ISSN 1000-0054 清华大学学报（自然科学版）2013 年 第 53 卷 第 6 期 39/40
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2013, Vol. 53, No. 6 917-920

基于空间声场扩散信息的混响抑制方法

王晓飞， 姜开宇， 国雁萌， 付 强， 颜永红

（中国科学院 声学研究所，语言声学与内容理解重点实验室，北京 100190）

摘 要：在远讲语音应用中，房间混响严重影响了语音的质量和主观听觉感受。该文利用双通道混响语音信号，根据语音直达声和混响声所反映出的声场扩散信息，提出一种基于空间声场扩散信息的时频递归平均混响功率谱估计方法，并通过谱增强的方法实现对混响的有效抑制。该文提出的算法在实录房间冲击响应（room impulse response，RIR）上实现了混响环境中语音增强，算法在分段信混比（segmental signal-to-reverberation ratio，SSRR）、对数谱距离（log spectral distortion，LSD）以及主观语音质量评估（perceptual evaluation of speech quality-mean opinion score，PESQ-MOS）方面都表现出性能的提升。

关键词：语音增强；混响抑制；声场扩散信息；功率谱

中图分类号：TN 912.3 **文献标志码**：A

文章编号：1000-0054(2013)06-0917-04

Reverberation suppression method based on diffuse information in a room sound field

WANG Xiaofei, JIANG Kaiyu, GUO Yanmeng, FU Qiang, YAN Yonghong

(Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Reverberation suppression can be used to improve speech quality and subjective hearing responses. This article describes a method to estimate the power spectrum of reverberation from reverberated signals of dual microphones based on diffuse information in the room sound field. This method takes full advantage of diffuse information reflected by direct sounds and reverberated sounds for speech enhancement. The system performance is evaluated based on indexes for the segmental signal-to-reverberation ratio (SSRR), log spectrum distance (LSD) and perceptual evaluation of speech quality-mean opinion score (PESQ-MOS) based on actual room impulse responses with the results showing the sound quality improvement.

Key words: speech enhancement; reverberation suppression; diffuse information of sound field; power spectrum

当前，随着手持设备、笔记本电脑、电视、VoIP 通信等免提语音应用的广泛发展，高质量的语音通信越来越受到重视。在远讲应用中，除了声学回波和环境噪声，因房间的反射而产生的语音混响也会严重降低语音的质量和可懂度。

混响所产生的效应在听感角度表现为盒子效应和说话人疏远效应[1]，在语谱图上表现为频谱染色效应和重叠掩蔽效应[2]，在时频上的拖尾会破坏语音信号的频谱包络和精细结构。现有混响抑制方法大致分为以下 3 类：1）波束形成方法，2）盲系统辨识方法，3）基于语音增强的方法。其中，基于语音增强的方法由于其在复杂环境中所表现出的鲁棒性，得到学者的青睐。

Habets 等[3]在 Lebart 等[4]的研究基础之上，建立广义统计混响模型，实现了晚期混响的谱增强，取得了不错的效果。该模型的缺点是依赖房间参数 T60（即房间混响时间）的准确估计[5]。Allen 等[6]最早利用双通道幅度平方相干函数（magnitude-squared coherence，MSC）的空间信息构建滤波器实现混响抑制，Jeub 等[7]结合 McCowan 等[8]的声场建模方法建立了直达声混响声能量比（direct-to-reverberant ratio，DRR）和 MSC 的关系，构建基于双耳的 Wiener 滤波器。Li[2]提出一种扩散程度估计方法，根据每个时频点受混响影响程度控制 Wiener 滤波器增益。

在此基础上，从实际应用出发，本文提出了一种采用双麦克风的谱增强方法，充分利用混响直达声和混响声在声场分类中所表现出的不同特性，引入空间

收稿日期：2013-04-27

基金项目：国家自然科学基金项目（10925419，90920302，61271426，61072124，11074275，11161140319，91120001）；
中国科学院战略性先导科技专项（XDA06030100，XDA06030500）；
国家“八六三”高技术项目（2012AA012503）；
中科院重点部署项目（KGZD-EW-103-2）

