AN EXPECTATION-MAXIMIZATION ALGORITHM FOR MULTICHANNEL ADAPTIVE SPEECH DEREVERBERATION IN THE FREQUENCY-DOMAIN

Dominic Schmid, Sarmad Malik, and Gerald Enzner

Institute of Communication Acoustics
Ruhr-Universität Bochum, 44780 Bochum, Germany
Email: {dominic.schmid, sarmad.malik, gerald.enzner}@rub.de

ABSTRACT

This paper presents an online dereverberation algorithm that is derived within the maximum-likelihood expectation-maximization (ML-EM) framework. We formulate an overlap-save observation model for the multichannel blind problem in the DFT-domain. The modeling of acoustic channel impulse responses as random variables with a first-order Markov property facilitates the ensuing algorithm to cope with time-varying conditions. We then show that the ML-EM learning rules for the multichannel state-space model at hand take the form of a recursive posterior estimator for the channels, followed by an equalization stage for recovering the speech signal subject to an expectation with respect to the estimated channel posterior. Our derivation thus results in an iterative ML algorithm for *b*lind *e*qualization a*n*d *ch*annel identification (ML-BENCH) which comprises two distinct and coupled subsystems. The dereverberation performance of the proposed system is evaluated by considering spectrograms and instrumental quality measures.

***Index Terms*—** Expectation-maximization, frequency-domain adaptive filtering, multichannel dereverberation, state-space model.

1. INTRODUCTION

The presence of reverberation can severely degrade the perceived quality and intelligibility of speech signals in hands-free communication systems. Moreover, it generally leads to a performance loss in applications such as automatic speech recognition or source localization. The task of recovering the undistorted speech signal, i.e., reducing the effects of reverberation, has thus evolved into a rapidly growing area of research over the recent years.

Many existing algorithms are based on spectral subtraction techniques which require the late reverberation to be estimated, e.g., by using statistical models [1] or long-term multi-step linear prediction [2]. Other approaches employ multichannel linear prediction [3, 4] or optimal information-theoretic inference [5] to directly estimate inverse filters that mitigate the reverberation effects. Channel-based dereverberation algorithms usually aim at blindly estimating the room impulse responses via adaptive algorithms [6] in order to recover the source signal in a subsequent multichannel equalization stage [7]. For the latter, the *m*ultiple-input/output *in*verse *t*heorem (MINT) has shown perfect dereverberation on the basis of precisely known channels [8], but it generally suffers from channel identification errors that can cause large distortions in the estimated signal. Consequently, a major focus of contemporary research has been on the design of robust equalizers [7, 9].

This work was supported by the German Research Foundation.

In this paper, we carry out a contained derivation within the expectation-maximization (EM) framework to formulate a multichannel DFT-domain dereverberation algorithm that is optimal in the maximum-likelihood (ML) sense. In our ML-EM formulation, we model the acoustic channel as a random variable with a first-order Markov property [10], which inherently enables the resulting algorithm to operate in time-varying conditions [11], whereas the speech signal is considered as an *a priori* unknown parameter. In addition to the resulting multichannel state-space model, a cost function in the form of a lower-bound on the log-likelihood is presented [12]. We then derive EM learning rules to tighten the bound and maximize the log-likelihood. The resulting ML algorithm for *b*lind *e*qualization a*n*d *ch*annel identification (ML-BENCH) eventually takes the form of an iterative system that comprises two coupled subsystems, i.e., a recursive posterior estimator for the channels in the expectation-step (E-step) and a multichannel equalization stage in the maximization-step (M-step), both of which are efficiently implementable using vector arithmetics and FFT / IFFT.

The remainder of this paper is organized as follows. In Sec. 2, we present our multichannel state-space observation model in the DFT-domain. The ML-BENCH algorithm is presented in Sec. 3, whereas the performance of the derived iterative system is evaluated in Sec. 4. Finally, we conclude our work in Sec. 5.

In our notation, 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. In addition, $\otimes$ represents the Kronecker product and $\mathcal{E}\{\cdot\}$ describes mathematical expectation. The superscript H denotes Hermitian transposition, $\mathbf{F}_M$ is the DFT-matrix of size $M \times M$, and $\mathbf{I}_R$ is an $R \times R$ identity matrix. Lowercase letters k and τ are sample- and frame-time indices, respectively, related via $k = \tau R + \nu$, $\nu = 0, 1, \ldots, R-1$, $\tau \in \mathbb{Z}$. The term R denotes the corresponding frame-shift, whereas M is the frame-size.

