神经网络的深度学习技术展现出极强建模能力。不同于盲反卷积和基于语音模型的去混响方法,基于深度学习的去混响算法无需构建特定的混响模型,而是通过大量的训练样本,直接学习混响语音和纯净语音之间的非线性映射关系。从先验信息的角度来看,基于深度学习的语音去混响方法使用了最多的先验知识,纯净语音信号和 RIR 信号均得到了充分的利用。

近期涌现出了很多基于各类人工神经网络架构的去混响算法，其中包括基于去噪自编码器(Denoising auto-encoder, DAE)的算法[46,48]，基于全连接深度神经网络(Deep neural networks, DNN)的算法[49]，考虑语音信号时序性的长短时记忆网络(Long short-term memory, LSTM)的算法[50]等。这些基于深度学习的去混响算法大致模型如图2所示。当高度非线性的映射函数，即深层神经网络训练完毕后，就可根据输入混响语音特征 $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_N$ 对纯净语音信号 $\tilde{\boldsymbol{X}}_N$ 进行估计，则

$$\tilde{\boldsymbol{X}}_N = F\{(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_N; \tilde{\theta}\} \tag{15}$$

式中：$F\{\cdot;\theta\}$ 为基于神经网络的以 θ 为参数的非线性变换函数。

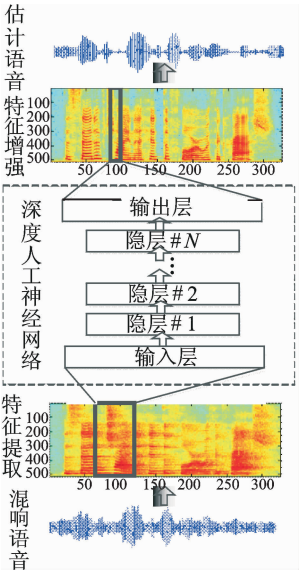


图2 基于深度人工神经网络的语音去混响算法

Fig. 2 Diagrams of deep artificial neural network based algorithms for speech dereverberation

比较有代表性的工作由 Han 提出[49]，首先将大量纯净语音与 RIR 信号卷积获得混响语音信号，然后使用具备3个隐层的神经网络学习混响语音谱和纯净语音谱之间的映射关系。在此基础上，Williamson 考虑语音谱的实部和虚部并分别进行映射，纯净语音信号和混响语音信号之间的映射关系由一个复数理想分数掩模(Complex ideal ratio mask, cIRM)来承载[51]。当使用 DAE 或是 DNN 时，语音信号的短时动态特性通过联合考虑邻近的语音帧来实现。需要指出的是，基于 DAE 或者 DNN 的语音去混响算法往往有延迟处理，因此当处理的语音帧序号为 k 时，以第 k 帧语音信号为中心的 $2p+1$ 个混响语音帧均被考虑作为特征向量以提升处理所得语音信号的连续性，则

$$\hat{\boldsymbol{Y}}_k = \{\boldsymbol{Y}_{k-p}, \boldsymbol{Y}_{k-p+1}, \cdots, \boldsymbol{Y}_k, \boldsymbol{Y}_{k+p-1}, \boldsymbol{Y}_{k+p}\} \tag{16}$$

式中：p 为单侧所考虑的邻近帧范围。

Wu 指出在不同混响时间条件下，STFT 的帧移和声音内容窗口 p 的选取对于去混响效果有一定的影响，并列出了不同混响时间条件下，上述参数的最优取值。该算法需要根据估计的混响时间查找相应的最优参数设定以获得较好的去混响效果[52]。Guzewich [29]重现了 Wu 的方法[52]，并指出在语音混响不严重、接近纯净语音的情况下，Han 的处理方法可能反而会使得处理得到的语音信号质量下降，而 Wu 的方法由于考虑了混响时间，使得上述问题得到了改善。此外，针对语音信号时序性的另一个解决思路是采用具备时序结构神经网络，LSTM 就是比较有代表性的时序网络包括递归神经网络[53]。不同于只进行层级堆叠的 DAE 与 DNN，LSTM 具备层间传递的时间递归结构，这些递归结构能够存储时序信息。此外，LSTM 还克服了循环神经网络 (Recurrent neural network, RNN) 时常出现的梯度消失问题(Vanishing gradient problem)且能够实现与 DNN 相仿甚至更好的效果[50]。图3给出了计算从输入序列 $\boldsymbol{x}=(x_1, \cdots, x_t)$ 到输出序列 $\boldsymbol{m}=(m_1, \cdots, m_t)$ 的映射关系的基本细胞单元，包括输入门(Input gate) i，遗忘门(Forget gate) f，输出门(Output gate) o，以及存储单元中存储的数据 s。