作者简介：王晓飞（1987—），男（汉），山东，博士研究生。

通信作者：付强，研究员，E-mail：qfu@hccl.ioa.ac.cn

声场扩散信息,根据时频点受混响影响程度,构建帧间的先验信息,采用时频递归平均方法更新混响声功率谱,并通过谱减法实现混响抑制。该方法不需要估计任何房间参数,因而不会因估计误差而导致性能下降,具有较强的鲁棒性。

1 问题描述

1.1 信号模型

双通道麦克风采集到的第 $m(m=1,2)$ 路语音信号用 $x_m(n)$ 表示, n 为样点数,

$$x_m(n)=\sum_{l=0}^{T_d f_s} s(n-l)h_m(l)+\sum_{l=T_d f_s+1}^{\infty} s(n-l)h_m(l)+v_m(n). \quad (1)$$

$s(n)$、$h(n)$和 $v(n)$分别表示目标语音、房间声学冲击响应和扩散噪声; T_d 表示直达声、混响声分界时刻; f_s 表示采样率。由此, $x_m(n)$前两部分别表示直达声部分 $x_{m,d}(n)$和混响声部分 $x_{m,r}(n)$。式(1)经过短时 Fourier 变换(STFT)为

$$X_m(l,k)=X_{m,d}(l,k)+X_{m,r}(l,k)+V_m(l,k). \quad (2)$$

l 和 k 分别表示第 l 帧和第 k 频带,并假设语音的直达声、混响声和扩散噪声之间两两不相关。

1.2 基于声场模型的直达声、混响声表达

Jeub 等[7]分析了直达声和混响声在 MSC 上的区别,直达声同混响声具有较强的区分性。

两通道复相干函数可以表示为

$$\Gamma_{X_1X_2}(e^{j\Omega})=\frac{\Phi_{X_1X_2}(e^{j\Omega})}{\sqrt{\Phi_{X_1X_2}(e^{j\Omega})\cdot\Phi_{X_1X_2}(e^{j\Omega})}}. \quad (3)$$

其中: Φ 为互(自)功率谱, Ω 为角频率。

对于直达声部分可以用相干声场进行建模,

$$\Gamma^{direct}_{X_1X_2}(e^{j\Omega})=e^{-j\Omega f_s d\cos\theta/c}. \quad (4)$$

d 为传声器间距, θ 为远场入射方向, c 为声音在空气中的传播速度。双通道直达声自、互功率谱分别表示为:

$$\Phi^{direct}_{X_1X_1}(e^{j\Omega})=\Phi^{direct}_{X_2X_2}(e^{j\Omega})=\Phi^{direct}(e^{j\Omega}), \quad (5)$$

$$\Phi^{direct}_{X_1X_2}(e^{j\Omega})=\Phi^{direct}(e^{j\Omega})e^{-j\Omega f_s d\cos\theta/c}. \quad (6)$$

对于混响声部分,由于其更接近扩散声场,同时假设加性噪声为各向同性的,因此近似使用扩散声场对扩散噪声和混响声统一建模[8],

$$\Gamma^{reverb}_{X_1X_2}(e^{j\Omega})=\mathrm{sinc}(\Omega f_s d/c). \quad (7)$$

双通道混响声自、互功率谱分别表示为:

$$\Phi^{reverb}_{X_1X_1}(e^{j\Omega})=\Phi^{reverb}_{X_2X_2}(e^{j\Omega})=\Phi^{reverb}(e^{j\Omega}), \quad (8)$$

$$\Phi^{reverb}_{X_1X_2}(e^{j\Omega})=\Phi^{reverb}(e^{j\Omega})\cdot\mathrm{sinc}(\Omega f_s d/c). \quad (9)$$

