LOW-COMPLEXITY KALMAN FILTER FOR MULTI-CHANNEL LINEAR-PREDICTION-BASED BLIND SPEECH DEREVERBERATION

Thomas Dietzen[1,2], *Simon Doclo*[3], *Ann Spriet*[2], *Wouter Tirry*[2], *Marc Moonen*[1], *Toon van Waterschoot*[1,4]

[1] KU Leuven, Dept. of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems,
Signal Processing and Data Analytics, Leuven, Belgium
[2] NXP Semiconductors Belgium N.V., Leuven, Belgium
[3] University of Oldenburg, Dept. of Medical Physics and Acoustics
and Cluster of Excellence Hearing4All, Oldenburg, Germany
[4] KU Leuven, Dept. of Electrical Engineering (ESAT-ETC), Geel, Belgium

ABSTRACT

Multi-channel linear prediction (MCLP) has been shown to be a suitable framework for tackling the problem of blind speech dereverberation. In recent years, a number of adaptive MCLP algorithms have been proposed, whereby the majority operates in the short-time Fourier transform (STFT) domain. In this paper, we focus on the STFT-based Kalman filter solution to the adaptive MCLP task. Similarly to all other available adaptive STFT-based MCLP algorithms, the Kalman filter exhibits a quadratic computational cost in the number of filter coefficients per frequency bin. Aiming at a reduced complexity, we propose to simplify to the Kalman filter solution by enforcing the state error correlation matrix to be block-diagonal, leading to a linear cost instead. Further, we apply a Wiener-gain spectral post-processor subsequent to MCLP, which is designed from readily available power spectral density (PSD) estimates. The convergence behavior of the standard and the simplified algorithm is evaluated by means of two objective measures, i.e. perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI), showing only a minor performance degradation for the simplified algorithm.

Index Terms— Speech dereverberation, Kalman filter, multi-channel linear prediction, low complexity

1. INTRODUCTION

It is well known that acoustic reverberation, caused by a multitude of reflections from room boundaries and objects, may have a deteriorating effect on the quality and intelligibility of speech signals recorded by a microphone. In recent years, a microphone-array-based framework known as multi-channel linear prediction (MCLP) [1–12] has gained increased popularity for blind speech dereverberation, since no prior knowledge on the room impulse responses (RIRs) between the speech source and the microphones is required. According to the multiple input/output inverse theorem (MINT) [13], multi-channel methods like MCLP are theoretically able to perfectly equalize the (presumed time-invariant) transfer functions between the speech source and the microphone array, provided that the individual transfer functions do not share common zeros. The majority of the recently proposed MCLP algorithms work in the short-time Fourier transform (STFT) domain [3–10, 12], where each frequency bin is treated independently.

In a practical scenario, adaptive filter estimation is required in order to equalize potentially time-varying RIRs. While most MCLP algorithms are based on either batch processing or iterative processing of individual, independent frames, a number of adaptive algorithms can be found [8–12]. In [8], the weighted recursive least squares (RLS) algorithm has been applied in the STFT domain. In [9], an RLS implementation of the STFT-based generalized weighted prediction error (GWPE) method [6] has been introduced. In order to prevent excessive cancellation of the speech signal, a constraint has been applied to the adaptive GWPE optimization problem in [10]. In our previous work, we have proposed a Kalman filter based on a partitioned-block frequency domain (PBFD) representation [11], while an STFT-based Kalman filter implementation has been proposed in [12].

In terms of complexity, all available STFT-based adaptive MCLP algorithms exhibit a quadratic computational cost in the number of filter coefficients per frequency bin. In this paper, we propose a simplification to the STFT-based Kalman filter leading to a linear computational cost in the number of filter coefficients instead. This simplification is conceptually equivalent to an assumption often made in PBFD-based Kalman filtering [11, 14, 15], namely that the variations in different filter partitions are mutually uncorrelated and have zero mean.

Further, in order to suppress residual reverberation, we utilize a Wiener-gain spectral post-processor subsequent to MCLP-based dereverberation. The MCLP Kalman filter requires an estimate of the target signal power spectral density (PSD), which we obtain using [16], and provides an estimate of the output signal PSD. We can therefore easily derive the Wiener gain from readily available PSD estimates.

