第 38 卷 第 1 期
2019 年 1 月

应用声学
Journal of Applied Acoustics

Vol. 38, No. 1
January, 2019

◇ 研究报告 ◇

偏度最大化多通道逆滤波语声去混响研究*

郭 颖[1,2] 彭任华[1] 郑成诗[1†] 李晓东[1]

(1 中国科学院噪声与振动重点实验室 (声学研究所) 北京 100190)

(2 中国科学院大学 北京 100049)

摘要 房间混响会降低语声质量和语声可懂度。高阶统计量是衡量非高斯性的重要参量，基于语声非高斯特性可实现语声去混响。该文提出一种基于高阶统计量的多通道语声去混响方法，该方法首次用多通道语声信号线性预测残差的三阶统计量偏度构造代价函数，以去混响重建信号线性预测残差的偏度最大化为目标自适应地更新逆滤波器，同时引入通道逆滤波和语声产生系统的联合估计。实验结果表明，该方法相较于已有的基于线性预测残差四阶统计量峰度的方法具有更好的去混响效果，且对噪声具有更强的鲁棒性。

关键词 高阶统计量，偏度，线性预测，房间脉冲响应，逆滤波

中图法分类号: TN912.35 **文献标识码:** A **文章编号:** 1000-310X(2019)01-0058-10

DOI: 10.11684/j.issn.1000-310X.2019.01.009

Maximum skewness-based multichannel inverse filtering for speech dereverberation

GUO Ying[1,2] PENG Renhua[1] ZHENG Chengshi[1] LI Xiaodong[1]

(1 *Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing* 100190, *China*)

(2 *University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract Room reverberation often leads to the reduction of speech quality and speech intelligibility. Speech dereverberation can be achieved by using non-Gaussian property of speech, where higher order statistics (HOS) are typical measurements. This paper presents a method based on HOS for multichannel speech dereverberation. The cost function is constructed using the third-order statistics, namely skewness, of multichannel speech signal linear prediction residuals, and then update the inverse filter adaptively by maximizing the skewness of the linear prediction residuals of the reconstructed speech signal. Meanwhile, we introduce the joint estimation of the channel's inverse filter and the speech production system. Experimental results show that the proposed method is superior to the method based on forth-order statistics, i.e. kurtosis, in terms of dereverberation and robustness to the noise.

Key words Higher order statistics, Skewness, Linear prediction, Room impulse response, Inverse filtering

2018-04-03 收稿; 2018-09-03 定稿

*国家自然科学基金项目 (61571435)

作者简介: 郭颖 (1992-), 女, 辽宁朝阳人, 硕士研究生, 研究方向: 信号与信息处理。

†通讯作者 E-mail: cszheng@mail.ioa.ac.cn

0 引言

在一个封闭空间中，传声器拾取的语声信号既包括直达声，也包括通过墙壁和天花板等反射的混响声。房间混响会引起谱染色，影响语声质量，降低语声可懂度，进而严重降低语声识别、语声分离等应用的性能。随着说话人与传声器距离的增加以及房间混响时间的增加，混响所带来的影响也会越严重。

去混响方法通常可以分为以下几类：(1) 波束形成[1]，该方法是一种空间滤波技术，广泛应用于雷达、声呐、远程通讯、声学、图像处理等多种领域[2]。在声学信号处理中，用于噪声环境下的声源提取以及混响抑制。该方法往往需要信号的波达方向 (Directions of arrival, DOAs) 作为先验信息，而且为了达到比较理想的去混响效果，需要相对较多的传声器个数以及较大的传声器阵列孔径，从而使直达方向的增益足够大。(2) 谱增强[3−4]，Lebart 等[5] 提出利用谱减法实现无噪声情况下的语声去混响。通常用于晚期混响抑制，该类方法需要根据房间的混响时间来估计混响的能量。Fang 等[6] 用基于相干函数的方法实现去混响。(3) 线性预测 (Linear prediction, LP) 残差增强，Peng 等[7] 的工作说明了晚期混响在 LP 残差域相对较白。文献 [8] 采用约束最小均方误差 LP 残差估计方法去除晚期混响和噪声，相比于传统的 LP 残差域处理方法和谱减法性能有很大的提升。文献 [9] 利用多级线性预测实现晚期混响抑制。(4) 逆滤波，该类方法直接估计引起房间混响的房间脉冲响应 (Room impulse response, RIR)，通过对观测信号进行解卷积得到原始信号。在实际应用场景中，房间脉冲响应通常是未知的，而且会随着声源移动或房间状态 (如温度和湿度等) 的改变而变化。因此，本文研究盲反卷积的方法。基于随机变量非高斯性极大的准则，混响信号可以假设为独立同分布 (i.i.d) 的语声信号进行延迟、加权的结果，依据中心极限定理[10]，混响信号可以近似为高斯分布。高阶统计量是衡量非高斯性的重要参量，语声信号是典型的非高斯信号，因此采用高阶统计量可实现语声分离和去混响。

文献 [11] 提出一种最大化线性预测残差四阶统计量峰度 (Kurtosis) 的去混响方法，证明了该方法比传统的波束形成方法具有更有效的去混响效果。文献 [12] 在此基础上提出单通道频域实现，通过实验说明该方法在 0.2∼0.4 s 的混响时间范围内有效，而在混响较强的环境下该方法失效。应用峰度准则的方法去混响性能有限，文献 [13] 采用三阶统计量偏度 (Skewness) 对具有不对称概率密度分布的信号进行盲反卷积，文献 [14] 提出最大化线性预测残差偏度的单通道逆滤波方法，通过实验说明了足够长的纯净语声信号概率密度分布呈现出明显的不对称特性，该方法相比于峰度准则在较强混响下性能更优，而且鲁棒性更强。直接对混响语声信号进行线性预测得到的线性预测系数存在一定的偏差，影响房间脉冲响应逆滤波的准确度，文献 [15] 从语声信号的产生模型出发，将混响语声信号的盲逆滤波分解为预测误差滤波器 (Prediction error filter, PEF) 的估计和房间脉冲响应逆滤波器的估计两部分。

