An Expectation-Maximization Algorithm for Multi-microphone Speech Dereverberation and Noise Reduction with Coherence Matrix Estimation

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

***Abstract*—In speech communication systems the microphone signals are degraded by reverberation and ambient noise. The reverberant speech can be separated into two components, namely, an early speech component that consists of the direct path and some early reflections, and a late reverberant component that consists of all late reflections. In this paper, a novel algorithm to simultaneously suppress early reflections, late reverberation and ambient noise is presented. The expectation-maximization (EM) algorithm is used to estimate the signals and spatial parameters of the early speech component and the late reverberation components. As a result, a spatially filtered version of the early speech component is estimated in the E-step. The power spectral density (PSD) of the anechoic speech, the relative early transfer functions (RETFs) and the PSD matrix of the late reverberation are estimated in the M-step of the EM algorithm. The algorithm is evaluated using real room impulse responses (RIRs) recorded in our acoustic lab with a reverberation time set to 0.36 s and 0.61 s and several signal-to-noise ratio (SNR) levels. It is shown that significant improvement is obtained and that the proposed algorithm outperforms baseline single-channel and multichannel dereverberation algorithms, as well as a state-of-the-art multichannel dereverberation algorithm.**

***Index Terms*—dereverberation, noise reduction, expectation-maximization.**

I. Introduction

DEREVERBERATION aims at the reduction of reverberation that is caused by a multitude of reflections from the room surfaces. Highly reverberant speech can be difficult to understand for both humans and machines, and can lead to listening fatigue [1]. Dereverberation has become a major research topic in the past decade due to theoretical advances in understanding the reverberation phenomenon and the available computational power.

The current work can be considered as a natural extension of the authors' previous work [2]. In [2], a multi-microphone minimum mean square error (MMSE) estimator of the early speech component was implemented by a minimum variance distortionless response (MVDR) beamformer followed by a postfilter. The room impulse response (RIR) was modeled by two components (that are assumed to be uncorrelated), namely the early reverberation (including the direct path and some early reflections) and the late reverberation [3]–[5]. In the short-time Fourier transform (STFT) domain, the early speech component was modeled as a multiplication of the transformed signal frame and the frequency response of the early reverberation of the RIR. The late reflections are usually dense, since they are a summation of many reflections arriving from all directions. Therefore, the late reverberation and ideal diffuse sound fields have very similar spatial properties. In the STFT domain, the late reverberation was modelled as a diffuse sound field with a time-varying level. The MVDR beamformer was implemented in a generalized sidelobe canceller (GSC) structure. The fixed beamformer (FBF) block of the GSC was implemented as a delay and sum (DS) beamformer to reduce the early reflections, and the blocking matrix (BM) of the GSC was designed to block the early speech component. Consequently, the branches of the GSC became nonorthogonal, unlike in the original GSC [6]. The early speech component was blocked using estimates of the relative early transfer functions (RETFs), for which a new identification procedure was proposed. The reverberation level was estimated by averaging the marginal reverberation levels at the microphones, obtained by using the single-channel estimator proposed in [7].

In the current paper, a procedure for simultaneous estimation of all relevant beamformer's parameters is proposed. First, the dereverberation problem is restated as a maximum likelihood (ML) estimation problem and then an expectation-maximization (EM) algorithm, which alternately estimates the beamformer's parameters and the early speech component, is presented. The anechoic speech is modelled as a Gaussian source and is subsequently multiplied by the early transfer functions (ETFs). The late reverberation is modelled as an additive Gaussian interference, with a time-invariant spatial coherence matrix multiplied by a time-varying power spectral density (PSD) level. The noise is also modelled as an additive Gaussian interference with known PSD matrix. The anechoic speech and the late reverberation are defined as the hidden data. Consequently, the estimation of the anechoic speech is obtained in the E-step using a multichannel Wiener filter (MCWF), while the PSD of the anechoic speech, the ETFs (actually, their corresponding normalized RETFs), the time-invariant spatial coherence matrix of the late reverberation and the PSD of the late reverberation are estimated in the M-step.

O. Schwartz and S. Gannot are with the Faculty of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel (e-mail: ofer.shwartz@live.biu.ac.il; sharon.gannot@biu.ac.il).

E. A. P. Habets is with the International Audio Laboratories Erlangen (a joint institution between 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.

The main contribution of our approach is the simultaneous estimation of the entire set of parameters (Coherence matrix, speech and reverberation PSDs and early relative transfer functions (RTFs)) using the ML criterion. This criterion is iteratively implemented by applying the EM procedure, to jointly estimate the desired speech and the parameters. The key component of the contribution is the observation that the perfect diffuse model is not accurate and should be substituted by the estimated coherence matrix of the late reverberation PSD. In previous works (e.g., [2], [8], [9]), the spatial coherence matrix was modelled as an ideal diffuse sound field, and was therefore assumed to be known. As this ideal diffuse sound field is not an accurate model of the reverberation phenomenon in many scenarios [10], we have incorporated the spatial coherence of the late reverberation into the set of estimated parameters, and show that by doing so, the dereverberation performance can indeed be improved.

Within the model, there are two gain ambiguity problems: 1) between the anechoic speech and the ETFs, and 2) between the time-invariant spatial coherence matrix and the time-varying PSD of the late reverberation. Methods to circumvent these problems are presented by normalizing the ETFs and the spatial coherence matrix. Due to the normalization of the ETFs (to form the RETFs), only a filtered version of the early speech components is estimated, rather than the anechoic speech.

The transition time between the early speech component and the late speech component is not well-defined within the framework of the EM algorithm. Accordingly, it is important to avoid a power exchange between these two components. A possible solution to this problem is proposed, based on the estimation of the late reverberation PSD with the maximum a posteriori (MAP)-EM [11] framework, to guarantee that the estimate would not differ significantly from its prior value. An alternative solution is to bound the estimator of the late reverberation level from above.

This remainder of this paper is organized as follows. In Section II, various solutions to the dereverberation task are presented, as a background to our approach. In Section III, we formulate the joint dereverberation and noise reduction problem. In Section IV, the EM solution for our statistical model is derived. Section IV-B is dedicated to the EM solution for the general noisy case, and Section IV-C is dedicated to the EM solution for the high signal-to-noise ratio (SNR) case. In Section IV-D, solutions for the gain ambiguity problems are presented and recommendations for initializations are presented in Section IV-E. In Section V, the performance of the proposed algorithm is evaluated. Section VI is dedicated to concluding remarks.

II. Background

Existing dereverberation methods can be broadly divided into two categories, namely *reverberation cancelation* and *reverberation suppression* [12]. The algorithm derived in this paper share attributes of both categories, since the RETFs are blindly estimated, as in the reverberation cancellation category, while the late reverberation is suppressed by a postfilter, as in the reverberation suppression category.

Reverberation cancelation algorithms are often based on a convolutive reverberation model. Reverberation cancelation can be achieved by directly inverting the acoustic system or by first identifying and then equalizing the acoustic system. Since clean speech is unobservable, these algorithms need to blindly estimate the acoustic system or its inverse directly. These acoustic systems may be very long in relation to the analysis window and therefore difficult to estimate.

In [13], [14], multichannel linear prediction techniques are used to blindly equalize the RIR without the need to first identify the RIRs. In [15], a dual-channel reconstruction method was presented, based on cepstrum techniques. A single-channel dereverberation method was presented in [16], based on the harmonic structure of the anechoic speech signal. The direct sound was approximated by extracting its harmonic parameters from the reverberant signal, and then the RIR was estimated by a division in the frequency domain. In [17], a two-stage multichannel dereverberation method was proposed. In the first stage, the RIRs were extracted from the null subspace of the data matrix. In the second stage, these estimates were used to equalize the microphone signals using the classical multichannel inverse theorem (MINT) method [18]. More recently, researchers proposed to apply channel shortening techniques to compute the inverse of the RIRs [19]–[22]. A technique that is based on explicit modelling of the vocal tract is the linear prediction (LP) residual enhancement method [23], [24]. The speech signal can be modelled as an LP residual excitation signal convolved with the vocal tract. The vocal tract is modeled by an all-pole filter and its time coefficients can be estimated by using common LP techniques. By inverse filtering of the speech signal, the LP residual can be restored. It turns out that it is easier to reduce the reverberation effects from the LP residual than from the regular reverberant signal. The LP residual enhancement can utilize single-channel [25] or multichannel [26] structures. In [27], the authors modelled the late reverberation as a convolution between past observations and a prediction filter. The weighted prediction error (WPE) algorithm was presented where the prediction filter coefficients are estimated in the ML sense. Then, using the estimated coefficients, the late reverberation is predicted and subtracted. When several channels are available, the late reverberation components from all the channels are predicted (using all the channels) and subtracted separately.

Reverberation suppression algorithms, which are often based on an additive reverberation model, circumvent the task of blindly identifying the acoustic system and employ instead spectral enhancement procedures, which can be implemented by using a simple reverberation model. Often, these spectral enhancement procedures result in residual noise, known as *musical noise*, which cannot be easily reduced.

Various techniques fall into the category of reverberation suppression. Polack, in [28], formulated the RIR as independent and identically distributed white Gaussian noise with an exponential decaying variance. This property was first utilized in [29] to show that the late reverberation PSD can be expressed as a delayed and attenuated version of the instantaneous reverberant PSD level. In [5], [29], a spectral subtraction algorithm was used to obtain an estimate of the

early speech component, using Polack's statistical model. This method was extended to the multi-microphone case in [30], by employing the single-channel spectral subtraction algorithm at the output of a DS beamformer. In this case, the estimation of the late reverberation PSD level was assisted by spatial averaging. In [9], [31], [32], both reverberation and noise were subtracted. In the authors' previous work [2], the MMSE estimation of the early speech component in a noisy environment was derived, as explained above. The late reverberation PSD level was estimated similarly to [29]. Although capable of significant reverberation suppression, these techniques may suffer from speech distortion, especially if the reverberation level is overestimated.

The EM algorithm has been used to perform dereverberation in the past. In [33], a single-channel algorithm for dereverberation and noise reduction was presented. The reverberant speech was modeled as an auto regressive (AR) process, while the anechoic speech was modeled as an all-pole model. In the E-step, the reverberant speech is estimated (without the noise component), while the M-step estimates the reverberation and speech parameters (the noise parameters assumed to be known). The anechoic speech is then estimated, externally to the EM iterations, by virtue of the MCWF. Since there is no closed-form solution for the maximization in the M-step, the expectation-conditional maximization (ECM) algorithm [34] was used. The ECM algorithm integrates the coordinate descent method within the M-step. In [35], an alternative multichannel EM-based dereverberation algorithm was presented. The RIRs were defined as stochastic processes and modeled as a first-order Markov chain, while the speech components were modeled as time-varying parameters. In the E-step, the RIRs were estimated using a virtue of the Kalman filter, and in the M-step, the speech parameters were estimated. In [36], a recursive multichannel EM-based algorithm was presented. The RIRs were defined as deterministic time-varying parameters and modeled using convolutive transfer functions (CTFs). The Kalman-EM procedure [37] was adopted: the Kalman filter was employed for estimating the anechoic speech in the E-step and the RIRs were recursively estimated in the M-step.

An EM framework, consisting of an MCWF, was used in [38], in order to employ source separation. Each source signal was modelled as a scalar process multiplied by a transfer function (TF), which yields a rank-1 PSD matrix for each source. The source signals were defined to be the hidden data and were separated in the E-step by the MCWF, while the TFs, the PSDs of the sources and the noise PSD matrix were estimated in the M-step. Note that TFs which are much longer than the window length of the STFT cannot be accurately estimated. In [39], a similar EM framework was used, with the reverberant sources modelled as a full-rank PSD matrix and a temporal gain. This modelling enabled the separation of reverberant sources as well. However, this algorithm does not aim at suppressing the reverberation, but rather at separating the reverberant sources. In [40], an EM-based approach for separating source signals, that also uses the MCWF, was proposed. In this method, the anechoic signals are modelled by the nonnegative matrix factorization (NMF) model. This approach was later extended to dynamic scenarios in [41].

Our proposed algorithm exhibits similarities to the source separation task of two sources, where in our case, the early speech component and the late reverberation component (associated with a single speaker) will be separated. Accordingly, the early speech component has a rank-1 PSD matrix and the late reverberation has a full-rank PSD matrix. Note, that in [38], [39], all sources were either modelled with full-rank PSD matrices or with rank-1 PSD matrices. It was stated in [40] that for rank-1 sources in the noiseless case, the EM procedure converges very slowly. It is therefore required to update the EM model for the noiseless case. In our paper, we propose a simplified model for the noiseless case, for which only the early speech component is defined as the hidden data, thus circumventing the slow convergence.

