ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10097198

SWITCHING KRONECKER PRODUCT LINEAR FILTERING FOR MULTISPEAKER ADAPTIVE SPEECH DEREVERBERATION

Gongping Huang[1] *Jacob Benesty*[2] *Israel Cohen*[3]
Emil Winebrand[4] *Jingdong Chen*[5] *Walter Kellermann*[1]

[1]LMS, University Erlangen-Nuremberg, 91058 Erlangen, Germany
[2]INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada
[3]Technion–Israel Institute of Technology, Haifa 3200003, Israel
[4]Insoundz Ltd., Tel Aviv 6473104, Israel
[5]CIAIC, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China

ABSTRACT

Dereverberation, a process to mitigate or eliminate the reverberation effect, plays an important role in hands-free speech communication and human-machine interfaces. Tremendous efforts have been devoted to this problem and various methods have been developed over the last three decades. Those methods generally assume that there is only a single speaker in the acoustic environment and, consequently, they suffer from significant performance degradation if multiple speakers participate in the conversation. How to deal with reverberation in multiple-speaker scenarios is still a challenging problem, which is studied in this work. We present a switching multichannel linear prediction filtering method, which designs multiple linear filters with each tracking one speaker. When some speaker is active, the corresponding filter and the weighted cross-correlation matrix are updated while the other filters are kept unchanged. To further improve the performance and reduce complexity, we apply the Kronecker product to decompose every linear prediction filter into a Kronecker product of two shorter filters: one is time-invariant and the other is time-varying. The former is estimated with a batch method (using only a few seconds of speech signal when the corresponding speaker starts to talk in the entire conversation) while a recursive least-squares algorithm is derived for identifying the time-varying set of Kronecker filters.

***Index Terms*—** Dereverberation, weighted-prediction-error, linear prediction, Kronecker product, switching filter.

1. INTRODUCTION

In voice communication and human-machine speech interfaces, a speech signal of interest picked up by microphones is inevitably contaminated by reverberation, which impairs the speech quality, intelligibility, and/or recognition performance [1–3]. As a result, dereverberation, a process to mitigate or even eliminate the reverberation effect, is needed in many applications [4–7]. This problem has been widely studied and several approaches have been investigated over the past decades [8–12]. Among those, the variance normalized delayed linear prediction method, also known as the weighted-prediction-error (WPE) based method [13–15], has demonstrated promising performance [16, 17]. This technique first estimates the reverberation effect due to the late reflections with a linear prediction filter and then subtracts the estimate from the observation [18–20]. In real-time applications, the linear prediction filter has to be updated in an adaptive manner, generally with the recursive least-squares (RLS) algorithm. The resulting method is called the adaptive WPE (AWPE) [21]. But the complexity of AWPE is rather high, making it challenging to implement in practical systems. To reduce the complexity, a Kronecker product AWPE (KAWPE) method was developed, which expresses the filter as a Kronecker product of two sets of shorter filters. This method is computationally much more efficient than AWPE since the filters that need to be identified are much shorter. Based on this framework, a partially time-varying Kronecker product AWPE is proposed, where the filter is expressed as a Kronecker product of one set of time-invariant filters and one set of time-varying filters. The complexity is further reduced since only the time-varying filters are updated on the fly. While it is very useful in dealing with reverberation, these methods generally suffer from significant performance degradation in situations where multiple speakers are participating in the conversation, which often happens in conferencing environments. One straightforward way of dealing with this problem is to detect the change in speaker position and re-initialize the filter and the corresponding statistics whenever a change in speaker position is detected [21]. Unfortunately, it takes some time for the algorithm to converge whenever a re-initialization happens, leading to inconsistent dereverberation performance and even disturbing artifacts. Consequently, how to deal with reverberation in multiple-speaker scenarios is still a challenging problem.

