A RECURSIVE GENERALIZED SIDELOBE CANCELER FOR MULTICHANNEL BLIND SPEECH DEREVERBERATION

Sarmad Malik[1], *Jacob Benesty*[1], *and Jingdong Chen*[2]

[1]INRS-EMT, University of Quebec, Montreal, Canada
[2]Northwestern Polytechnical University, Xi'an, China
{malik,benesty}@emt.inrs.ca, jingdongchen.wevoice@gmail.com

ABSTRACT

In this paper, we propose a generalized sidelobe canceler for multichannel blind speech dereverberation, which relies on recursive estimation of posterior distributions on the unknown acoustic channels and the adaptive interference canceler (AIC). Contrary to conventional design approaches where a fixed beamformer is employed, we consider a marginalized maximum-likelihood equalizer that is driven by the channel posterior estimator. It is shown that the first moment of the inferred channel posterior can also serve as a representation of an adaptive blocking matrix (ABM). Using the output of the blocking matrix, we estimate the AIC posterior to minimize the residual reverberation in the equalized signal. We demonstrate the efficacy of our approach by evaluating the algorithm in different degrees of observation noise and varying reverberation times.

Index Terms— Dereverberation, generalized sidelobe canceler, maximum likelihood, recursive Bayesian estimator

1. INTRODUCTION

Enhancement of reverberant speech has been a subject of active research over the years and it finds its application in hands-free communication and automatic speech recognition. Efforts have been made to cope with reverberant speech by means of single microphone [1] as well as multichannel approaches [2]. Multichannel methods allow the use of spatial information along with the spectral characteristics of the received signals.

In [3], a multichannel partial blind deconvolution scheme was presented for speech dereverberation by means of adaptively minimizing an information-theoretic cost function. A dereverberation approach using a time-varying Gaussian source model was formulated in [4], which was based on multichannel linear prediction. In [5], Evers *et al.* modeled the speech signal as an auto-regressive process and employed a Rao-Blackwellized particle filter to obtain an estimate of the source signal.

A frequency-domain approach was proposed in [6], which maximized the bin-wise signal-to-noise ratio (SNR) by solving a generalized eigenvalue problem. A structure resembling the generalized sidelobe canceler (GSC) [7] was outlined in [8] that relied on a fixed beamformer and a generalized eigenvector blocking matrix. This work was extended in [9] by incorporating estimated transfer function ratios within a GSC-like algorithm. The notion of adaptive blocking matrix (ABM) was also considered in [10] (and references therein), where multichannel adaptive filtering in the frequency domain was used to achieve robust speech signal acquisition.

In this work, we incorporate an ABM and a recursive interference cancellation (RIC) stage to extend the maximum-likelihood expectation-maximization blind equalization and channel identification (ML-BENCH) algorithm proposed by Schmid *et al.* in [11]. We posit that not only can the recursively estimated channel posterior be used to drive the ML-optimal equalizer but it can also be used to estimate signal components orthogonal to the desired source signal, which comprise residual reverberation and noise. Furthermore, similar to the acoustic channels we model the inference cancellation filters as mutually independent random variables with first-order Markov property. It is shown that the equalized signal and the estimate of orthogonal signal components can be used to recursively estimate the posterior distribution on the inference cancellation filters via the variationally diagonalized multichannel state-space frequency-domain adaptive filter (VD-MCSSFDAF) [12]. Simulation results show that the inclusion of the ABM and RIC stages enables the proposed recursive GSC (R-GSC) to achieve notable improvements as compared to the ML-BENCH algorithm.

In Sec. 2, we outline the signal and system model. Sec. 3 presents the formulation of the R-GSC algorithm. Simulation results are discussed in Sec. 4 followed by conclusions 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 L is the frame size. Superscripts T and H denote transposition and Hermitian transposition, respectively. $\mathbf{F}_L$ is the DFT matrix of size $L \times L$, whereas $\mathbf{I}_R$ is an $R \times R$ identity matrix. The symbol $\otimes$ denotes the Kronecker product. Letters t and n are sample- and frame-time indices, respectively. The notation $\mathcal{N}_c\left(\mathbf{b}\,|\,\widehat{\mathbf{b}}, \mathbf{\Psi_b}\right)$ is interpreted as a complex multivariate normal [13] distribution with $\widehat{\mathbf{b}}$ and $\mathbf{\Psi_b}$ as the mean vector and covariance matrix, respectively, i.e.,

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

such that $|\cdot|$ signifies the determinant of a matrix.

