VARIATIONALLY DIAGONALIZED MULTICHANNEL STATE-SPACE FREQUENCY-DOMAIN ADAPTIVE FILTERING FOR ACOUSTIC ECHO CANCELLATION

Sarmad Malik and Jacob Benesty

INRS-EMT, University of Quebec, Montreal, Canada
{malik,benesty}@emt.inrs.ca

ABSTRACT

In this contribution, we present a novel low-complexity state-space algorithm for multichannel acoustic echo cancellation. The reduction in complexity is brought about by means of top-down imposition of mutual independence on the respective acoustic echo paths within a variational Bayesian framework. This results in a fully diagonalized multichannel echo-path state estimator with a complexity that varies linearly with the channel order. The state estimator is augmented with learning rules for the model parameters that are optimal in the maximum-likelihood sense. We substantiate the efficacy of our formulation by means of simulation results in the presence of changes in the echo paths and continuous double-talk.

Index Terms— Adaptive filtering, multichannel acoustic echo cancellation, state-space estimation.

1. INTRODUCTION

The problem of multichannel acoustic echo cancellation (MAEC) [1, 2], and multichannel adaptive filtering in general [3, 4], has been a subject of considerable research over the years. As compared to the single-channel case the MAEC poses challenges of its own. Possible correlation between the respective channel excitation signals implies the absence of a unique solution [5, 6]. Consequently, approaches were proposed to weaken the relation between the channels that were based on directly altering the input signals [7, 8] or by introducing decorrelating external signals [9]. In [10], a hybrid of the two aforementioned approaches was also considered. Improvement in convergence rate of various adaptive filtering configurations was achieved in [11] by means of selective-tap update strategy.

Given an effective preprocessing stage for countering *non-uniqueness*, an MAEC approach still faces the issue of robust adaptation in the presence of changes in the echo path and continuous double-talk at the near-end. Here, the notion of optimal and adaptive step-size control becomes relevant. An overview regarding step-size control in relation to acoustic echo cancellation can be found in [12]. A rigorous frequency-domain derivation for the step-size factor is presented in [13] that has a dependence on the system misalignment covariance, which is to be estimated. Breining et al. [14] computed their system misalignment dependent step-size factor using in-filter coefficients, whereas Tourneret et al. [15] put forth an adaptation control mechanism exploiting the generalized likelihood-ratio test.

Efficient formulations in the frequency domain based on the recursive least-squares (RLS) criterion [16], which explicitly encompassed the diagonalization of the Kalman filter structure as well, opened the door for considering a frequency-domain state-space model for the purpose of single-channel acoustic echo cancellation [17]. Augmented with maximum-likelihood (ML) model parameter learning rules [18], the state-space frequency-domain adaptive filter (SSFDAF) [19] offered robust adaptation via an optimal step-size factor. In [20], the state-space methodology was extended to the multichannel case yielding the multichannel SSFDAF (MCSSFDAF).

Although the MCSSFDAF exploits submatrix-diagonality, it entails the computation of a fully populated submatrix-diagonal state-error covariance matrix exacting a complexity that varies as the cubic power of the channel order. It is important to note that the suboptimal fully-diagonalized state-space filter presented in [21] is a only a *low-complexity approximation* of MCSSFDAF and not analytically derived. In this paper, we propose the imposition of top-down mutual independence within a variational framework [22, 23] on the acoustic channels. It is shown that such an assumption renders the state-error covariance matrix fully diagonal and thus results in the variationally-diagonalized MCSSFDAF (VD-MCSSFDAF) with a complexity that varies linearly with the channel order. It is demonstrated that despite lower computational complexity as compared to the MCSSFDAF, the VD-MCSSFDAF maintains comparable convergence attributes and adapts robustly in continuous double-talk. Note that our adaptive front-end can, of course, be augmented with a preprocessing stage [5, 6] to counter the issue of *non-uniqueness*.

In Sec. 2, we introduce the frequency-domain signal model for the multichannel state-space structure. The fully diagonalized algorithm is derived in Sec. 3. Sec. 4 presents simulation results and our contribution, in the context of prior work, is concluded in Sec. 5.

We use non-bold lowercase letters for scalar quantities, bold lowercase letters for vectors, and bold uppercase letters for matrices. Frequency-domain quantities are distinguished by an underline and $\langle\cdot\rangle$ is the expectation operator. The frame shift is denoted by R, whereas M is the frame size. Superscripts T and H denote transposition and Hermitian transposition, respectively. $\mathbf{F}_M$ is the DFT matrix of size $M \times M$, whereas $\mathbf{I}_R$ is an $R \times R$ identity matrix. The symbol $\otimes$ denotes Kronecker product. Letters t and τ are sample- and frame-time indices, respectively. The notation $\mathcal{N}_c\left(\mathbf{b} \,|\, \widehat{\mathbf{b}}, \boldsymbol{\Psi}_{\mathbf{b}}\right)$ is interpreted as a complex multivariate normal [18, 24] distribution with $\widehat{\mathbf{b}}$ and $\boldsymbol{\Psi}_{\mathbf{b}}$ as the mean vector and covariance matrix, respectively, i.e.,

