Online speech dereverberation using Kalman filter and EM algorithm

Boaz Schwartz, Sharon Gannot, *Senior Member, IEEE,* and Emanuël A.P. Habets, *Senior Member, IEEE*

Abstract

Speech signals recorded in a room are commonly degraded by reverberation. In most cases, both the speech signal and the acoustic system of the room are unknown and time-varying. In this paper, a scenario with a single desired sound source and slowly time-varying and spatially-white noise is considered, and a multi-microphone algorithm that simultaneously estimates the clean speech signal and the time-varying acoustic system is proposed. The recursive expectation-maximization scheme is employed to obtain both the clean speech signal and the acoustic system in an online manner. In the expectation step, the Kalman filter is applied to extract a new sample of the clean signal, and in the maximization step, the system estimate is updated according to the output of the Kalman filter. Experimental results show that the proposed method is able to significantly reduce reverberation and increase the speech quality. Moreover, the tracking ability of the algorithm was validated in practical scenarios using human speakers moving in a natural manner.

I. Introduction

An acoustic sound that propagates in an enclosure is repeatedly reflected from the walls and other objects, this phenomenon, usually referred to as reverberation, degrades the speech quality and, in severe cases, the intelligibility. In recent years, due to advances in understanding the phenomenon and the availability of stronger computational resources, the interest in dereverberation increased and numerous methods were proposed.

Statistical room acoustics is widely used for dereverberation. The exponential decay of the late reverberant power was mathematically formulated in [1], and then utilized to derive an estimator for the late reverberant spectral variance in [2]. The estimated spectral variance was then used to suppress late reverberation using a spectral subtraction algorithm. The method was extended to the multi-microphone case in [3], and an improved estimator for the late reverberant spectral variance taking into account the reverberation time and the direct-to-reverberation ratio was proposed in [4].

Boaz Schwartz and Sharon Gannot are with the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel (e-mail: boazsh0@gmail.com; Sharon.Gannot@biu.ac.il).

E. A. P. Habets is with the International Audio Laboratories Erlangen, (a joint institution of the University of Erlangen-Nuremberg and Fraunhofer IIS), 91058 Erlangen, Germany (e-mail: emanuel.habets@audiolabs-erlangen.de).

This research was supported by a Grant from the GIF, the German-Israeli Foundation for Scientific Research and Development.

Dereverberation utilizing multiple microphones and system identification typically consists of two stages. Firstly, the acoustic system is blindly estimated from the observed signals (since the clean signal is unobservable) [5], [6]. Secondly, an equalizer is calculated and applied to the observed signals to obtain an estimate of the clean signal [7]–[9].

The problem of blind system identification (BSI) can be cast as a deterministic parameter estimation problem and hence solved using the maximum likelihood (ML) framework. Deriving the ML estimator can be a cumbersome task, and does not always result in a closed-form solution. The expectation-maximization (EM) procedure is frequently utilized to conveniently estimate the parameters that maximize the likelihood function. The EM procedure yields, as a byproduct, a signal estimate, i.e. it jointly estimates the required parameters and the desired signal.

In [10], an EM algorithm for dereverberation and noise reduction is presented. The room impulse response (RIR) is modelled as an auto-regressive (AR) process in each frequency band. In the E-step the Wiener filter, calculated by using the current values of the parameters, is applied to estimate the clean speech signal. In the M-step, the current estimated signal is used to update the parameters. The method was extended in [11] to simultaneously dereverberate and separate multiple speakers. Another EM algorithm for dereverberation and source separation is presented in [12], where the AR model is used only for the late reverberant part, while a finite impulse response (FIR) model is used for the early reflections. The E-step consists of linear filtering followed by a multichannel Wiener filter, and in the M-step, the acoustic system parameters are updated.

The Wiener filter can be replaced by the Kalman smoother or approximated by the Kalman filter [13]. The Kalman filter is also commonly utilized in speech processing problems [14], most commonly within the EM framework. The first application of Kalman filter to speech enhancement was proposed by Paliwal and Basu [15], where prior knowledge of the clean speech parameters was used. Gibson et al. presented a method that concurrently estimates the speech signal and the required parameters in [16]. Weinstein et al. [17] presented an algorithm for multi-microphone speech enhancement. They represented the signal model using linear dynamic state equations in the time-domain, and applied the EM algorithm to estimate system parameters. As the Kalman smoother is used in the E-step to estimate the speech and noise signals from the mixed measurements, and the RIR is updated in the M-step, we refer to this method as the Kalman-EM (KEM) method. The noise reduction capabilities of the KEM method were demonstrated in a simple two-microphone setup. In [18], the KEM scheme was applied to a single-microphone speech enhancement in the time-domain, where high-order statistics are considered to obtain a robust initialization for the parameter estimation stage.

The Kalman filter and its extensions were also utilized in the field of speech dereverberation. The authors have recently presented a KEM-based algorithm for dereverberation in [19]. In the E-step, the Kalman smoother is applied to extract the clean signal from the data utilizing the estimated parameters. In the M-step, the parameters are updated according to the output of the Kalman smoother. We refer to this algorithm as Kalman-EM for dereverberation (KEMD). Each EM iteration uses the entire measurement set, hence the KEMD is an iterative offline algorithm. Significant dereverberation capabilities of the proposed algorithm are demonstrated, while exhibiting only low speech distortion.

Under the Bayesian framework, the RIR filters are treated as stochastic processes. In [20], the E-step and the M-step objectives switch roles, namely the (stochastic) channel is identified in the E-step, and the (deterministic) clean speech in the M-step. It was proposed in [21] to use the unscented Kalman filter [22] to jointly estimate the RIR and the clean speech,

where both are treated as stochastic processes. The RIRs were modelled using FIRs, and simulation results demonstrates the convergence of the proposed method in a synthesized simple scenario with short RIRs. In [23], the Kalman filter is used to estimate the clean speech, and a particle filter is utilized to estimate the RIR of the reverberant room.

In many practical applications, the positions of the speaker and the microphones are dynamic, and the acoustic system is consequently time-varying. In these scenarios, the aforementioned solutions based on the Wiener or Kalman smoother cannot be straightforwardly applied. In order to enhance the reverberated signal in such conditions, algorithms must be able to update their parameters in an online fashion. To handle dynamic scenarios under the probabilistic framework described above, a recursive version of the EM procedure should be used. A recursive version for the EM algorithm was first formulated by Titterington [24], based on a Newton search for the maximum of the likelihood function. The convergence properties of Titterington's algorithm are discussed in [25] and a new recursive algorithm is proposed. It is shown that both algorithms converge with *probability one* to a stationary point of the likelihood function. An *almost surely* convergence of the Titterington's algorithm was proved by Wang and Zhao in [26], based on the results of Delyon [27]. Recursive algorithms based on the KEM scheme were proposed in [17], [18]. The convergence properties of the recursive KEM approach were demonstrated in [17] for a two-microphone speech enhancement task. An extensive experimental study for the single-microphone KEM and its recursive version was given in [18]. Recursive EM variants were also proposed in [28], with an application to multiple target tracking. The first variant is a Newton-based search, that is closely related to Titterington's algorithm, while the second is a KEM-based algorithm adapted to the specific model. Another algorithm is proposed in [28], where the parameter vector trajectory is modelled as an hidden Markov model (HMM) process, and a corresponding EM-HMM algorithm for parameter estimation is derived. A different online EM algorithm was proposed by Cappé and Moulines in [29]. A convergence proof under certain regularity conditions is also provided. In [30], the convergence speed of the batch EM algorithm and of three online variants is compared for various estimation tasks. The results show a better convergence speed of the online algorithms, and even an improved estimation accuracy in several cases. Titterington's and Cappé and Moulines' schemes were used for the multiple speaker tracking problem in [31], in which also a constrained version for Titterington's algorithm was proposed. To the best of our knowledge, no recursive EM (REM) algorithm for dereverberation has been reported in the literature.

To enable online dereverberation of a single speaker, we propose in this contribution a KEM-based algorithm in the short-time Fourier transform (STFT) domain. We show that this specific version of the KEM scheme can be defined as an REM algorithm and therefore possess the convergence properties proven in [29]. The acoustic system is modelled as an FIR in the STFT domain, and state-space equations are presented. In the E-step, a new sample of the speech signal is estimated by the Kalman filter, and in the M-step, the acoustical parameters are updated. The instantaneous power of the clean speech is predicted by a spectral enhancement (SE)-based method that utilizes the estimated parameters. This prediction is used in conjugation with the Kalman filter to estimate the clean speech signal. In this work, we assume a, possibly moving, desired sound source, and slowly time-varying and spatially-white noise.

This paper is organized as follows. A statistical model and an optimization criterion are given in Sec. II. Sec. III is dedicated to a brief summary of our previously proposed iterative-batch algorithm for dereverberation, KEMD. In Sec. IV, the proposed method is derived. Some practical considerations are given in Sec V. An extensive experimental study using speech recorded

in our lab (either reproduced by loudspeakers or uttered by human speakers) for both static and dynamic scenarios is presented in Sec. VI. Conclusions are drawn in Sec. VII.

II. Statistical Model and Optimization Criterion

A. Statistical Model

Let $x[n]$ be a clean speech signal in the time-domain. The noisy and reverberant speech signal received by the jth microphone is given by

$$z_j[n] = \sum_{l'=0}^{L'-1} h_{j,l'}[n]\ x[n-l'] + v_j[n], \tag{1}$$

where $h_{j,0}[n], h_{j,1}[n], \ldots, h_{j,L'-1}[n]$ are the coefficients of the, possibly time-varying, RIR relating the speaker and the jth microphone, and $v_j[n]$ is an additive noise at microphone j. In the STFT domain, $x(t,k)$ denotes the clean speech in time-frame t and frequency-bin k. Given the variance of the speech signal $\phi_x(t,k)$, the speech signal samples can be modelled as independent complex-Gaussian random variables [32]:

$$x(t,k) \sim \mathcal{N}_C\left\{0, \phi_x(t,k)\right\}. \tag{2}$$

In the STFT domain, the RIR can be approximately modelled by a convolutive transfer function (CTF) [33]. This approximation was successfully used for dereverberation in [4], and for relative transfer function (RTF) estimation in reverberant environments in [34]. Using this model, (1) can be expressed in the STFT domain as

$$z_j(t,k) = \sum_{l=0}^{L-1} h_{j,l}(t,k)\ x(t-l,k) + v_j(t,k). \tag{3}$$

Since the delay between the source and the microphones is unknown, we assume, without the loss of generality, that $h_j(k)$ are causal and of finite length. Typically, L is much shorter than L' in (1).

