ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 5,Sep. 2024,pp. 1135－1146
DOI:10. 16337/j. 1004-9037. 2024. 05. 007

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

利用互子带滤波器和稀疏特性的多通道线性预测语音去混响方法

康　瑶[1]，康　坊[2]，杨飞然[3]

(1. 国家开放大学数字化部,北京 100039;2. 奥卢大学机器视觉与信号分析中心，奥卢 90570;3. 中国科学院噪声与振动重点实验室(声学研究所),北京 100190)

摘　要：多通道线性预测是最为流行的语音去混响方法之一,现有相关研究大多利用子带谱减模型在每一个频带独立地获取期望信号,但这忽略了不同子带之间的相互影响。本文提出一种利用互子带谱减模型的多通道线性预测语音去混响方法。相比于大多数方法采用的子带谱减模型,本文方法采用的互子带谱减模型能够利用互子带滤波器来对不同子带之间的相互影响进行建模。本文方法利用复广义高斯分布建模期望信号,相比于常用的高斯分布,复广义高斯分布能够通过调整形状参数来描述语音信号的稀疏特性。在最大似然估计框架下,将语音去混响转化为关于互子带滤波器和子带滤波器的优化问题;并且基于替代最小化方法推导了保证收敛的优化算法。在不同混响时间、不同通道、不同声源和传声器距离情况下的一系列语音去混响实验验证了本文方法的性能显著优于传统去混响算法。

关键词：语音去混响;多通道线性预测;互子带滤波器;复广义高斯分布;替代最小化

中图分类号：TN912.3　　**文献标志码：**A

Multi-channel Linear Prediction for Speech Dereverberation Using Cross-Band Filters and Sparse Priors

KANG Yao[1], KANG Fang[2], YANG Feiran[3]

(1. Digitalization Department, Open University of China, Beijing 100039, China;2. Center for Machine Vision and Signal Analysis, University of Oulu, Oulu 90570, Finland;3. Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The multi-channel linear prediction (MCLP) is one of the most popular speech dereverberation methods. The band-to-band spectral subtraction model has been adopted by most existing studies to obtain the desired speech signal in each frequency band, but it neglects the interaction between different frequencies. This paper proposes a MCLP-based speech dereverberation method using the cross-band spectral subtraction model instead of the widely adopted band-to-band spectral subtraction model. The proposed model employs cross-band filters to account for the interactions between different frequencies. We model the desired signal using the complex generalized Gaussian (CGG) distribution. Compared with the Gaussian distribution, the CGG distribution can capture the sparse nature of speech signals using a suitable

基金项目：国家自然科学基金面上项目(62171438);北京市自然科学基金(4242013);中国科学院声学研究所自主部署“前沿探索”类项目(QYTS202111);2023年度国家开放大学重点科研项目(Z23C0007)。
收稿日期：2024-06-19;**修订日期：**2024-08-23

shape parameter. Within the maximum likelihood estimation framework, the speech dereverberation problem is formulated as an optimization problem involving the band-to-band and cross-band filters. An optimization algorithm with guaranteed convergence is derived based on the majorization-minimization method. A series of speech dereverberation experiments under various reverberation times, different channel numbers and different source-to-microphone distances demonstrate that the proposed method significantly outperforms traditional methods in terms of dereverberation performance.

Key words: speech dereverberation; multi-channel linear prediction; cross-band filter; complex generalized Gaussian distribution; majorization minimization

引　　言

在密闭空间中,传声器采集到的语音信号不仅包括直达声,还包括早期反射声和后期混响声[1-3]。语音去混响技术能够通过去除语音信号中的混响成分来提高语音质量和语音识别率、唤醒率等指标[4-5],因而通常被集成到如电话会议和智能语音助手等语音通信及人机语音交互系统中。

近年来,国内外研究学者在语音去混响领域取得了丰富的研究成果,提出了大量的语音去混响方法[1,6]。这些方法可分为两类:基于深度神经网络的方法和基于统计信号处理的方法。基于深度神经网络的语音去混响方法利用网络模型学习到的混响信号和期望语音信号之间的非线性映射关系直接恢复出期望信号[7-10],或者利用网络模型学习到的干净语音信号内在的分布特点间接地生成期望信号[10]。尽管这类方法能取得较好的去混响效果,但是网络模型需要较多的带标签数据进行监督训练,并且网络模型的训练过程需要占用较多的计算和存储资源[11]。基于统计信号处理的方法由于其无需监督学习及推理过程复杂度较低的特点也有其应用价值。这些研究大致可分为两类,其中一类先辨识声学多通道传递函数,然后再用反卷积滤波器对辨识的声学通道进行解卷积得到干净的语音信号[12-14]。因为多通道传递函数辨识本身是一个非常困难的任务,这类方法的难度较大[15-16]。为了克服通道辨识的难题,另一类方法无需对声学通道进行辨识,而是直接估计干净的语音信号。多通道线性预测(Multi-channel linear prediction, MCLP)便属于这类方法的范畴,并且已经被证明是一种非常有效的去混响方法[17-25]。在短时傅里叶变换域,MCLP方法通过谱减模型来获得期望信号,然后通过对期望信号进行先验分布建模来求解谱减模型的参数。在这类方法中,研究者非常关注期望信号的先验分布,通过先验分布建模语音信号的时变、低秩和稀疏等特性。加权预测误差(Weighted prediction error, WPE)[20-21]将期望信号建模为时变高斯分布,利用高斯分布的时变方差建模语音信号的时变特点。Jukic等[26]进一步将时变高斯分布的方差进行非负矩阵分解,利用非负矩阵分解模型来建模语音信号的低秩特性,有效提升了MCLP滤波器的去混响效果[25]。但是,非负矩阵分解模型的参数需要较多次的迭代才能够收敛。为了利用语音信号的稀疏特性,也有文献采用复广义高斯(Complex generalized Gaussian, CGG)分布建模MCLP的期望信号[22],其中CGG分布中的形状参数能够反映期望信号分布的稀疏程度。上述所有研究均是采用子带谱减模型在不同的频带上独立地建模期望信号。短时傅里叶变换的理论[27]表明,不同子带之间并非完全独立,当短时傅里叶分析窗的长度远小于混响时间时[28],用子带谱减模型获得的期望信号存在建模误差。为了解决这个问题,Cohen等[30]提出利用互子带滤波器建立谱减模型。但是,Cohen等所采用的互子带滤波器仅考虑了相邻的两个子带,并且建模期望信号所采用的时变高斯分布仅能利用语音信号的时变特点,这可能会限制该方法的性能。

本文基于互子带谱减模型提出一种考虑语音信号稀疏特点的多通道线性预测语音去混响方法。本

文方法采用互子带谱减模型获取期望信号,然后利用CGG声源模型来建模期望信号的分布。本文方法采用的互子带谱减模型考虑了不同子带之间的相互影响,因而比子带模型具有更小的建模误差。与Cohen等提出的方法相比,本文方法采用广义高斯分布来建模语音频谱的稀疏性,从而获得了更好的去混响性能。在最大似然估计的框架下,语音去混响问题被建模为子带和互子带多通道线性预测滤波器的优化问题。基于替代最小化方法推导了保证收敛的迭代优化算法。性能验证实验以及在不同混响时间、不同声源-传声器距离和不同通道数量的声学场景下的对比实验验证了本文方法性能的优越性。