需要注意的是，与独立计算各个输出的前馈网络独立不同，LSTM 的映射需要从 $t=1$ 到 T 以递归的方式进行计算，相应的计算公式为

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{ix}\boldsymbol{x}_t + \boldsymbol{W}_{im}\boldsymbol{m}_{t-1} + \boldsymbol{W}_{is}\boldsymbol{s}_{t-1} + \boldsymbol{b}_i) \tag{17}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{fx}\boldsymbol{x}_t + \boldsymbol{W}_{fm}\boldsymbol{m}_{t-1} + \boldsymbol{W}_{fs}\boldsymbol{s}_{t-1} + \boldsymbol{b}_f) \tag{18}$$

$$\boldsymbol{s}_t = \boldsymbol{f}_t \odot \boldsymbol{s}_{t-1} + \boldsymbol{i}_t \odot \tanh(\boldsymbol{W}_{sx}\boldsymbol{x}_t + \boldsymbol{W}_{sm}\boldsymbol{m}_{t-1} + \boldsymbol{b}_c) \tag{19}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{ox}\boldsymbol{x}_t + \boldsymbol{W}_{om}\boldsymbol{m}_{t-1} + \boldsymbol{W}_{os}\boldsymbol{s}_t + \boldsymbol{b}_c) \tag{20}$$

$$\boldsymbol{m}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{s}_t) \tag{21}$$

式中，$\boldsymbol{W}$ 为权重矩阵(例如 $\boldsymbol{W}_{ix}$ 表示从输入门到输入之间的权重矩阵)，偏置向量记作 $\boldsymbol{b}$(例如输入门)，$\boldsymbol{m}$

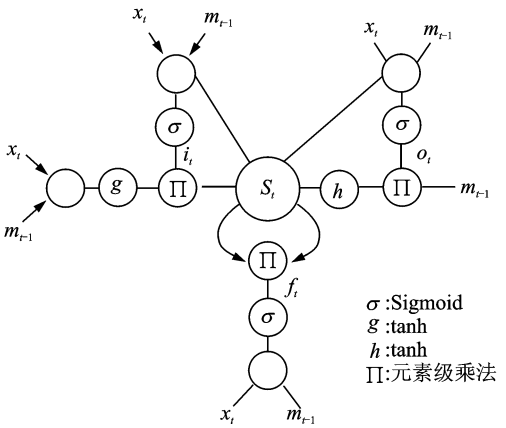


图 3 长短时记忆网络

Fig. 3 Long short-term memory (LSTM)

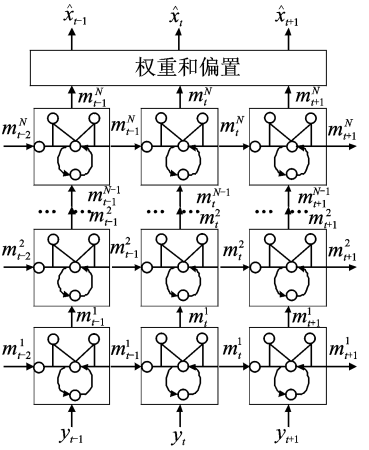


图 4 基于长短时记忆网络的特征级去混响算法

Fig. 4 Feature level speech dereverberation algorithm based on long short-term memory (LSTM)

为输出向量,σ 为 Sigmoid 函数,符号⊙表示元素对应相乘。LSTM 能够克服梯度消失问题的优势使得其能够处理具有极长时延的时序信号,并能够根据混响时长自适应做出调整。