本文提出一种基于高阶统计量的多通道语声去混响方法，该方法首次用多通道语声信号线性预测残差的偏度构造代价函数，以语声去混响重建信号线性预测残差的偏度最大化为目标，自适应地更新通道逆滤波器。同时为了得到更准确的通道逆滤波器估计，提出联合估计通道逆滤波器和语声产生系统逆滤波器的新方法。该方法相比于已有的线性预测残差域峰度最大化的多通道去混响方法，计算量更低，而且具有更好的去混响效果，特别是在混响时间较长的环境下性能更为突出，同时对噪声的鲁棒性更强。

1 偏度最大化多通道房间脉冲响应逆滤波

1.1 算法理论模型

混响语声模型可以表示为

$$x_m(n) = \sum_{l=0}^{L} h_m(l)s(n-l), \tag{1}$$

其中，传声器个数为 $M(M \geqslant 2)$，$x_m(n)$ 为第 m 个传声器拾取的混响语声信号，$s(n)$ 为目标语声信号，$\{h_m(l)\}_{l=0}^{L}$ 表示声源到第 m 个传声器的 $L+1$ 阶时不变的房间脉冲响应。

语声信号从产生、经过房间反射到被传声器拾取所经过的声学系统可认为是语声产生系统和房间声学系统的串联系统。其中语声信号的产生过程可建模成一个时变的自回归 (Autoregressive, AR) 过程[16]，考虑语声信号的短时平稳特性，第 i 帧的

声源信号可以表示为

$$s(n)=\sum_{p=1}^{P}b_i\left(p\right)s\left(n-p\right)+e_i(n), \tag{2}$$

其中，$\{b_i(p)\}_{p=1}^{P}$为P阶预测系数，语声产生系统的传递函数$B(z)$是$\{b_i(p)\}_{p=1}^{P}$的Z变换，可以用一个阶数为P的时变FIR滤波器来表示，其逆滤波器称为预测误差滤波器。房间声学系统的传递函数$H(z)$可以用一个阶数为L的时不变FIR滤波来表示。因此，观测信号$x(n)$是在$e(n)$激励下，经过语声产生系统$B(z)$和房间声学系统$H(z)$共同作用的输出结果。

盲去混响的目标是在无任何房间先验知识的前提下，仅通过传声器观测信号$x(n)$去除由房间声学系统$H(z)$所引起的混响，恢复声源信号$s(n)$。因此，一个重要的问题就是在盲逆滤波过程中，如何将房间脉冲响应的逆滤波从整个系统的逆滤波中分离出来，即去掉声道滤波对房间脉冲响应逆滤波所造成的偏差。一种常用的方法是首先对混响语声信号直接进行线性预测预白化处理，阶数一般取为10，然后在线性预测残差域进行逆滤波。考虑线性预测系数受语声信号中混响的影响，直接对混响信号进行线性预测求得的预测系数存在偏差，更为准确的方法可以采用预测误差滤波器与房间脉冲响应逆滤波器联合估计。图1 展示了联合估计算法的实现框图，考虑时域实现收敛较慢，甚至可能不收敛，因此本文采用频域方法实现。首先用时不变的房间脉冲响应逆滤波器在频域对观测信号进行滤波后，再通过时变的预测误差滤波器，得到线性预测残差信号，以残差信号的偏度最大化为目标，计算滤波器的更新梯度，进而更新房间脉冲响应逆滤波器，利用更新的逆滤波器对混响信号进行滤波，重构出逆滤波后的语声信号。算法记为基于偏度的预测误差滤波器与房间脉冲响应逆滤波器的联合估计方法，即MSJE-IF-MSD(Maximum-skewness joint estimation based-inverse filtering for multichannel speech dereverberation)，简化为MSJE。

$\boldsymbol{g}_m$表示通道m的L阶自适应房间脉冲响应逆滤波器系数，$\boldsymbol{g}_m=\left[g_m\left(0\right),\cdots,g_m\left(L-1\right)\right]^{\mathrm{T}}$；$G_m=\sum_{l=0}^{L-1}g_m(l)z^{-l}$为第$m$通道的房间脉冲响应逆滤波器系统传递函数。这里需要假设每个通道的房间传递函数$G_1\left(z\right),\cdots,G_M\left(z\right)$之间没有共同的零点。进而可以得到逆滤波后重构的语声信号：

$$y(n)=\sum_{m=1}^{M}\left(\boldsymbol{g}_m\right)^{\mathrm{T}}\boldsymbol{x}_m(n), \tag{3}$$

其中，$\boldsymbol{x}_m(n)=\left[x_m(n),\cdots,x_m\left(n-L+1\right)\right]^{\mathrm{T}}$。根据语声信号的短时平稳性，将逆滤波输出$y(n)$分帧后通过时变的预测误差滤波器$\{a_i(p)\}_{p=1}^{P}$，得到第$i$帧线性预测残差信号$d_i(n)$：

$$d_i(n)=y_i(n)-\sum_{p=1}^{P}a_i(p)y_i(n-p). \tag{4}$$

用向量形式表示：

$$d_i(n)=\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{y}_i(n), \tag{5}$$

其中，$\boldsymbol{a}_i=\left[1,-a_i\left(1\right),-a_i\left(2\right),\cdots,-a_i\left(P\right)\right]^{\mathrm{T}}$，$\boldsymbol{y}_i(n)=\left[y_i(n),y_i\left(n-1\right),\cdots,y_i\left(n-P\right)\right]^{\mathrm{T}}$，$P$为预测误差滤波器阶数。$A_i(z)=1-\sum_{p=1}^{P}a_i(p)z^{-p}$为预测误差滤波器系统传递函数。

理想情况下最终得到的$d_i(n)$与激励信号$e_i(n)$等价，只存在微小的延迟和幅度变化。因此，问题可退化为房间脉冲响应逆滤波器$\boldsymbol{g}$和预测误差滤波器$\boldsymbol{a}$的估计，$\boldsymbol{g}=\left[\boldsymbol{g}_1^{\mathrm{T}},\cdots\boldsymbol{g}_M^{\mathrm{T}}\right]^{\mathrm{T}}$，$\boldsymbol{a}=\left[\boldsymbol{a}_1^{\mathrm{T}},\cdots\boldsymbol{a}_S^{\mathrm{T}}\right]^{\mathrm{T}}$，$S$为线性预测总帧数。