1　基于子带谱减模型的多通道线性预测

1.1　子带谱减模型

考虑在一个房间中利用传声器阵列拾取语音信号的场景。第1个传声器拾取的信号$x_1(j)$在时域可以被表示为

$$x_1(j)=s(j)*h(j) \tag{1}$$

式中:$s(j)$表示声源信号;"*"表示卷积操作;$h(j)$表示房间脉冲响应。利用短时傅里叶变换,式(1)在频域通常被表示为子带卷积模型

$$x_{1,f,t}\approx\sum_{p=0}^{P-1}h_{f,p}^{*}s_{f,t-p} \tag{2}$$

式中:$h_{f,p}$为声源到参考传声器的P阶滤波器的第p个系数;上标*表示共轭;$s_{f,t}$表示声源信号$s(j)$的短时傅里叶变换,并且$f\in\{1,2,\cdots,F\}$表示频率索引,F表示频带数,$t\in\{1,2,\cdots,T\}$表示时间帧索引,T表示总时间帧数。子带卷积模型(2)可被近似为[20]

$$x_{1,f,t}=d_{f,t}+\sum_{m=1}^{M}\sum_{l=0}^{L-1}w_{m,f,l}^{*}x_{m,f,t-D-l} \tag{3}$$

式中:$d_{f,t}$为期望信号;M为传声器数量;$w_{m,f,l}$为参考通道的MCLP滤波器系数;L为延迟预测滤波器的阶数;D为时间延迟。时间延迟的作用是降低预测信号和直达信号之间的相关性[29]。式(3)中的第1项期望信号的频谱$d_{f,t}=\sum_{p=0}^{D-1}h_{f,p}s_{f,t-p}$,包含了声源到参考传声器的直达声部分和早期反射声部分,其中声源的早期反射声有助于提升语音质量。式(3)中的第2项是利用子带滤波器预测的晚期混响信号的频谱。

利用式(3),期望信号可以表示为

$$d_{f,t}=x_{1,f,t}-\boldsymbol{w}_f^{\mathrm{H}}\bar{\boldsymbol{x}}_{f,t-D} \tag{4}$$

式中:$\boldsymbol{w}_f=\left[\boldsymbol{w}_{1,f}^{\mathrm{H}},\boldsymbol{w}_{2,f}^{\mathrm{H}},\cdots,\boldsymbol{w}_{M,f}^{\mathrm{H}}\right]^{\mathrm{H}}\in\mathbf{C}^{ML\times1}$为子带MCLP滤波器,$\boldsymbol{w}_{m,f}=[w_{m,f,0}^{*},w_{m,f,1}^{*},\cdots,w_{m,f,L-1}^{*}]^{\mathrm{H}}\in\mathbf{C}^{L\times1}$,上标H表示共轭转置;$\bar{\boldsymbol{x}}_{f,t-D}=\left[\bar{\boldsymbol{x}}_{1,f,t-D}^{\mathrm{T}},\bar{\boldsymbol{x}}_{2,f,t-D}^{\mathrm{T}},\cdots,\bar{\boldsymbol{x}}_{M,f,t-D}^{\mathrm{T}}\right]^{\mathrm{T}}\in\mathbf{C}^{ML\times1}$;$\bar{\boldsymbol{x}}_{m,f,t-D}=[x_{m,f,t-D},x_{m,f,t-D-1},\cdots,x_{m,f,t-D-L+1}]^{\mathrm{T}}\in\mathbf{C}^{L\times1}$为第$m$个通道延迟的观测信号,上标T表示转置。因而只需求解出子带MCLP滤波器系数,便可以通过子带谱减模型(4)得到期望信号。

1.2　声源模型、代价函数及优化方法

通过对期望信号的先验分布进行建模,然后最大化期望信号的似然函数能够求解谱减模型中的滤波器参数。加权预测误差方法假设期望信号$d_{f,t}$服从均值为0、方差为$\lambda_{f,t}$的时变高斯分布,有

$$p(d_{f,t})=\frac{1}{\sqrt{2\pi\lambda_{f,t}}}\exp\left(-\frac{1}{2}\frac{d_{f,t}^{2}}{\lambda_{f,t}}\right) \tag{5}$$

将式(4)代入式(5),WPE在每一个子带最小化期望信号的负对数似然函数[18],有

$$\mathcal{L}(\boldsymbol{w}_f)=\sum_{t=1}^{T}\frac{\left|x_{1,f,t}-\boldsymbol{w}_f^{\mathrm{H}}\bar{\boldsymbol{x}}_{f,t}\right|^2}{\lambda_{f,t}}+\sum_{t=1}^{T}\log\lambda_{f,t} \tag{6}$$

得到MCLP滤波器的闭式解为

$$\boldsymbol{w}_f=\boldsymbol{R}_f^{-1}\boldsymbol{p}_f \tag{7}$$

式中

$$\boldsymbol{R}_f=\sum_{t=1}^{T}\frac{\bar{\boldsymbol{x}}_{f,t}\bar{\boldsymbol{x}}_{f,t}^{\mathrm{H}}}{\lambda_{f,t}} \tag{8}$$

$$\boldsymbol{p}_f=\sum_{t=1}^{T}\frac{\bar{\boldsymbol{x}}_{f,t}x_{1,f,t}^{*}}{\lambda_{f,t}} \tag{9}$$

声源方差$\lambda_{f,t}$的解通过令$\frac{\partial\mathcal{L}}{\partial\lambda_{f,t}}=0$得到

$$\lambda_{f,t}=\left|d_{f,t}\right|^2 \tag{10}$$

一般需要对声源方差进行预处理$\lambda_{f,t}=\max\left\{\lambda_{f,t},\varepsilon_{\min}\right\}$,以避免计算式(8)和式(9)时产生数值奇异,其中$\varepsilon_{\min}$是一个很小的正数。求解出子带滤波器$\boldsymbol{w}_f$之后,加权预测误差方法通过子带谱减模型(4)得到期望信号,也就是去混响信号。

2 基于互子带谱减模型和复广义高斯分布的多通道线性预测语音去混响

短时傅里叶变换的理论[27]表明,不同子带之间的信号并非完全独立,这意味着利用子带谱减模型(4)获得的期望信号存在逼近误差。为了解决这个问题,本文利用互子带谱减模型来获得期望信号。相比于文献[30]中仅考虑相邻两个子带的互子带谱减模型,本文所采用的互子带谱减模型通过把互子带数量参数化来考虑更多子带之间的相互影响。进一步地,采用复广义高斯分布而不是WPE所采用的高斯分布来建模期望信号的统计特性,能够有效地利用语音信号的稀疏特性。

