View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by INRIA a CCSD electronic archive server

Blind MultiChannel Identification and Equalization for Dereverberation and Noise Reduction based on Convolutive Transfer Function

Xiaofei Li, Sharon Gannot, Radu Horaud

▶ **To cite this version:**

Xiaofei Li, Sharon Gannot, Radu Horaud. Blind MultiChannel Identification and Equalization for Dereverberation and Noise Reduction based on Convolutive Transfer Function. 2017. hal-01568835

HAL Id: hal-01568835

https://hal.inria.fr/hal-01568835

Preprint submitted on 23 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Blind MultiChannel Identification and Equalization for Dereverberation and Noise Reduction based on Convolutive Transfer Function

Xiaofei Li, Sharon Gannot and Radu Horaud

***Abstract*—This paper addresses the problems of blind channel identification and multichannel equalization for speech dereverberation and noise reduction. The time-domain cross-relation method is not suitable for blind room impulse response identification, due to the near-common zeros of the long impulse responses. We extend the cross-relation method to the short-time Fourier transform (STFT) domain, in which the time-domain impulse responses are approximately represented by the convolutive transfer functions (CTFs) with much less coefficients. The CTFs suffer from the common zeros caused by the oversampled STFT. We propose to identify CTFs based on the STFT with the oversampled signals and the critical sampled CTFs, which is a good compromise between the frequency aliasing of the signals and the common zeros problem of CTFs. In addition, a normalization of the CTFs is proposed to remove the gain ambiguity across sub-bands. In the STFT domain, the identified CTFs is used for multichannel equalization, in which the sparsity of speech signals is exploited. We propose to perform inverse filtering by minimizing the ℓ_1-norm of the source signal with the relaxed ℓ_2-norm fitting error between the micophone signals and the convolution of the estimated source signal and the CTFs used as a constraint. This method is advantageous in that the noise can be reduced by relaxing the ℓ_2-norm to a tolerance corresponding to the noise power, and the tolerance can be automatically set. The experiments confirm the efficiency of the proposed method even under conditions with high reverberation levels and intense noise.**

I. Introduction

The goal of speech dereverberation is to remove the reverberations from the received microphone signals to improve the speech intelligibility for both human listening and machine recognition. The output of a dereverberation system could be an estimation of either the anechoic source signal or the source signal with some early reflections, since early reflections do not deteriorate the speech intelligibility [1]. Multi-channel dereverberation methods are generally more powerful than single-channel methods, due to the use of the extra spatial information and the relation across channels.

Multichannel dereverberation broadly includes the following techniques: (i) the spectral enhancement technique removes reverberation by spectral subtraction. Many techniques have been proposed to estimate the power spectral density of reverberation, such as statistical model [2], coherent-to-diffuse power ratio [3]; (ii) the linear prediction technique is used in the weighted prediction error (WPE) algorithm [4], [5], [**?**]: reverberation is first estimated by filtering the microphone signal with the prediction filter, and then subtracted from the microphone signal (in a way, the prediction filter could be considered as the inverse of the channel filter, which is estimated by maximizing a likelihood function derived from a *time-varying all-pole* model of source spectra); (iii) probabilistic techniques apply dereverberation by maximizing the likelihood of the generative model of the microphone signals, such as in [6], [7], and (iv) the multichannel equalization technique first blindly identifies the channel filters, then applies the inverse filtering on the microphone signals.

The focus of this paper is a multichannel equalization technique. For a single-input multiple-output (SIMO) system, the blind channel identification can be performed based on the second-order statistics, such as the subspace method [8] or the cross-relation method [9]. The cross-relation method identifies the channel filters by detecting the eigenvector corresponding to the unique zero eigenvalue of the covariance matrix of microphone signals. A noise subspace method proposed in [10] exploited the multiple eigenvectors corresponding to the zero eigenvalues of the over-modeled covariance matrix. These noise-subspace based methods, especially the cross-relation method, are vulnerable to noise interference and the filter length determination error. Thence they require prior knowledge of the exact filter length. However, in acoustic dereverberation, the room impulse response (RIR) is a time sequence with the variance exponentially asymptotically approaching to zero. The filter length is not measureable due to the long tails, in other words, the truncation point is difficult to determine. For the case of small tails, [11] proposed the channel under-modeling method that only considers the significant part of the filters, and a rank detection method was proposed in [12] to determine the length of the channel filter corresponding to the significant part. However, these methods are applicable only when a noticeable gap exists between the significant part and the small tails, which is obviously not the case of RIR. Based on the least mean squares method, the frequency-domain adaptive cross-relation method was proposed in [13], and was applied to speech separation and dereverberation in [14]. One of the identifiability conditions of the second-order statistics-based method is that the multiple channels are co-prime, namely they do not share any common zeros. It is shown in

X. Li and R. Horaud are with INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France. E-mail: `first.last@inria.fr`

Sharon Gannot is with Bar Ilan University, Faculty of Engineering, Israel. E-mail: `Sharon.Gannot@biu.ac.il`

This work was supported by the ERC Advanced Grant VHIA #340113.

[15] that a large number of near-common zeros exists for the long channel filters, which will deteriorate the performance of channel identification. A forced spectral diversity algorithm [16], [17] was proposed to mitigate the near-common zeros problem. In this method, the distance between the zeros of the channels is enlarged by filtering one channel with the spectral shaping filters, and then under-modeling this channel. Overall, the channel identification techniques mentioned above are hardly applicable to real-world RIR identification due to the filter length indetermination and the near-common zeros problem. As a result, experiments based on these methods are only carried out with channel filters that (i) do not have small tails, either the simulated filters or the RIRs with the tails being cut off, and (ii) are very short, e.g. from several taps to a few hundreds taps, which is not the case of RIRs that usually have thousands of taps.

For multichannel equalization using the known channel filters, ideally, the classical multiple-input/output inverse theorem (MINT) method can perfectly recover the source signal. In MINT, the inverse filters of the channel filters are obtained by using the least square method. However, MINT is sensitive to the filter perturbations and the additive noise in the microphone signals. To improve the robustness of MINT to the filter perturbations, many techniques have been proposed by preserving not only the direct-path impulse response but also the early reflections, such as channel shortening [18], infinity- and p-norm optimization-based channel shortening/reshaping [19], partial MINT [20], and relaxed multichannel least squares [21]. To improve the robustness of MINT to the noise interference, a filter energy regularization was adopted in [20], [22], since an inverse filter with small energy would avoid the amplification of noise. To jointly perform dereverberation and noise reduction, the regularization of the output noise power was proposed in [23] in the framework of multichannel equalization. The techniques mentioned above first estimate an inverse filter, being independent to the signals, and then apply the inverse filtering to the microphone signals.

This work aims to develop a blind multichannel equalization method for speech dereverberation in realistic scenarios. We propose to perform blind channel identification and multichannel equalization in the short-time Fourier transform (STFT) domain. The channel filters in each frequency band are much shorter than the time-domain filters, consequently they have much less near-common zeros. Moveover, the sparsity of speech signal in the STFT-domain is desirable for recovering the source signal, especially for the noisy case. It is known that STFT can be interpreted as the multi-rate filter banks [24]. The adaptive filtering in sub-bands has been widely studied [25], [26], and applied to acoustic echo cancellation. While the sub-band blind channel identification has been rarely studied. In [10], the noise subspace method was executed in sub-bands, however it was not applied to real scenarios. The channel filters are identified to a different gain factor for different sub-bands, which causes the gain ambiguity [10]. A gain ambiguity correction method was proposed in [27] based on the overlapping pass-band regions. The sparsity of speech in the STFT domain has been exploited in [28] by adding the signal sparsity function to the MINT function of time-domain filter as a regularization. However, the MINT function is heterogeneous with the regularization term, which makes it difficult to automatically set the regularization factor.

This paper proposes a blind convolutive transfer function (CTF) identification method, and a CTF-based inverse filtering method for dereverberation and noise reduction. In the STFT domain, the time-domain channel filters can be represented as the cross-band filter [29]. To simplify the analysis in practice, we consider the use of the CTF approximation, i.e. only the band-to-band filter, as used in [30]. Regarding the CTF indentification, this paper has the following contributions. First, the influence of the STFT configuration on the signal reconstruction, CTF approximation and common zeros issue are analyzed in II-B in detail. Briefly stated, in each frequency band, the cross-relation method [9] is extended for CTF identification, which is not trivial. CTF only takes the band-to-band filter, while it disregards the cross-band filters, thus suffers from an under-modeling error. The frequency response of the CTF will not be fully activated, i.e. zeros exist, for the oversampling case, namely the frame step is less than the window length. The sub-band filter is common to all channels, thus the zeros are also common to all channels, which is problematic for the cross-relation method. Conversely, critical sampling will lead to the severe frequency aliasing, and thus to poor signal reconstruction. Second, to achieve a good compromise, we propose a novel configuration, briefly, the signals are oversampled to avoid the frequency aliasing, and the channel filters are forced to be criticaly sampled to avoid the common zeros problem. Third, beyond exploiting the eigenvector corresponding to the unique zero eigenvalue, we propose to estimate the channel filters by solving a constrained least-squares problem, which is robust to the noise interference and the filter length determination error. Fourth, to remove the gain ambiguity accross sub-bands, we propose to use the channel filters normalized by the first coefficient of the first channel, as a result, the time-domain equalized signal is the original source signal filtered by the early (corresponding to the length of STFT window) impulse response of the first channel. In other words, the dereverberated signal preserves the early reflections.