2. PROBLEM FORMULATION

In MCLP, it is assumed that the reverberation component to be cancelled may be modeled as a linearly filtered version of the delayed microphone signals. The task at this point is to blindly estimate the filter coefficients by means of an adaptive filter. In the following, we

This research work was carried out in the frame of KU Leuven Research Council CoE PFV/10/002 (OPTEC), KU Leuven Impulse Fund IMP/14/037, KU Leuven Internal Funds C2-16-00449 'Distributed Digital Signal Processing for Ad-hoc Wireless Local Area Audio Networking', and the FP7-PEOPLE Marie Curie Initial Training Network 'Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS)', funded by the European Commission under Grant Agreement no. 316969. The scientific responsibility is assumed by its authors.

978-1-5386-1631-4/17/$31.00 ©2017 IEEE

briefly introduce the MCLP signal model in the STFT domain, presuming that background noise is absent (please note however that a low amount of background noise is added in the simulations in Section 6).

Let $x_m(l,k)$ with $m = 0, \ldots, M-1$ denote the STFT domain representation of the m^{th} microphone signal at frame l in frequency bin k, comprising a target component $x_{\text{t}|m}(l,k)$ typically including early reflections, and a reverberation component $x_{\text{r}|m}(l,k)$ to be cancelled,

$$x_m(l,k) = x_{\text{t}|m}(l,k) + x_{\text{r}|m}(l,k). \tag{1}$$

For the sake of simplicity, we will focus on the dereverberation of $x_0(l,k)$ only. Since we treat the frequency bins independently, the frequency bin index will be dropped for brevity. Further, let $\hat{w}_{p,m}(l)$ with $p = 0, \ldots, P-1$ denote the p^{th} STFT domain filter coefficient for microphone m at frame l in frequency bin k. We define the stacked representation,

$$\mathbf{x}(l) = \big(x_0(l) \;\cdots\; x_{M-1}(l)\big)^T \qquad \in \mathbb{C}^M, \tag{2}$$

$$\underline{\mathbf{x}}(l) = \big(\mathbf{x}^T(l) \;\cdots\; \mathbf{x}^T(l-P+1)\big)^T \qquad \in \mathbb{C}^{PM}, \tag{3}$$

$$\hat{\mathbf{w}}_p(l) = \big(\hat{w}_{p,0}(l) \;\cdots\; \hat{w}_{p,M-1}(l)\big)^T \qquad \in \mathbb{C}^M, \tag{4}$$

$$\underline{\hat{\mathbf{w}}}(l) = \big(\hat{\mathbf{w}}_0^T(l) \;\cdots\; \hat{\mathbf{w}}_{P-1}^T(l)\big)^T \qquad \in \mathbb{C}^{PM}, \tag{5}$$

wherein the superscript $(\cdot)^T$ denotes the transpose. With these definitions, we express the enhanced signal $e(l)$ as

$$\begin{aligned} e(l) &= x_0(l) - \sum_{p=0}^{P-1}\sum_{m=0}^{M-1} x_m(l-D-p)\hat{w}_{p,m}(l) \\ &= x_0(l) - \underline{\mathbf{x}}^T(l-D)\underline{\hat{\mathbf{w}}}(l). \end{aligned} \tag{6}$$

In (6), the prediction term $\underline{\mathbf{x}}^T(l-D)\underline{\hat{\mathbf{w}}}(l)$ is the estimate of the reverberation component $x_{\text{r}|0}(l)$ to be cancelled. The prediction delay D is a design parameter affecting the amount of early reflections maintained in the MCLP output $e(l)$, i.e. in the estimate of the target component $x_{\text{t}|0}(l)$. A corresponding block diagram is depicted in Fig. 1.

The per-frequency-bin filter operation in (6) may be considered an approximation of the time domain convolution [17], whereby each output frame is composed of P circular convolution terms, involving the P latest input signal frames. This is in contrast to PBFD processing, where one would add P linear convolution terms in an analogous formulation. In the STFT formulation, undesired circular convolution effects are alleviated by the use of appropriate weighted-overlap-add (WOLA) windowing.

3. KALMAN-FILTER-BASED MCLP

In the following, we define the state-space representation and derive the Kalman-filter-based update algorithm that produces the estimate $\underline{\hat{\mathbf{w}}}(l)$ of the presumed underlying state $\underline{\mathbf{w}}(l)$. The algorithm is then modified in order to reduce the computational complexity.