Φ^{direct}、Φ^{reverb}表示直达声、混响声功率谱。

1.3 DRR、MSC 及声场扩散程度之间的关系

由第 1.2 节, DRR 可以表示为

$$\mathrm{DRR}(e^{j\Omega})=\Phi^{direct}(e^{j\Omega})/\Phi^{reverb}(e^{j\Omega}). \quad (10)$$

根据 MSC 的定义,假设信号入射角度为 90°,那么可以由双通道观测信号建立 MSC 同 DRR 之间的关系[6],

$$\mathrm{MSC}(e^{j\Omega})=\frac{[\Phi^{direct}(e^{j\Omega})+\Phi^{reverb}(e^{j\Omega})\mathrm{sinc}(\Omega f_s d/c)]^2}{\Phi^{direct}(e^{j\Omega})+\Phi^{reverb}(e^{j\Omega})}. \quad (11)$$

每个时频点上 $\mathrm{DRR}(l,k)$的可以表示为

$$\mathrm{DRR}(l,k)=\frac{|\mathrm{sinc}(\Omega f_s d/c)|^2-\mathrm{MSC}(e^{j\Omega})}{\mathrm{MSC}(e^{j\Omega})-1}. \quad (12)$$

从式(12)可知, DRR 较大时,实际的声场更接近直达声场,而 DRR 较小时,实际声场则更接近扩散声场。Li[2]利用这样的信息给出了声场扩散程度的估计方法,利用扩散程度表征得到了每个时频点受混响的影响程度,扩散程度越大,受混响影响越强,如图 1 所示。

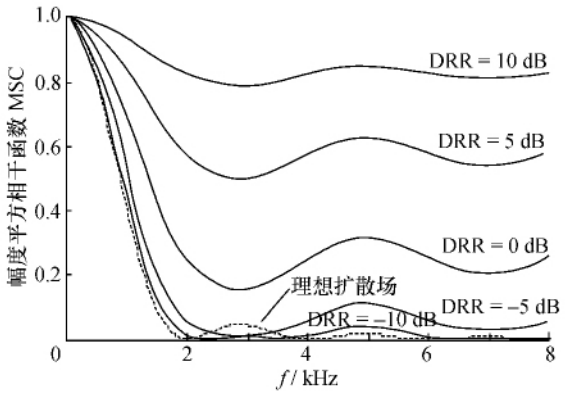

图 1 DRR 同 MSC 的关系[2]

2 基于声场扩散信息的谱增强

以带混响语音每个时频点受混响的影响程度作为先验信息可以有效地实现对混响声功率谱的估计,从而通过谱增强的方法实现混响声的抑制。

2.1 根据扩散信息的混响功率谱估计

采用式(13)所表述的时频递归平均方法构造混响声功率谱 λ。若为相干声场,按照 Direct 方式更新混响声功率谱;若为扩散声场,按照 Reverb 方式更新混响声功率谱:

$$\begin{cases}\mathrm{Direct}_{l,k}: \lambda^2_{l+1,k}=\alpha_r\lambda^2_{l,k},\\ \mathrm{Reverb}_{l,k}: \lambda^2_{l+1,k}=\alpha_d\alpha_r\lambda^2_{l,k}+(1-\alpha_d)|X_{l,k}|^2.\end{cases} \quad (13)$$

X 为双通道延迟相加结果, $\alpha_d(0<\alpha_d<1)$是帧间平滑

因子，α_r 为衰减因子。若当前帧的声场扩散程度为 $d_{l,k}$，那么：

$$\lambda^2_{l+1,k} = d_{l,k}(\alpha_d\alpha_r\lambda^2_{l,k} + (1-\alpha_d)\mid X_{l,k}\mid^2) + (1-d_{l,k})\alpha_r\lambda^2_{l,k}, \quad (14)$$

$$d_{l,k} = D(R_{l,k}\mid X_{l,k}, \xi_{l-1}). \quad (15)$$

ξ_{l-1} 表示在获得观测值 $X_{l,k}$ 之前的信息，D 表示扩散程度算子。利用 Bayes 定理，可得

$$d_{l,k} = \left[1 + \frac{(1-d_{l,k|l-1})d(X_{l,k}\mid D,\xi_{l-1})}{d_{l,k|l-1}d(X_{l,k}\mid R,\xi_{l-1})}\right]^{-1}. \quad (16)$$