$$\mathcal{N}_c\left(\mathbf{b} \,|\, \widehat{\mathbf{b}}, \boldsymbol{\Psi}_{\mathbf{b}}\right) = \frac{1}{\pi^M \left|\boldsymbol{\Psi}_{\mathbf{b}}\right|^M} \exp\left\{-\left(\mathbf{b}-\widehat{\mathbf{b}}\right)^H \boldsymbol{\Psi}_{\mathbf{b}}^{-1}\left(\mathbf{b}-\widehat{\mathbf{b}}\right)\right\},$$

such that $|\cdot|$ signifies the determinant of a matrix. The symbol $\boldsymbol{\partial}_{\boldsymbol{\Psi}_{\mathbf{b}}}$ denotes an $M \times M$ diagonal-differential operator such that

$$\boldsymbol{\partial}_{\boldsymbol{\Psi}_{\mathbf{b}}} = \frac{\boldsymbol{\partial}}{\boldsymbol{\partial}\boldsymbol{\Psi}_{\mathbf{b}}} \circ \mathbf{I}_M\,,$$

where $\circ$ is the element-wise Hadamard product.

2. FREQUENCY-DOMAIN MULTICHANNEL STATE-SPACE MODEL

Consider the multiple-input-single-out (MISO) case such that loudspeaker signals $x_{n,t}$ for $n = 1, \ldots, N$ are radiated into the loudspeaker-enclosure-microphone (LEM) system and convolve linearly with the respective acoustic echo-path vectors $\mathbf{w}_{n,t}$ to yield the echo signal d_t. Addition of observation noise s_t to d_t results in the microphone observation y_t that can be mathematically stated as

$$y_t = \sum_{n=1}^{N} x_{n,t} * \mathbf{w}_{n,t} + s_t \,, \tag{1}$$

where $*$ denotes linear convolution and $d_t = \sum_{n=1}^{N} x_{n,t} * \mathbf{w}_{n,t}$. In order to proceed with frequency-domain modeling, we introduce the following $M \times 1$ definitions:

$$\underline{\mathbf{s}}_\tau = \mathbf{F}_M \boldsymbol{\Upsilon} \left[s_{\tau R-R+1}\; s_{\tau R-R+2} \ldots s_{\tau R}\right]^T \tag{2}$$

$$\underline{\mathbf{y}}_\tau = \mathbf{F}_M \boldsymbol{\Upsilon} \left[y_{\tau R-R+1}\; y_{\tau R-R+2} \ldots y_{\tau R}\right]^T \tag{3}$$

representing the frequency-domain observation-noise vector $\underline{\mathbf{s}}_\tau$ and the frequency-domain observation vector $\underline{\mathbf{y}}_\tau$, respectively, followed by the diagonal $M \times M$ definition for the nth frequency-domain loudspeaker signal as

$$\underline{\mathbf{X}}_{n,\tau} = \operatorname{diag}\left\{\mathbf{F}_M \left[x_{n,\tau R-M+1}\; x_{n,\tau R-M+2} \ldots x_{n,\tau R}\right]^T\right\} . \tag{4}$$

Note that $\mathbf{F}_M$ is the DFT-matrix of size M, $\boldsymbol{\Upsilon} = [\mathbf{0}_{R\times L} \;\; \mathbf{I}_R]^T$, and diag $\{\cdot\}$ denotes diagonalization with $L = M - R$. Seeking an overlap-save convolution, we model L non-zero coefficients of the echo-path vector:

$$\mathbf{w}_{n,t} = \left[w_{0,n,t}\; w_{1,n,t} \ldots w_{L-1,n,t}\right]^T \tag{5}$$

to obtain the nth frequency-domain $M \times 1$ echo-path vector:

$$\underline{\mathbf{w}}_{n,\tau} = \mathbf{F}_M \left[\mathbf{w}_{n,\tau R}^T \;\; \mathbf{0}_{R\times 1}^T\right]^T , \tag{6}$$

where $\mathbf{0}_{R\times 1}$ is the padding of R zeros. Using (2)–(6), we express the frequency-domain representation of (1) using overlap-save constraining as

$$\underline{\mathbf{y}}_\tau = \mathbf{G} \sum_{n=1}^{N} \underline{\mathbf{X}}_{n,\tau} \underline{\mathbf{w}}_{n,\tau} + \underline{\mathbf{s}}_\tau = \mathbf{G} \underline{\mathbf{X}}_\tau \underline{\mathbf{w}}_\tau + \underline{\mathbf{s}}_\tau \tag{7}$$

such that $\mathbf{G} = \mathbf{F}_M \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^T \mathbf{F}_M^{-1}$ places the overlap save constraints and the following multichannel definitions apply

$$\underline{\mathbf{X}}_\tau = \left[\underline{\mathbf{X}}_{1,\tau}, \ldots, \underline{\mathbf{X}}_{N,\tau}\right] , \tag{8}$$

$$\underline{\mathbf{w}}_\tau = \left[\underline{\mathbf{w}}_{1,\tau}^T, \ldots, \underline{\mathbf{w}}_{N,\tau}^T\right]^T . \tag{9}$$