3.1. State-Space Model

We define the state $\underline{\mathbf{w}}(l)$ to be the filter that leads to perfect cancellation of the reverberation component $x_{\text{r}|0}(l)$, i.e. $\underline{\mathbf{x}}^T(l-D)\underline{\mathbf{w}}(l) = x_{\text{r}|0}(l)$, resulting in the so-called observation equation,

$$x_0(l) = \underline{\mathbf{x}}^T(l-D)\underline{\mathbf{w}}(l) + x_{\text{t}|0}(l). \tag{7}$$

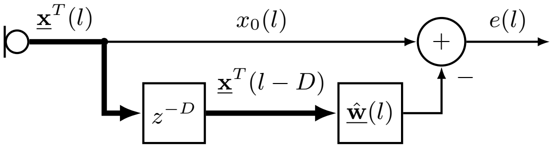
$\underline{\mathbf{x}}^T(l)$
$x_0(l)$
$+$
$e(l)$
$\underline{\mathbf{x}}^T(l-D)$
$-$
z^{-D}
$\underline{\hat{\mathbf{w}}}(l)$

Figure 1: MCLP in the STFT domain.

In Kalman filter terminology, we refer to $x_0(l)$ as the observable and to $x_{\text{t}|0}(l)$ as the observation noise. Of the latter, an estimate of its correlation $\psi_{x_{\text{t}|0}}(l)$ corresponding to its PSD is required. The estimation of $\psi_{x_{\text{t}|0}}(l)$ will be discussed in Section 4.

In order to derive the Kalman filter update equations, we further need to formulate the so-called process equation, which makes assumptions on the evolution of the underlying state $\underline{\mathbf{w}}(l)$ in the form of a first-order difference equation, i.e.

$$\underline{\mathbf{w}}(l) = \underline{\mathbf{A}}^T(l)\underline{\mathbf{w}}(l-1) + \underline{\mathbf{\Delta}}_w(l). \tag{8}$$

The matrix $\underline{\mathbf{A}}^T(l) \in \mathbb{C}^{PM\times PM}$ models the state transition from one frame to the next, and the process noise $\underline{\mathbf{\Delta}}_w(l) \in \mathbb{C}^{PM}$ models a random variation component of the state over time. As for the observation noise, an estimate of its correlation matrix $\underline{\mathbf{\Psi}}_{\Delta_w}(l)$ is required. Both $\underline{\mathbf{A}}^T(l)$ and $\underline{\mathbf{\Psi}}_{\Delta_w}(l)$ may be considered design parameters and are commonly chosen to be diagonal matrices.

3.2. Update Equations

From the observation and the process equation in (7)–(8), the Kalman filter update equations [18] can be derived as,

$$\underline{\hat{\mathbf{w}}}(l) = \underline{\mathbf{A}}^T(l)\underline{\hat{\mathbf{w}}}^+(l-1), \tag{9}$$

$$\underline{\mathbf{\Psi}}_w(l) = \underline{\mathbf{A}}^T(l)\underline{\mathbf{\Psi}}_w^+(l-1)\underline{\mathbf{A}}^*(l) + \underline{\mathbf{\Psi}}_{\Delta_w}(l), \tag{10}$$

$$e(l) = x_0(l) - \underline{\mathbf{x}}^T(l-D)\underline{\hat{\mathbf{w}}}(l), \tag{11}$$

$$\psi_e(l) = \underline{\mathbf{x}}^T(l-D)\underline{\mathbf{\Psi}}_w(l)\underline{\mathbf{x}}^*(l-D) + \psi_{x_{\text{t}|0}}(l), \tag{12}$$

$$\underline{\mathbf{k}}(l) = \underline{\mathbf{\Psi}}_w(l)\underline{\mathbf{x}}^*(l-D)\psi_e^{-1}(l), \tag{13}$$

$$\underline{\hat{\mathbf{w}}}^+(l) = \underline{\hat{\mathbf{w}}}(l) + \underline{\mathbf{k}}(l)e(l), \tag{14}$$

$$\underline{\mathbf{\Psi}}_w^+(l) = \underline{\mathbf{\Psi}}_w(l) - \underline{\mathbf{k}}(l)\underline{\mathbf{x}}^T(l-D)\underline{\mathbf{\Psi}}_w(l), \tag{15}$$