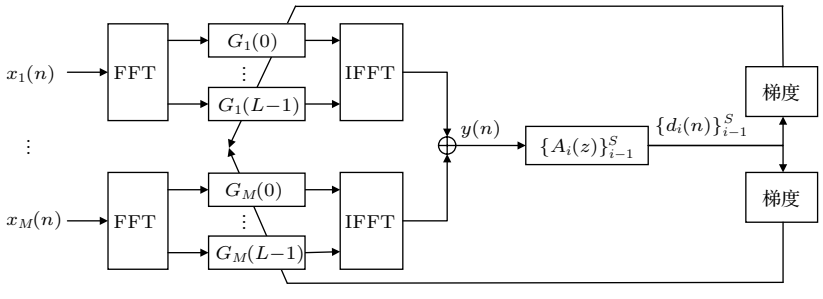


图1 MSJE算法框图

Fig. 1 Schematic diagram of MSJE

1.2 目标函数

根据上面的讨论，需要建立合适的目标函数来估计 $\boldsymbol{g}$ 和 $\boldsymbol{a}$。考虑逆滤波后残差信号 $\{d(n)\}_{n=1}^{W}$ 样本间的相关性最小，采用交互信息作为目标函数[15]：

$$\begin{aligned} J(n) &= \sum_{n=1}^{W} \mathrm{H}\left[d(n)\right] - \mathrm{H}\left(\boldsymbol{d}'\right) \\ &= -\sum_{n=1}^{W} \Gamma\left[d(n)\right] + \sum_{n=1}^{W} \lg \upsilon\left[d(n)\right] \\ &\quad - \lg\left|\det \sum\left(\boldsymbol{d}'\right)\right|, \end{aligned} \tag{6}$$

其中，W 为样本点数，$\mathrm{H}(\boldsymbol{\xi})$ 表示随机变量 $\boldsymbol{\xi}$ 的微分熵，$\boldsymbol{d}' = [d(W), \cdots, d(1)]^{\mathrm{T}}$，$\upsilon[d(n)]$ 表示 $d(n)$ 的方差，$\sum(\boldsymbol{d}') = E\left[\boldsymbol{d}'\boldsymbol{d}'^{\mathrm{T}}\right]$。$\Gamma[d(n)]$ 表示 $d(n)$ 的负熵，用来衡量信号的非高斯性，可以用高阶统计量表示，三阶统计量——偏度用来衡量概率密度分布的偏斜程度，定义为

$$\gamma = \frac{\mu_3}{\sigma^3}, \tag{7}$$

其中，μ_3 为三阶中心距，σ 为标准差。相对于四阶统计量峰度，偏度的优势主要体现在衡量一些概率密度分布具有不对称性的声源信号非高斯性上。

本文考虑偏度作为衡量语声信号非高斯性的准则，根据公式 (7)，目标函数可进一步表示为

$$\begin{aligned} J(n) = &-\sum_{n=1}^{W} \frac{E\left[d^3(n)\right]}{E^{\frac{3}{2}}\left[d^2(n)\right]} + \sum_{n=1}^{W} \lg \upsilon\left[d(n)\right] \\ &- \lg\left|\det \sum\left(\boldsymbol{d}'\right)\right|. \end{aligned} \tag{8}$$

因此可以建模为下面的优化问题：

$$\begin{cases} \{\boldsymbol{g}^*, \boldsymbol{a}^*\} = \arg\min\left(-\sum_{n=1}^{W} \frac{E\left[d^3(n)\right]}{E^{\frac{3}{2}}\left[d^2(n)\right]} + \sum_{n=1}^{W} \lg \upsilon\left[d(n)\right] - \lg\left|\det \sum\left(\boldsymbol{d}'\right)\right|\right), \\ \text{s.t. } \|\boldsymbol{g}\| = 1 \text{ 且 } \boldsymbol{a} \text{ 为最小相位.} \end{cases} \tag{9}$$

约束条件 $\|\boldsymbol{g}\| = 1$ 保证了房间脉冲响应逆滤波器的归一化。同时为了使系统稳定，应保证预测误差滤波器 $\boldsymbol{a}$ 的最小相位特性。

1.3 预测误差滤波器的估计

由于高阶统计量会使预测误差滤波器非最小相位，因此该部分的估计只考虑二阶项作为目标函数，表示为

$$J_1(n) = \sum_{n=1}^{W} \lg \upsilon\left[d(n)\right] - \lg\left|\det \sum\left(\boldsymbol{d}'\right)\right|, \tag{10}$$

其中，$\lg\left|\det \sum(\boldsymbol{d}')\right|$ 为常数项[15]，可忽略。由于语声信号的短时平稳性，预测误差滤波器系数在每一帧单独求取，对于第 i 帧残差信号 $d_i(n)$，目标函数为 $\sum_{n=N(i-1)+1}^{N\times i} \lg \upsilon\left[d(n)\right]$。假设 $d(n)$ 在一帧内是平稳的，则 $\sum_{n=N(i-1)+1}^{N\times i} \lg \upsilon\left[d(n)\right] = N \lg \upsilon\left[d(n)\right]$，且由于取对数操作为线性的，随着变量的增加而增加，因此有

$$N \lg \upsilon\left[d(n)\right] = N\upsilon\left[d(n)\right]. \tag{11}$$

公式 (11) 可以通过 $\sum_{n=N(i-1)+1}^{N\times i} d^2(n)$ 来估计，公式 (10) 最小化的问题变为使 $d(n)$ 的均方误差最小，可通过对 $\{y(n)\}_{N(i-1)+1\leqslant n\leqslant N\times i}$ 进行线性预测分析实现。

具体实现：首先对逆滤波后的输出信号 $y(n)$ 进行分帧得到 $y_i(n)$，逐帧通过线性预测估计 $y_i(n)$ 的预测误差滤波器系数 $\boldsymbol{a}_i$。而线性预测可以保证估计得到的预测误差滤波器的最小相位性。

1.4 房间脉冲响应逆滤波器的估计

通常语声信号的激励信号为超高斯分布，它的二阶矩相对于高阶矩可以忽略。因此，该部分只考虑公式 (8) 中的三阶项部分。目标函数可化简为

$$J_2(n) = \frac{E\left[d^3(n)\right]}{E^{\frac{3}{2}}\left[d^2(n)\right]}. \tag{12}$$

采用梯度下降法对每个通道的滤波器 $\boldsymbol{g}_m$ 进行单独更新，更新方程为

$$\boldsymbol{g}_m^{r+1} = \boldsymbol{g}_m^{r} + \mu \frac{\partial J_2(n)}{\partial \boldsymbol{g}_m^{r}}. \tag{13}$$

