PARTITIONED BLOCK FREQUENCY DOMAIN KALMAN FILTER FOR MULTI-CHANNEL LINEAR PREDICTION BASED BLIND SPEECH DEREVERBERATION

T. Dietzen[*] [‡], *A. Spriet*[*], *W. Tirry*[*], *S. Doclo*[†], *M. Moonen*[‡], *T. van Waterschoot*[‡] [§]

[*] NXP Software, Leuven, Belgium
[†] University of Oldenburg, Dept. of Medical Physics and Acoustics
and the Cluster of Excellence Hearing4All, Oldenburg, Germany
[‡] KU Leuven, Dept. of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems,
Signal Processing and Data Analytics, Leuven, Belgium
[§] KU Leuven, Dept. of Electrical Engineering (ESAT-ETC), Geel, Belgium

ABSTRACT

The multi-channel linear prediction framework for blind speech dereverberation has gained increased popularity over the recent years. While adaptive dereverberation is desirable, most multi-channel linear prediction algorithms are based on either batch or iterative frame-by-frame processing, where individual frames are treated independently. In this paper, we derive a partitioned block frequency domain Kalman filter that offers adaptive processing. The so-called excessive whitening problem is avoided by including an estimate of the target speech signal coloration in the filter update. The impact of constraining the state covariance matrix is discussed. The convergence behavior of the algorithm is evaluated in terms of the evolution of the room acoustical parameters direct-to-reverberant ratio, clarity index and early decay time, indicating good dereverberation performance.

***Index Terms*—** Dereverberation, multi-channel linear prediction, Kalman filter, partitioned block frequency domain

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, an array-based framework known as multi-channel linear prediction (MCLP) [1–11] has gained increased popularity for blind speech dereverberation, where no prior knowledge on the room impulse responses (RIRs) between the speech source and the microphone array is required. According to the multiple input/output inverse theorem (MINT) [12], 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. In a blind scenario, however, an ambiguity persists between the coloration of the clean speech signal and the room transfer functions, potentially causing undesired equalization of the inherent speech signal coloration at the output of the MCLP processing [2, 3]. This effect is known as excessive whitening. Several MCLP approaches alleviating the excessive whitening problem have been proposed, e.g., pre-whitening of the microphone signals [1], recovery of the speech signal coloration at the excessively whitened MCLP output (known as the LIME algorithm) [2], or delayed MCLP exploiting the limited autocorrelation width of speech signals [3]. Further, probabilistic approaches modeling the speech signal using a time-varying Gaussian distribution [4, 5] or using sparse priors [6] have been proposed. The majority of the proposed algorithms work in the STFT domain, as proposed in [7]. For the noiseless case, it has been shown that MCLP can be interpreted as data-dependent beamforming for speech dereverberation. [13].

In a practical scenario, adaptive filter estimation is required in order to equalize potentially time-varying RIRs. While adaptive processing is very common in data-dependent beamforming, it is rarely found in MCLP, where most algorithms are based on either batch processing or iterative processing of individual, independent frames. Yet, three exceptions can be found [8–10]. In [8], the weighted RLS algorithm has been applied in the STFT domain. In [9], an RLS implementation of the so-called weighted prediction error method [11] in the subband time domain has been proposed. In [10], a modification of LIME has been proposed and the performance for several standard adaptive schemes like NLMS and RLS have been compared in the time domain. In acoustic echo cancellation, which is a system identification problem as opposed to the system inversion problem corresponding to dereverberation, adaptive filter estimation based on the Kalman filter in the frequency domain has been applied successfully [14–17]. Hereby, the model-based formulation of the Kalman filter is advantageous, since on the one hand it allows to explicitly define statistical random walk models for the time-varying RIR, and on the other hand guarantees fast convergence. As the Kalman filter is computationally complex, it has been proposed to diagonalize the Kalman filter, yielding an NLMS-like algorithm with inherent step size control [14]. In order to avoid long algorithmic delays due to large FFT-sizes, the Kalman filter in [16, 17] has been formulated in the partitioned block frequency domain (PBFD) [18–20]. Recently, the Kalman filter has also been used in dereverberation [21] to estimate the acoustic transmission system and the clean speech signal directly.