2.1 互子带谱减模型

时域模型(1)在频域可以被表示为一系列互子带滤波器的和[28],即

$$x_{1,f,t}=\sum_{p=0}^{P-1}\sum_{f'=0}^{F-1}h_{f,f',p}s_{f',t-p} \tag{11}$$

式中$h_{f,f',l}$表示互子带滤波器。式(11)是式(1)的精确表达,但是式(11)中互子带滤波器和子带滤波器的参数总量为$F\times F\times P$,远大于式(2)中的子带滤波器数量$F\times P$。本文采用子带滤波器和$2Q$个子带之间的互子带滤波器来近似式(1),即

$$x_{1,f,t}\approx\sum_{p=0}^{P-1}h_{f,f,p}s_{f,t-p}+\sum_{q=0}^{P_{cb}-1}\sum_{f'=f-Q,f'\neq f}^{f+Q}h_{f,f'}s_{f',t-q} \tag{12}$$

相比于式(2)的子带卷积模型,式(12)由于采用了前P_{cb}个时刻的$2Q$个子带之间的互子带滤波器而能更加准确地近似时域模型(1)。本文所采用的互子带卷积模型(12)采用了$2Q$个互子带的滤波器可以被进一步表示为[20]

$$x_{1,f,t}=d_{f,t}+\sum_{m=1}^{M}\sum_{l_{bb}=0}^{L-1}w_{m,f,l}^{*}x_{m,f,t-D-l_{bb}}+\sum_{m=1}^{M}\sum_{l_{cb}=0}^{L_{cb}-1}\sum_{f'=f-Q,f'\neq f}^{f+Q}g_{m,f,f'}^{*}x_{m,f',t-l_{cb}} \tag{13}$$

式中:L_{cb}表示每两个子带之间互子带MCLP滤波器的阶数;$g_{m,f,f'}$表示子带f和子带f'之间的互子带多通道线性预测滤波器。式(13)可以被表示为利用互子带滤波器的谱减模型

$$d_{f,t}=x_{1,f,t}-\boldsymbol{w}_f^{\mathrm{H}}\bar{\boldsymbol{x}}_{f,t-D}-\boldsymbol{g}_f^{\mathrm{H}}\tilde{\boldsymbol{x}}_{f,t-D} \tag{14}$$

式中:$\boldsymbol{w}_f=\left[\boldsymbol{w}_{1,f}^{\mathrm{H}},\boldsymbol{w}_{2,f}^{\mathrm{H}},\cdots,\boldsymbol{w}_{M,f}^{\mathrm{H}}\right]^{\mathrm{H}}\in\mathbf{C}^{ML\times 1}$ 为子带 MCLP 滤波器;$\boldsymbol{w}_{m,f}=[w_{m,f,0}^{*},w_{m,f,1}^{*},\cdots,w_{m,f,L_b-1}^{*}]^{\mathrm{H}}\in\mathbf{C}^{L\times 1}$;$\bar{\boldsymbol{x}}_{f,t-D}=\left[\bar{\boldsymbol{x}}_{1,f,t-D}^{\mathrm{T}},\bar{\boldsymbol{x}}_{2,f,t-D}^{\mathrm{T}},\cdots,\bar{\boldsymbol{x}}_{M,f,t-D}^{\mathrm{T}}\right]^{\mathrm{T}}\in\mathbf{C}^{ML\times 1}$;$\bar{\boldsymbol{x}}_{m,f,t-D}=[x_{m,f,t-D},x_{m,f,t-D-1},\cdots,x_{m,f,t-D-L_b+1}]^{\mathrm{T}}\in\mathbf{C}^{L\times 1}$ 为第 m 个通道延迟的观测信号;$\boldsymbol{g}_f=\left[\bar{\boldsymbol{g}}_{1,f}^{\mathrm{H}},\bar{\boldsymbol{g}}_{2,f}^{\mathrm{H}},\cdots,\bar{\boldsymbol{g}}_{M,f}^{\mathrm{H}}\right]^{\mathrm{H}}\in\mathbf{C}^{2QML_{cb}\times 1}$ 为互子带 MCLP 滤波器;$\bar{\boldsymbol{g}}_{m,f}=[\boldsymbol{g}_{m,f,f-Q}^{\mathrm{H}},\boldsymbol{g}_{m,f,f-Q+1}^{\mathrm{H}},\cdots,\boldsymbol{g}_{m,f,f+Q}^{\mathrm{H}}]^{\mathrm{H}}\in\mathbf{C}^{2QL_{cb}\times 1}$;$\boldsymbol{g}_{m,f,f'}=[g_{m,f,f',0}^{*},g_{m,f,f',1}^{*},\cdots,g_{m,f,f',L_{cb}-1}^{*}]^{\mathrm{H}}\in\mathbf{C}^{L_{cb}\times 1}$;$\tilde{\boldsymbol{x}}_{f,t-D}=\left[\tilde{\boldsymbol{x}}_{1,f,t-D}^{\mathrm{T}},\tilde{\boldsymbol{x}}_{2,f,t-D}^{\mathrm{T}},\cdots,\tilde{\boldsymbol{x}}_{M,f,t-D}^{\mathrm{T}}\right]^{\mathrm{T}}\in\mathbf{C}^{2QML_{cb}\times 1}$;$\tilde{\boldsymbol{x}}_{m,f,t-D}=[\tilde{\tilde{\boldsymbol{x}}}_{m,f-Q,t-D}^{\mathrm{T}},\tilde{\tilde{\boldsymbol{x}}}_{m,f-Q+1,t-D}^{\mathrm{T}},\cdots,\tilde{\tilde{\boldsymbol{x}}}_{m,f,t-D}^{\mathrm{T}}]^{\mathrm{T}}\in\mathbf{C}^{2QL_{cb}\times 1}$;$\tilde{\tilde{\boldsymbol{x}}}_{m,f,t-D}=[\tilde{x}_{m,f,t-D-1},\tilde{x}_{m,f,t-D-2},\cdots,\tilde{x}_{m,f,t-D-L_{cb}+1}]^{\mathrm{T}}\in\mathbf{C}^{L_{cb}\times 1}$。相比于文献[30]所采用的仅考虑相邻两个子带之间互子带的谱减模型不同,本文互子带谱减模型(14)由于采用了 $2Q$ 个子带之间的互子带滤波器能更好地建模期望信号。

2.2　声源模型和代价函数

为了更加准确地估计子带和互子带 MCLP 滤波器,本文方法采用复广义高斯分布来建模期望信号

$$p(d_{f,t})=\frac{\beta}{2\pi\gamma^2\Gamma\left(2/\beta\right)}\exp\left(-\frac{\left|d_{f,t}\right|^{\beta}}{\gamma^{\beta}}\right) \tag{15}$$

式中:$\gamma>0$ 为幅度参数;β 为形状参数;$\Gamma(\cdot)$ 为 gamma 函数。当 $\beta=2$ 时,CGG 分布退化为高斯分布。当形状参数 $\beta=1$ 时,CGG 分布退化为拉普拉斯分布。形状参数反映了期望信号分布的稀疏程度,更小的形状参数导致了 $d_{f,t}=0$ 处出现更高的尖峰,这带来了更稀疏的先验分布[22]。广义高斯分布在语音去混响、基于波束形成的语音增强以及音频盲源分离任务中已经被证明能够有效地建模声源的概率分布,并且取得了优于高斯分布的效果。

