PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Binaural dereverberation based on delayed widely linear prediction in the time domain

Xin Leng[(1)], Jingdong Chen[(1)], and Jacob Benesty[(2)]

[(1)]CIAIC, Northwestern Polytechnical University, China, llengxin@126.com, jingdongchen@ieee.org
[(2)]INRS-EMT, University of Quebec, Canada, benesty@emt.inrs.ca

Abstract

Sound spatial information benefits human listeners in reverberant environments. This paper deals with the problem of binaural dereverberation, which reduces reverberation and meanwhile preserves the sound spatial information at the binaural outputs. A widely linear (WL) filtering framework is adopted where the multiple real microphone signals are merged into complex signals. The desired binaural outputs are also converted into complex signals with one channel being its real part, and the other channel being its imaginary part. By doing so, we transform the problem of binaural dereverberation to one of monaural dereverberation. In such a framework, the complex late reverberation is modeled using the multichannel delayed WL prediction by fully taking advantage of the noncircularity of the complex signals. A maximum likelihood method is then developed to estimate the optimal prediction filter with the speech signal of interest being modeled by a complex normal distribution. The relationship between the proposed method and the weighted prediction error (WPE) method is also discussed. Finally, simulation results are provided to justify the effectiveness of the developed method.

Keywords: Binaural dereverberation, delayed linear prediction, widely linear estimation, spatial information preservation.

1 INTRODUCTION

In various speech-related communication and human-machine interface systems, dereverberation is a critical problem since room reverberation may significantly degrade the speech quality and speech intelligibility as well as the recognition performance of automatic speech recognition (ASR) systems. As a result, tremendous research efforts have been devoted to this problem over the past few decades [1–8]. Most such efforts so far are focused on producing a single-channel dereverberated speech output. However, in hearing aid, virtual/augmented reality, and many other applications, in addition to reducing the reverberation effect, a key challenge is also to preserve the spatial information of the sound source such that after dereverberation, the human listeners can better localize the sound source with a binaural output. This problem is called binaural dereverberation [9–13]. The spatial information also benefits other signal processing techniques such as source separation, beamforming, and speech enhancement, to make them more effective in reverberant environments, where dereverberation is often used as a preprocessor [14–16].

It is known that early reflections (approximately within the first $40\,\mathrm{ms}$ of arrival) may help add warmth to sound due to the colorization effect, while the late reverberation smears the speech temporal and spectral information, thereby deteriorating the speech quality and intelligibility [13, 14]. To preserve the spatial information in binaural dereverberation, a straightforward way is to apply a same weighting factor at the two outputs while suppressing the reverberation with the spectral subtraction-based approach [9, 11] or based on a sigmoidal coherence-to-gain mapping [13] in the power spectral domain. In [9], a second stage coherence-based Wiener filter was also suggested to further attenuate the residual reverberation by taking into account the shadowing effects of the human head. However, this method requires good estimation of the spatial coherence and the dereverberation process corrupts the linear relationship between the source and microphone signals.

In this paper, a widely linear (WL) filtering approach is adopted to achieve binaural dereverberation using a microphone array in the time domain. The WL filtering framework has shown its potential to preserve the sound spatial information in binaural noise reduction [17, 18] and stereophonic acoustic echo cancellation [19–21]. In such a framework, the multiple real microphone signals and binaural outputs are merged into complex signals. The complex late reverberation is then modeled by a multichannel delayed WL prediction. The estimation of the WL prediction coefficients is achieved through a maximum likelihood method in which

the speech signal of interest is modeled by a complex normal distribution [22, 23]. Note that the optimal prediction filter can also be estimated in the real domain with a proper formulation where the binaural output speech signals are modeled with the multivariable normal distribution, which is also discussed in this paper. We also analyze the relationship between the proposed method and the popularly used weighted prediction error (WPE) method proposed in [24], where speech dereverberation is performed using the delayed linear prediction.

