Robust Dereverberation with Kronecker Product Based Multichannel Linear Prediction

Wenxing Yang, Gongping Huang, Jingdong Chen, Jacob Benesty, Israel Cohen, and Walter Kellermann

***Abstract*—Reverberation impairs not only the speech quality, but also intelligibility. The weighted-prediction-error (WPE) method, which estimates the late reverberation component based on a multichannel linear predictor, is by far one of the most effective algorithms for dereverberation. Generally, the WPE prediction filter in every short-time-Fourier-transform (STFT) subband has to be long enough to estimate accurately the late reverberation component. As a consequence, WPE is computationally expensive, which makes it difficult to implement into real-time embedded or edge computing devices. Moreover, WPE is sensitive to additive noise and its performance may suffer from dramatic degradation even in environments where the signal-to-noise ratio (SNR) is high. To address these drawbacks, this paper proposes to decompose the multichannel linear prediction filter as a Kronecker product of a temporal (interframe) prediction filter and a spatial filter. An iterative algorithm is then developed to optimize the two filters. In comparison with the original WPE algorithm, the presented method not only exhibits better performance in terms of dereverberation and robustness to additive noise, as there are fewer parameters to estimate for a given number of observation signal samples, but is also computationally more efficient, since the dimensions of the covariance matrices after Kronecker product decomposition are smaller.**

***Index Terms*—Dereverberation, noise robustness, weighted-prediction-error, beamforming, Kronecker product filter, speech enhancement.**

I. Introduction

Reverberation, which is the result of reflections of sound waves from surfaces and boundaries, typically impairs speech quality and intelligibility; and, therefore, affects significantly the performance of speech communication and human-machine speech interfaces [1]–[7]. Dereverberation, as its name indicates, is a process to mitigate the detrimental effect of reverberation. It has been intensively studied over the last couple of decades and numerous methods have been developed [8]–[10]. Broadly, those methods can be classified into four basic categories: channel equalization based methods [11], suppression based methods [12]–[14], beamforming [15], [16], and linear prediction based approaches [6], [7]. In applications with long acoustic impulse responses and large reverberation time, the most promising methods so far are the ones based on multichannel linear prediction, which first estimate the reverberation component due to late reflections with a delayed linear prediction filter and then subtract it from the observation signals [6], [7]. This principle can be formulated either in the time domain or in the short-time Fourier transform (STFT) domain [17], [18], resulting in different techniques, among which the so-called weighted-prediction-error (WPE) algorithm (implemented in the STFT domain) has demonstrated a great potential [19], [20].

While it exhibits promising performance in dereverberation, the WPE algorithm is found sensitive to additive noise. Moreover, the computational complexity of this algorithm is quite high, making it difficult to implement into embedded and edge computing devices. A great deal of efforts have been devoted to circumventing these drawbacks. For example, methods are proposed to combine WPE with differential microphone arrays (DMAs) [21], the generalized sidelobe canceler (GSC) [22], [23], the minimum variance distortionless constraint (MVDR) beamformer [24], and the weighted minimum power distortionless response (wMPDR) beamformer [25], [26] to improve the robustness of WPE with respect to additive noise. While it improves the resilience of WPE to additive noise, the existing methods of incorporating beamforming into WPE increase the computational complexity.

In this paper, we propose a new framework for dereverberation by expressing the multichannel linear prediction filter as a Kronecker product of a temporal (interframe) filter and a spatial filter, and the lengths of the two filters correspond, respectively, to the order of the prediction filter and the number of microphones. This is inspired by the recent works of using Kronecker product decomposition in system identification and beamforming [27]–[31], which demonstrated attractive properties in terms of performance, computational efficiency, and flexibility. The decomposition enables to reformulate the original optimization problem as two optimization sub-problems, and an iterative algorithm is then developed to optimize the two sub-filters. In this work, the spatial filter is optimized as a wMPDR beamformer, which is considered to be effective in reducing both background noise and reverberation, so the global filter can be more robust to noise. With the Kronecker product decomposition, the dereverberation problem is dealt with covariance matrices of much lower dimensions and, there-

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) Grant No. 61831019 and the NSFC and Israel Science Foundation (ISF) joint research program under Grant No. 61761146001 and 2514/17.

W. Yang is with the CIAIC, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China, and also with LMS, University Erlangen-Nuremberg, 91058 Erlangen, Germany (e-mail: yangwenxing521@mail.nwpu.edu.cn).