2. SIGNAL AND SYSTEM MODEL

Consider a source signal s_t that linearly convolves with M room impulse responses $\mathbf{w}_{m,t}$ inside a reverberant enclosure, where $m = 1, \ldots, M$. The convoluted signals are captured by M microphones in the presence of respective observation noise components $v_{m,t}$ to give the microphone observation $y_{m,t}$. We can express the formation of the mth observation signal $y_{m,t}$ as

$$y_{m,t} = \mathbf{w}_{m,t} * s_t + v_{m,t}\,, \tag{1}$$

978-1-4799-0972-8/13/$31.00 ©2013IEEE

where $*$ denotes linear convolution. In order to obtain a DFT-domain version of (1), we express $L \times 1$ frame-based definitions:

$$\underline{\mathbf{v}}_{m,n} = \mathbf{F}_L \boldsymbol{\Upsilon} \left[v_{m,nR-R+1} \; v_{m,nR-R+2} \dots v_{m,nR} \right]^T , \quad (2)$$

$$\underline{\mathbf{y}}_{m,n} = \mathbf{F}_L \boldsymbol{\Upsilon} \left[y_{m,nR-R+1} \; y_{m,nR-R+2} \dots y_{m,nR} \right]^T , \quad (3)$$

$$\underline{\mathbf{s}}_{n} = \mathbf{F}_L \left[s_{nR-L+1} \; s_{nR-L+2} \dots s_{nR} \right]^T , \quad (4)$$

$$\underline{\mathbf{w}}_{m,n} = \mathbf{F}_L \left[\mathbf{w}_{m,nR}^T \quad \mathbf{0}_{R\times 1}^T \right]^T , \quad (5)$$

for the DFT-domain mth observation noise vector, mth observation vector, source signal vector, and the mth acoustic transfer function, respectively, where $\boldsymbol{\Upsilon} = [\mathbf{0}_{R\times L-R} \;\; \mathbf{I}_R]^T$ is a padding matrix. Note that $\mathbf{w}_{m,nR} = [w_{0,m,n} \; w_{1,m,n} \dots w_{L-R-1,n,m}]^T$ in (5) is the frame-based time-domain representation of the mth room impulse response considering $L-R$ nonzero coefficients. Using (2)–(5), a DFT-domain observation model based on (1) can be expressed using overlap-save convolution as

$$\underline{\mathbf{y}}_{m,n} = \mathbf{G} \, \underline{\mathbf{W}}_{m,n} \underline{\mathbf{s}}_n + \underline{\mathbf{v}}_{m,n} , \quad (6)$$

$$= \mathbf{G} \, \underline{\mathbf{S}}_{n} \underline{\mathbf{w}}_{m,n} + \underline{\mathbf{v}}_{m,n} , \quad (7)$$

where $\mathbf{G} = \mathbf{F}_L \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^T \mathbf{F}_L^{-1}$ places the overlap-save constraint and

$$\underline{\mathbf{W}}_{m,n} = \operatorname{diag}\left\{ \underline{\mathbf{w}}_{m,n} \right\} , \quad (8)$$

$$\underline{\mathbf{S}}_{n} = \operatorname{diag}\left\{ \underline{\mathbf{s}}_{n} \right\} . \quad (9)$$

Expressions (6) and (7) are mathematically equivalent. We will use (6) for deriving an estimator for $\underline{\mathbf{s}}_n$, whereas (7) will be considered for inferring the posterior distribution on $\underline{\mathbf{w}}_{m,n}$. We model $\underline{\mathbf{w}}_{m,n}$ as a complex random vector with first-order Markov property [11] (and references therein):

$$\underline{\mathbf{w}}_{m,n} = A \, \underline{\mathbf{w}}_{m,n-1} + \Delta \underline{\mathbf{w}}_{m,n} , \quad (10)$$

where A is the state-transition coefficient and $\Delta\underline{\mathbf{w}}_{m,n}$ denotes the mth process noise term. Stacked definitions:

$$\underline{\mathbf{y}}_{n} = \left[\underline{\mathbf{y}}_{1,n}^H \cdots \underline{\mathbf{y}}_{M,n}^H \right]^H , \quad (11)$$