目标函数对 $\boldsymbol{g}_m^r$ 的偏导：

$$\frac{\partial J_2(n)}{\partial \boldsymbol{g}_m^r}=3\frac{E\left\{d^2(n)\frac{\partial\left[d(n)\right]}{\partial \boldsymbol{g}_m^r}\right\}E\left[d^2(n)\right]-E\left[d^3(n)\right]E\left\{d(n)\frac{\partial\left[d(n)\right]}{\partial \boldsymbol{g}_m^r}\right\}}{E^{\frac{5}{2}}\left[d^2(n)\right]}. \tag{14}$$

结合公式(3),对于第i帧残差信号:

$$\frac{\partial\left[d(n)\right]}{\partial \boldsymbol{g}_m^r}=\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{X}_m(n), \tag{15}$$

其中,$\boldsymbol{X}_m(n)=\left[x_m^{\mathrm{T}}(n),\cdots,x_m^{\mathrm{T}}(n-P)\right]^{\mathrm{T}}$。对第$i$帧残差信号的梯度进一步推导:

$$\frac{\partial J_2(n)}{\partial \boldsymbol{g}_m^r}=3\frac{E\left[d^2(n)\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{X}_m(n)\right]E\left[d^2(n)\right]-E\left[d^3(n)\right]E\left[d(n)\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{X}_m(n)\right]}{E^{\frac{3}{2}}\left[d^2(n)\right]}. \tag{16}$$

为了进一步简化,忽略式(16)的时间依赖性,令$\boldsymbol{r}_m(n)=\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{X}_m(n)$,梯度近似为

$$\begin{aligned}&\frac{\partial J_2(n)}{\partial \boldsymbol{g}_m^r}\\&\approx 3\left(\frac{d^2(n)E\left[d^2(n)\right]-d(n)E\left[d^3(n)\right]}{E^{\frac{5}{2}}\left[d^2(n)\right]}\right)\cdot\boldsymbol{r}_m(n)\\&=q(n)\cdot\boldsymbol{r}_m(n).\end{aligned} \tag{17}$$

逆滤波器在频域进行更新。将更新后的线性预测残差信号$\boldsymbol{r}_m(n)$分成长度为L的块,并将每一块补0至长度为$2L$,对每一块计算长度为$2L$的傅里叶变换(Fast Fourier transform, FFT)。将$q(n)$分成长度为$2L$的块,重叠50%,对每一块计算长度为$2L$的FFT。设分块个数为T,得到频域自适应更新方程:

$$\boldsymbol{G}_m'^{r+1}=\boldsymbol{G}_m^r+\frac{\mu}{T}\sum_{j=1}^{T}\boldsymbol{Q}_j\boldsymbol{R}_{mj}^{\mathrm{H}}, \tag{18}$$

$$\boldsymbol{G}^{r+1}=\frac{\boldsymbol{G}'^{r+1}}{\left|\boldsymbol{G}'^{r+1}\right|}, \tag{19}$$

其中,$\boldsymbol{G}_m^r$、$\boldsymbol{Q}_j$、$\boldsymbol{R}_{mj}$分别为第r次迭代的$\boldsymbol{g}_m$、$\boldsymbol{q}_j$、$\boldsymbol{r}_{mj}$的FFT。公式(19)对更新后的逆滤波器进行归一化,保证滤波器的收敛。$\boldsymbol{G}_m^0=[1,\cdots,1]^{\mathrm{T}}$。这里逆滤波器通过对20 s混响语声信号进行估计得到。

1.5 联合估计策略

上述目标函数的简化以及迭代估计两个逆滤波器需要基于如下假设:当$\boldsymbol{g}$固定时,最小化二阶项的同时也会使整体目标函数最小化;同理,当$\boldsymbol{a}$固定时,最大化三阶项也会使整体目标函数最小化。根据以上分析,迭代更新预测误差滤波器和房间脉冲响应逆滤波器。对观测信号,首先通过房间脉冲响应逆滤波器进行逆滤波后,再通过预测误差滤波器,得到更新后的残差信号;以残差信号的偏度最大化为目标,通过梯度下降法更新房间脉冲响应逆滤波器,迭代更新直至滤波器收敛,重构出逆滤波后的语声信号。

作为联合估计的替代,另外一种比较简单的实现可以直接对观测信号进行线性预测预白化处理,在线性预测残差域上求解房间脉冲响应逆滤波器。该方法可以认为近似于MSJE预测误差滤波器只迭代一次的情况。为了对比,将最大化线性预测残差偏度的多通道逆滤波语声去混响方法记为MLPRS-IF-MSD (Maximum linear prediction residual skewness-based inverse filtering for multichannel speech dereverberation),简化为MLPRS。

2 仿真和实验研究

2.1 仿真

采用镜像法[17]得到的4通道的RIR,声源信号由TIMIT数据库中选取的100段男声和100段女声语声段构成,将其与不同混响时间的RIR卷积得到混响语声信号。在模型中,4个传声器分布在尺寸为5.5 m × 4.5 m × 3.5 m的矩形房间内。声源(红色圆点)与传声器阵列(灰色圆点)在房间内的分布示意图如图2所示,传声器间隔0.2 cm按线型摆放,与声源距离$d_0=3.3$ m。

混响时间和声学比是影响混响声场中的语言清晰度的两个独立参量,混响声场中的清晰度与混响时间(RT_{60})和声学比乘积的对数成反比变化[18]。混响时间增加和声源距传声器距离增大都会独立地增加混响强度[19],RT_{60}会导致语声频谱模糊,而d的增加会引起谱染色。在本实验中,我们

考虑固定声源到传声器的距离 d，改变 RT_{60} 的大小，评价不同混响强度下的算法去混响性能。以下实验中帧长 N 取 512 (32 ms)，步长 μ 设为 e^{-9}。