III. PROBLEM FORMULATION

In the following section, the multichannel dereverberation and noise reduction problem is formulated. The observations consist of reverberant speech in a noisy environment. The reverberant speech is split into two components, namely early speech and late reverberation, such that the observations can be modelled in the STFT domain as:

$$Y_i(m,k) = X_{\mathrm{e},i}(m,k) + R_i(m,k) + V_i(m,k), \tag{1}$$

where $Y_i(m,k)$ denotes the ith microphone observation at time index m and frequency index k, $X_{\mathrm{e},i}(m,k)$ denotes the early speech component, $R_i(m,k)$ denotes the late reverberation and $V_i(m,k)$ denotes the ambient noise. Here $X_{\mathrm{e},i}(m,k)$ is modelled as a multiplication between the anechoic speech and the ETF, i.e.,

$$X_{\mathrm{e},i}(m,k) = G_{\mathrm{e},i}(k)\; S(m,k), \tag{2}$$

where $G_{\mathrm{e},i}(k)$ is the ETF and $S(m,k)$ is the anechoic speech.

Concatenating the N microphone signals in a vector form yields:

$$\mathbf{y}(m,k) = \mathbf{x}_{\mathrm{e}}(m,k) + \mathbf{r}(m,k) + \mathbf{v}(m,k) \tag{3}$$

$$\mathbf{x}_{\mathrm{e}}(m,k) = \mathbf{g}_{\mathrm{e}}(k)S(m,k), \tag{4}$$

where

$$\begin{aligned}
\mathbf{y}(m,k) &= \left[\begin{array}{cccc} Y_1(m,k) & Y_2(m,k) & \dots & Y_N(m,k) \end{array}\right]^T \\
\mathbf{x}_{\mathrm{e}}(m,k) &= \left[\begin{array}{cccc} X_{\mathrm{e},1}(m,k) & X_{\mathrm{e},2}(m,k) & \dots & X_{\mathrm{e},N}(m,k) \end{array}\right]^T \\
\mathbf{r}(m,k) &= \left[\begin{array}{cccc} R_1(m,k) & R_2(m,k) & \dots & R_N(m,k) \end{array}\right]^T \\
\mathbf{v}(m,k) &= \left[\begin{array}{cccc} V_1(m,k) & V_2(m,k) & \dots & V_N(m,k) \end{array}\right]^T \\
\mathbf{g}_{\mathrm{e}}(k) &= \left[\begin{array}{cccc} G_{\mathrm{e},1}(k) & G_{\mathrm{e},2}(k) & \dots & G_{\mathrm{e},N}(k) \end{array}\right]^T.
\end{aligned}$$

To circumvent the gain ambiguity problem, the desired signal can be recast as

$$\mathbf{x}_{\mathrm{e}}(m,k) = \bar{\mathbf{g}}_{\mathrm{e}}(k)S_{\mathrm{F}}(m,k), \tag{5}$$

where

$$S_{\mathrm{F}}(m,k) = \mathbf{q}^{\mathrm{H}}(k)\mathbf{g}_{\mathrm{e}}(k)S(m,k), \tag{6}$$

$$\bar{\mathbf{g}}_{\mathrm{e}}(k) = \frac{\mathbf{g}_{\mathrm{e}}(k)}{\mathbf{q}^{\mathrm{H}}(k)\mathbf{g}_{\mathrm{e}}(k)}, \tag{7}$$

and $\mathbf{q}(k)$ denotes a spatial filter. An example of such a spatial filter is the DS beamformer

$$\mathbf{h}_{\text{ds}}(k) = \frac{1}{N}\left[\begin{array}{cccc} 1 & \exp\left(j\frac{2\pi k}{K}\frac{\tau_2}{T_s}\right) & \dots & \exp\left(j\frac{2\pi k}{K}\frac{\tau_N}{T_s}\right)\end{array}\right]^T, \tag{8}$$

where τ_i is the time difference of arrival (TDOA) between the ith microphone and 1st microphone, T_s is the sampling time and K is the number of frequency bins. Hence, $S_{\text{F}}(m,k)$ is a spatially filtered version of the early speech components which enhances the direct arrival while incoherently adding the early reflections [2]. Alternatively, $\mathbf{q}(k)$ can be defined as $\mathbf{q}(k) = \left[\begin{array}{cccc} 1 & 0 & \dots & 0\end{array}\right]^T$. In this case, the output of the algorithm provides an estimate of the early speech component at the first microphone.

The late reverberation and the noise components are assumed to be undesired and have to be suppressed. The late reverberation and the noise vectors are assumed to be uncorrelated and may be modelled as zero-mean multi-dimensional Gaussian probabilities. In [42], the author shows that estimators based on super-Gaussian densities deliver an improved SNR. However, to simplify the derivations, we prefer to use the Gaussian density instead. The PSD matrix of the noise is assumed to be time-invariant and known. The late reverberation PSD matrix is time-variant, since the late reverberation arises from the speaker. However, the *spatial characteristic* of the late reverberation may be assumed to be time-invariant, as long as the microphone array geometry is fixed. Therefore, it is reasonable to model the PSD matrix of the late reverberation as a time-invariant spatial coherence matrix with time-variant PSD. Finally, the late reverberation probability density function (p.d.f.) is modelled as

$$f\left(\mathbf{r}(m,k);\phi_R(m,k),\mathbf{\Gamma}(k)\right) = \mathcal{N}^C(\mathbf{r}(m,k);\mathbf{0},\phi_R(m,k)\,\mathbf{\Gamma}(k)), \tag{9}$$

where

$$\mathcal{N}^C(\mathbf{x};\mathbf{0},\mathbf{\Phi}) = \frac{1}{\pi^N|\mathbf{\Phi}|}\exp\left(-\mathbf{x}^{\text{H}}\,\mathbf{\Phi}^{-1}\mathbf{x}\right), \tag{10}$$

$\mathbf{x}$ denotes Gaussian vector, $\mathbf{\Phi}$ is a PSD matrix and $|\cdot|$ denotes the matrix determinant operation. The time-invariant spatial coherence matrix $\mathbf{\Gamma}(k)$ describes the spatial characteristics of the late reverberant field, and $\phi_R(m,k)$ represents the time-variant PSD of the late reverberation. The spatial coherence matrix $\mathbf{\Gamma}(k)$ is normalized, such that $\frac{1}{N}\text{Tr}[\mathbf{\Gamma}(k)] = 1$, where $\text{Tr}[\,]$ denotes the trace operation[1]. As a result, the PSD of the late reverberation is given by $\phi_R(m,k) = \frac{1}{N}\sum_{i=1}^{N} E\left\{|R_i(m,k)|^2\right\}$, which represents the average late reverberation PSD across the different microphones. The filtered speech $S_{\text{F}}(m,k)$ can also be modelled as a zero-mean Gaussian process with variance $\phi_{S_{\text{F}}}(m,k) = E\{|S_{\text{F}}(m,k)|^2\}$[2]:

$$f\left(S_{\text{F}}(m,k);\phi_{S_{\text{F}}}(m,k)\right) = \mathcal{N}^C(S_{\text{F}}(m,k);0,\phi_{S_{\text{F}}}(m,k)). \tag{11}$$

[1]Strictly speaking the matrix $\mathbf{\Gamma}(k)$ can be defined as a coherence matrix only if the mean of its diagonal elements is equal to one.

[2]The variance of the Gaussian process is the PSD of the stochastic process, that can be obtained by the Fourier transform of the auto-correlation sequence, according to the Wiener-Khinchin theorem.

The ambient noise is modelled as a zero-mean Gaussian vector with known PSD matrix $\mathbf{\Phi}_{\mathbf{v}}(k)$.

Define $\boldsymbol{\phi}_R(k) = [\phi_R(1,k),\dots,\phi_R(M,k)]$ and $\boldsymbol{\phi}_{S_{\text{F}}}(k) = [\phi_{S_{\text{F}}}(1,k),\dots,\phi_{S_{\text{F}}}(M,k)]$, where M is the total number of observed frames, i.e., $m = 1,\dots,M$. The entire set of parameters of the problem is therefore given by:

$$\boldsymbol{\theta}(k) = \left\{\boldsymbol{\phi}_{S_{\text{F}}}(k),\bar{\mathbf{g}}_e(k),\boldsymbol{\phi}_R(k),\mathbf{\Gamma}(k)\right\}. \tag{12}$$

The number of the PSD parameters increases with the number of processed frames. The RETFs and the coherence are assumed to be time-invariant. Note, that as opposed to [2], where $\mathbf{\Gamma}(k)$ obeys a perfectly diffuse sound field, here it is an unknown parameter. This key component of the current contribution is based on our observation that the perfect diffuse model is not accurate. This observation has also been made by other researchers and has been reported in the literature (e.g., in [10]).

The early speech component, the late reverberation and the ambient noise are assumed to be mutually uncorrelated. Therefore, the observed signal vector is also a zero-mean Gaussian vector, and the PSD matrix of the observed signals is equal to the sum of the individual PSD matrices of the early speech component, late reverberation and ambient noise. Defining the set of all M available measurement vectors by:

$$\bar{\mathbf{y}}(k) = \left[\begin{array}{ccc}\mathbf{y}^T(1,k) & \dots & \mathbf{y}^T(M,k)\end{array}\right]^T, \tag{13}$$

their p.d.f. is given by:

$$f(\bar{\mathbf{y}}(k);\boldsymbol{\theta}(k)) = \prod_{m=1}^{M}\mathcal{N}^C\left(\mathbf{y}(m,k);\mathbf{0},\mathbf{\Phi}_{\mathbf{y}}(m,k)\right), \tag{14}$$

where it is further assumed that all measurements are independent. The PSD matrix of the observations is given by:

$$\mathbf{\Phi}_{\mathbf{y}}(m,k) = \phi_{S_{\text{F}}}(m,k)\bar{\mathbf{g}}_e(k)\bar{\mathbf{g}}_e^{\text{H}}(k) + \phi_R(m,k)\,\mathbf{\Gamma}(k) + \mathbf{\Phi}_{\mathbf{v}}(k). \tag{15}$$

The PSD matrix of the early speech component is obtained by $E\left\{\mathbf{g}_{\text{e}}(k)S_{\text{F}}(m,k)\left(\mathbf{g}_{\text{e}}(k)S_{\text{F}}(m,k)\right)^H\right\} = \phi_{S_{\text{F}}}(m,k)\bar{\mathbf{g}}_e(k)\bar{\mathbf{g}}_e^{\text{H}}(k)$.

Our goal now is to maximize the p.d.f. of the measurements with relation to the parameters, namely to apply the ML criterion yielding $\boldsymbol{\theta}(k)$, i.e.,

$$\boldsymbol{\theta}_{\text{ML}}(k) = \underset{\boldsymbol{\theta}}{\text{argmax}}\, f(\bar{\mathbf{y}}(k);\boldsymbol{\theta}(k)). \tag{16}$$

The maximization operation may be a cumbersome task. To simplify the derivations, the EM formulation is adopted in the following section.

IV. Expectation-Maximization Algorithm

In order to implement the EM algorithm, the *hidden data* should be defined. We are proposing to define $S_{\text{F}}(m,k)$ and the late reverberation $\mathbf{r}(m,k)$ as the hidden data. The expectation-step evaluates the auxiliary function (i.e., the expectation of the joint log-likelihood of the observations and the hidden data, conditional upon the observations and the

current estimate of the parameters) while the maximization-step maximizes the auxiliary function with relation to the set of parameters. This procedure converges into a local maximum of the likelihood function of the observation [11]. In the following, the frequency index k is omitted for brevity whenever possible.

