A BILINEAR FRAMEWORK FOR ADAPTIVE SPEECH DEREVERBERATION COMBINING BEAMFORMING AND LINEAR PREDICTION

Wenxing Yang[1,2], *Gongping Huang*[2], *Andreas Brendel*[2], *Jingdong Chen*[1],
Jacob Benesty[3], *Walter Kellermann*[2], *and Israel Cohen*[4]

[1]CIAIC, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[2]LMS, University Erlangen-Nuremberg, 91058 Erlangen, Germany
[3]INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada
[4]Faculty of Electrical and Computer Engineering,
Technion–Israel Institute of Technology, Technion City, Haifa 3200003, Israel

ABSTRACT

Speech dereverberation algorithms based on multichannel linear prediction (MCLP) are effective under various acoustic conditions. This paper proposes a bilinear form for the MCLP based dereverberation, where the MCLP filter is expressed as a Kronecker product of a spatial filter and a temporal filter. Then, a recursive least-squares (RLS)-based algorithm is derived for adaptive speech dereverberation. Compared with the original MCLP-based adaptive algorithm, the advantages of the proposed method are twofold: (1) the computational complexity is significantly reduced and is more suitable for dynamic scenarios, since fewer parameters have to be estimated per signal-block observation; and (2) it is more robust to noise by optimizing the spatial filter as a weighted minimum power distortionless response (wMPDR) beamformer. Simulation results validate the advantages of the proposed algorithm.

Index Terms— Dereverberation, multichannel linear prediction, beamforming, Kronecker product filtering, recursive least-squares (RLS) algorithm.

1. INTRODUCTION

Reverberation adversely affects the intelligibility and quality of speech signals [1–5]. Therefore, effective and robust dereverberation methods are in high demand. In the past couple of decades, dereverberation has been extensively studied and numerous techniques have been developed [5–11]. Among those, the methods based on multichannel linear prediction (MCLP) have been intensively studied, where the desired signal is recovered by subtracting late reverberation components estimated using delayed prediction filters from the microphone signals [5]. This principle can be formulated either in the time domain or in the short-time Fourier transform (STFT) domain [12], resulting in different algorithms, among which the so-called weighted-prediction-error (WPE) algorithm has demonstrated great potential [13, 14].

While the WPE method exhibits promising performance for dereverberation, the computational complexity is high and, therefore, renders implementation in some real-time embedded or edge computing devices difficult. Another problem arises when the acoustic environment contains additive noise since it will affect the correlations between observation signals and, therefore, will degrade the dereverberation performance. For the purpose of joint denoising and dereverberation, various beamforming techniques, including differential microphone arrays (DMAs) [15], the generalized sidelobe canceller (GSC) [16, 17], the minimum variance distortionless response (MVDR) beamformer [18], and the weighted minimum power distortionless response (wMPDR) beamformer [19, 20] have been combined with MCLP-based dereverberation in a cascaded or a unified manner. However, while they effectively improve the dereverberation performance in a noisy environment, such combined methods are computationally even more expensive to implement.

In [21], a new framework is proposed to decompose the MCLP filter into a wMPDR beamformer and a temporal (linear prediction) filter. This decomposition is beneficial for adaptive processing, e.g., improving computational efficiency, which is a crucial factor to be considered in online processing applications [22–29]. In this paper, a bilinear form of the MCLP model is derived, which decomposes the MCLP model into a combination of beamformer and linear prediction filter. Various adaptive algorithms can be applied in the proposed framework, among which the recursive least-squares (RLS) algorithm is used in this work to derive the joint online dereverberation and noise reduction algorithm. Compared with the original MCLP based algorithm, the proposed method achieves better performance in computational efficiency, statistics tracking, and noise reduction ability.

2. SIGNAL MODEL

Consider an acoustic scenario where M microphones capture a single speech source in a reverberant and noisy environment. In the STFT domain, the signal received by the

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2018AAA0102200 and in part by the Key Program of National Science Foundation of China (NSFC) under Grants 61831019 and 62192713.

mth microphone can be denoted by $Y_{m,n,k}$ with n and k indexing the time frame and frequency bin, respectively. We describe the stacked microphone signal vector as $\mathbf{y}_{n,k} = [Y_{1,n,k} \ \cdots \ Y_{M,n,k}]^T \in \mathbb{C}^M$, which can be formulated as

$$\mathbf{y}_{n,k} = S_{n,k}\mathbf{d}_k + \mathbf{r}_{n,k} + \mathbf{v}_{n,k}, \tag{1}$$

where $S_{n,k}$ is the STFT coefficient of the desired signal at the reference microphone, $\mathbf{d}_k \in \mathbb{C}^M$ is the (assumed time-invariant) signal propagation vector corresponding to the desired signal at the reference microphone, and $\mathbf{r}_{n,k}$ and $\mathbf{v}_{n,k}$ are vectors of the reverberant components and the additive noise, respectively, defined analogously to $\mathbf{y}_{n,k}$.