wherein the superscript $(\cdot)^*$ denotes the complex conjugate. Eq. (9)–(10) are referred to as the time update of the state estimate $\underline{\hat{\mathbf{w}}}(l)$ and the state error correlation matrix $\underline{\mathbf{\Psi}}_w(l) \in \mathbb{C}^{PM\times PM}$. In (11)–(13), the enhanced signal $e(l)$, its correlation $\psi_e(l)$, and the Kalman gain $\underline{\mathbf{k}}(l)$ are computed, which are then used in the so-called measurement update of the state estimate $\underline{\hat{\mathbf{w}}}^+(l)$ and the state error correlation matrix $\underline{\mathbf{\Psi}}_w^+(l)$ in (14)–(15). Both $\underline{\hat{\mathbf{w}}}^+(l)$ and $\underline{\mathbf{\Psi}}_w^+(l)$ need to be initialized at $l = 0$.

Within these update equations, the enhanced signal $e(l)$ in (11) represents the Kalman filter estimate of the target component $x_{\text{t}|0}(l)$. Note that its correlation $\psi_e(l)$ in (12) depends on the target component correlation $\psi_{x_{\text{t}|0}}(l)$. During convergence, the norm of the state error correlation matrix $\underline{\mathbf{\Psi}}_w(l)$ decreases, such that $\psi_e(l)$ converges to $\psi_{x_{\text{t}|0}}(l)$ and hence $e(l)$ converges to $x_{\text{t}|0}(l)$.

3.3. Complexity Reduction

The multi-channel frequency-domain filter $\hat{\mathbf{w}}_p(l)$ conceptually corresponds to what is referred to as a partition in the PBFD framework, in that it is multiplied with the frequency-domain representation of the p^{th} to last input signal frame. Subsequently, we therefore

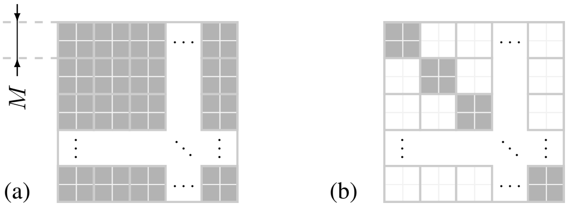

Figure 2: The structure of the state error correlation matrix $\underline{\boldsymbol{\Psi}}_w(l) \in \mathbb{C}^{PM \times PM}$ for (a) preserved and (b) omitted cross-partition error correlations, exemplarily shown for $M = 2$ microphones. The filled squares indicate non-zero matrix elements.

refer to $\hat{\mathbf{w}}_p(l)$ as the filter partition p at frequency bin k, despite the fact that this terminology is uncommon in literature on (WOLA-based) STFT processing.

Aiming to reduce the complexity of the Kalman filter, we assume that the cross-partition submatrices of the state error correlation matrix $\underline{\boldsymbol{\Psi}}_w(l)$ may be neglected in the estimation. This assumption directly corresponds to an assumption previously introduced in PBFD Kalman filtering [11, 14, 15], namely that the variations in different filter partitions are mutually uncorrelated and have zero mean, however has so far not been made in WOLA-based STFT Kalman filtering. Fig. 2 visualizes the structure of $\underline{\boldsymbol{\Psi}}_w(l)$ for both (a) preserved and (b) omitted cross-partition error correlations, with the latter resulting in a block-diagonal matrix.

In order to enforce block-diagonality of $\underline{\boldsymbol{\Psi}}_w(l)$ throughout the update equations (9)–(15), it is sufficient to simplify the product $\underline{\mathbf{k}}(l)\underline{\mathbf{x}}^T(l-D)$ in (15) as follows. Noting that $\underline{\mathbf{k}}(l)$ exhibits a vertical blockwise structure of P partitions $\mathbf{k}_p(l) \in \mathbb{C}^M$, i.e.