This work presents a switching multichannel linear prediction filtering method, which designs multiple linear filters: each tracking one speaker. When a certain speaker is active, the corresponding filter and weighted cross-correlation matrix are updated while the other filters are kept unchanged. To further improve the performance and reduce complexity, we adopt the Kronecker product filtering framework developed in [23, 24] and decompose every linear prediction filter as a Kronecker product of two shorter filters: one is time-invariant, and the other is time-varying. The time-invariant filters are estimated with a batch method using the first few seconds of speech signals from the respective speakers. For the time-varying filters, we follow the variance normalized delayed linear prediction method [14] to define the cost function and identify the optimal filters with RLS [25]. So, the proposed method only updates a set of shorter filters instead of updating long prediction filters as in the conventional AWPE method, leading to faster convergence. Furthermore, the proposed method is computationally more efficient than AWPE.

2. SIGNAL MODEL AND PROBLEM FORMULATION

We consider the signal model in which an array of M sensors is used to capture speech signals in some sound field where there are

This work was supported in part by the Alexander von Humboldt Foundation, 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 Grant No. 61831019 and 62192713.

Q speakers. We assume that there is only one active speaker every time. In the short-time-Fourier-transform (STFT) domain, the received signal at the mth ($m = 1, 2, \ldots, M$) microphone is expressed as

$$Y_m(n,\omega) = X_m(n,\omega) + V_m(n,\omega), \tag{1}$$

where n is the time-frame index, ω denotes the angular frequency, and $X_m(n,\omega)$ and $V_m(n,\omega)$ are the convolved speech and additive noise at the mth microphone, respectively. The convolved speech signals are coherent across the sensors. All signals are zero mean and broadband. For simplicity, we will omit ω in the remainder of this paper unless otherwise specified.

Multichannel linear prediction dereverberation estimates the late reflection components from the past L time frames and subtracts the estimate from the observation, thereby estimating the source signal. In real-time applications, the reverberation prediction filter is adaptively updated and the corresponding dereverberation process is expressed as [14]

$$\widehat{S}(n) = Y(n) - \mathbf{g}^H(n-1)\,\bar{\mathbf{y}}(n), \tag{2}$$

where $Y(n)$ is the reference signal, which can be the observation at any one of the microphones, $\mathbf{g}(n-1)$ is a time-varying prediction filter of length ML, which is obtained at the time frame $n-1$, the subscript ^{H} is the conjugate transpose operator, and

$$\bar{\mathbf{y}}(n) = \Big[\mathbf{y}^T(n-D) \;\; \mathbf{y}^T(n-D-1) \;\; \cdots \;\; \mathbf{y}^T(n-D-L+1)\Big]^T \tag{3}$$

is the stacked observation signal vector of length $L_M = ML$, with

$$\mathbf{y}(n-D-l) = [Y_1(n-D-l) \;\; Y_2(n-D-l) \;\; \cdots \;\; Y_M(n-D-l)]^T, \tag{4}$$

$l = 0, 1, \ldots, L-1$, being the observation signal vector of length M, the superscript ^{T} denotes the transpose of a vector or a matrix, and D is a delay parameter whose value depends on the system setup.

The problem of dereverberation is then to find the optimal filter, $\mathbf{g}(n)$, so that the late reflection components are suppressed as much as possible. One of the most widely used methods to estimate the filter is WPE, which obtains the dereverberation filter estimate through maximizing the likelihood function of the speech and channel models [14]. The solution is as follows:

$$\mathbf{g}(n) = \boldsymbol{\Phi}_{\bar{\mathbf{y}}}^{-1}(n)\boldsymbol{\rho}_{\bar{\mathbf{y}}}(n), \tag{5}$$

where

$$\boldsymbol{\Phi}_{\bar{\mathbf{y}}}(n) = \sum_{i=1}^{n} \alpha^{n-i}\,\frac{\bar{\mathbf{y}}(i)\,\bar{\mathbf{y}}^H(i)}{\lambda(i)},$$

$$\boldsymbol{\rho}_{\bar{\mathbf{y}}}(n) = \sum_{i=1}^{n} \alpha^{n-i}\,\frac{\bar{\mathbf{y}}(i)\,Y_m^*(i)}{\lambda(i)}$$

are the weighted correlation matrix and correlation vector, respectively, α ($0 < \alpha < 1$) is the forgetting factor, and $\lambda(i) = |\widehat{S}(i)|^2$ is the short time variance of the desired speech signal. By using the RLS method, the inverse weighted cross-correlation matrix and the prediction filters are recursively estimated, leading to the so-called AWPE method [21].