As a first step towards adaptive, fast converging MCLP with low algorithmic delay, we propose to apply the Kalman filter to the PBFD representation of the MCLP filter estimation problem. It is shown that the derived PBFD Kalman filter framework in principle allows

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, 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.

978-1-5090-2007-2/16/$31.00 ©2016 IEEE

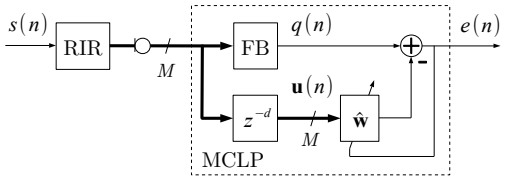

Fig. 1: MCLP in a reverberant but noise-free environment with a single speech source.

to avoid excessive whitening by including information on the target signal coloration. The PBFD Kalman filter differs from previous frequency domain derivations [14–17] in terms of the constraints included in the update equation. In section 2, the PBFD Kalman filter is derived. In Section 3, simulation results confirming the validity of the approach are presented.

2. PBFD KALMAN FILTER FOR MCLP

In order to derive the PBFD Kalman filter for MCLP, we first discuss the time domain relations, followed by the definition of the PBFD representation. Based on this, the Kalman filter state equations are formulated, from which the update equations are derived.

2.1. Time Domain Relations

The MCLP framework applied in a reverberant but noise-free environment with a single speech source is shown in Fig. 1. The microphone array composed of M microphones picks up the reverberant signals, which may be modeled as the convolution of the source signal $s(n)$ with the unknown RIRs between speech source and microphone array. In MCLP, traditionally the first microphone or the microphone closest to the source is selected to extract the linear prediction residual, but a fixed beamformer (FB) may be used as well [13]. Regardless of the actual implementation, we will refer to the output of the FB as the speech reference $q(n)$. We distinguish two components of this speech reference. Firstly, the target signal $q_t(n)$ to be maintained, which is composed of the direct component of $q(n)$ and early reflections up to a delay d, and secondly the remaining reverberation interference component $q_r(n)$ to be canceled,

$$q(n) = q_t(n) + q_r(n). \tag{1}$$

Let $u_m(n)$ denote the m^{th} microphone signal delayed by d samples, which in the MCLP framework serves as the input to the m^{th} prediction filter with filter coefficients $\hat{\mathbf{w}}_m$.

The difference between the speech reference $q(n)$ and the filter output is then given by

$$e(n) = q(n) - \sum_{m=0}^{M-1} \big(u_m * \hat{w}_m\big)(n), \tag{2}$$

where the symbol $(*)$ denotes convolution. For perfect reverberation cancellation, we seek the set of filter coefficients $\hat{\mathbf{w}}_m = \mathbf{w}_m$ that leads to $e(n) = q_t(n)$. In this case, the prediction filter output needs to equal $q_r(t)$, such that the target filter coefficients $\mathbf{w}_m$ satisfy the relation,

$$q(n) = \sum_{m=0}^{M-1} \big(u_m * w_m\big)(n) + q_t(n). \tag{3}$$

In order to estimate the filter coefficients, we transform (3) to the PBFD and apply the Kalman filter. In the Kalman filter framework, the speech reference $q(n)$ corresponds to the observation, the target filter coefficients $\mathbf{w}_m$ to the state to be estimated, and the target signal $q_t(n)$ to the observation noise.

2.2. PBFD Representation

In this subsection, we define the frame-based PBFD representation of the signals involved in (3). For details on the PBFD framework, we refer the reader to corresponding literature [18–20].

Assume that each prediction filter $\mathbf{w}_m$ consists of L_w coefficients, sectioned into B partitions of length L_{w_b} each, such that $B = \lceil L_w / L_{w_b} \rceil$. Let the vectors $\mathbf{u}_{m,b}(k) \in \mathbb{R}^N$ and $\mathbf{w}_{m,b}(k) \in \mathbb{R}^{L_{w_b}}$ respectively denote the b^{th} partition of the m^{th} input channel of the prediction filter and the prediction filter coefficients at frame k,

$$\mathbf{u}_{m,b}(k) = \big(u_m(kR - bL_{w_b} - N + 1) \ \cdots \ u_m(kR - bL_{w_b})\big)^T, \tag{4}$$