We model $\underline{\mathbf{s}}_\tau$ as a zero-mean complex Gaussian random vector with $\underline{\boldsymbol{\Psi}}_{\mathbf{s},\tau} = \left\langle \underline{\mathbf{s}}_\tau \underline{\mathbf{s}}_\tau^H \right\rangle$ as its diagonal covariance matrix. We augment (7) with the first-order Markov model for the nth frequency-domain echo-path vector [17, 20]:

$$\underline{\mathbf{w}}_{n,\tau} = A\, \underline{\mathbf{w}}_{n,\tau-1} + \Delta\underline{\mathbf{w}}_{n,\tau} \tag{10}$$

to complete our multichannel state-space formulation. In (10), $0 < A < 1$ is the state-transition coefficient. The process-noise vector $\Delta\underline{\mathbf{w}}_{n,\tau}$ is again modeled as a zero-mean complex Gaussian random vector with $\underline{\boldsymbol{\Psi}}_{\Delta,n,\tau} = \left\langle \Delta\underline{\mathbf{w}}_{n,\tau} \Delta\underline{\mathbf{w}}_{n,\tau}^H \right\rangle$ as its diagonal covariance matrix. We highlight that $\boldsymbol{\Theta}_\tau = \left\{\underline{\boldsymbol{\Psi}}_{\mathbf{s},\tau}, \underline{\boldsymbol{\Psi}}_{\Delta,1,\tau}, \cdots, \underline{\boldsymbol{\Psi}}_{\Delta,N,\tau}\right\}$ are the $N+1$ model parameters. It is essential to realize that our notion of mutual independence implies that a distribution over the multichannel echo-path vector can be factorized as

$$p(\underline{\mathbf{w}}_\tau) = \prod_{n=1}^{N} p(\underline{\mathbf{w}}_{n,\tau}) . \tag{11}$$

3. VARIATIONALLY DIAGONALIZED MULTICHANNEL STATE-SPACE ALGORITHM

As we have to learn N random variables, i.e., $\underline{\mathbf{w}}_{n,\tau}$, along with the model parameter set $\boldsymbol{\Theta}_\tau$, we revert to a variational Bayesian framework [25] for obtaining the learning rules. We formulate the objective function, which is the variational lower bound (VLB) on the log-likelihood distribution [18], as

$$\ln p(\underline{\mathbf{y}}_\tau|\boldsymbol{\Theta}_\tau) = \ln \int p(\underline{\mathbf{y}}_\tau, \underline{\mathbf{w}}_\tau|\boldsymbol{\Theta}_\tau) \mathrm{d}\underline{\mathbf{w}}_\tau \tag{12}$$

$$\geq \int \ln \left[\frac{p(\underline{\mathbf{y}}_\tau|\underline{\mathbf{w}}_\tau, \boldsymbol{\Theta}_\tau) p(\underline{\mathbf{w}}_\tau|\boldsymbol{\Theta}_\tau)}{q(\underline{\mathbf{w}}_\tau)}\right] q(\underline{\mathbf{w}}_\tau) \mathrm{d}\underline{\mathbf{w}}_\tau \tag{13}$$

$$= \mathcal{L}\left[q(\underline{\mathbf{w}}_\tau), \boldsymbol{\Theta}_\tau\right] , \tag{14}$$

where $\mathcal{L}[q(\underline{\mathbf{w}}_\tau), \boldsymbol{\Theta}_\tau]$ is the VLB, $q(\underline{\mathbf{w}}_\tau)$ is the posterior distribution on the multichannel echo path that is to be estimated, and (13) manifests the utilization of the Jensen's inequality [23] and employs Bayes' theorem to factorizes the joint distribution, i.e., $p(\underline{\mathbf{y}}_\tau, \underline{\mathbf{w}}_\tau|\boldsymbol{\Theta}_\tau) = p(\underline{\mathbf{y}}_\tau|\underline{\mathbf{w}}_\tau, \boldsymbol{\Theta}_\tau) p(\underline{\mathbf{w}}_\tau|\boldsymbol{\Theta}_\tau)$. The independence assumption of (11) enables the application of the mean-filed approximation [26] to the sought posterior distribution, i.e.,

$$q(\underline{\mathbf{w}}_\tau) \approx \prod_{n=1}^{N} q(\underline{\mathbf{w}}_{n,\tau}) , \tag{15}$$

which allows us to re-write VLB as

$$\mathcal{L}[q(\underline{\mathbf{w}}_\tau), \boldsymbol{\Theta}_\tau] \approx \mathcal{L}\left[\prod_{n=1}^{N} q(\underline{\mathbf{w}}_{n,\tau}), \boldsymbol{\Theta}_\tau\right] . \tag{16}$$

The application of variational calculus [27, 28] to the VLB yields learning rules for the estimated nth channel posterior $q^\star(\underline{\mathbf{w}}_{n,\tau})$ as

$$\ln q^\star(\underline{\mathbf{w}}_{n,\tau}) = \left\langle \ln p(\underline{\mathbf{y}}_\tau, \underline{\mathbf{w}}_\tau|\boldsymbol{\Theta}_\tau) \right\rangle_{\prod_{m=1, m\neq n}^{N} q^\star(\underline{\mathbf{w}}_{m,\tau-1})} + \kappa \tag{17}$$

$$\propto \left\langle \ln p(\underline{\mathbf{y}}_\tau|\underline{\mathbf{w}}_\tau, \boldsymbol{\Theta}_\tau) p(\underline{\mathbf{w}}_\tau|\boldsymbol{\Theta}_\tau) \right\rangle_{\prod_{m=1, m\neq n}^{N} q^\star(\underline{\mathbf{w}}_{m,\tau-1})} . \tag{18}$$