图 3 展示的仅仅是 LSTM 中最基本的一个处理细胞单元,将众多这样的单元以一定的层次结构相联结就能够得到 LSTM 网络。基于 LSTM 网络的线性映射就可以开发出相应的去混响算法[50],基本架构如图 4 所示。

LSTM 网络的训练通过跨越时间的反向传播(Back propagation through time, BPTT)算法来实现。目标函数由纯净语音特征 $\boldsymbol{x}$ 和增强特征 $\hat{\boldsymbol{x}}$ 之间的均方误差定义为

$$\boldsymbol{L}=\sum_{t=T-T_L+1}^{T}\|\boldsymbol{x}_t-\hat{\boldsymbol{x}}_t\|^2 \tag{22}$$

理论上,随机梯度下降算法使用的梯度被应用于从 $t=1$ 到 T 的所有语音帧,但这在实际计算中往往不可行,通常的处理是在计算时忽略过于遥远的语音帧,只对固定数量的 T_L 个语音帧进行计算。

在时间轴 t 上超过 T_L 的语音帧被认为不会影响当前帧的估计,因此 T_L 往往需要根据训练数据库的混响时间来设定。

深度人工神经网络提供了一种非常灵活的非线性映射途径,基于其所设计的去混响算法也不是孤立的独立系统。事实上,不仅不同神经网络能够很方便地进行堆砌,并且可以将深度学习与很多传统成熟的架构相融合。例如,Mimura 使用 DAE 进行前端处理,并使用 DNN 来建模后端的声学模型,提升语音识别的性能[54];Weninger[48] 提出将传统的谱减法与神经网络进行早期融合的方案,将混响的指数衰减模型与 LSTM 网络联合使用来改善原有算法。如何更加有效地将成熟的去混响算法和深度学习强大的非线性建模能力有机结合起来是需要进一步研究的问题。

现有的基于深度学习的去混响算法大多基于"见多识广"的思路,通过大量的纯净语音与各种 RIR 条件下进行组合得到海量的训练数据,实现对深层人工神经网络的众多参数进行优化。当训练数据和测试数据相匹配时,深层非线性映射的结构往往能给出令人满意的结果,然而当处理未遇到的语音信号以及未知房间声学特性时,神经网络能否有效地进行泛化仍然有待进一步研究。

为了使神经网络获得较为理想的泛化能力,最直接的解决思路是使用更大的语音库以及更多的 RIR 信号来训练神经网络,然而这就会对网络的训练提出挑战。为了解决这个问题,一个解决思路是研

究出更高效的软硬件算法来加速神经网络的学习效率;另一个解决思路是在算法层面研究如何更高效地进行映射。事实上,目前基于深度学习的去混响算法使用的映射并不是十分高效。例如,假设纯净语音库包含 1 h 的语音信号,混响数据库包含 100 个 RIR 信号,那么训练时就需要将纯净语音信号与 RIR 信号分别进行卷积,从而得到 100 h 的混响语音信号作为先验信息来对神经网络的参数进行优化。然而,实质上作为先验信息的信号仅长达 1 h 左右,显然其数量级明显小于实际的训练样本。即目前这种不太高效的映射关系将会使得训练样本的实际数量显著上升,远远高于实际信号数量的总和。此问题会随着语音信号和 RIR 信号数量的持续增加而变得日益突出。研究如何在训练时将语音信号和 RIR 信号进行某种程度地"去耦合",从而实现更加高效的映射,也是一个值得进一步研究和讨论的问题。

另外,近期兴起的生成式对抗网络(Generative adversarial networks, GAN)为语音去混响提供了一个新的思路。基于生成模型的 GAN 能够估测模拟数据样本的潜在分布并根据这个分布生成新的数据样本。在机器视觉领域,已经能够通过 GAN 生成非常逼真的复杂图像[55]。基于 GAN 架构的语音处理刚刚起步,一些语音增强算法目前刚刚被提出[56, 57],而基于 GAN 架构的去混响算法目前尚未出现,值得进一步研究。