Then, the task of joint dereverberation and noise reduction is to estimate $S_{n,k}$ from $\mathbf{y}_{n,k}$ in a blind manner while suppressing the late reverberation and noise. To simplify the notation in the following, we do not include the dependency on the frequency index k.

3. BILINEAR FORMS OF DEREVERBERATION

In this section, a dereverberation model expressed by a bilinear form is proposed, which combines a spatial filter and a temporal filter using the Kronecker product [30]. The desired signal is first estimated by applying a complex-valued beamforming filter, $\mathbf{h} \in \mathbb{C}^M$, to the observation signal vector, i.e.,

$$\begin{aligned} Z_{\mathbf{h},n} &= \mathbf{h}^H\mathbf{y}_n \\ &= \mathbf{h}^H\mathbf{d}S_n + \mathbf{h}^H(\mathbf{r}_n + \mathbf{v}_n), \end{aligned} \tag{2}$$

where the superscript ^{H} is the conjugate-transpose operator, and the distortionless constraint $\mathbf{h}^H\mathbf{d} = 1$ is needed.

Then, dereverberation is accomplished by subtracting the reverberant signal components estimated by a prediction filter of length L from the spatially filtered signal, i.e.,

$$\begin{aligned} \hat{S}_n &= Z_{\mathbf{h},n} - \sum_{l=\Delta}^{\Delta+L-1} G_l^* Z_{n-l} \\ &= Z_{\mathbf{h},n} - \mathbf{g}^H\mathbf{z}_{\mathbf{h},n-\Delta}, \end{aligned} \tag{3}$$

where $\{G_l\}_{\Delta}^{\Delta+L-1}$ is the prediction filter coefficient, the superscript * is the complex-conjugate operator, $\mathbf{g} = [G_\Delta \ \cdots \ G_{\Delta+L-1}]^T \in \mathbb{C}^L$ is the prediction filter, $\mathbf{z}_{\mathbf{h},n-\Delta} = [Z_{\mathbf{h},n-\Delta} \ \cdots \ Z_{\mathbf{h},n-\Delta-L+1}]^T \in \mathbb{C}^L$ contains beamforming output from the previous consecutive frames with $Z_{\mathbf{h},n-l} = \mathbf{h}^H\mathbf{y}_{n-l}$ and Δ is a prediction delay to avoid the removal of the correlation between the samples of the clean speech signals and prevent the excessive whitening problem. To further exploit the relationship between the beamformer and the dereverberation filter, we deduce that

$$\begin{aligned} \hat{S}_n &= Z_{\mathbf{h},n} - \sum_{l=\Delta}^{\Delta+L-1} G_l^*\left(\mathbf{h}^H\mathbf{y}_{n-l}\right) \\ &= Z_{\mathbf{h},n} - \mathbf{h}^H\mathbf{Y}_{n-\Delta}\mathbf{g}^*, \end{aligned} \tag{4}$$

where $\mathbf{Y}_{n-\Delta} = [\mathbf{y}_{n-\Delta} \ \cdots \ \mathbf{y}_{n-\Delta-L+1}] \in \mathbb{C}^{M\times L}$.

Obviously, the second term in (4) is bilinear in $\mathbf{h}^*$ and $\mathbf{g}^*$, i.e., for every fixed $\mathbf{h}^*$, it is a linear function of $\mathbf{g}^*$, and vice versa [31]. Moreover, we can write (4) as

$$\begin{aligned} \hat{S}_n &= Z_{\mathbf{h},n} - \operatorname{tr}\left(\mathbf{g}^*\mathbf{h}^H\mathbf{Y}_{n-\Delta}\right) \\ &= Z_{\mathbf{h},n} - \operatorname{vec}^H\left(\mathbf{h}\mathbf{g}^H\right)\operatorname{vec}\left(\mathbf{Y}_{n-\Delta}\right) \\ &= Z_{\mathbf{h},n} - (\mathbf{g}\otimes\mathbf{h})^H\,\bar{\mathbf{y}}_{n-\Delta}, \end{aligned} \tag{5}$$

where $\operatorname{tr}(\cdot)$ denotes the trace of a square matrix, $\operatorname{vec}(\cdot)$ is the vectorization operation which converts a matrix into a vector, $\otimes$ denotes the Kronecker product, and $\bar{\mathbf{y}}_{n-\Delta} = \operatorname{vec}(\mathbf{Y}_{n-\Delta}) \in \mathbb{C}^{ML}$.