J. Chen is with the CIAIC, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China (e-mail: jingdongchen@ieee.org).

G. Huang and I. Cohen are with Andrew and Erna Viterby Faculty of Electrical Engineering, Technion–Israel Institute of Technology, Technion City, Haifa 3200003, Israel (e-mail: gongping@campus.technion.ac.il; icohen@ee.technion.ac.il).

J. Benesty is with INRS-EMT, University of Quebec, 800 de la Gauchetiere Ouest, Montreal, QC H5A 1K6, Canada (e-mail: benesty@emt.inrs.ca).

W. Kellermann is with LMS, University Erlangen-Nuremberg, 91058 Erlangen, Germany (e-mail: walter.kellermann@fau.de).

fore, fewer computations are needed to estimate the covariance matrices and their inverses, which makes the proposed method computationally more efficient than WPE.

II. Signal Model and Problem Formulation

We consider a signal model in which an array with M microphones captures a convolved source signal in some noise field. The received signal at the mth ($m = 1, 2, \ldots, M$) microphone is expressed as

$$y_m(k) = h_m(k) * s(k) + v_m(k), \tag{1}$$

where k is the discrete-time index, $h_m(k)$ is the acoustic impulse response from the unknown speech source, $s(k)$, to the mth microphone, $*$ stands for the linear convolution, and $v_m(k)$ is additive noise at the mth microphone. The convolved speech signals are coherent across the microphones. We assume that the convolved speech and noise signals are uncorrelated, zero mean, real valued, and broadband.

The signal model (1) can be rewritten in the STFT domain as [7], [32]

$$Y_m(n,\omega) = \sum_{l=0}^{L_h-1} H_m(l,\omega) S(n-l,\omega) + V_m(n,\omega), \tag{2}$$

where n is the time-frame index, ω denotes the angular frequency, $H_m(l,\omega)$ of length L_h is the counterpart of $h_m(k)$ in the STFT domain, and $Y_m(n,\omega)$, $S(n,\omega)$, and $V_m(n,\omega)$ are the STFTs of $y_m(k)$, $s(k)$, and $v_m(k)$, respectively. To simplify the notation, we drop the dependence on ω in the rest of this paper.

The multichannel linear prediction dereverberation process consists of estimating the late reflection component from the past L consecutive frames and then subtracting it from the observations to get an estimate of the direct path plus early reflections. Mathematically, this process is expressed as

$$\widehat{S}(n) = Y_m(n) - \mathbf{g}^H \bar{\mathbf{y}}(n-D), \tag{3}$$

where $\bar{\mathbf{y}}(n-D) = \left[\mathbf{y}^T(n-D) \ \cdots \ \mathbf{y}^T(n-D-L+1)\right]^T$ is the stacked observation signal vector of length ML, $\mathbf{y}(n-D-l) = [Y_1(n-D-l) \ \cdots \ Y_M(n-D-l)]^T$ for $l = 0, 1, 2, \ldots L-1$ is the observation signal vector of length M at time frame $n-D-l$, $\mathbf{g}$ is the prediction filter of length ML, the superscript ^{T} is the transpose operator, the superscript ^{H} is the conjugate-transpose operator, and D is a predefined delay to avoid removing the correlation between clean speech signals and prevent the excessive whitening problem.

Now, the problem of dereverberation becomes one of finding the optimal prediction filter $\mathbf{g}$. In the WPE method [19], [20], the clean speech component in the STFT domain is modeled as a complex Gaussian random variable with zero mean and time-varying variance so that the cost function is defined as [7]

$$J(\mathbf{g}) = \sum_{n=1}^{N} \frac{\left|Y_m(n) - \mathbf{g}^H \bar{\mathbf{y}}(n-D)\right|^2}{\lambda(n)}, \tag{4}$$

where N is the number of time frames and $\lambda(n) = |\widehat{S}(n)|^2$ is the variance of the estimated desired speech signal. By minimizing the cost function, the solution to the multichannel linear prediction filter $\mathbf{g}$ is obtained as [7]

$$\mathbf{g} = \mathbf{R}_{\bar{\mathbf{y}}}^{-1} \mathbf{p}_{\bar{\mathbf{y}}}, \tag{5}$$