3 结束语

本文针对语音混响的不同方面进行了综述,包括混响声学模型、衡量混响声学特性的常用参数、自然界混响统计特性的最新发现、语音特征表示和评价去混响算法指标这 5 个方面。在此基础上,根据先验信息由少到多的发展规律,分 3 个层次对语音去混响的代表性算法分别进行论述,特别是针对近期出现的基于语音模型和深度学习的去混响算法,给出了比较系统的梳理和总结。尽管近年来,语音去混响问题取得了一些不错的效果,但是,仍然有一些问题没有能够很好地解决,未来针对语音去混响算法可能集中在以下几个方面:

(1)模型的泛化性。近年来,基于深度学习的语音增强算法展现出令人瞩目的性能。然而,常见的有监督模型在遇到未知的语音或者与训练样本不同的房间声学特性时,效果往往会下降。在匹配程度不好的情况下,如何解决有监督学习模型的泛化能力有待进一步研究。与此同时,以 GAN 为代表的生成模型能够逼近和模拟样本数据潜在的分布特性,为提升模型泛化能力提供了新的思路,值得进一步研究。

(2)训练样本的高效利用。当前的语音去混响算法常常是基于直接的非线性映射,这样的非线性映射对于训练样本的利用往往并不十分高效。同时,人类的听觉系统能够同时从混响声源中得出声源信息和房间混响特性,针对其机理的研究可能能够促进对训练样本进行更充分的利用。

(3)去混响系统的实时性研究。由于混响造成的时延,混响效应往往会跨越很多个语音帧。近年来大量涌现出的基于深度学习的语音去混响系统虽然展现出不错的性能,但是这些算法大多是有延迟或离线的算法,并不能对混响语音信号进行实时处理。这限制了去混响算法的实际应用。如何在不降低去混响效果的同时实现算法的实时处理,也值得继续探索。

参考文献:

[1] Wolfel M, McDonough J. Distant speech recognition [M]. Great Briton, Hoboken, NJ: Wiley, 2009.

[2] Sun M, Li Y, Gemmeke J F, et al. Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback-Leibler divergence [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(7): 1233-1242.

[3] 李轶南, 张雄伟, 贾冲, 等. 稀疏低秩噪声模型下无监督实时单通道语音增强算法[J]. 声学学报, 2015, 40(4): 607-614.
Li Yinan, Zhang Xiongwei, Jia Chong, et al. Unsupervised real-time single channel speech enhancement with sparse low-rank and noise model [J], ACTA Acustica, 2015, 40(4): 607-614.

[4] Yoshioka T, Sehr A, Delcroix M, et al, Making machines understand us in reverberant rooms: Robustness against reverber-

