文章编号:1002-8684(2006)05-0063-05

语音信号去混响原理与技术

·论文·

吴佳栋[1], 陈光冶[2]
(1. 上海交通大学 电子信息与电气工程学院, 上海 200030;
2. 上海交通大学 振动冲击与噪声国家重点实验室, 上海 200030)

【摘 要】语音信号去混响技术在通信、语言识别等方面有重要应用。介绍了国内外相关研究动态和方法,阐述了声音混响过程和倒谱法去混响原理,简要介绍了传声器阵列-倒谱法去混响技术。

【关键词】去混响; 语音信号; 阵列信号处理; 信号增强
【中图分类号】TN912.34 【文献标识码】A

Principle and Technologies of Speech Signal Dereverbration

WU Jia-dong[1], CHEN Guang-zhi[2]
(1. School of Electrical and Information Engineering, Shanghai Jiaotong University, Shanghai 200030, China;
2. State Key Laboratory of Vibration, Shock, and Noise, Shanghai Jiaotong University, Shanghai 200030, China)

【Abstract】 Dereverberation of speech signal is applied to many areas including communication and speech recognition. The research progress on the problem is reviewed. The phenomena of sound reverberation and the principle of dereverberation with cepstrum transform are addressed. A dereverberation method using microphone array in cepstrum domain is briefly introduced.

【Key words】dereverberation; speech signal; array signal processing; signal enhancement

1 引言

通常在声音信号采集或录制的情况下, 传声器除了接收到所需要的声源发射声波直接到达的部分外,还会接收声源发出的、经过其它途径传递而到达的声波,以及所在环境其它声源产生的不需要的声波(即背景噪声)。在声学上,延迟时间达到约 50 ms 以上的反射波称为回声,其余的反射波产生的效应称为混响。混响现象将对期望声信号的接收效果产生影响。一些建筑,如音乐厅和教堂,需要适度的混响作用而使音乐更加动听。但在许多场合,混响往往会带来干扰,导致声学接收系统性能变差。例如,混响会导致语音识别系统性能显著下降;在远程会议、免提电话、助听器和移动通信中,混响作用主要带来负面影响。当混响严重时,这些系统甚至无法正常发挥功能,因此,如何减少混响对声音接收系统的影响, 即去混响(dereverberation),是一个非常重要的课题。

2 混响消除技术发展概况

语音信号去混响可归为稳健语音识别涉及的问题之一[1],其与电话系统回声问题也有关联。在同时使用扬声器和传声器的场合,如远程会议和电话系统,扬声器和传声器之间存在声波传递途径, 这会引起送话方的回声干扰。针对电话系统的回声问题,国内有大量的研究报告,如参考文献[2],但消除回声干扰问题与从混响环境下检测声信号中还原所期望声源的原始信号问题不同。从所涉及的系统响应辨识问题而言,前者的系统输入是已知的,而后者的系统输入是未知的。

按照使用传声器数量分类, 去混响系统主要分为单传声器系统与传声器阵列系统。

2.1 单传声器系统

单传声器系统中的直接方法是采用逆向滤波[3]。若通过测量获得房间条件下扬声器到传声器之间的冲击响应, 使用一个逆向滤波器或者最小均方差解卷积算法就可以去除混响。但由于一般房间的冲击响应为非最小相位,具有非稳定的逆[4],因此基于逆滤波的方法可应用的实际场合很有限[5]。在实时应用中,例如,随着说话人的位置变化(如人的走动等)或者房间内物体位置变化(如门的开闭等),甚至对于每一帧信号,冲击响应都在改变,这就需要实时测量和跟踪,问题因此变得更加复杂。单传声器系统的另一种途径是倒谱技术[6],

依据的是时域解卷积相应于倒谱域的相减，而语音信号的复倒谱通常集中于倒谱域的原点附近，回声则由倒谱域扩展至远离原点的脉冲组成。因此，在倒谱域用的低时窗“滤波”可以削减混响。单传声器系统应用倒谱技术的主要问题是分段误差对复倒谱计算的影响，以及使用指数加权引起的数值误差，语音和回声在倒谱域重叠，用复倒谱进行重构时相位解卷绕困难。