While it has demonstrated great potential for dereverberation in the single-speaker scenario, AWPE generally suffers from significant performance degradation in situations where multiple speakers participate in the conversation. One straightforward way of dealing with this problem is through detecting the change in speaker position and re-initializing the filter and corresponding statistics whenever a change in speaker position is detected [21]. Unfortunately, it takes time for the algorithm to converge whenever a re-initialization happens, leading to inconsistent dereverberation performance and even disturbing artifacts.

3. SWITCHING KRONECKER PRODUCT ADAPTIVE MULTICHANNEL LINEAR FILTERING

In this work, we study the speech dereverberation problem in multiple-speaker scenarios. Let us assume that there is a total of Q speakers participating in the conversation (but only one speaker is active each time) and every speaker's position is either static or slowly varying. We consider designing Q linear filters to track the speakers separately.

Denote by $\mathbf{g}^{(q)}(n)$, $q = 1, 2, \ldots, Q$, the prediction filter corresponding to the qth speaker. When the qth speaker is active, we implement the dereverberation process with the prediction filter $\mathbf{g}^{(q)}(n)$ and update its corresponding variables. AWPE needs to update a linear prediction filter of length L_M in every frequency band, which involves high computational complexity. In [24], we proposed a Kronecker product filtering framework for speech dereverberation, where the linear prediction filter is formulated as the Kronecker product of two sets of shorter filters, and only some of the short filters are updated. This Kronecker product decomposition combined with partial filter adaptation can reduce the computational complexity without sacrificing dereverberation performance. In this work, we extend the principle in [24] to the design of switching linear prediction filters for speech dereverberation in multiple-speaker environments.

Let us write the linear prediction filter $\mathbf{g}^{(q)}(n)$ of length L_M as a partially time-varying Kronecker product of two sets of short filters [24]:

$$\mathbf{g}^{(q)}(n) = \sum_{p=1}^{P} \underbrace{\mathbf{g}_{2,p}^{(q)}}_{\text{time-invariant}} \otimes \underbrace{\mathbf{g}_{1,p}^{(q)}(n)}_{\text{time-varying}}, \; q = 1, 2, \ldots, Q, \tag{6}$$

where $\otimes$ is the Kronecker product, $\mathbf{g}_{2,p}^{(q)}$, $p = 1, 2, \ldots, P$ are the time-invariant filters of lengths L_2, $\mathbf{g}_{1,p}^{(q)}(n)$, $p = 1, 2, \ldots, P$ are time-varying filters of lengths L_1, with $L_M = L_1 L_2$. Since the position of every speaker is assumed to be static (or slowly varying), we can use the first few seconds of the speech signals from the qth speaker (when the speaker begins to speak for the first time) to compute the Kronecker filters $\mathbf{g}_{2,p}^{(q)}(n)$ and $\mathbf{g}_{1,p}^{(q)}(n)$, $p = 1, 2, \ldots, P$, respectively [24]. Then, the time-invariant filters $\mathbf{g}_{2,p}^{(q)}$, $p = 1, 2, \ldots, P$ are fixed, and we only need to update the optimal filters $\mathbf{g}_{1,p}^{(q)}$, $p = 1, 2, \ldots, P$.