Note that

$$p(\underline{\mathbf{y}}_\tau|\underline{\mathbf{w}}_\tau, \boldsymbol{\Theta}_\tau) = \mathcal{N}_c(\underline{\mathbf{y}}_\tau|\mathbf{G}\underline{\mathbf{X}}_\tau \underline{\mathbf{w}}_\tau, \underline{\boldsymbol{\Psi}}_{\mathbf{s},\tau}) \tag{19}$$

is the transmission distribution,

$$p(\underline{\mathbf{w}}_\tau|\boldsymbol{\Theta}_\tau) = \prod_{n=1}^{N} p(\underline{\mathbf{w}}_{n,\tau}|\boldsymbol{\Theta}_\tau) \tag{20}$$

is the prediction distribution [29] with

$$p(\underline{\mathbf{w}}_{n,\tau}|\mathbf{\Theta}_\tau)=\mathcal{N}_c(\underline{\mathbf{w}}_{n,\tau}|A\,\widehat{\underline{\mathbf{w}}}_{n,\tau-1},A^2\,\underline{\mathbf{P}}_{n,\tau-1}+\underline{\mathbf{\Psi}}_{\Delta,n,\tau})\,, \quad (21)$$

and $\langle\cdot\rangle_{\prod_{m=1, m\neq n}^{N} q^\star(\underline{\mathbf{w}}_{m,\tau-1})}$ and κ are expectation with respect to $\prod_{m=1, m\neq n}^{N} q^\star(\underline{\mathbf{w}}_{m,\tau-1})$ and the normalizing constant, respectively. Here, $\widehat{\underline{\mathbf{w}}}_{n,\tau-1}$ is the estimated nth state at time $\tau-1$ with

$$\underline{\mathbf{P}}_{n,\tau-1}=\left\langle\left(\underline{\mathbf{w}}_{n,\tau-1}-\widehat{\underline{\mathbf{w}}}_{n,\tau-1}\right)\left(\underline{\mathbf{w}}_{n,\tau-1}-\widehat{\underline{\mathbf{w}}}_{n,\tau-1}\right)^H\right\rangle \quad (22)$$

as the corresponding $M\times M$ state-error covariance matrix. The nth prediction distribution acts as the *pseudo-conjugate* prior and thus the estimated posterior $q^\star(\underline{\mathbf{w}}_{n,\tau})$ must also have a similar form, i.e.,

$$q^\star(\underline{\mathbf{w}}_{n,\tau})=\mathcal{N}_c(\underline{\mathbf{w}}_{n,\tau}|\widehat{\underline{\mathbf{w}}}_{n,\tau},\underline{\mathbf{P}}_{n,\tau})\,. \quad (23)$$

3.1. State Estimation

In order to obtain the recursion for the nth channel posterior, three essential steps have to be taken. First, the aforementioned normal forms of the transmission (19) and prediction (21) distributions are substituted into (18). Second, all first- and second-order expectations are resolved using the identities [29]:

$$\left\langle\underline{\mathbf{w}}_{m,\tau-1}\right\rangle_{q^\star(\underline{\mathbf{w}}_{m,\tau-1})}\doteq\widehat{\underline{\mathbf{w}}}_{m,\tau-1}\,, \quad (24)$$

$$\left\langle\underline{\mathbf{w}}_{m,\tau-1}\underline{\mathbf{w}}_{m,\tau-1}^H\right\rangle_{q^\star(\underline{\mathbf{w}}_{m,\tau-1})}\doteq\widehat{\underline{\mathbf{w}}}_{m,\tau-1}\widehat{\underline{\mathbf{w}}}_{m,\tau-1}^H+\underline{\mathbf{P}}_{m,\tau-1}\,. \quad (25)$$

Third, we compare the first- and the second-order terms in $\underline{\mathbf{w}}_{n,\tau}$ on the right-hand side of (18) with $q^\star(\underline{\mathbf{w}}_{n,\tau})$ in (23) to obtain the learning rules for the nth mean $\widehat{\underline{\mathbf{w}}}_{n,\tau}$ and the corresponding state-error covariance $\underline{\mathbf{P}}_{n,\tau}$ as

$$\widehat{\underline{\mathbf{w}}}_{n,\tau-1}^{+}=A\,\widehat{\underline{\mathbf{w}}}_{n,\tau-1}\,, \quad (26)$$

$$\underline{\mathbf{P}}_{n,\tau-1}^{+}=A^2\,\underline{\mathbf{P}}_{n,\tau-1}+\underline{\mathbf{\Psi}}_{\Delta,n,\tau}\,, \quad (27)$$

$$\underline{\boldsymbol{\mu}}_{n,\tau}=\frac{R}{M}\underline{\mathbf{P}}_{n,\tau-1}^{+}\left(\frac{R}{M}\underline{\mathbf{X}}_{n,\tau}\underline{\mathbf{P}}_{n,\tau-1}^{+}\underline{\mathbf{X}}_{n,\tau}^H+\underline{\mathbf{\Psi}}_{\mathbf{s},\tau}\right)^{-1}\,, \quad (28)$$

$$\widehat{\underline{\mathbf{y}}}_{n,\tau}=\underline{\mathbf{y}}_\tau-\mathbf{G}\sum_{\substack{m=1\\m\neq n}}^{N}\underline{\mathbf{X}}_{m,\tau}\widehat{\underline{\mathbf{w}}}_{m,\tau-1}\,, \quad (29)$$