$$\underline{\mathbf{k}}(l) = \left(\mathbf{k}_0^T(l) \;\cdots\; \mathbf{k}_{P-1}^T(l)\right)^T, \tag{16}$$

and that $\underline{\mathbf{x}}^T(l-D)$ exhibits an analogous horizontal blockwise structure as given in (3), we can simplify $\underline{\mathbf{k}}(l)\underline{\mathbf{x}}^T(l-D)$ by

$$\begin{aligned}\underline{\mathbf{k}}(l)\underline{\mathbf{x}}^T(l-D) := \operatorname{bdiag}\big\{&\mathbf{k}_0(l)\mathbf{x}^T(l-D),\ \ldots,\\ &\mathbf{k}_{P-1}(l)\mathbf{x}^T(l-D-P+1)\big\}.\end{aligned} \tag{17}$$

Herein, the operator bdiag$\{\cdot\}$ arranges its matrix arguments on the main diagonal of a block-diagonal matrix. This simplification will indeed cause $\underline{\boldsymbol{\Psi}}_w(l)$ to remain block-diagonal, provided that its initial value and the process equation parameters $\underline{\mathbf{A}}(l)$ and $\underline{\boldsymbol{\Psi}}_{\Delta_w}(l)$ are chosen to have the same form. The proposed simplified algorithm may then alternatively be implemented as P individual Kalman filters sharing the same error signal $e(l)$, whereby Kalman filter p with the underlying state $\mathbf{w}_p(l)$ processes the input signal frame $\mathbf{x}(l-D-p)$.

Table 1 provides an overview on the complexity per frequency bin of the update equations in terms of multiplications and divisions on different domains, assuming $\underline{\mathbf{A}}(l)$ and $\underline{\boldsymbol{\Psi}}_{\Delta_w}(l)$ to be diagonal. We find that the overall computational complexity is reduced from $\mathcal{O}(P^2M^2)$ to $\mathcal{O}(PM^2)$, i.e. from quadratic to linear in P, if the cross-partition error correlations are omitted. This complexity reduction naturally comes at the expense of a performance degradation, which will be investigated in Section 6.

Note that we may reduce the computational cost even further from $\mathcal{O}(PM^2)$ to $\mathcal{O}(PM)$ by omitting the cross-microphone error correlations, enforcing $\underline{\boldsymbol{\Psi}}_w(l)$ to be fully diagonal instead of block-diagonal, leading to further performance degradation. As we usually find $P \gg M$ in practical applications however, we limit our discussion to the $\mathcal{O}(PM^2)$ case in this paper.

Table 1: Complexity of the Kalman filter update equations for (a) preserved and (b) omitted cross-partition error correlations. The simplified version of (15) employing (17) is denoted by (15)′.

Eq. \ Domain		$\mathbb{R} \times \mathbb{R}$	$\mathbb{R} \times \mathbb{C}$	$\mathbb{C} \times \mathbb{C}$
(9)	(a,b)	0	PM	0
(10)	(a)	$2PM$	P^2M^2	0
	(b)		PM^2	
(11)	(a,b)	0	0	PM
(12)	(a)	0	PM	P^2M^2
	(b)			PM^2
(13)	(a)	0	$2PM$	$P^2M^2 - PM$
	(b)			$PM^2 - PM$
(14)	(a,b)	0	0	PM
(15)	(a)	0	PM	$1.5P^2M^2 - 0.5PM$
(15)′	(b)			$1.5PM^2 - 0.5PM$
$\sum$	(a)	$2PM$	$P^2M^2 + 5PM$	$3.5P^2M^2 + 0.5PM$
	(b)		$PM^2 + 5PM$	$3.5PM^2 + 0.5PM$

4. TARGET COMPONENT PSD ESTIMATION

The Kalman filter requires an estimate of the correlation $\psi_{x_{\mathrm{t}|0}}(l)$, which corresponds to the PSD of the target component $x_{\mathrm{t}|0}(l)$. We model the late reverberation as an isotropic sound field [16], i.e. we assume the reverberation components $x_{\mathrm{r}|m}(l)$ share the same PSD $\psi_{x_\mathrm{r}}(l)$. The target component $x_{\mathrm{t}|0}(l)$ and the reverberation component $x_{\mathrm{r}|0}(l)$ are further presumed to be uncorrelated, i.e.

$$\psi_{x_0}(l) = \psi_{x_{\mathrm{t}|0}}(l) + \psi_{x_\mathrm{r}}(l). \tag{18}$$