$$\underline{\mathbf{W}}_{n} = \left[\underline{\mathbf{W}}_{1,n}^H \cdots \underline{\mathbf{W}}_{M,n}^H \right]^H , \quad (12)$$

can be used to express the DFT-domain single-input multiple-output (SIMO) observation model as

$$\underline{\mathbf{y}}_{n} = \mathbf{G}_L \, \underline{\mathbf{W}}_{n} \underline{\mathbf{s}}_n + \underline{\mathbf{v}}_{n} , \quad (13)$$

where $\mathbf{G}_L = \mathbf{I}_L \otimes \mathbf{G}$ and $\underline{\mathbf{v}}_n$ is defined analogous to (11). We model the noise terms $\Delta\underline{\mathbf{w}}_{m,n}$ in (10) and $\underline{\mathbf{v}}_n$ in (13) as normally distributed complex random vectors with diagonal covariance matrices, i.e., $\underline{\boldsymbol{\Psi}}_{\Delta,m,n} = \langle \Delta\underline{\mathbf{w}}_{m,n} \Delta\underline{\mathbf{w}}_{m,n}^H \rangle$ and $\underline{\boldsymbol{\Psi}}_{\mathbf{v},n} = \langle \underline{\mathbf{v}}_n \underline{\mathbf{v}}_n^H \rangle$, respectively.

In order to motivate the formulations in the ensuing section, we draw the reader's attention towards the system diagram in Fig. 1 that lays out the derivational tasks for us. First, given an estimate of the posterior distribution $q_{\underline{\mathbf{w}}}^{\star}$ on the unknown acoustic channels $\underline{\mathbf{w}}_{m,t}$, we will derive the marginalized ML equalizer such that a *raw* estimate of the speech signal $\widehat{s}_{e,t}$ can be obtained. In the context of DFT-domain Gaussian state-space modeling, the acoustic channel posterior $q_{\underline{\mathbf{w}}}^{\star}$ implies

$$q_{\underline{\mathbf{w}}}^{\star} = \prod_{m=1}^{M} \mathcal{N}_c \left(\underline{\mathbf{w}}_{m,n} \,|\, \widehat{\underline{\mathbf{w}}}_{m,n}, \underline{\mathbf{P}}_{m,n} \right) , \quad (14)$$

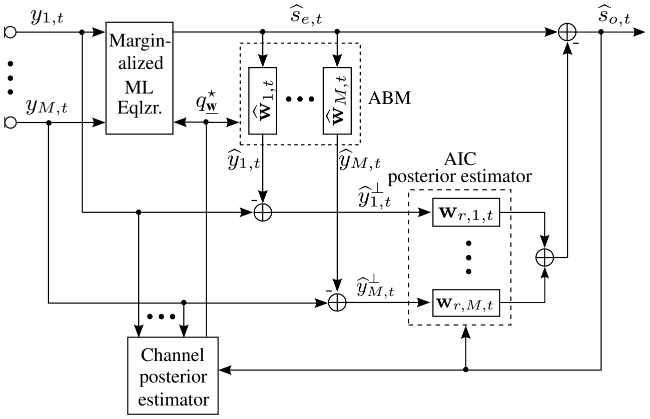

Figure 1: Time-domain depiction of the recursive generalized sidelobe canceler comprising the marginalized ML equalizer, adaptive blocking matrix (ABM), and recursive AIC and channel posterior estimators.

where $\widehat{\underline{\mathbf{w}}}_{m,n}$ and $\underline{\mathbf{P}}_{m,n}$ are the mean and state-error covariance for the mth channel [13]. Thereafter, we reuse the mean of the channel posterior as the ABM, denoted in the time domain as $\widehat{\underline{\mathbf{w}}}_{m,t}$, to compute the estimated observation signals $\widehat{y}_{m,t}$ that can be subtracted from the observation signals $y_{m,t}$ to yield a representation of orthogonal signals $\widehat{y}_{m,t}^{\perp}$. The orthogonal signals $\widehat{y}_{m,t}^{\perp}$ in conjunction with the equalized signal $\widehat{s}_{e,t}$ can be used to set up a state-space posterior estimator for recursive estimation of the interference cancellation filters $\underline{\mathbf{w}}_{r,m,t}$ to minimize the residual reverberation and noise in $\widehat{s}_{e,t}$. The output of the system $\widehat{s}_{o,t}$, which is the dereverberated source signal, can then be employed along with the microphone observation $y_{m,t}$ to update the channel posterior.