A. Definition of the Auxiliary Function

Concatenate the two components of the hidden data:

$$\mathbf{d}(m) \triangleq \left[\begin{array}{cc} S_{\mathrm{F}}(m) & \mathbf{r}^T(m) \end{array} \right]^T. \tag{17}$$

Using this definition, the measurement equation (4) can be rewritten as:

$$\mathbf{y}(m) = \mathbf{H}\,\mathbf{d}(m) + \mathbf{v}(m), \tag{18}$$

where

$$\mathbf{H} \triangleq \left[\begin{array}{cc} \bar{\mathbf{g}}_{\mathrm{e}} & \mathbf{I}_{N\times N} \end{array} \right] \tag{19}$$

and $\mathbf{I}_{N\times N}$ is the identity matrix. Now, by concatenating the hidden data vectors of time frames $1,\ldots,M$, i.e.,

$$\bar{\mathbf{d}} = \left[\begin{array}{ccc} \mathbf{d}^T(1) & \ldots & \mathbf{d}^T(M) \end{array} \right]^T, \tag{20}$$

the auxiliary function, i.e., the conditional expectation of the log-likelihood function, can be deduced as:

$$Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(\ell)}\right) = E\left\{\log f(\bar{\mathbf{y}},\bar{\mathbf{d}};\boldsymbol{\theta})|\bar{\mathbf{y}};\boldsymbol{\theta}^{(\ell)}\right\}, \tag{21}$$

where $\boldsymbol{\theta}^{(\ell)}$ is the parameter-set estimate at iteration ℓ. The joint probability of the observed data and the hidden data is Gaussian and given by Bayes rule, where the p.d.f. of $\bar{\mathbf{d}}$ is defined using (9) and (11):

$$\begin{aligned} f(\bar{\mathbf{y}},\bar{\mathbf{d}};\boldsymbol{\theta}) &= f(\bar{\mathbf{y}}|\bar{\mathbf{d}};\boldsymbol{\theta})\; f(\bar{\mathbf{d}};\boldsymbol{\theta}) = \\ &\prod_{m=1}^{M} \mathcal{N}^C(\mathbf{y}(m) - \mathbf{H}\mathbf{d}(m);\mathbf{0},\boldsymbol{\Phi}_{\mathbf{v}}) \\ &\times \mathcal{N}^C(\mathbf{r}(m);\mathbf{0},\phi_R(m)\;\boldsymbol{\Gamma}) \times \mathcal{N}^C(S_{\mathrm{F}}(m);0,\phi_{S_{\mathrm{F}}}(m)). \end{aligned} \tag{22}$$

According to (15), the variance of the observations consists of a summation of three main components: the early speech, the late reverberation and the noise. The variances of the early speech and the late reverberation are both inferred by the EM algorithm. According to our model, the boundary between the early component and the late component is not well-defined. Accordingly, it is important to avoid leakage of energy between these two component. A possible solution for this indeterminacy is to replace the ML estimation procedure of $\phi_R(m)$ with the MAP-EM [11] procedure, and hence to guarantee that the estimate will remain close to the prior mean of $\phi_R(m)$. In the Bayesian inference literature, it is customary to model the variance matrix of a Gaussian vector using the inverse Wishart probability, the so-called *conjugate prior* probability of the Gaussian variance (e.g., see [43]). Since $\phi_R(m)$ is a 1-dimensional variable, a degenerated version of the inverse Wishart probability, the inverse Gamma probability, can be used:

$$\begin{aligned} &\mathcal{IG}(\phi_R(m);\psi_R(m),\nu) = \\ &\qquad \frac{\psi_R^{\nu}(m)}{G(\nu)}\left(\phi_R(m)\right)^{-(\nu+1)} \exp\left(-\frac{\psi_R(m)}{\phi_R(m)}\right), \end{aligned} \tag{23}$$

where $\psi_R(m)$ denotes the scale parameter, ν denotes the shape parameter and $G(\nu)$ denotes the Gamma function. The maximum value of the inverse Gamma probability is $\phi_R(m) = \frac{\psi_R(m)}{\nu+1}$. Under the MAP framework, the joint p.d.f. of the observed data, the hidden data and $\phi_R(m)$ can be expressed as:

$$f(\bar{\mathbf{y}},\bar{\mathbf{d}},\boldsymbol{\phi}_R;\tilde{\boldsymbol{\theta}}) = f(\bar{\mathbf{y}},\bar{\mathbf{d}}|\boldsymbol{\phi}_R;\tilde{\boldsymbol{\theta}}) \prod_{m=1}^{M} \mathcal{IG}(\phi_R(m);\psi_R(m),\nu), \tag{24}$$

where $\tilde{\boldsymbol{\theta}}$ is the parameter set excluding $\boldsymbol{\phi}_R$. In practice, a weighted version of the MAP criterion is commonly used, namely:

$$\begin{aligned} &f(\bar{\mathbf{y}},\bar{\mathbf{d}},\boldsymbol{\phi}_R;\tilde{\boldsymbol{\theta}}) = \\ &f^{(1-\gamma)}(\bar{\mathbf{y}},\bar{\mathbf{d}}|\boldsymbol{\phi}_R;\tilde{\boldsymbol{\theta}}) \left(\prod_{m=1}^{M} \mathcal{IG}(\phi_R(m);\psi_R(m),\nu)\right)^{\gamma}, \end{aligned} \tag{25}$$

where γ $(0 \le \gamma < 1)$ is a weighting factor.

The augmented auxiliary function can now be defined as a weighted combination of the ML auxiliary function and the prior p.d.f.:

$$\begin{aligned} Q_{\mathrm{MAP}}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(\ell)}\right) &= (1-\gamma)Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(\ell)}\right) + \\ &\gamma \sum_{m=1}^{M} \log \mathcal{IG}(\phi_R(m)|\psi_R(m),\nu). \end{aligned} \tag{26}$$

In the experimental study below, we will examine the applicability of the MAP estimation (and other potential solutions) to the indeterminacy problem at hand.

B. Derivation of the E-step and M-step

For implementing the E-step, it is sufficient to estimate the following:

1) $\widehat{\mathbf{d}}(m)$
2) $\widehat{\boldsymbol{\Psi}}_{\mathbf{d}}(m) \triangleq \widehat{\mathbf{d}(m)\;\mathbf{d}^{\mathrm{H}}(m)}$,

where $\widehat{\mathbf{d}}(m) \triangleq E\left\{\mathbf{d}(m)|\mathbf{y}(m);\boldsymbol{\theta}^{(\ell)}\right\}$ is the expected first-order statistic of the hidden-data given the measurements, and $\widehat{\boldsymbol{\Psi}}_{\mathbf{d}} \triangleq E\left\{\mathbf{d}\mathbf{d}^H|\mathbf{y}(m);\boldsymbol{\theta}^{(\ell)}\right\}$ is the expected second-order statistic of the hidden-data given the measurements. Note, that the hidden data is an independent stochastic process (see (22)). Hence, $\widehat{\mathbf{d}}(m)$ and $\widehat{\boldsymbol{\Psi}}_{\mathbf{d}}(m)$ only depend on the measurement vector at frame m, namely $\mathbf{y}(m)$, and all other time frames can be excluded from the estimation procedure. This is in line with standard beamforming techniques.

Note, that the required sufficient statistics of $\mathbf{r}(m)$ and $S_{\mathrm{F}}(m)$ can be deduced from the above terms, i.e.,

$$\widehat{\mathbf{r}}(m) = \widehat{\mathbf{d}}_{\{2:N+1\}}(m), \tag{27a}$$

$$\widehat{S}_{\mathrm{F}}(m) = \widehat{\mathbf{d}}_{\{1\}}(m), \tag{27b}$$

$$\widehat{\mathbf{r}(m)\mathbf{r}^{\mathrm{H}}(m)} = \widehat{\boldsymbol{\Psi}}_{\mathbf{d},\{2:N+1,2:N+1\}}(m), \tag{27c}$$

$$\widehat{|S_{\mathrm{F}}(m)|^2} = \widehat{\boldsymbol{\Psi}}_{\mathbf{d},\{1,1\}}(m), \quad \text{and} \tag{27d}$$

$$\widehat{S_{\mathrm{F}}^*(m)\;\mathbf{r}(m)} = \widehat{\boldsymbol{\Psi}}_{\mathbf{d},\{2:N+1,1\}}(m). \tag{27e}$$

Since $\mathbf{y}(m)$ and $\mathbf{d}(m)$ in (18) are Gaussian random vectors, $\widehat{\mathbf{d}}(m)$ can be estimated by the optimal linear estimator, which in our case, is given by the MCWF

$$\begin{aligned}\widehat{\mathbf{d}}(m) &= E\left\{\mathbf{d}(m)\mathbf{y}^{\mathrm{H}}(m)\right\} \times \left(E\left\{\mathbf{y}(m)\mathbf{y}^{\mathrm{H}}(m)\right\}\right)^{-1} \mathbf{y}(m) \\ &= \mathbf{\Phi}_{\mathbf{d}}^{(\ell)}(m)\left(\mathbf{H}^{(\ell)}\right)^{\mathrm{H}}\left(\mathbf{\Phi}_{\mathbf{y}}^{(\ell)}(m)\right)^{-1}\mathbf{y}(m) \quad (28)\end{aligned}$$

with

$$\mathbf{\Phi}_{\mathbf{d}}^{(\ell)}(m) = \begin{bmatrix} \phi_{S_{\mathrm{F}}}^{(\ell)}(m) & \mathbf{0}_{1\times N} \\ \mathbf{0}_{N\times 1} & \phi_R^{(\ell)}(m)\,\mathbf{\Gamma}^{(\ell)} \end{bmatrix}, \quad (29)$$

$\mathbf{0}_{M\times N}$ an all-zeros matrix of dimension $M\times N$ and $\mathbf{\Phi}_{\mathbf{y}}^{(\ell)}(m)$ being the PSD matrix of the observations. The latter can be calculated using $\mathbf{\Phi}_{\mathbf{y}}^{(\ell)}(m) = \mathbf{H}^{(\ell)}\mathbf{\Phi}_{\mathbf{d}}^{(\ell)}(m)\left(\mathbf{H}^{(\ell)}\right)^{\mathrm{H}} + \mathbf{\Phi}_{\mathbf{v}}$.

The matrix $\widehat{\mathbf{\Psi}}_{\mathbf{d}}(m)$ can be estimated using the following relation:

$$\widehat{\mathbf{\Psi}}_{\mathbf{d}}(m) = \widehat{\mathbf{d}}(m)\,\widehat{\mathbf{d}}^{\mathrm{H}}(m) + \mathrm{Cov}\left\{\mathbf{d}(m)|\mathbf{y}(m);\boldsymbol{\theta}^{(\ell)}\right\}, \quad (30)$$

where $\mathrm{Cov}\left\{\mathbf{d}(m)|\mathbf{y}(m);\boldsymbol{\theta}^{(\ell)}\right\}$ is the covariance matrix of $\mathbf{d}(m)$ given $\mathbf{y}(m)$. Since $\mathbf{d}(m)$ and $\mathbf{y}(m)$ are Gaussian random vectors, the conditional covariance matrix is generally given by [44]:

$$\mathrm{Cov}\left\{\mathbf{d}(m)|\mathbf{y}(m)\right\} = \mathbf{\Phi}_{\mathbf{d}}(m) - \mathbf{\Phi}_{\mathbf{dy}}(m)\mathbf{\Phi}_{\mathbf{y}}^{-1}(m)\mathbf{\Phi}_{\mathbf{yd}}(m), \quad (31)$$

where $\mathbf{\Phi}_{\mathbf{d}}(m), \mathbf{\Phi}_{\mathbf{dy}}(m), \mathbf{\Phi}_{\mathbf{yd}}(m)$ and $\mathbf{\Phi}_{\mathbf{y}}(m)$ are the corresponding PSD matrices. The conditional covariance at the ℓth iteration is therefore given by

$$\begin{aligned}\mathrm{Cov}\left\{\mathbf{d}(m)|\mathbf{y}(m);\boldsymbol{\theta}^{(\ell)}\right\} &= \mathbf{\Phi}_{\mathbf{d}}^{(\ell)}(m) - \\ &\mathbf{\Phi}_{\mathbf{d}}^{(\ell)}(m)\left(\mathbf{H}^{(\ell)}\right)^{\mathrm{H}}\left(\mathbf{\Phi}_{\mathbf{y}}^{(\ell)}(m)\right)^{-1}\mathbf{H}^{(\ell)}\mathbf{\Phi}_{\mathbf{d}}^{(\ell)}(m). \quad (32)\end{aligned}$$

Maximizing $Q_{\mathrm{MAP}}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(\ell)}\right)$ with relation to the problem parameters constitutes the M-step:

1. $$\phi_S^{(\ell+1)}(m) = \widehat{|S_{\mathrm{F}}(m)|^2} \quad (33)$$

2. $$\bar{\mathbf{g}}_e^{(\ell+1)} = \frac{\sum_{m=1}^{M}\widehat{S}_{\mathrm{F}}^{*}(m)\mathbf{y}(m) - \widehat{S_{\mathrm{F}}^{*}(m)\mathbf{r}(m)}}{\sum_m \widehat{|S_{\mathrm{F}}(m)|^2}} \quad (34)$$

3. $$\mathbf{\Gamma}^{(\ell+1)} = \frac{1}{M}\sum_{m=1}^{M}\left(\phi_R^{(\ell)}(m)\right)^{-1}\widehat{\mathbf{r}(m)\mathbf{r}^{\mathrm{H}}(m)} \quad (35)$$

4. $$\phi_R^{(\ell+1)}(m) = \frac{\gamma\frac{\psi_R(m)}{N} + (1-\gamma)\bar{\phi}_R^{(\ell+1)}(m)}{\gamma\frac{\nu+1}{N} + (1-\gamma)}, \quad (36)$$

where

$$\bar{\phi}_R^{(\ell+1)}(m) = \frac{1}{N}\mathrm{Tr}\left[\widehat{\mathbf{r}(m)\mathbf{r}^{\mathrm{H}}(m)}\left(\mathbf{\Gamma}^{(\ell+1)}\right)^{-1}\right]. \quad (37)$$

This algorithm is a generalized form of the EM algorithm (referred to as the ECM algorithm in [34]). Since the maximization operations of $\phi_R^{(\ell+1)}(m)$ and $\mathbf{\Gamma}^{(\ell+1)}(m)$ are interlaced, each iteration of the M-step does not maximize the likelihood but merely increases it.