It should be noted that these two linear filters can be decoupled for the following optimization process by [32]

$$\begin{aligned} \mathbf{g}\otimes\mathbf{h} &= (\mathbf{I}_L\otimes\mathbf{h})\,\mathbf{g} \qquad (6) \\ &= (\mathbf{g}\otimes\mathbf{I}_M)\,\mathbf{h}, \qquad (7) \end{aligned}$$

where $\mathbf{I}_M \in \mathbb{R}^{M\times M}$ and $\mathbf{I}_L \in \mathbb{R}^{L\times L}$ are the identity matrices. Therefore, by using the relationship in (6), the dereverberated signal in (5) can be written as

$$\begin{aligned} \hat{S}_n &= Z_{\mathbf{h},n} - \mathbf{g}^H(\mathbf{I}_L\otimes\mathbf{h})^H\,\bar{\mathbf{y}}_{n-\Delta} \\ &= Z_{\mathbf{h},n} - \mathbf{g}^H\bar{\mathbf{y}}_{[\mathbf{h}],n-\Delta}, \end{aligned} \tag{8}$$

where $\bar{\mathbf{y}}_{[\mathbf{h}],n-\Delta} = (\mathbf{I}_L\otimes\mathbf{h})^H\,\bar{\mathbf{y}}_{n-\Delta} \in \mathbb{C}^L$ is the observation signal vector filtered by $\mathbf{h}$.

Similarly, we can also write (5) based on (7) as

$$\begin{aligned} \hat{S}_n &= \mathbf{h}^H\mathbf{y}_n - \mathbf{h}^H(\mathbf{g}\otimes\mathbf{I}_M)^H\,\bar{\mathbf{y}}_{n-\Delta} \\ &= \mathbf{h}^H\bar{\mathbf{y}}_{[\mathbf{g}],n-\Delta}, \end{aligned} \tag{9}$$

where $\bar{\mathbf{y}}_{[\mathbf{g}],n-\Delta} = \mathbf{y}_n - (\mathbf{g}\otimes\mathbf{I}_M)^H\,\bar{\mathbf{y}}_{n-\Delta} \in \mathbb{C}^M$ is the observation signal vector filtered by $\mathbf{g}$.

As seen, the MCLP-based dereverberation problem can be reformulated as a problem of optimizing two shorter filters, whose coefficients are combined according to (5).

4. RLS-BASED ADAPTIVE DEREVERBERATION OF BILINEAR FORM

In this section, we derive an online dereverberation algorithm with the RLS algorithm [23] based on the proposed dereverberation model.

The two adaptive filters, $\mathbf{h}_n$ and $\mathbf{g}_n$, can be iteratively optimized by defining the following weighted cost functions:

$$\mathcal{J}\left[\mathbf{g}_n|\mathbf{h}_{n-1}\right] = \sum_{i=1}^{n}\alpha^{n-i}\frac{|Z_{[\mathbf{h}_{n-1}],i} - \mathbf{g}_n^H\bar{\mathbf{y}}_{[\mathbf{h}_{n-1}],i-\Delta}|^2}{\lambda_i}, \tag{10}$$

$$\mathcal{J}\left[\mathbf{h}_n|\mathbf{g}_{n-1}\right] = \sum_{i=1}^{n}\alpha^{n-i}\frac{|\mathbf{h}_n^H\bar{\mathbf{y}}_{[\mathbf{g}_{n-1}],i-\Delta}|^2}{\lambda_i}, \tag{11}$$

where $\lambda_i = |\hat{S}_i|^2$ is the variance of the a *priori* estimate of the desired signal, i.e., $\hat{S}_i = \mathbf{h}_{n-1}^H\bar{\mathbf{y}}_{[\mathbf{g}_{n-1}],i-\Delta}$, $i = 1,\ldots,n$, and α is the forgetting factor.