$$\mathbf{w}_{m,b}(k) = \big(w_m(bL_{w_b}, k) \ \cdots \ w_m((b+1)L_{w_b} - 1, k)\big)^T, \tag{5}$$

where $w_m(\nu, k)$ refers to the ν^{th} filter coefficient of $\mathbf{w}_m$ at frame k. In (4)–(5), the parameters $N \geq L_{w_b}$ and R denote the size of the discrete Fourier transform (DFT) and the frame shift, respectively. We further define the speech reference vector $\mathbf{q}(k) \in \mathbb{R}^V$ at frame k as

$$\mathbf{q}(k) = \big(q(kR - V + 1) \ \cdots \ q(kR)\big)^T, \tag{6}$$

and the target signal vector $\mathbf{q}_t(k)$ analogously. The parameter $V = N - L_{w_b} + 1 \geq R$ denotes the number of valid samples after fast convolution using overlap-and-save processing. Let $\mathbf{F} \in \mathbb{C}^{N \times N}$ denote the DFT-matrix. We define the frequency domain representations $\mathbf{U}_{m,b}(k)$, $\mathbf{W}_{m,b}(k)$, and $\mathbf{Q}(k)$ as

$$\mathbf{U}_{m,b}(k) = \operatorname{diag}\big\{\mathbf{F}\mathbf{u}_{m,b}(k)\big\}, \tag{7}$$

$$\mathbf{W}_{m,b}(k) = \mathbf{F}\begin{pmatrix}\mathbf{w}_{m,b}(k) \\ \mathbf{0}^{[N-L_{w_b}]}\end{pmatrix}, \tag{8}$$

$$\mathbf{Q}(k) = \mathbf{F}\begin{pmatrix}\mathbf{0}^{[N-V]} \\ \mathbf{q}(k)\end{pmatrix}. \tag{9}$$

The operation $\operatorname{diag}\{\cdot\}$ in (7) arranges the elements of its vector argument in a diagonal matrix. The notation $\mathbf{0}$ in (8)–(9) refers to a zero vector and the superscript to its dimension. The frequency domain representation $\mathbf{Q}_t(k)$ of the target signal vector $\mathbf{q}_t(k)$ is defined analogously to (9). To achieve a compact representation, we stack the matrices $\mathbf{U}_{m,b}(k)$ and vectors $\mathbf{W}_{m,b}(k)$ over all M channels and B partitions in $\mathbf{U}(k) \in \mathbb{C}^{N \times BMN}$ and $\mathbf{W}(k) \in \mathbb{C}^{BMN}$, respectively,

$$\mathbf{U}_b(k) = \big(\mathbf{U}_{0,b}(k) \ \cdots \ \mathbf{U}_{M-1,b}(k)\big), \tag{10}$$

$$\mathbf{U}(k) = \big(\mathbf{U}_0(k) \ \cdots \ \mathbf{U}_{B-1}(k)\big), \tag{11}$$

$$\mathbf{W}_b(k) = \big(\mathbf{W}_{0,b}^T(k) \ \cdots \ \mathbf{W}_{M-1,b}^T(k)\big)^T, \tag{12}$$

$$\mathbf{W}(k) = \big(\mathbf{W}_0^T(k) \ \cdots \ \mathbf{W}_{B-1}^T(k)\big)^T. \tag{13}$$

2.3. Constraining and State Equations

The objective of the Kalman filter is to estimate the filter coefficients $\mathbf{W}(k)$ that lead to the required dereverberation, i.e. resulting in the target signal $\mathbf{Q}_t(k)$ at the MCLP output. When estimating these parameters, we need to ensure that the estimated vectors comply with

the zero-padded form defined in (8) and (9). We note that $\mathbf{W}(k)$ and $\mathbf{Q}_t(k)$ fulfill the relations

$$\mathbf{W}(k) = \tilde{\mathbf{G}}\mathbf{W}(k), \tag{14}$$
$$\mathbf{Q}_t(k) = \mathbf{G}\mathbf{Q}_t(k), \tag{15}$$

where the matrices $\mathbf{G} \in \mathbb{R}^{N\times N}$ and $\tilde{\mathbf{G}} \in \mathbb{C}^{BMN\times BMN}$ are Hermitian projections, usually referred to as constraining matrices, which are defined as