将式(14)代入式(15),根据最大似然准则得到期望信号在每一个子带的负对数似然函数为

$$\mathcal{L}\left(\boldsymbol{w}_f,\boldsymbol{g}_f\right)=\sum_{t=1}^{T}\frac{\left|x_{1,f,t}-\boldsymbol{w}_f^{\mathrm{H}}\bar{\boldsymbol{x}}_{f,t-D}-\boldsymbol{g}_f^{\mathrm{H}}\tilde{\boldsymbol{x}}_{f,t-1}\right|^{\beta}}{\gamma^{\beta}}+2\log\gamma \tag{16}$$

式(15)和 WPE 的代价函数具有相似的形式,优化式(15)等价于最小化被稀疏约束加权的预测误差。相比于 WPE 的似然函数(6)是关于期望信号的二次方,代价函数(16)是关于期望信号 β 次方,当 $0<\beta<2$ 时,能够有效地利用语音信号的稀疏特性来约束期望信号。

2.3　优化方法

代价函数(16)引入了稀疏参数,所以不能直接采用 MCLP 滤波器的求解方式解得未知参数。本文采用 MM 算法[31]对代价函数(16)进行优化。在 MM 算法的框架中,首先根据目标函数设计一个具有闭式解的 majorization 函数(即替代函数),然后优化 majorization 函数而非目标函数。由于 majorization 函数被保证从不在目标函数的下边,所以 MM 算法的收敛性是被保证的[31]。

首先利用 MM 算法优化式(16)。考虑 $\beta\in(0,2)$ 的情况,根据加权算术平均和几何平均不等式[32],得到

$$\left|d_{f,t}\right|^{\beta}\leqslant\frac{\beta}{2\xi_{f,t}^{2-\beta}}\left|d_{f,t}\right|^{2}+\left(1-\frac{\beta}{2}\right)\xi_{f,t}^{\beta} \tag{17}$$

式中 $\xi_{f,t}>0$ 为辅助变量。当且仅当 $\xi_{f,t}=\left|d_{f,t}\right|$ 时,式(17)中的等号成立。利用式(17),本文构造一个式(16)的辅助函数 $\mathcal{L}^{+}(\boldsymbol{w}_f,\boldsymbol{g}_f)$,并且该函数满足

$$\mathcal{L}(\boldsymbol{w}_f,\boldsymbol{g}_f)\leqslant\mathcal{L}^{+}(\boldsymbol{w}_f,\boldsymbol{g}_f)=\sum_{f=1}^{F}\sum_{t=1}^{T}\frac{1}{\gamma^{\beta}}\left(\frac{\beta}{2\xi_{f,t}^{2-\beta}}\left|x_{1,f,t}-\boldsymbol{w}_f^{\mathrm{H}}\bar{\boldsymbol{x}}_{f,t-D}-\boldsymbol{g}_f^{\mathrm{H}}\tilde{\boldsymbol{x}}_{f,t-1}\right|^2-\left(1-\frac{\beta}{2}\right)\xi_{f,t}^{\beta}\right) \tag{18}$$

式中当且仅当$\xi_{f,t}=\left|x_{1,f,t}-\boldsymbol{w}_f^{\mathrm{H}}\bar{\boldsymbol{x}}_{f,t-D}-\boldsymbol{g}_f^{\mathrm{H}}\tilde{\boldsymbol{x}}_{f,t-1}\right|$时等号成立。辅助函数$\mathcal{L}^{+}(\boldsymbol{w}_f,\boldsymbol{g}_f)$是关于$\left[\boldsymbol{w}_f^{\mathrm{H}},\boldsymbol{g}_f^{\mathrm{H}}\right]^{\mathrm{H}}$的二次函数，因而$\left[\boldsymbol{w}_f^{\mathrm{H}},\boldsymbol{g}_f^{\mathrm{H}}\right]^{\mathrm{H}}$具有闭式解。联合求解子带MCLP滤波器$\boldsymbol{w}_f$和互子带MCLP滤波器$\boldsymbol{g}_f$，令$\dfrac{\partial\mathcal{L}^{+}(\boldsymbol{w},\boldsymbol{g}_f)}{\partial[\boldsymbol{w}_f^{\mathrm{H}},\boldsymbol{g}_f^{\mathrm{H}}]}=0$得

$$\left[\boldsymbol{w}_f^{\mathrm{H}},\boldsymbol{g}_f^{\mathrm{H}}\right]^{\mathrm{H}}=\boldsymbol{R}_f^{-1}\boldsymbol{p}_f \tag{19}$$

式中

$$\boldsymbol{R}_f=\sum_{t=1}^{T}\frac{\left[\bar{\boldsymbol{x}}^{\mathrm{H}}{}_{f,t-D},\tilde{\boldsymbol{x}}^{\mathrm{H}}{}_{f,t-1}\right]^{\mathrm{H}}\left[\bar{\boldsymbol{x}}_{f,t-D},\tilde{\boldsymbol{x}}_{f,t-1}\right]}{\sigma_{f,t}} \tag{20}$$

$$\boldsymbol{p}_f=\sum_{t=1}^{T}\frac{\left[\bar{\boldsymbol{x}}^{\mathrm{H}}{}_{f,t-D},\tilde{\boldsymbol{x}}^{\mathrm{H}}{}_{f,t-1}\right]^{\mathrm{H}}x_{m,f,t}^{*}}{\sigma_{f,t}} \tag{21}$$

$$\sigma_{f,t}=\max\left\{d_{f,t}^{2-\beta},\varepsilon_{\min}\right\} \tag{22}$$

相比于WPE和CB-WPE(Cross-band WPE)方法中的归一化因子$d_{f,t}^{2}$，所提方法采用的归一化因子$d_{f,t}^{2-\beta}$通过形状参数能够有效地考虑语音信号的稀疏特性。

另外，观察到本文方法采用的归一化因子$d_{f,t}^{2-\beta}$是WEP方法中归一化因子$d_{f,t}^{2}$的广义形式。当$\beta=0$，并且$Q=0$时，本文方法退化为传统的WPE方法，所以传统的WPE方法没有利用语音信号的稀疏特性；当$\beta=0$，并且$Q=1$时，本文方法退化为CB-WEP方法，因而本文方法相比于CB-WPE方法能够考虑更多子带之间的互子带滤波器。总的来说，本文方法通过迭代地估计子带和互子带MCLP滤波器，能够最终通过互子带谱减模型(14)得到期望信号，本文方法被称为CB-CGG-WPE算法，算法1给出了所提方法的伪代码。

算法1 CB-CGG-WPE算法

参数设置：L, D, β, K, Q, L_{cb}, I, J, $\varepsilon_{\min}$

输入：$x_{m,f,t}$

初始化：$\boldsymbol{w}_f\leftarrow 0_{ML\times 1}$, $\boldsymbol{g}_f\leftarrow 0_{2Q\times 1}$, $i\leftarrow 1$