As the algorithm uses the entire set of available measurements, i.e., $m = 1,\ldots,M$, for estimating $\bar{\mathbf{g}}_e^{(\ell+1)}$ and $\mathbf{\Gamma}^{(\ell+1)}$, it should be applied in *batch* mode. Recursive solutions of the problem at hand are beyond the scope of this contribution.

Note, that when $\gamma = 0$ the contribution of the prior p.d.f. is discarded, and when $\gamma = 1$ the PSD $\phi_R^{(\ell+1)}(m) = \frac{\psi_R(m)}{\nu+1}$. The scale parameter $\psi_R(m)$ may be set by utilizing the initial value of $\phi_R(m)$, i.e.,

$$\psi_R(m) = \phi_R^{(0)}(m)\,(\nu+1). \quad (38)$$

To simplify the denominator in the calculation of $\phi_R^{(\ell+1)}(m)$, ν was set to $N-1$. Consequently, (36) can be written as

$$\phi_R^{(\ell+1)}(m) = \gamma\phi_R^{(0)}(m) + (1-\gamma)\bar{\phi}_R^{(\ell+1)}(m). \quad (39)$$

C. High SNR Case

In cases where the noise PSD matrix tends to zero, the resulting update rule for the other parameters might converge slowly or not converge at all [40]. Therefore, it is worthwhile to define an alternative model when the SNR is high. When the SNR is very high, the signal model in (1) can be simplified to:

$$Y_i(m) = \bar{G}_{\mathrm{e},i}\,S_{\mathrm{F}}(m) + R_i(m). \quad (40)$$

We can now simplify the hidden data to consist only of the filtered speech $S_{\mathrm{F}}(m)$, i.e., without the reverberation term. Similar to the joint p.d.f. of the observed data and the hidden data given by (22), the joint p.d.f. of the observations and the anechoic speech is given by:

$$\begin{aligned}f(\mathbf{y}(m), S_{\mathrm{F}}(m);\boldsymbol{\theta}) &= \mathcal{N}^{C}(\mathbf{y}(m) - \bar{\mathbf{g}}_{\mathrm{e}}S_{\mathrm{F}}(m), 0, \phi_R(m)\,\mathbf{\Gamma}) \\ &\times \mathcal{N}^{C}(S_{\mathrm{F}}(m), 0, \phi_{S_{\mathrm{F}}}(m)). \quad (41)\end{aligned}$$

The E-step can be derived similarly to the noisy case:

$$\widehat{S}_{\mathrm{F}}(m) = \phi_{S_{\mathrm{F}}}^{(\ell)}(m)\left(\bar{\mathbf{g}}_{\mathrm{e}}^{(\ell)}\right)^{\mathrm{H}}\left(\mathbf{\Phi}_{\mathbf{y}}^{(\ell)}(m)\right)^{-1}\mathbf{y}(m) \quad (42)$$

and

$$\begin{aligned}\widehat{|S_{\mathrm{F}}(m)|^2} &= |\widehat{S_{\mathrm{F}}}(m)|^2 + \phi_{S_{\mathrm{F}}}^{(\ell)}(m) - \\ &\left(\phi_{S_{\mathrm{F}}}^{(\ell)}(m)\right)^2\left(\bar{\mathbf{g}}_{\mathrm{e}}^{(\ell)}\right)^{\mathrm{H}}\left(\mathbf{\Phi}_{\mathbf{y}}^{(\ell)}(m)\right)^{-1}\bar{\mathbf{g}}_{\mathrm{e}}^{(\ell)}, \quad (43)\end{aligned}$$

where $\mathbf{\Phi}_{\mathbf{y}}(m) = \bar{\mathbf{g}}_{\mathrm{e}}\bar{\mathbf{g}}_{\mathrm{e}}^{\mathrm{H}}\phi_{S_{\mathrm{F}}}(m) + \phi_R(m)\,\mathbf{\Gamma}$. For $\nu = N-1$, the M-step is obtained by:

1. $$\phi_{S_{\mathrm{F}}}^{(\ell+1)}(m) = \widehat{|S_{\mathrm{F}}(m)|^2} \quad (44)$$

2. $$\bar{\mathbf{g}}_e^{(\ell+1)} = \frac{\sum_{m=1}^{M}\left(\phi_R^{(\ell)}(m)\right)^{-1}\hat{S}_{\mathrm{F}}^{*}(m)\mathbf{y}(m)}{\sum_{m=1}^{M}\left(\phi_R^{(\ell)}(m)\right)^{-1}\widehat{|S_{\mathrm{F}}(m)|^2}} \quad (45)$$

3. $$\mathbf{\Gamma}^{(\ell+1)} = \frac{1}{M}\sum_{m=1}^{M}\left(\phi_R^{(\ell)}(m)\right)^{-1}\widehat{\mathbf{Z}}^{(\ell+1)}(m) \quad (46)$$

4. $$\phi_R^{(\ell+1)}(m) = \gamma\phi_R^{(0)}(m) + (1-\gamma)\bar{\phi}_R^{(\ell+1)}(m), \quad (47)$$

where

$$\begin{aligned}\widehat{\mathbf{Z}}^{(\ell+1)}(m) &\triangleq \widehat{(\mathbf{y}(m)-\bar{\mathbf{g}}_{\mathrm{e}}S_{\mathrm{F}}(m))(\mathbf{y}(m)-\bar{\mathbf{g}}_{\mathrm{e}}S_{\mathrm{F}})^{\mathrm{H}}(m)}\\ &= \mathbf{y}(m)\mathbf{y}^{\mathrm{H}}(m)-\bar{\mathbf{g}}_{\mathrm{e}}^{(\ell)}\widehat{S}_{\mathrm{F}}(m)\mathbf{y}^{\mathrm{H}}(m)\\ &\quad-\mathbf{y}(m)\widehat{S}_{\mathrm{F}}^{*}(m)\left(\bar{\mathbf{g}}_{\mathrm{e}}^{(\ell)}\right)^{\mathrm{H}}\\ &\quad+\widehat{|S_{\mathrm{F}}(m)|^2}\,\bar{\mathbf{g}}_{\mathrm{e}}^{(\ell)}\left(\bar{\mathbf{g}}_{\mathrm{e}}^{(\ell)}\right)^{\mathrm{H}},\end{aligned}\tag{48}$$

and

$$\bar{\phi}_R^{(\ell+1)}(m)=\frac{1}{N}\mathrm{Tr}\left[\widehat{\mathbf{Z}}^{(\ell+1)}(m)\,\left(\mathbf{\Gamma}^{(\ell+1)}\right)^{-1}\right].\tag{49}$$

It is evident that the algorithm for the noiseless case is also applied in batch mode.

D. Practical Considerations

In the following, several practical aspects are discussed. These aspects are relevant to both the low and high SNR cases.

1) Normalization: The RETF $\bar{\mathbf{g}}_e$ and filtered speech $\phi_{S_{\mathrm{F}}}(m)$ are both estimated by the algorithm, such that $\bar{\mathbf{g}}_e\bar{\mathbf{g}}_e^{\mathrm{H}}\phi_{S_{\mathrm{F}}}(m)$ represents the total PSD matrix of the early speech component. Therefore, each $a\bar{\mathbf{g}}_e$ and $\frac{1}{|a|^2}\ \phi_{S_{\mathrm{F}}}(m)$ satisfies the same likelihood score, where a is an arbitrary frequency-dependent gain. Similarly, each $\frac{1}{b}\phi_R(m)$ and $b\mathbf{\Gamma}$ satisfies the same likelihood, where b is an arbitrary frequency-dependent gain. Since we know that $\mathbf{q}^{\mathrm{H}}\bar{\mathbf{g}}_e=1$ (see (7)) and $\frac{1}{N}\mathrm{Tr}\left[\mathbf{\Gamma}\right]=1$, the following normalization operations may be employed to resolve these gain ambiguity issues:

$$\bar{\mathbf{g}}_e^{(\ell+1)}\leftarrow\frac{\bar{\mathbf{g}}_e^{(\ell+1)}}{\mathbf{q}^{\mathrm{H}}\bar{\mathbf{g}}_e^{(\ell+1)}}\tag{50}$$

$$\phi_{S_{\mathrm{F}}}^{(\ell+1)}(m)\leftarrow|\mathbf{q}^{\mathrm{H}}\bar{\mathbf{g}}_e^{(\ell+1)}|^2\phi_{S_{\mathrm{F}}}^{(\ell+1)}(m)\tag{51}$$

$$\mathbf{\Gamma}^{(\ell+1)}\leftarrow\frac{\mathbf{\Gamma}^{(\ell+1)}}{\frac{1}{N}\mathrm{Tr}\left[\mathbf{\Gamma}^{(\ell+1)}\right]}.\tag{52}$$

The PSD $\phi_R(m)$ is then automatically normalized, due to (37).

2) Avoiding speech distortion: To avoid speech distortion in estimating $S_{\mathrm{F}}(m)$, it is proposed to replace the regular E-step at the last (Lth) iteration by an alternative step. The optimal MCWF in (28) can be split into a multichannel MVDR beamformer and a subsequent single-channel Wiener filter as shown in [45], [46]. The output of the MVDR beamformer is computed using:

$$\hat{S}_{\mathrm{MVDR}}^{(L)}(m)=\left(\mathbf{w}_{\mathrm{MVDR}}^{(L)}(m)\right)^{\mathrm{H}}\mathbf{y}(m),\tag{53}$$

where

$$\mathbf{w}_{\mathrm{MVDR}}^{(L)}(m)=\frac{\left(\phi_R^{(L)}(m)\,\mathbf{\Gamma}^{(L)}+\mathbf{\Phi}_{\mathbf{v}}\right)^{-1}\bar{\mathbf{g}}_{\mathrm{e}}^{(L)}}{\left(\bar{\mathbf{g}}_{\mathrm{e}}^{(L)}\right)^{\mathrm{H}}\left(\phi_R^{(L)}(m)\,\mathbf{\Gamma}^{(L)}+\mathbf{\Phi}_{\mathbf{v}}\right)^{-1}\bar{\mathbf{g}}_{\mathrm{e}}^{(L)}}.\tag{54}$$

Then, only at the Lth iteration, the single-channel Wiener filter is substituted by the following single-channel postfilter (with $\xi^{(L)}(m)$ denoting the *a priori* signal-to-reverberation plus noise ratio (SRNR)):

$$H_{\mathrm{W}}^{(L)}(m)=\min\left\{\frac{\xi^{(L)}(m)}{\xi^{(L)}(m)+1},H_{\min}\right\},\tag{55}$$

i.e., a single-channel Wiener filter with a lower-bound constraint $H_{\min}$. Since the SRNR is unobservable, we recursively estimate it using the posterior SRNR, as proposed in [5]:

$$\begin{aligned}\xi^{(L)}(m)=\beta_r|H_{\mathrm{W}}^{(L)}(m-1)|^2\eta^{(L)}(m-1)+\\(1-\beta_r)\max\left\{\eta^{(L)}(m)-1,0\right\},\end{aligned}\tag{56}$$

where β_r is a weighting factor and $\eta^{(L)}(m)$ is the *a posteriori* SRNR at the MVDR output given by:

$$\eta^{(L)}(m)=\frac{\left|\hat{S}_{\mathrm{MVDR}}^{(L)}(m)\right|^2}{\tilde{\phi}_R^{(L)}(m)+\tilde{\phi}_V^{(L)}(m)}.\tag{57}$$

The residual reverberation $\tilde{\phi}_R^{(L)}(m)$ and the residual noise $\tilde{\phi}_V^{(L)}$ at the output of the MVDR stage are, respectively, given by

$$\tilde{\phi}_R^{(L)}(m)=\left(\mathbf{w}_{\mathrm{MVDR}}^{(L)}(m)\right)^{\mathrm{H}}\phi_R^{(L)}(m)\,\mathbf{\Gamma}^{(L)}\mathbf{w}_{\mathrm{MVDR}}^{(L)}(m)\tag{58}$$

and

$$\tilde{\phi}_V^{(L)}(m)=\left(\mathbf{w}_{\mathrm{MVDR}}^{(L)}(m)\right)^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}\mathbf{w}_{\mathrm{MVDR}}^{(L)}(m).\tag{59}$$

The output of the algorithm is finally given by:

$$\widehat{S}_{\mathrm{O}}(m)=H_{\mathrm{W}}^{(L)}(m)\,\hat{S}_{\mathrm{MVDR}}^{(L)}(m).\tag{60}$$