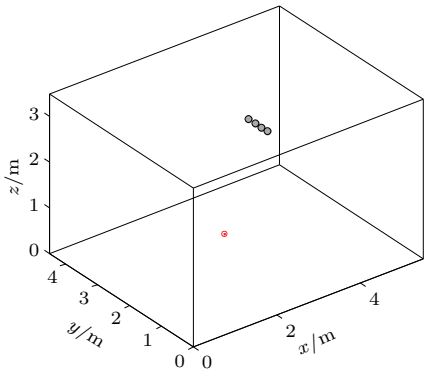


图 2　传声器位置示意图

Fig. 2　Diagram of the microphone position

2.1.1　滤波器阶数选择

滤波器的阶数 L 理论上应与混响时间 (RIR 的样本点数) 对应，即 $L = \mathrm{RT}_{60}(\mathrm{s}) \times f_s(\mathrm{Hz})$，其中采样率 $f_s = 16000$ Hz。混响时间越长，滤波器长度也相应的增加。而且，滤波器阶数增加会导致计算复杂度增加；滤波器阶数增加，逆滤波后信号的延迟也会增加 (RIR 与逆滤波器的卷积会使逆滤波后的冲激响应与原 RIR 之间存在近似 L 的延迟)。因此滤波器阶数的选择应该在理论值的基础上，结合实际效果选择尽量小的值且能保证逆滤波的性能。本文通过实验验证，给出一定混响时间范围的最小滤波器阶数。文献 [14] 给出了单通道线性预测残差偏度逆滤波算法的最小滤波器阶数。表 1 给出本文提出的 MLPRS 和 MSJE 算法的最小滤波器阶数。

表 1　不同混响时间下的滤波器阶数选择

Table 1　Selection of filter order for different reverberation times

算法类型	RT_{60}/ms		
	100∼500	600∼900	1000∼1500
单通道偏度	2000	4000	6000
多通道峰度	1000	2000	3000
MLPRS	1000	1500	2000
MSJE	800	1000	1500

对比本文方法和已有的单通道偏度准则方法，利用多通道数据可以有效减少滤波器阶数，而且算法对滤波器阶数的选择不敏感；同时，采用偏度准则相比于峰度准则最小滤波器阶数也有明显的下降；采用联合估计的 MSJE 方法可以进一步减少滤波器阶数。滤波器阶数越少，算法的计算复杂度也会降低。

2.1.2　混响抑制性能分析

为了评估本文算法的混响抑制性能，这里采用直达-反射路径能量比 (Direct-to-reverberation ratio, DRR) 和主观语声质量评估 [20] (Perceptual evaluation of speech quality, PESQ) 作为衡量指标，用于比较本文方法和 Gillespie 等 [11] 提出的峰度最大化多通道逆滤波语声去混响方法 (以下简称峰度算法)。DRR 可以用公式 (20) 进行计算：

$$\mathrm{DRR} = 10\lg\left(\frac{\sum\limits_{n=n_d-n_0}^{n_d+n_0} h^2(n)}{\sum\limits_{n=0}^{n_d-n_0} h^2(n) + \sum\limits_{n=n_d+n_0}^{\infty} h^2(n)}\right), \tag{20}$$

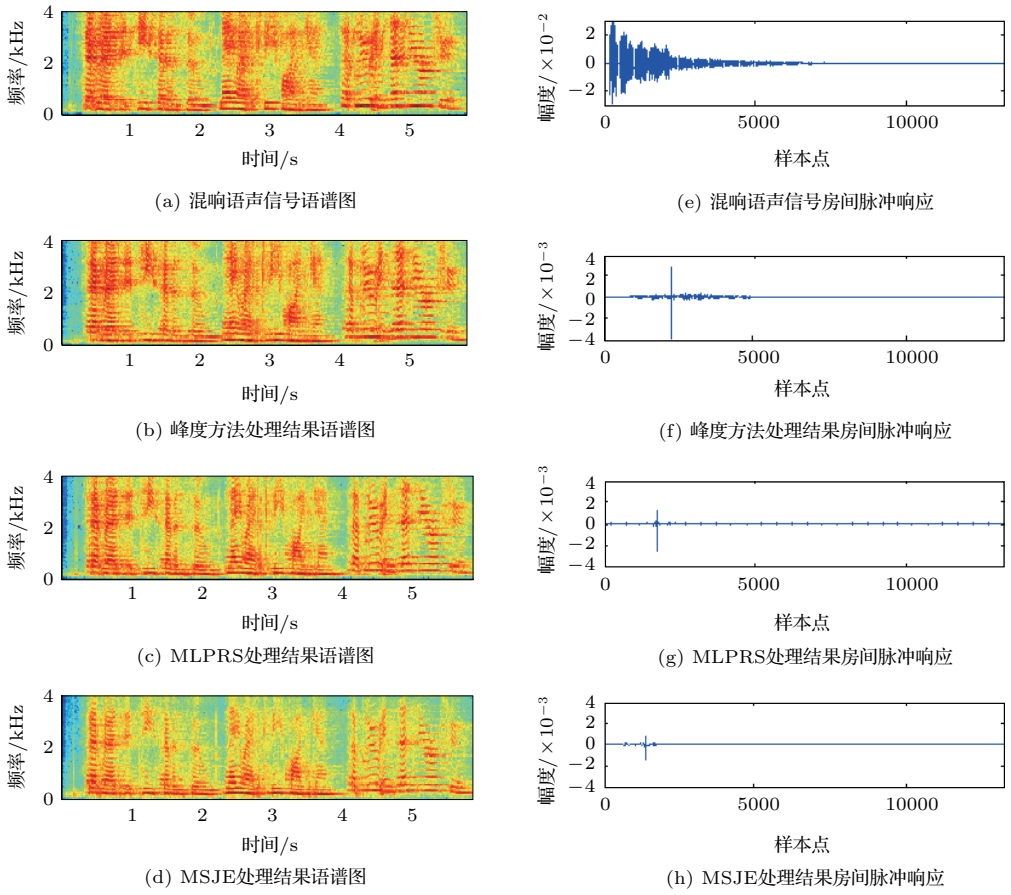
其中，直达信号在第 n_d 个采样点到达，直达路径的能量用冲激响应峰值周围 8 ms (即 $n_0 = 128$ 个采样点) 的信号能量计算。因此，DRR 通过直达路径能量与反射路径的总能量的比值来计算。图 3 为 $\mathrm{RT}_{60} = 1$ s 时，测试语声信号在 0∼4 kHz 部分的语谱图及逆滤波后的房间脉冲响应。