由于单传声器系统具有硬件简单的优点，因此不断有改进的方法提出。文献[7]描述了一种修正的倒谱方法估计房间响应，通过对有语音的段加指数窗，使信号具有最小相位特性，并对一些这样的帧作平均。文献[8]提出了一种通过检测浊音的基音的延迟变化来估计混响时间，然后利用得到的混响时间设计算法来消除混响。文献[9]提出了一种单传声器盲解卷积算法，其中谐波结构被用来近似混响声音中的直达声部分。近似直达声与混响声的平均比例被用作去混响运算符，可以得出逆传递函数。

2.2 传声器阵列系统

单传声器系统去混响技术只利用声场中接收位置一点的声信号时间和变换域的特性。而多传声器阵列系统能利用声场的空间特性，其主要优点是由阵列带来的接收方向性除了能直接提高信号与混响声能比之外，同时还对本底噪声有显著的抑制作用。但阵列系统的硬件复杂度高，数据处理量成倍增加，对计算速度有较高要求。但随着计算机技术的发展，采用阵列的去混响技术受到更多重视。

早期应用阵列主要通过应用延时-求和方法形成波束，将波束最大方向指向期望声源，波束零点朝向噪声源传播方向，来抑制噪声和混响[10]。在语音宽带范围形成一致波束指向是其中比较困难的问题。文献[11]提出了两个传声器系统去混响的算法，利用不同位置上延时混响声的非相关性质通过补偿两个传声器通道之间的相位和幅度差，将两路信号相关叠加。文献[12]提出了利用两个相隔足够远传声器信号的倒谱系数差，进行双倒谱迭代重构，获得通道的倒谱系数并重构源信号的方法。但是，由于仅用波束形成技术对相干性噪声抑制的能力有限，而混响前期反射波更接近相干，因此，需要针对阵列波束形成技术之外的其它信号处理技术进行研究。

多通道盲解卷积的研究比较活跃。文献[13]提出了一种确定性时域算法，在最小方差意义上估计各通道的冲击响应，获得对源信号的一个估计，再在最小范数意义下计算各解卷积滤波器。文献[14]采用了基于贝叶斯方法的盲解卷积算法估计声源信号；针对房间冲击响应渐近趋向零导致盲通道识别病态问题，采取贝叶斯最大后验估计直接计算源信号，避开对未知通道参数的要求，所需的初始样本采用随机马尔科夫-蒙特卡罗方法生成。

基于语音模型去混响方法是一种能很好适应语言非平稳性的方法，其基本点是：混响导致接收的浊音激励脉冲性质变化，从而影响语音清晰度。所采用的模型为：语音产生用全极点模型描述，同时认为混响和加性噪声在整个系统引入了零点，使接收语音信号中浊音激励脉冲性质发生变化，而不影响全极点滤波器；同时还假设各个通道之间的噪声和激励脉冲序列增加的回波脉冲互不相关，而声源本身的激励脉冲序列是不受环境影响的。方法是从一组受污损的激励信号中辨识干净的语言激励序列，然后只用增强的序列重构语音信号。具体步骤是：利用从各个通道导出的 LPC 残差，按照基音同步猝发准则（pitch-synchronous clustering criterion）辨别干净语音激励信号[15]。通过对脉冲在各个通道极大值的定位，获取隐藏的原来未受混响语音的激励脉冲结构。文献[16]进一步在系统中加入混响通道的一个粗的估计模型，获取初始反射波的近似时间和幅度，用以对各通道的 PLC 残差进行加权计算，强化信号。这种方法利用期望信号的性质，来抑制混响和加性噪声，具有随信号帧变化的适应性。

针对阵列的倒谱法具有很明显的实际效果[17]。基于最小相位分量较少受混响影响的实验观察，方法中将信号分解为一个最小相位分量（minimum-phase）和一个全通分量（all-pass），分别处理后再综合还原。在恢复声源语音的最小相位分量时，首先在倒谱域进行低时窗滤波，再对各个通道的最小相位分量进行空间平均；保留全通分量的相位信息，将传统波束形成技术用于传声器信号的全通分量来恢复原语声的全通分量；最终，综合复原的最小相位分量和全通分量获得去混响语声信号。