2 SIGNAL MODEL

Consider a linear microphone array consisting of $2M$ sensors, which is used to capture a signal of interest in some reverberant and noisy environment. The received signal at the mth ($m = 1, 2, \ldots, 2M$) microphone can then be expressed as

$$\begin{aligned} x_m(t) &= g_m(t) * s(t) + v_m(t) \\ &= \sum_{k=0}^{L_g-1} g_{m,k} s(t-k) + v_m(t), \end{aligned} \tag{1}$$

where $*$ denotes linear convolution, $s(t)$ is the unknown source signal, $g_m(t)$ is the room impulse response from the source to the mth microphone, and $v_m(t)$ is the additive noise at the mth microphone. We assume that all the signals $x_m(t)$ and $v_m(t)$ are real, broadband, and zero mean, and $g_m(t)$, $m = 1, 2, \ldots, 2M$ share no common zeros.

Let us first neglect the noise term $v_m(t)$ for simplicity as in [24, 25]. The reverberant speech signal $x_m(t)$ can then be written as the sum of two components, i.e.,

$$x_m(t) = e_m(t) + r_m(t), \tag{2}$$

where

$$e_m(t) \triangleq \sum_{k=0}^{D-1} g_{m,k} s(t-k), \tag{3}$$

$$r_m(t) \triangleq \sum_{k=D}^{L_g-1} g_{m,k} s(t-k), \tag{4}$$

are, respectively, the direct sound plus early reflections and late reverberation, and D is the time duration of the early reflections. In dereverberation, $e_m(t)$ and $r_m(t)$ are often assumed to be uncorrelated with each other [14].

Following (4), we can rewrite $x_m(t)$ as [24]

$$\begin{aligned} x_m(t) &= e_m(t) + \mathbf{g}_{\mathrm{l},m}^T \mathbf{s}(t-D) \\ &= e_m(t) + \mathbf{g}_{\mathrm{l},m}^T \mathbf{G}^+ \mathbf{G}\mathbf{s}(t-D) \\ &= e_m(t) + \mathbf{a}_m^T \mathbf{x}(t-D), \end{aligned} \tag{5}$$

where the superscript ^{T} is the transpose operator, $^+$ denotes the Moore-Penrose pseudo-inverse,

$$\mathbf{g}_{\mathrm{l},m} \triangleq \begin{bmatrix} g_{m,D} & g_{m,D+1} & \cdots & g_{m,L_g-1} & 0 & \cdots & 0 \end{bmatrix}^T, \tag{6}$$

$$\mathbf{s}(t) \triangleq \begin{bmatrix} s(t) & s(t-1) & \cdots & s(t-L_s+1) \end{bmatrix}^T, \tag{7}$$

$$\begin{aligned} \mathbf{x}(t) &\triangleq \begin{bmatrix} \mathbf{x}_1^T(t) & \mathbf{x}_2^T(t) & \cdots & \mathbf{x}_{2M}^T(t) \end{bmatrix}^T \\ &= \mathbf{G}\mathbf{s}(t), \end{aligned} \tag{8}$$

$$\mathbf{x}_m(t) \triangleq \begin{bmatrix} x_m(t) & x_m(t-1) & \cdots & x_m(t-L+1) \end{bmatrix}^T, \tag{9}$$

$$\mathbf{a}_m = (\mathbf{G}^T)^+ \mathbf{g}_{\mathrm{l},m}, \tag{10}$$

$\mathbf{g}_{\mathrm{l},m}$ and $\mathbf{s}(t)$ are, respectively, the late impulse response and source signal, both of length L_s, $\mathbf{x}_m(t)$, of length L, is the vector form of $x_m(t)$, $\mathbf{x}(t)$ and $\mathbf{a}_m$ are, respectively, the observation signal vector and prediction filter, both of length $2ML$,

$$\mathbf{G} \triangleq \begin{bmatrix} \mathbf{G}_1^T & \mathbf{G}_2^T & \cdots & \mathbf{G}_{2M}^T \end{bmatrix}^T, \tag{11}$$

$$\mathbf{G}_m \triangleq \begin{bmatrix} \mathbf{g}_m^T & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{g}_m^T & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \mathbf{g}_m^T \end{bmatrix}, \tag{12}$$

$$\mathbf{g}_m \triangleq \begin{bmatrix} g_{m,0} & g_{m,1} & \cdots & g_{m,L_g-1} \end{bmatrix}^T, \tag{13}$$

$\mathbf{g}_m$, of length L_g, is the vector form of $g_m(t)$, and $\mathbf{G}$ and $\mathbf{G}_m$ are matrices of size $2ML \times L_s$ and $L \times L_s$ respectively, $L_s = L + L_g - 1$. Here, we assume $L_s \leqslant 2ML$. Assume that $g_m(t),\ m = 1, 2, \ldots, 2M$ share no common zeros, then $\mathrm{rank}(\mathbf{G}) = L_s$, and $\mathbf{G}^+\mathbf{G} = \mathbf{I}_{L_s}$, where $\mathbf{I}_{L_s}$ is the identity matrix of size $L_s \times L_s$. From (5), one can see that the late reverberation $r_m(t)$ can be modeled by a delayed linear prediction with the prediction filter $\mathbf{a}_m$.