An estimate of $\psi_{x_{\mathrm{t}|0}}(l)$ may then be obtained from an estimate of $\psi_{x_\mathrm{r}}(l)$. In [16], an estimation procedure for $\psi_{x_\mathrm{r}}(l)$ relying on the eigenvalue decomposition (EVD) has been proposed. Unlike other methods such as e.g. [19, 20], the EVD-based procedure does not require any knowledge on the direction of arrival or early relative transfer functions and will briefly be reviewed in the following. Let $\boldsymbol{\Psi}_x(l) = E\{\mathbf{x}(l)\mathbf{x}^H(l)\}$ denote the multi-channel correlation matrix of the microphone signals, with $(\cdot)^H$ and $E\{\cdot\}$ denoting the complex conjugate transpose and the expected value operation, respectively. The matrix $\boldsymbol{\Psi}_x(l)$ can be estimated from the microphone signals directly. Let $\boldsymbol{\Gamma}$ denote the (time-invariant) spatial coherence matrix of an isotropic sound field, which may be computed analytically given the geometry of the microphone array. We then perform the EVD of the pre-whitened correlation matrix $\boldsymbol{\Psi}_x(l)\boldsymbol{\Gamma}^{-1}$,

$$\boldsymbol{\Psi}_x(l)\boldsymbol{\Gamma}^{-1} = \mathbf{V}(l)\boldsymbol{\Lambda}_x(l)\mathbf{V}^{-1}(l), \tag{19}$$

with $\mathbf{V}(l)$ a matrix of eigenvectors and $\boldsymbol{\Lambda}_x(l)$ the corresponding diagonal matrix of M eigenvalues $\lambda_{x|m}(l)$. As shown in [16], all eigenvalues except their maximum theoretically correspond to the reverberation component PSD, while the maximum eigenvalue additionally depends on the target component. Hence, the late reverberation PSD $\psi_{x_r}(l)$ can be computed as

$$\psi_{x_r}(l) = \frac{1}{M-1}\Big(\sum_{m=0}^{M-1} \lambda_{x|m}(l) - \max_m \lambda_{x|m}(l)\Big). \tag{20}$$

From an estimate of $\psi_{x_\mathrm{r}}(l)$ based on (20), we can estimate $\psi_{x_{\mathrm{t}|0}}(l)$ by means of an a-priori signal-to-reverberation ratio (SRR) estimate based on the decision-directed approach [21].

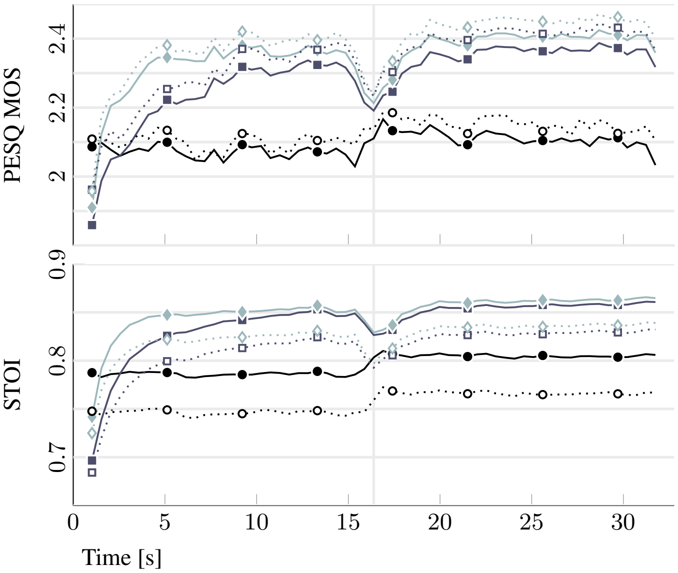

Figure 3: Averaged PESQ MOS and STOI scores for the microphone signal $\mathbf{x}_0(l)$ [—●—], the MCLP output signal $e(l)$ employing the quadratic- [—◆—] and the linear-cost [—■—] Kalman filter, and the respective spectrally post-processed signals [···○···, ···◇···, ···□···]. The vertical line indicates the time instant of the RIR change.