结合阵列的去混响方法除上述介绍的以外，还有匹配滤波法[18]和自适应滤波法[19]。

3 倒谱法去混响技术

3.1 混响环境下声音检测与倒谱域去混响原理

在房间内发声者和接收者之间的声传递基本过程如图 1 所示。直达声为接收者所期望听到的，在房间

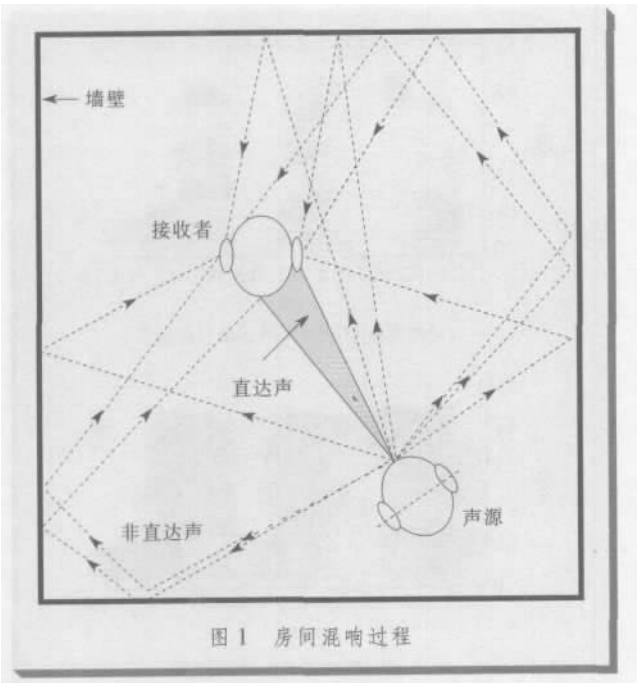


图 1 房间混响过程

的环境内，实际上，接收者只能听到直达声和非直达声混合在一起的声音。非直达部分即混响声，通过去混响技术中将其从接收端消除。这种混响声，可以看作是房间的冲击响应与声源发出声音（激励）的卷积结果，房间典型的冲击响应持续时间长度大约为 250 ms。

实际环境中，除混响的影响外，通常还有背景噪声干扰，它直接叠加在接收声信号上。所以，房间内传声器接收的声信号 $y(t)$ 的一般表达式为

$$y(t)=h(t)*x(t)+n(t) \quad (1)$$

其中，$x(t)$ 是声源发出的信号；$n(t)$ 是房间的其它噪声；$h(t)$ 是从发声者（声源）位置到接受者（传声器）位置内两点之间房间的冲击响应；t 表示时间，$*$ 表示卷积运算。

加性噪声干扰可以通过减谱等方法来消除。去混响主要考虑由于 $h(t)$ 与 $x(t)$ 卷积中的混响部分。忽略加性噪声的影响，以离散形式表示，传声器接收到的信号为

$$y(m)=\sum_{k=0}^{P-1}h(k)x(m-k) \quad (2)$$

其中 P 为采样长度。去混响问题就是上面卷积的逆问题，即期望从获得的 $h(m)$，解出 $x(m)$，而房间的冲击响应 $h(m)$ 并不预先知道。所以，去混响问题本质是一个盲解卷积问题。将式(2)变换到频域，可表示为

$$Y(\omega)=H(\omega)X(\omega) \quad (3)$$

两边再取对数，作逆傅里叶变换，给出复倒谱关系

$$\hat{y}(n)=F^{-1}[\ln Y(\omega)]=$$
$$F^{-1}[\ln H(\omega)]+F^{-1}[\ln H(\omega)]=\hat{h}(n)+\hat{x}(n) \quad (4)$$

实际测试结果表明，倒谱域语音信号的复倒谱 $\hat{x}(n)$ 通常主要分布在靠近原点的附近，因此，可以采用“低时窗”滤除混响对应部分 $\hat{h}(n)$ 的影响，然后再通过逆向操作，获得混响受到消减的声源信号。