3) Using only relevant time frames: In practice, a poor estimate of $\mathbf{r}(m)$ may be obtained when the late reverberation power is much smaller than the power of the early speech component. Therefore, we propose to evaluate $\phi_R(m)$ and $\mathbf{\Gamma}$ in (36) and (35), respectively, based on the late-reverberation-to-early-speech ratio (RER) and the late-reverberation-to-noise ratio (RNR), and to select only segments for which both ratios are higher than a predefined threshold, i.e.,

$$\mathrm{RER}\triangleq\frac{\phi_R^{(\ell)}(m)}{\phi_S^{(\ell)}(m)}>\lambda_{RER},\tag{61}$$

$$\mathrm{RNR}\triangleq\frac{\phi_R^{(\ell)}(m)}{\frac{1}{N}\mathrm{Tr}\left[\mathbf{\Phi}_{\mathbf{v}}\right]}>\lambda_{RNR}.\tag{62}$$

Similarly, the estimation of $\phi_{S_{\mathrm{F}}}(m)$ and $\bar{\mathbf{g}}_e$ in (33) and (34), respectively, should be carried out only in segments where the early-speech-to-late-reverberation ratio (ERR) and the early-speech-to-noise ratio (ENR) are higher than a predefined threshold, i.e.,

$$\mathrm{ERR}\triangleq\frac{1}{\mathrm{RER}}=\frac{\phi_S^{(\ell)}(m)}{\phi_R^{(\ell)}(m)}>\lambda_{ERR},\tag{63}$$

$$\mathrm{ENR}\triangleq\frac{\mathrm{RNR}}{\mathrm{RER}}=\frac{\phi_S^{(\ell)}(m)}{\frac{1}{N}\mathrm{Tr}\left[\mathbf{\Phi}_{\mathbf{v}}\right]}>\lambda_{ENR}.\tag{64}$$

4) Upper-bound for $\phi_R^{(\ell)}(m)$: To avoid overestimation of $\phi_R^{(\ell)}(m)$ by the EM algorithm, we propose to bound from above its estimated value by $\frac{1}{N}\mathbf{y}^H(m)\mathbf{y}(m)$. The latter is an instantaneous estimate of the a posteriori PSD level of the observations, and can hence serve as an upper-bound of the late reverberation component. This heuristic step may alleviate the need for the EM-MAP procedure, as will be examined in the experimental study.

E. Initialization and Summary

It is well known that the performance of the EM algorithm depends on the initialization of the parameters. In the following, parameter initialization procedures are provided.

1) Late reverberation PSD: The late reverberation PSD level can be initialized by averaging the PSD level at each microphone, $\phi_{R_i}(m)$, which can be obtained using Polack's model [28] (c.f. [4], [5], [7], [12], [47]), after compensating for the noise level $\widehat{\phi}_{V_i}$ at each microphone:

$$\widehat{\phi}_{R_i}(m) \;=\; \exp(-2\alpha RJ) \times \left[\widehat{\phi}_{Y,i}(m-J) - \widehat{\phi}_{V_i}\right], \quad (65)$$

where $\alpha = \frac{3\log(10)}{T_{60}f_s}$, J is the time in frames (measured with respect to the arrival time of the direct sound) indicating the beginning of late reverberation, R is the number of samples between two subsequent STFT frames, T_{60} is the reverberation time, and f_s is the sampling frequency in Hz. The PSD of $Y_i(m)$ can be estimated recursively from the microphone signals, using:

$$\widehat{\phi}_{Y,i}(m) = \beta_y\widehat{\phi}_{Y,i}(m-1) + (1-\beta_y)|Y_i(m)|^2, \quad (66)$$

where β_y is a forgetting factor. An estimate of the late reverberation level is obtained by averaging the PSD estimates across all channels:

$$\phi_R^{(0)}(m) = \frac{1}{N}\sum_{i=1}^{N}\widehat{\phi}_{R,i}(m). \quad (67)$$

2) Relative early transfer functions: Initialization of the RETFs $\bar{\mathbf{g}}_e$ is a necessary step for constructing the beamformer. We propose the following procedure that will be applied to the microphone signals prior to the application of a beamforming step.

From (5), the (normalized) early speech components at the microphone signals are the output of filtering the (normalized) desired signal by the RETFs, namely $X_{\mathrm{e},i}(m,k) = \bar{G}_{\mathrm{e},i}(k)S_{\mathrm{F}}(m,k)$; $i = 1,\ldots,N$. Multiplying both sides by $S_{\mathrm{F}}^*(m)$ and taking the expectation yields:

$$\phi_{X_{\mathrm{e},i},S_{\mathrm{F}}}(m) = \bar{G}_{\mathrm{e},i}\phi_{S_{\mathrm{F}}}(m), \quad (68)$$

which can be used to formulate a least squares (LS) optimization criterion for the estimation of the RETF. Assuming that the RETF are slowly time-varying, and hence may be considered time-invariant during the M time frames, the LS estimate of $\bar{G}_{\mathrm{e},i}$ can be used to initialize the RETFs:

$$\bar{G}_{\mathrm{e},i}^{(0)} = \frac{\sum_{m'=1}^{M}\phi_{X_{\mathrm{e},i},S_{\mathrm{F}}}(m')\,\phi_{S_{\mathrm{F}}}(m')}{\sum_{m'=1}^{M}\phi_{S_{\mathrm{F}},S_{\mathrm{F}}}^2(m')}. \quad (69)$$

The auto- and cross-PSDs are, respectively, recursively estimated using:

$$\widehat{\phi}_{S_{\mathrm{F}}}(m) \;=\; \beta_{\mathrm{e}}\widehat{\phi}_{S_{\mathrm{F}}}(m-1) + (1-\beta_{\mathrm{e}})|\widehat{S}_{\mathrm{F}}(m)|^2 \quad (70)$$

and

$$\widehat{\phi}_{X_{\mathrm{e},i},S_{\mathrm{F}}}(m) = \beta_{\mathrm{e}}\widehat{\phi}_{X_{\mathrm{e},i},S_{\mathrm{F}}}(m-1) + (1-\beta_{\mathrm{e}})\widehat{X}_{\mathrm{e},i}(m)\widehat{S}_{\mathrm{F}}^*(m). \quad (71)$$

Note, that both $X_{\mathrm{e},i}(m,k)$ and $S_{\mathrm{F}}(m,k)$ are unavailable at this stage. We therefore propose the following procedure to obtain a preliminary estimate that can be utilized to estimate $\bar{\mathbf{g}}_{\mathrm{e}}^{(0)}$. The early speech component at the ith microphone $\widehat{X}_{\mathrm{e},i}(m)$ can be obtained by applying a single-channel dereverberation algorithm based on the Wiener filter as presented in [2], and the preliminary estimate of the desired signal can be obtained as (see (5)-(6)):

$$\widehat{S}_{\mathrm{F}}(m) = \mathbf{q}^{\mathrm{H}}\widehat{\mathbf{x}}_{\mathrm{e}}(m). \quad (72)$$

3) Spatial coherence matrix of the late reverberation: The normalized spatial coherence matrix of the late reverberation can be initialized as a spherically isotropic sound field [48], [49] plus diagonal loading:

$$\mathbf{\Gamma}^{(0)}(k) = \begin{bmatrix} \mathrm{sinc}\left(\frac{2\pi f_s k d_{1,1}}{Kc}\right) & \cdots & \mathrm{sinc}\left(\frac{2\pi f_s k d_{1,N}}{Kc}\right) \\ \vdots & \ddots & \vdots \\ \mathrm{sinc}\left(\frac{2\pi f_s k d_{N,1}}{Kc}\right) & \cdots & \mathrm{sinc}\left(\frac{2\pi f_s k d_{N,N}}{Kc}\right) \end{bmatrix} + \varepsilon\mathbf{I}, \quad (73)$$

where $\mathrm{sinc}(x) = \sin(x)/x$, K is the number of frequency bins, $d_{i,j}$ is the inter-distance between microphones i and j, c is the sound velocity, and ϵ is a small positive number.

4) PSD of the filtered early speech component: In order to initialize $\phi_{S_{\mathrm{F}}}(m)$, an initial estimate of $\hat{S}_{\mathrm{F}}(m)$, denoted as $\hat{S}_{\mathrm{INIT}}(m)$, may be used. Note, that the estimate in (70) is only a preliminary estimate, since it uses the received microphone signals but not any beamformer output. A better estimate of $\hat{S}_{\mathrm{INIT}}(m)$ is obtained by:

$$\hat{S}_{\mathrm{INIT}}(m) = H_{\mathrm{W}}^{(0)}(m)\;\hat{S}_{\mathrm{MVDR}}^{(0)}(m), \quad (74)$$

where $H_{\mathrm{W}}^{(0)}(m)$ was defined in (55) and $\hat{S}_{\mathrm{MVDR}}^{(0)}(m)$ is similar to (53) with L substituted by 0. Then $\phi_{S_{\mathrm{F}}}(m)$ can be estimated as follows:

$$\phi_{S_{\mathrm{F}}}^{(0)}(m) = \beta_s\;\phi_{S_{\mathrm{F}}}^{(0)}(m-1) + (1-\beta_s)\;|\hat{S}_{\mathrm{INIT}}(m)|^2, \quad (75)$$

with β_s a forgetting factor.

5) Noise PSD matrix: The noise PSD matrix $\mathbf{\Phi}_{\mathbf{v}}$ can be estimated during speech-absence segments by using an estimate of the speech presence probability (c.f. [50]–[53]). Estimating the noise PSD matrix is beyond the scope of this contribution.

The EM based dereverberation and noise reduction algorithm for the noisy case is summarized in Algorithm 1.

Algorithm 1: EM based dereverberation and noise reduction algorithm.

Initialize $\bar{\mathbf{g}}_e^{(0)}$ by (69), $\phi_S^{(0)}(m)$ by (75), $\mathbf{\Gamma}^{(0)}$ by (73) and $\phi_R^{(0)}(m)$ by (67).
for $\ell = 1, \ldots, L$ **do**
  E-step:
  if *SNR is low* **then**
    Calculate $\widehat{\mathbf{d}}(m)$ by (28) and $\widehat{\mathbf{\Psi}}_{\mathbf{d}}(m)$ by (30).
  else
    Calculate $\widehat{S}(m)$ by (42) and $\widehat{|S(m)|^2}$ by (43).
  end
  M-step:
  Calculate $\phi_S^{(\ell+1)}(m)$, $\bar{\mathbf{g}}_e^{(\ell+1)}$, $\phi_R^{(\ell+1)}(m)$ and $\mathbf{\Gamma}^{(\ell+1)}$ by (33)-(36) for low SNR or by (44) for high SNR.
  Normalize $\mathbf{\Gamma}^{(\ell+1)}$, $\bar{\mathbf{g}}_e^{(\ell+1)}$ and $\phi_S^{(\ell+1)}(m)$ by (50).
end
Calculate $\widehat{S}_{\mathrm{O}}(m)$ using (60).

V. Performance Evaluation

The performance of the proposed algorithm is evaluated in terms of two objective measures that are commonly used in the speech enhancement community, namely perceptual evaluation of speech quality (PESQ) [54] and log-spectral distance (LSD). The experiments consist of reverberant signals plus directional noise or diffuse noise with various SNR levels.

For comparison, we also evaluated the performance of the single-channel dereverberation algorithm, proposed in [5], and the performance of $\hat{S}_{\mathrm{INIT}}(m)$, as defined in (74). $\hat{S}_{\mathrm{INIT}}(m)$ is the MMSE estimate of the early speech component given the initial parameter-set. $\hat{S}_{\mathrm{INIT}}(m)$ can be considered as the output of the multichannel MMSE dereverberation and noise reduction algorithm proposed in [2], since they share an identical parameter-set. The only difference between the two lies in the implementation of the MVDR stage. Whereas in [2], it was implemented in a (non-orthogonal) GSC structure, as explained in Sec. I, here we use a direct MVDR implementation.

To demonstrate the effectiveness of the proposed algorithm, the results of a competing state-of-the-art algorithm [27] known as WPE[3] are also presented. The results are reported only for the reverberant and noiseless case, since the WPE algorithm was designed only for this case.

A. Setup

For all considered scenarios, a loudspeaker was positioned at a distance of 2 m in front of a non-uniform linear array with 4 microphones, such that no delay compensation was required. Thus, $\mathbf{q} = \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^T$. Anechoic speech signals were convolved by RIRs which were downloaded from an open-source RIRs database. Details about the database and RIR estimation method can be found in [55]. An illustration of the geometric setup is given in Fig. 1. Further details about the speaker-microphones constellation can be found in [55].