当 $i\leqslant I$ 执行：

对于每一个频带和每一个时间帧，按照式(14)计算 $d_{f,t}$

对于每一个频带和每一个时间帧，按照式(22)计算 $\sigma_{f,t}$

对于每一个频带，按照式(20)计算 $\boldsymbol{R}_f$

对于每一个频带，按照式(21)计算 $\boldsymbol{p}_f$

对于每一个频带，按照式(19)计算 $\boldsymbol{w}_f$ 和 $\boldsymbol{g}_f$

$i\leftarrow i+1$

结束

输出：$d_{f,t}$

3 实验和结果评价

本文通过一系列实验来验证所提方法的性能。实验中从TIMIT数据集中随机选取了30段12 s长的源信号组成测试集。观测信号通过将测试集中的源信号和房间脉冲响应做卷积运算得到。本文利用镜

像法[33]生成房间脉冲响应,其中1个声源和由4个传声器组成的间距为8 cm标准线阵被放置在7 m × 5 m × 2.5 m(长×宽×高)的房间中。采用声源-传声器间距0.5 m和2 m,分别表示近场和远场情景。因而一共有4种场景,如表1所示,分别为近场双通道、近场四通道、远场双通道和远场四通道。在每一个场景中,设置混响时间从0.5 s以0.1 s的步长变化到1 s。

表1　不同场景的传声器数量和声源-传声器间距

Table 1　Microphone numbers and source-to-microphone distances in different scenarios

场景	传声器数量	声源-传声器间距/m
场景1	2	0.5
场景2	4	0.5
场景3	2	2.0
场景4	4	2.0

本文将所提方法与WPE[20]和CB-WPE[30]进行比较。在所有的实验中,采样率为16 kHz,短时傅里叶变换的分析窗为32 ms的Hamming窗,帧移为窗长的一半。时间延迟$D=1$,滤波器的长度与混响时间及通道数有关[23]。随着混响时间从0.5 s增加到1 s,四通道场景的子带滤波器长度为$L\in\{15, 17, 19, 21, 23, 25\}$。相比于四通道场景,双通道场景中的MCLP滤波器长度增加一倍。MCLP滤波器被初始化为零向量。根据文献[22],本文设置所有算法的迭代次数为6,最小值$\varepsilon_{\min}=10^{-8}$。本文采用PESQ(Perceptual evaluation of speech quality),归一化SRMR(Speech-to-reverberation modulation energy ratio),CD(Cepstral distance),FWSNR(Frequency-weighted segmental signal-to-noise ratio)和ESTOI(Extended Short-time objective intelligibility)来评价算法的性能[6],越高的PESQ、SRMR、FWSNR和ESTOI值以及越低的CD值表明越好的去混响性能。

3.1　性能验证实验

为了证明本文方法的有效性,首先研究了所采用的互子带滤波器和复广义高斯分布的有效性。图1展示了所提方法处理场景4中混响时间为0.5 s的数据时取得的平均PESQ随着互子带数量(2Q)和互子带MCLP滤波器长度(L_{cb})变化的情况,其中形状参数$\beta=0.5$。FWSNR、SRMR、ESTOI和CD值的变化情况与PESQ相似,因而考虑到空间限制并未在文中展示。从图1可以看出,互子带数量及每一个互子带滤波器的长度对于所提方法的效果均有显著的影响。随着互子带数量的增加以及每一个互子带滤波器数量的增加,本文方法取得的指标均是先变好再变坏,并在$Q=2$和$L_{cb}=3$附近取得最好的效果,这证明了所提互子带谱减模型以及引入更多互子带的有效性。

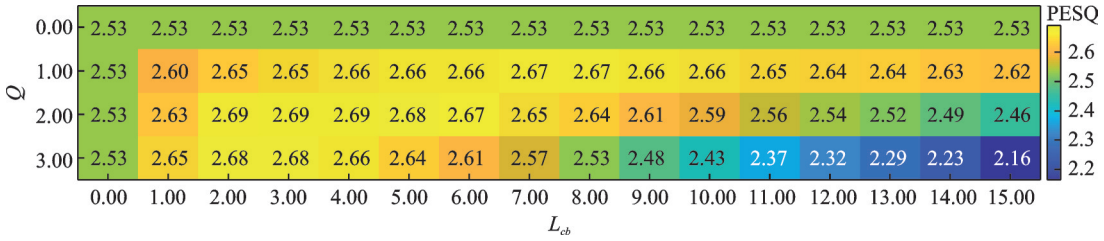

图1　本文方法随着互子带MCLP滤波器参数Q和L_{cb}变化时取得的平均PESQ得分

Fig.1　Average PESQ scores obtained by the proposed method as a function of Q and L_{cb}

图2展示了所提方法处理场景4中混响时间为0.5 s的数据时取得的平均PESQ随着形状参数β变化的情况,其中$Q=2$和$L_{cb}=3$。FWSNR、SRMR、ESTOI和CD值的变换情况和PESQ相似,因而未被展示。可以看出,当形状参数从2逐步减小为0时,本文方法取得的5个指标的平均值均是先变好再变坏,在$\beta=0.5$和0.6附近取得最优值。这是因为适当的形状参数能够更好地描述干净语音信号的分布,进而带来更好的去混响效果。当$\beta=0$时,本文方法退化为传统的CB-WPE方法,性能也显著下降。这证明了本文方法引入复广义高斯分布的有效性。

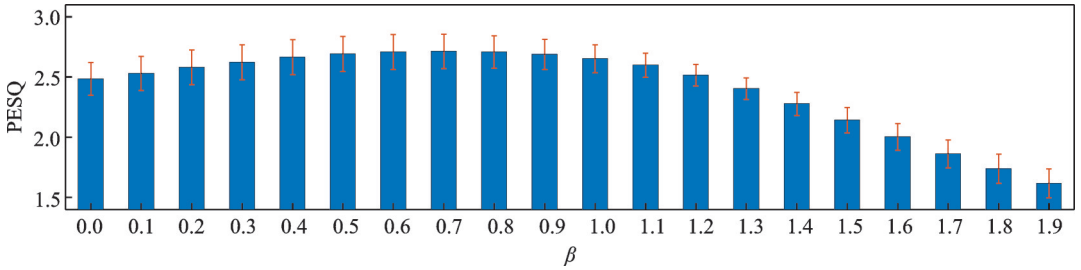

图2 本文方法随着形状参数β变化时取得的平均PESQ得分

Fig.2 Average PESQ scores obtained by the proposed method as a function of β

3.2 性能对比实验

图3～5分别展示了3种方法在4种场景下取得的PESQ、FWSNR和CD提升的平均值，其中横轴为混响时间，误差线表示标准差。3种方法在SRMR、ESTOI指标上的对比结果与在PESQ、FWSNR和CD指标上的对比结果类似，因而不再展示。提升值表示输出值和输入值之间的差值，其中输入值利用第1个通道的观测信号进行计算，并在图中顶部（底部）直接给出。可以看出在大部分的场景下CB-WPE方法能取得优于WPE方法的效果，这是因为利用互子带滤波器的互子带谱减模型比仅利用子带滤波器的子带谱减模型更加精确。在4个场景下的对比实验中，本文方法CB-CGG-WPE均取得了最高的PESQ、FWSNR和CD下降值。这是因为本文方法不仅能够利用互子带谱减模型获得误差更小的期望信号，还通过考虑语音信号的稀疏特性能够更准确地建模期望信号的分布特点。相比于双通道的场景1和场景3，所有方法在四通道的场景2和场景4下均取得了更加明显的性能提升，这表明所提