3 WIDELY LINEAR MODEL

To perform binaural dereverberation, it is necessary to produce at least two output signals from the observations at the microphone array. This can be done by using the WL filtering framework, where the multiple real microphone signals are merged into complex signals [17, 18], i.e.,

$$y_i(t) \triangleq x_i(t) + \jmath x_{M+i}(t) = h_i(t) * s(t), \tag{14}$$

where $\jmath = \sqrt{-1}$ is the imaginary unit, and $h_i(t) \triangleq g_i(t) + \jmath g_{M+i}(t)$ is the complex room impulse response of the ith ($i = 1, 2, \ldots, M$) complex channel. It follows immediately according to (5) that

$$\begin{aligned} y_i(t) &= d_i(t) + \mathbf{h}_{\mathrm{l},i}^T \mathbf{s}(t - D) \\ &= d_i(t) + \mathbf{h}_{\mathrm{l},i}^T \mathbf{H}^+ \mathbf{H}\mathbf{s}(t - D) \\ &= d_i(t) + \widetilde{\mathbf{h}}_i^H \widetilde{\mathbf{y}}(t - D), \end{aligned} \tag{15}$$

where the superscript ^{H} stands for the conjugate-transpose operator, $d_i(t) \triangleq e_i(t) + \jmath e_{M+i}(t)$ is the complex direct sound plus early reflections, $\mathbf{h}_{\mathrm{l},i} \triangleq \mathbf{g}_{\mathrm{l},i} + \jmath \mathbf{g}_{\mathrm{l},M+i}$ is the complex late impulse response,

$$\widetilde{\mathbf{h}}_i = (\mathbf{H}^H)^+ \mathbf{h}_{\mathrm{l},i}^* = \frac{1}{2} \mathbf{T}_{ML} (\mathbf{G}^T)^+ \mathbf{h}_{\mathrm{l},i}^*, \tag{16}$$

$$\widetilde{\mathbf{y}}(t) \triangleq \begin{bmatrix} \mathbf{y}(t) \\ \mathbf{y}^*(t) \end{bmatrix} = \mathbf{H}\mathbf{s}(t) = \mathbf{T}_{ML}\mathbf{x}(t), \tag{17}$$

$$\mathbf{y}(t) \triangleq \begin{bmatrix} \mathbf{y}_1^H(t) & \mathbf{y}_2^H(t) & \cdots & \mathbf{y}_M^H(t) \end{bmatrix}^H, \tag{18}$$

the superscript * denotes complex conjugation, $\mathbf{y}_i(t)$ is defined analogously to $\mathbf{x}_i(t)$,

$$\mathbf{H} = \mathbf{T}_{ML}\mathbf{G}, \tag{19}$$

and

$$\mathbf{T}_{ML} \triangleq \begin{bmatrix} \mathbf{I}_{ML} & \jmath\mathbf{I}_{ML} \\ \mathbf{I}_{ML} & -\jmath\mathbf{I}_{ML} \end{bmatrix} \tag{20}$$

is the real-to-complex transformation matrix, $\mathbf{H}^+\mathbf{H} = \mathbf{I}_{L_s}$.

Inspecting (15) and (17), one can see that both $\mathbf{y}(t - D)$ and $\mathbf{y}^*(t - D)$ are used to model the late reverberation. Without loss of generality, we choose $d_1(t) = e_1(t) + \jmath e_{M+1}(t)$ as the desired signal to recover. An estimate of $d_1(t)$ can then be obtained as

$$\widehat{d}_1(t) = y_1(t) - \widetilde{\mathbf{h}}^H \widetilde{\mathbf{y}}(t - D), \tag{21}$$

where $\widetilde{\mathbf{h}}$ is the complex prediction filter to be estimated. From (21), one can see that the objective of binaural dereverberation is to find an optimal filter $\widetilde{\mathbf{h}}$ that can best estimate $d_1(t)$ from $y_1(t)$ and $\widetilde{\mathbf{y}}(t - D)$.