$$\underline{\mathbf{e}}_{n,\tau}=\widehat{\underline{\mathbf{y}}}_{n,\tau}-\mathbf{G}\underline{\mathbf{X}}_{n,\tau}\widehat{\underline{\mathbf{w}}}_{n,\tau-1}^{+}\,, \quad (30)$$

$$\widehat{\underline{\mathbf{w}}}_{n,\tau}=\widehat{\underline{\mathbf{w}}}_{n,\tau-1}^{+}+\underline{\boldsymbol{\mu}}_{n,\tau}\underline{\mathbf{X}}_{n,\tau}^H\underline{\mathbf{e}}_{n,\tau}\,, \quad (31)$$

$$\underline{\mathbf{P}}_{n,\tau}=\underline{\mathbf{P}}_{n,\tau-1}^{+}-\frac{R}{M}\underline{\boldsymbol{\mu}}_{n,\tau}\underline{\mathbf{X}}_{n,\tau}^H\underline{\mathbf{X}}_{n,\tau}\underline{\mathbf{P}}_{n,\tau-1}^{+}\,, \quad (32)$$

where the superscript "+" signifies the predicted quantities. In (26)–(32), $\underline{\boldsymbol{\mu}}_{n,\tau}$, $\widehat{\underline{\mathbf{y}}}_{n,\tau}$, and $\underline{\mathbf{e}}_{n,\tau}$ are the $M\times M$ Kalman step size, $M\times 1$ effective-observation vector, and $M\times 1$ error signal, respectively, for the nth channel. It is important to note that except for (29) and (30), the approximations [1, 17, 20]

$$\mathbf{G}\underline{\mathbf{X}}_{n,\tau}\approx\frac{R}{M}\underline{\mathbf{X}}_{n,\tau} \quad (33)$$

$$\mathbf{G}\underline{\mathbf{X}}_{n,\tau}\underline{\mathbf{P}}_{n,\tau-1}^{+}\underline{\mathbf{X}}_{n,\tau}^H\mathbf{G}^H\approx\frac{R}{M}\underline{\mathbf{X}}_{n,\tau}\underline{\mathbf{P}}_{n,\tau-1}^{+}\underline{\mathbf{X}}_{n,\tau}^H \quad (34)$$

have been applied to attain a diagonalized implementation using vector arithmetic. Thereafter, given a diagonally initialized $\underline{\mathbf{P}}_{n,\tau-1}$ the recursion (26)–(32) perpetually remains diagonal, and the $M\times M$ matrix inverse in (28) boils down to simple inversion of a diagonal matrix. Using the following multichannel definitions:

$$\widehat{\underline{\mathbf{w}}}_\tau=\left[\widehat{\underline{\mathbf{w}}}_{1,\tau}^T,\ldots,\widehat{\underline{\mathbf{w}}}_{n,\tau}^T,\ldots,\widehat{\underline{\mathbf{w}}}_{N,\tau}^T\right]^T\,, \quad (35)$$

$$\widehat{\underline{\mathbf{y}}}_\tau=\left[\widehat{\underline{\mathbf{y}}}_{1,\tau}^T,\ldots,\widehat{\underline{\mathbf{y}}}_{n,\tau}^T,\ldots,\widehat{\underline{\mathbf{y}}}_{N,\tau}^T\right]^T\,, \quad (36)$$

$$\widetilde{\underline{\mathbf{X}}}_\tau=\begin{bmatrix}\underline{\mathbf{X}}_{1,\tau} & \ldots & \mathbf{0} & \ldots & \mathbf{0}\\ \vdots & \ddots & \vdots & & \vdots\\ \mathbf{0} & \ldots & \underline{\mathbf{X}}_{n,\tau} & \ldots & \mathbf{0}\\ \vdots & & \vdots & \ddots & \vdots\\ \mathbf{0} & \ldots & \mathbf{0} & \ldots & \underline{\mathbf{X}}_{N,\tau}\end{bmatrix}\,, \quad (37)$$

$$\underline{\mathbf{P}}_\tau=\begin{bmatrix}\underline{\mathbf{P}}_{1,\tau} & \ldots & \mathbf{0} & \ldots & \mathbf{0}\\ \vdots & \ddots & \vdots & & \vdots\\ \mathbf{0} & \ldots & \underline{\mathbf{P}}_{n,\tau} & \ldots & \mathbf{0}\\ \vdots & & \vdots & \ddots & \vdots\\ \mathbf{0} & \ldots & \mathbf{0} & \ldots & \underline{\mathbf{P}}_{N,\tau}\end{bmatrix}\,, \quad (38)$$

we express the variationally-diagonalized multichannel state-space frequency-domain adaptive filter (VD-MCSSFDAF) as

$$\widehat{\underline{\mathbf{w}}}_{\tau-1}^{+}=A\,\widehat{\underline{\mathbf{w}}}_{\tau-1}\,, \quad (39)$$

$$\underline{\mathbf{P}}_{\tau-1}^{+}=A^2\,\underline{\mathbf{P}}_{\tau-1}+\underline{\mathbf{\Psi}}_{\Delta,\tau}\,, \quad (40)$$

$$\underline{\boldsymbol{\mu}}_\tau=\frac{R}{M}\underline{\mathbf{P}}_{\tau-1}^{+}\left(\frac{R}{M}\widetilde{\underline{\mathbf{X}}}_\tau\underline{\mathbf{P}}_{\tau-1}^{+}\widetilde{\underline{\mathbf{X}}}_\tau^H+\mathbf{I}_N\otimes\underline{\mathbf{\Psi}}_{\mathbf{s},\tau}\right)^{-1}\,, \quad (41)$$