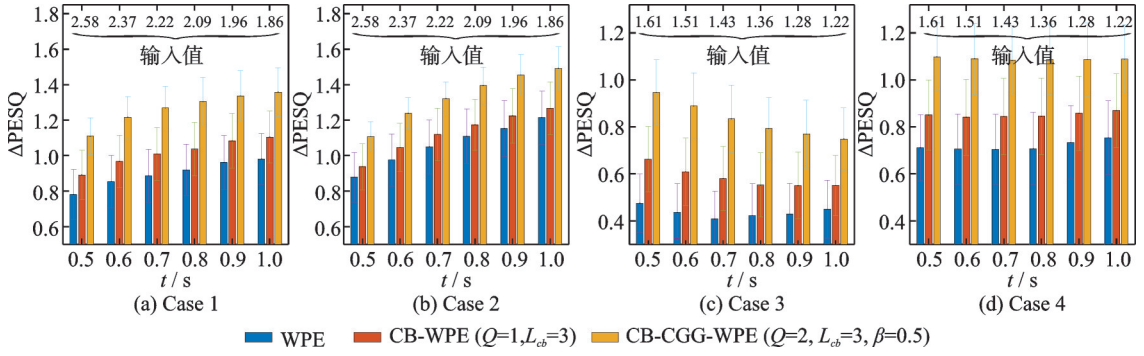

图3 3种方法在4种场景下处理不同混响时间的语音信号时取得的平均PESQ提升值

Fig.3 Average PESQ score improvements obtained by three algorithms for different reverberation time in four cases

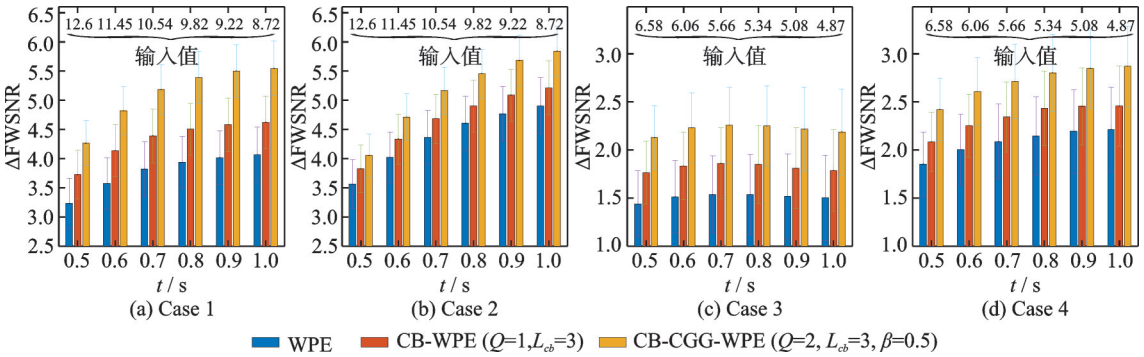

图4 3种方法在4种场景下处理不同混响时间的语音信号时取得的平均FWSNR提升值

Fig.4 Average FWSNR improvements obtained by three algorithms for different reverberation time in four cases

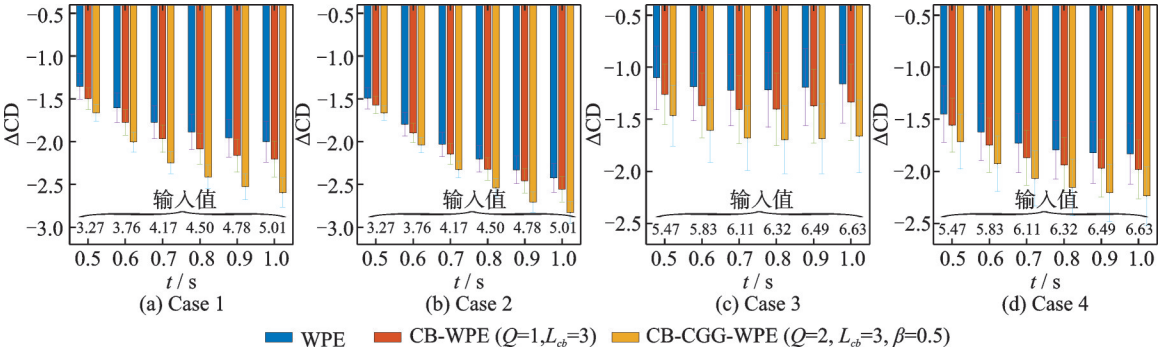

图5　3种方法在4种场景下处理不同混响时间的语音信号时取得的平均值CD提升值

Fig.5　Average CD improvements obtained by three algorithms for different reverberation time in four cases

方法能从更多通道的观测信号中受益。这种现象在文献[22-23]中也被观察到。相比于近场的场景1和场景2,本文方法在远场的场景3和场景4下取得的性能提升则均有所下降,这说明随着声源-传声器间距变大,去混响任务更加困难。

图6展示了一个语音样例信号加混响前后的时频谱,以及利用3种方法对该信号去混响之后得到的时频谱。该样例的混响信号通过场景4中500 ms混响时间的房间脉冲响应生成。从图6中标识的方框区域可以看出,相比于干净语音信号的时频谱,观测信号中由于存在混响成分,所以其时频谱在时间