ation for automatic speech recognition [J]. IEEE Signal Processing Magazine, 2012, 29(6): 114-126.
[5] 赵力, 梁瑞宇, 谢跃, 等. 语音测谎技术研究现状与展望[J]. 数据采集与处理, 2017, 32(2):246-258.
Zhao Li, Liang Ruiyu, Xie Yue, et al. Progress and outlook of lie detection technique in speech [J]. Journal of Data Acquisition and Processing, 2017, 32(2): 246-258.
[6] Sun M, Zhang X, van Hamme H, et al. Unseen noise estimation using separable deep auto encoder for speech enhancement [J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2016, 24(1): 93-104.
[7] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups [J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.
[8] Mosayyebpour S, Esmaeili M, Gulliver T A. Single-microphone early and late reverberation suppression in noisy speech [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(2): 322-335.
[9] Kinoshita K, Delcroix M, Habets E, et al. The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech [C] // IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz, NY: IEEE Press, 2013:1-4.
[10] Kinoshita K, Delcroix D, Gannot S, et al. A summary of the REVERB challenge: State-of-the-are and remaining challenges in reverberant speech processing research [J]. EURASIP Journal on Advances in Signal Processing, 2012, 60(6): 2882-2898.
[11] Kuttruff H. Room acoustics [M]. 5th ed. New York, USA: Taylor & Francis, 2009.
[12] Tsilfidis A, Mporas I, Mourjopoulos J, et al. Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing [J]. Computer Speech and Language, 2013, 27(1): 380-395.
[13] Parada P P, Sharma D, Lainez J, et al. A single-channel non-intrusive C50 estimator correlated with speech recognition performance [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(4): 719-732.
[14] Traer J, McDermott J H. Statistics of natural reverberation enable perceptual separation of sound and space[J]. Proceedings of the National Academy of Sciences, 2016, 113(48): E7856-E7865.
[15] Delfarah M, Wang D L. Features for masking-based monaural speech separation in reverberant conditions [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(5): 1085-1094.
[16] Wang D L, Brown G L. Computational auditory scene analysis: Principles, algorithms, and applications [M]. Hoboken, NJ, USA: Wiley IEEE Press, 2006.
[17] Shao Y, Srinivasan S, Wang D L. Incorporating auditory feature uncertainties in robust speaker identification [C] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hawaii, USA: IEEE Press, 2007:277-280.
[18] Chen J, Wang Y, Wang D L. A feature study for classificationbased speech separation at low signal-to-noise ratios [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1993-2002.
[19] Maganti H K, Matassoni M. An auditory based modulation spectral feature for reverberant speech recognition[C] // INTERSPEECH 2010. Brighton, U K: International Speech Communication Association (ISCA) Press, 2010: 570-573.
[20] Zhang X, Zhang H, Nie S, 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.
[21] Hermansky H. Perceptual linear predictive (PLP) analysis of speech [J]. Journal of the Acoustical Society of America, 1990, 87(4): 1738-1752.
[22] Hermansky H, Morgan N. RASTA processing of speech[J]. IEEE Transactions on Audio, Speech, and Language Processing, 1994, 2(4): 578-589.
[23] Kim G, Lu Y, Hu Y, et al. An algorithm that improves speech intelligibility in noise for normal-hearing listeners [J]. Journal of the Acoustical Society of America, 2009, 126(3): 1486-1494.
[24] Schadler M R, Meyer B T, Kollmeier B. Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition[J]. Journal of the Acoustical Society of America, 2012, 131(5): 4134-4151.
[25] Yuo K H, Wang H C. Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences [J]. Speech Communication, 1999, 28(1): 13-24.
[26] Ikbal S, Misra H, Bourlard H. Phase autocorrelation (PAC) derived robust speech features [C] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hong Kong, China: IEEE Press, 2003: II-133-II-136.
[27] 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.
[28] Kim C, Stern R M. Power-normalized cepstral coefficients (PNCC) for robust speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(7): 1315-1329.

[29] Guzewich P, Zahorian S. Improving speaker verification for reverberant conditions with deep neural network dereverberation processing [C] // INTERSPEECH. Stockholm, Sweden: International Speech Communication Association (ISCA) Press, 2017: 171-175.

[30] Hu Y, Loizou P C. Evaluation of objective quality measures for speech enhancement [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(1): 229-238.

[31] Falk T H, Zheng C, Chan W Y. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(7): 1766-1774.

[32] Rix A, Beerends J, Hollier M, et al. Perceptual evaluation of speech quality (PESQ) —A new method for speech quality assessment of telephone networks and codes [C] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Salt Lake City, Utah, USA: IEEE, 2001: 749-752.

[33] Taal C H, Hendriks R C, Heusdens R H, et al. An algorithm for intelligibility prediction of time-frequency weighted noisy speech [J]. IEEE Transactions on Audio, Speech and Language Processing, 2011(1): 2125-2136.

[34] Mowlace P, Saeidi R, Christensen M G I. Subjective and objective quality assessment of single-channel speech separation algorithms [C] // Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. [S. l.]: IEEE, 2012: 69-72.

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

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

[37] Yoshioka T, Nakatani T, Miyoshi M, et al. Blind separation and dereverberation of speech mixtures by joint optimization [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(1): 69-84.

[38] Smaragdis P, Fevotte C, Mysore G J, et al. Static and dynamic source separation using nonnegative matrix factorizations: A unified view [J]. IEEE Signal Processing Magazine, 2014, 31(3): 66-75.