2. DFT-DOMAIN DYNAMICAL MODELING

We consider a speech signal s_k that is transmitted through an acoustic system and captured by P microphones at discrete time k. The ith microphone signal $y_{i,k}$ can then be expressed as

$$y_{i,k} = \mathbf{h}_{i,k}^T \, \mathbf{s}'_k + n_{i,k} \;, \quad i = 1 \ldots P \;, \tag{1}$$

where

$$\mathbf{h}_{i,k} = [\, h_{i,k,0} \; h_{i,k,1} \; \ldots \; h_{i,k,L-1} \,]^T \tag{2}$$

is the time-varying channel impulse response of length L between the source and the ith microphone. The vector $\mathbf{s}'_k = [\, s_k \; s_{k-1} \; \ldots \; s_{k-L+1} \,]^T$ contains the L most recent samples of s_k and $n_{i,k}$ represents additive observation noise of the ith channel.

2.1. Multichannel Observation Model

In order to formulate an efficient algorithm in the DFT-domain, we now consider frame-based definitions of the quantities involved in the linear convolution in (1). The frame-based source vector

$$\mathbf{s}_\tau = [\, s_{\tau R-M+1}\ s_{\tau R-M+2}\ \dots\ s_{\tau R} \,]^T \tag{3}$$

of length M can be transformed to the DFT-domain by multiplying it with the DFT-matrix $\mathbf{F}_M$, i.e.,

$$\underline{\mathbf{s}}_\tau = \mathbf{F}_M \mathbf{s}_\tau \ . \tag{4}$$

Next, we consider a frame-based version $\mathbf{w}_{i,\tau} = \mathbf{h}_{i,k=\tau R}$ of the channel impulse response $\mathbf{h}_{i,k}$. By modeling $L = M - R$ non-zero coefficients of $\mathbf{w}_{i,\tau}$ and applying the DFT-matrix $\mathbf{F}_M$, we obtain a DFT-domain representation

$$\underline{\mathbf{w}}_{i,\tau} = \mathbf{F}_M \Big[\, \mathbf{w}_{i,\tau}^T\ \mathbf{0}_{R\times 1}^T \,\Big]^T , \tag{5}$$

where $\mathbf{0}_{R\times 1}$ denotes the padding of R zeros. The ith DFT-domain channel matrix $\underline{\mathbf{W}}_{i,\tau}$ of size $M \times M$ is obtained by applying diagonalization to $\underline{\mathbf{w}}_{i,\tau}$,

$$\underline{\mathbf{W}}_{i,\tau} = \mathrm{diag}\{\underline{\mathbf{w}}_{i,\tau}\} \ . \tag{6}$$

According to (1), the overlap-save convolution of $\underline{\mathbf{W}}_{i,\tau}$ with $\underline{\mathbf{s}}_\tau$ in the presence of an additive frame-based noise vector

$$\mathbf{n}_{i,\tau} = [\, n_{i,\tau R-R+1}\ n_{i,\tau R-R+2}\ \dots\ n_{i,\tau R} \,]^T \tag{7}$$

of length R then describes the time-domain observation vector $\mathbf{y}_{i,\tau}$,

$$\mathbf{y}_{i,\tau} = \mathbf{Q}^T \mathbf{F}_M^{-1} \underline{\mathbf{W}}_{i,\tau} \underline{\mathbf{s}}_\tau + \mathbf{n}_{i,\tau} \ , \tag{8}$$

which is defined analogous to $\mathbf{n}_{i,\tau}$. Here, $\mathbf{Q} = [\, \mathbf{0}_{R\times L}\ \mathbf{I}_R \,]^T$ is an $M \times R$ matrix required for linearizing the cyclic convolution in the DFT-domain. By padding the observation vector $\mathbf{y}_{i,\tau}$ with $L = M - R$ zeros and applying $\mathbf{F}_M$, i.e., $\underline{\mathbf{y}}_{i,\tau} = \mathbf{F}_M \mathbf{Q} \mathbf{y}_{i,\tau}$, we arrive at the DFT-domain observation model for the ith channel,

$$\underline{\mathbf{y}}_{i,\tau} = \mathbf{F}_M \mathbf{Q} \mathbf{Q}^T \mathbf{F}_M^{-1} \underline{\mathbf{W}}_{i,\tau} \underline{\mathbf{s}}_\tau + \mathbf{F}_M \mathbf{Q} \mathbf{n}_{i,\tau} \ . \tag{9}$$

The constant matrix $\mathbf{T} = \mathbf{F}_M \mathbf{Q} \mathbf{Q}^T \mathbf{F}_M^{-1}$ can be combined with the channel matrix $\underline{\mathbf{W}}_{i,\tau}$ to obtain an overlap-save constrained version $\underline{\boldsymbol{\mathcal{W}}}_{i,\tau} = \mathbf{T}\, \underline{\mathbf{W}}_{i,\tau}$. This enables us to compactly express (9) as

$$\underline{\mathbf{y}}_{i,\tau} = \underline{\boldsymbol{\mathcal{W}}}_{i,\tau} \underline{\mathbf{s}}_\tau + \underline{\mathbf{n}}_{i,\tau} \ . \tag{10}$$