where

$$\mathbf{R}_{\bar{\mathbf{y}}} = \sum_{n=1}^{N} \frac{\bar{\mathbf{y}}(n-D)\bar{\mathbf{y}}^H(n-D)}{\lambda(n)},$$

$$\mathbf{p}_{\bar{\mathbf{y}}} = \sum_{n=1}^{N} \frac{\bar{\mathbf{y}}(n-D)Y_m^*(n)}{\lambda(n)}$$

are the weighted covariance matrix of size $ML \times ML$ and the weighted covariance vector of length ML, respectively, and the superscript * is the complex-conjugate operator. In practice, the estimation of $S(n)$ and $\mathbf{g}$ follows an iterative process, i.e., with an initial setup of the filter $\mathbf{g}$, the prediction filter in (5) and the desired speech component in (3) are updated repeatedly till convergence.

III. Kronecker Product Multichannel Prediction

The WPE method, though demonstrated promising performance, is computationally expensive and sensitive to additive noise [21]. Consequently, new methods to make WPE more robust to noise and computationally efficient are highly desirable, one of which is presented in the following.

A. Kronecker Product Filter

In this work, we propose the following Kronecker product multichannel linear prediction model:

$$\widehat{S}(n) = \mathbf{g}_2^H \mathbf{y}(n) - (\mathbf{g}_1 \otimes \mathbf{g}_2)^H \bar{\mathbf{y}}(n-D), \tag{6}$$

where $\otimes$ denotes the Kronecker product, and $\mathbf{g}_1$ and $\mathbf{g}_2$ are two sub-filters of length L and M, respectively. Now the original prediction filter is decomposed as a Kronecker product of two short sub-filters (one is a temporal filter and the other a spatial filter), which deal with the interframe correlation and spatial correlation, respectively. Note that the lengths of the filters $\mathbf{g}_1$ and $\mathbf{g}_2$ can be much smaller than ML and, as a result, this reformulation of the multichannel linear prediction filter can help reduce the computational complexity, which will become clear soon. Moreover, this new model has the potential to better deal with additive noise since it inherits the property of the conventional beamforming technique to reduce both noise and reverberation without affecting the correlation between consecutive frames of the observations.

We use the following relationships of the Kronecker product [33]:

$$\begin{aligned} \mathbf{g}_1 \otimes \mathbf{g}_2 &= (\mathbf{I}_L \otimes \mathbf{g}_2)\mathbf{g}_1 \ =: \mathbf{G}_2\mathbf{g}_1 \\ &= (\mathbf{g}_1 \otimes \mathbf{I}_M)\mathbf{g}_2 =: \mathbf{G}_1\mathbf{g}_2, \end{aligned} \tag{7}$$

where $\mathbf{I}_L$ and $\mathbf{I}_M$ are the identity matrices of sizes $L \times L$ and $M \times M$, respectively, and $\mathbf{G}_2 = \mathbf{I}_L \otimes \mathbf{g}_2$ and $\mathbf{G}_1 = \mathbf{g}_1 \otimes \mathbf{I}_M$ are matrices of sizes $ML \times L$ and $ML \times M$, respectively. Substituting (7) into (6), the dereverberated signal can be written as

$$\begin{aligned} \widehat{S}(n) &= \mathbf{g}_2^H \mathbf{y}(n) - \mathbf{g}_1^H \mathbf{G}_2^H \bar{\mathbf{y}}(n-D) \\ &= \widehat{Y}(n) - \mathbf{g}_1^H \bar{\mathbf{y}}_{\mathbf{G}_2}(n-D), \end{aligned} \tag{8}$$

TABLE I
THE PROPOSED KP-WPE ALGORITHM.

Initialize: $\mathbf{g}_1,\ \mathbf{g}_2$	
Repeat	
Step 1:	Calculate the estimated speech signal $\widehat{S}(n)$ using (6), and its variance $\lambda(n)$ for all the N time frames.
Step 2:	Calculate the two partial covariance matrices $\mathbf{R}_{\bar{\mathbf{y}}\|\mathbf{G_2}}$ and $\mathbf{R}_{\bar{\mathbf{y}}\|\mathbf{G_1}}$ using (13) and (15), and the partial covariance vector $\mathbf{p}_{\bar{\mathbf{y}}\|\mathbf{G}_2}$ using (12).
Step 3:	Calculate the two filters $\mathbf{g}_2$ and $\mathbf{g}_1$ using (17) and (19).
Until convergence	

where $\widehat{Y}(n)=\mathbf{g}_2^H\mathbf{y}(n)$ and $\bar{\mathbf{y}}_{\mathbf{G}_2}(n-D)=\mathbf{G}_2^H\bar{\mathbf{y}}(n-D)$.