$$\mathbf{G} = \mathbf{F}\mathbf{g}\mathbf{F}^{-1}, \tag{16}$$
$$\mathbf{g} = \text{blkdiag}\big\{\mathbf{0}^{[N-V\times N-V]},\ \mathbf{I}^{[V\times V]}\big\}, \tag{17}$$
$$\tilde{\mathbf{G}} = \text{blkdiag}\big\{\mathbf{F}\tilde{\mathbf{g}}\mathbf{F}^{-1},\ \ldots,\ \mathbf{F}\tilde{\mathbf{g}}\mathbf{F}^{-1}\big\}, \tag{18}$$
$$\tilde{\mathbf{g}} = \text{blkdiag}\big\{\mathbf{I}^{[L_{w_b}\times L_{w_b}]},\ \mathbf{0}^{[N-L_{w_b}\times N-L_{w_b}]}\big\}. \tag{19}$$

The operation $\text{blkdiag}\{\cdot\}$ in (17)–(19) arranges its matrix arguments in a diagonal block matrix and $\mathbf{I}$ denotes the identity matrix.

Let us now define the so-called state and observation equation, which form the basis for the derivation of the Kalman filter update equations. The state equation is generally formulated as a random walk model of the state $\mathbf{W}(k)$, while the observation equation is given as the PBFD counterpart to the time domain relation in (3), relating the definitions (9),(11), and (13). In accordance with (14)–(15), we derive,

$$\mathbf{W}(k) = \tilde{\mathbf{G}}\big(\mathbf{A}\mathbf{W}(k-1) + \mathbf{\Delta}_W(k)\big), \tag{20}$$
$$\mathbf{Q}(k) = \mathbf{C}(k)\mathbf{W}(k) + \mathbf{Q}_t(k), \tag{21}$$

with the constrained input $\mathbf{C}(k)$,

$$\mathbf{C}(k) = \mathbf{G}\mathbf{U}(k). \tag{22}$$

The random walk model in the state equation (20) with the transition matrix $\mathbf{A}$ and the random process $\mathbf{\Delta}_W(k)$ with given covariance matrix $\mathbf{\Psi}_{\Delta_W}(k)$ may be used to include statistical assumptions on the time variation of the target filter $\mathbf{W}(k)$. In order to limit the scope of this paper, we do not discuss the choice of the random walk parameters $\mathbf{A}$ and the covariance $\mathbf{\Psi}_{\Delta_W}(k)$, but instead refer to corresponding literature, e.g. [14–17]. While a time-varying model is certainly required in practice, we will assume the RIRs to be time-invariant for demonstration purposes in the simulations in section 3, i.e. we assume $\mathbf{W}(k) = \mathbf{W}(k-1)$. Note that in this particular case, the resulting Kalman filter update equations may also be interpreted as a regularized RLS algorithm [22].

The observation equation (21) is the PBFD counterpart to (3). Note that the observation $\mathbf{Q}(k)$ in the observation equation (21) will always be constrained, as it is constructed according to (9) from the observed signal $\mathbf{q}(k)$. In order to ensure that the estimation of $\mathbf{Q}_t(k)$ is constrained according to (15), we therefore only need to constrain the filter output $\mathbf{U}(k)\mathbf{W}(k)$, as shown in (21–22).

2.4. Kalman Filter Update Equations

Using Kalman filter theory, we can derive a set of update equations adaptively estimating the filter coefficients $\mathbf{W}(k)$ in the state space model (20)–(21). The Kalman filter update is given by the set of equations [23],

$$\hat{\mathbf{W}}(k) = \tilde{\mathbf{G}}\mathbf{A}\hat{\mathbf{W}}^{+}(k-1), \tag{23}$$
$$\mathbf{P}(k) = \tilde{\mathbf{G}}\big(\mathbf{A}\mathbf{P}^{+}(k-1)\mathbf{A}^{H} + \mathbf{\Psi}_{\Delta_W}(k)\big)\tilde{\mathbf{G}}^{H}, \tag{24}$$
$$\mathbf{E}(k) = \mathbf{Q}(k) - \mathbf{C}(k)\hat{\mathbf{W}}(k), \tag{25}$$
$$\mathbf{\Psi}_E(k) = \mathbf{C}(k)\mathbf{P}(k)\mathbf{C}^{H}(k) + \mathbf{\Psi}_{Q_t}(k), \tag{26}$$
$$\mathbf{K}(k) = \mathbf{P}(k)\mathbf{C}(k)^{H}\mathbf{\Psi}_E^{-1}(k), \tag{27}$$
$$\hat{\mathbf{W}}^{+}(k) = \hat{\mathbf{W}}(k) + \mathbf{K}(k)\mathbf{E}(k), \tag{28}$$
$$\mathbf{P}^{+}(k) = \big(\mathbf{I} - \mathbf{K}(k)\mathbf{C}(k)\big)\mathbf{P}(k). \tag{29}$$