4 DELAYED WIDELY LINEAR PREDICTION

The dereverberation filter proposed in [24] assumes that $e_m(t)$ is a Gaussian process with a time-varying variance. Since we deal with complex signals, the complex normal distribution should be exploited to model $d_1(t)$ [22, 23]. Therefore, the probability density function (pdf) of $d_1(t)$ is

$$p\big[\mathbf{d}_1(t)\big] = \frac{1}{\pi[\det \mathbf{\Gamma}(t)]^{1/2}} \exp\left[-\frac{1}{2}\mathbf{d}_1^H(t)\mathbf{\Gamma}^{-1}(t)\mathbf{d}_1(t)\right], \tag{22}$$

where

$$\mathbf{d}_1(t) = \left[d_1(t) \;\; d_1^*(t)\right]^T, \tag{23}$$

$$\mathbf{\Gamma}(t) = \sigma_{d_1}^2(t) \begin{bmatrix} 1 & \gamma_{d_1}(t) \\ \gamma_{d_1}^*(t) & 1 \end{bmatrix}, \tag{24}$$

$\sigma_{d_1}^2(t) \triangleq E\left[|d_1(t)|^2\right]$ is the variance of $d_1(t)$, $E[\cdot]$ denotes mathematical expectation, and

$$\gamma_{d_1}(t) = \frac{E\left[d_1^2(t)\right]}{E\left[|d_1(t)|^2\right]} \tag{25}$$

is the second-order circularity quotient [17, 18] of $d_1(t)$. The absolute value of $\gamma_{d_1}(t)$ satisfies $0 \leqslant |\gamma_{d_1}(t)| \leqslant 1$. If $\gamma_{d_1}(t) = 0$, $d_1(t)$ is second-order circular since its pdf can be well modeled by its variance. Otherwise, it is non-circular.

Let us check the circularity of $d_1(t)$. According to (25), we have

$$\gamma_{d_1}(t) \triangleq \frac{E\left[e_1^2(t)\right] - E\left[e_{M+1}^2(t)\right] + 2\jmath E\left[e_1(t)e_{M+1}(t)\right]}{E\left[|d_1(t)|^2\right]}. \tag{26}$$

According to the signal model given in (1)–(3) and the assumption that $e_m(t)$, $m = 1, 2, \ldots, 2M$, are from the same source $s(t)$, we should have $E\left[e_1(t)e_{M+1}(t)\right] \neq 0$. Therefore, with our signal model, we can state that $\gamma_{d_1}(t) \neq 0$, and $d_1(t)$ is noncircular [17, 18].

4.1 Delayed Widely Linear Prediction Method

With the pdf of $d_1(t)$, the parameters $\boldsymbol{\theta} = \left\{\widetilde{\mathbf{h}}, \mathbf{\Gamma}(t)\right\}$ can be estimated by maximizing the log likelihood function [24]:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{\mathcal{T}} \log p\left[\mathbf{d}_1(t)\right], \tag{27}$$

where $\mathcal{T}$ is the time duration of the signal of interest. The solution can be derived iteratively as follows [24].

1) Initialize

$$\widehat{\mathbf{\Gamma}}(t) = \sum_{\tau=-L_f}^{L_f} \mathbf{d}_1(t+\tau)\mathbf{d}_1^H(t+\tau) \tag{28}$$

with a frame length of $2L_f + 1$ and $d_1(t) = y_1(t)$.

2) Repeat $\widetilde{\mathbf{h}}$ and $\widehat{\mathbf{\Gamma}}(t)$ until convergence

(a) Update $\widetilde{\mathbf{h}}$

$$\begin{bmatrix} \widetilde{\mathbf{h}} \\ \widetilde{\mathbf{h}}^* \end{bmatrix} = \begin{bmatrix} \mathbf{R} & -\mathbf{C} \\ -\mathbf{C}^H & \mathbf{R}^T \end{bmatrix}^+ \begin{bmatrix} \mathbf{r} - \mathbf{c} \\ \mathbf{r}^* - \mathbf{c}^* \end{bmatrix}, \tag{29}$$