Alternately, one can rewrite (6) using the relationship in (7) as

$$\begin{aligned}\widehat{S}(n)&=\mathbf{g}_2^H\mathbf{y}(n)-\mathbf{g}_2^H\mathbf{G}_1^H\bar{\mathbf{y}}(n-D)\\&=\mathbf{g}_2^H\mathbf{y}(n)-\mathbf{g}_2^H\bar{\mathbf{y}}_{\mathbf{G}_1}(n-D)=\mathbf{g}_2^H\widetilde{\mathbf{y}}_{\mathbf{G}_1}(n-D),\end{aligned}\quad(9)$$

where $\bar{\mathbf{y}}_{\mathbf{G}_1}(n-D)=\mathbf{G}_1^H\bar{\mathbf{y}}(n-D)$ and $\widetilde{\mathbf{y}}_{\mathbf{G}_1}(n-D)=\mathbf{y}(n)-\bar{\mathbf{y}}_{\mathbf{G}_1}(n-D)$.

B. Optimal Sub-Filters

Following the above formulation, the problem of dereverberation becomes one of finding the two optimal filters, $\mathbf{g}_1$ and $\mathbf{g}_2$. We define the cost function as

$$\mathcal{J}(\mathbf{g}_1,\mathbf{g}_2)=\sum_{n=1}^{N}\frac{\left|\mathbf{g}_2^H\mathbf{y}(n)-(\mathbf{g}_1\otimes\mathbf{g}_2)^H\bar{\mathbf{y}}(n-D)\right|^2}{\lambda(n)}.\quad(10)$$

Using (8), we can write the above cost function as

$$\begin{aligned}\mathcal{J}(\mathbf{g}_1|\mathbf{g}_2)&=\sum_{n=1}^{N}\frac{\left|\widehat{Y}(n)-\mathbf{g}_1^H\bar{\mathbf{y}}_{\mathbf{G}_2}(n-D)\right|^2}{\lambda(n)}\\&=\phi_{\widehat{Y}}-2\Re(\mathbf{g}_1^H\mathbf{p}_{\bar{\mathbf{y}}|\mathbf{G_2}})+\mathbf{g}_1^H\mathbf{R}_{\bar{\mathbf{y}}|\mathbf{G_2}}\mathbf{g}_1,\end{aligned}\quad(11)$$

where $\phi_{\widehat{Y}}=\sum_{n=1}^{N}\frac{|\widehat{Y}(n)|^2}{\lambda(n)}$ is the weighted variance of $\widehat{Y}(n)$,

$$\mathbf{p}_{\bar{\mathbf{y}}|\mathbf{G}_2}=\sum_{n=1}^{N}\frac{\bar{\mathbf{y}}_{\mathbf{G_2}}(n-D)\widehat{Y}^*(n)}{\lambda(n)},\quad(12)$$

$$\mathbf{R}_{\bar{\mathbf{y}}|\mathbf{G_2}}=\sum_{n=1}^{N}\frac{\bar{\mathbf{y}}_{\mathbf{G}_2}(n-D)\bar{\mathbf{y}}_{\mathbf{G_2}}^H(n-D)}{\lambda(n)}\quad(13)$$

are the partial covariance vector and partial covariance matrix of sizes $L\times 1$ and $L\times L$, respectively, and $\Re(\cdot)$ denotes the real part of a complex number.

Alternately, using (9), one can write the cost function as

$$\mathcal{J}(\mathbf{g}_2|\mathbf{g}_1)=\sum_{n=1}^{N}\frac{\left|\mathbf{g}_2^H\widetilde{\mathbf{y}}_{\mathbf{G}_1}(n-D)\right|^2}{\lambda(n)}=\mathbf{g}_2^H\mathbf{R}_{\widetilde{\mathbf{y}}|\mathbf{G_1}}\mathbf{g}_2,\quad(14)$$

where

$$\mathbf{R}_{\widetilde{\mathbf{y}}|\mathbf{G_1}}=\sum_{n=1}^{N}\frac{\widetilde{\mathbf{y}}_{\mathbf{G}_1}(n-D)\widetilde{\mathbf{y}}_{\mathbf{G_1}}^H(n-D)}{\lambda(n)}\quad(15)$$

is the partial covariance matrix of size $M\times M$.