5. SPECTRAL POST-PROCESSING

Residual reverberation may be suppressed by applying a spectro-temporal gain $g(l) \in \mathbb{R}$ to the MCLP output signal $e(l)$, yielding the spectrally post-processed output signal $\tilde{e}(l) = g(l)e(l)$. It is well known that the so-called Wiener gain is given by the ratio of the target signal PSD to the input signal PSD of the post-processor, i.e. in our case $g(l) = \psi_{x_{\mathrm{t}|0}}(l)/\psi_e(l)$. As noted previously, the Kalman filter relates $\psi_{x_{\mathrm{t}|0}}(l)$ and $\psi_e(l)$ via (12). The Wiener gain can hence be implemented using these readily available PSD estimates. To achieve smoother gain transitions from one frame to the next, we apply exponential smoothing in our implementation, i.e. instead of using $g(l) = \psi_{x_{\mathrm{t}|0}}(l)/\psi_e(l)$ we compute the gain as

$$g(l) = \beta g(l-1) + (1-\beta)\frac{\psi_{x_{\mathrm{t}|0}}(l)}{\psi_e(l)}, \tag{21}$$

wherein $\beta \in (0, 1]$ denotes a smoothing factor.

6. SIMULATIONS

In our simulations, the STFT analysis and synthesis is based on square-root Hann windows of 512 samples with 50% overlap at 16 kHz. The prediction delay D is set to one. The adaptive filter has $P = 19$ partitions. We initialize the filter coefficients as $\hat{\underline{\mathbf{w}}}^+(0) = \mathbf{0}$, while the initial state error correlation matrix $\underline{\mathbf{\Psi}}_w^+(0)$ is chosen to be diagonal in all simulations. Expecting lower values for later coefficients of $\underline{\mathbf{w}}(l)$, we chose the diagonal elements of $\underline{\mathbf{\Psi}}_w^+(0)$ corresponding to partition $p+1$ to have 3 dB less power than those of partition p. We set the process noise correlation matrix to $\underline{\mathbf{\Psi}}_{\Delta_w}(l) = \alpha \underline{\mathbf{\Psi}}_w^+(0)$ with $\alpha = -25$ dB. For the state transition matrix $\underline{\mathbf{A}}(l)$, we choose an identity matrix scaled by $\sqrt{1-\alpha}$. The spectral gain smoothing factor is set to $\beta = 0.85$.

Measured RIRs [22] of $M = 3$ microphones with 8 cm spacing and 610 ms reverberation time are used. In order to investigate the adaptive behavior, we simulate a transition between two different source positions. Initially, the speech source is positioned at 2 m distance in the broadside direction of the microphone array, and then shifted by 15° after 16 s of simulation. Male speech [23] is chosen for the source signal. We simulate 128 realizations, each using two randomly selected 15.5 s long segments of the speech file. Despite additive noise not being modelled explicitly in (1), we add a low amount of incoherent white noise to the synthesized microphone signals at a signal-to-noise ratio of 50 dB for the sake of realism.

For each of the 128 realizations, MCLP output signals $e(l)$ are computed employing the standard quadratic-cost and the proposed linear-cost Kalman filter. A spectral gain according to (21) is applied to the MCLP output $e(l)$, while the spectral gain for the microphone signal $x_0(l)$ is computed directly from the a-priori SRR in the decision-directed approach. Objective measures are computed within windows of 2 s and 75% overlap for the microphone signal, the MCLP output signals, and the respective spectrally post-processed signals. The results are averaged over all realizations.

As objective measures, we select the perceptual evaluation of speech quality (PESQ) measure [24] with mean opinion scores (MOS) $\in [-0.5, 4.5]$ and the short-time objective intelligibility measure (STOI) [25] $\in [0, 1]$. We choose the direct component of $x_0(l)$ as a clean reference signal, defined from a window of 1 ms around the maximum peak of the corresponding RIR.

The simulation results are shown in Fig. 3. For both measures, a significant improvement can be seen for each of the two MCLP output signals [—■—, —◆—] over the microphone signal [—●—]. At convergence, quadratic-cost MCLP [—◆—] reaches an improvement of roughly 0.3 and 0.06 in terms of PESQ MOS and STOI, respectively. Compared to quadratic-cost MCLP, a small degradation can be seen for linear-cost MCLP [—■—] in both measures. The latter algorithm further shows a somewhat slower convergence after initialization, however not after the RIR change. While PESQ MOS predicts a slight improvement for all spectrally post-processed signals [···○···, ···□···, ···◇···] of about 0.05 with respect to the unprocessed signals, STOI predicts a significant degradation.