The superscript $(\cdot)^H$ denotes the Hermitian transpose. Equations (23)–(24) are called the time update of the state estimate $\hat{\mathbf{W}}(k)$ and the state error covariance matrix $\mathbf{P}(k) \in \mathbb{C}^{BMN\times BMN}$. In (25)–(27), the error signal (or, in Kalman filter terminology, the innovation) $\mathbf{E}(k)$, its covariance matrix $\mathbf{\Psi}_E(k)$, and the Kalman gain $\mathbf{K}(k)$ are computed, which are then used in the so-called measurement update of the state estimate and its covariance matrix in (28)–(29). The measurement update is denoted by the superscript $(\cdot)^+$. Both $\hat{\mathbf{W}}^{+}(k)$ and $\mathbf{P}^{+}(k)$ are initialized at $k = 0$. In the remainder of this subsection, we discuss a couple of aspects of the algorithm in more detail.

Target Signal Coloration: The error signal $\mathbf{E}(k)$ in (25) is the Kalman filter estimate of the target signal $\mathbf{Q}_t(k)$, where it should be noted that its covariance matrix $\mathbf{\Psi}_E(k)$ in (26) depends on the target signal covariance matrix $\mathbf{\Psi}_{Q_t}(k)$. During convergence, the state error covariance matrix $\mathbf{P}(k)$ decreases, such that $\mathbf{\Psi}_E(k)$ converges to $\mathbf{\Psi}_{Q_t}(k)$ and hence $\mathbf{E}(k)$ converges to $\mathbf{Q}_t(k)$. In this way, the PBFD Kalman filter in principle allows to include information on the coloration of the target signal $\mathbf{Q}_t(k)$ in the filter update, thereby circumventing the excessive whitening problem. Unfortunately, the covariance $\mathbf{\Psi}_{Q_t}(k)$ is not known and so needs to be estimated. We assume $\mathbf{\Psi}_{Q_t}(k)$ to be a diagonal matrix, where the diagonal is proportional to the power spectral density (PSD) of $q_t(n)$. This assumption is commonly made and justified by the decorrelation properties of the DFT. We further note that the PSD of $q_t(n)$ is roughly given as a somewhat smoother version of the PSD of the reverberant signal $q(n)$ [24]. We therefore approximate $\mathbf{\Psi}_{Q_t}(k)$ in a simple manner as

$$\hat{\mathbf{\Psi}}_{Q_t}(k) = \text{diag}\big\{\mathbf{F}\mathbf{\Lambda}\mathbf{F}^{-1}\big(\mathbf{Q}(k)\circ\mathbf{Q}^{*}(k)\big)\big\}, \tag{30}$$

where the superscript $(\cdot)^*$ and the symbol $(\circ)$ respectively denote the conjugate of a matrix and the Hadamard product. The matrix $\mathbf{\Lambda}$ denotes a triangular window function applied in the time domain, accounting for the smoothing of the PSD estimate.