Now, we present an iterative method to estimate $\mathbf{g}_1$ and $\mathbf{g}_2$. Suppose that we have an initial estimate of $\mathbf{g}_1$, denoted $\mathbf{g}_1^{(0)}$. The spatial filter $\mathbf{g}_2$ can be optimized from different perspectives. Here, we compute it by minimizing the cost function in (14) with a distortionless constraint at the desired look direction. Therefore, the partial optimization problem for this filter (which can be viewed as a beamformer) can be formulated as

$$\min_{\mathbf{g}_2}\mathcal{J}(\mathbf{g}_2|\mathbf{g}_1^{(0)})\ \text{ s. t. }\ \mathbf{g}_2^H\mathbf{d}=1,\quad(16)$$

where $\mathbf{d}$ is the steering-like vector corresponding to the desired look direction. The solution to this problem is

$$\mathbf{g}_2^{(1)}=\frac{\mathbf{R}_{\widetilde{\mathbf{y}}|\mathbf{G}_1}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{R}_{\widetilde{\mathbf{y}}|\mathbf{G}_1}^{-1}\mathbf{d}},\quad(17)$$

which is the so-called wMPDR beamformer [25], [26]. Similarly, with the value of $\mathbf{g}_2^{(0)}$, the dereverberation filter $\mathbf{g}_1$ can be derived by minimizing the partial cost function defined in (10), i.e.,

$$\min_{\mathbf{g}_1}\mathcal{J}(\mathbf{g}_1|\mathbf{g}_2^{(0)}),\quad(18)$$

whose solution is

$$\mathbf{g}_1^{(1)}=\mathbf{R}_{\bar{\mathbf{y}}|\mathbf{G}_2}^{-1}\mathbf{p}_{\bar{\mathbf{y}}|\mathbf{G}_2}.\quad(19)$$

By repeating the iterations alternatively $\mathcal{I}$ times, we get the solution for the subfilters, $\mathbf{g}_1=\mathbf{g}_1^{(\mathcal{I})}$ and $\mathbf{g}_2=\mathbf{g}_2^{(\mathcal{I})}$, and the dereverberated speech $\widehat{S}(n)$ is obtained according to (6). This Kronecker product based WPE (KP-WPE) algorithm is summarized in Table I.

IV. PERFORMANCE STUDY

A. Computational Complexity Analysis

We first analyze the computational complexity of the proposed KP-WPE method and the conventional WPE method [7]. Table II shows the computational complexity for the three steps of the KP-WPE and WPE methods, where the three steps, as shown in Table I, are prediction filtering, estimation of the covariance matrices and vector, and computation of the prediction filter. The complexity reduction factor is defined as the ratio between the amount of computation required for KP-WPE and that for WPE. Notice that the KP-WPE is computationally slightly more expensive than WPE in Step 1. However, for the two other steps, KP-WPE is much more computationally efficient than WPE. As seen in Table II, the complexity reduction factors are approximately $(M+L)/ML$ and $(M^2+L^2)/M^2L^2$ (in terms of complex-valued multiplications) for Steps 2 and 3 respectively, and a reduction factor of $(M^3+L^3)/M^3L^3$ can be achieved in Step 3 for matrix inversion. This is mainly because the matrices $\mathbf{R}_{\bar{\mathbf{y}}|\mathbf{G_2}}$ and $\mathbf{R}_{\widetilde{\mathbf{y}}|\mathbf{G_1}}$ are much smaller than $\mathbf{R}_{\bar{\mathbf{y}}}$. Table III presents the values of $\beta_{(\times)}$ in Step 2 and $\beta_{(/)}$ in Step 3 with $M=8$ for different values of L, where $L\in\{5,10,15,20,25\}$. It is seen that the complexity reduction factor is much smaller than 1 and decreases as the value of L increases. Therefore, one can conclude that KP-WPE is much more computationally efficient than WPE.