The dereverberated signal can be written as

$$\widehat{S}(n) = Y(n) - \sum_{q=1}^{Q} \gamma_q \left(\mathbf{g}^{(q)}(n-1)\right)^H \bar{\mathbf{y}}(n), \tag{7}$$

where

$$\gamma_q = \begin{cases} 1, & \text{if the } q\text{th speaker active} \\ 0, & \text{else} \end{cases}$$

indicates which speaker is active, $\mathbf{g}^{(q)}(n)$ is the linear prediction filters of length L_M corresponding to the q speaker, and $q = 1, 2, \ldots, Q$.

For the Kronecker product, we have the following relationships [26]:

$$\mathbf{g}_{2,p}^{(q)} \otimes \mathbf{g}_{1,p}^{(q)}(n) = \left[\mathbf{g}_{2,p}^{(q)} \otimes \mathbf{I}_{L_1}\right] \mathbf{g}_{1,p}^{(q)}(n) = \mathbf{G}_{2,p}^{(q)} \mathbf{g}_{1,p}^{(q)}(n), \; p = 1, 2, \ldots, P, \tag{8}$$

where $\mathbf{I}_{L_1}$ is the identity matrix of size $L_1 \times L_1$ and

$$\mathbf{G}_{2,p}^{(q)} = \mathbf{g}_{2,p}^{(q)} \otimes \mathbf{I}_{L_1} \tag{9}$$

is a matrix of size $L_1 L_2 \times L_1$. Consequently, the linear prediction filters can be written as

$$\mathbf{g}^{(q)}(n) = \sum_{p=1}^{P} \mathbf{G}_{2,p}^{(q)} \mathbf{g}_{1,p}^{(q)}(n),\ q = 1, 2, \ldots, Q. \tag{10}$$

Substituting (10) into (7) gives the dereverberated signal, i.e.,

$$\begin{aligned}\widehat{S}(n) &= Y(n) - \sum_{q=1}^{Q} \gamma_q \sum_{p=1}^{P} \left(\mathbf{g}_{1,p}^{(q)}(n-1)\right)^H \left(\mathbf{G}_{2,p}^{(q)}\right)^H \bar{\mathbf{y}}(n) \\ &= Y(n) - \sum_{q=1}^{Q} \gamma_q \sum_{p=1}^{P} \left(\mathbf{g}_{1,p}^{(q)}(n-1)\right)^H \mathbf{y}_{2,p}^{(q)}(n) \\ &= Y(n) - \sum_{q=1}^{Q} \gamma_q \left(\underline{\mathbf{g}}_1^{(q)}(n-1)\right)^H \underline{\mathbf{y}}_2^{(q)}(n),\end{aligned} \tag{11}$$

where

$$\mathbf{y}_{2,p}^{(q)}(n) = \left(\mathbf{G}_{2,p}^{(q)}\right)^H \bar{\mathbf{y}}(n),\ p = 1, 2, \ldots, P \tag{12}$$

are vectors of length L_1, and

$$\underline{\mathbf{g}}_1^{(q)}(n-1) = \left[\ \left(\mathbf{g}_{1,1}^{(q)}(n-1)\right)^T \quad \left(\mathbf{g}_{1,2}^{(q)}(n-1)\right)^T \quad \cdots \quad \left(\mathbf{g}_{1,P}^{(q)}(n-1)\right)^T\right]^T, \tag{13}$$

$$\underline{\mathbf{y}}_2^{(q)}(n) = \left[\ \left(\mathbf{y}_{2,1}^{(q)}(n)\right)^T \quad \left(\mathbf{y}_{2,2}^{(q)}(n)\right)^T \quad \cdots \quad \left(\mathbf{y}_{2,P}^{(q)}(n)\right)^T\right]^T \tag{14}$$

are vectors of length PL_1. Substituting (9) into (12), one can express the vector $\mathbf{y}_{2,p}^{(q)}(n)$ as

$$\mathbf{y}_{2,p}^{(q)}(n) = \mathbf{Y}(n) \left(\mathbf{g}_{2,p}^{(q)}(n-1)\right)^*,\ p = 1, 2, \ldots, P, \tag{15}$$