本文研究的方法均为在房间脉冲响应未知情况下的多通道盲逆滤波算法，这里给出房间脉冲响应仅为了分析和比较逆滤波的结果。从图 3 中的语谱图可以看出，对于 $\mathrm{RT}_{60} = 1$ s 混响时间比较长的情况，已有的多通道峰度准则方法表现一般。而用本文提出的多通道偏度准则方法 (图 3(c)，图 3(d)) 语谱图的模糊程度明显下降，模糊的频谱结构变得清晰，采用联合估计的多通道偏度算法表现出了更好的结果。从房间脉冲响应的逆滤波结果来看，三种方法逆滤波后的 RIR 均有比较明显的单一峰值。

图 4 给出了本文提出算法在不同混响时间下的平均 DRR 及 PESQ 得分。在混响时间较短时，基于峰度的方法与本文提出的基于偏度的方法结果相近；而当混响时间较长时，本文提出的基于偏度的方法要明显优于基于峰度的方法，且随着混响时间的增加，这种优势会越明显。且本文提出的 MSJE 在不同混响时间下的 DRR 整体优于 MLPRS。

比较本文提出的两个算法与峰度算法的 PESQ

得分，可以看出，在不同混响时间下本文提出的基于多通道偏度的去混响算法(MSJE，MLPRS)都较已有的基于多通道峰度的去混响算法性能有很大提升。基于峰度的方法对于混响时间较长的情况效果不理想。从整体上看，对于该组仿真的混响数据，除了 RT_{60} = 200 ms时MLPRS方法的PESQ得分更高一些，其他情况下MSJE较MLPRS算法的PESQ得分都略有提升。



(a) 混响语声信号语谱图　(e) 混响语声信号房间脉冲响应

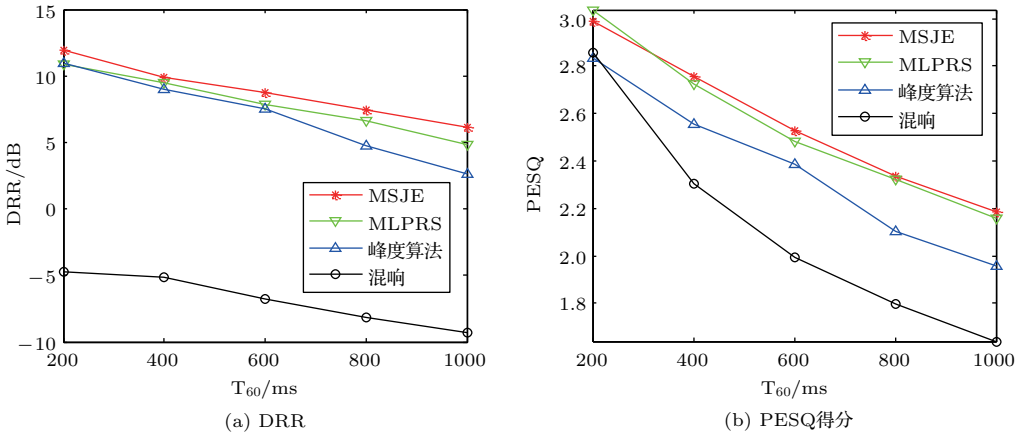
(b) 峰度方法处理结果语谱图　(f) 峰度方法处理结果房间脉冲响应

(c) MLPRS处理结果语谱图　(g) MLPRS处理结果房间脉冲响应

(d) MSJE处理结果语谱图　(h) MSJE处理结果房间脉冲响应

图3　RT_{60} = 1 s时逆滤波后的语谱图及房间脉冲响应

Fig. 3　Equalized speech spectrogram and impulse response with RT_{60} = 1 s



(a) DRR　(b) PESQ得分

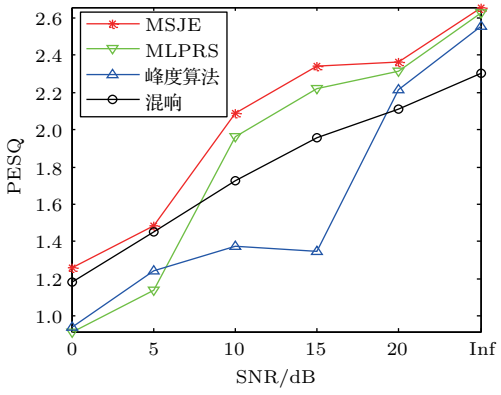
图4　提出算法在不同混响时间下的DRR以及PESQ得分

Fig. 4　DRR and PESQ score of the proposed algorithms for different reverberation times

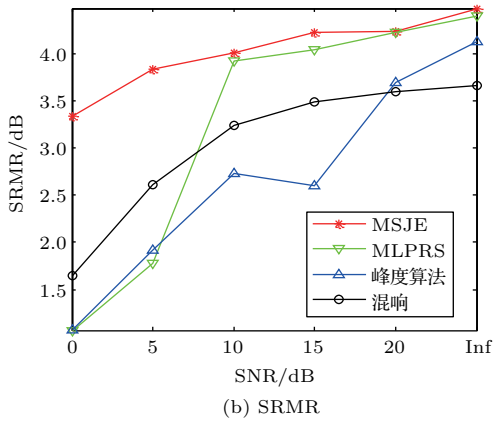
2.1.3　高斯噪声环境下算法鲁棒性

该实验测试本文提出算法在加性高斯白噪声环境下的去混响性能。用PESQ得分和语声-混响调制能量比(Speech-to-reverberation modulation energy ratio, SRMR)[21]作为评价指标。

图5给出了对$\mathrm{RT}_{60}=400$ ms时的传声器阵列信号加入不同信噪比的高斯白噪声，逆滤波后的信号平均PESQ得分和平均SRMR。峰度算法对信噪比低于20 dB输入信号失效，对房间脉冲响应的逆滤波无法得到单一峰值的结果；本文提出的MLPRS 算法对低于10 dB的输入信号失效，但去混响效果明显优于峰度方法；而采用联合估计的MSJE算法对测试的所有信噪比下的数据都能达到比较好的效果。本文提出的基于偏度的多通道逆滤波方法在高斯白噪声环境下的去混响性能比已有的基于峰度的多通道逆滤波方法有很大提升，提出方法对高斯噪声的鲁棒性更强。



(a) PESQ得分



(b) SRMR

图5　提出算法在$\mathrm{RT}_{60}=400$ ms不同噪声环境下的PESQ得分以及SRMR

Fig. 5　PESQ score and SRMR of the proposed algorithms for different noisy conditions with $\mathrm{RT}_{60}=400$ ms