We further assume that $v_j(t,k)$ are complex-Gaussian random variables:

$$v_j(t,k) \sim \mathcal{N}_C\left\{0, \phi_{v_j}(t,k)\right\}. \tag{4}$$

In addition, we assume that the noise is uncorrelated in all channels, i.e., $E\left\{v_i(t,k)v_j^*(t,k)\right\} = 0$ for $j \neq i$.

B. State-Space Representation

Eq. (3) can be expressed in a vector form as

$$z_j(t,k) = \mathbf{h}_j^T(t,k)\mathbf{x}_t(k) + v_j(t,k), \tag{5}$$

where

$$\mathbf{h}_j(t,k) = [h_{j,L-1}(t,k), \ldots, h_{j,0}(t,k)]^T, \tag{6}$$

$$\mathbf{x}_t(k) = [x(t-L+1,k), \ldots, x(t,k)]^T, \tag{7}$$

and $(\cdot)^T$ is the transpose operator. The multi-microphone state-space equations are given by (when appropriate, the frequency index k is henceforth omitted for brevity):

$$\begin{aligned} \mathbf{x}_t &= \mathbf{\Phi}\mathbf{x}_{t-1} + \mathbf{w}_t, \\ \mathbf{z}_t &= \mathbf{H}_t\mathbf{x}_t + \mathbf{v}_t, \end{aligned} \tag{8}$$

where the innovation process is given by

$$\mathbf{w}_t \equiv [0, \ldots, x(t)]^T,$$

the measurement and noise vectors are

$$\begin{aligned} \mathbf{z}_t &\equiv [z_1(t), \ldots, z_J(t)]^T, \\ \mathbf{v}_t &\equiv [v_1(t), \ldots, v_J(t)]^T, \end{aligned}$$

and J is the number of microphones. The process and measurement matrices are, respectively, equal to

$$\mathbf{\Phi} \equiv \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix},$$

$$\mathbf{H}_t \equiv [\mathbf{h}_1(t), \ldots, \mathbf{h}_J(t)]^T.$$

Note that unlike the time-domain state-space representation in [17], [18], here the process is not modelled as an AR signal, as evident from the absence of regression parameters in $\mathbf{\Phi}$. In the model presented previously, we assumed no statistical dependency of adjacent time-frames of $x(t)$, given the variance of speech signal $\phi_x(t)$. For this assumption to hold, it is required that the overlap between STFT frames is sufficiently small, as discussed in Sec. VI.

Finally, the second-order statistics matrices of the innovation noise and the measurement noise processes are defined as:

$$\mathbf{F}_t \equiv E\left\{\mathbf{w}_t\mathbf{w}_t^H\right\} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_x(t) \end{bmatrix}$$

$$\mathbf{R}_t \equiv E\left\{\mathbf{v}_t\mathbf{v}_t^H\right\} = \begin{bmatrix} \phi_{v_1}(t) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_{v_J}(t) \end{bmatrix},$$

where $(\cdot)^H$ is the complex conjugate operator, and the matrix $\mathbf{R}_t$ is diagonal since the noise is assumed uncorrelated between the channels.

C. Optimization Criterion

Let $\mathcal{Z}$ be a set of measurements

$$\mathcal{Z} = \{z_j(t,k):\ j \in \mathcal{J},\ t \in \mathcal{T},\ k \in \mathcal{K}\},$$

where $\mathcal{J} = \{1 \dots J\}$ are the microphone indices, $\mathcal{T} = \{1 \dots T\}$ are time indices in the STFT domain, and $\mathcal{K} = \{1 \dots K\}$ are the frequency indices. Our goal is to estimate the clean speech signal

$$\mathcal{X} = \{x(t,k):\ t \in \mathcal{T}, k \in \mathcal{K}\}, \tag{9}$$

given the measurements set $\mathcal{Z}$. For the solution of this dereverberation task we adopt the ML approach and estimate the following acoustic parameters:

$$\boldsymbol{\Theta} \equiv \{\boldsymbol{\Theta}_x, \boldsymbol{\Theta}_h, \boldsymbol{\Theta}_v\} \tag{10a}$$

$$\boldsymbol{\Theta}_x \equiv \{\phi_x(t,k) : t \in \mathcal{T}, k \in \mathcal{K}\} \tag{10b}$$

$$\boldsymbol{\Theta}_h \equiv \{\mathbf{h}_j(t,k) : j \in \mathcal{J}, t \in \mathcal{T}, k \in \mathcal{K}\} \tag{10c}$$

$$\boldsymbol{\Theta}_v \equiv \left\{\phi_{v_j}(t,k) : j \in \mathcal{J}, t \in \mathcal{T}, k \in \mathcal{K}\right\}. \tag{10d}$$

Since the spectral coefficients of the clean speech in (9) are unobservable, the ML estimator of (10a)-(10d) can be obtained from the given measurements by defining $\mathcal{X}$ as a latent data set, and by applying the EM algorithm.

The statistical model in Sec. II-A assumes that adjacent time frames of speech and noise are statistically independent, and that noise and speech signals are uncorrelated. Therefore, the log-likelihood of the complete data is:

$$\log f(\mathcal{X},\mathcal{Z};\mathbf{\Theta}) = C - \frac{1}{2}\sum_{\tau=1}^{T}\left[\log\phi_x(\tau) + \frac{|x(\tau)|^2}{\phi_x(\tau)}\right] - \frac{1}{2}\sum_{\tau=1}^{T}\sum_{j=1}^{J}\left[\log\phi_{v_j}(\tau) + \frac{1}{\phi_{v_j}(\tau)}\left|z_j(\tau) - \mathbf{h}_j^T(\tau)\mathbf{x}_\tau\right|^2\right], \quad (11)$$

where C is a constant value independent of the parameters. Note that the first summation term is the log-likelihood of clean speech signal, and the second summation term is related to the noise signal.

III. Iterative EM algorithm

In this section we briefly summarize the iterative KEMD algorithm proposed in [19], that is based on the EM algorithm proposed by Dempster, Laird, and Rubin [35]. In the derivation of the KEMD, both the acoustic systems and the noise were assumed to be time-invariant.

The EM consists of two steps, repeated iteratively until convergence. In the E-step of the p-th iteration, the auxiliary function

$$Q\left[\mathbf{\Theta}\middle|\widehat{\mathbf{\Theta}}^{(p-1)}\right] \equiv E\left\{\log f(\mathcal{Z},\mathcal{X};\mathbf{\Theta})\middle|\mathcal{Z};\widehat{\mathbf{\Theta}}^{(p-1)}\right\}, \quad (12)$$

is calculated using the entire data set $\mathcal{Z}$ and the latest parameter estimate $\widehat{\mathbf{\Theta}}^{(p)}$. In the M-step, the parameters are re-estimated by maximizing the auxiliary function, i.e.,

$$\widehat{\mathbf{\Theta}}^{(p)} = \arg\max_{\mathbf{\Theta}} Q\left[\mathbf{\Theta}\middle|\widehat{\mathbf{\Theta}}^{(p-1)}\right]. \quad (13)$$

By iteratively repeating the E- and M-steps, the convergence of $\widehat{\mathbf{\Theta}}^{(p)}$ to a local maximum of the likelihood function is guaranteed.

Applying the EM scheme to the likelihood function in (11) yields the following auxiliary function [19]:

$$Q\left(\mathbf{\Theta}\middle|\widehat{\mathbf{\Theta}}^{(p-1)}\right) = -\sum_{\tau=1}^{T}\left[\log\phi_x(\tau) + \frac{1}{\phi_x(\tau)}\,\widehat{|x(\tau)|^2}^{(p-1)}\right] - \sum_{j=1}^{J}\sum_{\tau=1}^{T}\left[\log\phi_{v_j} + \frac{1}{\phi_{v_j}}\left\{|z_j(\tau)|^2 - 2\Re\left(\mathbf{h}_j^T\widehat{\mathbf{x}}_\tau^{(p-1)}z_j^*(\tau)\right) + \mathbf{h}_j^T\,\widehat{\mathbf{x}_\tau\mathbf{x}_\tau^H}^{(p-1)}\,\mathbf{h}_j^*\right\}\right], \quad (14)$$

where $\Re(\cdot)$ is the real part, $(\cdot)^*$ is the (scalar) complex conjugate, and

$$\widehat{\mathbf{x}}_t^{(p-1)} \equiv E\left\{\mathbf{x}_t\middle|\mathcal{Z};\widehat{\mathbf{\Theta}}^{(p-1)}\right\}, \quad (15a)$$

$$\widehat{\mathbf{x}_t\mathbf{x}_t^H}^{(p-1)} \equiv E\left\{\mathbf{x}_t\mathbf{x}_t^H\middle|\mathcal{Z};\widehat{\mathbf{\Theta}}^{(p-1)}\right\}, \quad (15b)$$

$$\widehat{|x(t)|^2}^{(p-1)} \equiv E\left\{|x(t)|^2\middle|\mathcal{Z};\widehat{\mathbf{\Theta}}^{(p-1)}\right\}. \quad (15c)$$

As in [17] [18], the Kalman smoother was used in [19] to obtain the first- and second-order statistics depicted in (15a)-(15c).