where

$$\mathbf{Y}(n) = \begin{bmatrix} Y_1(n) & Y_{L_1+1}(n) & \cdots & Y_{(L_2-1)L_1+1}(n) \\ Y_2(n) & Y_{L_1+2}(n) & \cdots & Y_{(L_2-1)L_1+2}(n) \\ \vdots & \vdots & \ddots & \vdots \\ Y_{L_1}(n) & Y_{2L_1}(n) & \cdots & Y_{L_2 L_1}(n) \end{bmatrix}$$

is a matrix of size $L_1 \times L_2$ folded from the vector $\bar{\mathbf{y}}(n)$ defined in (3), $Y_i(n)$, $i = 1, 2, \ldots, L_2 L_1$, is the ith element of the vector $\bar{\mathbf{y}}(n)$. Clearly, computing $\mathbf{y}_{2,p}^{(q)}(n)$ with (15) can be more efficient than (12) since it needs only $L_1 L_2$ complex-valued multiplications.

The problem of dereverberation then becomes one of identifying the optimal filter $\underline{\mathbf{g}}_1^{(q)}(n)$. When the qth speaker is active, i.e., $\gamma_q = 1$, we follow the WPE method [14] and the RLS algorithm developed in [25] to define the cost function under the least-squares (LS) error criterion as

$$\mathcal{J}\left[\underline{\mathbf{g}}_1^{(q)}(n)\right] = \sum_{i=1}^{n} \alpha^{n-i} \frac{\left|Y(i) - \left(\underline{\mathbf{g}}_1^{(q)}(n)\right)^H \underline{\mathbf{y}}_2^{(q)}(i)\right|^2}{\lambda(i)}, \tag{16}$$

where α ($0 < \alpha < 1$) is the forgetting factor, and $\lambda(i) = \left|\widehat{S}(i)\right|^2$, $i = 1, 2, \ldots, n$, is an estimate of the short-time variance of the desired speech signal [14].

The minimization of $\mathcal{J}\left[\underline{\mathbf{g}}_1^{(q)}(n)\right]$ with respect to $\underline{\mathbf{g}}_1^{(q)}(n)$ leads to the Wiener solution:

$$\underline{\mathbf{g}}_1^{(q)}(n) = \left(\mathbf{\Phi}_{\underline{\mathbf{y}}_2}^{(q)}(n)\right)^{-1} \boldsymbol{\rho}_{\underline{\mathbf{y}}_2}^{(q)}(n), \tag{17}$$

where $\mathbf{\Phi}_{\underline{\mathbf{y}}_2}^{(q)}(n)$ and $\boldsymbol{\rho}_{\underline{\mathbf{y}}_2}^{(q)}(n)$ are the weighted correlation vector and weighted cross-correlation matrix that are estimated recursively as

$$\mathbf{\Phi}_{\underline{\mathbf{y}}_2}^{(q)}(n) = \alpha \mathbf{\Phi}_{\underline{\mathbf{y}}_2}^{(q)}(n-1) + \frac{\underline{\mathbf{y}}_2^{(q)}(n) \left(\underline{\mathbf{y}}_2^{(q)}(n)\right)^H}{\lambda(n)}, \tag{18}$$

$$\boldsymbol{\rho}_{\underline{\mathbf{y}}_2}^{(q)}(n) = \alpha \boldsymbol{\rho}_{\underline{\mathbf{y}}_2}^{(q)}(n-1) + \frac{\underline{\mathbf{y}}_2^{(q)}(n) Y^*(n)}{\lambda(n)}. \tag{19}$$