在高斯白噪声环境下，影响本文算法去混响性能的因素有如下两个方面：一方面，当信号信噪比过低时，会引起线性预测模型谱密度产生畸变，使谱估计的质量受到损失，LP系数的估计变得不准确。另一方面，加性噪声的存在使信号的概率密度分布更趋于高斯分布，会改变自适应滤波过程中高阶统计量局部极大值点的位置，相比于没有噪声的情况使目标函数收敛到次极大值点，从而降低逆滤波的性能。提出的偏度方法相较于峰度方法对高斯噪声的鲁棒性更强，其原理可以通过以带有加性噪声的信号作为输入，计算两种算法的梯度来直观解释。峰度方法的梯度中受加性噪声影响的项更多，不稳定因素更多，因此峰度方法相比于偏度方法对加性噪声更加敏感。MSJE方法采用预测误差滤波器与房间脉冲响应逆滤波器的联合估计方法，使LP系数的估计更为准确，减弱了上述第一个因素的影响，因此相较于MLPRS方法对噪声的鲁棒性更强一些，在信噪比较低的情况下能更准确地估计逆滤波器。

2.1.4　计算复杂度

采用峰度和偏度准则的计算复杂度差别主要体现在梯度上，基于偏度准则的更新梯度表示为公式(17)，峰度准则的更新梯度可以进行类似的推导，最终表示为

$$\begin{aligned}&\frac{\partial Q'(n)}{\partial \boldsymbol{g}_m^r}\\ &\approx 4\left(\frac{d_i^3(n)E\left\{d_i^2(n)\right\}-d_i(n)E\left\{d_i^4(n)\right\}}{E^3\left\{d_i^2(n)\right\}}\right)\cdot \boldsymbol{r}_{mi}(n)\\ &=q'(n)\cdot \boldsymbol{r}_{mi}(n).\end{aligned}\tag{21}$$

在计算梯度过程中$q'(n)$相比于$q(n)$在计算时多一次乘法，因此偏度方法相比于峰度方法计算量更低。另一方面，表1给出了两种算法在不同混响时间情况下所需的最小滤波器阶数，偏度算法所需的滤波器阶数更少，也同时降低了算法的计算复杂度。

2.2　实际环境录音仿真测试

为了更合理地评估提出算法的去混响性能，本实验采用实际环境录音的多通道房间脉冲响应数据库[22−23]与TIMIT数据库的20 s纯净语声信号进行卷积作为测试信号，测试算法对不同声学比位置处(改变d)拾声信号的去混响性能。房间大小6 m × 6 m × 3 m，混响时间$\mathrm{RT}_{60}\approx 0.4$ s，混响半径r_c为1.02 m。声源位置与传声器距离d分别为1 m、

2 m、4 m，对应拾声位置处的声学比分别为1.06、0.26、0.07。选取角度$\theta=-80°,\cdots,80°$，对间隔40°测试的RIR进行处理，所用传声器个数为4，传声器间隔8 cm摆放。为了更全面地评估，实验对混响半径以内($d=1$ m)的信号也进行了测试。改变声源与传声器阵列的距离d，对相同的距离每隔40°测试一组数据，用PESQ得分作为去混响性能的评估指标。表2为实验结果。图6为改变声源与传声器阵列距离，对每组不同方向的实验结果取平均值得到的柱状图。

表2　实际环境录音测试结果

Table 2　Recording test results in real rooms

		角度				
		−80°	−40°	0°	40°	80°
$d=1$ m	混响	2.05	2.05	2.08	2.15	2.15
	峰度算法	2.28	**2.18**	2.22	2.37	2.36
	MLPRS	**2.37**	2.13	**2.24**	**2.39**	2.40
	MSJE	2.36	**2.18**	**2.24**	2.38	**2.42**
$d=2$ m	混响	1.68	1.72	1.74	1.77	1.79
	峰度算法	2.08	2.00	1.91	1.93	2.09
	MLPRS	2.12	2.03	2.01	**2.03**	2.16
	MSJE	**2.15**	**2.06**	**2.05**	2.01	**2.20**
$d=4$ m	混响	—	1.63	1.60	1.57	—
	峰度算法	—	1.90	1.87	1.65	—
	MLPRS	—	1.92	2.02	**2.02**	—
	MSJE	—	**2.06**	**2.07**	**2.02**	—

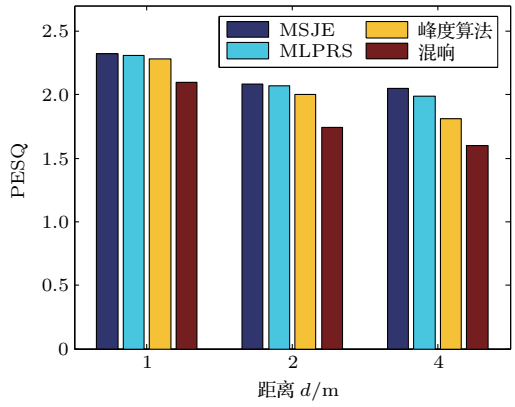


图6　声源距传声器不同距离时算法平均PESQ得分

Fig. 6　Average PESQ score of the proposed algorithms for different distances between source and microphone array

该实验验证了提出算法在不同混响强度下的去混响性能均优于峰度算法，且对于在声学比远小于1位置拾声的强混响信号，本文算法的优势更为明显。

为了进一步验证算法对汉语的有效性，采用20 s “GSBM 6001-89”国家标准样件中的有代表性的两段分别由男女声朗诵的《美谈不美》纯净语声信号与上述RIR数据的卷积作为测试数据，随着声源与传声器距离的改变对标准样件添加了不同强度的混响，得到的结果如图7所示。该实验验证了算法对于处理汉语以及男女声信号的有效性。