Informal listening tests indicate only faint differences between the two MCLP output signals, but a major advantage over the microphone signal. The spectral post-processing is considered advantageous for all signals. Audio examples are available online [26].

7. CONCLUSION

In this paper, we have presented a simplification to the Kalman filter solution for adaptive MCLP-based speech dereverberation formulated in the STFT domain. The simplification leads to a reduced computational cost that is linear in the number of filter coefficients instead of quadratic. Residual reverberation is suppressed using a Wiener-gain spectral post-processor subsequent to MCLP, whereby the gain computation relies on PSD estimates readily available from the Kalman filter update equations.

Simulation results indicate overall good speech quality of the enhanced signal for both the quadratic- and the linear-cost Kalman filter, with only a minor performance degradation for the latter. Spectral post-processing further improves the perceptual quality.

8. ACKNOWLEDGMENT

The authors express their gratitude towards Ina Kodrasi for providing an implementation of the PSD estimator presented in [16].

9. REFERENCES

[1] M. Triki and D. T. M. Slock, "Blind dereverberation of a single source based on multichannel linear prediction," in *Proc. Int. Workshop Acoustic Echo Noise Control (IWAENC 2005)*, Eindhoven, Netherlands, Sep. 2005, pp. 173–176.

[2] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 430–440, Jan. 2007.

[3] 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. Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2008)*, Las Vegas, USA, Apr. 2008, pp. 85–88.

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

[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, Aug. 2010.

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

[7] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, June 2015.

[8] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2009)*, Taipei, Taiwan, Apr. 2009, pp. 3733–3736.

[9] T. Yoshioka, "Dereverberation for reverberation-robust microphone arrays," in *Proc. 21st European Signal Process. Conf. (EUSIPCO 2013)*, Marrakech, Morocco, Sep. 2013, pp. 1 – 5.

[10] A. Jukić, T. van Waterschoot, and S. Doclo, "Adaptive speech dereverberation using constrained sparse multichannel linear prediction," *IEEE Signal Process. Letters*, vol. 24, no. 1, pp. 101–105, Jan. 2017.

[11] T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Partitioned block frequency domain Kalman filter for multi-channel linear prediction based blind speech dereverberation," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC 2016)*, Xi'An, China, Sep. 2016, pp. 1–5.

[12] S. Braun and E. A. P. Habets, "Online dereverberation for dynamic scenarios using a Kalman filter with and autoregressive model," *IEEE Signal Process. Letters*, vol. 23, no. 12, pp. 1741–1745, Dec. 2016.

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

[14] F. Kuech, E. Mabande, and G. Enzner, "State-space architecture of the partitioned-block-based acoustic echo controller," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2014)*, Florence, Italy, July 2014, pp. 1295–1299.

[15] M. L. Valero, E. Mabande, and E. A. P. Habets, "A state-space partitioned-block adaptive filter for echo cancellation using inter-band correlations in the Kalman gain computation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2015)*, Brisbane, Australia, Apr. 2015, pp. 599–603.

[16] I. Kodrasi and S. Doclo, "Late reverberant power spectral density estimation based on an eigenvalue decomposition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc. (ICASSP 2017)*, New Orleans, USA, Mar. 2017, pp. 611–615.

[17] 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, Apr. 2007.

[18] S. Haykin, *Adaptive Filter Theory*. Prentice-Hall, 2002, vol. 4th edition.

[19] S. Braun and E. A. P. Habets, "A multichannel diffuse power estimator for dereverberation in the presence of multiple sources," *EURASIP Journal Applied Signal Process.*, pp. 1–14, Dec. 2015.

[20] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum Likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1595–1608, Sep. 2016.

[21] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[22] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC 2014)*, Antibes – Juan les Pins, France, Sept. 2014, pp. 313–317.

[23] Bang and Olufsen, "Music for Archimedes," Compact Disc B&O, 1992.

[24] ITU-T, "Perceptual evaluation of of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," in *ITU-T Recpmmendation P.862, Int. Telecommun. Union*, Geneva, Switzerland, Feb. 2001.

[25] 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, Sep. 2011.

[26] T. Dietzen, "Audio examples for WASPAA 2017," ftp://ftp.esat.kuleuven.be/pub/SISTA/tdietzen/reports/waspaa17/audio, Apr. 2017.