3.2 阵列-倒谱分解处理去混响技术

若将传声器检测信号分解为最小相位和全通分量，由计算机对房间混响仿真数据以及实际测试显示，混响对全通分量的影响比对最小相位分量的影响要显著得多[17]，并且全通分量包含了声源位置的信息。所以考虑将信号分解为最小相位分量和全通分量分别进行不同的处理。从各个通道的最小相位分量通过低时窗去混响操作，和所有通道取均值操作去非相干噪声，来获取声源语音信号的时变部分；将传统波束形成技术用于各个通道的全通分量；来恢复声源语音信号的全通分量；最终，合成经过修正的最小相位分量和全通分量获得去混响语音信号。图 2 是方法的流程框图，$x_1(n)$ 即第 1 通道的输入，经 FFT 得到 $X_1(\omega)$。然后经 DHT 算出 $\hat{x}_{1,\min}(n)$，进而得到全通分量 $X_{1,all}(\omega)$ 和最小相位分量 $X_{1,\min}(\omega)$。其它通道也用类似方法计算，然后将 M 个通道的全通分量以及最小相位分量分别进行处理并合成为 $Y_{all}(\omega)$ 和 $Y_{\min}(\omega)$，最后得到的去混响输出 $y(n)$。通道的时延是使阵列形成对准声源方向的波束，需要采用时延估计算法确定[20]。如果进一步需要控制波束的宽度和带宽，还需要在延时后进行加权。要实现阵列在几个倍频程范围有一致的方向性，比较简便的方法

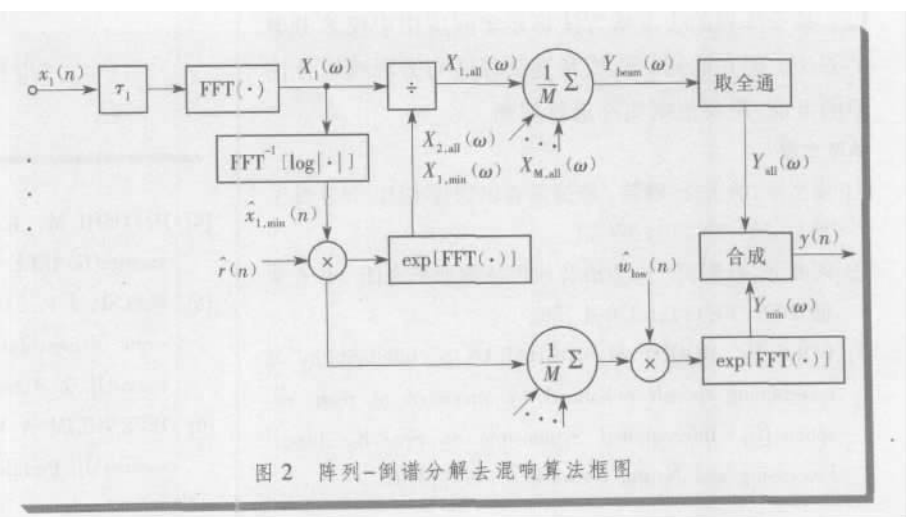


图 2 阵列-倒谱分解去混响算法框图

是采用传声器简谐布置的办法。图 2 中

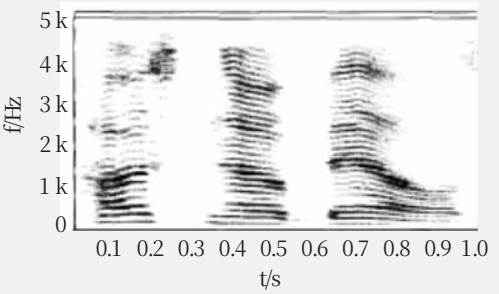
$$\hat{r}(n)=\begin{cases}2, & \frac{N}{2}>n>1\\ 1, & n=0,\frac{N}{2}\\ 0, & \frac{N}{2}<n<N-1\end{cases} \tag{5}$$

其中，N 为序列长度，作用是使复倒谱为单边序列。处理步骤中，对所有通道的全通分量求和并取平均的操作，实际上是一个经典的波束形成运算，可以使最后得到的全通分量获得阵列的方向性增益；对所有通道的最小相位分量在倒谱域求和并取平均的操作，相当于 Z 域求几何平均，因而结果保持最小相位性质，这种操作在整个倒谱“频率”范围削减了非相干的混响成分，而用低时窗序列$\hat{W}_{low}(n)$加权进一步消除在高倒谱“频率”范围内的混响成分。