The solution for the temporal filter $\mathbf{g}_n$ can be obtained from the minimization of $\mathcal{J}\left[\mathbf{g}_n|\mathbf{h}_{n-1}\right]$. We get

$$\mathbf{g}_n = \mathbf{R}_{\mathbf{h},n}^{-1}\mathbf{p}_{\mathbf{h},n}, \tag{12}$$

where

$$\begin{aligned}\mathbf{R}_{\mathbf{h},n} &= \sum_{i=1}^{n} \alpha^{n-i}\frac{\bar{\mathbf{y}}_{[\mathbf{h}_{n-1}],i-\Delta}\bar{\mathbf{y}}_{[\mathbf{h}_{n-1}],i-\Delta}^{H}}{\lambda_i} \\ &= \alpha\mathbf{R}_{\mathbf{h},n-1} + \frac{\bar{\mathbf{y}}_{[\mathbf{h}_{n-1}],n-\Delta}\bar{\mathbf{y}}_{[\mathbf{h}_{n-1}],n-\Delta}^{H}}{\lambda_n}\end{aligned} \tag{13}$$

is the weighted covariance matrix of $\bar{\mathbf{y}}_{[\mathbf{h}_{n-1}],n-\Delta}$, and

$$\mathbf{p}_{\mathbf{h},n} = \sum_{i=1}^{n} \alpha^{n-i}\frac{\bar{\mathbf{y}}_{[\mathbf{h}_{n-1}],i-\Delta}Z_{[\mathbf{h}_{n-1}],i}^{*}}{\lambda_i} \tag{14}$$

is the weighted correlation vector between $\bar{\mathbf{y}}_{[\mathbf{h}_{n-1}],n-\Delta}$and $Z_{[\mathbf{h}_{n-1}],n}$.

The spatial beamformer can be optimized by minimizing $\mathcal{J}\left[\mathbf{h}_n|\mathbf{g}_{n-1}\right]$ with the distortionless constraint [20]:

$$\min_{\mathbf{h}_n} \mathcal{J}\left[\mathbf{h}_n|\mathbf{g}_{n-1}\right] \quad \text{s.t.} \quad \mathbf{h}_n^H\mathbf{d} = 1, \tag{15}$$

whose solution is the wMPDR beamformer [19,20]:

$$\mathbf{h}_n = \frac{\mathbf{R}_{\mathbf{g},n}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{R}_{\mathbf{g},n}^{-1}\mathbf{d}}, \tag{16}$$

with

$$\begin{aligned}\mathbf{R}_{\mathbf{g},n} &= \sum_{i=1}^{n} \alpha^{n-i}\frac{\bar{\mathbf{y}}_{[\mathbf{g}_{n-1}],i-\Delta}\bar{\mathbf{y}}_{[\mathbf{g}_{n-1}],i-\Delta}^{H}}{\lambda_i} \\ &= \alpha\mathbf{R}_{\mathbf{g},n-1} + \frac{\bar{\mathbf{y}}_{[\mathbf{g}_{n-1}],n-\Delta}\bar{\mathbf{y}}_{[\mathbf{g}_{n-1}],n-\Delta}^{H}}{\lambda_n}\end{aligned} \tag{17}$$

being the weighted covariance matrix of $\bar{\mathbf{y}}_{[\mathbf{g}_{n-1}],n-\Delta}$.

Using the matrix inversion lemma [33], the updates of $\mathbf{R}_{\mathbf{h},n}^{-1}$ and $\mathbf{R}_{\mathbf{g},n}^{-1}$ are obtained by

$$\mathbf{R}_{\mathbf{h},n}^{-1} = \frac{\mathbf{I}_L - \mathbf{k}_{\mathbf{h},n}\bar{\mathbf{y}}_{[\mathbf{h}_{n-1}],n-\Delta}^{H}}{\alpha}\mathbf{R}_{\mathbf{h},n-1}^{-1}, \tag{18}$$

$$\mathbf{R}_{\mathbf{g},n}^{-1} = \frac{\mathbf{I}_M - \mathbf{k}_{\mathbf{g},n}\bar{\mathbf{y}}_{[\mathbf{g}_{n-1}],n-\Delta}^{H}}{\alpha}\mathbf{R}_{\mathbf{g},n-1}^{-1}, \tag{19}$$

where

$$\mathbf{k}_{\mathbf{h},n} = \frac{\mathbf{R}_{\mathbf{h},n-1}^{-1}\bar{\mathbf{y}}_{[\mathbf{h}_{n-1}],n-\Delta}}{\alpha\lambda_n + \bar{\mathbf{y}}_{[\mathbf{h}_{n-1}],n-\Delta}^{H}\mathbf{R}_{\mathbf{h},n-1}^{-1}\bar{\mathbf{y}}_{[\mathbf{h}_{n-1}],n-\Delta}}, \tag{20}$$

$$\mathbf{k}_{\mathbf{g},n} = \frac{\mathbf{R}_{\mathbf{g},n-1}^{-1}\bar{\mathbf{y}}_{[\mathbf{g}_{n-1}],n-\Delta}}{\alpha\lambda_n + \bar{\mathbf{y}}_{[\mathbf{g}_{n-1}],n-\Delta}^{H}\mathbf{R}_{\mathbf{g},n-1}^{-1}\bar{\mathbf{y}}_{[\mathbf{g}_{n-1}],n-\Delta}} \tag{21}$$

are the Kalman gains.