Here, we model $\underline{\mathbf{n}}_{i,\tau} = \mathbf{F}_M \mathbf{Q} \mathbf{n}_{i,\tau}$ as a zero-mean and normally distributed DFT-domain observation noise term of the ith channel, with $\underline{\boldsymbol{\Psi}}_{i,\tau}^{\mathbf{n}} = \mathcal{E}\{\underline{\mathbf{n}}_{i,\tau} \underline{\mathbf{n}}_{i,\tau}^H\}$ as its diagonal covariance matrix. By defining the following stacked quantities,

$$\underline{\mathbf{y}}_\tau = \big[\, \underline{\mathbf{y}}_{1,\tau}^H\ \underline{\mathbf{y}}_{2,\tau}^H\ \cdots\ \underline{\mathbf{y}}_{P,\tau}^H \,\big]^H \ , \tag{11}$$

$$\underline{\boldsymbol{\mathcal{W}}}_\tau = \big[\, \underline{\boldsymbol{\mathcal{W}}}_{1,\tau}^H\ \underline{\boldsymbol{\mathcal{W}}}_{2,\tau}^H\ \cdots\ \underline{\boldsymbol{\mathcal{W}}}_{P,\tau}^H \,\big]^H \ , \tag{12}$$

our multichannel observation model can be expressed as

$$\underline{\mathbf{y}}_\tau = \underline{\boldsymbol{\mathcal{W}}}_\tau \underline{\mathbf{s}}_\tau + \underline{\mathbf{n}}_\tau \ , \tag{13}$$

where $\underline{\mathbf{n}}_\tau$ is defined analogous to (11). If we model the noise terms $\underline{\mathbf{n}}_{i,\tau}$ to be channel-wise uncorrelated, i.e., $\mathcal{E}\{\underline{\mathbf{n}}_{i,\tau} \underline{\mathbf{n}}_{j,\tau}^H\} = \mathbf{0}_{M\times M}$, $\forall\, i \neq j$, the multichannel noise covariance matrix $\underline{\boldsymbol{\Psi}}_\tau^{\mathbf{n}} = \mathcal{E}\{\underline{\mathbf{n}}_\tau \underline{\mathbf{n}}_\tau^H\}$ maintains full diagonality. An equivalent representation of the observation model in (13) is given by

$$\underline{\mathbf{y}}_\tau = \underline{\boldsymbol{\mathcal{S}}}_\tau \underline{\mathbf{w}}_\tau + \underline{\mathbf{n}}_\tau \ , \tag{14}$$

where $\underline{\boldsymbol{\mathcal{S}}}_\tau = \mathbf{I}_P \otimes \mathbf{T} \underline{\mathbf{S}}_\tau$ is a $PM \times PM$ matrix with $\underline{\mathbf{S}}_\tau = \mathrm{diag}\{\underline{\mathbf{s}}_\tau\}$ and $\underline{\mathbf{w}}_\tau$ denotes a stacked vector defined analogous to (11) containing the channels $\underline{\mathbf{w}}_{i,\tau}$.

2.2. First-Order Markov Model for Time-Varying Channels

For incorporating the time-variability of the acoustic channels, we model $\underline{\mathbf{w}}_{i,\tau}$ as a random variable with a first-order Markov property,

$$\underline{\mathbf{w}}_{i,\tau} = A \cdot \underline{\mathbf{w}}_{i,\tau-1} + \Delta \underline{\mathbf{w}}_{i,\tau} \ . \tag{15}$$

The scalar A denotes the state-transition coefficient in the range $0 < A < 1$ and $\Delta \underline{\mathbf{w}}_{i,\tau}$ is a zero-mean and frame-wise uncorrelated process noise vector. The covariance of $\Delta \underline{\mathbf{w}}_{i,\tau}$ is then given by the diagonal matrix $\underline{\boldsymbol{\Psi}}_{i,\tau}^{\Delta} = \mathcal{E}\{\Delta \underline{\mathbf{w}}_{i,\tau} \Delta \underline{\mathbf{w}}_{i,\tau}^H\}$. Analogous to (11), we define the stacked quantity

$$\Delta \underline{\mathbf{w}}_\tau = \big[\, \Delta \underline{\mathbf{w}}_{1,\tau}^H\ \Delta \underline{\mathbf{w}}_{2,\tau}^H\ \cdots\ \Delta \underline{\mathbf{w}}_{P,\tau}^H \,\big]^H \tag{16}$$

to express (15) in the multichannel form

$$\underline{\mathbf{w}}_\tau = A \cdot \underline{\mathbf{w}}_{\tau-1} + \Delta \underline{\mathbf{w}}_\tau \tag{17}$$