[39] 张雄伟, 李轶南, 时文华, 等. 非负组合模型及其在声源分离中的应用[J]. 数据采集与处理, 2017, 32(2): 266-277.
Zhang Xiongwei, Li Yinan, Shi Wenhua, et al. Non-negative compositional models and its application in acoustic source separation [J]. Journal of Data Acquisition and Processing, 2017, 32(2): 266-277.

[40] Kameoka H, Nakatani T, Yoshioka T. Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms [C] // IEEE International Conference on Acoustics, Speech and Signal Processing. [S. l.]: IEEE, 2009: 45-48.

[41] Mohammadiha N, Smaragdis P, Doclo S. Joint acoustic and spectral modeling for speech dereverberation using non-negative representations [C] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brisbane, Australia: IEEE, 2015: 4410-4414.

[42] Mohammadiha N, Doclo S. Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(2): 276-289.

[43] Baby D, Van Hamme H. Supervised speech dereverberation in noisy environments using exemplar-based sparse representations [C] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China: IEEE Press, 2016: 156-160.

[44] Mohanan N, Velmurugan R, Rao P. Speech dereverberation using NMF with regularized room impulse response [C] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, USA: IEEE Press, 2017: 4955-4959.

[45] Liang D, Hoffman M D, Mysore G J. Speech dereverberation using a learned speech model [C] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brisbane, Australia: IEEE Press, 2015: 1871-1875.

[46] Feng X, Zhang Y, Glass J. Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition [C] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE Press, 2014: 1759-1763.

[47] Verwimp L, Pelemans J, Van hamme H, et al. Character-word LSTM language models [C] // European Chapter of the Association for Computational Linguistics. Valencia, Spain: EACL Press, 2017: 417-442.

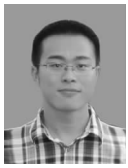
[48] Weninger F, Watanabe S, Tachioka Y, et al. Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition [C] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, 2014: 4623-4627.

[49] Han K, Wang Y, Wang D L, 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.

[50] Mimura M, Sakai S, Kawahara T. Speech dereverberation using long short-term memory [C] // INTERSPEECH. Dresden, Germany: International Speech Communication Association (ISCA) Press, 2015: 2435-2439.
[51] Williamson D S, Wang D L. Time-frequency masking in the complex domain for speech dereverberation and denoising [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(7): 1492-1501.
[52] Wu B, Li K, Yang M, 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): 102-111.
[53] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.
[54] Mimura M, Sakai S, Kawahara T. Deep autoencoders augmented with phone-class feature for reverberant speech recognition [C] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brisbane, Australia: IEEE Press, 2015: 4365-4369.
[55] 王坤峰, 苟超, 段艳杰, 等. 生成式对抗网络 GAN 的研究进展与展望[J]. 自动化学报, 2017, 43(3): 321-332.
Wang Kunfeng, Gou Chao, Duan Yanjie, et al. Generative adversarial networks: The state of the art and beyond [J]. Acta Automatica Sinica, 2017, 43(3): 321-332.
[56] Pascual S, Bonafonte A, Serrà J. SEGAN: Speech enhancement generative adversarial network [C] // Inter Speech, Stockholm. Sweden: International Speech Communication Association (ISCA) Press, 2017: 3642-3646.
[57] Michelsanti D, Tan Z. Conditional generative adversarial networks for speech Enhancement and noise-robust speaker verification [C] // INTERSPEECH. Stockholm, Sweden: International Speech Communication Association (ISCA) Press, 2017: 2008-2012.

作者简介:

张雄伟(1965-),男,教授,博士生导师,研究方向:语音与图像处理、多媒体信息处理,E-mail: xwzhang9898@163.com。



李轶南(1988-),男,博士研究生,研究方向:多媒体信息处理、模式识别。

郑昌艳(1990-),女,博士研究生,研究方向:多媒体信息处理、深度学习。

曹铁勇(1971-),男,教授,研究方向:智能信息处理。

孙 蒙(1984-),男,博士,讲师,研究方向:多媒体信息处理、模式识别。

闵 刚(1983-),男,博士,讲师,研究方向:语音信息处理、数字通信。