3. R-GSC ALGORITHM

3.1. Marginalized ML Equalization

The ML equalization stage can be derived by maximizing the marginalized log-likelihood function $\left\langle \ln \mathcal{N}_c \left(\underline{\mathbf{y}}_n \,|\, \mathbf{G}_L \underline{\mathbf{W}}_n \underline{\mathbf{s}}_n, \underline{\boldsymbol{\Psi}}_{\mathbf{v},n} \right) \right\rangle_{q_{\underline{\mathbf{w}}}^{\star}}$ with respect to the unknown speech signal vector $\underline{\mathbf{s}}_n$ [11], i.e.,

$$\frac{\partial}{\partial \underline{\mathbf{s}}_n^{*}} \left\langle \ln \mathcal{N}_c \left(\underline{\mathbf{y}}_n \,|\, \mathbf{G}_L \underline{\mathbf{W}}_n \underline{\mathbf{s}}_n, \underline{\boldsymbol{\Psi}}_{\mathbf{v},n} \right) \right\rangle_{q_{\underline{\mathbf{w}}}^{\star}} = \mathbf{0}_{L\times 1} , \quad (15)$$

where $\langle \cdot \rangle_{q_{\underline{\mathbf{w}}}^{\star}}$ denotes expectation with respect to $q_{\underline{\mathbf{w}}}^{\star}$ and $\partial/\partial\underline{\mathbf{s}}_n^{*}$ is the conjugate differential operator. Considering homogenous noise field [14] and diagonal approximation for the constraining matrix $\mathbf{G}_L$ [13], the marginalized ML equalizer takes the form [11]

$$\widehat{\underline{\mathbf{s}}}_{e,n} = \left[\sum_{m=1}^{M} \left(\widehat{\underline{\mathbf{W}}}_{m,n}^H \widehat{\underline{\mathbf{W}}}_{m,n} + \underline{\mathbf{P}}_{m,n} \right) \right]^{-1} \widehat{\underline{\mathbf{W}}}_{n}^H \underline{\mathbf{y}}_n , \quad (16)$$

where $\widehat{\underline{\mathbf{W}}}_{m,n} = \operatorname{diag}\left\{ \widehat{\underline{\mathbf{w}}}_{m,n} \right\}$.

3.2. Recursive Posterior Estimation for Adaptive Interference Canceler

As indicated in Fig. 1, the orthogonal signals $\widehat{y}_{m,t}^{\perp}$ can be computed by filtering the equalized signal $\widehat{s}_{e,t}$ through the estimated unknown acoustic channels $\widehat{\mathbf{w}}_{m,t}$ and subtracting the filtered output from the microphone observation $y_{m,t}$. In order to minimize the residual reverberation and noise in the output signal $\widehat{s}_{o,t}$, we devise a convolution model relating the orthogonal signals $\widehat{y}^{\perp}_{m,t}$ to the equalized signal $\widehat{s}_{e,t}$ via the interference cancellation filters $\mathbf{w}_{r,m,t}$ as

$$\widehat{s}_{e,t} = \sum_{m=1}^{M} \widehat{y}^{\perp}_{m,t} * \mathbf{w}_{r,m,t} + v^{\parallel}_{t} \,, \tag{17}$$

where $v^{\parallel}_{t}$ represents the desired signal components. It is interesting to see that the desired signal components $v^{\parallel}_{t}$ are akin to the near-end speech in multiple-input single-output (MISO) acoustic echo cancellation [12], where near-end speech is to be preserved while eliminating the unwanted echo component. Analogous to the development in (2)–(7), where a DFT-domain version of a time-domain linear convolution equation was obtained, we summarily write the DFT-domain version of (17) as [12]

$$\widehat{\ddot{\underline{\mathbf{s}}}}_{e,n} = \mathbf{G} \sum_{m=1}^{M} \widehat{\underline{\mathbf{Y}}}^{\perp}_{m,n} \underline{\mathbf{w}}_{r,m,n} + \underline{\mathbf{v}}^{\parallel}_{n} \,, \tag{18}$$