Using the matrix inversion lemma, the inverse weighted cross-correlation matrix corresponding to the qth speaker is estimated recursively as

$$\begin{aligned}\left(\mathbf{\Phi}_{\underline{\mathbf{y}}_2}^{(q)}(n)\right)^{-1} &= \frac{1}{\alpha}\left(\mathbf{\Phi}_{\underline{\mathbf{y}}_2}^{(q)}(n-1)\right)^{-1} \\ &\quad - \frac{1}{\alpha}\boldsymbol{\kappa}_2^{(q)}(n)\left(\underline{\mathbf{y}}_2^{(q)}(n)\right)^H \left(\mathbf{\Phi}_{\underline{\mathbf{y}}_2}^{(q)}(n-1)\right)^{-1},\end{aligned} \tag{20}$$

where

$$\begin{aligned}&\boldsymbol{\kappa}_2^{(q)}(n) \\ &= \frac{\left(\mathbf{\Phi}_{\underline{\mathbf{y}}_2}^{(q)}(n-1)\right)^{-1} \underline{\mathbf{y}}_2^{(q)}(n)}{\alpha\lambda(n) + \left(\underline{\mathbf{y}}_2^{(q)}(n)\right)^H \left(\mathbf{\Phi}_{\underline{\mathbf{y}}_2}^{(q)}(n-1)\right)^{-1} \underline{\mathbf{y}}_2^{(q)}(n)}\end{aligned} \tag{21}$$

is the gain vector corresponding to the qth speaker. Substituting (19) and (20) into (17) gives the rule to update the filter $\underline{\mathbf{g}}_1^{(q)}(n)$, i.e.,

$$\underline{\mathbf{g}}_1^{(q)}(n) = \underline{\mathbf{g}}_1^{(q)}(n-1) + \boldsymbol{\kappa}_2^{(q)}(n) \widehat{S}^*(n). \tag{22}$$

The overall strategy is to update the filter $\underline{\mathbf{g}}_1^{(q)}(n)$ and its corresponding inverse weighted cross-correlation matrix when the qth speaker is active and freeze the other $Q-1$ filters and their corresponding inverse weighted cross-correlation matrices until a speaker change is detected. In the proposed partially time-varying switching KAWPE (PTV-SKAWPE) algorithm, the time-invariant filters $\mathbf{g}_{2,p}^{(q)}$, $p = 1, 2, \ldots, P$, $q = 1, 2, \ldots, Q$ are computed using the first few-second speech signal from the qth speaker with the KAWPE method [24].

Table 1. Computational Complexity in Terms of Complex-Valued Multiplications of the AWPE and PTV-SKAWPE Methods.

Method	Complex-valued Multiplications
AWPE	$4L_1^2L_2^2 + 4L_1L_2 + 3$
PTV-SKAWPE	$4P^2L_1^2 + PL_1L_2 + 4PL_1 + 3$

As summarized in Table 1, the AWPE method with an RLS adaptive method has a computational complexity proportional to $O(L^2) = O(L_1^2 L_2^2)$. In comparison, the proposed PTV-SKAWPE method has a computational complexity proportional to $O(P^2 L_1^2)$ (without considering the complexity for speaker position change detection), so the computational complexity of the proposed method is expected to be much lower if $P \ll L_2$.

4. EXPERIMENTS AND ANALYSIS

In this section, we study the performance of the proposed PTV-SKAWPE method and compare it to AWPE. We consider a uniform linear array consisting of 4 omnidirectional microphones lo-

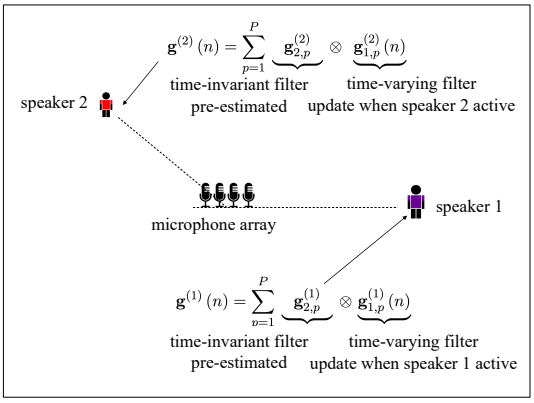

Fig. 1. An illustration of the proposed PTV-SKAWPE method for speech dereverberation with two speakers.