$d_{l,k|l-1} = D(R_{l,k}\mid\xi_{l-1})$ 即为先验的声场扩散程度的度量因子，将在下一节详细介绍。

$$\Lambda = \frac{d(X_{l,k}\mid D,\xi_{l-1})}{d(X_{l,k}\mid R,\xi_{l-1})}. \quad (17)$$

式(17)得到的 Λ 即为似然比。对直达声和混响声的直达声场和扩散声场的建模可由式(18)得到：

$$\Lambda = \frac{\widehat{\mathrm{DRR}} - \mathrm{DRR}_{\min}}{\mathrm{DRR}_{\max} - \mathrm{DRR}_{\min}}, \quad \widehat{\mathrm{DRR}} = \mathrm{mean}\{\mathrm{DRR}(l,k)\}. \quad (18)$$

mean 表示取算术平均。

2.2 先验声场扩散程度估计

下面利用软决策信息，给出扩散程度估计方法。对式(12)采用一阶递归平滑，

$$\zeta(l,k) = \beta\zeta(l,k) + (1-\beta)\mathrm{DRR}(l,k). \quad (19)$$

利用频域窗 $b_\rho(i)$ 平滑，窗长为 $2w_\rho+1$，其中 ρ 分别代表 local 和 global。

$$\widehat{\zeta(l,k)^\rho} = \sum_{i=-w_\rho}^{w_\rho} b_\rho(i)\zeta(l,k-i). \quad (20)$$

$$d^\rho_{l,k} = \begin{cases} 0, & \widehat{\zeta(l,k)^\rho} > \zeta_{\max} \\ 1, & \widehat{\zeta(l,k)^\rho} < \zeta_{\min} \\ \dfrac{\lg(\widehat{\zeta(l,k)^\rho}/\zeta_{\min})}{\lg(\zeta_{\max}/\zeta_{\min})}, & \text{其他} \end{cases} \quad (21)$$

$$d_{l,k|l-1} = d^{\mathrm{local}}_{l,k}d^{\mathrm{global}}_{l,k}. \quad (22)$$

$d_{l,k|l-1}$ 获得之后，由式(16)得到 $d_{l,k}$ 的估计。

2.3 计算谱减法增益

在获得混响声功率谱之后，利用谱减法在频域构造滤波器，图 2 为该混响抑制方法框图。

$$S(l,k) = X(l,k)G(l,k), \quad (23)$$

$$G(l,k) = \max\left(\mathrm{G\_min}, 1-\sqrt{\frac{\lambda^2_{l,k}}{X^2_{l,k}}}\right). \quad (24)$$

其中，G_min 为每个频带的最大抑制量。

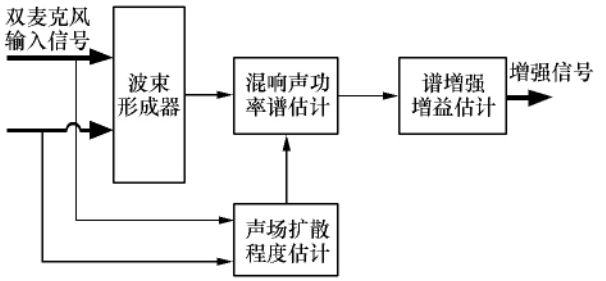

图 2 基于声场扩散信息混响抑制方法框图

3 实验结果及讨论

为了测试算法性能，本文采用纯净语音卷积不变房间冲击响应数据(RIR)的方法构造测试数据库。其中：语音是从 TIMIT 数据库中随机抽取的 10 女 10 男数据，RIR 数据库来自德国亚琛(Aachen)工业大学[9]。为了测量不同类型房间和不同混响时间下算法的性能，分别从该数据库中选取 RIR，如表 1 所示。