Constraining: Note that the PBFD Kalman filter equations derived in (23)–(29) slightly differ from previously presented frequency domain derivations as in [14–17], where the constraint $\tilde{\mathbf{G}}$ was not included in the state equation (20), and hence was not reflected in the time update (23–24) of $\hat{\mathbf{W}}(k)$ and $\mathbf{P}(k)$. Instead, the constraint was included in the measurement update of $\hat{\mathbf{W}}^{+}(k)$ in (28), which is not necessary in our derivation. Mathematically, it does not make a substantial difference whether the state estimate constraint of $\hat{\mathbf{W}}^{+}(k)$ is applied in the measurement update (28) or in the time update (23). The major difference between our and previous derivations therefore consists in the application of the constraint $\tilde{\mathbf{G}}$ in the time update of $\mathbf{P}(k)$ in (29). Previous derivations may hence be seen as simplifications of (23–29), where the constraint in the time update of $\mathbf{P}(k)$ is dropped. In fact, we can easily verify from (23–29) that the constraining carries over from one update to the next, i.e. that $\hat{\mathbf{W}}^{+}(k)$ and $\mathbf{P}^{+}(k)$ will be constrained if $\hat{\mathbf{W}}^{+}(k-1)$, $\mathbf{A}$, $\mathbf{\Psi}_{\Delta_W}(k)$, and $\mathbf{P}^{+}(k-1)$ are constrained. Therefore, in theory the constraint $\tilde{\mathbf{G}}$ is actually not needed at all during adaptation as long as the initial values $\hat{\mathbf{W}}^{+}(0)$ and $\mathbf{P}^{+}(0)$ are constrained, i.e. if they are chosen such

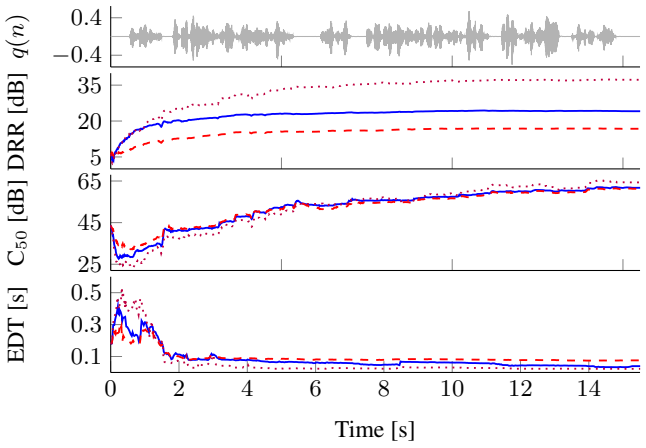

Fig. 2: The signal $q(n)$ and the evolution of the DRR, C_{50}, and EDT over time in case of (——) constrained and (- - -) unconstrained time update of $\mathbf{P}(k)$ while employing an estimate of $\mathbf{\Psi}_{Q_t}(k)$ according to (30), as well as in case of (·····) constrained time update while employing an estimate of $\mathbf{\Psi}_{Q_t}(k)$ based on the true, instantaneous PSD of $\mathbf{Q}_t(k)$.

that they satisfy $\hat{\mathbf{W}}^+(0) = \tilde{\mathbf{G}}\hat{\mathbf{W}}^+(0)$ and $\mathbf{P}^+(0) = \tilde{\mathbf{G}}\mathbf{P}^+(0)\tilde{\mathbf{G}}^H$. However, in case of error accumulation due to finite precision or simplifications of the update equations as proposed in [14–17], the constrained form of $\hat{\mathbf{W}}^+(k)$ and $\mathbf{P}^+(k)$ is no longer perfectly maintained from update to update. In these cases, constraining of the time update of both $\hat{\mathbf{W}}(k)$ and $\mathbf{P}(k)$ should be included.

Complexity: The Kalman filter is computationally extremely demanding in case of long state vectors as required here for dereverberation. For computational reasons, we therefore assume that the individual partitions $\mathbf{W}_b(k)$ may be estimated independently [16, 17]. Noting that $\mathbf{K}(k) \in \mathbb{C}^{BMN\times N}$ exhibits a vertical blockwise structure, and $\mathbf{C}(k) \in \mathbb{C}^{N\times BMN}$ a horizontal blockwise structure, with each subblock $\mathbf{K}_b(k) \in \mathbb{C}^{MN\times N}$ and $\mathbf{C}_b(k) \in \mathbb{C}^{N\times MN}$ corresponding to the b^{th} partition, this assumption corresponds to approximating the product $\mathbf{K}(k)\mathbf{C}(k)$ in (29) by a diagonal block matrix composed of the products $\mathbf{K}_b(k)\mathbf{C}_b(k)$ on the main diagonal, i.e. we assume that the off-diagonal cross-products $\mathbf{K}_b(k)\mathbf{C}_{b'}(k)$ with $b \neq b'$ may be neglected. The matrix $\mathbf{P}^+(k)$ must be initialized accordingly, i.e. all its cross-partition sub-blocks should be zero.