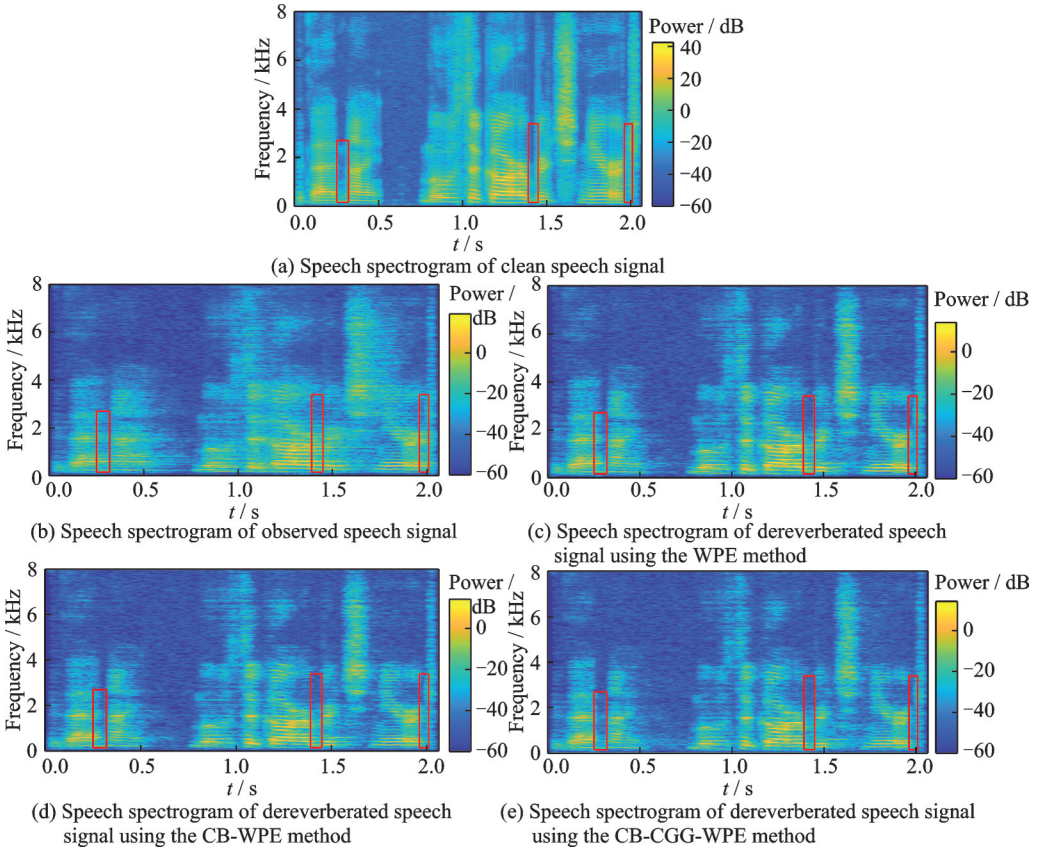

图6　干净信号、观测的混响信号和3种方法去混响得到语音的时频谱

Fig.6　Speech spectrograms of clean, observed, and dereverberated speech signals using three different methods

维度上存在严重的拖尾现象。3种去混响方法均能有效地去除一些混响成分。相比于其他两种方法，利用本文方法得到的去混响语音信号的时频谱最为清晰，在时间维度上的拖尾现象最弱，这也证明了本文方法的有效性。

本文在实录数据集上对比了几种去混响算法的性能。实录数据集包含了30条12 s长的带混响数据，这些数据是在包含两面玻璃墙的会议室由间距为8 cm的4元线阵录制，并且声源和线阵中心的距离为2 m。除了上文提到的两种基于统计信号处理的方法，本实验还增加了一种基于深度神经网络的SGMSE+方法[34]，该方法采用了6.5×10^7参数量的深度神经网络架构。本文直接采用文献[34]开源的预训练模型进行推理。考虑到SGMSE+是一种单通道语音去混响算法，表2同时展示了3种基于统计信号处理的方法在单通道场景下的性能。从表2可以看出，在单通道场景，SGMSE+方法取得的指标均显著优于基于统计信号处理的方法。但是，传统方法通过利用多通道的观测信号可以获得指标的显著提升，本文方法在四通道场景下取得的FWSRN、SRMR、CD和ESTOI值均显著优于SGMSE+在单通道下取得的相应指标。现有的大部分用于去混响的深度神经网络架构针对单通道场景进行设计，所以如何将利用多通道信息的统计信号处理方法与深度神经网络方法相结合是值得研究的方向。

表2 不同方法在实录数据集上的性能(均值±标准差)

Table 2 Performance of different methods on the real-world dataset (mean ± standard deviation)

方法	PESQ	FWSNR	SRMR	CD	ESTOI
观测信号	0.98±0.15	1.82±1.07	1.25±0.10	8.29±0.37	0.24±0.02
WPE ($M=1$)	1.08±0.12	2.20±1.23	1.50±0.13	8.10±0.40	0.28±0.02
CB-WPE ($M=1$)	1.08±0.13	2.22±1.22	1.50±0.13	8.10±0.40	0.29±0.02
CB-CGG-WPE ($M=1$)	1.10±0.14	2.25±1.22	1.53±0.12	8.07±0.40	0.31±0.02
SGMSE+ ($M=1$)	1.56+0.11	2.84+1.23	2.31+1.18	7.83+0.47	0.35+0.04
WPE ($M=4$)	1.37±0.18	2.81±1.35	2.12±0.27	7.77±0.45	0.41±0.04
CB-WPE ($M=4$)	1.40±0.20	2.92±1.37	2.19±0.29	7.72±0.46	0.42±0.04
CB-CGG-WPE ($M=4$)	1.53±0.24	3.08±1.36	2.40±0.33	7.62±0.47	0.47±0.03

3.3 计算复杂度对比实验

表3对比了3种算法在四通道场景下利用不同长度的预测滤波器时每迭代一次所需的平均计算时间。所有算法均在基于Intel的i7-8550U 1.80 GHz CPU上的Matlab平台上运行。从表3中可以看出，相比于WPE方法，CB-WPE方法的计算增加约29%，这是因为CB-WPE比WPE多采用了$2ML_{cb}$个MCLP滤波器系数。相比于CB-WPE，所提方法的计算时间增加约23%，原因在于本文CB-CGG-WPE方法比CB-WPE方法多采用了$2ML_{cb}$个MCLP滤波器系数。

表3 每次迭代所需时间

Table 3 Computational time per iteration s

子带滤波器长度	方法		
	WPE	CB-WPE	CB-CGG-WPE
$L_b=15$	1.43	1.85	2.43
$L_b=17$	1.58	2.12	2.63
$L_b=19$	1.70	2.28	2.79
$L_b=21$	1.85	2.44	2.95
$L_b=23$	2.13	2.64	3.12
$L_b=25$	2.29	2.80	3.33

4 结束语

本文提出了一种同时利用互子带滤波器和语音稀疏特性的多通道线性预测语音去混响方法。首先利用互子带谱减模型获得期望信号，进而采用复广义高斯分布来建模语音谱的稀疏特性。在最大似然估计框架下，语音去混响问题被转化为关于子带和互子带滤波器的最优化问题。基于辅助函数技术，本文推导了保证收敛的优化算法来迭代地估计子带和互子带滤波器参数，发现本文方法具有更加广义的形式，传统的WPE和CB-WPE方法可以看作是所提方法的两个特例。实验验证了本文方法采

用的互子带滤波器以及稀疏声源模型的有效性。在不同混响时间、不同声源-传声器距离、不同通道数量等场景下的一系列实验验证了本文方法比现有算法具有更好的性能。

参考文献:

[1] NAYLOR P A, NIKOLAY D. Speech dereverberation[M]. London: Springer-Verlag, 2010.

[2] 齐园蕾. 语音去混响关键技术研究[M]. 北京:中国科学院声学研究所, 2020.
QI Yuanlei. Key technologies for speech dereverberation research[M]. Beijing: Institute of Acoustics, Chinese Academy of Sciences, 2020.

[3] 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.
ZHANG Xiongwei, LI Yinan, ZHENG Changyan, et al. Speech dereverberation: Review of state-of-the-arts and prospects[J]. Journal of Data Acquisition and Processing, 2017, 32(6): 1069-1081.