In the M-step, the updated parameters were calculated according to

$$\widehat{\phi}_x^{(p)}(t) = \widehat{|x(t)|^2}^{(p-1)} \tag{16a}$$

$$\left(\widehat{\mathbf{h}}_j^{(p)}\right)^* = \left[\widehat{\mathbf{R}}_{xx}^{(p-1)}\right]^{-1} \widehat{\mathbf{r}}_{xz_j}^{(p-1)} \tag{16b}$$

$$\widehat{\phi}_{v_j}^{(p)} = \widehat{r}_{z_j z_j} - 2\,\Re\left[\left(\widehat{\mathbf{h}}_j^{(p)}\right)^T \widehat{\mathbf{r}}_{xz_j}^{(p-1)}\right] + \left(\widehat{\mathbf{h}}_j^{(p)}\right)^T \widehat{\mathbf{R}}_{xx}^{(p-1)} \left(\widehat{\mathbf{h}}_j^{(p)}\right)^*, \tag{16c}$$

where we have defined:

$$\widehat{\mathbf{R}}_{xx}^{(p-1)} \equiv \sum_{\tau=1}^{T} \widehat{\mathbf{x}_\tau \mathbf{x}_\tau^H}^{(p-1)}, \quad \widehat{\mathbf{r}}_{xz_j}^{(p-1)} \equiv \sum_{\tau=1}^{T} \widehat{\mathbf{x}}_\tau^{(p-1)} z_j^*(\tau),$$

$$\widehat{r}_{z_j z_j} \equiv \sum_{\tau=1}^{T} |z_j(\tau)|^2 .$$

It was shown in [19] that the KEMD algorithm is able to significantly dereverberate the input signal without distorting the speech signal. However, the KEMD algorithm is an iterative algorithm and is not suitable for online applications. Moreover, in the iterative scheme it is assumed that the RIRs are time-invariant, rendering this method inappropriate for scenarios where the speaker and/or the microphones are moving. The newly proposed recursive algorithm described in Sec. IV is extending the KEMD algorithm to online applications and dynamic scenarios.

IV. Recursive EM Algorithm

We now derive a recursive version for the KEMD algorithm, where the Kalman smoother is substituted by the Kalman filter in the E-step, and the acoustic system is updated utilizing the recursive M-step proposed by Cappé and Moulines [29]. The algorithm is nicknamed recursive Kalman-EM for dereverberation (RKEMD), and it is summarized in Fig. 1. As opposed to the KEMD, we now use the more general time-varying signal model in (3) and (4). The REM scheme for the problem at hand is described in Sec. IV-A, the E- and M-steps are detailed in Sec. IV-B and Sec. IV-C, respectively, and the variance estimator of the clean speech signal is given in Sec. IV-D.

A. Recursive EM Scheme

Applying the REM scheme presented in [29] to the problem at hand, the auxiliary function (14) is replaced by a recursive one:

$$Q\left(\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t)\right) = - \frac{1-\beta}{2} \sum_{\tau=1}^{t} \beta^{t-\tau} \left[\log \phi_x(\tau) + \frac{1}{\phi_x(\tau)} \widehat{|x(\tau)|^2}\right] - \frac{1-\beta}{2} \sum_{j=1}^{J} \sum_{\tau=1}^{t} \beta^{t-\tau} \left[\log \phi_{v_j} + \frac{1}{\phi_{v_j}} \left\{|z_j(\tau)|^2 - 2\Re\left(\mathbf{h}_j^T \widehat{\mathbf{x}}_{\tau|\tau} z_j^*(\tau)\right) + \mathbf{h}_j^T \widehat{\mathbf{x}_\tau \mathbf{x}_\tau^H} \mathbf{h}_j^*\right\}\right], \tag{17}$$

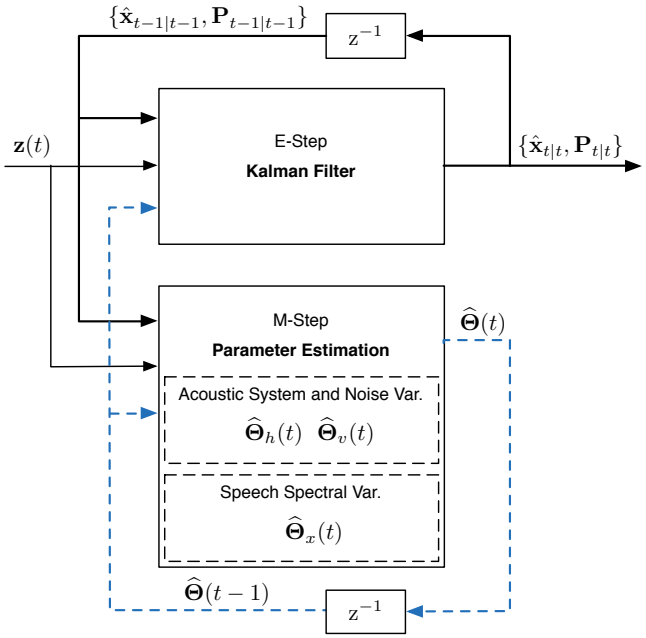


Fig. 1. Block diagram of the proposed algorithm.

where

$$\widehat{\mathbf{x}}_{t|t} \equiv E\left\{\mathbf{x}_t \middle| \mathcal{Z}_t; \widehat{\mathbf{\Theta}}(t)\right\}, \tag{18a}$$

$$\widehat{\mathbf{x}_t \mathbf{x}_t^H} \equiv E\left\{\mathbf{x}_t \mathbf{x}_t^H \middle| \mathcal{Z}_t; \widehat{\mathbf{\Theta}}(t)\right\}, \tag{18b}$$

$$\widehat{|x(t)|^2} \equiv E\left\{|x(t)|^2 \middle| \mathcal{Z}_t; \widehat{\mathbf{\Theta}}(t)\right\}, \tag{18c}$$

are the first- and second-order statistics of the clean speech signal and $\mathcal{Z}_t$ is the set of available measurements

$$\mathcal{Z}_t = \{z_j(\tau, k) : \ j \in \mathcal{J}, \ \tau \in [1, t], \ k \in \mathcal{K}\}.$$

A detailed derivation of (17) and (18) is available in Appendix A. In the M-step, the updated parameters $\widehat{\mathbf{\Theta}}(t+1)$ are obtained, similarly to (13), by the maximization:

$$\widehat{\mathbf{\Theta}}(t+1) = \arg\max_{\mathbf{\Theta}} \left\{Q\left[\mathbf{\Theta} \middle| \widehat{\mathbf{\Theta}}(t)\right]\right\}. \tag{19}$$

The convergence of the REM algorithm is proven in [29] under the assumption that the likelihood function is a member of the exponential distribution family. This condition is verified in Appendix B for the statistical model defined in Sec. II. Note that the convergence properties of the EM and REM algorithms are essentially different. The series of the EM estimators $\widehat{\mathbf{\Theta}}^{(p)}$ converges to a local maximum of the ML function defined by the observed data. Conversely, the series of REM estimators $\widehat{\mathbf{\Theta}}(t)$ converges to a stationary point of the Kullback-Leibler divergence between the actual probability distribution function (PDF) of the measurements and the parametric PDF that incorporates the estimated parameters $\widehat{\mathbf{\Theta}}(t)$.

B. E-Step: Kalman Filter

In the E-step, the auxiliary function (17) should be calculated, for which an estimate of the first- and second-order statistics of the clean speech signal (18) is required. The Kalman filter is providing both a recursive minimum mean square error (MMSE) estimator of the first-order statistics of the clean signal and the respective error covariance matrix, from which the second-order statistics of the clean signal can be easily calculated, as shown below. Due to its recursive nature and its optimality, the Kalman filter constitutes the E-step of the REM algorithm described in Sec. IV-A. The Kalman filtering equations are summarized in Algorithm 1.

The outcome of the Kalman filter is the state-vector estimator, $\widehat{\mathbf{x}}_{t|t}$, which is the required first-order statistics estimator (18a), and the respective estimation covariance matrices, namely $\mathbf{P}_{t|t}$. For the M-step described in Sec. IV-C, an *instantaneous* estimate of the second-order statistics is obtained by [18]:

$$\widehat{\mathbf{x}_t\mathbf{x}_t^H} = E\left\{\mathbf{x}_t\mathbf{x}_t^H \;\middle|\; \mathcal{Z}_t; \widehat{\boldsymbol{\Theta}}(t)\right\} = \widehat{\mathbf{x}}_{t|t}\widehat{\mathbf{x}}_{t|t}^H + \mathbf{P}_{t|t} \ . \tag{20}$$

Note that each of the elements of the state-vector $\widehat{\mathbf{x}}_{t|t}$ corresponds to a different frame of the estimated speech signal (see (7)). A fixed-lag Kalman smoother [18] can be obtained by selecting one of the delayed elements as the algorithm output. Selecting the first element, $\hat{x}(t-L+1|t)$, will most likely yield a more accurate solution than selecting the last one, $\hat{x}(t|t)$, but will result in a latency of few frames. In the experimental study in Sec. VI, we preferred the accuracy and sacrificed the latency of the algorithm.

C. M-Step: Acoustical System Estimation

In the M-step, defined in (19), the maximization of (17) with respect to (w.r.t.) $\mathbf{h}_j$ and $\sigma_{v_j}^2$ results in:

$$\widehat{\mathbf{h}}_j^*(t+1) = \left[\widehat{\mathbf{R}}_{xx}^{(t)}\right]^{-1} \widehat{\mathbf{r}}_{xz_j}^{(t)} \tag{21}$$

$$\widehat{\phi}_{v_j}(t+1) = \frac{1-\beta}{1-\beta^t}\left\{\widehat{r}_{z_jz_j}^{(t)} - 2\,\Re\left[\widehat{\mathbf{h}}_j^T(t)\,\widehat{\mathbf{r}}_{xz_j}^{(t)}\right] + \widehat{\mathbf{h}}_j^T(t)\,\widehat{\mathbf{R}}_{xx}^{(t)}\,\widehat{\mathbf{h}}_j^*(t)\right\}, \tag{22}$$

Algorithm 1: Kalman Filtering.