$$\underline{\mathbf{e}}_\tau=\widehat{\underline{\mathbf{y}}}_\tau-(\mathbf{I}_N\otimes\mathbf{G})\,\widetilde{\underline{\mathbf{X}}}_\tau\widehat{\underline{\mathbf{w}}}_{\tau-1}^{+}\,, \quad (42)$$

$$\widehat{\underline{\mathbf{w}}}_\tau=\widehat{\underline{\mathbf{w}}}_{\tau-1}^{+}+\underline{\boldsymbol{\mu}}_\tau\widetilde{\underline{\mathbf{X}}}_\tau^H\underline{\mathbf{e}}_\tau\,, \quad (43)$$

$$\underline{\mathbf{P}}_\tau=\underline{\mathbf{P}}_{\tau-1}^{+}-\frac{R}{M}\underline{\boldsymbol{\mu}}_\tau\widetilde{\underline{\mathbf{X}}}_\tau^H\widetilde{\underline{\mathbf{X}}}_\tau\underline{\mathbf{P}}_{\tau-1}^{+}\,. \quad (44)$$

The $MN\times MN$ dimensional process noise covariance and step-size matrices, i.e., $\underline{\mathbf{\Psi}}_{\Delta,\tau}$ and $\underline{\boldsymbol{\mu}}_\tau$ respectively, are defined analogously to (38). It is evident from (38) that the VD-MCSSFDAF in (39)–(44) is fully diagonal and, unlike the submatrix-diagonal MCSSFDAF in [20] that has the complexity on the order $\mathcal{O}(N^3M+NM\log(M))$, it attains a complexity on the order $\mathcal{O}(NM+NM\log(M))$ [21], i.e., *linear* with respect to the channel order N.

3.2. Parameter Learning

Learning of the model parameters $\mathbf{\Theta}_\tau$ can be carried out in accordance with the maximum-likelihood scheme presented in [18], which entails the application of a suitable differential operator to the VLB [28]. Due to the assumption of independence, learning of the nth process noise covariance matrix remains contained in the discussion presented in [18]. Owing to the change in the observation model, however, the observation-noise covariance $\underline{\mathbf{\Psi}}_{\mathbf{s},\tau}$ requires attention. We substitute (19) and (21) into (16) and solve [30]:

$$\boldsymbol{\partial}_{\underline{\mathbf{\Psi}}_{\mathbf{s},\tau}}\mathcal{L}\left[q^\star(\underline{\mathbf{w}}_\tau),\mathbf{\Theta}_\tau\right]=\mathbf{0}_M \quad (45)$$

using (24) and (25) to obtain the learning rule for the estimate of the observation-noise covariance, (cf. (8)):

$$\widehat{\underline{\mathbf{\Psi}}}_{\mathbf{s},\tau}=\frac{R}{M}\underline{\mathbf{X}}_\tau\underline{\mathbf{P}}_\tau\underline{\mathbf{X}}_\tau^H+\widetilde{\underline{\mathbf{e}}}_\tau\widetilde{\underline{\mathbf{e}}}_\tau^H\circ\mathbf{I}_M\,, \quad (46)$$

where $\widetilde{\underline{\mathbf{e}}}_\tau=\underline{\mathbf{y}}_\tau-\mathbf{G}\underline{\mathbf{X}}_\tau\widehat{\underline{\mathbf{w}}}_\tau$ is the composite error signal. The Hadamard product in (46) follows from the definition of the diagonal-differential operator in Sec. 1.

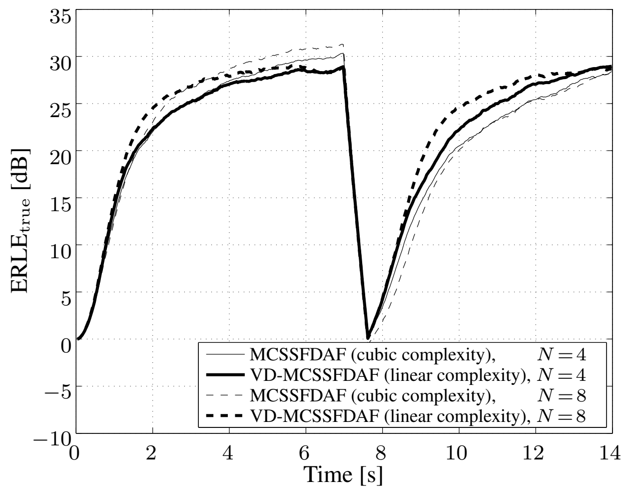

Fig. 1. Performance at ESR $=$ 30 dB: A step change is applied at time 7.0 s via regeneration of near-end room impulse responses.

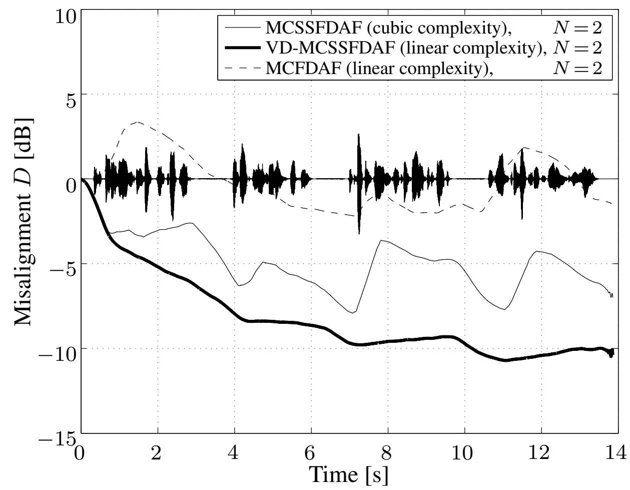

Fig. 2. Performance at ESR $=$ 0 dB: Continuous speech-speech double-talk with the depicted speech signal as the near-end signal.