with $\underline{\boldsymbol{\Psi}}_\tau^{\Delta} = \mathcal{E}\{\Delta \underline{\mathbf{w}}_\tau \Delta \underline{\mathbf{w}}_\tau^H\}$ as the $PM \times PM$ diagonal multichannel process noise covariance matrix. Full diagonality of $\underline{\boldsymbol{\Psi}}_\tau^{\Delta}$ is assumed for derivational ease, i.e., $\mathcal{E}\{\Delta \underline{\mathbf{w}}_{i,\tau} \Delta \underline{\mathbf{w}}_{j,\tau}^H\} = \mathbf{0}_{M\times M}$, $\forall\, i \neq j$. The expression (17) together with (13), or equivalently with (14), describe our multichannel state-space model.

3. ML-BENCH ALGORITHM

For ensuring *online* attributes of the resulting algorithm, we formulate our ML-EM framework considering a likelihood function of the form $p(\underline{\mathbf{y}}_\tau |\, \underline{\mathbf{y}}_{1:\tau-1}, \underline{\mathbf{s}}_\tau)$ which is conditioned on the parameter $\underline{\mathbf{s}}_\tau$ and all previous observations $\underline{\mathbf{y}}_{1:\tau-1}$. For the sake of brevity, hereinafter, we will refer to this likelihood function as $p(\underline{\mathbf{y}}_\tau |\, \underline{\mathbf{s}}_\tau)$. By inserting the unknown channel impulse responses $\underline{\mathbf{w}}_\tau$ into the likelihood function via marginalization and by using Jensen's inequality, we formulate the lower-bound [12] on the log-likelihood

$$\ln\, p(\underline{\mathbf{y}}_\tau |\, \underline{\mathbf{s}}_\tau) = \ln \int p(\underline{\mathbf{y}}_\tau, \underline{\mathbf{w}}_\tau |\, \underline{\mathbf{s}}_\tau)\, \mathrm{d}\underline{\mathbf{w}}_\tau \tag{18}$$

$$\geq \left\langle \ln \left(\frac{p(\underline{\mathbf{y}}_\tau, \underline{\mathbf{w}}_\tau |\, \underline{\mathbf{s}}_\tau)}{q(\underline{\mathbf{w}}_\tau)} \right) \right\rangle_{q(\underline{\mathbf{w}}_\tau)} \tag{19}$$

$$= \mathcal{F}(q(\underline{\mathbf{w}}_\tau), \underline{\mathbf{s}}_\tau) \ , \tag{20}$$

where $q(\underline{\mathbf{w}}_\tau)$ is an arbitrary distribution over the unknown channels and $\langle \cdot \rangle_{q(\underline{\mathbf{w}}_\tau)}$ denotes an expectation with respect to $q(\underline{\mathbf{w}}_\tau)$. Optimization of (20), to first fit and then maximize the log-likelihood, leads to the joint estimation of $\underline{\mathbf{w}}_\tau$ and $\underline{\mathbf{s}}_\tau$, respectively.

3.1. E-Step: Recursive Channel Identification

The joint distribution $p(\underline{\mathbf{y}}_\tau, \underline{\mathbf{w}}_\tau |\, \underline{\mathbf{s}}_\tau) = p(\underline{\mathbf{w}}_\tau |\, \underline{\mathbf{y}}_\tau, \underline{\mathbf{s}}_\tau)\, p(\underline{\mathbf{y}}_\tau |\, \underline{\mathbf{s}}_\tau)$ in (19) can be factorized into the posterior distribution $p(\underline{\mathbf{w}}_\tau |\, \underline{\mathbf{y}}_\tau, \underline{\mathbf{s}}_\tau)$ and the likelihood distribution $p(\underline{\mathbf{y}}_\tau |\, \underline{\mathbf{s}}_\tau)$. Thereafter, the functional differentiation [12] of $\mathcal{F}(q(\underline{\mathbf{w}}_\tau), \underline{\mathbf{s}}_\tau)$ with respect to $q(\underline{\mathbf{w}}_\tau)$, i.e., $\partial \mathcal{F}(q(\underline{\mathbf{w}}_\tau), \underline{\mathbf{s}}_\tau) \,/\, \partial q(\underline{\mathbf{w}}_\tau)$, reveals $q(\underline{\mathbf{w}}_\tau) = p(\underline{\mathbf{w}}_\tau |\, \underline{\mathbf{y}}_\tau, \underline{\mathbf{s}}_\tau)$ as the distribution that optimally tightens the lower-bound [11]. The mean $\widehat{\underline{\mathbf{w}}}_{i,\tau}$ and the covariance $\underline{\mathbf{P}}_{i,\tau}$ of the posterior $p(\underline{\mathbf{w}}_\tau |\, \underline{\mathbf{y}}_\tau, \underline{\mathbf{s}}_\tau)$ can be learned via the efficiently implementable state-space frequency-domain adaptive filter (SSFDAF) [11]. By considering the multichannel state-space model in (14) and (17), we can conveniently ex-