Predict:
$\widehat{\mathbf{x}}_{t|t-1} = \boldsymbol{\Phi}\ \widehat{\mathbf{x}}_{t-1|t-1}$
$\mathbf{P}_{t|t-1} = \boldsymbol{\Phi}\ \mathbf{P}_{t-1|t-1}\ \boldsymbol{\Phi}^T + \mathbf{F}_t$
Update:
$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}_t^H\left[\mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^H + \mathbf{R}_t\right]^{-1}$
$\mathbf{e}_t = \mathbf{z}_t - \mathbf{H}_t\widehat{\mathbf{x}}_{t|t-1}$
$\widehat{\mathbf{x}}_{t|t} = \widehat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t\ \mathbf{e}_t$
$\mathbf{P}_{t|t} = [\mathbf{I} - \mathbf{K}_t\mathbf{H}_t]\ \mathbf{P}_{t|t-1}$

where we define the *accumulation* of the instantaneous second-order statistics as:

$$\widehat{\mathbf{R}}_{xx}^{(t)} \equiv \sum_{\tau=1}^{t} \beta^{t-\tau} \widehat{\mathbf{x}_\tau \mathbf{x}_\tau^H} = \beta\, \widehat{\mathbf{R}}_{xx}^{(t-1)} + \widehat{\mathbf{x}_t \mathbf{x}_t^H}$$
$$\widehat{\mathbf{r}}_{xz_j}^{(t)} \equiv \sum_{\tau=1}^{t} \beta^{t-\tau} \widehat{\mathbf{x}}_{\tau|\tau} z_j^*(\tau) = \beta\, \widehat{\mathbf{r}}_{xz_j}^{(t-1)} + \widehat{\mathbf{x}}_{t|t} z_j^*(t)$$
$$\widehat{r}_{z_j z_j}^{(t)} \equiv \sum_{\tau=1}^{t} \beta^{t-\tau} \left|z_j(\tau)\right|^2 = \beta\, \widehat{r}_{z_j z_j}^{(t-1)} + \left|z_j(t)\right|^2 .$$

D. Recursive Estimation of Speech Variance

Unlike $\mathbf{H}$ and ϕ_v, the speech variance $\phi_x(t)$ cannot be assumed slowly time-varying. Unfortunately, the REM scheme described in Sec. IV-A is inappropriate for estimating $\phi_x(t)$, which can be shown by repeating the derivations of the M-step for estimating of the clean speech variance. Writing only the components of (17) involving ϕ_x yields:

$$Q\left(\boldsymbol{\Theta}_x \middle| \widehat{\boldsymbol{\Theta}}(t)\right) = - \frac{1-\beta}{2} \sum_{\tau=1}^{t} \beta^{t-\tau} \left[\log \phi_x(\tau) + \frac{\widehat{\left|x(\tau)\right|^2}}{\phi_x(\tau)} \right] . \quad (23)$$

Now, taking into account the time-variations of $\phi_x(t)$, and according to (19), the derivative of $Q\left(\boldsymbol{\Theta}_x \middle| \widehat{\boldsymbol{\Theta}}(t)\right)$ w.r.t. $\phi_x(t+1)$ equals zero. Hence, $\phi_x(t+1)$ cannot be resolved. In contrast, in the iterative algorithm (Sec. III) that uses the entire time segment, calculating the derivative of (14) w.r.t. $\phi_x(\tau)$ does not vanish $\forall \tau \in [1, T]$.

Since $\phi_x(t)$ cannot be determined by the REM procedure, we propose a different solution. Given the clean signal and the statistical model in (2), the ML estimator of $\phi_x(t)$ simplifies to the periodogram, i.e. $\hat{\phi}_x(t) = \left|x(t)\right|^2$. Since $x(t)$ is unobservable, we estimate $\phi_x(t)$ using $E\{\left|x(t)\right|^2 | \mathcal{Z}_t\}$ as in (16a). A suitable variance estimator of the clean speech component at channel j, i.e. $\widehat{\phi}_{x_j}(t)$, can be obtained by using the method presented in [36], utilizing the instantaneous power at the respective microphone,

$$\widehat{\phi}_{x_j}(t) = \left|\widehat{h}_{j,0}(t)\right|^{-2} G_j^2(t) \left|z_j(t)\right|^2 \approx E\{\left|x_j(t)\right|^2 | \mathcal{Z}_t\} , \quad (24)$$

where $G_j^2(t) \left|z_j(t)\right|^2$ is a variance estimator of the early speech component $x_j^e(t) = h_{j,0}(t) x(t)$. The estimator is given by

$$G_j^2(t) = \frac{\zeta_{\text{prior},j}(t)}{\zeta_{\text{prior},j}(t)+1} \left(\frac{1+\nu_j(t)}{\zeta_{\text{post},j}(t)} \right) , \quad (25)$$

where the a priori signal to interference ratio (SIR), the a posteriori SIR, and $\nu_j(t)$ are, respectively, defined as:

$$\zeta_{\text{prior},j}(t) \equiv \frac{\phi_{x_j^e}(t)}{\phi_{r_j}(t) + \phi_{v_j}(t)} , \quad \zeta_{\text{post},j}(t) \equiv \frac{\left|z_j(t)\right|^2}{\phi_{r_j}(t) + \phi_{v_j}(t)} ,$$
$$\nu_j(t) = \frac{\zeta_{\text{prior},j}(t)}{1+\zeta_{\text{prior},j}(t)} \zeta_{\text{post},j}(t) .$$

The calculation of the gain function (25) requires an estimate of the a priori SIR $\widehat{\zeta}_{\text{prior},j}(t)$, the reverberation variance

$\widehat{\phi}_{r_j}(t)$, and the noise variance $\widehat{\phi}_{v_j}(t)$ for each channel. In this work, the a priori SIR is obtained using the decision-directed approach [37]:

$$\widehat{\zeta}_{\text{prior},j}(t) = \alpha_{\text{sir}}\ G_j^2(t-1)\ \zeta_{\text{post},j}(t-1) + [1-\alpha_{\text{sir}}] \max\left\{\zeta_{\text{post},j}(t)-1, \zeta_{\text{min}}\right\}, \tag{26}$$

where α_{sir} is a smoothing factor, and ζ_{min} is a predefined minimum SIR. The spectral variances can now be computed using the estimates derived in Sections IV-B and IV-C.

The reverberation variance $\widehat{\phi}_{r_j}(t)$ can be estimated in two steps. In the first step, we use the acoustical system estimator at frame t (21) and the output of the prediction stage of the second-order statistics to estimate the instantaneous power of the reverberation component denoted by $\widehat{\psi}_{r_j}(t)$:

$$\widehat{\psi}_{r_j}(t) = \widehat{\mathbf{h}}_j^T(t) \left(\widehat{\mathbf{x}_{t|t-1}\mathbf{x}_{t|t-1}^H}\right) \widehat{\mathbf{h}}_j^*(t)\ , \tag{27}$$

where

$$\widehat{\mathbf{x}_{t|t-1}\mathbf{x}_{t|t-1}^H} = \mathbf{\Phi}\left(\widehat{\mathbf{x}}_{t-1|t-1}\widehat{\mathbf{x}}_{t-1|t-1}^H + \mathbf{P}_{t-1|t-1}\right)\mathbf{\Phi}^H\ . \tag{28}$$

Here, it should be stressed that the first coefficient of $\widehat{\mathbf{h}}_j(t)$ is excluded from the calculation of $\widehat{\psi}_{r_j}(t)$ by the definition of $\mathbf{\Phi}$. As a consequence, only the reverberant tail is taken into account. In the second step, the variance $\widehat{\phi}_r$ is computed from $\widehat{\psi}_{r_j}(t)$ by time smoothing and by spatial averaging, assuming that the reverberant field is slowly time-varying and homogeneous:

$$\widehat{\phi}_r(t) = \alpha_r\ \widehat{\phi}_r(t-1) + (1-\alpha_r)\ \frac{1}{J}\sum_{j=1}^{J}\widehat{\psi}_{r_j}(t), \tag{29}$$

with $0 < \alpha_r < 1$.

Finally, the spectral variance $\widehat{\phi}_x(t)$ is obtained by averaging the individual channel estimates, i.e.,

$$\widehat{\phi}_x(t) = \frac{1}{J}\sum_{j=1}^{J}\widehat{\phi}_{x_j}(t). \tag{30}$$

The reverberant model in (5) suffers from an inherent *gain ambiguity* problem, which is evident from the following equation:

$$\mathbf{h}_j^T(t,k)\mathbf{x}_t(k) = \left[g(k)\mathbf{h}_j^T(t,k)\right]\left[\tfrac{1}{g(k)}\mathbf{x}_t(k)\right],$$

where $g(k)$ is an arbitrary frequency-dependent gain. Since the algorithm is independently applied to each frequency bin, this can result in undesired fluctuations in the spectral envelope of the output speech signal. In order to mitigate this problem, we substitute $|h_{j,0}(t)| = 1\ , \forall j$ in (24).

The entire procedure is summarized in Algorithm 2.

V. Practical Considerations

A. Gain Control

Due to estimation errors of the RIRs, some frequency bands may suffer from unnatural attenuation or amplification. As a practical cure to this problem, we constrained the power profile of the system output to match the respective averaged power at a reference input microphone. The output of the algorithm with gain normalization, $\hat{x}_{\text{GN}}$, is finally given by:

$$\hat{x}_{\text{GN}}(t-L+1,k) = b(t-L+1,k)\ \hat{x}(t-L+1|t,k),$$

where

$$b^2(t,k) = \frac{\sum_{\tau=0}^{t} \alpha_b^{\tau}\ |z_1(t-\tau,k)|^2}{\sum_{\tau=0}^{t} \alpha_b^{\tau}\ \widehat{\phi}_x(t-\tau,k)}, \tag{31}$$

and $0 < \alpha_b < 1$ is a smoothing factor. To save memory resources, the numerator and denominator in (31) are calculated recursively. Application of this procedure guarantees the preservation of the average spectral profile of the input signal without affecting the convergence of the algorithm. In the current paper, we focus on dereverberation in a relatively low noise scenarios, hence the contribution of the noise component to $|z_1(t-\tau,k)|^2$ can be ignored in the normalization procedure. In higher-noise scenarios, the noise variance should be subtracted from $|z_1(t-\tau,k)|^2$.