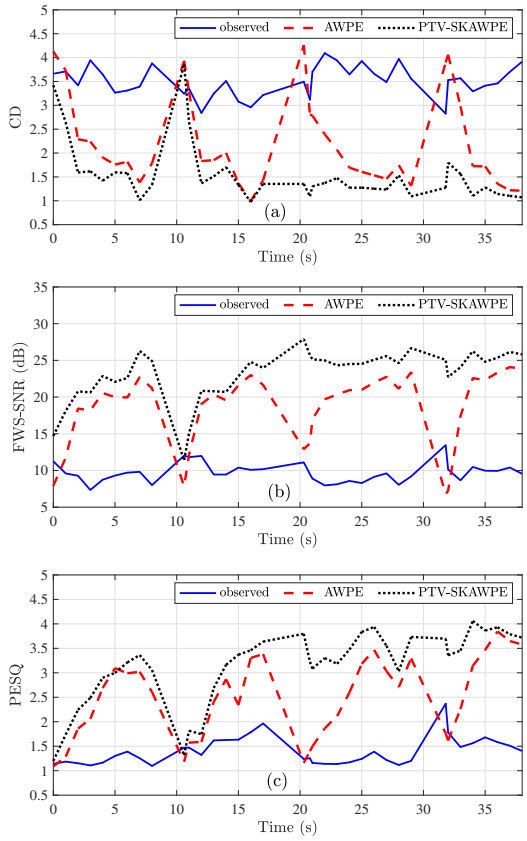

Fig. 2. Performance of AWPE and PTV-SKAWPE: (a) CD, (b) FWS-SNR, and (c) PESQ. The speaker position change happens approximately at the 10th, 20th, and 32th second.

cated in a room of size 6 m × 8 m × 4 m. For ease of exposition, the 3-dimensional Cartesian coordinate system is used to specify the position of a point in the room. The positions of the four microphones are, respectively, at (x, 4.0, 1.5), where $x = 2.9362, 2.9787, 3.0213, 3.0638$. The acoustic channel impulse responses from the source to the microphones are generated using the image model method [27], where the reverberation time, T_{60}, is approximately 400 ms. The clean source speech signal used in the experiment is recorded in a quiet office room with a sampling frequency of 16 kHz. The microphone observation signals are generated by convolving the source signal with the corresponding impulse responses. The dereverberation process is implemented in the STFT domain, where the observation signals are divided into overlapping

Table 2. Average Performance of AWPE and PTV-SKAWPE.

	CD	FWS-SNR (dB)	PESQ
observed	3.499	9.705	1.398
AWPE	2.210	18.563	2.472
PTV-SKAWPE	1.543	23.129	3.123

frames with a frame size of 512 samples and an overlapping factor of 75%. A Kaiser window is applied to every frame before transforming it into the STFT domain. After processing, the dereverberated signal is transformed to the time domain with the overlap-add method. To evaluate the dereverberation performance, the signals are separated into short segments of approximately 2-seconds duration each. Then, the cepstral distance (CD), the frequency-weighted segmental SNR (FWS-SNR), and the perceptual evaluation of speech quality (PESQ) are evaluated for each segment [28]. Note that for CD, a smaller value typically indicates better speech dereverberation performance, while for FWS-SNR and PESQ, larger values correspond to better performance.