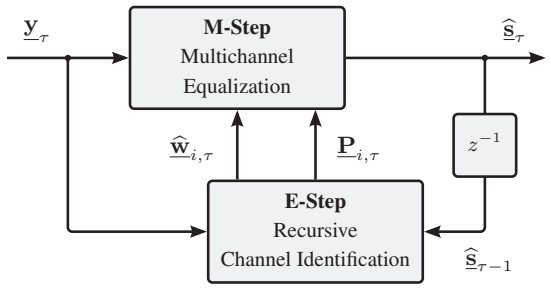

Fig. 1. Proposed ML-BENCH dereverberation system.

press the diagonalized SSFDAF recursions for the ith channel as

$$\widehat{\underline{\mathbf{w}}}^{+}_{i,\tau-1} = A \cdot \widehat{\underline{\mathbf{w}}}_{i,\tau-1} \tag{21}$$

$$\underline{\mathbf{P}}^{+}_{i,\tau-1} = A^2 \cdot \underline{\mathbf{P}}_{i,\tau-1} + \underline{\mathbf{\Psi}}^{\Delta}_{i,\tau} \tag{22}$$

$$\underline{\boldsymbol{\mu}}_{i,\tau} = \underline{\mathbf{P}}^{+}_{i,\tau-1}\left[\widehat{\underline{\mathbf{S}}}_{\tau-1}\underline{\mathbf{P}}^{+}_{i,\tau-1}\widehat{\underline{\mathbf{S}}}^{H}_{\tau-1} + \frac{M}{R}\underline{\mathbf{\Psi}}^{\mathbf{n}}_{i,\tau}\right]^{-1} \tag{23}$$

$$\underline{\mathbf{e}}_{i,\tau} = \underline{\mathbf{y}}_{i,\tau} - \mathbf{T}\,\widehat{\underline{\mathbf{S}}}_{\tau-1}\widehat{\underline{\mathbf{w}}}^{+}_{i,\tau-1} \tag{24}$$

$$\widehat{\underline{\mathbf{w}}}_{i,\tau} = \widehat{\underline{\mathbf{w}}}^{+}_{i,\tau-1} + \underline{\boldsymbol{\mu}}_{i,\tau}\widehat{\underline{\mathbf{S}}}^{H}_{\tau-1}\underline{\mathbf{e}}_{i,\tau} \tag{25}$$

$$\underline{\mathbf{P}}_{i,\tau} = \left[\mathbf{I}_M - \frac{R}{M}\underline{\boldsymbol{\mu}}_{i,\tau}\widehat{\underline{\mathbf{S}}}^{H}_{\tau-1}\widehat{\underline{\mathbf{S}}}_{\tau-1}\right]\underline{\mathbf{P}}^{+}_{i,\tau-1}\,, \tag{26}$$

where $\underline{\boldsymbol{\mu}}_{i,\tau}$, $\underline{\mathbf{e}}_{i,\tau}$, and $\underline{\mathbf{P}}_{i,\tau}$ are the Kalman stepsize, the error signal and the state-error covariance for the ith channel. The superscript $+$ denotes prediction terms, whereas $\widehat{\underline{\mathbf{S}}}_{\tau-1} = \mathrm{diag}\{\widehat{\underline{\mathbf{s}}}_{\tau-1}\}$ represents the estimated speech signal being injected into the E-step from the previous M-step at $\tau-1$ as shown in Fig. 1. The covariances $\underline{\mathbf{\Psi}}^{\Delta}_{i,\tau}$ and $\underline{\mathbf{\Psi}}^{\mathbf{n}}_{i,\tau}$ are estimated using the EM-based rules described in [11].

3.2. M-Step: Multichannel Equalization

For the resolution of the M-step, we factorize the joint distribution $p(\underline{\mathbf{y}}_\tau, \underline{\mathbf{w}}_\tau|\, \underline{\mathbf{s}}_\tau) = p(\underline{\mathbf{y}}_\tau|\, \underline{\mathbf{w}}_\tau, \underline{\mathbf{s}}_\tau)\, p(\underline{\mathbf{w}}_\tau|\, \underline{\mathbf{s}}_\tau)$ in (19) in terms of the distributions $p(\underline{\mathbf{y}}_\tau|\, \underline{\mathbf{w}}_\tau, \underline{\mathbf{s}}_\tau)$ and $p(\underline{\mathbf{w}}_\tau|\, \underline{\mathbf{s}}_\tau)$. The application of the conjugate derivative $\partial\,/\,\partial\underline{\mathbf{s}}^{*}_{\tau}$ [13] to $\mathcal{F}(q(\underline{\mathbf{w}}_\tau), \underline{\mathbf{s}}_\tau)$ in order to obtain an optimal estimate $\widehat{\underline{\mathbf{s}}}_\tau$ results in