Therefore, the temporal filter can be updated using the derived Kalman gain:

$$\mathbf{g}_n = \mathbf{g}_{n-1} + \mathbf{k}_{\mathbf{h},n}\hat{S}_n^*, \tag{22}$$

and the beamformer can be updated using (16), where the covariance matrix is inverted by (19) and (21). The RLS-based dereverberation algorithm of bilinear forms, which we term as "RLS-KP-WPE," is summarized in Algorithm 1.

Algorithm 1 The RLS-KP-WPE algorithm.

```
Initialization: $\mathbf{g}_0, \mathbf{h}_0, \mathbf{R}_{\mathbf{g},0}, \mathbf{R}_{\mathbf{y}_{\mathbf{h}},0}$
for $n = 1, 2, \ldots$ do
    $\bar{\mathbf{y}}_{[\mathbf{g}_{n-1}],n-\Delta} = \mathbf{y}_n - \left[\mathbf{g}_{n-1} \otimes \mathbf{I}_M\right]^H \bar{\mathbf{y}}_{n-\Delta}$
    $\bar{\mathbf{y}}_{[\mathbf{h}_{n-1}],n-\Delta} = \left[\mathbf{I}_L \otimes \mathbf{h}_{n-1}\right]^H \bar{\mathbf{y}}_{n-\Delta}$
    $\lambda_n = |\mathbf{h}_{n-1}^H\bar{\mathbf{y}}_{[\mathbf{g}_{n-1}],n-\Delta}|^2$
    $\mathbf{u}_{\mathbf{g},n} = \mathbf{R}_{\mathbf{g},n-1}^{-1}\bar{\mathbf{y}}_{[\mathbf{g}_{n-1}],n-\Delta}$
    $\mathbf{u}_{\mathbf{h},n} = \mathbf{R}_{\mathbf{h},n-1}^{-1}\bar{\mathbf{y}}_{[\mathbf{h}_{n-1}],n-\Delta}$
    $\mathbf{k}_{\mathbf{g},n} = \frac{\mathbf{u}_{\mathbf{g},n}}{\alpha\lambda_n+\bar{\mathbf{y}}_{[\mathbf{g}_{n-1}],n-\Delta}^H\mathbf{u}_{\mathbf{g},n}}$
    $\mathbf{k}_{\mathbf{h},n} = \frac{\mathbf{u}_{\mathbf{h},n}}{\alpha\lambda_n+\bar{\mathbf{y}}_{[\mathbf{h}_{n-1}],n-\Delta}^H\mathbf{u}_{\mathbf{h},n}}$
    $\mathbf{R}_{\mathbf{g},n}^{-1} = \frac{\mathbf{R}_{\mathbf{g},n-1}^{-1}-\mathbf{k}_{\mathbf{g},n}\mathbf{u}_{\mathbf{g},n}^H}{\alpha}$
    $\mathbf{R}_{\mathbf{h},n}^{-1} = \frac{\mathbf{R}_{\mathbf{h},n-1}^{-1}-\mathbf{k}_{\mathbf{h},n}\mathbf{u}_{\mathbf{h},n}^H}{\alpha}$
    $\mathbf{g}_n = \mathbf{g}_{n-1} + \mathbf{k}_{\mathbf{h},n}\hat{S}_n^*$
    $\mathbf{h}_n = \frac{\mathbf{R}_{\mathbf{g},n}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{R}_{\mathbf{g},n}^{-1}\mathbf{d}}$
end for
```

We also analyze the computational complexity of the proposed RLS-KP-WPE method and RLS-based WPE (RLS-WPE) method [34,35]. Table 1 shows the computational complexity of the RLS-KP-WPE and RLS-WPE methods in terms of the number of complex-valued multiplications. It can be seen that the computational complexity is reduced by a factor of approximately $(M^2L^2)/(M^2 + L^2)$ by the proposed method as compared to RLS-WPE.