4. RESULTS

For our simulations, N loudspeaker signals $x_{n,t}$ were generated by convolving a common source signal with N far-end impulse responses. Loudspeaker signals were then convolved with corresponding near-end impulse responses $\mathbf{w}_{n,t}$ and summed together with additive near-end disturbance s_t to generate the observation signal y_t. All impulse responses were randomly generated with an exponential decay corresponding to $T_{60} = 0.2\,\text{s}$. A sampling frequency of $f_\text{s} = 8$ kHz was used. The frame size and frame shift were selected as $M = 1024$ and $R = 256$, respectively, which implied an echo-path length of $M - R = 768$ samples. The true echo return loss enhancement [20]:

$$\text{ERLE}_\text{true} = 10\log_{10}\left(\frac{\sigma^2_{d_t}}{\sigma^2_{d_t-\widehat{d}_t}}\right) \tag{47}$$

and the misalignment [31]

$$D = 10\log_{10}\left(\frac{\sum_{n=1}^{N}\|\mathbf{w}_{n,t}-\widehat{\mathbf{w}}_{n,t}\|_2^2}{\sum_{n=1}^{N}\|\mathbf{w}_{n,t}\|_2^2}\right) \tag{48}$$

were employed to measure performance, where $\widehat{d}_t$ and $\widehat{\mathbf{w}}_{n,t}$ are the estimated echo signal and the estimated nth echo path, respectively.

In Fig. 1, we compare the performance of the low-complexity VD-MCSSFDAF with the submatrix-diagonal MCSSFDAF of [20], with white noise excitation selected as the source signal. The near-end white noise disturbance was added to the echo signal at an echo-to-near-end-signal ratio:

$$\text{ESR} = 10\log_{10}\left(\frac{\sigma^2_{d_t}}{\sigma^2_{s_t}}\right) \tag{49}$$

of 30 dB. The contending state-space algorithms were operated with $A = 0.9997$. It is evident for $N = 4$ as well as for $N = 8$ that despite considerable complexity reduction the derived VD-MCSSFDAF offers convergence and re-convergence properties comparable to the computationally demanding MCSSFDAF.

In order to examine the robustness of the derived algorithm, we consider a stereophonic scenario, i.e., $N = 2$, using speech excitation for both the far- and near-end signals with continuous double-talk at ESR $=$ 0 dB. Input signals were passed through positive and negative half-wave rectifiers [31] using the distortion parameter $\alpha_\text{r} = 0.4$. We consider a block-least-mean-square based multichannel frequency-domain adaptive filter (MCFDAF) [20, 32] as an additional anchor, with its error and update equations given as

$$\underline{\mathbf{e}}_\tau = \underline{\mathbf{y}}_\tau - \mathbf{G}\sum_{n=1}^{N}\underline{\mathbf{X}}_{n,\tau}\underline{\widehat{\mathbf{w}}}_{n,\tau-1} \tag{50}$$

$$\underline{\widehat{\mathbf{w}}}_{n,\tau} = \underline{\widehat{\mathbf{w}}}_{n,\tau-1} + \underline{\boldsymbol{\mu}}_\tau\underline{\mathbf{X}}^H_{n,\tau}\underline{\mathbf{e}}_\tau\,. \tag{51}$$

In (51), the step size $\underline{\boldsymbol{\mu}}_\tau = \alpha\underline{\boldsymbol{\Psi}}^{-1}_{\mathbf{x},\tau}$ is computed using the estimated frequency-domain power spectral density

$$\underline{\boldsymbol{\Psi}}_{\mathbf{x},\tau} = \gamma\underline{\boldsymbol{\Psi}}_{\mathbf{x},\tau-1} + (1-\gamma)\,\underline{\mathbf{X}}_\tau\underline{\mathbf{X}}^H_\tau \tag{52}$$

of the multichannel input signal $\underline{\mathbf{X}}_\tau$. The adaptation and the smoothing constants were set to $\alpha = 0.15$ and $\gamma = 0.9$, respectively. We can observe in Fig. 2 that the VD-MCSSFDAF, due to the incorporation of the estimated near-end noise covariance in the adaptation and mutual independence assumption, outperforms the traditional MCFDAF as well as the computationally demanding MCSSFDAF.

5. RELATION TO PRIOR WORK AND CONCLUSIONS

In [1, 16] (and references therein), efficient RLS-based multichannel frequency-domain formulations were presented for acoustic echo cancellation, which facilitated the diagonalization of the transform-domain Kalman filter in [17]. ML-optimal parameter learning rules for the single-channel state-space algorithm [19], i.e., SSFDAF, were derived in [18]. The submatrix-diagonal multichannel state-space adaptive filter, i.e., the MCSSFDAF, was presented in [20]. Motivated by the spatio-temporal decorrelation exploited in [4], this paper presents the derivation of a novel variationally diagonalized multichannel state-space algorithm for acoustic echo cancellation. The derived algorithm was evaluated in the presence of changes in the acoustic echo paths and continuous double-talk.