where $\widehat{\underline{\mathbf{Y}}}^{\perp}_{m,n}$ is the $L \times L$ DFT-domain representation of the mth orthogonal signal $\widehat{y}^{\perp}_{m,t}$ [c.f. (4) and (9)], $\underline{\mathbf{w}}_{r,m,n}$ signifies the mth DFT-domain interference cancellation filter defined in accordance with (5), and $\widehat{\ddot{\underline{\mathbf{s}}}}_{e,n}$ is the $L \times 1$ DFT-domain equalized signal vector defined according to (3) using $\widehat{s}_{e,t}$. Furthermore, we again impose the first-order Markov property on the unknown filters $\underline{\mathbf{w}}_{r,m,n}$, i.e.,

$$\underline{\mathbf{w}}_{r,m,n} = A\, \underline{\mathbf{w}}_{r,m,n-1} + \Delta \underline{\mathbf{w}}_{r,m,n} \,, \tag{19}$$

where $\Delta \underline{\mathbf{w}}_{r,m,n}$ denotes the process noise. The noise terms $\underline{\mathbf{v}}^{\parallel}_{n}$ and $\Delta \underline{\mathbf{w}}_{r,m,n}$ are modeled as normally distributed complex random vectors with diagonal covariance matrices $\underline{\mathbf{\Psi}}_{\mathbf{v}^{\parallel},n} = \left\langle \underline{\mathbf{v}}^{\parallel}_{n} \underline{\mathbf{v}}^{\parallel H}_{n} \right\rangle$ and $\underline{\mathbf{\Psi}}_{\Delta_r,m,n} = \left\langle \Delta \underline{\mathbf{w}}_{r,m,n} \Delta \underline{\mathbf{w}}^{H}_{r,m,n} \right\rangle$, respectively.

Additionally, we consider $\underline{\mathbf{w}}_{r,m,n}$ to be mutually independent random variables, which implies $p(\underline{\mathbf{w}}_{r,1,n}, \cdots, \underline{\mathbf{w}}_{r,M,n}) = \prod_{m=1}^{M} p(\underline{\mathbf{w}}_{r,m,n})$. Recursive posterior estimation for the state-space model described in (18) and (19) under mutual independence assumption on $\underline{\mathbf{w}}_{r,m,n}$, can be carried out using the variationally diagonalized state-space frequency-domain adaptive filter (VD-MCSSFDAF) [12]. The VD-MCSSFDAF recursion for the mth filter $\underline{\mathbf{w}}_{r,m,n}$ is then given as

$$\widehat{\underline{\mathbf{w}}}^{+}_{r,m,n-1} = A\, \widehat{\underline{\mathbf{w}}}_{r,m,n-1} \,, \tag{20}$$

$$\underline{\mathbf{P}}^{+}_{r,m,n-1} = A^2\, \underline{\mathbf{P}}_{r,m,n-1} + \underline{\mathbf{\Psi}}_{\Delta_r,m,n} \,, \tag{21}$$

$$\underline{\boldsymbol{\mu}}_{r,m,n} = \underline{\mathbf{P}}^{+}_{r,m,n-1} \left(\widehat{\underline{\mathbf{Y}}}^{\perp}_{m,n} \underline{\mathbf{P}}^{+}_{r,m,n-1} \widehat{\underline{\mathbf{Y}}}^{\perp H}_{m,n} + \frac{L}{R} \underline{\mathbf{\Psi}}_{\mathbf{v}^{\parallel},n} \right)^{-1}, \tag{22}$$

$$\underline{\mathbf{e}}_{r,m,n} = \widehat{\ddot{\underline{\mathbf{s}}}}_{e,n} - \mathbf{G} \sum_{\substack{i=1 \\ i \neq m}}^{M} \widehat{\underline{\mathbf{Y}}}^{\perp}_{i,n} \widehat{\underline{\mathbf{w}}}_{r,i,n-1} - \mathbf{G}\, \widehat{\underline{\mathbf{Y}}}^{\perp}_{m,n} \widehat{\underline{\mathbf{w}}}^{+}_{r,m,n-1} \,, \tag{23}$$

$$\widehat{\underline{\mathbf{w}}}_{r,m,n} = \widehat{\underline{\mathbf{w}}}^{+}_{r,m,n-1} + \underline{\boldsymbol{\mu}}_{r,m,n} \widehat{\underline{\mathbf{Y}}}^{\perp H}_{m,n} \underline{\mathbf{e}}_{r,m,n} \,, \tag{24}$$