Table 1. Computational complexity of RLS-KP-WPE and RLS-WPE methods.

	Number of multiplications
RLS-KP-WPE	$4M^2 + 3L^2 + 2ML + 5M + 3L + 8$
RLS-WPE	$3(ML)^2 + 4ML + 5$

5. SIMULATIONS

In this section, we study the performance of the proposed dereverberation method. The clean speech signals were taken from the TIMIT database with a sampling rate of 16 kHz, where the clean signals are concatenated such that the length of each signal exceeded 30 seconds. A small uniform linear array of 8 omnidirectional microphones with an interelement spacing of 2 cm is used. The source is placed at the endfire direction and 2 m away from the array center. The acoustic channel impulse responses from the source to the microphones are generated using the image model method with a room of size 6 m × 5 m × 4 m [36]. The dereverberation algorithm is implemented in the STFT domain. The observation signals are divided into overlapping frames of 512 samples using 75% overlap using a Kaiser window with a window-shape parameter of 1.9π. The evaluation is performed in two

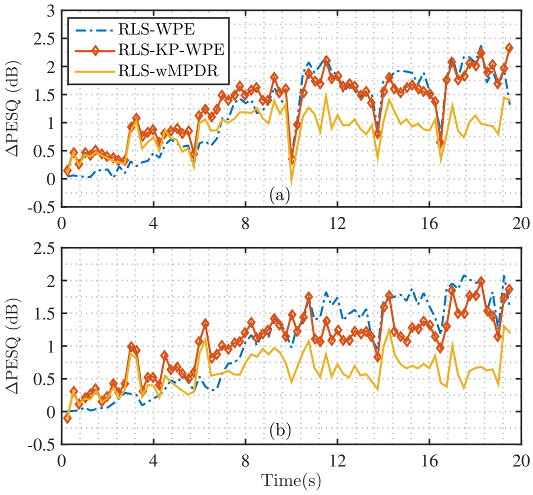

Fig. 1. Segmental performance, ΔPESQ, of the RLS-WPE, RLS-KP-WPE, and RLS-wMPDR in a reverberant and noise-free environment: (a) REVB1 and (b) REVB2.

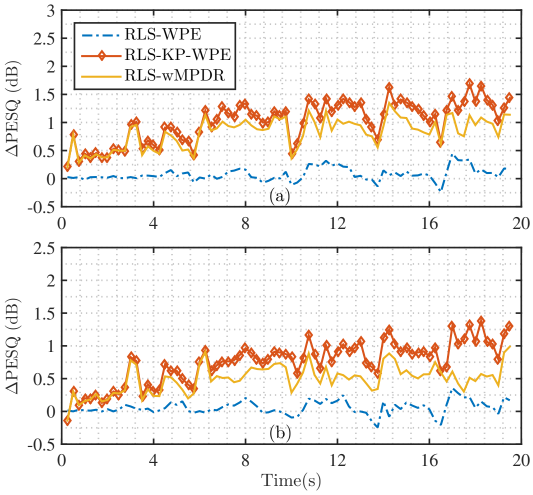

Fig. 2. Segmental performance, ΔPESQ, of the RLS-WPE, RLS-KP-WPE, and RLS-wMPDR in reverberant and noisy environment (SNR = 20 dB): (a) REVB1 and (b) REVB2.

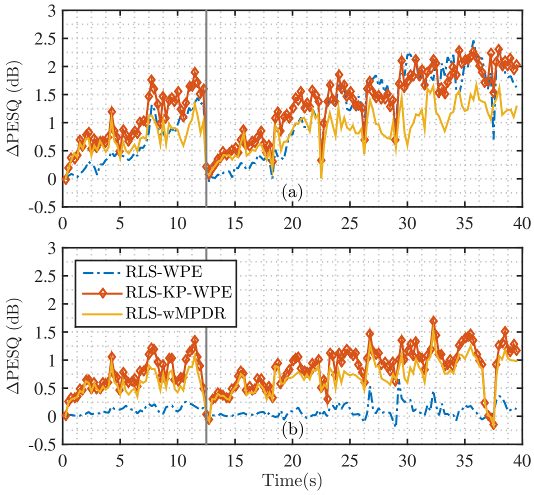

Fig. 3. Segmental performance, ΔPESQ, of the RLS-WPE, RLS-KP-WPE, and RLS-wMPDR for REVB1 in noise free and noisy environment: (a) noise-free and (b) noisy with SNR = 20 dB.