where

$$\mathbf{R} = \sum_{t=1}^{\mathcal{T}} \frac{\widetilde{\mathbf{y}}(t-D)\widetilde{\mathbf{y}}^H(t-D)}{\widehat{\sigma}_{d_1}^2(t)(1 - |\gamma_{d_1}(t)|^2)}, \tag{30}$$

$$\mathbf{C} = \sum_{t=1}^{\mathcal{T}} \frac{\widehat{\gamma}_{d_1}^*(t)\widetilde{\mathbf{y}}(t-D)\widetilde{\mathbf{y}}^T(t-D)}{\widehat{\sigma}_{d_1}^2(t)(1 - |\gamma_{d_1}(t)|^2)}, \tag{31}$$

$$\mathbf{r} = \sum_{t=1}^{\mathcal{T}} \frac{\widetilde{\mathbf{y}}(t-D)y_1^*(t)}{\widehat{\sigma}_{d_1}^2(t)(1 - |\gamma_{d_1}(t)|^2)}, \tag{32}$$

$$\mathbf{c} = \sum_{t=1}^{\mathcal{T}} \frac{\widehat{\gamma}_{d_1}^*(t)\widetilde{\mathbf{y}}(t-D)y_1(t)}{\widehat{\sigma}_{d_1}^2(t)(1 - |\gamma_{d_1}(t)|^2)}. \tag{33}$$

(b) Update (28) using

$$\widehat{d}_1(t) = y_1(t) - \widetilde{\mathbf{h}}^H\widetilde{\mathbf{y}}(t-D). \tag{34}$$

4.2 Delayed Linear Prediction Method

The previous subsection derived the prediction filter $\widetilde{\mathbf{h}}$ by using the WL filtering framework in the complex domain. In this subsection, we derive an equivalent method in the real domain. Let us denote $\underline{\mathbf{x}}(t) = [x_1(t) \;\; x_{M+1}(t)]^T$ and $\underline{\mathbf{e}}(t) = [e_1(t) \;\; e_{M+1}(t)]^T$. Using (5), we have

$$\underline{\mathbf{x}}(t) = \underline{\mathbf{e}}(t) + \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_{M+1}^T \end{bmatrix} \mathbf{x}(t-D), \tag{35}$$

$$\mathbf{d}_1(t) = \begin{bmatrix} 1 & \jmath \\ 1 & -\jmath \end{bmatrix} \underline{\mathbf{e}}(t). \tag{36}$$

Therefore, an estimate of $\underline{\mathbf{e}}(t)$ can then be obtained as

$$\widehat{\underline{\mathbf{e}}}(t) = \underline{\mathbf{x}}(t) - \mathbf{A}^T \mathbf{x}(t-D), \tag{37}$$

where $\mathbf{A} = [\widehat{\mathbf{a}}_1 \;\; \widehat{\mathbf{a}}_{M+1}]$ is the estimated prediction filter.
Substituting (36) into (22), we get

$$p\left[\underline{\mathbf{e}}(t)\right] = \frac{1}{2\pi[\det \mathbf{\Lambda}(t)]^{1/2}} \exp\left[-\frac{1}{2}\underline{\mathbf{e}}^T(t)\mathbf{\Lambda}^{-1}(t)\underline{\mathbf{e}}(t)\right], \tag{38}$$

where $\mathbf{\Lambda}(t) = E[\underline{\mathbf{e}}(t)\underline{\mathbf{e}}^T(t)]$ is the correlation matrix of $\underline{\mathbf{e}}(t)$. The estimation of $\mathbf{A}$ and $\mathbf{\Lambda}(t)$ can be achieved in a similar way as in [15, 16] by maximizing the log likelihood function (27).

4.3 Special Case

Let us consider the particular case where $\gamma_{d_1}(t) = 0$. In such a scenario, we have

$$p\big[d_1(t)\big] = \frac{1}{\pi\sigma_{d_1}^2(t)} \exp\left[-\frac{|d_1(t)|^2}{\sigma_{d_1}^2(t)}\right]. \tag{39}$$

