A Recursive Expectation-Maximization Algorithm for Online Multi-Microphone Noise Reduction

Ofer Schwartz
CEVA DSP, Audio Department
Herzliya, Israel
Ofer.Schwartz@ceva-dsp.com

Sharon Gannot
Bar-Ilan University, Faculty of Engineering
Ramat-Gan, Israel
Sharon.Gannot@biu.ac.il

***Abstract*—Speech signals, captured by a microphone array mounted to a smart loudspeaker device, can be contaminated by ambient noise. In this paper, we present an online multichannel algorithm, based on the recursive EM (REM) procedure, to suppress ambient noise and enhance the speech signal. In the E-step of the proposed algorithm, a multichannel Wiener filter (MCWF) is applied to enhance the speech signal. The MCWF parameters, that is, the power spectral density (PSD) of the anechoic speech, the steering vector, and the PSD matrix of the noise, are estimated in the M-step. The proposed algorithm is specifically suitable for online applications since it uses only past and current observations and requires no iterations.**

To evaluate the proposed algorithm we used two sets of measurements. In the first set, static scenarios were generated by convolving speech utterances with real room impulse responses (RIRs) recorded in our acoustic lab with reverberation time set to 0.16 s and several signal to directional noise ratio (SDNR) levels. The second set was used to evaluate dynamic scenarios by using real recordings acquired by CEVA "smart and connected" development platform.

Two practical use cases were evaluated: 1) estimating the steering vector with a known noise PSD matrix and 2) estimating the noise PSD matrix with a known steering vector. In both use cases, the proposed algorithm outperforms baseline multichannel denoising algorithms.

I. Introduction

Noisy speech can be difficult to understand for both humans and machines, and can lead to listening fatigue. Multichannel noise reduction has become a major research topic in the past decade due to available multichannel arrays (frequently used in smart loudspeaker devices) and computational power.

The minimum variance distortionless response (MVDR) beamformer (BF) is a popular noise reduction algorithm [1]–[4]. The MVDR BF requires two parameters: 1) the steering vector of the speakers and 2) the noise PSD matrix. Usually, the noise PSD matrix is estimated using speech absence segments. These segments can be detected using perfect voice activity detector (VAD). The steering vector of the speaker can be estimated using a speaker localization algorithm like the steered response power (SRP)-phase transform (PHAT) [5]. In the presence of directional noise, the SRP-PHAT tends to be biased.

In the past, the EM algorithm [6] was used for noisy speech enhancement [7]–[10]. In [4], the multichannel joint dereverberation and denoising problem is addressed, and a procedure for simultaneous estimation of all relevant beamformer parameters is proposed. 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 MCWF, while the PSD of the anechoic speech, the early transfer functions (ETFs) (actually, their corresponding normalized relative early transfer functions (RETFs)), the time-invariant spatial coherence matrix of the late reverberation, and the PSD of the late reverberation are estimated in the M-step. The algorithm operates in a batch mode and is not appropriate for online requirements.

An REM algorithm is an implementation of the EM iterations in a recursive manner. Two versions of the REM algorithm are described in the literature, one of which was proposed by Titterington [11] and the other by Cappé and Moulines [12]. The Titterington recursive EM (TREM) version is based on implementing the M-step by the Newton method; the Cappé and Moulines recursive EM (CREM) version is based on the time-smoothing of the auxiliary function obtained through the E-step. Both variants were utilized in [13] to derive speaker tracking schemes by processing the pair-wise relative phases between each pair of microphones. The problem of speech enhancement and noise reduction was not addressed in this paper.

In this paper, an online REM-based multichannel algorithm for suppressing the ambient noise and for enhancing the speech signal is presented. In the E-step of the proposed algorithm, an MCWF is applied to enhance the speech signal. The MCWF parameters (that is, the PSD of the anechoic speech), the steering vector, and the time-varying PSD matrix of the noise, are estimated in the M-step. These steps are recursively implemented by applying the REM procedure, which uses only past and current frames, and requires only a single iteration per time instant. Experimentally, we have realized that the proposed REM procedure cannot converge well if an estimate of the entire set of parameters is required. The cases that were found to be practical are: 1) estimating the steering vector with a known noise PSD matrix and 2) estimating the noise PSD matrix with a known steering vector. These cases were tested using real RIR recordings from the Bar Ilan university (BIU) acoustic lab and real recordings using the CEVA-DSP "smart and connected" development platform. It is shown that the proposed algorithm outperforms baseline multichannel

ISBN 978-90-827970-1-5 © EURASIP 2018

denoising algorithms.