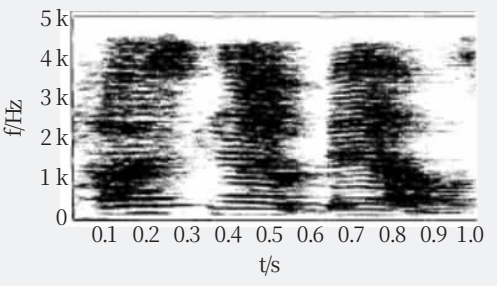
图 3 给出同一语音信号语谱图的比较。从图中可以看到，经过阵列-倒谱分解处理的语音信号语谱图与消声室干净声源语音信号语谱图非常接近，比传统的阵列方法有显著的改进。

4 结论

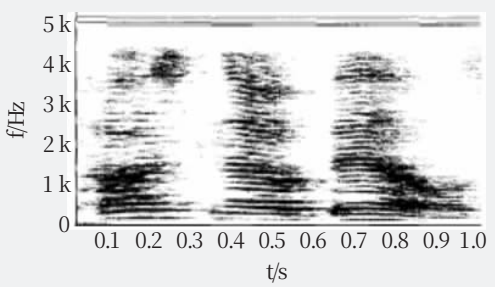
语音信号去混响在通信和语音识别等应用方面都具有重要意义。单个传声器系统的房间响应进行逆向滤波和倒谱处理的方法，由于房间冲击响应并不总是可逆的，或响应的时变跟踪困难，或重构的相位缠绕问题而受到限制。而所有基于传声器阵列的去混响方法，如基于语音模型的方法和倒谱方法等，不仅可以获得由使用阵列带来的方向性增益，而且对非相干噪声的抑制作用具有较好的鲁棒性并且在去混响效果和期望语音信号的失真度均衡方面更大的灵活性等优点。所以，基于阵列的去混响方法将是实际应用中应该考虑的方法，但在阵列的规模和复杂程度与去混响综合性能两方面，需要根据实际进行权衡。



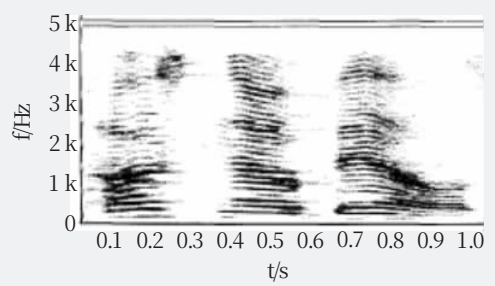
(a) 消声室干净声源语音信号语谱图



(b) 混响语音信号语谱图
(混响时间 0.3 s, 房间墙面反射系数 0.8)



(c) 经传统阵列处理的语谱图



(d) 采用阵列-倒谱分解处理的语谱图

图 3 语谱图比较[20]

参考文献

[1] 姚文冰，姚天任，韩涛. 稳健语音识别技术[J]. 计算机工程与应用，2002(7)：69-71.

[2] 阎兆立，杜利民. 电话语音回声消除的研究[J]. 电子学报，2002，30(11)：1 726-1 728.

[3] COLE D, MOODY M, SRIDHARAN S. Intelligibility of reverberant speech enhanced by inversion of room response[C]. International Symposium on Speech, Image Processing and Neural Networks, 1994: 13-16.

[4] MIYOSHI M, KANEDA Y. Inverse filtering of room acoustics[J]. IEEE Trans. ASSP, 1988, 36(2): 145-152.

[5] WALSH J P. On limitations of minimum mean-square error deconvolution in deriving impulse response of rooms[J]. J. Acoust. Soc. Amer., 1985, 77(2): 547-556.

[6] OPPENHEIM A V, SCHAFER R W. Digital signal processing[M]. Prentice Hall Inc., 1975.

[7] BEES D, BLOSTEIN M, KABAL P. Reverberant speech enhancement using cepstral processing[C]. IEEE ICASSP-91, 1991: 977-980.