As for the inverse filtering method based on identified CTFs, this paper has the following contributions. First, an optimization problem is proposed for both dereverberation and noise reduction. In detail, we propose to recover the source signal by exploiting the ℓ_2 fit between the microphone signals and the channel filtered source signal, rather than explicitly estimating the inverse filters. The sparsity of the estimated signal is exploited by minimizing its ℓ_1 norm. To reduce the noise caused by the filter perturbations and the microphone noise interference, the ℓ_2 fitting error is relaxed to a tolerance corresponding to the noise power. Overall, the source signal is obtained by solving an ℓ_1 minimization problem with an ℓ_2 fit constraint. Second, the convex optimization algorithm primal-dual interior-point method [31] is adopted to solve the optimization problem. The proposed inverse filtering algorithm is advantageous in that the sparsity promotion is efficient for noise reduction, and the tolerance for the ℓ_2 fitting error can be automatically set according to the noise power, while the regularization term in [22], [23], [28] has to be empirically set even if the noise power is known. Finally, the time-domain signal is obtained by inverse STFT.

The remainder of this paper is organized as follows. The blind channel identification in the STFT-domain is given in Section II. The inverse filtering method is presented for dereverberation and noise reduction in Section III. Experiments with binaural simulation data and with multichannel real recordings are presented in Section IV and V, respectively. Section VI concludes the work.

II. STFT-domain Channel Identification

We consider a two channel system. In the time domain, the noise-free microphone signals $x(n)$ and $y(n)$ are

$$x(n) = s(n) \star a(n), \quad y(n) = s(n) \star b(n), \tag{1}$$

where $\star$ denotes convolution, $s(n)$ is a non-stationary source signal, e.g., speech, and $a(n)$ and $b(n)$ are the RIRs. The cross-relation method [9] is not practical for the time-domain RIR identification due to the near common zeros problem and the length of the filters. In this section, we propose a STFT-domain channel filter identification algorithm.

A. Problem Formulation in the STFT Domain

The STFT representation of the signal $x(n)$ is

$$x_{p,k} = \sum_{n=-\infty}^{+\infty} x(n)\tilde{w}(n-pL)e^{-j\frac{2\pi}{N}k(n-pL)}, \tag{2}$$

where $p = 0, \dots, P-1$ and $k = 0, \dots, N-1$ denote the frame index and the frequency index, respectively, $\tilde{w}(n)$ denotes an analysis window, and N and L denote the frame (window) length, and the frame step, respectively. Equation (2) gives the overlap-add view of STFT. In the filter bank interpretation, the analysis window is considered as the low-pass filter, and L as the decimation factor.

The cross-band filter model consists in representing the STFT coefficient $x_{p,k}$ as a summation over multiple convolutions (between the STFT-domain source signal and filter) across frequency bins. Mathmatically, the linear time invariant system (1) can be written in the STFT domain as

$$x_{p,k} = \sum_{k'=0}^{N-1} \sum_{p'=-C}^{Q-1} s_{p-p',k'} \, a_{p',k,k'}, \tag{3}$$

and note that there is an equivalent expression for $y_{p,k}$. From [29], if $L < N$, then $a_{p',k,k'}$ is non-causal, with $C = \lceil N/L \rceil - 1$ non-causal coefficients. The number of causal filter coefficients Q is related to the reverberation time. Let $w(n)$ denote the STFT synthesis window. The STFT-domain impulse response $a_{p',k,k'}$ is related to the time-domain impulse response $a(n)$ by:

$$a_{p',k,k'} = (a(n) \star \zeta_{k,k'}(n))|_{n=p'L}, \tag{4}$$

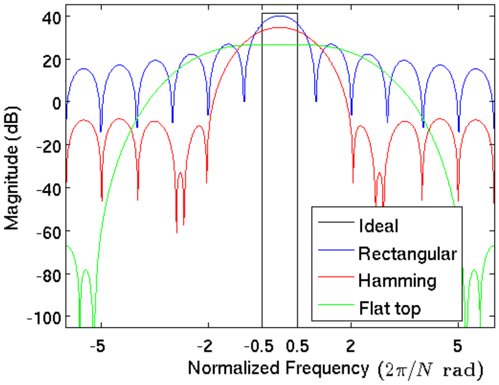


Fig. 1: The frequency response of STFT windows.

which represents the convolution with respect to the time index n evaluated at frame steps, with

$$\zeta_{k,k'}(n) = e^{j\frac{2\pi}{N}k'n} \sum_{m=-\infty}^{+\infty} \tilde{w}(m)\, w(n+m)\, e^{-j\frac{2\pi}{N}m(k-k')}. \tag{5}$$

To simplify the analysis, we consider the convolutive transfer function (CTF) approximation, i.e., only band-to-band filters with $k = k'$ are considered:

$$x_{p,k} \approx \sum\nolimits_{p'=-C}^{Q-1} s_{p-p',k} a_{p',k} = s_{p,k} \star a_{p,k}, \tag{6}$$

and similarly $y_{p,k} \approx s_{p,k} \star b_{p,k}$.

To identify the filters $a_{p,k}$ and $b_{p,k}$, the cross-relation between the two channels

$$x_{p,k} \star b_{p,k} = s_{p,k} \star a_{p,k} \star b_{p,k} = y_{p,k} \star a_{p,k} \tag{7}$$

can be used. This relation was originally proposed for the time-domain filter identification and here is extended to the CTF domain. The conditions that this identification problem has a unique solution are given in [9], namely that (i) the linear complexity of the source signal $s_{p,k}$ should be sufficiently large to fully excite the filters, and that (ii) the two filters $a_{p,k}$ and $b_{p,k}$ are co-prime, i.e. they do not share any common zeros. The first condition can be satisfied by increasing the length of source signal $s(n)$. The second condition is related to the configuration of the STFT. Prior to the detailed filter identification algorithm, below we analyze the influence of the STFT configuration on the signal reconstruction, CTF approximation and the common zeros issue.

B. Analysis of STFT Configuration

The filter bank interpretation of the STFT is that the time domain signal is first modulated (frequency shifted) by $e^{-j\frac{2\pi}{N}kn}$, then low-pass filtered with the analysis window $\tilde{w}(n)$, and down-sampled by L. Let $\tilde{W}(\omega)$ denote the frequency responses of $\tilde{w}(n)$, where ω denotes the angular frequency. Fig. 1 shows the frequency responses of three typical windows, i.e. rectangular, Hamming and flat-top windows. In addition, the ideal low-pass filter that has the bandwidth of $2\pi/N$ is shown with a black rectangle. The down sampling operation will fold the frequency response with the cycle of $2\pi/L$. For the ideal filter, there is no frequency aliasing up to the critical down-sampling, i.e. $L = N$. While for the three practical windows, frequency aliasing appears when $L > 1$. Although the frequency aliasing can be cancelled by the synthesis windows as long as the completeness condition is met, we still expect the aliasing to be as small as possible to guarantee a high robustness for spectral modifications. The dereverberation algorithm intends to recover the source signal $s_{p,k}$ from the microphone signals $x_{p,k}$ and $y_{p,k}$. In other words, the microphone signals are modified to a large extent, which will disturb the aliasing cancellation. To this end, we should use the window with a high side-lobe attenuation, such as the Hamming or flat-top windows. In addition, the cutoff frequency of the lowpass filter should not be larger than the folding frequency π/L. Here we consider the cutoff frequency to be at the first zero-crossing, namely the main lobe. Taking the Hamming window as an example, one can see from Fig. 1 that the first zero-crossing is at about $4\pi/N$, as a result, the decimation factor should be set as $L \leq N/4$. We denote with L_{max} the maximum decimation factor that guarantees a small frequency aliasing, and $L_{max} = N/4$ for the Hamming window.

To illustrate the reliability of the CTF approximation, we analyze the significance of the cross-band filters for various windows. According to (4), the undecimated STFT-domain filter $\tilde{a}_{n,k,k'}$ is obtained by setting $L = 1$. Applying the discrete-time Fourier transform to $\tilde{a}_{n,k,k'}$ with respect to the time index n, the frequency response is given as, [29]:

$$\tilde{A}_{k,k'}(\omega) = A(\omega)\tilde{W}\left(\omega - \frac{2\pi}{N}k\right)W\left(\omega - \frac{2\pi}{N}k'\right) \qquad (8)$$

where $A(\omega)$ and $W(\omega)$ are the frequency responses of $a(n)$, $w(n)$, respectively. Without loss of generality, the synthesis window is assumed to possess the same frequency response as the analysis window. The product of $\tilde{W}(\omega - \frac{2\pi}{N}k)$ and $W(\omega - \frac{2\pi}{N}k')$ indicates the power of the cross-band filters. For the ideal filter, the cross band filters $\tilde{A}_{k,k'}(\omega)$ with $k' \neq k$ are equal to zero. In other words, the band-to-band filter $\tilde{A}_{k,k}(\omega)$ is a perfect STFT representation of the filter $a(n)$. However, for the three practical windows, it can be seen from Fig. 1 that $\tilde{A}_{k,k'}(\omega)$ are not zero for $k' \neq k$. As a result, the band-to-band filter, i.e. CTF, can only approximately represent the filter $a(n)$ in the STFT domain. The approximation error depends on the power of the cross-bands filters with $k' \neq k$ relative to the power of the band-to-band filter. Consider a specific frequency band $\omega = 0$, it can be seen from Fig. 1 that the rectangular window has the least approximation error, since the overlap between $\tilde{W}(0)$ and $W(\frac{2\pi}{N}k')|_{k'\neq 0}$ is the smallest among the three windows. The flat-top window has the largest approximation error.