We consider two speakers located at $(5.0, 4.0, 1.5)$ and $(2.0, 6.0, 1.5)$, respectively. The overall length of the source signal is approximately 40 seconds, where the speaker position change happens approximately at the 10th, 20th, and 32th second (when a change happens, there is about 0.1-second silence between switching from one speaker to the other). In this simulation, we assume that the change of speaker position is given as the *a priori* information (note that for the developed method, the exact location of the target speaker is not needed, so different techniques such as direction-of-arrival estimation and speaker identification can be used to detect the speaker position change in practice). For the implementation of the PTV-SKAWPE algorithm, we design two linear prediction filters to track the two speakers, where the parameters are set to: $D = 5$, $L_1 = 8$, $L_2 = 8$, $P = 4$, and $\alpha = 0.99$. We use the first 5-second speech signal from each speaker to compute the time-invariant filters $\left\{\mathbf{g}_{2,p}^{(q)}(n)\right\}_p$. The AWPE method is designed under the same conditions with $L = 16$ and $M = 4$ (the overall lengths of the filters of PTV-SKAWPE and AWPE are the same) and $\alpha = 0.99$.

Figure 2 plots the CD, FWS-SNR, and PESQ obtained by the AWPE and PTV-SKAWPE methods, all as a function of time. As seen, at the initialization stage, all methods need some time to converge. After convergence, when the speaker position changes, the PTV-SKAWPE methods maintain a good performance while AWPE fails. Table 2 presents the CD, FWS-SNR, and PESQ averaged over time for these methods. As seen, the PTV-SKAWPE yields better performance. Note that the computational complexity of SKAWPE decreases with the value of P. So, it can reduce the computational complexity and still achieve a good dereverberation performance if the value of P is properly chosen.

5. CONCLUSIONS

This paper studied the problem of speech dereverberation in multiple-speaker environments. A switching multichannel linear prediction filtering method was developed, which involves multiple linear filters with each tracking one speaker. The Kronecker product is applied to decompose every linear prediction filter into two shorter filters: one is time-invariant, capturing the static properties of the reverberation, while the other is time-varying. The time-invariant filters are estimated with a batch method using a short-segment of speech signals from the respective speakers and an RLS algorithm is derived for identifying the time-varying set of Kronecker filters in an adaptive manner. Analysis and simulation results showed that the developed PTV-SKAWPE method outperforms AWPE in terms of both complexity and dereverberation performance.

6. REFERENCES

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

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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*, pp. 96–100, IEEE, 2019.

[9] 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.

[10] J. S. Erkelens and R. Heusdens, "Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1746–1765, 2010.

[11] 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.

[12] F. D. Aprilyanti, H. Saruwatari, S. Nakamura, and T. Takatani, "Optimized joint noise suppression and dereverberation based on blind signal extraction for hands-free speech recognition system," in *hscma*, pp. 182–186, IEEE, 2014.

[13] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multichannel linear prediction based on short time fourier transform representation," in *Proc. IEEE ICASSP*, pp. 85–88, IEEE, 2008.

[14] 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.

[15] 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.

[16] R. Ikeshita, N. Kamo, and T. Nakatani, "Blind signal dereverberation based on mixture of weighted prediction error models," *IEEE Signal Process. Lett.*, vol. 28, pp. 399–403, 2021.

[17] 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.

[18] T. Nakatani and K. Kinoshita, "Maximum likelihood convolutional beamformer for simultaneous denoising and dereverberation," in *EUSIPCO*, pp. 1–5, IEEE, 2019.

[19] S. Song, L. Cheng, S. Luan, D. Yao, J. Li, and Y. Yan, "An integrated multichannel approach for joint noise reduction and dereverberation," *Applied Acoustics*, vol. 171, p. 107526, 2021.

[20] S. Hashemgeloogerdi and S. Braun, "Joint beamforming and reverberation cancellation using a constrained Kalman filter with multichannel linear prediction," in *Proc. IEEE ICASSP*, pp. 481–485, IEEE, 2020.

[21] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *Proc. IEEE ICASSP*, pp. 3733–3736, IEEE, 2009.

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

[23] 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.*, vol. 28, pp. 101–105, 2021.

[24] 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, 2022.

[25] 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/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 5, pp. 903–918, 2019.

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

[27] 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.

[28] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, *et al.*, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE WASPAA*, pp. 1–4, IEEE, 2013.