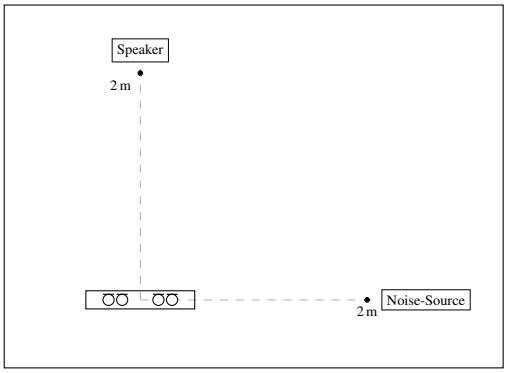


Fig. 1: Geometric setup (adopted from [55]).

The sampling frequency was 16 kHz and the frame length of the STFT was 32 ms with 8 ms between successive time frames (i.e., 75% overlap). We set the starting point of the late reverberation to 32 ms after the arrival of the direct-path, namely $J = 4$ in (65). The spatial coherence matrix $\mathbf{\Gamma}$ was estimated using frames where the RER and RNR are above 0 dB (i.e., $\lambda_{RER} = \lambda_{RNR} = 1$), and $\phi_R(m)$ was estimated using only frames with RER and RNR above -10 dB (i.e., $\lambda_{RER} = \lambda_{RNR} = 10^{-1}$). The RETF $\bar{\mathbf{g}}_e$ and the filtered speech PSD $\phi_{S_{\mathrm{F}}}(m)$ were estimated using frames where the ERR is above 0 dB and the ENR is above 0 dB (i.e., $\lambda_{ERR} = \lambda_{ENR} = 1$).

The designated threshold values, λ_{RER}, λ_{RNR}, λ_{ERR}, and λ_{ENR} are the nominal values used throughout the simulation study unless otherwise stated. As part of the study, we will examine a range of values for these thresholds and their influence on the performance of the algorithm.

The role of the MAP weighting factor γ in (26) will also be examined as part of the simulation study. Its nominal value was set to zero, i.e., the contribution of the a priori knowledge was discarded.

The nominal values of all other parameters of the algorithm is summarized in Table I. Note, that all parameter values are independent of the SNR level (apart from β_{e} and β_y) and of the noise field (either directional or diffuse noise).

In Sec. V-C we will explicitly examine the influence of changing the value of some of the important parameters (i.e., L, T_{60}, γ and the thresholds) to a range around their nominal values.

B. Performance Measures

The speech quality was evaluated by computing the PESQ score and LSD. Both the PESQ and the LSD were measured by comparing $\widehat{S}_{\mathrm{F}}(m)$ with $S_{\mathrm{F}}(m)$, where $S_{\mathrm{F}}(m)$ was obtained by filtering the anechoic speech $S(m)$ with the average ETF $\frac{1}{N}\sum_{i=1}^{N} G_{\mathrm{e},i}$. The first 32 ms (measured from the arrival time of the direct-path) of the RIRs were assumed to be the ETF.

[3]The results for the WPE method were obtained using the implementation available at http://www.kecl.ntt.co.jp/icl/signal/wpe/

TABLE I: The simulation setup.

Parameter	Setting
β_r	0.3
β_y, β_{e}	0.9 noisy case; 0.2 noiseless case
β_s	0.1
$H_{\min}$	-10 dB
L	2
Microphone-speaker distance	2 m
J	4
T_{60}	0.36 s, 0.61 s
Inter-microphone distance	$[0.03, 0.08, 0.03]$ m
f_s	16 kHz
Room dimensions	$6 \times 6 \times 2.4$ m
FFT size, overlap, analysis win.	1024, 75%, 32 ms
ϵ	0.1

The LSD between $\widehat{S}_{\mathrm{F}}(m,k)$ and $S_{\mathrm{F}}(m,k)$ is obtained using

$$\mathrm{LSD} = \frac{1}{M}\sum_{m}\sqrt{\frac{1}{K}\sum_{k}\left[20\log_{10}\left(\frac{\max\{|S_{\mathrm{F}}(m,k)|,\epsilon\}}{\max\{|\widehat{S}_{\mathrm{F}}(m,k)|,\hat{\epsilon}\}}\right)\right]^2}, \tag{76}$$

where

$$\epsilon = 10^{-A_{\mathrm{dB}}/10}\max_{m,k}\{|S_{\mathrm{F}}(m)|\}$$
$$\hat{\epsilon} = 10^{-A_{\mathrm{dB}}/10}\max_{m,k}\{|\widehat{S}_{\mathrm{F}}(m)|\}.$$

The parameter A_{dB} was set to the desired dynamic range, which was chosen, in our case, to be 60 dB. The PESQ scores and LSDs were computed by averaging the results obtained using 100 sentences, 50 by male speaker and 50 by female speaker, with each sentence being 3-5 s long and consisting of 2-8 words. As for the single-channel dereverberation algorithm proposed in [5], the algorithm was applied to $Y_1(m)$ and the output was compared with $S(m)G_{\mathrm{e},1}$.

C. Experimental Results

In the following sections, the results for the proposed algorithm and the competing algorithms are presented. The influence of the hyper-parameters on the performance of the proposed algorithm are then analyzed. Next, the results for the noisy case are presented (including directional noise and diffuse noise) and then the results for the noiseless case. The latter results also include a comparison of the proposed algorithm and the WPE [27]. Unless otherwise stated, the nominal parameters discussed in Sec. V-A are used.

1) Preliminary test of the EM algorithm: We start by examining the convergence of the EM algorithm along the iteration index. For that, microphone observations, obeying the exact model with a predefined set of parameters $\boldsymbol{\theta}$, were synthesized. A total of $N = 4$ microphones with $M = 10000$ samples and $K = 1$ frequency bins were synthesised and $L = 1000$ iterations were carried out. The log-likelihood of the observations as a function of the iteration index is depicted in Fig. 2. It is evident that the log-likelihood indeed increases monotonically with the number of iterations. As for the complexity of the algorithm, only two iterations were carried out in each time-frequency bin. Each iteration consists of an inversion of a $N \times N$ matrix $\boldsymbol{\Phi}_{\mathbf{y}}^{(\ell)}(m)$ in the E-step and the inversion of a $N \times N$ matrix $\boldsymbol{\Gamma}^{(\ell)}(k)$ in the M-step.

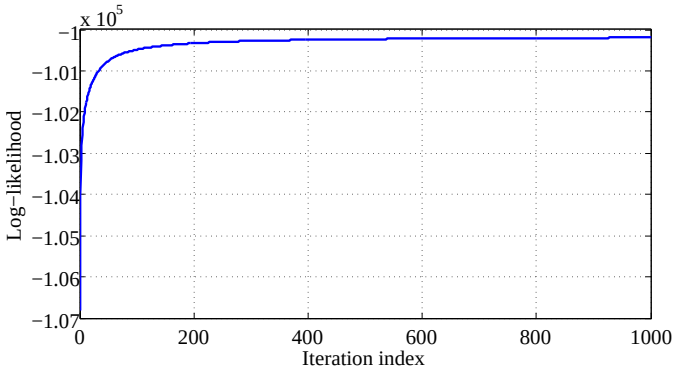


Fig. 2: Log-likelihood as a function of the iteration index.

2) Influence of the hyper-parameters: Secondly, we examine the influence of the hyper-parameters controlling the algorithm. A representative scenario with 15 dB SNR and reverberation time of 0.61 s was examined. LSD results and PESQ scores were calculated for all 100 signals and their mean value is depicted in the following tables. Specifically, the following parameters are examined: 1) L, the number of iterations; 2) T_{60}, the value of the reverberation given to the algorithm; 3) λ_{RER}, λ_{RNR}, λ_{ENR} and λ_{ERR}, the parameters that determine which segments should be taken into account; 4) γ, the weighting factor of the EM-MAP extension. Apart from the examined parameter, all other parameters are fixed to their nominal value.

First, the iteration number L is examined. In Table II, the results for $\widehat{S}_{\mathrm{O}}(m)$ are presented for $L = 0, \ldots, 7$. It can be seen that two iterations are sufficient to obtain satisfactory results. Although the results do not show monotonic improvement with relation to the iteration index, this non-monotonic behaviour is not severe.

The value of T_{60} is used by the algorithm for initializing the PSD of the late reverberation in (65). In this experiment, the true value of T_{60} is 0.61 s. We examine the sensitivity of the algorithm to the T_{60} parameter by varying it in the range $0.1, 0.2, \ldots, 1.5$ s. In Table III, the results for $\widehat{S}_{\mathrm{INIT}}(m)$ and $\widehat{S}_{\mathrm{O}}(m)$ are presented. As expected, the best result was obtained for $T_{60} = 0.7$ s. Nevertheless, the algorithm is only marginally sensitive to the value of T_{60}.

We turn now to the examination of the influence of the threshold levels on the performance of the algorithm. As explained in Sec. IV-D3, λ_{RNR} and λ_{RER} control the number of frames used for estimating $\boldsymbol{\Gamma}$ and $\phi_R(m)$. We first allow these thresholds to control only the number of frames used for estimating $\boldsymbol{\Gamma}$ (and hence its accuracy), while keeping the nominal values of the thresholds in estimating $\phi_R(m)$. Table IV depicts the performance of the algorithm for various values of $\lambda_{RNR} = \lambda_{RER}$. It is evident from the table that the proper setting of λ_{RER} and λ_{RNR} is critical and that the nominal values $\lambda_{RER} = \lambda_{RNR} = 1$ yield the best performance.

In Table V, we use λ_{RNR} and λ_{RER} to control the number of frames used for estimating $\phi_R(m)$, while keeping the threshold values used for estimating $\boldsymbol{\Gamma}$ at their nominal value.

We see only weak dependency on these thresholds with best results obtained for the nominal values $\lambda_{RER} = \lambda_{RNR} = 10^{-1}$.

To complete this examination, we test the influence of the values of λ_{ERR} and λ_{ENR} on the performance of the algorithm. These thresholds determine the number of frames used for the estimation of $\phi_S(m)$ and $\mathbf{g}_e$. It is evident from Table VI, that choosing $\lambda_{ERR} = \lambda_{ENR} = 10$ achieves the best results and that the results for the nominal values $\lambda_{ERR} = \lambda_{ENR} = 1$ are comparable. Finally, the contribution of the EM-MAP extension (versus the standard EM) is examined. Three representative values of γ were chosen: 0, 0.5 and 1. Note, that setting $\gamma = 0$ discards the contribution of the priors p.d.f. and setting $\gamma = 1$ fixes $\phi_R(m)$ to its prior estimation and discards the contribution of the EM iterations. In Table VII, the results for the proposed EM algorithm are presented for the various values of γ. As an alternative, we have also examined the setting $\gamma = 0$, together with bounding $\phi_R(m)$ from above to $\frac{1}{N}\mathbf{y}^H(m)\mathbf{y}(m)$. Since the performance measures only exhibit marginal sensitivity to the various options, we have decided to choose the simpler upper-bound option for the remaining experiments to avoid unreasonable reverberation level estimates. Note, that in this option, the EM-MAP extension is inactive.

3) Results for the noisy case: For the following experiment, RIRs recorded in a room with a reverberation time of approximately $T_{60} = 0.36$ s and $T_{60} = 0.61$ s were used. Directional noise was added to the speech signals with various SNR levels. The noise source was placed at 90 degrees relative to the microphone array at a distance of 2 m. The directional noise was generated by convolving a noise signal with the corresponding RIR. All the parameters of the algorithm were tuned to their nominal value.

Another set of experiments was carried on in the presence of diffuse noise. Rather than adding the directional noise, an artificial diffuse noise[4] was added to the speech signals with various SNR levels.

White noise was added in the two noise-field cases to simulate sensor noise. The directional noise-to sensor noise ratio and diffuse noise-to-sensor noise ratio were set to 20 dB. The PSD matrix $\mathbf{\Phi_v}$, which is non-diagonal, was estimated using periods during which the desired speech source was inactive.

In Tables VIII and IX, the PESQ and LSD scores are presented for several SNR levels for the directional noise case and the diffuse noise case, respectively. The advantage of using the EM algorithm is demonstrated for all SNR levels. We have added an additional row to the tables, depicting the results obtained by employing the MCWF with true parameters. We refer to the proposed algorithm with the true parameters as the *oracle* algorithm. The true late reverberation coherence was obtained by computing the normalized coherence of the synthesized late reverberant signals, which was computed by convolving the speech signal with the tails of the RIRs. The spatially averaged PSD of the late reverberation was obtained by computing the level of the synthesized late reverberant signals. The ETFs were computed by applying the LS technique to the early speech components. These results can be considered as the best achievable results of the algorithm. It can be observed that the results obtained by our algorithm are quite close to the oracle algorithm. In our opinion, the results of the oracle algorithm are limited due to the following reasons: 1) in every MMSE estimation there is an estimation error, even while using the oracle parameters; 2) the late component is modelled in our approach with a fixed normalized PSD matrix and a time-varying level. It seems that this model does not accurately reflect real recordings.