The decimation will fold the frequency response, as for the band-to-band filter $a_{p,k}$ with the decimation factor L, the frequency response is

$$\tilde{A}_{k,k}(\omega)_{\downarrow L} = \frac{1}{L}\sum_{l=0}^{L-1}\tilde{A}_{k,k}\left(\frac{\omega}{L} - \frac{2\pi}{L}l\right), \qquad (9)$$

where $\tilde{A}_{k,k}(\omega)$ is defined in (8) with $k' = k$. To simplify the analysis, the magnitude of the side lobes is assumed to be zero. The decimation factor is set to be less than the maximum decimation factor L_{max} to guarantee a small frequency aliasing. Without loss of generality, we consider the case that N/L is an integer, and the frequency band $\bar{k}$ being the integer multiple of N/L. Then (9) is simplified as

$$\tilde{A}_{\bar{k},\bar{k}}(\omega)_{\downarrow L} = \frac{1}{L}A\left(\frac{\omega}{L} - \frac{2\pi}{N}\bar{k}\right)\tilde{W}\left(\frac{\omega}{L}\right)W\left(\frac{\omega}{L}\right). \qquad (10)$$

The filter $\tilde{W}(\frac{\omega}{L})$ involves only the mainlobe if $L = L_{max}$, and involves some sidelobes if $L < L_{max}$. For the ideal filter, the mainlobe is fully excited, and thence the filter $\tilde{W}(\frac{\omega}{L})$ is fully excited when $L = L_{max} = N$. However, for the three practical windows, the magnitude of $\tilde{W}(\frac{\omega}{L})$ is dramatically decreasing along with the increase of frequency, even in the mainlobe. Thence, the magnitudes of $\tilde{W}(\frac{\omega}{L})W(\frac{\omega}{L})$ are close to zero in the high frequency region. Note that for the case that N/L is not an integer, and/or the frequency band being not the integer multiple of N/L, the zero-magnitude region also exsits possibly in a different frequency region. This zero-magnitude region is caused by the filter banks, and thence is also present in the frequency response of filter $b_{p,k}$. This common zeros of two channels are problematic for the cross-relation method (7). It can be seen from Fig. 1 that the flat-top window has the least common zeros among the three windows, since its mainlobe is excited the most.

To summarize, extending the time-domain cross-relation method to the STFT domain is not a trivial task. It suffers from the problems of frequency aliasing, CTF approximation and common zeros. The windows (low-pass filters) and the frame step (decimation factor) are crucial for circumventing these problems. Briefly, the frequency aliasing can be suppressed by using a window with a high side lobe attenuation, and a small frame step. A small CTF approximation error requires the window to have a narrow main lobe. To avoid the common zeros, the window is better to have a rectangular main lobe. Otherwise, a larger frame step (even critical sampling) is needed to fully excite the CTF filter. Unfortunately, these requirements can only be fulfilled by the ideal filter with critical sampling. For the practical filters, as discussed above, one specific STFT configuration can only fit parts of the conditions.

In this work, to achieve a good compromise between these requirements, we propose a novel STFT configuration. The Hamming window with frame step $L = N/4$ is adopt for the STFT of the signals, which has the negligible frequency aliasing and acceptable CTF approximation error (slightly larger than the approximation error of the rectangular window). In addition, to avoid the common zeros problem, the STFT-domain filters $a_{p,k}$ and $b_{p,k}$ are forced to take the critical sampling, i.e. we set $L_f = N$, where L_f denotes the frame step of the STFT-domain filters. The details will be presented in the next subsection. As for the window length, a relatively large one could be selected, e.g. 64 ms in this work. A large window size leads to a small number of channel filter taps, which is beneficial to channel identification. In addition, the spectral differences between the neighboring frequency bands get smaller with the increasing of the frequency resolution (window length), which will mitigate the cross-bands effect.

C. Channel Identification

Note that since the channel identification algorithm is applied frequency-wise, hereafter the frequency index k is omitted unless necessary. In (6) and (7), the frame index p ($p \in [0, P-1]$ for the signals s_p, x_p and y_p, and $p \in [-C, Q]$ for the filters a_p and b_p) corresponds to the frame step $L = N/4$. Denote the filters in vector form as $\mathbf{a}$ and $\mathbf{b}$. To avoid the common zeros problem, we further down sample the filters by a factor 4. The down sampled filters $a_{p\downarrow 4}$ and $b_{p\downarrow 4}$ correspond to the critical sampling, which have larger frequency aliasing than the original filters. There will be no non-causal coefficients anymore for the case of critical sampling. The down sampled filters start with the zeroth tap, and have a length of $\tilde{Q} = \lceil Q/4 \rceil$. They are written in vector form as

$$\begin{aligned} \tilde{\mathbf{a}} &= [a_0, a_4, \dots, a_{4(\tilde{Q}-1)}]^\top, \\ \tilde{\mathbf{b}} &= [b_0, b_4, \dots, b_{4(\tilde{Q}-1)}]^\top. \end{aligned} \tag{11}$$

where $^\top$ is the transpose operator.

Define the matrix $\mathbf{X}$ from the signal x_p as

$$\mathbf{X} = \begin{bmatrix} x_0 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ x_p & x_{p-4} & \ddots & \ddots & x_{p-4(\tilde{Q}-1)} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ x_{P-1} & x_{P-5} & \cdots & \cdots & x_{p-1-4(\tilde{Q}-1)} \end{bmatrix} \tag{12}$$

with the size of $P \times \tilde{Q}$. The row number is the frame number of the oversampled signal, while the column number is the length of the critical sampled filter. This indicates that $\mathbf{X}\tilde{\mathbf{b}}$ is the convolution between the oversampled signal and the critical sampled filter. The matrix $\mathbf{Y}$ is defined from y_p following the same principle. Then (7) can be rewritten as $\mathbf{X}\tilde{\mathbf{b}} = \mathbf{Y}\tilde{\mathbf{a}}$, which is equivalent to

$$\mathbf{Z}\mathbf{c} = 0 \tag{13}$$

where $\mathbf{Z} = [\mathbf{Y}, -\mathbf{X}]$ and $\tilde{\mathbf{c}} = [\tilde{\mathbf{a}}^\top, \tilde{\mathbf{b}}^\top]^\top$. In [9], [13], [17], the filter vector $\tilde{\mathbf{c}}$ is estimated by taking an eigenvector of $\mathbf{Z}$ corresponding to a zero eigenvalue. This method is available only for the case that the filter length is exactly known. However, in practice, the length of the RIR cannot be exactly measured. In addition, the one-dimensional null space of $\mathbf{Z}$ could be easily contaminated even by a mild noise interference. As a result, this eigenvector method is hardly applicable for RIR identification.

Instead of the eigenvector method, we estimate the filter vector $\tilde{\mathbf{c}}$ by solving the following least-square problem

$$\min \parallel \mathbf{Z}\tilde{\mathbf{c}} \parallel^2, \quad \text{s.t. } \mathbf{g}^\top \tilde{\mathbf{c}} = 1, \tag{14}$$

where $\mathbf{g}$ is a constant vector to constrain $\tilde{\mathbf{c}}$. Empirically, $\mathbf{g}$ is set as

$$\mathbf{g} = [1, \underbrace{0, \dots, 0}_{\tilde{Q}-1}, 1, \underbrace{0, \dots, 0}_{\tilde{Q}-1}]^\top. \tag{15}$$

Here we constrain the sum of the first entries of the two filters to 1. The constraint can be written as $\mathbf{g}^\top \tilde{\mathbf{c}} = a_0 + b_0 = 1$. Constraining the scale of the first entries will remove the delay ambiguity, namely the direct-path responses is enforced to start with the first entries. Moveover, the first entries of the RIRs usually have a larger amplitude than the others in $\tilde{\mathbf{c}}$, and the minimization of the objective function tends to suppress the scale of the unconstrained entries. Thence, constraining the scale of the first entries may promote a reasonable estimation of $\tilde{\mathbf{c}}$. The solution of (14) is

$$\hat{\mathbf{c}} = \frac{(\mathbf{Z}^H \mathbf{Z})^{-1} \mathbf{g}}{\mathbf{g}^\top (\mathbf{Z}^H \mathbf{Z})^{-1} \mathbf{g}} \tag{16}$$

where ^{H} denotes complex transpose. Consequently, the estimations of $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$ are respectively $\hat{\mathbf{a}} = \hat{\mathbf{c}}_{1:\tilde{Q}}$ and $\hat{\mathbf{b}} = \hat{\mathbf{c}}_{\tilde{Q}+1:2\tilde{Q}}$. This channel filter identification method can be extended to the multichannel case. The filter vector $\tilde{\mathbf{c}}$ would then be the concatenation of all the channels, and the signal matrix $\mathbf{Z}$ can be organized as proposed in [9], [10] that concatenates the signal matrices of each microphone pair. Correspondingly, in the constraint vector $\mathbf{g}$, we set the entries corresponding to the first entries of each channel to 1, and the others to 0. The solution is still (16).

It is obvious that $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ are the estimations of a normalized version of $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$, and the normalization factor is $a_0 + b_0$. However, the normalization factor varies along the frequency bands, which leads to the gain ambiguity. To remove the gain ambiguity we propose to further normalize both the filters by the first entry of one of the filters, e.g. $\hat{a}_0$. Formally, the normalized filters are defined as $\bar{\mathbf{a}} = \hat{\mathbf{a}}/\hat{a}_0$ and $\bar{\mathbf{b}} = \hat{\mathbf{b}}/\hat{a}_0$, where $\bar{a}_0 = 1$. Obviously, $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ are the estimations of $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$ normalized by a_0. In the frequency band k, the source signal corresponding to $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$ is $s_{p,k}$, thence the one corresponding to $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ is $a_{0,k} s_{p,k}$. From (4), $a_{0,k}$ can be represented as

$$a_{0,k} = \sum_{n=0}^{N-1} a(n) \nu(n) e^{-j \frac{2\pi}{N} kn}, \tag{17}$$

where $\nu(n) = \sum_m \tilde{w}(m) w(m-n)$ is a window function. Therefore, $a_{0,k}|_{k=0}^{N-1}$ can be interpreted as the Fourier transform of the impulse response segment $a(n)|_{n=0}^{N-1}$ windowed by $\nu(n)$. As a result, the time-domain signal corresponding to $a_{0,k} s_{p,k}$ will be the convolution between $s(n)$ and $a(n)|_{n=0}^{N-1}$. The gain ambiguity is removed by consistently normalizing the filters.

III. Inverse Filtering for dereverberation

In this section we propose to estimate the source signals in the STFT domain by inverse filtering. In addition, to exploiting the spectral sparsity of the source signal in the STFT-domain, the ℓ_1-norm minimization is proposed to suppress the noise caused by the filter perturbations and the ambient noise interference.