different reverberation conditions with the reverberation time T_{60} being 300 ms and 400 ms, which are labeled as REVB1 and REVB2, respectively. The background noise is diffuse noise generated according to [37]. We compare the performance of RLS-WPE, RLS-KP-WPE, and RLS-based wMPDR (RLS-wMPDR) in terms of the perceptual evaluation of speech quality (PESQ) [38, 39]. In our implementation, the prediction delay was set as $D = 2$. For REVB1 and REVB2, the length of the prediction filters are set as $L \in \{12, 10, 8\}$ and $L \in \{16, 14, 12\}$, respectively, for frequency ranges from 0 to 1, 1 to 3 and 3 to 8 kHz. The temporal filter and the spatial filter are initialized as a zero vector and a delay-and-sum filter, i.e., $\mathbf{g}_0 = [0 \ 0 \ \cdots \ 0]^T$ and $\mathbf{h}_0 = \mathbf{d}/M$, respectively. Since the source is placed at the endfire direction and the spacing between the two microphones is known, the steering vector $\mathbf{d}$ was calculated accordingly (as a free-field steering vector). The covariance matrices are initialized as $\mathbf{R}_{\mathbf{g},0} = \delta_{\mathbf{g}}\mathbf{I}_M$ and $\mathbf{R}_{\mathbf{h},0} = \delta_{\mathbf{h}}\mathbf{I}_L$, with $\delta_{\mathbf{h}} = 10^{-4}$ and $\delta_{\mathbf{g}} = 10^{-2}$. The forgetting factor is set as $\alpha = 0.993$.

To assess how the performance varies with time, we divide the signal into overlapping segments (each segment is 1 s long and the overlapping rate is 50%) and evaluate the performance for each segment. Figure 1 presents the segmental performance improvement in PESQ, i.e., ΔPESQ, of the three studied methods in a reverberant and noise-free environment. The RLS-KP-WPE and RLS-wMPDR methods achieve better performance than the RLS-WPE in the first few seconds. After convergence, the RLS-KP-WPE and RLS-WPE have a better performance than the RLS-wMPDR method. Figure 2 presents the ΔPESQ of the three studied methods in a reverberant and noisy environment with an SNR level of 20 dB. It is seen that the RLS-KP-WPE method performs better than other methods in the presence of noise.

To further highlight the advantage of the RLS-KP-WPE, we add another set of simulations where the position of the source signal changes abruptly to the opposite direction at 12.5 seconds (assume the time of the position change is known). Figure 3 plots the ΔPESQ of the three studied methods under reverberant only and reverberant-plus-noise environments. It is seen that the RLS-KP-WPE achieves the best performance in all cases.

6. CONCLUSIONS

This paper presents a bilinear framework for adaptive speech dereverberation by combining beamforming and linear prediction. In such a framework, the MCLP filter is expressed as a Kronecker product of a spatial filter and a temporal filter. Based on this formulation, an iterative RLS-based algorithm is derived for speech dereverberation. Compared with the original MCLP-based adaptive WPE algorithm, the presented method exhibits better dereverberation performance and robustness to additive noise and involves a much lower computational complexity.

7. REFERENCES

[1] T. Yoshioka, *et al.*, "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.

[2] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[3] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin, "Variational Bayesian inference for multichannel dereverberation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 8, pp. 1320–1335, Aug. 2014.

[4] E. A. Habets and P. A. Naylor, "Dereverberation," *Audio Source Separation and Speech Enhancement*, pp. 317–343, 2018.

[5] 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, Sept. 2010.

[6] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 1006–1018, Jun. 2015.

[7] G. Huang, J. Benesty, and J. Chen, "On the design of frequency-invariant beampatterns with uniform circular microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 1140–1153, 2017.

[8] H. W. Löllmann, A. Brendel, and W. Kellermann, "Generalized coherence-based signal enhancement," in *Proc. IEEE ICASSP*, 2020, pp. 201–205.

[9] G. Huang, J. Chen, and J. Benesty, "Insights into frequency-invariant beamforming with concentric circular microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 12, pp. 2305–2318, Dec. 2018.

[10] R. Ikeshita, K. Kinoshita, N. Kamo, and T. Nakatani, "Online speech dereverberation using mixture of multichannel linear prediction models," *IEEE Signal Process. Lett.*, vol. 28, pp. 1580–1584, 2021.

[11] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Blind and neural network-guided convolutional beamformer for joint denoising, dereverberation, and source separation," in *Proc. IEEE ICASSP*, 2021, pp. 6129–6133.