[8] WU M. A One-microphone algorithm for reverberant speech enhancement[J]. Proc. of ICASSP, 2003, 1: 92-95.

[9] NAKATANI T. Blind dereverberation of single channel speech signal based on harmonic structure[J]. Proc. of ICASSP, 2003, 1: 892-895.

[10] LANAGAN J L F, JOHNSTON J D, ZAHN R. Computer-steered microphone arrays for sound transduction in large room[J]. J.Acoust. Soc. Amer., 1985, 78(5): 1 508-1 518.

[11] ALLEN J B. Short term spectral analysis, synthesis, and modification by discrete Fourier transform[J]. IEEE Trans. Acoust. Speech Signal Process., 1977, 25: 235-238.

[12] PETROPULU A P, NIKIAS C L N. Blind deconvolution using signal reconstruction from partial higher order cepstral information[J]. IEEE Trans. On Signal Processing, 1993, 41(6): 2 088-2 095.

[13] LAURENT C. Blind deconvolution for multi-microphone speech[C]. Proc. 3rd IEEE Benelux Signal Processing Symposium (SPS-2002), Leuven, Belgium, March, 2002: 1-4.

[14] DALY M J, REILLY J P. Blind deconvolution using bayesian methods with application to the dereverberation of speech[J]. Proc. of ICASSP 2004, 2: 1 009-1 012.

[15] BRANDSTEIN M. On the use of explicit speech modeling in microphone array applications[J]. Pro. of ICASSP, 1998, 6: 3 613-3 616.

[16] GRIEBEL S M, BRANDSTEIN M S. Microphone array speech dereverberation using coarse channel modeling [J]. Proc. of ICASSP, 2001, 1: 201-204.

[17] LIU Q G, CHAMPAGNE B, KABAL P. A microphone array processing technique for speech enhancement in a reverberant space[J]. Speech Communication, 1996, 18: 317-334.

[18] FLANAGAN J, SURENDRAN A, JAN E. Spatially selective sound capture for speech and audio processing [J]. Speech Communication, 1993, 13(1-2): 207-222.

[19] GILLESPIE B W. Speech deverberation via maximum-kurtosis subband adaptive filtering[J]. Proc. of ICASSP, 2001, 6: 3 701-3 704.

[20] CARTER C C. Coherence and Time Delay Estimation [M]. New York: IEEE Press, 1993.

[**收稿日期**] 2006-03-22

(**上接第** 62 **页**) Northern Ireland, 2000: 189-194.

[6] WHITESIDE SP. Simulated emotions: an acoustic study of voice and perturbation measures[C]. In: Proceedings of the 5th International Conference on Spoken Language Processing ICSLP 98, Sydney, Australia, 1998, 3: 699-703.

[7] PEREIRA C, WATSON C. Some acoustic characteristics of emation[C]. In: Proceedings of the 5th International Conference on Spoken Language Processing ICSLP 98, Sydney, Australia, 1998, 3: 927-930.

[8] AMIR N, RON S. Towards an automatic classification of emotions in speech[C]. In: Proceedings of the 5th International Conference on Spoken Language Processing ICSLP 98, Sydney, Australia, 1998, 3: 555-558.

[9] CAHN J E. Generating expression in synthesized speech [D]. Cambridge: Massachusetts Institute of Technology, 1989.

[10] MURRAY I R, ARNOTT J L. Implementation and testing of a system for producing emotion-by-rule in synthetic speech[J]. Speech Communication. 1995, 16: 369-390.

[11] IIDA A, CAMPBELL N, HIGUCHI F. A corpus-based speech synthesis system with emotion[J]. Speech Communication, 2003, 40: 161-187.

[12] NAKATSU R, NICHOLSON J, TOSA N. Emotion recognition and its application to computer agents with spontaneous interactive capabilities[J]. Knowledge-Based Systems, 2000, 13: 497-504.

[13] BOU-Ghazale S, HANSEN JHL, HMM-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress[J]. IEEE Trans. on Speech and Audio Processing, 1998, 6(3): 201-216.

作者简介

韩纪庆, 教授, 博士生导师, 主要研究领域为语音信号处理; 邵艳秋, 博士研究生, 主要研究方向为语音合成。

[**收稿日期**] 2006-01-25