A. Inverse Filtering

Let us still consider the two channel case. The filter $\bar{\mathbf{a}}$ is an estimation of the critical sampled and frequency aliased filter $\tilde{\mathbf{a}}$, from which we could reconstruct an estimation of the original filter $\mathbf{a}$ by interpolation. The *first-order hold* interpolation,[1] namely linear interpolation, is applied. Note that the perfect reconstruction cannot be achieved due to the frequency aliasing of $\bar{\mathbf{a}}$. From the interpolated filter $\mathbf{a}$, we construct the filter matrix $\mathbf{A}$ as

$$\mathbf{A} = \begin{bmatrix} a_0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ a_1 & a_0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ a_{Q-1} & a_{Q-2} & \ddots & a_0 & 0 & \ddots & 0 \\ 0 & a_{Q-1} & a_{Q-2} & \ddots & a_0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{Q-1} & a_{Q-2} & \cdots & a_0 \end{bmatrix}$$

of size of $P \times P$. In $\mathbf{A}$, the transpose of the filter $\mathbf{a}$ is duplicated as the row vectors, with one element shift per row. The filter matrix $\mathbf{B}$ is defined from $\mathbf{b}$ following the same principle. Then we concatenate the two matrices to yield $\mathbf{C} = [\mathbf{A}^\top, \mathbf{B}^\top]^\top$.

We define the microphone signals x_p and y_p in vector form as $\mathbf{x} = [x_0, \cdots, x_{P-1}]^\top$ and $\mathbf{y} = [y_0, \cdots, y_{P-1}]^\top$, respectively. In the previous section we assumed that the microphone signals were noise free. In this section, we explicitly introduce additive noise to the microphone signals as $\mathbf{x} = \mathbf{x}_c + \mathbf{e}_x$, where $\mathbf{x}_c$ and $\mathbf{e}_x$ denote the noise-free signal and noise, respectively. The noise-free signal and noise are assumed to be uncorrelated. The noise signal obeys an i.i.d. complex Gaussian distribution. Similarly, define $\mathbf{y} = \mathbf{y}_c + \mathbf{e}_y$. Then we concatenate them as $\mathbf{z}_c = [\mathbf{x}_c^\top, \mathbf{y}_c^\top]^\top$, $\mathbf{e} = [\mathbf{e}_x^\top, \mathbf{e}_y^\top]^\top$ and $\mathbf{z} = [\mathbf{x}^\top, \mathbf{y}^\top]^\top = \mathbf{z}_c + \mathbf{e}$. Note that the proposed channel identification algorithm was directly applied to the noisy case.

We define the source signal s_p in vector form as $\mathbf{s} = [s_0, \cdots, s_{P-1}]^\top$. The complex source signal can be estimated by minimizing the ℓ_2-norm $\parallel \mathbf{Cs} - \mathbf{z} \parallel$. Due to the CTF approximation error and the frequency aliasing of the critical sampled filters, the filter estimations are inaccurate. We found that the amplitude of the filters is reliable, while the phase information is very inaccurate. Therefore, we only use the amplitude of the filters to recover the amplitude of the source signal. We define $\tilde{\mathbf{z}} = |\mathbf{z}|$ and $\tilde{\mathbf{C}} = |\mathbf{C}|$, where $|\cdot|$ denotes the entry-wise absolute value of a matrix or vector. The amplitude filtering is defined as $\tilde{\mathbf{C}}\mathbf{s}|_{\mathbf{s}\succeq\mathbf{0}}$, where $\succeq$ denotes entry-wise vector inequality and $\mathbf{0}$ is the P-dimensional vector with all entries equal to 0. The amplitude of the source signal can be recovered by minimizing the following cost function

$$\bar{\mathbf{s}} = \underset{\mathbf{s},\ \text{s.t. } \mathbf{s}\succeq\mathbf{0}}{\operatorname{argmin}} \parallel \tilde{\mathbf{C}}\mathbf{s} - \tilde{\mathbf{z}} \parallel^2 . \tag{18}$$

For the ideal noise-free case with the well-estimated filter $\mathbf{C}$ and the true source signal $\mathbf{s}$, the relation $\mathbf{Cs} \approx \mathbf{z}$ holds. Based on the triangle inequality theorem, we have $|\mathbf{C}||\mathbf{s}| \geq |\mathbf{z}|$. Thence, the solution of (18) would be an underestimation of $|\mathbf{s}|$. The underestimated source singal is smaller than the proper estimation, thus it will possibly suffer from some distortions, but it will not include more reverberations than the proper estimation. In addition, as mentioned in Section II-C, corresponding to the normalized filters, the source signal involves some early reflections, which provides more tolerance to the underestimation. Briefly stated, the lost information of the direct-path signal due to the underestimation could possibly be preserved in its early reflections, and vice versa. The experiments using both the estimated CTFs and the true CTFs computed by (4) demonstrate that the distortion of source signal is very low (for details see Section **??**).

In addition, the STFT-domain signal is sparse, which can be exploited by minimizing the ℓ_1 norm of the source signal to suppress the noise. For the positive amplitude vector $\mathbf{s}$, the ℓ_1 norm is $|\mathbf{s}|_1 = \mathbf{1}^\top \mathbf{s}, \mathbf{s} \succeq \mathbf{0}$, where $\mathbf{1}$ is the P-dimensional vector with all entries set to 1. The ℓ_1 minimization can be realized by adding the ℓ_1 regularization on (18), however the regularization factor is difficult to be automatically set. We realize the ℓ_1 minimization by solving the constrained optimization problem:

$$\begin{aligned} \tilde{\mathbf{s}} &= \underset{\mathbf{s}}{\operatorname{argmin}}\, \mathbf{1}^\top \mathbf{s} \\ \text{s.t.} \quad & \mathbf{s} \succeq \mathbf{0},\ \parallel \tilde{\mathbf{C}}\mathbf{s} - \tilde{\mathbf{z}} \parallel^2 \leq \delta. \end{aligned} \tag{19}$$

The ℓ_2 fit $\parallel \tilde{\mathbf{C}}\mathbf{s} - \tilde{\mathbf{z}} \parallel^2$ is relaxed to at most the tolerance δ.

The relaxing tolerance δ is related to the noise power in the microphone signal. Let σ_{ex}^2 and σ_{ey}^2 denote the noise PSD in the two channels, which can be estimated from the pure noise signal for stationary noise, or estimated by a noise PSD estimator for non-stationary noise, e.g. [32], [33]. The squared ℓ_2 norm of the noise signal $\parallel \mathbf{e}_x \parallel^2$ (or $\parallel \mathbf{e}_y \parallel^2$) follows an Erlang distribution with mean $P\sigma_{ex}^2$ (or $P\sigma_{ey}^2$) and variance $P\sigma_{ex}^4$ (or $P\sigma_{ey}^4$). Assume the noise signal in the two channels are uncorrelated, then $\parallel \mathbf{e} \parallel^2$ has the mean $P(\sigma_{ex}^2 + \sigma_{ey}^2)$ and variance $P(\sigma_{ex}^4 + \sigma_{ey}^4)$. To relax the ℓ_2 fit to the noise power, we set the noise relaxing term as

$$\delta_e = P(\sigma_{ex}^2 + \sigma_{ey}^2) - 2\sqrt{P(\sigma_{ex}^4 + \sigma_{ey}^4)}. \tag{20}$$

Subtracting two times the standard deviation makes the probability that the ℓ_2 fit being larger than $\parallel \mathbf{e} \parallel^2$ very small. When the ℓ_2 fit is allowed to be larger than $\parallel \mathbf{e} \parallel^2$, the minimization of $\mathbf{1}^\top \mathbf{s}$ will distort the source signal. Here we prefer less source signal distortion on more noise reduction. As a result, some noise remains in the estimated source signal.

Besides, the ℓ_2 fit should also be relaxed with respect to the filter perturbations, since the amplitude filtering $\tilde{\mathbf{C}}\mathbf{s}$ is not accurate to fit $|\mathbf{z}_c|$ by definition. The inaccuracy is akin to the level of the noise-free signal, i.e. $\lambda_c = \parallel \mathbf{z}_c \parallel^2$. It can be estimated by spectral subtraction as

$$\hat{\lambda}_c = \max(\parallel \mathbf{z} \parallel^2 - P(\sigma_{ex}^2 + \sigma_{ey}^2),\ 0). \tag{21}$$

Empirically, the relaxing term with respect to the noise signal is set to $\delta_c = 0.05\hat{\lambda}_c$.

[1] Some other interpolation methods have also been tested, such as *zero-order hold* interpolation, piecewise cubic spline interpolation. They achieve almost the same results.

TABLE I: The specifications of the optimization problems (18) and (19), where $\mathbf{I} \in \mathbb{R}^{P\times P}$ denotes the identity matrix, $\mathbf{0}_{P\times P} \in \mathbb{R}^{P\times P}$ denotes the matrix with all entries set to 0.