To demonstrate the effectiveness of the algorithm, an example of the cross PSD between the first and the second microphones along the frequency index $\Gamma_{1,2}^{(L)}(k)$ is depicted in Fig. 3. In addition, the coherence of an ideal (spherical) diffuse sound field, and the true coherence of the late reverberation field are depicted. It is evident that the diffuse noise modeling is only a rough estimate of the true late reverberation field, while the coherence obtained by the proposed algorithm is closer to the true sound field. The good fit between the estimated sound field and the true sound field is reflected in the improved dereverberation performance of the proposed algorithm.

4) Results for the noiseless case: In the noiseless case, only the dereverberation task is carried out as elaborated in Section IV-C. As a benchmark to the proposed algorithm, the WPE algorithm was applied to the same data. The length of the prediction filter was set to 3, when $\mathrm{T}_{60} = 0.36$ s, and to 9, when $\mathrm{T}_{60} = 0.61$ s. The lengths were optimized in order to achieve the best possible results. The range of lengths which was examined was $1, \ldots, 7$ when $\mathrm{T}_{60} = 0.36$ and $5, \ldots, 13$ when $\mathrm{T}_{60} = 0.61$. The other parameters of the WPE algorithm remained fixed according to the nominal values provided on the website. The N outputs of the WPE algorithm were delayed and summed similarly to (72). By this, the early components were also reduced in a similar way to our approach.

In Table X, the results for the algorithms under examination are presented. In terms of LSD results, our algorithm outperforms the WPE algorithm for both reverberation levels. In terms of PESQ scores, the WPE outperforms the proposed algorithm for $T_{60} = 0.61$ s, while the proposed algorithm exhibits better scores for $\mathrm{T}_{60} = 0.36$ s.

Sonogram examples of the various signals for the noiseless case are depicted in Fig. 4. Fig. 4a depicts the early reverberation. The presence of reverberation is evident in Fig. 4b, depicting the observed signal. Fig. 4c depicts the result of the single-channel algorithm, and Fig. 4e depicts the output signal after the initialization stage $\hat{S}_{\mathrm{INIT}}(m)$. The output of the proposed algorithm $\widehat{S}_{\mathrm{O}}(m)$ is depicted in Fig. 4f. The output of WPE is depicted in Fig. 4d. By careful examination of the sonograms, it can be verified that the algorithm is capable of dereverberating the signal and that it outperforms $\hat{S}_{\mathrm{INIT}}(m)$. The output of WPE has some residual reverberation which can be noticed. Audio examples for both the noisy and noiseless

[4]Details on the diffuse noise generator can be found in [49] and the software can be freely downloaded from https://www.audiolabs-erlangen.de/fau/professor/habets/software/noise-generators

TABLE II: PESQ scores (top) and LSD results (bottom) for the proposed algorithm along the iterations index.

Alg.\ L	0	1	2	3	4	5	6	7
proposed EM	2.17	2.19	2.19	2.19	2.19	2.18	2.17	2.17
Alg.\ L	0	1	2	3	4	5	6	7
proposed EM	3.86	3.66	3.63	3.62	3.62	3.62	3.64	3.64

TABLE III: PESQ scores (top) and LSD results (bottom) for $T_{60} = 0.61$ s with T_{60} in (65) set in the range $0.1, 0.3, \ldots, 1.5$ s.

Alg.\ T_{60}	0.1	0.3	0.5	0.7	0.9	1.1	1.3	1.5
4-channel derev. (init.)	2.05	2.15	2.17	2.17	2.17	2.16	2.16	2.15
proposed EM	2.08	2.17	2.18	2.19	2.19	2.19	2.18	2.18
Alg.\ T_{60}	0.1	0.3	0.5	0.7	0.9	1.1	1.3	1.5
4-channel derev. (init.)	4.10	3.90	3.87	3.86	3.87	3.87	3.88	3.89
proposed EM	3.86	3.68	3.64	3.62	3.62	3.63	3.63	3.63

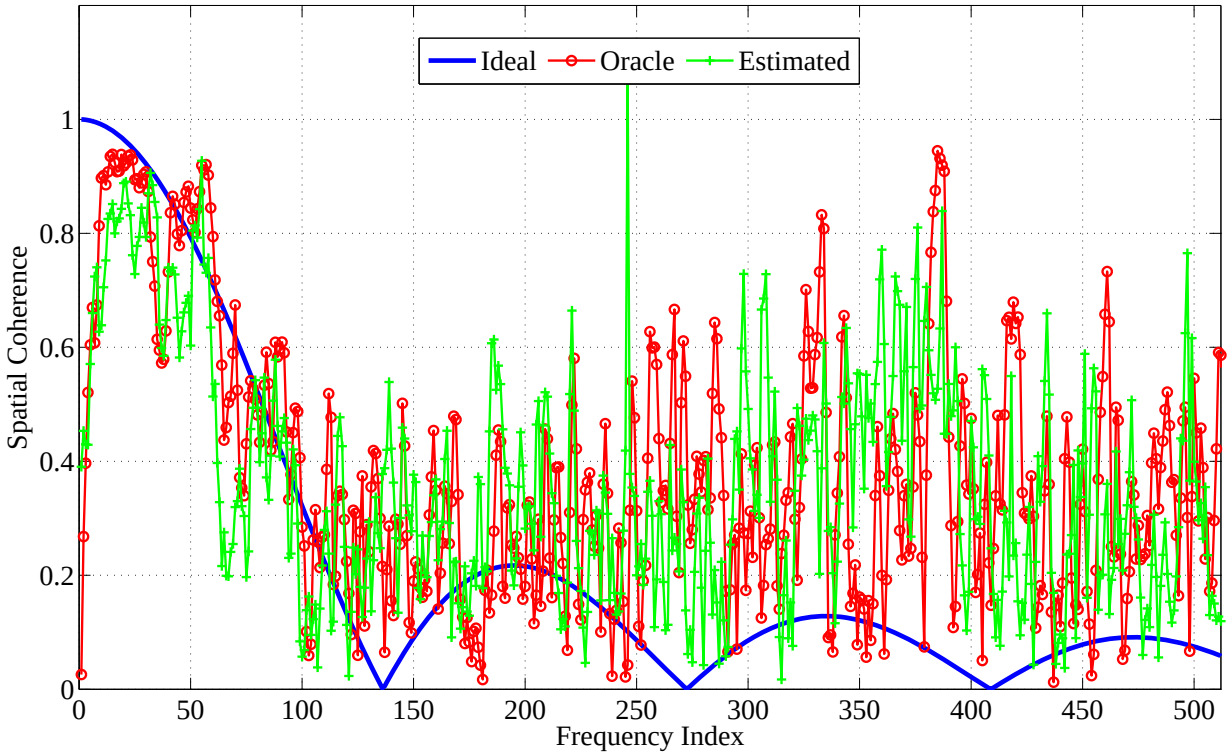


Fig. 3: Spatial coherence between the first and second microphone of i) an ideal diffuse field, ii) the oracle late reverberation field, and iii) the estimated reverberant field.

TABLE IV: PESQ scores (top) and LSD results (bottom) for the proposed EM with $\lambda_{RNR} = \lambda_{RER} = 10^{-2}, \ldots, 10^{2}$, determining the frames used for the estimation of $\mathbf{\Gamma}$.

Alg.\ $\lambda_{RNR}, \lambda_{RER}$	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}
Proposed EM	2.16	2.18	2.19	2.18	2.19
Alg.\ $\lambda_{RNR}, \lambda_{RER}$	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}
Proposed EM	3.74	3.68	3.63	3.65	3.74

TABLE V: PESQ scores (top) and LSD results (bottom) for the proposed EM where $\lambda_{RNR} = \lambda_{RER} = 10^{-2}, \ldots, 10^{2}$, determining the frames used for the estimation of $\phi_R(m)$.

Alg.\ $\lambda_{RNR}, \lambda_{RER}$	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}
Proposed EM	2.18	2.19	2.19	2.18	2.18
Alg.\ $\lambda_{RNR}, \lambda_{RER}$	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}
Proposed EM	3.64	3.63	3.63	3.65	3.64

cases are available on our website[5]. By listening to these examples, it is evident that the proposed algorithm produces

[5] http://www.eng.biu.ac.il/gannot/speech-enhancement/

TABLE VI: PESQ scores (top) and LSD results (bottom) for the proposed EM where $\lambda_{ERR} = \lambda_{ENR} = 10^{-2} \dots 10^{2}$, determining the frames used for the estimation of $\phi_S(m)$ and $\mathbf{g}_e$.

Alg.\ $\lambda_{ERR}, \lambda_{ENR}$	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}
Proposed EM	2.20	2.20	2.20	2.21	2.19
$\lambda_{ERR}, \lambda_{ENR}$	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}
Proposed EM	3.63	3.63	3.62	3.60	3.69

TABLE VII: PESQ scores (top) and LSD results (bottom) for the proposed EM with $\gamma = 0, 0.5, 1$ and using upper limitation of $\phi_R(m)$

Alg.\ γ	0	0.5	1	U.L.
Proposed EM	2.20	2.19	2.18	2.20
Alg.\ γ	0	0.5	1	U.L.
Proposed EM	3.61	3.61	3.63	3.62

the most natural dereverberated speech, when compared with the baseline and competing algorithms.

VI. Conclusions

In this contribution, a novel algorithm was presented to obtain an estimate of a spatially filtered version of the early speech component, thereby suppressing early reflections, late reverberation and ambient noise. The EM algorithm was used to estimate the spatial parameters of the early speech and the late reverberation components. The early speech component was modelled as anechoic speech multiplied by an early TF, while the late reverberation was modelled as additive interference with time-invariant spatial characteristics and a time-varying level. The PSD of the anechoic speech, the ETFs and the PSD matrix of the late reverberation (modelling the time-varying level and time-invariant spatial characteristics) were estimated by the M-step of the EM algorithm. The hidden data was defined to be the anechoic speech and the late reverberation signals. The parameters-based estimation of the anechoic speech was obtained in the E-step of each iteration. As a result of a gain ambiguity problem between the anechoic speech and the ETFs, only a filtered version of the early speech components was estimated. The algorithm was tested in a room with a reverberation time of 0.36 s and 0.61 s for several signal-to-noise levels. In terms of the objective performance measures as well as an informal listening test, the proposed algorithm outperforms baseline and competing single-channel and multichannel dereverberation algorithms for the considered scenarios.

References

[1] A. Kjellberg, "Effects of reverberation time on the cognitive load in speech communication: Theoretical considerations," *Noise and Health*, vol. 7, no. 25, pp. 11–21, 2004.

[2] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, Feb. 2015.

[3] J. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912–915, 1977.

[4] K. Lebart, J.-M. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[5] E. A. P. Habets, "Single-channel speech dereverberation based on spectral subtraction," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, 2004, pp. 250–254.

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

[7] 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, 2009.

[8] E. A. P. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 945–958, 2013.

[9] E. A. P. H. Habets, "Towards multi-microphone speech dereverberation using spectral enhancement and statistical reverberation models," in *Asilomar Conference on Signals, Systems and Computers*. IEEE, 2008, pp. 806–810.

[10] M. Hodgson, "When is diffuse-field theory applicable?" *Applied Acoustics*, vol. 49, no. 3, pp. 197–207, 1996.

[11] 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.

[12] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. Thesis, Technische Universiteit Eindhoven, Jun. 2007.

[13] M. Triki and D. T. Slock, "Iterated delay and predict equalization for blind speech dereverberation," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2006.

[14] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 430–440, 2007.

[15] S. Subramaniam, A. P. Petropulu, and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 392–396, 1996.

[16] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. IEEE, 2003, pp. 92–95.

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

[18] M. Miyoshi and Y. Kenda, "Inverse filtering of room acoustics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 36, no. 2, pp. 145–152, 1988.

[19] M. Kallinger and A. Mertins, "Multi-channel room impulse response shaping - A study," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[20] A. Mertins, T. Mei, and M. Kallinger, "Room impulse response shortening/reshaping with infinity- and p-norm optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 249–259, Feb. 2010.

[21] W. Zhang, E. A. P. Habets, and P. A. Naylor, "On the use of channel shortening in multichannel acoustic system equalization," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2010.

[22] I. Kodrasi and S. Doclo, "Robust partial multichannel equalization techniques for speech dereverberation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 537–540.

[23] B. Yegnanarayana, P. Satyanarayana Murthy, C. Avendano, and H. Hermansky, "Enhancement of reverberant speech using lp residual," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. IEEE, 1998, pp. 405–408.

[24] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "On the use of linear prediction for dereverberation of speech," in *Proc. Int. Workshop Acoust. Echo Noise Control*, vol. 1, 2003, pp. 99–102.

[25] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using lp residual signal," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 267–281, 2000.

[26] B. W. Gillespie, H. Malvar, and D. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 6, 2001, pp. 3701–3704.