Using (39), (27) can be rewritten as

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_{t=1}^{\mathcal{T}} \left[\frac{\left|y_1(t) - \widetilde{\mathbf{h}}^H \widetilde{\mathbf{y}}(t-D)\right|^2}{\sigma_{d_1}^2(t)} + \log \sigma_{d_1}^2(t)\right] + \text{const.}, \tag{40}$$

and (29) becomes

$$\widetilde{\mathbf{h}} = \mathbf{R}^{+}\mathbf{r}. \tag{41}$$

Combining (17) and (41) and substituting the result into (21), we get

$$\begin{aligned} \widehat{d}_1(t) &= y_1(t) - \widehat{\mathbf{a}}_1^T \mathbf{x}(t-D) - \jmath \widehat{\mathbf{a}}_{M+1}^T \mathbf{x}(t-D) \\ &= \widehat{e}_1(t) + \jmath \widehat{e}_{M+1}(t), \end{aligned} \tag{42}$$

where

$$\widehat{\mathbf{a}}_m = \left[\sum_{t=1}^{\mathcal{T}} \frac{\mathbf{x}(t-D)\mathbf{x}^T(t-D)}{\widehat{\sigma}_{e_m}^2(t)}\right]^{+} \sum_{t=1}^{\mathcal{T}} \frac{\mathbf{x}(t-D)x_m(t)}{\widehat{\sigma}_{e_m}^2(t)}, \tag{43}$$

with $\sigma_{e_m}^2(t) \triangleq E[e_m^2(t)]$ being the variance of $e_m(t)$. From (42), one can see that this output fully corresponds to the delayed linear prediction as in [24]. In other words, the method developed in [24] is a particular case of the method developed in this paper.

5 SIMULATIONS

Simulations are conducted using the impulse responses measured in the Bell Labs Varechoic Chamber [26, 27]. Two microphones at (3.237, 0.500, 1.400) (in meters) and (3.437, 0.500, 1.400) are used, and the clean speech source is placed at the positions (1.337:1.000:5.337, 1.938, 1.600) to simulate a moving source, i.e., the source stays at one position for 2 s and then moves to the next position. This moving source is used to evaluate the performance of spatial information preservation of the developed dereverberation method. We consider two sets of reverberation conditions: a light reverberation condition with the reverberation time T_{60} being approximately 0.24 s and a moderate reverberation condition with T_{60} being approximately 0.58 s.

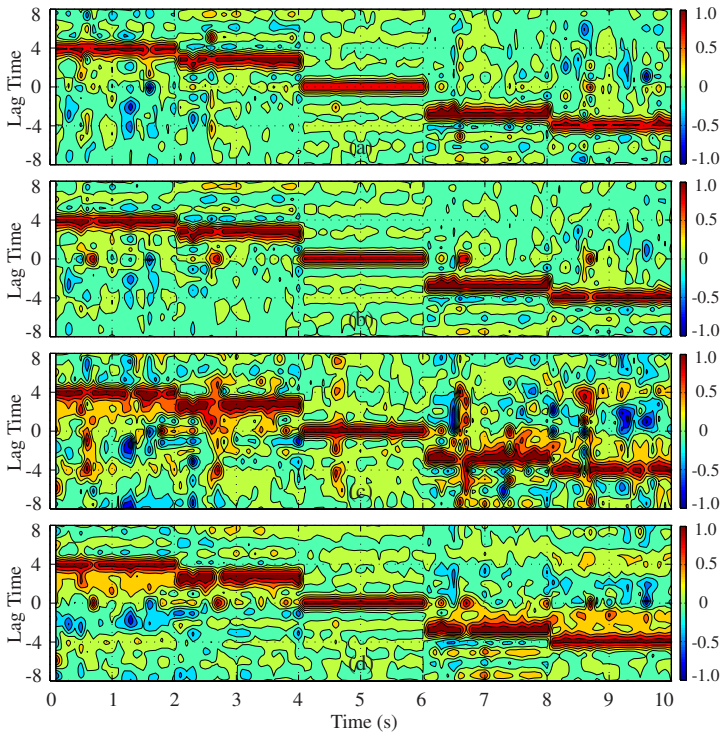

Figure 1. Contours of the computed short-time CCFs in two room reverberant conditions: (a) between the two reverberant signals with $T_{60} = 0.24\,\mathrm{s}$, (b) between the signals after dereverberation of the signals in (a), (c) between the two reverberant signals with $T_{60} = 0.58\,\mathrm{s}$, and (d) between the signals after dereverberation of the signals in (c).

The clean speech signals used are taken from the TIMIT database [28]. We consider dereverberation of narrowband speech with $4\,\mathrm{kHz}$ bandwidth and $8\,\mathrm{kHz}$ sampling rate. So, all the signals are resampled from its original sampling frequency to $8\,\mathrm{kHz}$. The reverberant signals are generated by convolving the source signal with the corresponding measured room impulse responses.