B. Minimum Noise Variance

In high signal-to-noise ratio (SNR) scenarios, estimation errors might result in a negative noise variance estimate in (22). To avoid this negative estimate, the constraints $\phi_{v_j} \geq \phi_m$ can be incorporated in the optimization, where ϕ_m denotes the lower bound on the noise variances. Following [38], we obtain the auxiliary function for the constrained problem by adding

Algorithm 2: Kalman-EM for Dereverberation summary.

for *t=1 to T* **do**

1) Calculate $\widehat{\phi}_{v_j}(t)$, $\widehat{\phi}_{r_j}(t)$, and $\widehat{\phi}_{z_j}(t)$ for all j.
2) Estimate the variance of speech $\widehat{\phi}_x(t)$.
3) Execute one step of Kalman filtering to get $\widehat{\mathbf{x}}_{t|t}$ and the respective estimation error $\mathbf{P}_{t|t}$.
4) Update the accumulated second-order statistics: $\widehat{\mathbf{R}}_{xx}^{(t)}$, $\widehat{\mathbf{r}}_{xz_j}^{(t)}$, and $\widehat{r}_{z_jz_j}^{(t)}$.
5) Re-estimate the acoustic parameters: $\widehat{\mathbf{h}}_j(t+1)$ and $\widehat{\phi}_{v_j}(t+1)$.

end

Lagrange multipliers λ_j with slack variables ζ_j to (17):

$$F\left(\phi_{v_1},\ldots,\phi_{v_J}\lambda_1,\ldots,\lambda_J,\zeta_1,\ldots,\zeta_J\right) = \\ -\frac{1-\beta}{2}\sum_{j=1}^{J}\sum_{\tau=1}^{t}\beta^{t-\tau}\left[\log\phi_{v_j}+\frac{1}{2\phi_{v_j}}\widehat{\left|z_j(\tau)-\mathbf{h}_j^T\mathbf{x}_\tau\right|^2}\right] \\ +\sum_{j=1}^{J}\lambda_j\left[\phi_{v_j}-\phi_m-\zeta_j^2\right]+C, \quad (32)$$

where C is independent of all ϕ_{v_j}. Calculating the derivatives w.r.t. λ_j and ζ_j, and setting the result to zero, we conclude that either λ_j or ζ_j equals to zero, for every $1 \le j \le J$. If λ_j is zero, (32) reduces to the unconstrained maximization problem w.r.t ϕ_{v_j}. If ζ_j is zero we get $\phi_{v_j} = \phi_m$. Therefore, the constrained solution is obtained by adding a lower bound to the unconstrained solution given by (22):

$$\widehat{\phi}_{v_j}(t,k) \leftarrow \max\left[\widehat{\phi}_{v_j}(t,k)\ ,\ \phi_m(t,k)\right]. \quad (33)$$

In the experimental study described in Sec. VI, the lower bound $\phi_m(t,k)$ was set to a fraction, determined by A_{m} dB, of the smoothed value of the spatial average of the instantaneous power of each of the microphones:

$$\phi_m(t,k) = 10^{A_{\mathrm{m}}/10}\left[(1-\alpha_m)\sum_{\tau=0}^{t}\alpha_m^{\tau}\frac{1}{J}\sum_{j=1}^{J}|z_j(t-\tau,k)|^2\right]$$

where $0 < \alpha_m < 1$ is a smoothing factor.

VI. Performance Evaluation

The RKEMD algorithm was evaluated in both static and dynamic scenarios. Experiments were conducted in the Speech & Acoustic Lab of the Faculty of Engineering at Bar-Ilan University, with controllable reverberation time. The room dimensions are $6 \times 5.9 \times 2.3$ m (length $\times$ width $\times$ height).

The STFT analysis window for both scenarios was set to a 32 ms Hamming window, with 50% overlap. Higher percentage of overlap will result in a significant dependency between adjacent frames, rendering the statistical model of Sec. II-A inaccurate, and hence leading to performance degradation. The system length L should be chosen in accordance with the sampling rate, the length of the RIR in the time-domain, the analysis window length, and the overlap between successive frames. For the T_{60} values tested in Sec. VI-A and VI-B, L should be chosen between 30 and 60 frames. However, we have found that when L increases, the estimation error increases as well, and test results show that choosing L to be lower than the actual systems length may improve the performance, in addition to the reduction in the computational load. Therefore, L was set to 20 frames. The code was implemented in MATLAB, and the processing was performed on an Intel Core i7-3770 CPU at 3.4 GHz with four cores, and using 8 GB of RAM. Since the algorithm processes each frequency band independently, the frequency bands were processed in parallel using eight threads to reduce the processing time. It required 4.88 seconds of computation to process 10 seconds of four-channel signal sampled at 16 kHz.

Some of the parameters defined in previous sections should be determined in advance, and the values chosen for this

TABLE I
VALUES OF VARIOUS ALGORITHM PARAMETERS.

Parameter	Value	Parameter	Value
ϵ	10^{-10}	β	0.95
G_{min}	0.2	α_r	0.5
ζ_{min}	0.03	α_b	0.99
A_{m}	-20	α_m	0.99

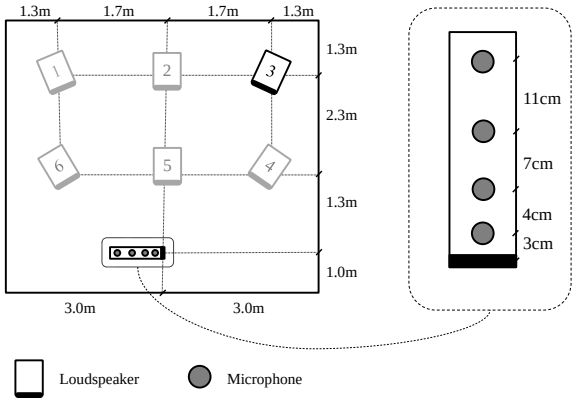


Fig. 2. Lab and microphone array setup in the static experiment.

experimental study are depicted in Table I. These parameters were identical for both static and dynamic experiments.

A. Experiments Using Loudspeakers

For the static scenario, three different reverberation times (T_{60}) were tested: 480, 630, and 940 ms. For each T_{60} value, the room was adjusted to the required reverberation level, which was verified by the calculation of energy decay curves (EDCs) that were extracted from several RIR measurements. Different speech signals related to eight different speakers from the TIMIT database were played from one of six positions in the room using Fostex 6301B loudspeakers. The reverberant signal was captured by a linear array with four AKG CK32 omni-directional microphones. For performance evaluation, a reference signal was also measured at a distance of 5 cm from the active loudspeaker. The sources were positioned at 150 cm height. The setup is depicted in Fig. 2.

Eight different clean speech signals were recorded from each position and for each reverberation time. Each signal is 60 s long and belongs to a different human speaker. The total number of experiments therefore equals 144, comprising 2 hours and 24 minutes of reverberant speech.

We used three objective measures to evaluate the performance of the proposed algorithm, namely the speech to reverberation modulation energy ratio (SRMR) [39], the log-spectral distance (LSD) and the frequency-weighted signal to interference ratio (WSIR) [40]. The LSD between x and $\tilde{z} \in \{z_1, \hat{x}\}$ in frame t is defined as

$$\text{LSD}(t) = \sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left[10\log_{10}\left(\frac{\max\left\{|x(t,k)|, \epsilon(x)\right\}}{\max\left\{|\tilde{z}(t,k)|, \epsilon(\tilde{z})\right\}}\right)\right]^2}, \tag{34}$$

TABLE II
AVERAGE OBJECTIVE MEASURES IN THE NOISELESS CASE, ACCORDING TO THE REVERBERATION TIME T_{60} IN THE ROOM. THE NOISY AND REVERBERANT (INPUT) THE RKEMD RESULT (OUTPUT) AND THE IMPROVEMENT ARE DISPLAYED.

Measure	T_{60} (ms)	Input	Output	Improvement
SRMR	480	4.66	7.67	3.02
	630	4.37	7.64	3.27
	940	3.57	6.66	3.10
	All	4.25	7.33	3.07
WSIR	480	-0.57	5.44	6.02
	630	-0.77	5.55	6.32
	940	-1.67	5.12	6.79
	All	-0.88	5.37	6.26
LSD	480	2.39	1.86	0.53
	630	2.53	1.89	0.64
	940	2.97	2.05	0.92
	All	2.59	1.93	0.66

where the minimum value is calculated as

$$\epsilon(y) = 10^{-A_{\text{LSD}}/10} \max_{t,k} |y(t,k)| \,,$$

and A_{LSD} is set to the desired dynamic range, which was chosen to be 60 dB. The interference component in the WSIR is defined as the reverberation plus noise, and the measure is calculated as in [41]:

$$\text{WSIR}(t) = \frac{\sum_{\mathcal{k}=0}^{\mathcal{K}-1} w(t,\mathcal{k}) \log_{10} \frac{|\mathcal{x}(t,\mathcal{k})|^2}{|\tilde{z}(t,\mathcal{k}) - \mathcal{x}(t,\mathcal{k})|^2}}{\sum_{\mathcal{k}=0}^{\mathcal{K}-1} w(t,\mathcal{k})} \tag{35}$$

where $\mathcal{x}(t,\mathcal{k})$ is the clean speech signal split into bands in accordance with the human auditory system. The evaluated signals $\tilde{z}(t,\mathcal{k}) \in \{z_1(t,\mathcal{k}), \hat{\mathcal{x}}(t,\mathcal{k})\}$, were obtained from the noisy-reverberant and the estimated signals in a same procedure. The weighting factors $w(t,\mathcal{k})$ are determined from the clean signal

$$w(t,\mathcal{k}) = |\mathcal{x}(t,\mathcal{k}))|^{0.2} \,.$$

While a reduction in reverberation is indicated by a higher SRMR value, better speech estimates would be indicated by lower LSD and higher WSIR values. The LSD, WSIR, and SRMR average results for the noiseless case are summarized in Table II. The direct-to-reverberant ratio (DRR) is an important measure to the quality of a reverberant signal, and is defined as follows:

$$\text{DRR} = 10 \log_{10} \frac{\sum_{l'=0}^{L'_d-1} |h_{1,l'}|^2}{\sum_{l'=L'_d}^{\infty} |h_{1,l'}|^2} \tag{36}$$

where L_d is the number of coefficients in time domain dominated by the direct path. In all our experiments L_d was set to 120 coefficients.

The setup depicted in Fig. 2 is comprised of various source-microphone distances, and hence different DRR values for each reverberation time. For each of the loudspeaker positions and for every reverberation time, the RIRs were measured, and

TABLE III
IMPROVEMENT VERSUS DRR VALUES IN THE NOISELESS CASE. THE RECORDINGS FROM THE DIFFERENT T_{60} VALUES AND DIFFERENT SOURCE POSITIONS WERE CLASSIFIED TO FOUR GROUPS, ACCORDING TO THE DRR VALUES. THE AVERAGE RESULTS FOR EACH GROUP ARE DISPLAYED.

Measure	DRR Range	Input	Output	Improvement
SRMR	{-2,3}	4.34	7.55	3.21
	{-4.5,-2}	4.73	8.48	3.75
	{-7,-4.5}	4.20	6.97	2.76
	{-10,-7}	3.51	6.49	2.97
	All	4.25	7.33	3.07
WSIR	{-2,3}	0.50	6.66	6.16
	{-4.5,-2}	-0.12	6.41	6.53
	{-7,-4.5}	-1.59	4.66	6.25
	{-10,-7}	-2.52	4.11	6.63
	All	-0.88	5.37	6.26
LSD	{-2,3}	2.43	1.81	0.62
	{-4.5,-2}	2.49	1.75	0.75
	{-7,-4.5}	2.61	1.99	0.62
	{-10,-7}	2.98	2.15	0.84
	All	2.59	1.93	0.66

the DRR values were calculated. The experiments were segmented according to their input DRR values. The average results per segment are displayed in Table III. As expected, the values of the SRMR and the WSIR at the input decrease for lower DRR values, while the input LSD increases. For all the tested DRR values, and for all the calculated measures, the algorithm achieves approximately the same improvement.