	$f_0(\mathbf{s})$	$f_i(\mathbf{s})$	λ	$\nabla f_0(\mathbf{s})$	$\nabla^2 f_0(\mathbf{s})$	$Df(\mathbf{s})$	$\nabla^2 f_i(\mathbf{s})$
(18)	$\parallel \tilde{\mathbf{C}}\mathbf{s} - \tilde{\mathbf{z}} \parallel^2$	$-s_i, i \in [1, P]$	$\lambda_i, i \in [1, P]$	$\tilde{\mathbf{C}}^\top(\tilde{\mathbf{C}}\mathbf{s} - \tilde{\mathbf{z}})$	$\tilde{\mathbf{C}}^\top\tilde{\mathbf{C}}$	$-\mathbf{I}$	$\mathbf{0}_{P\times P}, \forall i$
(19)	$\mathbf{1}^\top\mathbf{s}$	$-s_i, i \in [1, P]$; $\parallel \tilde{\mathbf{C}}\mathbf{s} - \tilde{\mathbf{z}} \parallel^2 -\delta, i = P+1$	$\lambda_i, i \in [1, P]$; λ_{P+1}	$\mathbf{1}$	$\mathbf{0}_{P\times P}$	$[-\mathbf{I}, \tilde{\mathbf{C}}^\top(\tilde{\mathbf{C}}\mathbf{s} - \tilde{\mathbf{z}})]^\top$	$\mathbf{0}_{P\times P}, i \in [1, P]$; $\tilde{\mathbf{C}}^\top\tilde{\mathbf{C}}$, $i = P+1$

The relaxing tolerance can be set by $\delta_e + \delta_c$. However, for this quantity, the ℓ_2 norm constraint in (19) is not definitely feasible, since both δ_e and δ_c are set to be relatively small to avoid the source signal distortion. The minimum ℓ_2 norm fitting error is defined in (18), i.e. $\parallel \tilde{\mathbf{C}}\bar{\mathbf{s}} - \tilde{\mathbf{z}} \parallel^2$, which is taken as the lower bound of the relaxing tolerance. Overall, the relaxing tolerance is set as

$$\delta = \max(\delta_e + \delta_c,\ 1.05 \parallel \tilde{\mathbf{C}}\bar{\mathbf{s}} - \tilde{\mathbf{z}} \parallel^2) \tag{22}$$

where 1.05 is a slack factor.

The extension of this inverse filtering model to the multichannel case is straightforward. The filter matrix $\mathbf{C}$ and signal vector $\mathbf{z}$ are constructed by concatenating the filter matrices and the microphone signal vectors of all the channels, respectively.

The optimal solution $\tilde{\mathbf{s}}$ is an estimation of the amplitude of the source signal. The elements in $\tilde{\mathbf{s}}$ are written with the frame and the frequency index as $\tilde{s}_{p,k}$. The phase of one of the microphone signals is taken as the phase of the estimated source signal, we have $\hat{s}_{p,k} = \tilde{s}_{p,k}e^{j\arg[x_{p,k}]}$, where $\arg[\cdot]$ is the phase of complex number. The time-domain source signal $\hat{s}(n)$ can be obtained Apply the inverse STFT to $\hat{s}_{p,k}$. As mentioned in Sec. II-C, based on the $a_{0,k}$-normalized filters, the output of the inverse filtering $\hat{s}_{p,k}$ is an estimation of $a_{0,k}s_{p,k}$. Consequently, the time-domain signal $\hat{s}(n)$ is an estimation of $s(n) \star a(n)|_{n=0}^{N-1}$, where $a(n)$ starts with the direct-path impulse response. The window size N is generally far larger than the duration of the direct-path impulse response, thus $\hat{s}(n)$ involves the early reflections, e.g. 64 ms in this work.

B. Convex Optimization

Both (18) and (19) are convex optimization problems with an inequality constraint. We adopt the primal-dual interior-point method (PDIPM) [31] to solve them. The book [31] provides a general optimization algorithm for a convex objective function $f_0(x)$ with a set of inequality constraints of the form $f_i(x) \leq 0, i \in [1, m]$ and an affine equality constraint. Here x denotes the optimization variable and m is the number of inequality constraints. Note that there is no affine equality constraint in the presented problems. Define the vector $f(x) = [f_1(x), \cdots, f_m(x)]^\top$ including all the inequality functions, and its derivative matrix $Df(x) = [\nabla f_1(x), \cdots, \nabla f_m(x)]^\top$, where ∇ denotes gradient operator. Let λ_i denote the dual variable corresponding to the inequality constraint $f_i(x) \leq 0$. The dual variable vector is $\lambda = [\lambda_1, \cdots, \lambda_m]$. In PDIPM, the inequality constraint is approximately formulated as an equality constraint by the logarithmic barrier function. The parameter t sets the accuracy of the logarithmic barrier approximation, the larger t, the better the approximation.

The PDIPM is summarized in Algorithm 1, with variable update in Step 4 given by:

$$\begin{bmatrix}\Delta x\\ \Delta\lambda\end{bmatrix} = -\begin{bmatrix}\nabla^2 f_0(x) + \sum_i \lambda_i \nabla^2 f_i(x) & Df(x)^\top \\ -\mathrm{diag}(\lambda)Df(x)^\top & -\mathrm{diag}(f(x))\end{bmatrix}^{-1} \times \begin{bmatrix}\nabla f_0(x) + Df(x)^\top \\ -\mathrm{diag}(\lambda)f(x) - (1/t)\mathbf{1}\end{bmatrix}. \tag{23}$$

In Algorithm 1, the so-called surrogate duality gap $\hat{\eta}^{(n)}$ is decreasing with the iterations, thence the parameter t is increased by the factor μ (a positive value of the order of 10). The goal of the line search (Step 3) is to find the largest step-length $\zeta^{(n)}$ under the conditions that (i) the updated dual variable λ is nonnegative, (ii) the inequality constraint is feasible with the updated variables, and (iii) the problem is optimized. We refer to [31] for more details. The convergence criterion is briefly set as (i) the surrogate duality gap $\hat{\eta}^{(n)}$ is less than a small tolerance ϵ to guarantee a high optimization, (ii) the dual residual $\nabla f_0(x) + Df(x)^\top$ is less than a small tolerance ϵ_{feas} to guarantee the feasibility of the variables.

To apply the PDIPM to the problems (18) and (19), the general quantities in Algorithm 1 should be accordingly specified. Table I gives the specifications for both (18) and (19). For solving (18), a good initialization is to set $\mathbf{s}^{(0)} = |\mathbf{x}|$ and $\lambda^{(0)}$ is set to an arbitrary positive vector ($10 \cdot \mathbf{1}_S$ in this paper). For solving (19), a feasible initialization is to set $\mathbf{s}^{(0)}$ and $\lambda^{(0)}$ as the solution of (18).

Algorithm 1 Primal-dual interior-point method

```
Iteration step $n = 0$. Initialize $f_i(x^{(0)}) \leq 0, \lambda_i^{(0)} \geq 0, \forall i$.
repeat
  1 Compute $\hat{\eta}^{(n)} = -f(x)^\top\lambda$,
  2 Set $t^{(n)} := \mu m/\hat{\eta}^{(n)}$,
  3 Line search the step-length $\zeta^{(n)}$,
  4 Update variables $x^{(n+1)} = x^{(n)} + \zeta^{(n)}\Delta x^{(n)}$,
    and $\lambda^{(n+1)} = \lambda^{(n)} + \zeta^{(n)}\Delta\lambda^{(n)}$.
until Convergence.
```

IV. Experiments with Simulated Binaural Data

In this section, we present a series of experiments with simulated binaural data. A set of BRIRs were generated with the ROOMSIM simulator [34] combined with the head-related impulse responses (HRIRs) of the KEMAR dummy head [35]. The simulated room is of dimension 5 m × 8 m × 3 m. The dummy head is located at (1 m, 4 m, 1.5 m).

TABLE II: The performance measures for the noise-free microphone signals as a function of filter length. The filter length is set up to approximately T_{60}. The 'unproc.' column is the corresponding performance measure of the microphone signal. The best performance measures for each condition are shown in **bold**.

			Filter Length (ms)										
	T_{60} (s)	unproc.	128	192	256	320	384	448	512	576	640	704	768
SRMR	0.5	2.42	3.11	3.46	3.56	3.59	3.62	3.64	**3.66**	-	-	-	-
	0.79	1.93	2.60	3.05	3.21	3.28	3.32	3.35	3.37	3.38	3.39	3.40	**3.41**
PESQ	0.5	2.44	2.60	2.70	**2.71**	2.68	2.62	2.56	2.51	-	-	-	-
	0.79	2.14	2.30	2.42	2.47	**2.49**	2.48	2.45	2.40	2.36	2.32	2.28	2.24
LSD	0.5	3.95	3.02	**2.93**	2.99	3.10	3.22	3.31	3.40	-	-	-	-
	0.79	5.15	4.03	3.63	3.48	**3.44**	3.47	3.53	3.60	3.68	3.75	3.80	3.87

Sound sources are placed in front of the dummy head with azimuths (relative to the dummy head center) varying from $-90°$ to $90°$, spaced by $5°$ (hence 37 azimuths), and an elevation of $0°$. The dummy-head-to-source distances were always 2 m. Two reverberation times, i.e. $T_{60} = 0.5$ s and 0.79 s, are simulated by adjusting the absorption coefficients of the walls. The speech signals from the TIMIT dataset [36] are taken as the source signals, with the duration about 4 s. For each reverberation time, 50 TIMIT signals are convolved with the BRIRs with a random azimuth as the reverberated microphone signals. To generate the noisy microphone signal, the spatially uncorrelated stationary speech-like noise is added to the microphone signals with signal-to-noise ratio (SNR) of 0, 5, 10, 15, 20 dB, respectively.

The sampling rate is 16 kHz. As already metioned in II-B, the STFT takes the Hamming window, with the window length and frame step of $N = 1,024$ (64 ms) and $L = N/4 = 256$, respectively. In the convex optimization algorithm, the parameters are set to $\mu = 20$, $\epsilon = \epsilon_{feas} = 10^{-6}$. Since the early reflections are preserved in the proposed method, the reference (target) signal of the proposed method is set as the early (64 ms) reverberated signal, which is generated by convolving the source signal with the first 64 ms (starting from the direct-path) of the first-channel RIRs. The noise PSD is estimated from the pure noise signals for various SNRs.

Two STFT-based state-of-the-art methods are compared, namely the weighted prediction error (WPE) [5] method, and the coherent-to-diffuse power ration (CDR) method [3]. For the WPE method, under the conditions with $T_{60} = 0.5$ s and 0.79 s, the number of filter coefficient is set to 50 and 80, respectively, which correspond $0.8T_{60}$ with respect to an 8 ms frame step. The source signal is taken as the reference signal. For the CDR method, the estimator with known DOA and unknown noise coherence is adopted, and the true DOA is used. The early (50 ms) reverberated signal is taken as the reference signal as used in [3].