3. SIMULATIONS

We assume that the RIRs and hence the set of prediction filter coefficients $\mathbf{W}(k)$ in the state equation (20) are time-invariant, i.e. $\mathbf{A} = \mathbf{I}$ and $\mathbf{\Psi}_{\Delta_W}(k) = \mathbf{0}$. For simulations purposes, we use measured RIRs [25] with 360 ms reverberation time, downsampled to 16 kHz and truncated after 8000 samples. The speech source is positioned at 2 m distance in the broadside direction of the microphone array composed of $M = 3$ microphones spaced by 8 cm. The filter input delay is set to $d = 1$ and the speech reference $q(n)$ is computed using a delay-and-sum beamformer. As the source signal, a 15.5 s speech signal of a male talker is chosen [26]. The DFT-size, the frame shift and the filter partition length are set to $N = 256$ and $R = L_{w_b} = 128$, and the number of partitions to $B = 16$, yielding $L_w = 2048$ coefficients per filter channel. To illustrate the convergence behavior, we compute a number of room acoustical parameters of the overall impulse response of the room and MCLP at every frame k. In particular, we employ the direct-to-reverberant energy ratio (DRR) [27], the clarity index (C_{50}) [28], and the early decay time (EDT) [28], which correlate with both the perceived amount of reverberation and the performance of automatic speech recognition [29, 30]. For the direct component energy of the DRR, 5 ms around the maximum peak of the impulse response are considered.

In the following, we consider three different setups. The algorithm in (23–30) is evaluated firstly with and secondly without application of the constraint $\tilde{\mathbf{G}}$ in the time update of $\mathbf{P}(k)$ in (24), where the latter case corresponds to previous derivations. In order to exemplify the impact caused by omitting the constraint, we initialize $\mathbf{P}^+(k)$ as a diagonal matrix, which does *not* satisfy the constrained initialization condition discussed in subsection 2.4. The state estimate $\hat{\mathbf{W}}^+(k)$ is initialized as a zero vector. Thirdly, we additionally evaluate the fully constrained version of the algorithm using the true instantaneous PSD of the target signal $\mathbf{Q}_t(k)$, i.e. instead of estimating $\mathbf{\Psi}_{Q_t}(k)$ by (30), we employ $\hat{\mathbf{\Psi}}_{Q_t}(k) = \text{diag}\{(\mathbf{Q}_t(k) \circ \mathbf{Q}_t^*(k))\}$, which is obviously not possible in practice. The results of this simulation will provide an idea of the potential gain that may be reached by improving the estimation of the target signal PSD.

The signal $q(n)$, as well as the evolution of the DRR, C_{50}, and the EDT over time are shown for all three cases in Fig. 2. In the (——) constrained case using the estimate $\hat{\mathbf{\Psi}}_{Q_t}(k)$ as defined in (30), the DRR and C_{50} respectively increase by 18.1 dB and 18.2 dB during convergence, while the EDT drops from 180 to 40 ms, indicating good dereverberation performance. The algorithm still performs well in the (- - -) unconstrained case, but worse than in the constrained case, in particular in terms of DRR, which increases by 10.8 dB only, and the EDT, which decreases to 76 ms. Based on our simulations, the unconstrained algorithm does not appear to be able to suppress very early reflections, hence performing worse in terms of DRR, while no significant difference between constrained and unconstrained version is observed in the later reverberation. As expected, the (·····) constrained version using the true instantaneous PSD of the target signal $\mathbf{Q}_t(k)$ shows the best performance, indicating some room for improvement in the estimation of $\mathbf{\Psi}_{Q_t}(k)$. The potential improvement is mostly prominent for the DRR, where 13.1 dB may be gained additionally. Further, the EDT drops to 23 ms as opposed to 40 ms when using the estimated PSD. In all three cases, C_{50} behaves rather similar.

Audio files of the simulated signals are available at [31]. The dereverberated signals show a moderate loss in the low frequency content below about 500 Hz, caused by the diagonal approximation errors of $\hat{\mathbf{\Psi}}_{Q_t}(k)$ in this frequency range, where the diagonalization assumption that neighbouring frequency bins may be considered to be uncorrelated is less well justified due to the limited frame length.