$$\frac{\partial}{\partial\underline{\mathbf{s}}^{*}_{\tau}}\mathcal{F}(q(\underline{\mathbf{w}}_\tau), \underline{\mathbf{s}}_\tau) = \left\langle \frac{\partial}{\partial\underline{\mathbf{s}}^{*}_{\tau}} \ln\, p(\underline{\mathbf{y}}_\tau|\, \underline{\mathbf{w}}_\tau, \underline{\mathbf{s}}_\tau)\right\rangle_{q(\underline{\mathbf{w}}_\tau)}. \tag{27}$$

The transition distribution $p(\underline{\mathbf{w}}_\tau|\, \underline{\mathbf{s}}_\tau) = p(\underline{\mathbf{w}}_\tau|\, \boldsymbol{\mathcal{H}}_\tau)$ gets eliminated from (27) as it is not a function of $\underline{\mathbf{s}}_\tau$, rather it is a multivariate Gaussian which is conditioned on the *belief-state* $\boldsymbol{\mathcal{H}}_\tau = \left[(\widehat{\underline{\mathbf{w}}}_{1,\tau-1}, \underline{\mathbf{P}}_{1,\tau-1})\ \ldots\ (\widehat{\underline{\mathbf{w}}}_{P,\tau-1}, \underline{\mathbf{P}}_{P,\tau-1})\right]$ [11]. Considering the observation model in (13), which is equivalent to (14), the multivariate Gaussian log-transmission distribution is expressed as

$$\ln p(\underline{\mathbf{y}}_\tau|\, \underline{\mathbf{w}}_\tau, \underline{\mathbf{s}}_\tau) = - \left[\underline{\mathbf{y}}_\tau - \underline{\boldsymbol{\mathcal{W}}}_\tau\underline{\mathbf{s}}_\tau\right]^{H}\underline{\mathbf{\Psi}}^{\mathbf{n}^{-1}}_{\tau}\left[\underline{\mathbf{y}}_\tau - \underline{\boldsymbol{\mathcal{W}}}_\tau\underline{\mathbf{s}}_\tau\right] + \mathrm{const.} \tag{28}$$

Next, we substitute (28) into (27) and equate the expression to zero,

$$\left\langle \frac{\partial}{\partial\underline{\mathbf{s}}^{*}_{\tau}}\left[\underline{\mathbf{y}}_\tau - \underline{\boldsymbol{\mathcal{W}}}_\tau\underline{\mathbf{s}}_\tau\right]^{H}\underline{\mathbf{\Psi}}^{\mathbf{n}^{-1}}_{\tau}\left[\underline{\mathbf{y}}_\tau - \underline{\boldsymbol{\mathcal{W}}}_\tau\underline{\mathbf{s}}_\tau\right]\right\rangle_{q(\underline{\mathbf{w}}_\tau)} = \mathbf{0}\,. \tag{29}$$

The expansion of (29) followed by differentiation then yields

$$\left\langle \underline{\boldsymbol{\mathcal{W}}}^{H}_{\tau}\underline{\boldsymbol{\mathcal{W}}}_\tau\right\rangle_{q(\underline{\mathbf{w}}_\tau)}\widehat{\underline{\mathbf{s}}}_\tau = \left\langle \underline{\boldsymbol{\mathcal{W}}}^{H}_{\tau}\right\rangle_{q(\underline{\mathbf{w}}_\tau)}\underline{\mathbf{y}}_\tau\,. \tag{30}$$

The expectations in (30) are resolved after invoking the approximations $\mathbf{T} \approx (R/M)\,\mathbf{I}_M$ and $\mathbf{T}^H\mathbf{T} \approx (R/M)\,\mathbf{I}_M$ [14], which lead to $\underline{\boldsymbol{\mathcal{W}}}^{H}_{i,\tau} \approx (R/M)\,\underline{\mathbf{W}}^{H}_{i,\tau}$ and $\underline{\boldsymbol{\mathcal{W}}}^{H}_{i,\tau}\underline{\boldsymbol{\mathcal{W}}}_{i,\tau} \approx (R/M)\,\underline{\mathbf{W}}^{H}_{i,\tau}\underline{\mathbf{W}}_{i,\tau}$, to get [11]