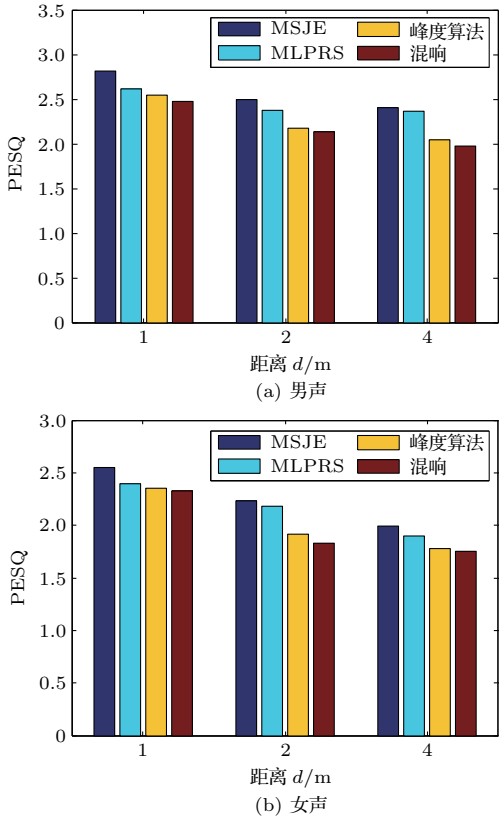


图7　声源距传声器不同距离时算法对汉语信号处理的PESQ得分

Fig. 7　PESQ score of the proposed algorithms for Chinese language signals at different distances between source and microphone array

3　结论

本文提出了基于偏度的多通道房间脉冲响应逆滤波方法。该方法不需要已知房间脉冲响应或波达方向的先验知识，采用非高斯性极大的准则实现盲逆滤波。实验结果表明，相比于基于四阶统计量峰度的方法，本文提出方法具有更好的去混响效果，

尤其在混响较强的情况下优势更为明显，且算法复杂度更低，对高斯噪声的鲁棒性更强。应该指出的是，本文所提的方法主要用于抑制早期混响所引起的谱染色现象，而对较长混响时间所引起的拖尾现象抑制不明显，结合谱减法等后处理方法可以对残余晚期混响进行抑制，进而进一步提升可懂度。其次，在研究中发现，在混响较强情况下，相比于多通道方法，单通道算法表现出了明显的局限性。另外，在实际应用中，本文所提方法的实时处理问题也是值得进一步深入研究的。

参 考 文 献

[1] Flanagan J L, Johnston J D, Zahn R, et al. Computer-steered microphone arrays for sound transduction in large rooms[J]. The Journal of the Acoustical Society of America, 1985, 78(5): 1508–1518.

[2] 肖栋, 向阳, 卓瑞岩, 等. 基于波束形成的多类型多声源定位研究 [J]. 应用声学, 2017, 36(3): 220–227.
Xiao Dong, Xiang Yang, Zhuo Ruiyan, et al. Location of multiple sound source with multi-type based on beamforming[J]. Journal of Applied Acoustics, 2017, 36(3): 220–227.

[3] Boll S F. Suppression of acoustic noise in speech using spectral subtraction[C]//Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP. IEEE, 1979: 200–203.

[4] Li R, Bao C, Xia B, et al. Speech enhancement using the combination of adaptive wavelet threshold and spectral subtraction based on wavelet packet decomposition[C]//Signal Processing (ICSP), 2012 IEEE 11th International Conference on. IEEE, 2012, 1: 481–484.

[5] 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.

[6] Fang Y, Feng H, Chen Y. A robust interaural time differences estimation and dereverberation algorithm based on the coherence function[J]. Applied Acoustics, 2018, 129: 126–134.

[7] Peng R, Tan Z H, Li X, et al. A perceptually motivated LP residual estimator in noisy and reverberant environments[J]. Speech Communication, 2018, 96: 129–141.

[8] Zheng C, Peng R, Li J, et al. A constrained MMSE LP residual estimator for speech dereverberation in noisy environments[J]. IEEE Signal Processing Letters, 2014, 21(12): 1462–1466.

[9] 赵红, 李双田. 改进的多级线性预测晚期混响抑制算法 [J]. 信号处理, 2014, 30(6): 674–682.
Zhao Hong, Li Shuangtian. Improved late reverberation suppression algorithm using multiple-step linear prediction[J]. Journal of Signal Processing, 2014, 30(6): 674–682.

[10] Papoulis A, Hoffman J G. Probability, random variables, and stochastic processes[M]. New York: McGraw-Hill, 2013.

[11] Gillespie B W, Malvar H S, Florêncio D A F. Speech dereverberation via maximum-kurtosis subband adaptive filtering[C]//Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on. IEEE, 2001, 6: 3701–3704.

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

[13] Pääjärvi P, Leblanc J. Skewness maximization for impulsive sources in blind deconvolution[C]//Nordic Signal Processing Symposium: 09/06/2004-11/06/2004. Helsinki University of Technology, 2004: 304–307.

[14] Mosayyebpour S, Sheikhzadeh H, Gulliver T A, et al. Single-microphone LP residual skewness-based inverse filtering of the room impulse response[J]. IEEE Transactions on Audio, Speech & Language Processing, 2012, 20(5): 1617–1632.

[15] Yoshioka T, Hikichi T, Miyoshi M, et al. Robust decomposition of inverse filter of channel and prediction error filter of speech signal for dereverberation[C]//Signal Processing Conference, 2006 14th European. IEEE, 2006: 1–5.

[16] Rabiner L R, Schafer R W. Digital processing of speech signals[M]. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[17] Allen J B, Berkley D A. Image method for efficiently simulating small-room acoustics[J]. The Journal of the Acoustical Society of America, 1979, 65(4): 943–950.

[18] 饶宇安. 关于声学比 -混响时间 -语言清晰度关系的实验与理论计算 [J]. 声学学报, 1981, 17(1): 20–33.
Rao Yu'an. Experimental and theoretical calculation of the relationship among acoustic ratio, reverberation time and speech intelligibility[J]. Acta Acustica, 1981, 17(1): 20–33.

[19] Habets E A P. Single- and multi-microphone speech dereverberation using spectral enhancement[J]. Dissertation Abstracts International, 2007, 68(4): 10.6100/IR627677.

[20] 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]. Proceedings of the Acoustics, Speech, and Signal Processing, 2001, 2: 749–752.

[21] Fall 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 & Language Processing, 2010, 18(7): 1766–1774.

[22] Schwarz A, Kellermann W. Coherent-to-diffuse power ratio estimation for dereverberation[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2015, 23(6): 1006–1018.

[23] Zheng C, Tan Z H, Peng R, et al. Guided spectrogram filtering for speech dereverberation[J]. Applied Acoustics, 2018, 134(5): 154–159.