The remainder of this paper is organized as follows. In Section II, we formulate the noise reduction problem. In Section III, the REM procedure for our statistical model is derived. In Section IV, the performance of the proposed algorithm is evaluated. Section V is dedicated to concluding remarks.

II. Problem Formulation

In the following section, the multichannel noise reduction problem is formulated. The observations consist of speech in a noisy environment and is modeled in the short-time Fourier transform (STFT) domain as:

$$Y_i(m,k) = X_i(m,k) + V_i(m,k), \tag{1}$$

where $Y_i(m,k)$ denotes the ith microphone observation at time index m and frequency index k, $X_i(m,k)$ denotes the speech component, and $V_i(m,k)$ denotes the ambient noise. Here $X_i(m,k)$ is modeled as a multiplication between the speech received by the first microphone and the direct transfer function (DTF), that is:

$$X_i(m,k) = G_i(m,k)\ X_1(m,k), \tag{2}$$

where $G_i(k)$ is the DTF. Note that, in the general case, the relative transfer function (RTF) or the RETF can be used as in [4]. The DTF is constructed by the time difference of arrival (TDOA) between the microphones:

$$G_i(m,k) = \exp\left(-j\frac{2\pi k}{K}\frac{\tau_i(m)}{T_s}\right) \tag{3}$$

where $\tau_i(m)$ is the TDOA between the ith microphone and first microphone, T_s is the sampling time, and K is the number of frequency bins. For uniform linear array (ULA), the TDOA equals $\tau_i = (i-1)\frac{d\cos(\vartheta(m))}{c}$, where d is the microphones inter-distance, c is the sound velocity, and $\vartheta(m)$ is the angle related to the RTF. Concatenating the N microphone signals in a vector form yields:

$$\mathbf{y}(m,k) = \mathbf{x}(m,k) + \mathbf{v}(m,k) \tag{4}$$

$$\mathbf{x}(m,k) = \mathbf{g}(m,k)X_1(m,k), \tag{5}$$

where

$$\mathbf{y}(m,k) = \begin{bmatrix} Y_1(m,k) & Y_2(m,k) & \dots & Y_N(m,k) \end{bmatrix}^T$$
$$\mathbf{x}(m,k) = \begin{bmatrix} X_1(m,k) & X_2(m,k) & \dots & X_N(m,k) \end{bmatrix}^T$$
$$\mathbf{v}(m,k) = \begin{bmatrix} V_1(m,k) & V_2(m,k) & \dots & V_N(m,k) \end{bmatrix}^T$$
$$\mathbf{g}(m,k) = \begin{bmatrix} G_1(m,k) & G_2(m,k) & \dots & G_N(m,k) \end{bmatrix}^T.$$

The ambient noise is modeled as a zero-mean Gaussian vector with PSD matrix $\mathbf{\Phi_v}(k)$:

$$f\left(\mathbf{v}(m,k);\mathbf{\Phi_v}(m,k)\right) = \mathcal{N}^C(\mathbf{v}(m,k);0,\mathbf{\Phi_v}(m,k)). \tag{6}$$

where:

$$\mathcal{N}^C(\mathbf{z};\mathbf{0},\mathbf{\Phi}) = \frac{1}{\pi^N|\mathbf{\Phi}|}\exp\left(-\mathbf{z}^{\mathrm{H}}\ \mathbf{\Phi}^{-1}\mathbf{z}\right), \tag{7}$$

where $\mathbf{z}$ denotes a Gaussian vector, $\mathbf{\Phi}$ is a PSD matrix, and $|\cdot|$ denotes the matrix determinant operation. The signal at the first microphone $X_1(m,k)$ can also be modeled as a zero-mean Gaussian process with variance $\phi_X(m,k) = E\{|X_1(m,k)|^2\}$. The parameter set for the k-th frequency bin is:

$$\theta(k) = \{\phi_X(m,k), \mathbf{g}(m,k), \mathbf{\Phi_v}(m,k)\quad \forall m\}. \tag{8}$$