The parameters for all simulations are set as follows: $D = 240$ and $L_f = 160$. L is set to 1600 for the light reverberation condition and 3200 for the moderate reverberation condition. The number of iterations is limited to 1 as in [24] in all cases. Note that the pre-whitening technique was also exploited before estimating the filter coefficients as in [14].

The cross-correlation function (CCF) is computed to visualize the spatial sound information before and after dereverberation. The CCF is computed using a short time average every $100\,\mathrm{ms}$ without any overlap as in [17], whose maximal value stands for the current source direction. Figure 1 plots the contours of the computed CCF between the two reverberant inputs as well as that between the two dereverberant outputs in the two studied reverberantion conditions. Comparing Fig. 1(a) and (c), one can clearly see that reverberation deteriorates the source spatial information. But the spatial information is well recovered by the developed method as seen in Fig. 1(b) and (d).

To further evaluate the dereverberation performance of the developed method, we compute the average cepstral distance (CD) over short-time frames according to [24]. Specifically, the CD (in dB) between two signals at each short-time frame is defined as

$$\mathrm{CD} = \frac{10}{\ln 10}\sqrt{(\hat{\beta}_0 - \beta_0)^2 + 2\sum_{k=1}^{12}(\hat{\beta}_k - \beta_k)^2}, \tag{44}$$

where $\hat{\beta}_k$ and β_k are the real cepstral coefficients [29] of the speech signal under evaluation and desired signal (direct sound plus early reverberation), respectively. We use 20 sentences from different speakers (10 male and 10 female), and compute the average CDs for the left and right output channels separately with a frame length of $32\,\mathrm{ms}$, and 75% overlap. The length of speech are all $2\,\mathrm{s}$, and the room reverberation condition is $T_{60} = 0.58\,\mathrm{s}$. With our experimental setup, the average CDs of the left- and right-channel reverberant speech are 2.69 and 2.72, respectively, while that of the dereverberant speech are 2.43 and 2.45, respectively.

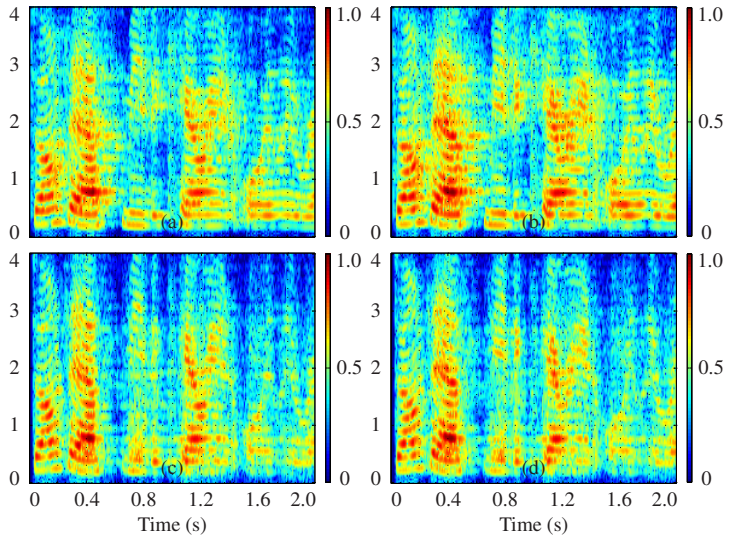

Figure 2. Spectrograms of the reverberant signals [(a) left channel, (b) right channel] and the dereverberated signals [(c) left channel, (d) right channel]. The room reverberation time is $T_{60} = 0.58\,\mathrm{s}$.

To give an example, Figure 2 plots the spectrograms of the left- and right-channel reverberant and dereverberated speech signals. One can see from this figure that the developed method has effectively reduced the reverberation component from the reverberant speech. Meanwhile, the speech spectrum and the spatial information are well preserved.

6 CONCLUSIONS

This paper studied the binaural dereverberation problem using a microphone array in the time domain. By adopting the WL framework, we merge the multiple real microphone inputs and binaural outputs into complex signals. The late reverberation is then modeled using the delayed WL prediction, which takes into account the noncircular property of the complex speech signals. The optimal prediction filter coefficients are estimated by maximizing the log likelihood function where the complex normal distribution is used to model the speech signal of interest. The relation of the developed method with the WPE method developed in the literature was also discussed. Simulation results showed that this dereverberation filter can effectively reduce the reverberation effect and meanwhile preserves the spatial information of the sound source.

ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation of China (NSFC) and the Israel Science Foundation (ISF) joint research program (grant No. 61761146001), the NSFC key program (grant No. 61831019), the NSFC Distinguished Young Scientists Fund (grant No. 61425005).

REFERENCES

[1] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, 1988.

[2] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation*. Berlin, Germany: Springer Science & Business Media, 2010.

[3] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[4] J.-M. Yang and H.-G. Kang, "Online speech dereverberation algorithm based on adaptive multichannel linear prediction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 608–619, Mar. 2014.

[5] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, Jun. 2015.

[6] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, Jan 2016.

[7] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1492–1501, Jul. 2017.

[8] S. Watanabe, M. Delcroix, F. Metze, and J. R. Hershey, *New Era for Robust Speech Recognition*. Berlin, Germany: Springer-Verlag, 2017.

[9] M. Jeub, M. Schafer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1732–1745, Sep. 2010.

[10] M. Jeub and P. Vary, "Binaural dereverberation based on a dual-channel Wiener filter with optimized noise field coherence," in *Proc. IEEE ICASSP*, 2010, pp. 4710–4713.

[11] A. Tsilfidis, E. Georganti, and J. Mourjopoulos, "Binaural extension and performance of single-channel spectral subtraction dereverberation algorithms," in *Proc. IEEE ICASSP*, 2011, pp. 1737–1740.

[12] A. Tsilfidis, A. Westermann, J. M. Buchholz, E. Georganti, and J. Mourjopoulos, "Binaural dereverberation," in *The technology of binaural listening*. Berlin, Germany: Springer, 2013, pp. 359–396.

[13] A. Westermann, J. M. Buchholz, and T. Dau, "Binaural dereverberation based on interaural coherence histograms," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. 2767–2777, May 2013.

[14] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 534–545, May 2009.

[15] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 69–84, Jan. 2011.

[16] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[17] J. Benesty, J. Chen, and Y. Huang, "Binaural noise reduction in the time domain with a stereo setup," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2260–2272, Nov. 2011.

[18] J. Chen and J. Benesty, "On the time-domain widely linear LCMV filter for noise reduction with a stereo system," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1343–1354, Jul. 2013.

[19] J. Benesty, C. Paleologu, T. Gänsler, and S. Ciochină, *A perspective on stereophonic acoustic echo cancellation*. Springer Science & Business Media, 2011, vol. 4.

[20] C. Stanciu, J. Benesty, C. Paleologu, T. Gänsler, and S. Ciochină, "A widely linear model for stereophonic acoustic echo cancellation," *Signal Process.*, vol. 93, no. 2, pp. 511–516, 2013.

[21] C. Paleologu, J. Benesty, and S. Ciochină, "Widely linear general Kalman filter for stereophonic acoustic echo cancellation," *Signal Process.*, vol. 94, pp. 570–575, 2014.

[22] A. van den Bos, "The multivariate complex normal distribution-a generalization," *IEEE Trans. Inform. Theory*, vol. 41, no. 2, pp. 537–539, Mar. 1995.

[23] B. Picinbono, "Second-order complex random vectors and normal distributions," *IEEE Trans. Signal Process.*, vol. 44, no. 10, pp. 2637–2640, Oct. 1996.

[24] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[25] A. Jukic and S. Doclo, "Speech dereverberation using weighted prediction error with Laplacian model of the desired signal," in *Proc. IEEE ICASSP*, 2014, pp. 5172–5176.

[26] W. C. Ward, G. Elko, R. Kubli, and W. McDougald, "The new varechoic chamber at AT&T Bell Labs," in *Proc. Wallace Clement Sabine Centenn. Symp.*, 1994.

[27] A. Härmä, "Acoustic measurement data from the varechoic chamber," *Agere Systems, Tech. Memo.*, Nov. 2001.

[28] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, *TIMIT acoustic-phonetic continuous speech corpus*. Linguistic Data Consort., 1993.

[29] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. Commun. Comput. Signal Process.*, vol. 1, 1993, pp. 125–128.