$$\left\langle \underline{\boldsymbol{\mathcal{W}}}^{H}_{i,\tau}\right\rangle_{q(\underline{\mathbf{w}}_\tau)} \approx \frac{R}{M}\,\widehat{\underline{\mathbf{W}}}^{H}_{i,\tau} \tag{31}$$

and

$$\left\langle \underline{\boldsymbol{\mathcal{W}}}^{H}_{i,\tau}\underline{\boldsymbol{\mathcal{W}}}_{i,\tau}\right\rangle_{q(\underline{\mathbf{w}}_\tau)} \approx \frac{R}{M}\left(\widehat{\underline{\mathbf{W}}}^{H}_{i,\tau}\widehat{\underline{\mathbf{W}}}_{i,\tau} + \underline{\mathbf{P}}_{i,\tau}\right), \tag{32}$$

where $\widehat{\underline{\mathbf{W}}}_{i,\tau} = \mathrm{diag}\{\widehat{\underline{\mathbf{w}}}_{i,\tau}\}$. We substitute the results obtained in (31) and (32) into (30) to finally obtain the estimated speech signal

$$\widehat{\underline{\mathbf{s}}}_\tau = \underline{\mathbf{G}}^{H}_{\tau}\underline{\mathbf{y}}_\tau\,, \tag{33}$$

where the equalizer

$$\underline{\mathbf{G}}^{H}_{\tau} = \left[\sum_{i=1}^{P}\left(\widehat{\underline{\mathbf{W}}}^{H}_{i,\tau}\widehat{\underline{\mathbf{W}}}_{i,\tau} + \underline{\mathbf{P}}_{i,\tau}\right)\right]^{-1}\widehat{\underline{\mathbf{W}}}^{H}_{\tau} \tag{34}$$

comprises P equalization filters and $\widehat{\underline{\mathbf{W}}}_\tau$ is defined analogous to (12). The common term $\sum_{i=1}^{P}\left(\widehat{\underline{\mathbf{W}}}^{H}_{i,\tau}\widehat{\underline{\mathbf{W}}}_{i,\tau} + \underline{\mathbf{P}}_{i,\tau}\right)$ in (34) is fully diagonal and hence an $M \times M$ matrix inverse is avoided. In practice, this term is normalized to unity to prevent the mutually coupled system in Fig. 1 from approaching a trivial solution. Additional constraining eventually needs to be applied to (33) to avoid the effect of cyclic convolution in the DFT-domain.

4. RESULTS

For our simulations, we generated impulse responses $\mathbf{h}_{i,k}$ corresponding to a single source and a linear array with $P = 10$ microphones inside a room with dimensions 7 m $\times$ 5 m $\times$ 4 m (length $\times$ width $\times$ height) using a modified version [15] of the image method [16]. The source was located at (5 m, 1.5 m, 1.5 m), whereas the first microphone of the array was positioned at (2 m, 4 m, 1.5 m). The other microphone positions can be obtained by successively adding 0.1 m to the x-coordinate of the first microphone. The reverberation time T_{60} of the room was varied from 0.2 s to 1.0 s in steps of 0.2 s. For the generation of the microphone signals $y_{i,k}$, the channel length L in (2) was chosen as $T_{60} \cdot f_s$ at a sampling rate of $f_s = 16$ kHz.

As the source signal s_k, we used ten different speech signals corresponding to five female and five male speakers. Reverberant signals $y_{i,k}$ were obtained by convolving the generated channels $\mathbf{h}_{i,k}$ with each of the source signals s_k and adding zero-mean white Gaussian noise $n_{i,k}$ at a signal-to-noise ratio (SNR) of 30 dB. We consider the last microphone of the array as the reverberant reference signal, since it is the closest to the source. For our algorithm, we use a state-transition coefficient $A = 0.9997$, a frame-size $M = 1024$, and a frame-shift $R = 512$, thus resulting in the modeled impulse response length of $M - R = 512$ coefficients.

As an anchor for our evaluation, we consider the output $\overline{s}_k = (1/P)\sum_{i=1}^{P} y_{i,k-d_i}$ of an ideal delay-and-sum beamformer (DSB), where the time delays d_i were exactly determined from the known source and microphone positions. Note that this *a priori* information is not available to our fully blind ML-BENCH algorithm.