Three metrics are taken to quantitatively evaluate the performance, (i) a non-intrusive metric, normalized speech-to-reverberation modulation energy ratio (SRMR) [37], and two intrusive metrics (ii) perceptual evaluation of speech quality (PESQ) [38] and (iii) log-spectral distance (LSD) [7] with the desired dynamic range of 50 dB. The dereverberated signal of various methods is an estimation of their corresponding reference signal up to a time shift and/or a gain factor. Therefore, as is for PESQ, the signals are aligned and amplitude normalized prior to the computation of the LSD. For the SRMR and PESQ, the higher the better, and for LSD, the lower the better.

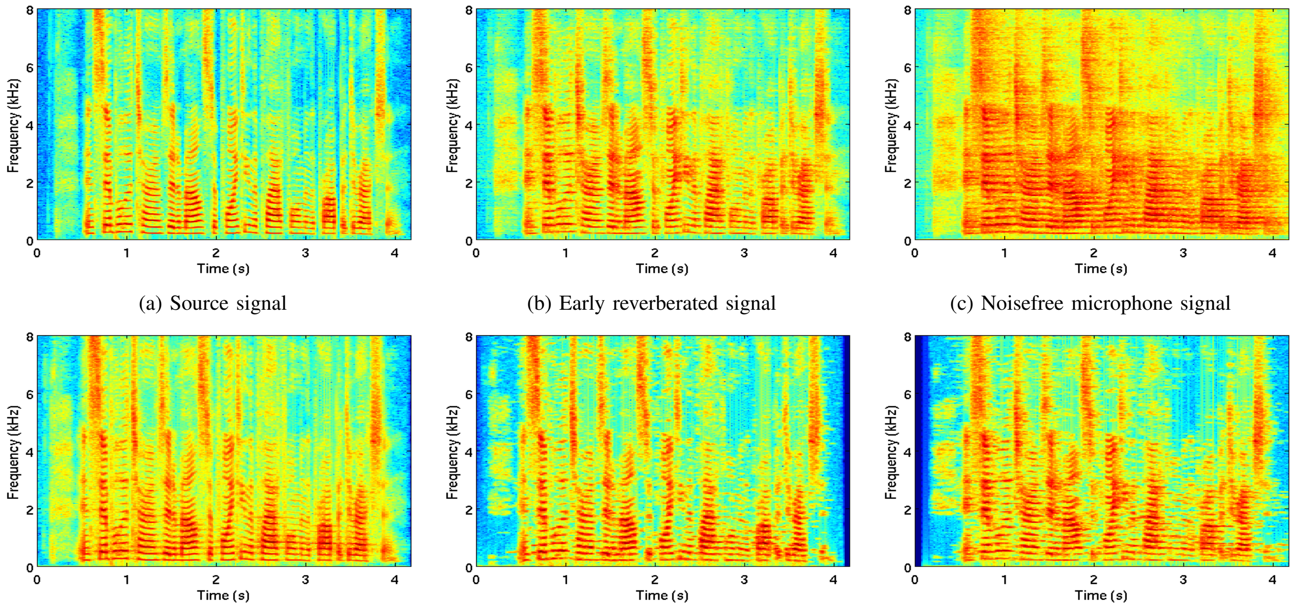
A. Selection of Channel Filter Length

The length of the channel filter Q (or $\tilde{Q}$) corresponding to the frame step L (or L_f) is a key parameter to be set. On the one hand, it influences the performance measures of the proposed method to a large extent. On the other hand, it is the only one prior knowledge that the proposed method requires, since it is related to the reverberation time.

Table II shows the performance measures of the proposed method as a function of filter length. It is observed that SRMR increases along with the increasing of filter length. The reason is that SRMR mainly measures the amount of the reverberation, and a long filter leads to a more sparse source signal estimation. However, when the channel filter is too long, the tail of the estimated filters will be noisy, which will distort the early reverberated signal. As a result, the PESQ score and LSD score degrade for the case of very long filter. When the filter length is set to be small, such as 128 ms, the filters can not cover most of the energy of the real RIRs. Consequently, the fitler estimation is inaccurate, and all the three metrics performs poorly. As a good compromise of the different metrics, the filter length is set to approximately 0.5 times T_{60}, e.g. 256 ms for $T_{60} = 0.5$ s and 384 ms for $T_{60} = 0.79$ s. With this setting, the channel filters cover major part of the energy of the RIRs, and avoid the long tail effect. Correspondingly, Q ($\tilde{Q}$) is set to 16 and 24 (4 and 6) for the cases of $T_{60} = 0.5$ s and $T_{60} = 0.79$ s, respectively. Note that although the performance scores are considerably affected by the selection of Q, the sound qualities by human listening are quite similar for the filter length selections from $0.3T_{60}$ to T_{60}.

B. Comparison with the Knwon CTFs Case

This paper proposes an STFT domain blind channel identification method. As mentioned in the methodology part, the identified frequency-wise filter suffers from some errors i) the CTF approximation error, namely the loss of cross-band information, ii) the frequency aliasing of filters caused by the critical sampling, iii) the amplitude approximation error, namely the loss of phase information. To evaluate



(a) Source signal

(b) Early reverberated signal

(c) Noisefree microphone signal

(d) Output: true CTFs

(e) Output: amplitude of true CTFs

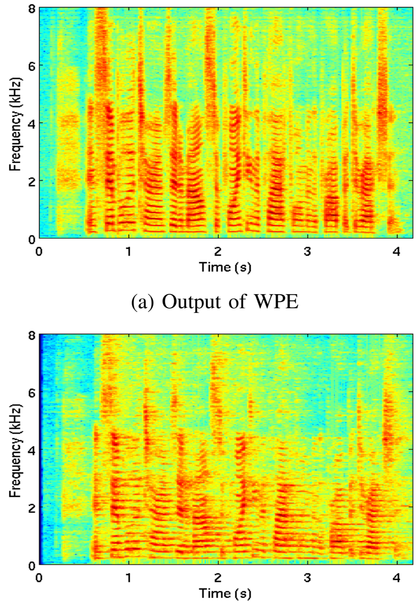
(f) Output: amplitude of identified CTFs

Fig. 2: Sonogram examples for the dereverberation algorithm using the true CTFs and identified CTFs. $T_{60} = 0.79$ s.

the applicability of the proposed method, we compare the dereverberated signals obtained by different approaches, i) using the true CTFs computed by (4) based on the known time domain filter, with the frame step $L = N/4$. The inverse filtering is carried out by minimizing the regualarized ℓ_2-norm cost function $\| \mathbf{C}\mathbf{s} - \mathbf{z} \| + \lambda |\mathbf{s}|$, where the regualarization term $|\mathbf{s}|$ is used to impose the sparsity of the source signal, and λ is set to 10^{-3}. The optimization method was proposed in [?] for the multi-source case, and is directly used for the single source case in this experiment. This filter only has the CTF approximation error, but not the frequency aliasing and the loss of phase, ii) using the true CTFs and the inverse filtering method proposed in Sec. III, namely only the amplitude of the true CTFs is used, iii) the proposed blind dereverberation method.

Fig. **??** depicts sonogram examples for these three cases. Fig. **??**a, b and c respectively illustrates the source signal, early reverberated signal and microphone singal. The smearing effect is evident by observing the microphone signal (Fig. **??**c). From Fig. **??**d, e and f, it can be seen that the signals of all the three approaches are much less reverberated than the microphone signal, showing the efficiency of the CTF-based inverse filtering method. The CTFs computed from the time domain filter do not have the gain ambiguity across frequencies. In other words, the frequency-wise dereverberated signals are consistent with the source signal, thus the output of true CTFs is an estimation of the source signal. We can observe that the dereverberated signal of true CTFs (Fig. **??**d) is close to the source signal, namely it does not include early reflections. This can also be verified by listening to the audio file, however, we can perceive a small delayed replica of the original source signal. This small replica is not obvious in the sonogram, and possibly caused by the loss of cross-band information. By only using the amplitude of true CTFs (Fig. **??**e), it can be seen that it recovers the main structure of the source signal. There are some spectral distortions for the low power region due to the loss of phase information, however we could not clearly perceive these distortations by listening. The small replica still exists, but is perceively less natural than the output of true CTFs, possibly also due to the loss of phase information. Compared to Fig. **??**d, it is observed that Fig. **??**e has more speech distortions, but not more reverberations. This verifies the assertion that the underestimated source singal will only suffer from more distortions, but not from more reverberations.

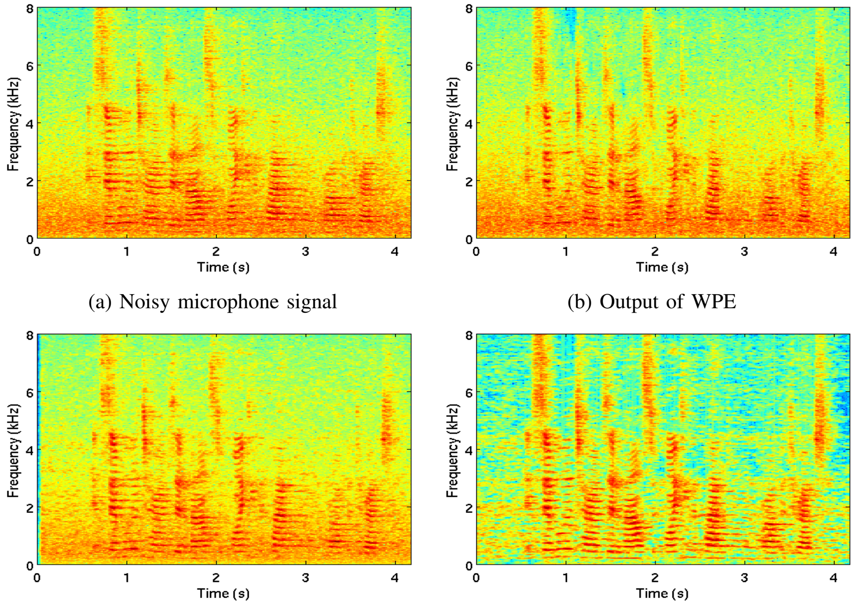
The output of the proposed blind method is shown in Fig. **??**f, which recovers the main structure of the early reverberated signal (Fig. **??**b). This verifies that the STFT configuration with oversampled signals and critical-sampled filters is suitable for the STFT-domain cross-relation method. This signal involves the early reflections, and also has some spectral distortions for the low power region. The sparsity regularization can partially reduce the residual noise caused by the filter aliasing and perturbations, and reduce the low power spectra in the replica. As a result, the overall residual noise is smaller, but more unnatural, and even sounds like musical noise. Compared to the singals in Fig. **??**d and e, the output signal of the proposed blind method sounds less noisy, since the early reflections enhance the desired signal. Fig. **??**f looks less distorted than Fig. **??**e, which verifies the assertion that the early reflections provide more tolerance to the underestimation. As will be shown in the following experiments, for the low input SNR case, the residual noise caused by the input noise will mask the musical noise. In addition, the residual noise can be mitigated by increasing the number of microphones.