$$\underline{\mathbf{P}}_{r,m,n} = \underline{\mathbf{P}}^{+}_{r,m,n-1} - \frac{R}{L} \underline{\boldsymbol{\mu}}_{r,m,n} \widehat{\underline{\mathbf{Y}}}^{\perp H}_{m,n} \widehat{\underline{\mathbf{Y}}}^{\perp}_{m,n} \underline{\mathbf{P}}^{+}_{r,m,n-1} \,, \tag{25}$$

where the superscript "+" signifies the predicted quantities. In (20)–(25), $\underline{\mathbf{P}}_{r,m,n}$, $\underline{\boldsymbol{\mu}}_{r,m,n}$, and $\underline{\mathbf{e}}_{r,m,n}$ are the $L \times L$ state-error covariance, $L \times L$ Kalman step size, and $L \times 1$ error signal, respectively, for the mth AIC filter. After executing the VD-MCSSFDAF recursion for each channel in a given iteration, the DFT-domain vector of the dereverberated output $\widehat{s}_{o,t}$ of the R-GSC as shown in Fig. 1 is computed using

$$\widehat{\underline{\mathbf{s}}}_{o,n} = \widehat{\ddot{\underline{\mathbf{s}}}}_{e,n} - \mathbf{G} \sum_{m=1}^{M} \widehat{\underline{\mathbf{Y}}}^{\perp}_{m,n} \widehat{\underline{\mathbf{w}}}_{r,m,n} \,. \tag{26}$$

3.3. Recursive Posterior Estimation for Unknown Acoustic Channels

In order to recursively estimate the posterior $q^{\star}_{\underline{\mathbf{w}}}$ on the unknown acoustic channels, we modify the observation model of (7) as

$$\underline{\mathbf{y}}_{m,n} = \mathbf{G}\, \widehat{\underline{\mathbf{S}}}_{o,n} \underline{\mathbf{w}}_{m,n} + \underline{\mathbf{v}}_{m,n} \,, \tag{27}$$

where

$$\widehat{\underline{\mathbf{S}}}_{o,n} = \mathrm{diag} \left\{ \mathbf{F}_L \left[\widehat{s}_{o,nR-L+1}\ \widehat{s}_{o,nR-L+2} \ldots \widehat{s}_{o,nR} \right]^T \right\} \,. \tag{28}$$

Thus we have altered the observation model to incorporate the dereverberated output of the R-GSC, rather than using the raw equalized signal as proposed in [11]. Considering the state-space model described by (27) and (10), the posterior distribution for the mth channel can be recursively estimated using the single-channel state-space frequency-domain adaptive filter (SSFDAF) [13, 11]. The SSFDAF recursion for the mth channel is thus given as

$$\widehat{\underline{\mathbf{w}}}^{+}_{m,n-1} = A\, \widehat{\underline{\mathbf{w}}}_{m,n-1} \,, \tag{29}$$

$$\underline{\mathbf{P}}^{+}_{m,n-1} = A^2\, \underline{\mathbf{P}}_{m,n-1} + \underline{\mathbf{\Psi}}_{\Delta,m,n} \,, \tag{30}$$

$$\underline{\boldsymbol{\mu}}_{m,n} = \underline{\mathbf{P}}^{+}_{m,n-1} \left(\widehat{\underline{\mathbf{S}}}_{o,n} \underline{\mathbf{P}}^{+}_{m,n-1} \widehat{\underline{\mathbf{S}}}^{H}_{o,n} + \frac{L}{R} \underline{\mathbf{\Psi}}_{\mathbf{v},m,n} \right)^{-1} \,, \tag{31}$$

$$\underline{\mathbf{e}}_{m,n} = \underline{\mathbf{y}}_{m,n} - \mathbf{G}\, \widehat{\underline{\mathbf{S}}}_{o,n} \widehat{\underline{\mathbf{w}}}^{+}_{m,n-1} \,, \tag{32}$$

$$\widehat{\underline{\mathbf{w}}}_{m,n} = \widehat{\underline{\mathbf{w}}}^{+}_{m,n-1} + \underline{\boldsymbol{\mu}}_{m,n} \widehat{\underline{\mathbf{S}}}^{H}_{o,n} \underline{\mathbf{e}}_{m,n} \,, \tag{33}$$

$$\underline{\mathbf{P}}_{m,n} = \underline{\mathbf{P}}^{+}_{m,n-1} - \frac{R}{L} \underline{\boldsymbol{\mu}}_{m,n} \widehat{\underline{\mathbf{S}}}^{H}_{o,n} \widehat{\underline{\mathbf{S}}}_{o,n} \underline{\mathbf{P}}^{+}_{m,n-1} \,. \tag{34}$$