表 1 选取的房间冲击响应(RIR)描述

房间类型	声源距麦克距离/m	T60/s
会议室	1.45，2.8	0.23
办公室	2.0	0.43
报告厅	2.25，4.0	0.78

同时为了对比，考察了同为基于谱增强方法的 Allen 等[6]的基于幅度相干函数的去混响算法、Habets 等[3]的基于广义统计模型的去混响算法，分别在分段信混比(segmental signal-to-reverberation ratio，SSRR)、对数谱距离(log spectral distortion，LSD)以及主观语音质量评估(perceptual evaluation of speech quality-mean opinion score，PESQ-MOS)方面给出了对比实验结果。采样率为 8 kHz，帧长为 256 点，帧移为 128 点，麦克风间距为 17 cm，T_d 为 5 ms。其他实验参数如下：$\alpha=0.85$，$\alpha_d=0.5$，$\alpha_r=0.9$，$\beta=0.5$，$w_{\mathrm{local}}=3$，$w_{\mathrm{global}}=15$，$\mathrm{DRR}_{\min}=-10$ dB，$\mathrm{DRR}_{\max}=10$ dB，$\zeta_{\min}=-5$ dB，$\zeta_{\max}=10$ dB，G_min $=-15$ dB。

3.1 SSRR 性能

分段信混比 SSRR 可以表示为式(25)，$\hat{s}$ 和 s_d 分别表示待测信号和直达声[10]，

$$\mathrm{SSRR} = \frac{1}{N_{\mathrm{seg}}}\sum_{l=0}^{N_{\mathrm{seg}}-1} 10\lg\left(\frac{\|s_d(l)\|_2^2}{\|\hat{s}(l)-s_d(l)\|_2^2}\right)\mathrm{dB}. \quad (25)$$

分段信混比的分析结果见表 2。实验证明，本文算法能够有效提高信混比，较其他算法具有明显

的优势。

表 2 不同算法分段信混比的结果对比

方法	SSRR/dB				
	会议室		办公室	报告厅	
	1.45 m	2.8 m	2 m	2.25 m	4 m
未处理	−0.07	−3.77	−9.28	−6.77	−10.01
Allen 等	0.97	−2.87	−8.11	−5.32	−8.51
Habets 等	0.80	−2.70	−7.24	−4.96	−6.46
本文	1.89	−0.01	−3.29	−2.27	−4.65

3.2 LSD 性能

对数谱距离(LSD)[11]反映的是语音失真情况,计算的是直达声和待测信号对数谱差的均方根,

$$\mathrm{LSD}(l)=\left(\frac{1}{K}\sum_{k=0}^{\frac{k}{2}-1}|L\{S(l,k)\}-L\{S_{\mathrm{d}}(l,k)\}|^2\right)^{\frac{1}{2}}\ \mathrm{dB}. \quad (26)$$

其中,$L\{S(l,k)\}=\max\{20\lg|S(l,k)|,\delta\}$表示约束在 50 dB 动态范围内的对数谱。研究结果见表 3。DRR 较低时,由于 DRR 估计的准确性不够,本文算法考虑到混响抑制和失真的折中,失真略大于 Allen 等的算法。而相对于 Habets 等的算法,在混响较大时,本文算法更加满足混响声为扩散声场的假设,加之不依赖房间参数 T60 的估计,本文算法表现出更好的性能。

表 3 不同算法对数谱距离(LSD)的结果对比

方法	LSD/dB				
	会议室		办公室	报告厅	
	1.45 m	2.8 m	2 m	2.25 m	4.0 m
未处理	1.00	1.06	1.17	1.34	1.45
Allen 等	0.90	0.98	1.11	1.21	1.31
Habets 等	0.94	1.01	1.10	1.27	1.34
本文	0.87	0.95	1.09	1.21	1.36