[4] 齐园蕾, 杨飞然, 杨军. 基于卡尔曼滤波的低复杂度去混响算法[J]. 应用声学, 2018, 37(4): 559-566.
QI Yuanlei, YANG Feiran, YANG Jun. Kalman filter based low-complexity dereverberation algorithm[J]. Applied Acoustics, 2018, 37(4): 559-566.

[5] YOSHIOKA T, SEHR A, DELCROIX M, et al. Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition[J]. IEEE Signal Processing Magazine, 2012, 29(6): 114-126.

[6] KINOSHITA K, DELCROIX M, GANNOT S, et al. A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research[J]. EURASIP Journal on Advances in Signal Processing, 2016, 2016(1): 1-19.

[7] ZHANG J, PLUMBLEY M D, WANG W. Weighted magnitude-phase loss for speech dereverberation[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, ON, Canada: IEEE, 2021: 5794-5798.

[8] 戴礼荣, 张仕良. 深度语音信号与信息处理: 研究进展与展望[J]. 数据采集与处理, 2014, 29(2): 171-179.
DAI Lirong, ZHANG Shiliang. Deep speech signal and information processing: Research progress and prospect[J]. Journal of Data Acquisition and Processing, 2014, 29(2): 171-179.

[9] QI Y, YANG F, YANG J. A late reverberation power spectral density aware approach to speech dereverberation based on deep neural networks[C]//Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Lanzhou, China: APSIPA, 2019: 1700-1703.

[10] RICHTER J, WELKER S, LEMERCIER K, et al. Speech enhancement and dereverberation with diffusion-based generative models[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2351-2364.

[11] 张鹏程, 郭海燕, 王婷婷, 等. 基于联合图学习的多通道语音增强方法[J]. 数据采集与处理, 2023, 38(2): 283-292.
ZHANG Pengcheng, GUO Haiyan, WANG Tingting, et al. Multi-channel speech enhancement based on joint graph learning [J]. Journal of Data Acquisition and Processing, 2023, 38(2): 283-292.

[12] MALIK S, SCHMID D, ENZNER G . A state-space cross-relation approach to adaptive blind SIMO system identification[J]. IEEE Signal Processing Letters, 2012, 19(8): 511-514.

[13] SCHMID D, ENZNER G, MALIK S, et al. Variational Bayesian inference for multichannel dereverberation and noise reduction[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(8): 1320-1335.

[14] 张航, 赵尚, 林志斌, 等. 基于多通道解卷积的车内声重放系统优化设计[J]. 南京大学学报(自然科学), 2021, 57(6): 1023-1031.
ZHANG Hang, ZHAO Shang, LIN Zhibin, et al. Optimal design of automotive audio sound reproduction system based on multi-channel deconvolution[J]. Journal of Nanjing University(Natural Sciences), 2021, 57(6): 1023-1031.

[15] KHONG W H, LIN X, NAYLOR P A. Algorithms for identifying clusters of near-common zeros in multichannel blind system identification and equalization[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, NV, USA: IEEE, 2008: 389-392.

[16] SCHMID D, ENZNER G. Cross-relation-based blind SIMO identifiability in the presence of near-common zeros and noise[J]. IEEE Transactions on Signal Processing, 2012, 60(1): 60-72.

[17] KINOSHITA K, DELCROIX M, NAKATANI T, et al. Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction[J]. IEEE Transactions on Audio, Speech and Language Processing, 2009, 17(4): 534-545.

[18] YOSHIOKA T, NAKATANI T. Generalization of multi-channel linear prediction methods for blind MIMO impulse response

shortening[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(10): 2707-2720.
[19] NAKATANI T, JUANG B H, YOSHIOKA T, et al. Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(8): 1512-1527.
[20] NAKATANI T, YOSHIOKA T, KINOSHITA K, et al. Speech dereverberation based on variance-normalized delayed linear prediction[J]. IEEE Transactions on Audio Speech, and Language Processing, 2010, 18(7): 1717-1731.
[21] NAKATANI T, JUANG B H, YOSHIOKA T, et al. Importance of energy and spectral features in Gaussian source model for speech dereverberation[C]//Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing. New Paltz, NY, USA: IEEE, 2007: 299-302.
[22] JUKIC A, VAN WATERSCHOOT T, GERKMANN T, et al. Multi-channel linear prediction-based speech dereverberation with sparse priors[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(9): 1509-1520.
[23] WITKOWSKI M, KOWALCZYK K. Split Bregman approach to linear prediction based dereverberation with enforced speech sparsity[J]. IEEE Signal Processing Letters, 2021, 28: 942-946.
[24] CHETUPALLI S, SREENIVAS T. Late reverberation cancellation using Bayesian estimation of multi-channel linear predictors and student's t-source prior[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(6): 1007-1018.
[25] JUKIC A, MOHAMMADIHA N, WATERSCHOOT T V, et al. Multi-channel linear prediction-based speech dereverberation with low-rank power spectrogram approximation[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. South Brisbane, QLD, Australia: IEEE, 2015: 96-100.
[26] LEE D D, SEUNG H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755): 788-791.
[27] AVARGEL Y, COHEN I. System identification in the short-time Fourier transform domain with crossband filtering[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(4): 1305-1319.
[28] 刘杨, 杨飞然, 梁兆杰, 等. 基于卡尔曼滤波的STFT域回声抵消算法[J]. 声学技术, 2022, 41(5): 757-762.
LIU Yang, YANG Feiran, LIANG Zhaojie, et al. Kalman filter based acoustic echo cancellation in the STFT domain[J]. Technical Acoustics, 2022, 41(5): 757-762.
[29] LOHMANN T, WATERSCHOOT T, BITZER J, et al. Dereverberation in acoustic sensor networks using weighted prediction error with microphone-dependent prediction delays[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: IEEE, 2023: 1-5.
[30] ROSENBAUM T, COHEN I, WINEBRAND E. Crossband filtering for weighted prediction error-based speech dereverberation[J]. Applied Sciences, 2023, 13(7): 9537.
[31] SUN Y, BABU P, PALOMAR D P. Majorization-minimization algorithms in signal processing, communications, and machine learning[J]. IEEE Transactions on Signal Processing, 2016, 65(3): 794-816.
[32] CVETKOVSKI Z. Inequalities: Theorems, techniques and selected problems[M]. Berlin Heidelberg: Springer, 2012: 74-75.
[33] 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.
[34] JULIUS R, SIMON W, JEAN-MARIE L, et al. Speech enhancement and dereverberation with diffusion-based generative models[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2351-2364.

作者简介:

康瑶(1995-),女,硕士,研究实习员,研究方向:统计信号处理、雷达信号处理、音频信号处理,E-mail: kangyao@ouchn.edu.cn。

康坊(1993-),女,博士,助理教授,研究方向:音频信号处理,E-mail:fang.s.kang@gmail.com。

杨飞然(1982-),**通信作者**,男,博士,研究员,研究方向:自适应滤波、传声器阵列信号处理和声场重建等,E-mail: feiran@mail.ioa.ac.cn。

(编辑:张黄群)