TABLE II
COMPUTATIONAL COMPLEXITY OF KP-WPE AND WPE. $(\times)$, $(+)$, AND $(\div)$ DENOTE MULTIPLICATION, ADDITION, AND DIVISION OF COMPLEX-VALUED NUMBERS, RESPECTIVELY, AND $(/)$ DENOTES MATRIX INVERSION.

KP-WPE				
Step	$(\times)$	$(+)$	$(\div)$	$(/)$
1	$N(2ML+M+1)$	$N(ML+M-1)$	–	–
2	$N(M^2L+ML^2+M^2+L^2+L)+M^2L+ML^2$	$N(M^2L+ML^2+M^2+L^2)$	$N(M+L+1)$	–
3	M^2+L^2+M	M^2+L^2-L-1	1	$\mathcal{O}(M^3+L^3)$
WPE				
Step	$(\times)$	$(+)$	$(\div)$	$(/)$
1	$N(ML+1)$	$N \cdot ML$	–	–
2	$N(M^2L^2+ML)$	$N(M^2L^2+ML)$	$N(ML+1)$	–
3	M^2L^2	M^2L^2-ML	–	$\mathcal{O}(M^3L^3)$
Complexity Reduction Factor				
Step	$\beta_{(\times)} \approx$	$\beta_{(+)} \approx$	$\beta_{(\div)} \approx$	$\beta_{(/)} \approx$
1	2	1	–	–
2	$(M+L)/ML$	$(M+L)/ML$	$(M+L)/ML$	–
3	$(M^2+L^2)/M^2L^2$	$(M^2+L^2)/(M^2L^2-ML)$	–	$(M^3+L^3)/M^3L^3$

TABLE III
VALUES OF $\beta_{(\times)}$ IN STEP 2 AND $\beta_{(/)}$ IN STEP 3 WITH FIXED $M=8$ FOR DIFFERENT VALUE OF L.

L	5	10	15	20	25
$\beta_{(\times)}$in step 2	0.3250	0.2250	0.1917	0.1750	0.1650
$\beta_{(/)}$ in step 3	0.0100	0.0030	0.0022	0.0021	0.0020

B. Simulations

The clean source speech signals used in the simulations are from the TIMIT database, where they are recorded with a sampling rate of 16 kHz. The acoustic channel impulse responses from the source to the microphones are generated using the image method with a room of size $7 \times 7 \times 4$ (in meters) [34]. A uniform linear array of 8 omnidirectional microphones with an inter-element spacing of 8 cm is used. The source is placed at the endfire direction and 2 m away from the array center. The dereverberation algorithm is implemented in the STFT domain, where the observation signals are divided into overlapping frames of 512 samples with 75% overlap using a Kaiser window. The evaluation is performed in two different reverberation conditions with the reverberation time T_{60} being 300 ms and 210 ms, which are labeled as REVB1 and REVB2, respectively. The background noise is diffuse noise [35]. We compare the performance of WPE, wMPDR, wMPDR followed by single-channel WPE (wMPDR-WPE), and KP-WPE in terms of three metrics: perceptual evaluation of speech quality (PESQ) [36], frequency-weighted segmental SNR (SNRseg) [37] and cepstral distance (CD) [37]. Note that for PESQ and SNRseg, the higher the score, the better is the performance, while for CD, the smaller the better. In the implementation of all the four methods, we set L to 20 and 10 for REVB1 and REVB2, respectively, $D=2$, and three iterations are performed to ensure convergence. The two sub-filters, $\mathbf{g}_1$ and $\mathbf{g}_2$, are initialized as a zero vector and a delay-and-sum filter, i.e., $\mathbf{g}_1 = [0 \;\; 0 \;\; \cdots \;\; 0]^T$ and $\mathbf{g}_2 = \mathbf{d}/M$, respectively. Note that we assume the direction-of-arrival is known and the steering vector $\mathbf{d}$ can be calculated accordingly. Table IV presents the results for the studied methods in the conditions of REVB1 and REVB2 with signal-to-noise ratio (SNR) $\in \{20, 30\}$ dB and without additive noise. It is clearly seen that the proposed KP-WPE has better performance in most conditions.

TABLE IV
PERFORMANCE OF WPE, WMPDR, WMPDR-WPE AND KP-WPE IN REVERBERANT ENVIRONMENT FOR VARIOUS NOISE LEVELS.