The speech component and the ambient noise are assumed to be mutually uncorrelated. This means that the observed signal vector is also a zero-mean Gaussian vector, and that the PSD matrix of the observed signals is equal to the sum of the individual PSD matrices of the speech component and ambient noise. The PSD matrix of the observations is given by:

$$\mathbf{\Phi_y}(m,k) = \phi_X(m,k)\mathbf{g}(m,k)\mathbf{g}^{\mathrm{H}}(m,k) + \mathbf{\Phi_v}(m,k). \tag{9}$$

Our goal now is to maximize the probability density function (p.d.f.) of the measurements with respect to the parameters, that is, to apply the maximum likelihood (ML) criterion yielding $\theta(k)$:

$$\theta_{\mathrm{ML}}(k) = \underset{\theta}{\mathrm{argmax}}\prod_m f(\mathbf{y}(m,k);\theta(k)) \tag{10}$$

where $f(\cdot)$ denotes p.d.f. and statistical independence is assumed between each time-instance. The maximization operation might be a cumbersome task. To simplify the derivations, the expectation-maximization (EM) formulation is adopted in the following section. Moreover, to achieve online estimation of the parameter set and to maintain smooth estimates over time of the speech signal, the REM algorithm is adopted.

III. The REM Algorithm

In this paper, we adopt a recursive procedure based on the CREM algorithm [12]. The CREM is a recursive version of the batch EM algorithm. To implement the EM algorithm, the *hidden data* should be defined. We are proposing to define $X_1(m,k)$ as the hidden data. The E-step evaluates the auxiliary function, while the maximization-step maximizes the auxiliary function with respect to the set of parameters. This batch EM procedure converges to a local maximum of the likelihood function of the observation [6]. To track time-varying parameters and to satisfy the online requirements, the CREM [12] algorithm is adopted. This algorithm is based on the time-smoothing of the auxiliary function obtained through the E-step and employing a single maximization per time-instance. In the following sections, the frequency index k is omitted for brevity whenever no ambiguity arises.

A. Recursive Expectation Maximization steps

CREM is based on smoothing the auxiliary function along the time axis and executing single maximization per time instance. The smoothing operation is given by [12, Eq. (10)]:

$$Q_R\left(\theta;\hat{\theta}(m)\right) = \alpha Q_R\left(\theta;\theta(m-1)\right) + (1-\alpha)Q\left(\theta;\hat{\theta}(m)\right), \tag{11}$$

ISBN 978-90-827970-1-5 © EURASIP 2018

where the instantaneous auxiliary function of the m-th observation is given by:

$$Q\left(\theta;\hat{\theta}(m)\right) = E\left\{\log f(\mathbf{y}(m), X_1(m);\theta)|\mathbf{y}(m);\hat{\theta}(m)\right\}, \tag{12}$$

where $Q_R\left(\theta;\hat{\theta}(m)\right)$ is the recursive auxiliary function, and $\hat{\theta}(m)$ is the estimate of θ at the m-th time instance. The $(m+1)$th parameter set estimate is obtained by maximizing $Q_R\left(\theta;\hat{\theta}(m)\right)$ w.r.t. θ.

The joint p.d.f. of the observations and the clean speech (that is, the complete data) is given by:

$$f(\mathbf{y}, X_1;\theta) = \mathcal{N}^C(\mathbf{y}-\mathbf{g}X_1, 0, \mathbf{\Phi}_\mathbf{v})\mathcal{N}^C(X_1, 0, \phi_X). \tag{13}$$

where the independence between the speech and noise was invoked.

The E-step in the m-th time instance boils down to the calculation of $Q\left(\theta;\hat{\theta}(m)\right)$. Similarly to [4], it is sufficient to estimate the following sufficient statistics:

$$1.\quad \widehat{X}_1(m) = \phi_X(m)\mathbf{g}^\mathrm{H}(m)\mathbf{\Phi}_\mathbf{y}^{-1}(m)\,\mathbf{y}(m) \tag{14a}$$

$$2.\quad \widehat{|X_1(m)|^2} = |\widehat{X}_1(m)|^2 + \phi_X(m) \tag{14b}$$

$$- \phi_X^2(m)\mathbf{g}^\mathrm{H}(m)\mathbf{\Phi}_\mathbf{y}^{-1}(m)\mathbf{g}(m), \tag{14c}$$

where $\mathbf{\Phi}_\mathbf{y}(m)$ is defined by (9).