(a) Output of WPE

(b) Output of CDR

Fig. 3: Sonogram examples of WEP and CDR. $T_{60} = 0.79$ s. The source signal and microphone signal are the ones in the example Fig. **??**.



(a) Noisy microphone signal

(b) Output of WPE

(c) Output of CDR

(d) Output of the proposed method

Fig. 4: Sonogram examples for the noisy case. The noisy microphone signal is generated by adding the speech-like noise to the noise-free microphone signal in Fig. **??**c with SNR of 5 dB.

C. Experimental Results for the Noise-free Case

For the proposed method and two competing methods, the quantitative performance measures, i.e. the averaged score over the 50 utterances, for the case of noise-free microphone signal are given in Table **??**. For each method, the performance measures for the unprocessed signal is computed based on the corresponding reference signal. Thence the unprocessed signals have the different PESQ and LSD scores for the three methods. For all the three methods, it is not surprising that all the performance scores for the case of $T_{60} = 0.79$ s are worse than the scores for the case of $T_{60} = 0.5$ s. In terms of SRMR, the proposed method achieves the highest scores among the three methods, with respect to the same baseline. This indicates that the reverberation residual of the proposed method is very low. For PESQ, the WPE method performs the best in terms of both the PESQ scores and the PESQ improvement (the increment from the PESQ score of the unprocessed signal to the score of the processed signal), and the proposed method outperforms the CDR method. The proposed method achieves the smallest LSD scores with the early reverberated signal as the reference. WPE has the highest LSD improvement with respect to the worst baseline. Overall, for the noise-free case, the proposed method and WPE achieves good performance measures, their outputs are fairly close to the early (64 ms) reverberated signal and the source signal, respectively. As for the CDR method, the spectral magnitude of the diffuse reverberations cannot be accurately estimated using only the known DOA.

TABLE III: The performance measures of the three methods for the case of noise-free microphone signal.

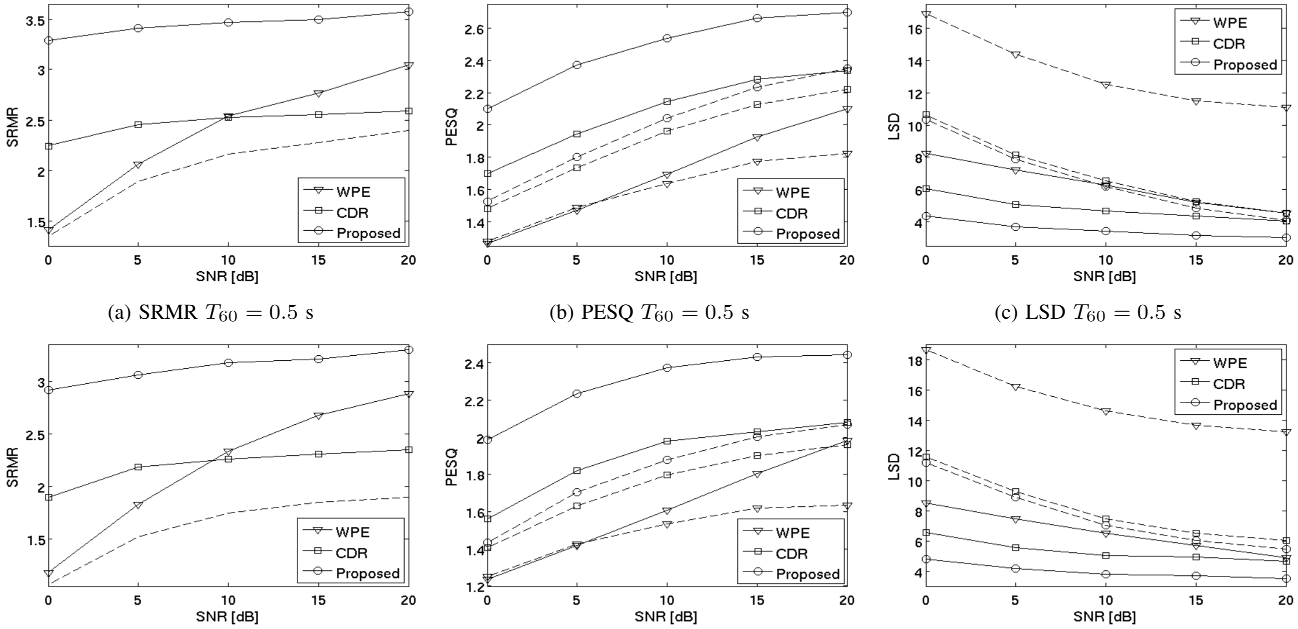
		WPE		CDR		Proposed	
	T_{60} (s)	unproc.	proc.	unproc.	proc.	unproc.	proc.
SRMR	0.5	2.42	3.23	2.42	2.61	2.42	3.56
	0.79	1.93	3.12	1.93	2.34	1.93	3.32
PESQ	0.5	1.90	3.18	2.31	2.52	2.44	2.71
	0.79	1.69	3.01	2.03	2.15	2.14	2.48
LSD	0.5	8.28	3.37	4.49	4.00	3.95	2.99
	0.79	12.78	6.01	5.84	4.96	5.15	3.47

A sonogram example for the proposed method was shown in Fig. **??**f, which intuitively illustrated the behavior of the proposed method. For comparison, Fig. **??** shows the sonogram examples for the WPE and CDR methods. It can be seen that the output of WPE is well dereverberated and less distorted, while there are late reverberations remained in the output of CDR.

D. Experimental Results for noisy case

To evaluate the dereverberation and noise reduction capabilities of the proposed method, the noisy signals with various SNRs are tested. WPE is not designed for the noisy case, however, it is worthwhile to test its sensitivity to noise. Fig. 3 depicts the performance measures of the three methods as a function of SNR. It can be seen that the performance measures for the cases of $T_{60} = 0.5$ s and $T_{60} = 0.79$ s are influenced by the noise signal in an identical way.

From Fig. 3a and 3d, it can be seen that the SRMR score of the unprocessed signal decreases with the decreasing of SNR. The proposed method achieves the much higher SRMR scores than the other two methods. This confirms the efficiency



(a) SRMR $T_{60} = 0.5$ s

(b) PESQ $T_{60} = 0.5$ s

(c) LSD $T_{60} = 0.5$ s

(d) SRMR $T_{60} = 0.79$ s

(e) PESQ $T_{60} = 0.79$ s

(f) LSD $T_{60} = 0.79$ s

Fig. 5: The performance measures for noisy signals as a function of SNR. The solid lines denote the scores of the processed signal for the three methods. The dashed lines with the same mark are the scores of the unprocessed signal taking the reference signal of the corresponding methods.

of the proposed method for jointly dereverberation and noise reduction. Along the decreasing of SNR, for the proposed method, the intense noise leads to more filter perturbations, noise residual and speech distortion. For the CDR method, the noise leads to the larger variance of the CDR estimation, and consequently more noise residual. As a result, the SRMR scores of CDR and the proposed method slowly decrease with the decreasing of SNR. However, WPE has a much larger degradation than the other two, which indicates the proposed method and CDR are less sensitivie to noise than WPE. From Fig. 3b and 3e, it is observed that the proposed method achieves the highest PESQ scores and PESQ improvement. The PESQ improvement of WPE is rapidly approaching to 0 with the decreasing of SNR. From Fig. 3e and 3f, it is seen that, the proposed method achieves the lowest LSD, which means that the outputs of the proposed method are close to the noise-free reference signal in terms of the logarithmic distance.

To intuitively illustrate the behaviors of three methods for noisy case, the sonogarm examples are given in Fig. **??**. As seen, WPE achieves a less reverberated speech signal than the microphone signal. However, it even amplifies the noise signal, as can be observed from the low frequency region where the WPE output is more noisy than the microphone signal. It is observed from Fig. **??**c that the CDR output is less noisy than the WPE output. Fig. **??**d shows that the output of the proposed method is much less noisy than the other two methods. This confirms that the proposed channel identification method (14) is robust to the spatially uncorrelated microphone noise. As metioned in Section III-A, the tolerance δ is set to be relatively small to avoid the signal distortion, thence we still can observe some residual noise, and the low power speech spectra are still masked by the residual noise. By informal listening test, the small replica and residual musical noise that present in the noise-free case are not clearly audible for this case, since the residual noise caused by the input noise is dominant, and this residual noise sounds close to the input noise.