4. CONCLUSION AND FUTURE WORK

In this paper, a PBFD Kalman filter for multi-channel linear prediction based blind speech dereverberation has been derived. The required target speech signal covariance matrix is estimated by smoothing the PSD of the reverberant speech signal. Simulations have compared the convergence behavior for both the constrained and unconstrained state error covariance in terms of room acoustical parameter improvements of the overall impulse response. The results indicate a good dereverberation performance in both cases with advantage for the constrained version. The quality of the PSD estimate has been shown to have a strong impact on the performance. Future work will focus on improving the estimation of the target signal PSD, the appropriate design of the random walk model for the filter coefficients, and further complexity reduction.

5. REFERENCES

[1] M. Triki and D. T. M. Slock, "Blind dereverberation of a single source based on multichannel linear prediction," in *Proc. 2005 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, 2007.

[3] 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, 2009.

[4] 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.

[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, 2010.

[6] 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, 2015.

[7] 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.*, Las Vegas, NV, 2008, pp. 85–88.

[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 09)*, Taipei, Taiwan, April 2009, pp. 3733–3736.

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

[10] J. M. Yang and H. G. Kang, "Online speech dereverberation algorithm based on adaptive multichannel linear prediction," *Audio, Speech, Lang. Process., IEEE/ACM Trans. on*, vol. 22, pp. 608–619, 2014.

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

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

[13] T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "On the relation between data-dependent beamforming and multichannel linear prediction for dereverberation," in *Proc. AES 60th Conf. on Dereverberation and Reverberation of Audio, Music, and Speech*, Leuven, Belgium, 2016.

[14] G. Enzner and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Processing*, vol. 86, no. 6, pp. 1140–1156, 2006.

[15] G. Enzner, *A Model-Based Optimum Filtering Approach to Acoustic Echo Control: Theory and Practice*, Ph.D. thesis, RWTH Aachen University, Aachen, Germany, 2006.

[16] 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, 2014, pp. 1295–1299.

[17] 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, 2015, pp. 599–603.

[18] J. S. Soo and K. Pang, "Multidelay block frequency domain adaptive filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, pp. 373–376, 1990.

[19] E. Moulines, O. A. Amrane, and Grenier Y., "The generalized multidelay adaptive filter: Structure and convergence analysis," *IEEE Trans. Signal Process.*, vol. 43, no. 1, pp. 14–28, 1995.

[20] P. C. W. Sommen, *Adaptive Filtering Methods: On methods to use a priori information in order to reduce complexity while maintaining convergence properties*, Ph.D. thesis, Technical University of Eindhoven, 1992.

[21] B. Schwartz, S. Gannot, and E. A. P. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 2, pp. 394–406, 2015.

[22] A. H. Sayed and T. Kailath, "A state-space approach to adaptive RLS filtering," *IEEE Signal Process. Mag.*, vol. 11, no. 3, pp. 18–60, 1994.

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

[24] Schroeder M. R. and K. H. Kuttruff, "On frequency response curves in rooms. Comparison of experimental, theoretical, and Monte Carlo results for the average frequency spacing between maxima," *The Journal of the Acoustical Society of America*, vol. 34, no. 1, pp. 76–80, 1962.

[25] 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.

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

[27] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *The Journal of the Acoustical Society of America*, vol. 112, pp. 2110–2117, 2002.

[28] "ISO 3382–1:2009. Measurement of room acoustic parameters – Part 1: Performance spaces," 2009.

[29] J. S. Bradley, "Review of objective room acoustics measures and future needs," *Applied Acoustics*, vol. 72, pp. 713–720, 2011.

[30] P. P. Parada, D. Sharma, J. Lainez, D. Barreda, T. van Waterschoot, and P. Naylor, "A single-channel non-intrusive C50 estimator correlated with speech recognition performance," *IEEE Trans. Audio Speech Lang. Process.*, vol. 24, no. 4, pp. 719–732, 2016.

[31] T. Dietzen, "Audio examples for IWAENC 2016," `ftp://ftp.esat.kuleuven.be/pub/SISTA/tdietzen/reports/iwaenc16/audio`, 2016.