After some algebraic steps, the implementation of (11) is summarized according to the following recursive equations:

$$1.\quad \eta_R(m) = \alpha_\eta\eta_R(m-1) + (1-\alpha_\eta)\ \widehat{|X_1(m)|^2} \tag{15a}$$

$$2.\quad \zeta_R(m) = \alpha_\zeta\zeta_R(m-1) + (1-\alpha_\zeta)\mathbf{y}(m)\widehat{X}_1^*(m) \tag{15b}$$

$$3.\quad \mathbf{Z}_R(m) = \alpha_\mathbf{Z}\mathbf{Z}_R(m-1) + (1-\alpha_\mathbf{Z})\widehat{\mathbf{Z}}(m), \tag{15c}$$

where $\eta_R(m)$, $\zeta_R(m)$, and $\widehat{\mathbf{Z}}_R(m)$ are recursive sufficient statistics, and

$$\widehat{\mathbf{Z}} \triangleq \widehat{(\mathbf{y}-\mathbf{g}X_1)(\mathbf{y}-\mathbf{g}X_1)}^\mathrm{H}$$

$$= \ \mathbf{y}\mathbf{y}^\mathrm{H} - \mathbf{g}\widehat{X}_1\mathbf{y}^\mathrm{H} - \mathbf{y}\widehat{X}_1^*\mathbf{g}^\mathrm{H} + \widehat{|X_1|^2}\ \mathbf{g}\mathbf{g}^\mathrm{H}. \tag{16}$$

Similarly to the batch mode EM, the M-step is obtained by maximizing $Q_R\left(\theta;\hat{\theta}(m)\right)$ with respect to the problem parameters:

$$1.\quad \phi_X(m+1) = \eta_R(m) \tag{17a}$$

$$2.\quad \mathbf{g}(m+1) = \frac{\zeta_R(m)}{\eta_R(m)} \tag{17b}$$

$$3.\quad \mathbf{\Phi}_\mathbf{v}(m+1) = \mathbf{Z}_R(m). \tag{17c}$$

B. Practical Considerations

Several practical aspects are discussed in the sequel.

1) Fitting the Steering Vector: The estimated DTF $\mathbf{g}$ can be tuned to the closest feasible steering vector by a least squares (LS) fitting:

$$\mathbf{g}(m) \leftarrow \underset{\tilde{\mathbf{g}}}{\operatorname{argmin}}\,||\mathbf{g}(m)-\tilde{\mathbf{g}}||^2 \tag{18}$$

where $\tilde{\mathbf{g}}$ is a feasible steering vector selected from a predefined set of directions. For example, for ULA, the steering vectors has the following shape:

$$\tilde{\mathbf{g}} = \left[\begin{array}{cccc} 1 & \exp\left(-j\frac{2\pi k}{K}\frac{\tau}{T_s}\right) & \dots & \exp\left(-j\frac{2\pi k}{K}\frac{(N-1)\tau}{T_s}\right) \end{array}\right]^T \tag{19}$$

where τ is the time delay between two successive microphones.

2) Avoiding Speech Distortion: The MCWF typically distorts the speech signal [14]. To avoid speech distortion in estimating X_1, the output of the MVDR BF can be used instead of the output of the MCWF. The MCWF in (14a) can be split into a multichannel MVDR beamformer $\mathbf{w}_\mathrm{MVDR}$ and a subsequent single-channel Wiener filter H_WF, as shown in [15], [16]:

$$\widehat{X}_1 = \phi_X\mathbf{g}^\mathrm{H}\mathbf{\Phi}_\mathbf{y}^{-1}\,\mathbf{y} \tag{20}$$

$$= \underbrace{\frac{\phi_X}{\phi_X + (\mathbf{g}^H\mathbf{\Phi}_\mathbf{v}^{-1}\mathbf{g})^{-1}}}_{\text{Post}-\text{filter}}\ \underbrace{\frac{\mathbf{g}^H\mathbf{\Phi}_\mathbf{v}^{-1}}{\mathbf{g}^H\mathbf{\Phi}_\mathbf{v}^{-1}\mathbf{g}}\mathbf{y}}_{\text{MVDR}\quad\text{estimate}}\ . \tag{21}$$

In our experimental study, only the MVDR output is evaluated. Note that the MVDR BF is totally determined by $\mathbf{\Phi}_\mathbf{v}$ and $\mathbf{g}$, which are the main parameters of our study. We stress that the REM recursion is always applied with the outcome of the E-step, namely the MCWF.