6. REFERENCES

[1] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*, Springer, Berlin, Germany, 2001.

[2] S. Emura, Y. Haneda, and S. Makino, "Enhanced frequency-domain adaptive algorithm for stereo echo cancellation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Orlando, FL, May 2002, vol. 2, pp. 1901–1904.

[3] K. Halwani, H. Buchner, and S. Spors, "Source-domain adaptive filtering for MIMO systems with application to acoustic echo cancellation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Dallas, TX, Mar. 2010, pp. 321–324.

[4] H. Buchner, S. Spors, and W. Kellermann, "Wave-domain adaptive filtering: Acoustic echo cancellation for full-duplex systems based on wave-field synthesis," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Montreal, QC, May 2004, pp. 117–120.

[5] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation-an overview of the fundamental problem," *IEEE Signal Process. Lett.*, vol. 2, no. 8, pp. 148–151, Aug. 1995.

[6] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 156–165, Mar. 1998.

[7] A. Sugiyama, Y. Joncour, and A. Hirano, "A stereo echo canceler with correct echo-path identication based on an input-sliding technique," *IEEE Trans. Signal Process.*, vol. 49, no. 11, pp. 2577–2587, Nov. 2001.

[8] D.-Q. Nguyen, W.-S. Gan, and A. W. H. Khong, "Time-reversal approach to the stereophonic acoustic echo cancellation problem," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 385–395, Feb. 2011.

[9] J. Herre, H. Buchner, and W. Kellermann, "Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Honolulu, HI, Apr. 2007, vol. 1, pp. 17–20.

[10] L. Romoli, S. Cecchi, L. Palestini, P. Peretti, and F. Piazza, "A novel approach to channel decorrelation for stereo acoustic echo cancellation based on missing fundamental theory," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Dallas, TX, Mar. 2010, pp. 329–332.

[11] A. W. H. Khong and P. A. Naylor, "Stereophonic acoustic echo cancellation employing selective-tap adaptive algorithms," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 785–796, May 2006.

[12] A. Mader, H. Puder, and G. U. Schmidt, "Step-size control for acoustic echo cancellation filters - an overview," *Signal Process.*, vol. 80, no. 9, pp. 1697–1719, Sep. 2000.

[13] B. H. Nitsch, "A frequency-selective stepfactor control for an adaptive filter algorithm working in the frequency domain," *Signal Process.*, vol. 80, no. 9, pp. 1733–1745, Sep. 2000.

[14] C. Breining and T. Schertler, "Delay-free low-cost step-gain estimation for adaptive filters in acoustic echo cancellation," *Signal Process.*, vol. 80, no. 9, pp. 1697–1719, Sep. 2000.

[15] J.-Y. Tourneret, N. Bershad, and J. C. M. Bermudez, "Echo cancellation - the generalized likelihood ratio test for double-talk vs. channel change," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 916–926, Mar. 2009.

[16] H. Buchner, J. Benesty, and W. Kellermann, "Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication," *Signal Process.*, vol. 85, no. 3, pp. 549–570, Mar. 2005.

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

[18] S. Malik and G. Enzner, "Online maximum-likelihood learning of time-varying dynamical models in block-frequency domain," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Dallas, TX, Mar. 2010, pp. 3822–3825.

[19] S. Malik and G. Enzner, "Model-based vs. traditional frequency-domain adaptive filtering in the presence of continuous double-talk and acoustic echo path variability," in *Proc. Int. Workshop, Acoustic Echo, Noise Control*, Seattle, WA, Sep. 2008.

[20] S. Malik and G. Enzner, "Recursive bayesian control of multichannel acoustic echo cancellation," *IEEE Signal Process. Lett.*, vol. 18, no. 11, pp. 619–622, Nov. 2011.

[21] S. Malik and G. Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 2065–2079, Sept. 2012.

[22] M. Beal and Z. Ghahramani, "The variational Kalman smoother," Tech. Rep. GCNU TR 2001-003, Gatsby Computational Neuroscience Unit, London, U.K., Apr. 2001.

[23] C. M. Bishop, *Pattern recognition and machine learning*, Springer, NY, 2006.

[24] N. R. Goodman, "Statistical analysis based on a certain multivariate complex Gaussian distribution," *Annals Math. Statist.*, vol. 34, no. 1, pp. 152–177, Mar. 1963.

[25] S. Malik and G. Enzner, "Variational Bayesian inference for nonlinear acoustic echo cancellation using adaptive cascade modeling," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Kyoto, JP, Mar. 2012.

[26] G. Parisi, *Statistical Field Theory*, Addison Wesley, New York, NY, 1988.

[27] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, London, UK, 2003.

[28] S. Malik, *Bayesian learning of linear and nonlinear acoustic system models in hands-free communication*, Ph.D. thesis, Institute of Communication Acoustics, Ruhr-University Bochum, Bochum, Germany, Oct. 2012.

[29] L. L. Scharf, *Statistical Signal Processing*, Addison-Wesley, Reading, MA, 1991.

[30] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," 2008.

[31] T. Gansler and J. Benesty, "New insights into the stereophonic acoustic echo cancellation problem and an adaptive nonlinearity solution," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 257–267, Jul. 2002.

[32] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Upper Saddle River, NJ, 2002.