In (29)–(34), $\underline{\mathbf{P}}_{m,n}$, $\underline{\boldsymbol{\mu}}_{m,n}$, $\underline{\mathbf{\Psi}}_{\mathbf{v},m,n}$, and $\underline{\mathbf{e}}_{m,n}$ are the $L \times L$ state-error covariance, $L \times L$ Kalman step size, $L \times L$ observation noise covariance, and $L \times 1$ error signal, respectively, for the mth channel.

4. SIMULATION RESULTS

We considered utterances from four male and four female speakers at a sampling frequency $f_{\mathrm{s}} = 16$ kHz. Room impulse responses (RIRs) were generated for $M = 8$ microphones using the modified image method [15]. Room size was selected as $7\,\mathrm{m} \times 5\,\mathrm{m} \times 4\,\mathrm{m}$ ($x \times y \times z$), with the source and reference microphone located at $(5\,\mathrm{m} \times 1.5\,\mathrm{m} \times 1.5\,\mathrm{m})$ and $(2\,\mathrm{m} \times 4\,\mathrm{m} \times 1.5\,\mathrm{m})$, respectively. Other microphones were positioned by successively subtracting 0.1 m from the x-coordinate of the reference microphone. For obtaining the reverberant signals, the impulse response length was selected as $T_{60} \cdot f_{\mathrm{s}}$, where reverberation times were selected in the range $0.2\,\mathrm{s} \leq T_{60} \leq 1.0\,\mathrm{s}$ in steps of $0.2\,\mathrm{s}$. The direct-to-reverberant ratio (DRR) of the generated RIRs was in the range

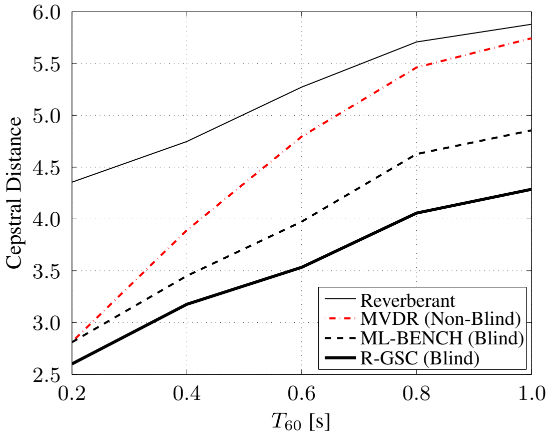

Figure 2: Cepstral distances (CDs) measured at SNR $=$ 30 dB and different reverberation conditions.

$-3.19\,\text{dB} \geq \text{DRR} \geq -15.74\,\text{dB}$. Microphone signals were generated by corrupting the reverberant signals with diffusive white noise [14] at two different SNRs, i.e., 10 and 30 dB. We considered the minimum variance distortionless response (MVDR) beamformer [16] and the ML-BENCH algorithm as the reference approaches to evaluate the proposed R-GSC. All algorithms were operated with a frame size $L = 2048$ and frame shift $R = 1024$ with the state-transition coefficient $A = 0.9997$. Noise covariance matrices were computed using the rules discussed in [11, 12, 13]. For an objective evaluation, cepstral distances (CDs) and log-likelihood ratios (LLRs) to the clean speech signal were computed after processing, according to the definitions given in [17].

In Fig. 2, we can observe that the CDs measured at SNR $=$ 30 dB for varying T_{60}. It is evident that the R-GSC considerably outperforms the MVDR and ML-BENCH approaches. For $T_{60} = 1.0\,\text{s}$, ML-BENCH alleviates the CD of the reverberant microphone signal by almost 1.02, whereas R-GSC causes an improvement of 1.59 in terms of the CD. Thus, R-GSC enhances the performance by almost 50 % as compared to the ML-BENCH algorithm. In Table 1, we again see that the R-GSC consistently achieves the best performance as compared to the contending approaches in terms of the measured LLRs for all considered reverberation times and at two different noise levels.