We also investigated the influence of the number of microphones on the algorithm performance. For that, eight microphone signals were recorded, 4 of which were used in the above experiments as depicted in Fig. 2. To evaluate the performance of the algorithm with different number of inputs, we used 8, 4, 2 and 1 microphone signals from the database. The objective measures for this comparison are displayed in Figure Fig. 3. While only marginal change in objective measures is demonstrated, informal listening tests indicate a significant decrease in musical noise when more microphones are used. When eight microphone are used the musical noise is hardly noticeable[1].

We compared the RKEMD algorithm with the KEMD algorithm [19], and a multichannel spectral enhancement (MCSE) algorithm for dereverberation [42]. The MCSE comprises a nonlinear spatial processor, followed by a single-channel spectral enhancement algorithm. The spatial processor first aligns the observed signals according to the direction of arrival (DOA) of the direct arrival. Then, the averaged instantaneous power of the aligned signal is computed according to:

$$\widehat{\psi}_z(t,k) = \frac{1}{J}\sum_{j=1}^{J}\left|z_j(t)e^{j\omega_k\tau_{1j}}\right|^2 \ , \tag{37}$$

where $\omega_k = \frac{2\pi jk}{K}$, and τ_{1j} denotes the time difference of arrival (TDOA) of the desired source signal between the j-th and

[1]Signals available at http://www.eng.biu.ac.il/gannot/speech-enhancement.

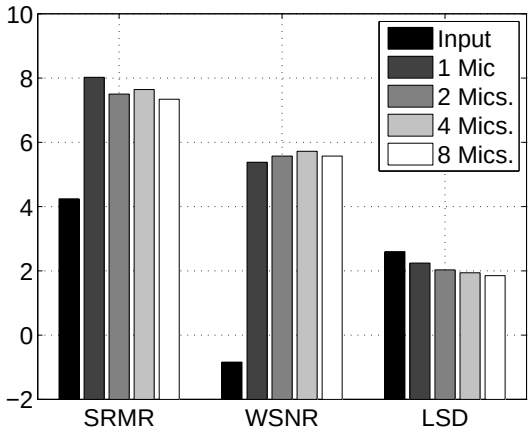


Fig. 3. Objective measures for different number of microphones in the noiseless case.

the first microphone. The phase is extracted from the average of the aligned signals:

$$\varphi(t,k) = \arg\left\{ \frac{1}{J} \sum_{j=1}^{J} z_j(t) e^{j\omega_k \tau_{1j}} \right\} . \tag{38}$$

Finally, the output of the spatial processor is given by $\sqrt{\widehat{\psi}_z(t,k)} e^{j\varphi(t,k)}$. Now, a single-channel spectral enhancement algorithm, based on a statistical model for the reverberation is applied. A comparison is presented in Table IV.

It is clear from Table IV that the RKEMD algorithm achieves a better performance than the MCSE and KEMD algorithms. While the MCSE makes use of the decision-directed spectral enhancement scheme, and the KEMD makes use of the linear Kalman filtering, the RKEMD comprises both spectral enhancement and Kalman filtering and hence obtains better results.

The enhanced performance of the RKEMD algorithm with respect to the KEMD algorithm can be attributed to the advantages of the REM scheme over the EM scheme, as was reported also in [30]. To verify this hypothesis, we compared the precision of the iterative-EM and the recursive-EM estimators of $\mathbf{H}$, using synthesized signals. The results showed that the REM estimator of $\mathbf{H}$, which by definition uses the data set only once, had the same precision as the EM estimator obtained by numerous iterations. It may be suggested that these results are a consequence of the large number of parameter updates (M-steps) in the recursive scheme, as compared with the iterative scheme.

The proposed algorithm exhibits noise reduction capabilities as well. To evaluate the performance of both dereverberation and noise reduction, we added sensors noise to the measurements in several different levels. The sensor noise was generated independently for each channel using a first-order AR process, and as a consequence was uncorrelated between the channels, as assumed in Sec. II-A. The reverberated-signal to noise ratio (RSNR) is defined as the ratio of noise-free reverberant signal power and the additive noise power:

$$\text{RSNR} = 10 \log_{10} \frac{\sum_{t,k} |z(t,k) - v(t,k)|^2}{\sum_{t,k} |v(t,k)|^2} . \tag{39}$$

Note that the average RSNR of the loudspeaker recordings is 40 dB, due to sensors noise. Sonograms for the clean, noisy-reverberant, and output signals, obtained using $J = 4$ microphones, are depicted at Fig. 4, and the average measures for

TABLE IV
COMPARISON BETWEEN THE SUGGESTED METHOD AND THE ALGORITHMS PRESENTED IN [19] AND [42]. THESE ARE THE RESULTS FOR EVERY DRR RANGE OBTAINED BY PROCESSING THE ENTIRE STATIC DATABASE WITHOUT ADDITIONAL NOISE.

Measure	DRR Range	Input	MCSE	KEMD	RKEMD
SRMR	{-2,3}	4.34	6.03	5.34	7.55
	{-4.5,-2}	4.73	6.41	5.74	8.48
	{-7,-4.5}	4.20	5.83	5.00	6.97
	{-10,-7}	3.51	5.07	4.41	6.49
	All	4.25	5.84	5.11	7.33
WSIR	{-2,3}	0.50	4.72	5.50	6.66
	{-4.5,-2}	-0.12	3.73	4.89	6.41
	{-7,-4.5}	-1.59	2.81	3.56	4.66
	{-10,-7}	-2.52	2.63	2.73	4.11
	All	-0.88	3.47	4.10	5.37
LSD	{-2,3}	2.43	2.08	1.96	1.81
	{-4.5,-2}	2.49	2.15	1.99	1.75
	{-7,-4.5}	2.61	2.25	2.13	1.99
	{-10,-7}	2.98	2.48	2.36	2.15
	All	2.59	2.24	2.11	1.93

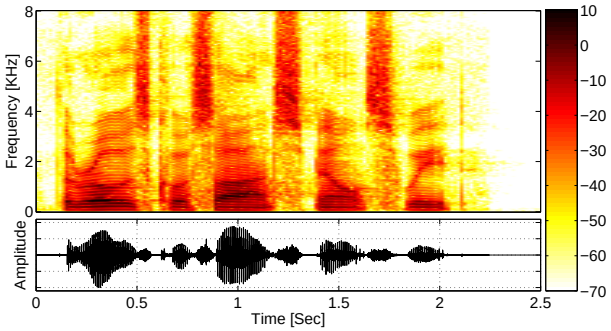
different RSNR values are depicted in Fig. 5.

B. Experiments Using Human Speakers

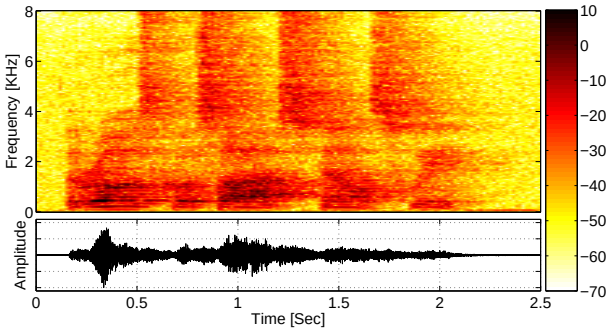
The major attribute of the online EM algorithm is its ability to track time-varying parameters. This ability is crucial in the practical case of dynamic scenario where the RIRs are time-varying. To demonstrate the algorithm's tracking ability, we recorded a reverberant dynamic speech database. Subjects were requested to read an article from the New-York Times while moving in the room according to predefined instructions. The room and the microphone array for this experiment are depicted in Fig. 6. The height of the array in this experiment was 130 cm, and the reverberation time was set to $T_{60} = 750$ ms. Since the power of natural voice is lower than the power of the loudspeakers, the RSNR in the dynamic scenario is only 20 dB.

Two types of experiments were conducted. The first type involved speaking in different locations in the room, and walking naturally between them. For example - speaking a few sentences sitting in chair 1, another sentence while walking to point 2, and some other sentences standing in point 2. The second type consists of only slight movements - head turning, sitting down and standing up. For example - speaking one paragraph facing the microphone array, then turning the head to the opposite direction and speaking another paragraph. We stress that unlike the static scenario involving loudspeakers, in the dynamic scenario the sentences were uttered by human speakers that were not absolutely static even if requested to stand or sit in a single position, due to inevitable natural behavior. Four different dynamic experiments were defined and each was conducted with four native English speakers (two female and two male speakers), while every experiment lasted about 3 minutes. The total length of the database is hence 48 minutes of natural speaking speech.

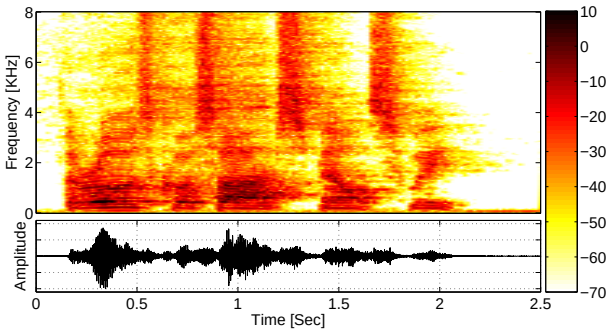
The performance in the human speakers scenario was evaluated by splitting each experiment to the *static* parts, where the subjects were standing or sitting, and the *dynamic* parts, where the subjects were moving. Average results for both parts are depicted in Table V. A significant improvement is obtained in both parts, where, as expected, better performance is achieved



(a) Clean signal



(b) Noisy and reverberant signal at the first microphone.



(c) Output signal ($J = 4$)

Fig. 4. Sonograms and waveforms for signal played from loudspeaker #1, $T_{60} = 940$ ms and RSNR of 5 dB.