	REVB1 | no noise			REVB2 | no noise		
	PESQ	SNRseg	CD	PESQ	SNRseg	CD
observed	1.759	10.570	3.158	2.276	14.562	2.341
WPE	3.275	**15.728**	**1.719**	3.686	**18.494**	**1.238**
wMPDR	2.542	12.984	2.767	3.145	14.137	2.264
wMPDR-WPE	3.301	14.215	2.001	3.628	14.566	1.825
KP-WPE	**3.404**	14.494	1.823	**3.688**	14.708	1.711
	REVB1 | SNR = 30 dB			REVB2 | SNR = 30 dB		
	PESQ	SNRseg	CD	PESQ	SNRseg	CD
observed	1.644	7.836	3.969	2.024	10.314	3.634
WPE	2.562	12.571	3.393	2.763	14.112	3.274
wMPDR	2.368	12.261	3.457	2.867	13.656	3.273
wMPDR-WPE	2.849	13.407	3.216	3.115	14.013	3.181
KP-WPE	**2.967**	**13.621**	**3.107**	**3.170**	**14.126**	**3.114**
	REVB1 | SNR = 20 dB			REVB2 | SNR = 20 dB		
	PESQ	SNRseg	CD	PESQ	SNRseg	CD
observed	1.410	4.535	4.764	1.564	6.053	4.615
WPE	1.665	7.556	4.527	1.721	8.182	4.450
wMPDR	2.028	10.986	4.057	2.330	12.672	3.887
wMPDR-WPE	2.215	11.981	3.956	2.408	**13.006**	3.855
KP-WPE	**2.260**	**12.046**	**3.922**	**2.424**	13.002	**3.839**

V. CONCLUSIONS

This paper presented a Kronecker product based multi-channel linear prediction dereverberation method, in which the multichannel linear prediction filter is expressed as a Kronecker product of two sub-filters: one serves as a temporal prediction filter and the other serves as a spatial filter. In this work, the temporal prediction filter was formulated as a weighted Wiener filter and the spatial filter is formulated as a wMPDR beamformer. An iterative algorithm was developed to optimize the two sub-filters. In comparison with the widely-used WPE algorithm, the presented method not only exhibits better performance in terms of dereverberation and robustness to additive noise, but also has a much lower computational complexity, since the covariance matrices that need to be inverted are significantly smaller. Simulations verified the properties of the proposed method.

REFERENCES

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

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

[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, 2014.

[4] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[5] 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.

[6] T. Nakatani, B.-H. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1512–1527, 2008.

[7] 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, 2010.

[8] O. Schwartz, S. Gannot, and E. A. Habets, "An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1495–1510, 2016.

[9] S. Inoue, H. Kameoka, L. Li, S. Seki, and S. Makino, "Joint separation and dereverberation of reverberant mixtures with multichannel variational autoencoder," in *Proc. IEEE ICASSP*, 2019, pp. 96–100.

[10] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1369–1380, 2013.

[11] I. Kodrasi and S. Doclo, "Joint dereverberation and noise reduction based on acoustic multi-channel equalization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 680–693, 2016.

[12] I. Kodrasi and S. Doclo, "Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1106–1118, 2018.

[13] 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, 2015.

[14] S. Braun, A. Kuklasinski, O. Schwartz, O. Thiergart, E. A. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1052–1067, 2018.

[15] 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, 2018.

[16] G. Huang, J. Benesty, I. Cohen, and J. Chen, "A simple theory and new method of differential beamforming with uniform linear microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, no. 1, pp. 1079–1093, 2020.

[17] 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, 2018.

[18] A. Jukić, Z. Wang, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Constrained multi-channel linear prediction for adaptive speech dereverberation," in *Proc. IWAENC*, 2016, pp. 1–5.

[19] 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.

[20] 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, 2009.

[21] 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*, pp. 376–380, 2018.

[22] 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," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 544–558, 2018.

[23] T. Dietzen, S. Doclo, M. Moonen, and T. Van Waterschoot, "Joint multi-microphone speech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction," in *Proc. IWAENC*, 2018, pp. 221–225.

[24] 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.

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

[26] 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.

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

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

[29] 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.

[30] 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.

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

[32] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, 2007.

[33] D. A. Harville, *Matrix Algebra from a Statistician's Perspective*. Berlin, Germany: Springer-Verlag,1998.

[34] 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.

[35] 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.

[36] 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, 2007.

[37] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, *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, pp. 7–26, 2016.