SNR $=$ 30 dB					
$T_{60} =$	0.2 s	0.4 s	0.6 s	0.8 s	1.0 s
Reverberant	0.57	0.57	0.74	0.82	0.87
MVDR	0.25	0.41	0.57	0.70	0.74
ML-BENCH	0.26	0.32	0.40	0.51	0.56
R-GSC	**0.22**	**0.27**	**0.32**	**0.41**	**0.43**

SNR $=$ 10 dB					
$T_{60} =$	0.2 s	0.4 s	0.6 s	0.8 s	1.0 s
Reverberant	1.38	1.40	1.41	1.43	1.44
MVDR	0.99	1.07	1.16	1.21	1.20
ML-BENCH	0.72	0.73	0.80	0.83	0.88
R-GSC	**0.66**	**0.71**	**0.76**	**0.80**	**0.85**

Table 1: Log-likelihood ratios (LLRs) measured at different SNR and reverberation conditions.

5. CONCLUSIONS

In this work, we have formulated a multichannel approach based on generalized sidelobe cancellation for blind speech dereverberation. Unlike conventional approaches, the proposed recursive generalized sidelobe canceler (R-GSC) consists of a data-dependent beamformer, i.e., marginalized ML equalizer, an adaptive blocking matrix, and two recursive Bayesian estimators for inferring posteriors on the unknown acoustic channels and the interference cancellation filters. We show by means of simulations conducted in different degrees of observation noise and varying reverberation times that the R-GSC achieves notable improvement as compared to the considered blind and non-blind reference approaches.

6. REFERENCES

[1] E. A. P. Habets, S. Gannot, and I. Cohen, “Late reverberant spectral variance estimation based on a statistical model,” *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–773, Sep. 2009.

[2] S. Doclo and M. Moonen, “GSVD-based optimal filtering for single and multimicrophone speech enhancement,” vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[3] H. Buchner, R. Aichner, and W. Kellermann, “TRINICON: A versatile framework for multichannel blind signal processing,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, May 2004, pp. 889–892.

[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, Nov. 2008.

[5] C. Evers and J. R. Hopgood, “Multichannel online blind speech dereverberation with marginalization of static observation parameters in a Rao-Blackwellized particle filter,” *J. Signal Process. Syst.*, vol. 63, no. 3, pp. 315–332, 2011.

[6] E. Warsitz and R. Haeb-Umbach, “Blind acoustic beamforming based on generalized eigenvalue decomposition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1529–1539, Jul. 2007.

[7] L. J. Griffiths and C. W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Oct. 1982.

[8] E. Warsitz, A. Krueger, and R. Haeb-Umbach, “Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, Apr. 2008, pp. 73–76.

[9] A. Krueger, E. Warsitz, and R. Haeb-Umbach, “Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 206–219, 2011.

[10] W. Herbordt, H. Buchner, S. Nakamura, and W. Kellermann, “Multichannel binwise robust frequency-domain adaptive filtering and its application to adaptive beamforming,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1340–1351, May 2007.

[11] D. Schmid, S. Malik, and G. Enzner, “An expectation-maximization algorithm for multichannel adaptive speech dereverberation in the frequency-domain,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, Mar. 2012, pp. 17–20.

[12] S. Malik and J. Benesty, “Variationally diagonalized multichannel state-space frequency-domain adaptive filtering for acoustic echo cancellation,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, May. 2013.

[13] 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. (ICASSP)*, Mar. 2010, pp. 3822–3825.

[14] E. A. P. Habets, I. Cohen, and S. Gannot, “Generating nonstationary multisensor signals under a spatial coherence constraint,” *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

[15] E. Lehmann and A. Johansson, “Prediction of energy decay in room impulse responses simulated with an image-source model,” *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 269–277, Jul. 2008.

[16] M. S. Brandstein and D. B. Ward, Eds., *Microphone arrays: Signal processing techniques and applications*, New York: Springer-Verlag, 2001.

[17] P. C. Loizou, *Speech Enhancement: Theory and Practice*, Boca Raton, FL: CRC Press, 2007.