V. Experiments with Real Recordings

To evaluate the proposed method in real world, and with multiple microphones, we conducted two set of experiments using i) the multichannel impulse response data [39], and ii) the real recordings in REVERB challenge [**?**]. The multichannel impulse response dataset is recorded using a 8-channel linear microphone array in the speech and acoustic lab of Ba Ilan University, with room size of 6 m × 6 m × 2.4 m. The reverberation time is controlled by 60 panels covering the room facets. The acoustic configuration of the impulse response dataset used in this experiment is (i) reverberation time $T_{60} = 0.61$ s, (ii) the microphone-to-source distance is 2 m, (iii) source direction is from $-90°$ to $90°$. The microphone array with 2-channel, 4-channel (the central two or four microphones) are tested for the proposed method. For each case, 50 TIMIT signals are convolved with the impulse responses with a random direction. The spatially uncorrelated stationary speech-like noise is added to the microphone signals with high SNR (20 dB) and low SNR (5 dB), respectively. The RealData in REVERB challenge is recorded in a room with T_{60} of 0.7 s. It contains 2 types of microphone-to-speaker distances, namely *near* (1 m) and *far* (2.5 m), which respectively have 90 and

TABLE IV: The performance measures for the multichannel impulse response dataset.

		CDR		WPE			Proposed		
	noise	unproc.	2-ch	unproc.	2-ch	4-ch	unproc.	2-ch	4-ch
SRMR	20 dB	2.55	2.79	2.55	3.05	3.12	2.55	3.47	3.48
	5 dB	1.97	2.62	1.97	2.12	2.14	1.97	3.25	3.32
PESQ	20 dB	2.55	2.83	2.00	2.36	2.42	2.67	2.92	2.96
	5 dB	1.91	2.32	1.59	1.60	1.58	1.97	2.47	2.59
LSD	20 dB	3.21	2.92	7.96	3.50	3.33	2.86	2.45	2.41
	5 dB	7.34	3.78	11.48	6.50	6.49	7.12	3.15	3.07

89 recordings with different directions. We test the 2-channel and 8-channel RealData for development (dev).

The same parameters are used as for the simulation dataset in Section IV, except that, according to the reverberation time, the filter length is approximately set to 320 ms, i.e. 20 taps, for both the two datasets. The noise PSD is estimated from the pure noise signals for the multichannel impulse response dataset, and is estimated using the noise PSD estimator [32] for the REVERB challenge dataset.

The perfomance measures for the multichannel impulse response dataset are shown in Table **??**. Note that CDR is only applicable for the 2-channel case, the filter length for WPE is set to 60 and 20 for the 2-channel and 4-channel cases, respectively. For the 2-channel case, it can be seen that the perfomance measures of all the three methods are almost consistent to the results obtained on the simulation data. The proposed method achieves better scores than the others. The WPE method performs worse for the low SNR case. WPE achieves better scores with 4-channel than the 2-channel case for the high SNR case, while not for the low SNR case. The proposed method achieves better scores with the 4-channel than the 2-channel case for both the two SNR cases. This indicates the efficiency of the multichannel extension of the proposed method. Both the channel identification method presented in Section II and the inverse filtering method presented in Section III benifit from a larger number of microphones. For the channel identification method, the identification of each channel is carried out by using the cross-relations with all the other channels, thence a more robust identification can be achieved by increasing the number of channels. For the inverse filtering method, a larger number of channels will give a larger data size of the ℓ_2 norm fitting problem in (18) and (19), which leads to a smaller error covariance of the least square estimation [40]. In addition, by listening test, the musical noise presented in the 2-channel case is noticeably suppressed in the 4-channel case.

The perfomance measures for the REVERB challenge dataset are shown in Table **??**. CDR needs the prior knowledge of either the DOA or the noise coherence, or both, which are not available for this real recording case, thus CDR is not tested for this experiment. Only the SRMR score is given due to the lack of reference signal. The proposed method and WPE both require the prior knowledge of reverberation time. The filter length for WPE is set to 70 and 10 for the 2-channel and 8-channel cases, respectively. We can observe that the proposed method achieves better SRMR scores. In the dereverberated signal of WPE, the residual noise is noticeable. The *near* case has a larger direct-to-reverberation ratio than the *far* case, in other words, the desired direct-path signal (and early reverberations) is less contaminated by the late reverberations. As a result, the SRMR scores of the *near* case are higher.

TABLE V: The performance measures for the REVERB challenge dataset.

			WPE		Proposed	
	dis	unproc.	2-ch	8-ch	2-ch	8-ch
SRMR	near	2.07	2.54	2.61	3.20	3.20
	far	1.90	2.46	2.54	3.03	3.07

The audio examples for all the datasets in this paper are available in our website.[2]

VI. Conclusion

In this paper, a blind multichannel speech dereverberation and noise reduction method has been proposed. The cross-relation method was extended to the STFT domain to circumvent the problem of near-common zeros for the long channel filters. The common zeros caused by the oversampling of STFT is solved by forcing the channel filters to perform the critical sampling. A constrained least-square problem was proposed to estimate the filters, which is robust to the noise interference and the filter length determination error. The channel filters were normalized by the first coefficient of one unique channel to remove the gain ambiguity accross subbands. The STFT domain inverse filtering can benifit from the sparsity of the source signal. An optimization problem with respect to the ℓ_1 norm of the source signal and the ℓ_2 norm fitting error between the microphone signals and the channel filtered source signal was proposed to reduce the noise caused by both the filter perturbations and microphone noise interference.

A series of experiments using binaural data and multichannel data have been conducted. It is confirmed that the proposed channel identification method is efficient for the real RIRs, even for the high reverberant case. In addition, the channel identification method is robust to the spatially uncorrelated stationary noise, even for the low SNR case. In the proposed inverse filtering method, the automatically set tolerance for the ℓ_2 norm fit works well for noise reduction. Overall, this paper proposes a practical real-world RIRs identification method, and thus a novel blind dereverberation and noise reduction method in the family of multichannel equalization technique.

References

[1] I. Arweiler and J. M. Buchholz, “The influence of spectral characteristics of early reflections on speech intelligibility,” *The Journal of the Acoustical Society of America*, vol. 130, no. 2, pp. 996–1005, 2011.

[2] E. A. Habets, S. Gannot, and I. Cohen, “Late reverberant spectral variance estimation based on a statistical model,” *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, 2009.

[2] https://team.inria.fr/perception/research/ctf-dereverberation

[3] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1006–1018, 2015.

[4] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 85–88, 2008.

[5] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 69–84, 2011.

[6] N. Ito, S. Araki, and T. Nakatani, "Probabilistic integration of diffuse noise suppression and dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5167–5171, 2014.

[7] O. Schwartz, S. Gannot, and E. Habets, "An expectation-maximization algorithm for multi-microphone speech dereverberation and noise reduction with coherence matrix estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1495–1510, 2016.

[8] E. Moulines, P. Duhamel, J.-F. Cardoso, and S. Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters," *IEEE Transactions on signal processing*, vol. 43, no. 2, pp. 516–525, 1995.

[9] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Transactions on signal processing*, vol. 43, no. 12, pp. 2982–2993, 1995.

[10] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 1074–1090, 2003.

[11] A. P. Liavas, P. A. Regalia, and J.-P. Delmas, "Robustness of least-squares and subspace methods for blind channel identification/equalization with respect to effective channel undermodeling/overmodeling," *IEEE transactions on signal processing*, vol. 47, no. 6, pp. 1636–1645, 1999.

[12] A. P. Liavas, P. A. Regalia, and J.-P. Delmas, "Blind channel approximation: effective channel order determination," *IEEE Transactions on Signal Processing*, vol. 47, no. 12, pp. 3336–3344, 1999.

[13] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Transactions on Signal Processing*, vol. 51, no. 1, pp. 11–24, 2003.

[14] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 882–895, 2005.

[15] X. Lin, N. D. Gaubitch, and P. A. Naylor, "Two-stage blind identification of SIMO systems with common zeros," in *14th European Signal Processing Conference*, pp. 1–5, 2006.

[16] P. A. Naylor, X. Lin, and A. W. Khong, "Near-common zeros in blind identification of simo acoustic systems," in *IEEE Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, pp. 21–24, 2008.

[17] X. Lin, A. W. Khong, and P. A. Naylor, "A forced spectral diversity algorithm for speech dereverberation in the presence of near-common zeros," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 888–899, 2012.

[18] M. Kallinger and A. Mertins, "Multi-channel room impulse response shaping-a study," in *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP)*, vol. 5, pp. V101–V104, 2006.

[19] A. Mertins, T. Mei, and M. Kallinger, "Room impulse response shortening/reshaping with infinity-and-norm optimization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 249–259, 2010.

[20] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1879–1890, 2013.

[21] F. Lim, W. Zhang, E. A. Habets, and P. A. Naylor, "Robust multichannel dereverberation using relaxed multichannel least squares," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1379–1390, 2014.

[22] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 1–12, 2007.

[23] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 680–693, 2016.

[24] M. Vetterli, "A theory of multirate filter banks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 3, pp. 356–372, 1987.

[25] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation," *IEEE Transactions on Signal Processing*, vol. 40, no. 8, pp. 1862–1875, 1992.

[26] J. P. Reilly, M. Wilbur, M. Seibert, and N. Ahmadvand, "The complex subband decomposition and its application to the decimation of large adaptive filtering problems," *IEEE Transactions on Signal Processing*, vol. 50, no. 11, pp. 2730–2743, 2002.

[27] B. Castro, N. Gaubitch, E. Habets, S. Gannot, P. Naylor, and S. Grant, "Subband scale factor ambiguity correction using multiple filterbanks," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, vol. 17, pp. 1061–1070, 2010.

[28] I. Kodrasi, A. Juki, and S. Doclo, "Robust sparsity-promoting acoustic multi-channel equalization for speech dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 166–170, 2016.

[29] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[30] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[31] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[32] X. Li, L. Girin, S. Gannot, and R. Horaud, "Non-stationary noise power spectral density estimation based on regional statistics," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 181–185, 2016.

[33] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[34] D. Campbell, "The roomsim user guide (v3. 3)," 2004.

[35] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR dummy-head microphone," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.

[36] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, 1988.

[37] J. F. Santos and T. H. Falk, "Updating the SRMR-CI metric for improved intelligibility prediction for cochlear implant users," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2197–2206, 2014.

[38] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 749–752, 2001.

[39] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 313–317, 2014.

[40] D. G. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing*, vol. 46. Artech House Norwood, 2005.