[27] 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, 2011.

TABLE VIII: PESQ scores (top) and LSD results (bottom) for reverberant signals plus directional noise for a reverberation time of 0.36 s (left) and 0.61 s (right).

Alg.\SNR	10 dB	15 dB	20 dB	25 dB	30 dB	10 dB	15 dB	20 dB	25 dB	30 dB
Unprocessed	1.48	1.72	2.00	2.25	2.43	1.48	1.65	1.82	1.93	2.00
1-channel derev.	1.67	2.00	2.33	2.59	2.77	1.64	1.87	2.04	2.15	2.22
4-channel derev. (oracle)	2.62	2.94	3.18	3.33	3.40	2.14	2.35	2.51	2.63	2.70
4-channel derev. (init.)	2.46	2.78	2.98	3.10	3.17	2.01	2.18	2.29	2.35	2.39
Proposed EM 2 ite.	**2.56**	**2.90**	**3.13**	**3.27**	**3.33**	**2.05**	**2.21**	**2.31**	**2.37**	**2.42**

Alg.\SNR	10 dB	15 dB	20 dB	25 dB	30 dB	10 dB	15 dB	20 dB	25 dB	30 dB
Unprocessed	10.06	7.39	5.32	3.93	3.05	10.75	8.20	6.33	5.14	4.41
1-channel derev.	7.25	5.21	3.81	2.96	2.53	7.77	5.85	4.58	3.84	3.45
4-channel derev. (oracle)	2.83	2.24	1.91	1.76	1.70	3.76	2.07	2.69	2.50	2.41
4-channel derev. (init.)	3.47	2.73	2.34	2.12	2.00	4.63	3.76	3.33	3.13	3.04
Proposed EM 2 ite.	**3.01**	**2.37**	**2.07**	**1.93**	**1.87**	**4.11**	**3.38**	**3.04**	**2.88**	**2.80**

TABLE IX: PESQ scores (top) and LSD results (bottom) for reverberant signals plus diffuse noise for a reverberation time of 0.36 s (left) and 0.61 s (right).

Alg.\SNR	10 dB	15 dB	20 dB	25 dB	30 dB	10 dB	15 dB	20 dB	25 dB	30 dB
Unprocessed	1.49	1.74	2.02	2.26	2.43	1.50	1.67	1.83	1.93	1.99
1-channel derev.	1.70	2.00	2.31	2.57	2.75	1.66	1.87	2.03	2.14	2.21
4-channel derev. (oracle)	2.24	2.63	2.96	3.23	3.38	2.05	2.31	2.52	2.65	2.73
4-channel derev. (init.)	2.16	2.49	2.77	2.98	3.12	1.92	2.11	2.24	2.32	2.37
Proposed EM 2 ite.	**2.20**	**2.54**	**2.84**	**3.09**	**3.26**	**1.95**	**2.14**	**2.28**	**2.37**	**2.43**

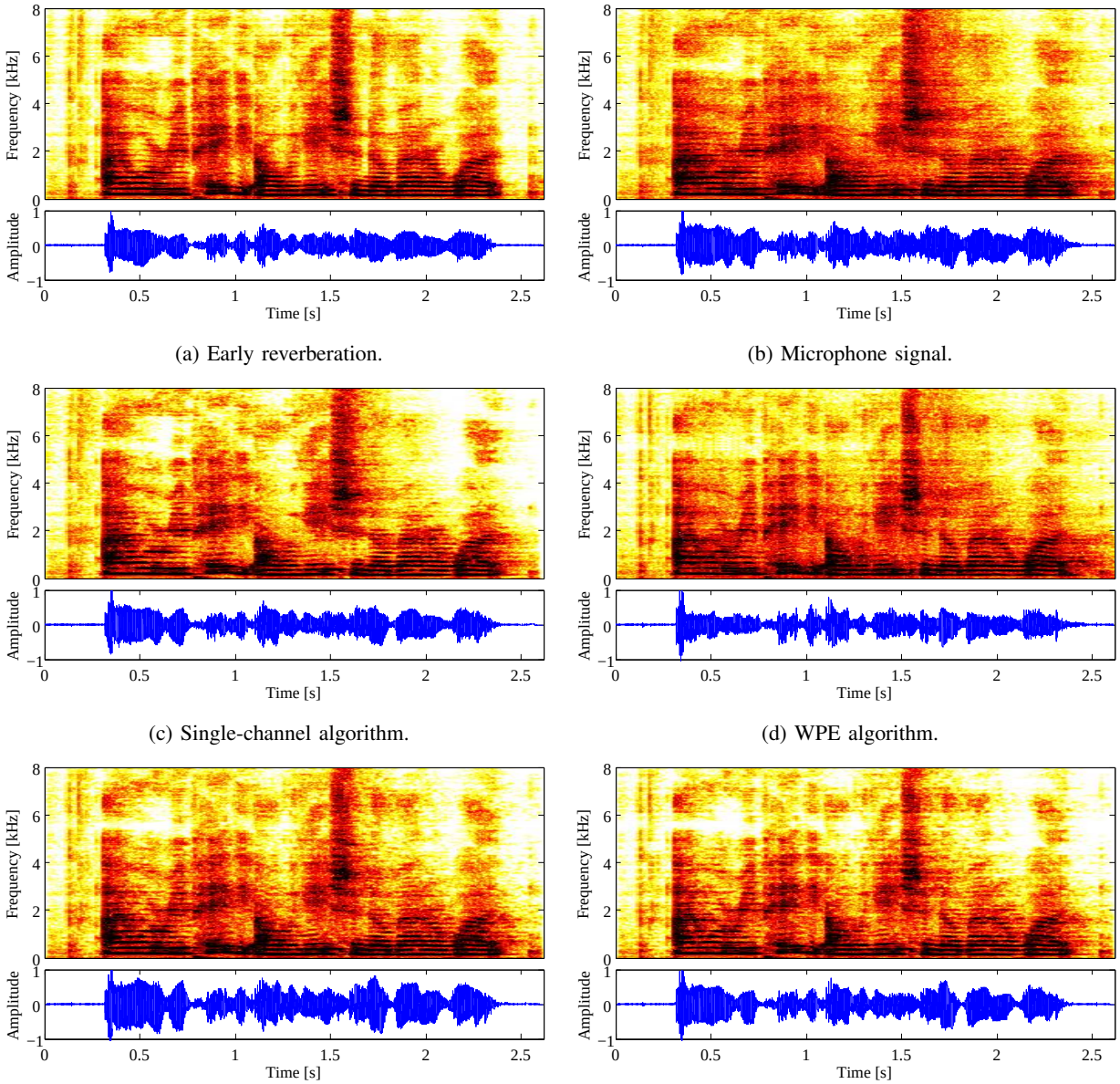
Alg.\SNR	10 dB	15 dB	20 dB	25 dB	30 dB	10 dB	15 dB	20 dB	25 dB	30 dB
Unprocessed	15.52	11.71	8.1	5.10	3.17	15.63	12.00	8.67	6.06	4.49
1-channel derev.	11.19	7.82	5.14	3.44	2.60	11.43	8.22	5.76	4.22	3.49
4-channel derev. (oracle)	6.19	3.99	2.63	1.97	1.71	6.11	4.14	3.03	2.54	2.39
4-channel derev. (init.)	6.72	4.35	2.95	2.31	2.07	7.24	4.87	3.58	3.07	2.92
Proposed EM 2 ite.	**6.48**	**4.18**	**2.80**	**2.15**	**1.92**	**6.93**	**4.66**	**3.43**	**2.92**	**2.78**

TABLE X: PESQ scores (top) and LSD results (bottom) for reverberant signals for two different reverberation times.

Alg.\ T_{60}	0.36 s	0.61 s
Unprocessed	2.68	2.07
1-channel derev.	2.99	2.28
4-channel derev. (oracle)	3.44	2.75
4-channel derev. (WPE)	3.09	**2.65**
4-channel derev. (init.)	3.28	2.42
Proposed EM 2 ite.	**3.39**	2.46

Alg.\ T_{60}	0.36 s	0.61 s
Unprocessed	2.24	3.77
1-channel derev.	2.25	3.20
4-channel derev. (oracle)	1.69	2.35
4-channel derev. (WPE)	2.11	2.69
4-channel derev. (init.)	1.72	2.78
Proposed EM 2 ite.	**1.71**	**2.67**

[28] J. D. Polack, "La transmission de l'énergie sonore dans les salles," Ph.D. dissertation, Université du Maine, Le Mans, France, 1988.
[29] 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.
[30] E. A. P. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Mar. 2005, pp. 173–176.
[31] E. A. P. Habets, S. Gannot, and I. Cohen, "Dual-microphone speech dereverberation in a noisy environment," in *IEEE International Symposium on Signal Processing and Information Technology*, 2006, pp. 651–655.
[32] H. W. Löllmann and P. Vary, "Low delay noise reduction and dereverberation for hearing aids," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 1, 2009.
[33] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 231–246, 2009.
[34] X.-L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: A general framework," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.
[35] D. Schmid, S. Malik, and G. Enzner, "An expectation-maximization algorithm for multichannel adaptive speech dereverberation in the frequency-domain," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 17–20.
[36] B. Schwartz, S. Gannot, and E. A. P. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 394–406, Feb 2015.
[37] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 373–385, Jul. 1998.
[38] J.-F. Cardoso, H. Snoussi, J. Delabrouille, and G. Patanchon, "Blind separation of noisy Gaussian stationary sources. Application to cosmic microwave background imaging," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Toulouse, France, Sep. 2002.
[39] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.
[40] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.
[41] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. P. Horaud, "A variational EM algorithm for the separation of moving sound sources," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, Oct. 2015.
[42] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 845–856, 2005.
[43] N. Q. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for gaussian model based reverberant audio source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, pp. 1–11, 2013.



(a) Early reverberation.

(b) Microphone signal.

(c) Single-channel algorithm.

(d) WPE algorithm.

(e) Initial estimate using (60).

(f) Proposed EM algorithm after 2 iterations.

Fig. 4: Spectrograms of a real noiseless recording with $T_{60} = 0.61$ s.

[44] M. L. Eaton, *Multivariate statistics: A vector space approach*, ser. Lecture Notes–Monograph Series. Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2007, vol. 53.

[45] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*. Berlin Heidelberg: Springer, 2001, pp. 39–60.

[46] R. Balan and J. Rosca, "Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase," in *IEEE Sensor Array and Multichannel Signal Processing Workshop*, 2002, pp. 209–213.

[47] E. A. P. Habets, S. Gannot, and I. Cohen, "Speech dereverberation using backward estimation of the late reverberant spectral variance," in *Proc. IEEE Convention of Electrical & Electronics Engineers in Israel (IEEEI)*, 2008, pp. 384–388.

[48] N. Dal Degan and C. Prati, "Acoustic noise analysis and speech enhancement techniques for mobile radio applications," *Signal Processing*, vol. 15, no. 1, pp. 43–56, 1988.

[49] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Am.*, vol. 122, pp. 3464–3470, Dec. 2007.

[50] E. A. P. Habets, "A distortionless subband beamformer for noise reduction in reverberant environments," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Tel-Aviv, Israel, Aug. 2010.

[51] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2159–2169, 2011.

[52] M. Taseska and E. A. P. Habets, "MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Aachen, Germany, Sep. 2012.

[53] R. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 223–233, Jan. 2012.

[54] ITU-T, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, International Telecommunications Union (ITU-T) Recommendation P.862, Feb. 2001.

[55] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 313–317.

Ofer Schwartz received his B.Sc. (Cum Laude) and M.Sc. degrees in Electrical Engineering from Bar-Ilan University, Israel in 2010 and 2013, respectively. He is now pursuing his Ph.D. in Electrical Engineering at the Speech and Signal Processing laboratory of the Faculty of Engineering at Bar-Ilan University. His research interests include statistical signal processing and in particular dereverberation and noise reduction using microphone arrays and speaker localization and tracking.

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 a Full Professor at the Faculty of Engineering, Bar-Ilan University, Israel, where he is heading the Speech and Signal Processing laboratory and the Signal Processing Track. 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 several 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. Currently, he serves as the committee vice-chair. 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 was a member of the IEEE Signal Processing Society Standing Committee on Industry Digital Signal Processing Technology (2013-2015), and a Guest Editor for the IEEE Journal of Selected Topics in Signal Processing and the EURASIP Journal on Advances in Signal Processing. He is the recipient, with S. Gannot and I. Cohen, of the 2014 IEEE Signal Processing Letters Best Paper Award. Currently, he is a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing (2011-2016), vice-chair of the EURASIP Special Area Team on Acoustic, Sound and Music Signal Processing, an Associate Editor of the IEEE Signal Processing Letters, and Editor in Chief of the EURASIP Journal on Audio, Speech, and Music Processing.