3.3 PESQ-MOS 分数

PESQ-MOS[12]从总体上反映了语音质量,并且同主观听感具有较强的相关性。不同环境下 PESQ-MOS 的分析结果如表 4 所示。本文算法在不同房间、不同混响时间下都表现出了良好的性能。

表 4 不同算法的主观语音质量评估分数(PESQ-MOS)的结果对比

方法	PESQ-MOS				
	会议室		办公室	报告厅	
	1.45 m	2.8 m	2 m	2.25 m	4.0 m
未处理	2.56	2.36	1.99	2.04	1.81
Allen 等	2.67	2.45	2.05	2.15	1.92
Habets 等	2.63	2.40	2.05	2.09	1.89
本文	2.66	2.47	2.09	2.19	1.95

4 结 论

本文根据直达声和混响声在声场上所反映出的扩散信息,对二者分别建模,构建 DRR 和 MSC 之间的关系。利用 DRR 作为指导,反映声场扩散程度,并以此为依据得到混响声功率谱的估计值。通过时频域递归平均的方式按照先验的扩散信息更新混响声功率谱。最终,通过谱增强的方法进一步提升了算法的性能。实验结果表明,本文算法相比对 Allen 等和 Habets 等的算法在 SSRR、LSD 以及 PESQ-MOS 等评价指标上都表现出明显的优势。

参考文献 (References)

[1] Naylor P A, Gaubitch N D. Speech Dereverberation [M]. London, UK: Springer, 2010.

[2] 李凯. 复杂环境下基于传声器阵列的语音增强方法 [D]. 北京: 中国科学院声学研究所, 2012.
LI Kai. Microphone Array for Speech Enhancement in Complex Environments [D]. Beijing: Institute of Acoustics, Chinese Academy of Sciences, 2012. (in Chinese)

[3] Habets E A P, Gannot S, Cohen I. Late reverberant spectral variance estimation based on a statistical model [J]. *Signal Processing Letters, IEEE*, 2009, **16**(9): 770-773.

[4] Lebart K, Boucher J M, Denbigh P N. A new method based on spectral subtraction for speech dereverberation [J]. *Acta Acustica United with Acustica*, 2001, **87**(3): 359-366.

[5] Ratnam R, Jones D L, Wheeler B C, et al. Blind estimation of reverberation time [J]. *The Journal of the Acoustical Society of America*, 2003, **114**: 2877-2892.

[6] Allen J B, Berkley D A, Blauert J. Multimicrophone signal-processing technique to remove room reverberation from speech signals [J]. *The Journal of the Acoustical Society of America*, 1977, **62**: 912-915.

[7] Jeub M, Schafer M, Esch T, et al. Model-based dereverberation preserving binaural cues [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, **18**(7): 1732-1745.

[8] McCowan I A, Bourlard H. Microphone array post-filter based on noise field coherence [J]. *IEEE Transactions on Speech and Audio Processing*, 2003, **11**(6): 709-716.

[9] Jeub M, Schafer M, Vary P. A binaural room impulse response database for the evaluation of dereverberation algorithms [C]// 16th International Conference on Digital Signal Processing. Santorini, Greece, 2009: 1-5.

[10] Naylor P A, Gaubitch N D. Speech dereverberation [C]// Proc Int Workshop Acoust Echo Noise Control. Eindhoven, Netherlands, 2005.

[11] Erell A, Weintraub M. Estimation using log-spectral-distance criterion for noise-robust speech recognition [C]// International Conference on Acoustics, Speech, and Signal Processing. Albuquerque, NM, USA, 1990: 853-856.

[12] 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]// International Conference on Acoustics, Speech, and Signal Processing. Salt Lake, UT, USA, 2001: 749-752.