[12] S. Braun and E. A. Habets, "Linear prediction-based online dereverberation and noise reduction using alternating Kalman filters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1115–1125, Jun. 2018.

[13] 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 *Proc. IEEE ICASSP*, 2008, pp. 85–88.

[14] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[15] W. Yang, G. Huang, W. Zhang, J. Chen, and J. Benesty, "Dereverberation with differential microphone arrays and the weighted-prediction-error method," in *Proc. IEEE IWAENC*, 2018, pp. 376–380.

[16] T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. Van Waterschoot, "Comparative analysis of generalized sidelobe cancellation and multi-channel linear prediction for speech dereverberation and noise reduction," in *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 544–558, Mar. 2018.

[17] T. Dietzen, S. Doclo, A. Spriet, W. Tirry, M. Moonen, and T. van Water-choot, "Low-complexity Kalman filter for multi-channel linear-prediction-based blind speech dereverberation," in *Proc. IEEE WASPAA*, 2017, pp. 1–5.

[18] A. Cohen, G. Stemmer, S. Ingalsuo, and S. Markovich-Golan, "Combined weighted prediction error and minimum variance distortionless response for dereverberation," in *Proc. IEEE ICASSP*, 2017, pp. 446–450.

[19] C. Boeddeker, T. Nakatani, K. Kinoshita, and R. Haeb-Umbach, "Jointly optimal dereverberation and beamforming," in *Proc. IEEE ICASSP*, 2020, pp. 216–220.

[20] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 903–907, 2019.

[21] W. Yang, G. Huang, J. Chen, J. Benesty, I. Cohen, and W. Kellermann, "Robust dereverberation with Kronecker product based multichannel linear prediction," *IEEE Signal Process. Lett.*, pp. 101–105, 2020.

[22] X. Wang, G. Huang, I. Cohen, J. Benesty, and J. Chen, "Kronecker product adaptive beamforming for microphone arrays," *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp.49–54.

[23] C. Elisei-Iliescu, C. Paleologu, J. Benesty, C. Stanciu, C. Anghel, and S. Ciochină, "Recursive least-squares algorithms for the identification of low-rank systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 5, pp. 903–918, May. 2019.

[24] C. Paleologu, J. Benesty, and S. Ciochină, "Linear system identification based on a Kronecker product decomposition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 1793–1809, Oct. 2018.

[25] J. Benesty, I. Cohen, and J. Chen, *Array Processing–Kronecker Product Beamforming*. Berlin, Germany: Springer-Verlag, 2019.

[26] I. Cohen, J. Benesty, and J. Chen, "Differential Kronecker product beamforming," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 892–902, May. 2019.

[27] W. Yang, G. Huang, J. Benesty, I. Cohen, and J. Chen, "On the design of flexible Kronecker product beamformers with linear microphone arrays," in *Proc. IEEE ICASSP*, 2019, pp. 441–445.

[28] G. Huang, J. Benesty, J. Chen, and I. Cohen, "Robust and steerable Kronecker product differential beamforming with rectangular microphone arrays," in *Proc. IEEE ICASSP*, 2020, pp. 211–215.

[29] G. Huang, J. Benesty, I. Cohen, and J. Chen, "Kronecker product multichannel linear filtering for adaptive weighted prediction error-based speech dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1277–1289, Mar. 2022.

[30] J. Benesty, C. Paleologu, L.-M. Dogariu and S. Ciochină, "Identification of linear and bilinear systems: A unified study," *Electronics*, vol. 10, no. 15(33 pages), Jul. 2021.

[31] J. Benesty, C. Paleologu and S. Ciochină, "On the identification of bilinear forms with the Wiener filter," *IEEE Signal Process. Lett.*, vol. 24, pp. 653–657, 2017.

[32] D. A. Harville, "Matrix algebra from a statistician's perspective," New York: Springer-Verlag,1997.

[33] S. Haykin, Adaptive filter theory. 4th ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.

[34] T. Yoshioka, "Speech enhancement in reverberant environments," Kyoto University, 2010.

[35] T. Xiang, J. Lu, and K. Chen, "Multi-channel adaptive dereverberation robust to abrupt change of target speaker position," *J. Acoust. Soc. Am.*, vol. 145, no. 3, pp. EL250–EL256, 2019.

[36] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[37] E. A. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Am.*, vol. 122, no. 6, pp. 3464–3470, 2007.

[38] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2007.

[39] K. Kinoshita, *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adva. Signal Process.*, vol. 2016, no. 1, p. 7, Jan. 2016.