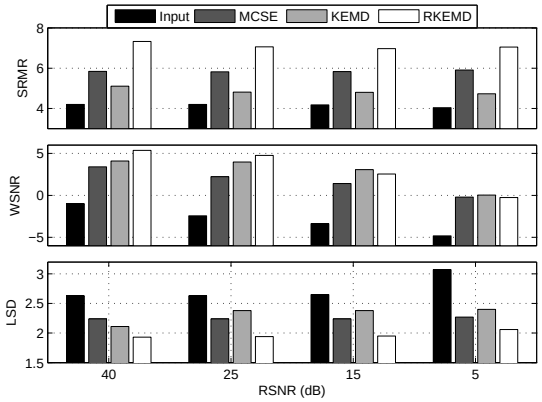


Fig. 5. Average objective measures as a function of the RSNR input level.

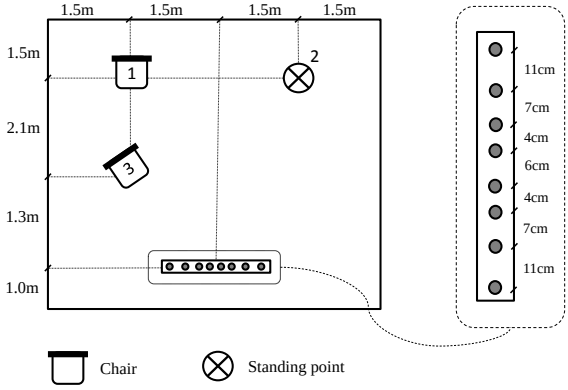


Fig. 6. Lab and microphone array setup used for the dynamic reverberant database. In the first type of experiments, subjects were requested to walk from one position to another, where in the second type, only minor movements were involved.

Measure	Case	Input	Output	Improvement
SRMR	Static	3.41	5.85	2.44
	Dynamic	3.42	5.24	1.83
	Average	3.41	5.55	2.14
WSIR	Static	-1.30	3.76	5.06
	Dynamic	-2.14	2.46	4.60
	Average	-1.72	3.11	4.83
LSD	Static	3.11	2.36	0.75
	Dynamic	3.12	2.48	0.64
	Average	3.11	2.42	0.70

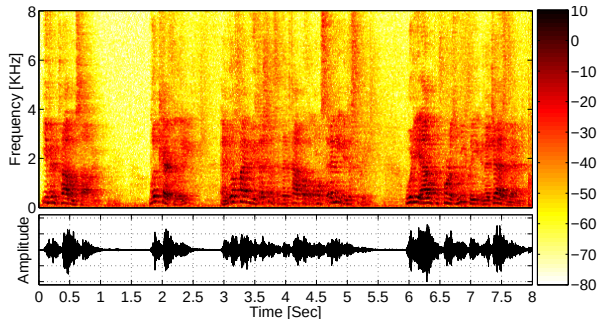
TABLE V
RKEMD PERFORMANCE IN THE HUMAN SPEAKERS SCENARIO.

in the *static* parts. It can be seen that the performance in the *static* parts of the human speakers scenario (Table V) is inferior to the performance in the loudspeakers recordings (Table II). The lower scores can be explained by the inevitable movements of natural speakers even in the *static* parts. Sonograms, waveforms, and the frame-wise WSIR values are depicted in Fig. 7, where the robustness of the algorithm to natural movements is depicted. A median filter with 15 frames was applied to smooth the WSIR estimate (35) for both the reverberant and output signals.

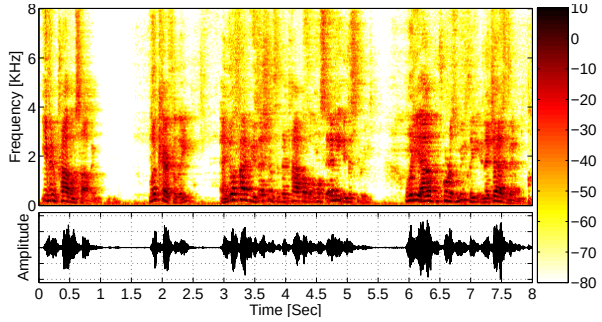
Informal listening tests revealed a significant dereverberation and improvement of the sound quality by the proposed algorithm. Some quality degradation was noticeable when the speaker was walking from one point to another (first type of experiments), with fast recovery of the algorithm after the speaker arrived to its destination. In the second type of experiments (involving only minor movements), almost no degradation is perceived during movements.

VII. CONCLUSION

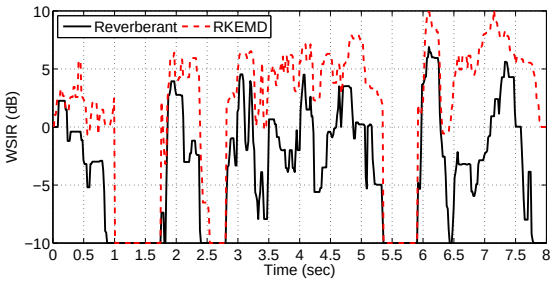
A recursive EM algorithm for speech dereverberation was presented, where the acoustic parameters and the enhanced signal are estimated simultaneously in an online manner. We assumed a, possibly moving, single desired sound source, and slowly time-varying and spatially-white noise. For the considered scenarios with an RSNR between 5 and 40 dB and reverberation times between 0.48 and 0.94 s, the proposed algorithm was able to improve the WSIR by up to 5 dB and the SRMR by up to 3. For these scenarios, the proposed RKEMD algorithm provided similar or better results compared with the previously



(a) Reverberant input signal



(b) RKEMD output signal



(c) WSIR values of the reverberant and RKEMD output signals

Fig. 7. Sonograms, waveforms, and median smoothed WSIR values for a moving speaker, $T_{60} = 750$ ms. The speaker was first standing at point 2 (0-1.5 sec), then started walking to point 3 (1.5-8 sec).

proposed KEMD algorithm that is not suitable for online processing and not able to handle time-varying acoustic systems and noise. Finally, similar performance was obtained by the RKEMD algorithm in terms of SRMR, WSIR and LSD using both a static and dynamic sound source positions.

The proposed method can be viewed as a recursive extension of [19] with an improved speech variance estimator. However, the recursive EM algorithm outperforms the accuracy of the iterative EM without the need to process the same data more than once. These results are in agreement with the conclusions in [30], in which the iterative and recursive EM approaches are compared.

APPENDIX A
ONLINE EM ALGORITHM

The REM scheme defined in [29] is an online version of the original EM [35], and has a similar structure. Following the notation in Sec. II, the E-step of the online algorithm is

$$Q\left[\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t)\right] = Q\left[\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t-1)\right] + \gamma_t \cdot \left\{E\left\{\log f\left(\mathbf{x}_t, \mathbf{z}_t; \boldsymbol{\Theta}\right) \;\middle|\; \mathbf{z}_t, \widehat{\boldsymbol{\Theta}}(t)\right\} - Q\left[\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t-1)\right]\right\}, \tag{40}$$

where $\widehat{\boldsymbol{\Theta}}(t)$ is the parameter estimation at time t, and $0 < \gamma_t < 1$ is a smoothing factor. As compared to (12), where the entire data set $\mathcal{Z}$ was used, only the latest observation $\mathbf{z}_t$ is used in (40). In the M-step, the updated parameters $\widehat{\boldsymbol{\Theta}}(t+1)$ are obtained, similarly to (13), by the maximization:

$$\widehat{\boldsymbol{\Theta}}(t+1) = \arg\max_{\boldsymbol{\Theta}} \left\{Q\left[\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t)\right]\right\}. \tag{41}$$

In order to develop a solution to the problem formulated in Sec. II, we define

$$q\left[\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t)\right] \equiv E\left\{\log f\left[\mathbf{x}_t, \mathbf{z}_t; \boldsymbol{\Theta}\right] \;\middle|\; \mathcal{Z}_t, \widehat{\boldsymbol{\Theta}}(t)\right\}, \tag{42}$$

and for a constant smoothing factor $\gamma_t = 1 - \beta$, such that the recursion in (40) can be written as:

$$\begin{aligned} Q\left[\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t)\right] &= \beta \cdot Q\left[\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t-1)\right] + (1-\beta)\; q\left[\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t)\right] \\ &= (1-\beta)\sum_{\tau=1}^{t} \beta^{t-\tau}\; q\left[\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(\tau)\right]. \end{aligned} \tag{43}$$

Note that the expectation in (40) is only taking the last measurement $\mathbf{z}_t$ into account, while in (42) the expectation takes into account all the previous measurements $\mathcal{Z}_t$. Although, apparently different, it can be straightforwardly shown that all derivations leading to the proof of convergence of Cappè and Moulines REM procedure [29] are still valid also for (42). The proof of this claim is beyond the scope of this contribution.

The complete log-likelihood function (11) is separable in t, i.e.,

$$\log f(\mathcal{X}, \mathcal{Z}; \boldsymbol{\Theta}) = \sum_{t=1}^{T} \log f\left[\mathbf{x}_t, \mathbf{z}_t; \boldsymbol{\Theta}\right], \tag{44}$$

where

$$\begin{aligned} \log f\left[\mathbf{x}_t, \mathbf{z}_t; \boldsymbol{\Theta}\right] = &-\frac{1}{2}\left[\log \phi_x(t) + \frac{|x(t)|^2}{\phi_x(t)}\right] \\ &- \frac{1}{2}\sum_{j=1}^{J}\left[\log \phi_{v_j} + \frac{1}{\phi_{v_j}}\left|z_j(t) - \mathbf{h}_j^T \mathbf{x}_t\right|^2\right]. \end{aligned} \tag{45}$$

Substituting (45) and (42), in the recursive auxiliary function (43), the recursive auxiliary function (17) is obtained.