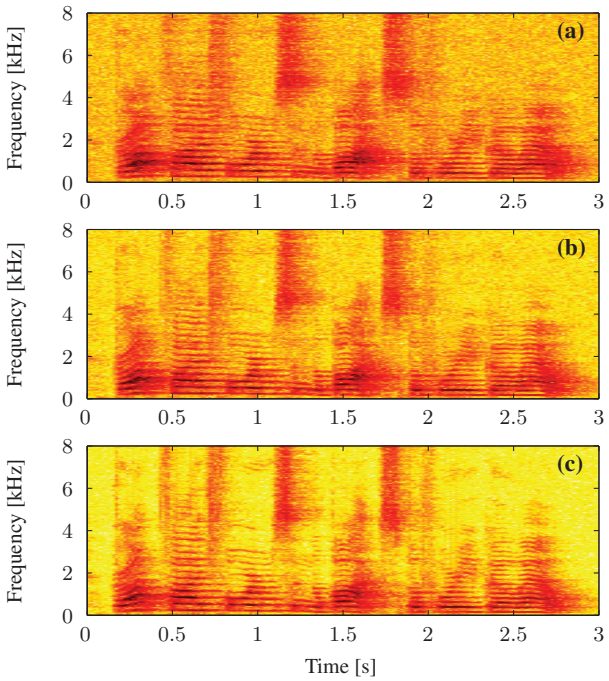

Fig. 2. Spectrograms of (a) the reverberant reference, (b) the DSB output, and (c) the ML-BENCH output. $T_{60} = 0.6$ s, SNR $= 30$ dB.

Fig. 2 shows the spectrograms of the reverberant reference signal, the output $\overline{s}_k$ of the DSB, and the estimate $\widehat{s}_k$ of the proposed algorithm for a female speaker and a reverberation time of $T_{60} = 0.6$ s. It can be seen that the DSB mainly reduces the observation noise, whereas the proposed ML-BENCH algorithm also sharpens the temporal edges and restores the fine-structure such that the speech harmonics become visible again.

For an objective evaluation, we consider spectral distances (SD) and cepstral distances (CD) to the clean speech signal as defined in [3] and [17], respectively. Table 1 and Fig. 3 depict both of the measures averaged over all ten speakers for different reverberation times T_{60}. In all cases, the proposed blind algorithm shows significant improvements as compared to the ideal DSB. On average, the DSB achieves SD and CD gains of 2.1 dB and 0.7 dB, respectively, as compared to the reverberant signal. The ML-BENCH algorithm surpasses the DSB with corresponding gains of 3.1 dB and 1.6 dB.

5. CONCLUSIONS

In this contribution, we proposed an online multichannel blind dereverberation algorithm in the DFT-domain that was derived within the maximum-likelihood expectation-maximization framework. The resulting ML-BENCH algorithm comprises coupled channel identification and equalization subsystems, both of which are efficiently implementable. We finally evaluated the derived algorithm by considering instrumental performance measures for different reverberation times and by analyzing speech spectrograms. The obtained results substantiate the efficacy of our approach.

$T_{60} =$	0.2 s	0.4 s	0.6 s	0.8 s	1.0 s
Reverberant	8.7	10.9	12.1	12.9	13.4
Ideal DSB	6.2	8.6	10.0	10.9	11.6
ML-BENCH	5.6	7.5	9.1	10.0	10.4

Table 1. Spectral distances [dB] for five reverberation times T_{60}.

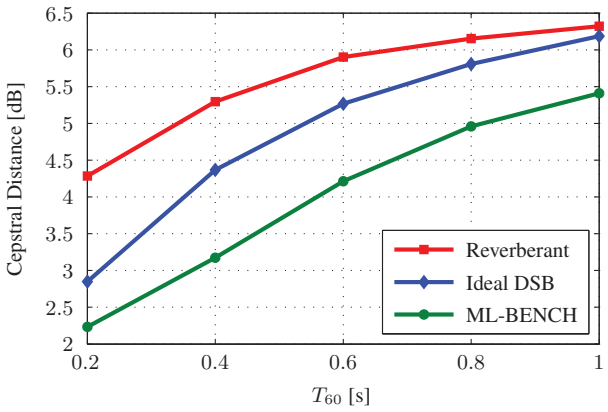

Fig. 3. Cepstral distances for five reverberation times T_{60}.

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, Sept. 2009.

[2] 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, May 2009.

[3] 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, Feb. 2007.

[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] H. Buchner and W. Kellermann, "TRINICON for dereverberation of speech and audio signals," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds., chapter 10, pp. 311–385. Springer, London, 2010.

[6] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[7] M. A. Haque, T. Islam, and Md. K. Hasan, "Robust speech dereverberation based on blind adaptive estimation of acoustic channels," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 775–787, May 2011.

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

[9] D. Schmid and G. Enzner, "A parametric least-squares approximation for multichannel equalization of room acoustics," in *Proc. Int. Workshop on Acoust. Echo and Noise Control*, Sept. 2010.

[10] 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, June 2006.

[11] 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, and Signal Processing*, Dallas, USA, Mar. 2010, pp. 3822–3825.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, NY, 1st edition, 2006.

[13] S. S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Upper Saddle River, NJ, 4th edition, 2002.

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

[15] E. A. Lehmann and A. M. 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, July 2008.

[16] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[17] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.