APPENDIX B
CONDITIONS FOR THE CONVERGENCE OF THE ONLINE EM ALGORITHM

The convergence properties of the REM algorithm proved in [29] requires a few assumptions regarding to the statistical model of the complete data. The primary requirement is that the complete data likelihood function will be of the exponential family:

$$\log f\left(\mathcal{Y};\boldsymbol{\Theta}\right) = H(\mathcal{Y}) - \Psi(\boldsymbol{\Theta}) + \left\langle \Phi(\boldsymbol{\Theta}), \mathbf{S}(\mathcal{Y}) \right\rangle . \tag{46}$$

The likelihood function in (45) can be written in the required form by defining:

$$\Psi(\boldsymbol{\Theta}) = \tfrac{1}{2}\sum_{j=1}^{J} \log \phi_{v_j}, \tag{47}$$

$$\Phi_j(\boldsymbol{\Theta}) = \phi_{v_j}^{-1} \left[h_{j,0}\mathbf{h}_j^H, \ldots, h_{j,L-1}\mathbf{h}_j^H, \mathbf{h}_j^H, \mathbf{h}_j^T, 1\right]^T, \tag{48}$$

$$\mathbf{S}_j(\mathcal{Y}) = \left[x_t^*\mathbf{x}_t^T, \ldots, x_{t-L+1}^*\mathbf{x}_t^T, z_j^*(t)\mathbf{x}_t^T, z_j(t)\mathbf{x}_t^H, |z_j(t)|^2\right]^T, \tag{49}$$

and $H(\mathcal{Y}) = 0$. Finally, the structure in (46) is obtained by defining $\Phi(\boldsymbol{\Theta})$ and $\mathbf{S}(\mathcal{Y})$ using the concatenation $\left[\Phi_1^T(\boldsymbol{\Theta}), \ldots, \Phi_J^T(\boldsymbol{\Theta})\right]$ and $\left[\mathbf{S}_1^T(\mathcal{Y}), \ldots, \mathbf{S}_J^T(\mathcal{Y})\right]$, respectively.

The *regularity* conditions mentioned in [29] require that (46) is twice continuously differentiable w.r.t. $\boldsymbol{\Theta}$, and have a single global maximum that is obtained by a continuously differentiable function of the sufficient statistic. These requirements are satisfied as can be seen in the derivation of (21). Another requirement is related to the sufficient statistics (49), and assumes its expected value is bounded. This assumption is also satisfied, since (49) is a combination of Gaussian random variables with final variances.

REFERENCES

[1] J. D. Polack, "La transmission de l'energie sonore dans les salles," Ph.D. dissertation, Université du Maine, Le Mans, 1988.

[2] K. Lebart, J. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, pp. 359–366, 2001.

[3] E. A. P. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4. IEEE, Mar. 2005, pp. 173–176.

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

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

[6] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, pp. 1074–1090, 2003.

[7] M. Miyoshi and Y. Kenda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 2, pp. 145–152, 1988.

[8] W. Zhang, E. A. Habets, and P. A. Naylor, "A system-identification-error-robust method for equalization of multichannel acoustic systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2010, p. 109–112.

[9] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1879–1890, 2013.

[10] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[11] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 69–84, Jan. 2011.

[12] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1369–1380, Jul. 2013.

[13] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[14] S. Gannot, "Speech processing utilizing the kalman filter," *Instrumentation & Measurement Magazine, IEEE*, vol. 15, no. 3, p. 10–14, 2012.

[15] K. K. Paliwal and A. Basu, "A speech enhancement method based on kalman filtering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 12. IEEE, Apr. 1987, pp. 177– 180.

[16] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Transactions on Signal Processing*, vol. 39, no. 8, pp. 1732–1742, Aug. 1991.

[17] E. Weinstein, A. Oppenheim, M. Feder, and J. Buck, "Iterative and sequential algorithms for multisensor signal enhancement," *IEEE Transactions on Signal Processing*, vol. 42, pp. 846–859, Apr. 1994.

[18] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential kalman filter-based speech enhancement algorithms," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 373–385, 1998.

[19] B. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation using expectation-maximization and kalman smoothing," in *European Signal Processing Conference (EUSIPCO)*, Marakech, Morocco, Sep. 2013.

[20] D. Schmid, S. Malik, and G. Enzner, "An expectation-maximization algorithm for multichannel adaptive speech dereverberation in the frequency-domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 17–20.

[21] S. Gannot and M. Moonen, "On the application of the unscented kalman filter to speech processing," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2001.

[22] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, p. 401–422, 2004.

[23] C. Evers and J. R. Hopgood, "Marginalization of static observation parameters in a rao-blackwellized particle filter with application to sequential blind speech dereverberation," in *European Signal Processing Conference (EUSIPCO)*, 2009, pp. 1437–1441.

[24] D. Titterington, "Recursive parameter estimation using incomplete data," *Journal of the Royal Statistical Society*, vol. 46, no. 2, 1984.

[25] P.-J. Chung and J. Bohme, "Recursive EM and SAGE-inspired algorithms with application to DOA estimation," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 2664–2677, Aug. 2005.

[26] S. Wang and Y. Zhao, "Almost sure convergence of titterington's recursive estimator for mixture models," *Statistics & Probability Letters*, vol. 76, no. 18, pp. 2001–2006, Dec. 2006.

[27] B. Delyon, "General results on the convergence of stochastic algorithms," *IEEE Transactions on Automatic Control*, vol. 41, no. 9, p. 1245–1255, 1996.

[28] L. Frenkel and M. Feder, "Recursive expectation-maximization (EM) algorithms for time-varying parameters with applications to multiple target tracking," *IEEE Transactions on Signal Processing*, vol. 47, no. 2, pp. 306–320, 1999.

[29] O. Capp'e and E. Moulines, "On-line expectation-maximization algorithm for latent data models," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 71, no. 3, pp. 593–613, 2009.

[30] P. Liang and D. Klein, "Online EM for unsupervised models," in *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, 2009, p. 611–619.

[31] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 392–402, Feb. 2014.

[32] I. Cohen, "Modeling speech signals in the time–frequency domain using GARCH," *Signal Processing*, vol. 84, no. 12, pp. 2453–2459, 2004.

[33] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[34] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, May 2009.

[35] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[36] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the ephraim and malah suppression rule for audio signal enhancement," *EURASIP Journal on Advances in Signal Processing, Special Issue on Digital Audio for Multimedia Communications*, vol. 2003, no. 10, p. 1043–1051, 2003.

[37] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[38] S. S. Rao, *Engineering Optimization: Theory and Practice*. John Wiley & Sons, Jul. 2009.

[39] T. H. Falk, C. Zheng, and W. Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Speech and Audio Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[40] J. M. Tribolet, P. Noll, B. McDermott, and R. Crochiere, "A study of complexity and quality of speech waveform coders," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, 1978, p. 586–590.

[41] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, p. 229–238, 2008.

[42] E. A. P. Habets, "Speech dereverberation using statistical reverberation models," in *Speech Dereverberation*. Springer, 2010, pp. 57–93.

Boaz Schwartz (M14) was born in Rehovot, Israel, in 1982. He received a BSc and MSc degree in Electrical Engineering from Bar-Ilan University, Israel, in 2010 and 2013, respectively. In the years 2008-2010 he participated in the development of the WiMAX protocol in Intel, Israel. He is currently a PhD student at the Speech and Signal Processing laboratory of the Faculty of Engineering at Bar-Ilan. His research interests are speech dereverberation and binaural hearing aids.

Sharon Gannot (S'92-M'01-SM'06) received his B.Sc. degree (summa cum laude) from the Technion Israel Institute of Technology, Haifa, Israel in 1986 and the M.Sc. (cum laude) and Ph.D. degrees from Tel-Aviv University, Israel in 1995 and 2000 respectively, all in electrical engineering. In 2001 he held a post-doctoral position at the department of Electrical Engineering (ESAT-SISTA) at K.U.Leuven, Belgium. From 2002 to 2003 he held a research and teaching position at the Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. Currently, he is an Associate Professor at the Faculty of Engineering, Bar-Ilan University, Israel, where he is heading the Speech and Signal Processing laboratory. Prof. Gannot is the recipient of Bar-Ilan University outstanding lecturer award for 2010 and 2014.

Prof. Gannot has served as an Associate Editor of the EURASIP Journal of Advances in Signal Processing in 2003-2012, and as an Editor of two special issues on Multi-microphone Speech Processing of the same journal. He has also served as a guest editor of ELSEVIER Speech Communication and Signal Processing journals. Prof. Gannot has served as an Associate Editor of IEEE Transactions on Speech, Audio and Language Processing in 2009-2013. Currently, he is a Senior Area Chair of the same journal. He also serves as a reviewer of many IEEE journals and conferences. Prof. Gannot is a member of the Audio and Acoustic Signal Processing (AASP) technical committee of the IEEE since Jan., 2010. He is also a member of the Technical and Steering committee of the International Workshop on Acoustic Signal Enhancement (IWAENC) since 2005 and was the general co-chair of IWAENC held at Tel-Aviv, Israel in August 2010. Prof. Gannot has served as the general co-chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in October 2013. Prof. Gannot was selected (with colleagues) to present a tutorial sessions in ICASSP 2012, EUSIPCO 2012, ICASSP 2013 and EUSIPCO 2013. Prof. Gannot research interests include multi-microphone speech processing and specifically distributed algorithms for ad hoc microphone arrays for noise reduction and speaker separation; dereverberation; single microphone speech enhancement and speaker localization and tracking.

Emanuël A.P. Habets (S'02-M'07-SM'11) is an Associate Professor at the International Audio Laboratories Erlangen (a joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg and Fraunhofer IIS), and Head of the Spatial Audio Research Group at Fraunhofer IIS, Germany. He received the B.Sc. degree in electrical engineering from the Hogeschool Limburg, The Netherlands, in 1999, and the M.Sc. and Ph.D. degrees in electrical engineering from the Technische Universiteit Eindhoven, The Netherlands, in 2002 and 2007, respectively.

From 2007 until 2009, he was a Postdoctoral Fellow at the Technion - Israel Institute of Technology and at the Bar-Ilan University, Israel. From 2009 until 2010, he was a Research Fellow in the Communication and Signal Processing Group at Imperial College London, U.K.

His research activities center around audio and acoustic signal processing, and include spatial audio signal processing, spatial sound recording and reproduction, speech enhancement (dereverberation, noise reduction, echo reduction), and sound localization and tracking.

Dr. Habets was a member of the organization committee of the 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC) in Eindhoven, The Netherlands, a general co-chair of the 2013 International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in New Paltz, New York, and general co-chair of the 2014 International Conference on Spatial Audio (ICSA) in Erlangen, Germany. He is a Senior Member of the IEEE, a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing (2011-2016) and a member of the IEEE Signal Processing Society Standing Committee on Industry Digital Signal Processing Technology (2013-2015). Currently, he is an Associate Editor of the IEEE Signal Processing Letters, and a Guest Editor for the IEEE Journal of Selected Topics in Signal Processing.