IV. Performance Evaluation

According to our preliminary experimental study, joint estimation of the entire parameter set in (8) is a cumbersome task and may suffer from convergence problems. We have therefore limited the experimental study to two practical cases: 1) a known noise PSD matrix and an unknown DOA, and 2) a known DOA and an unknown noise PSD matrix. Case #1 is evaluated for both static speakers, using real room impulse responses (RIRs) recorded at the BIU acoustic lab with reverberation time set to 0.16 s and several SDNR levels, and for dynamic speakers, using real recordings using the CEVA "smart and connected" development platform. In case #2, the algorithm was tested for dynamic noise source and static speakers, only using the CEVA platform. These cases, although limited, are practically important. Case #1 represents a very common case in which the noise source is static, e.g. air-condition, and the speaker is free to move. In this case, the noise characteristics can be easily estimated by employing a simple VAD. Case #2 represents a case where the speaker direction of arrival (DOA) with respect to the array is known, e.g. a moving person wearing a headset or a bluetooth headphones.

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE I: PESQ scores (left) and LSD results (right) for the DOA tracking.

Alg.\SNR	0 dB	10 dB	20 dB	0 dB	10 dB	20 dB
Unprocessed	0.09	0.85	2.27	14.90	9.62	5.31
Proposed MVDR	**2.73**	2.77	**2.85**	**4.96**	**5.81**	**5.82**
SRP based MVDR	2.56	**2.86**	2.77	5.79	5.99	6.44

A. DOA Estimation with a Known Noise PSD Matrix

In this section, the results of the proposed algorithm are shown for an unknown speaker DOA and a known noise PSD matrix (that is, only ϕ_X and $\mathbf{g}$ were estimated in the M-step).

First, the DOA estimation was tested for static speakers using real recorded RIRs. A loudspeaker was positioned at a distance of 2 m in front of a non-uniform linear array with 8 microphones. Further details about the RIR database can be found in [17]. Anechoic speech signals (20 signals, each 20 sec long) evenly distributed between male and female speakers, were convolved by the RIRs. The reverberation time was $T_{60} = 0.16$. Directional noise with various SDNR levels was added. Sensor noise was also added with 40 dB signal-to-noise ratio (SNR). The sampling frequency was 16 kHz and the frame length of the STFT was 32 ms with 8 ms between successive time frames (that is, 75% overlap).

The speaker was positioned at 105^o and the noise source at 15^o. The smoothing parameters α_η and α_ζ were set to 0.8 and 0.99, respectively. As a comparison, the DOA was estimated using the SRP-PHAT technique [5]. The observed signals were then filtered with an MVDR beamformer steered towards the estimated DOA.

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) [18] and log-spectral distance (LSD). The speech enhancement quality measures of the EM-based MVDR estimates, the SRP-based MVDR estimates, and the unprocessed signals are presented in Table I for several directional noise SNR levels. The best achieved scores are depicted in boldface. The proposed algorithm usually outperforms the competing algorithm since it is based on a more accurate DOA estimate.

Second, the DOA estimation was tested for dynamic speaker using the CEVA "smart and connected" development platform (see Figure 1). The platform has 13 digital microphones in a circular array. The signals are captured using pulse-width modulation (PDM) in 1.5 MHz and transformed to pulse-code modulation (PCM) in 16 kHz using a Cascaded Integrator Comb (CIC) filter.[1]

The noise source was positioned at 300^o w.r.t. the array. The speaker moved in a circle from 0^o to 270^o, and started to speak after eight seconds. It can be qualitatively deduced from Figure 2 that the SRP focuses on the noise DOA while the proposed REM tracks the speaker.

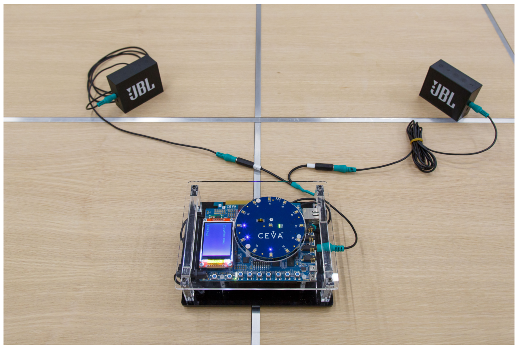


Fig. 1: CEVA "smart and connected" development platform.

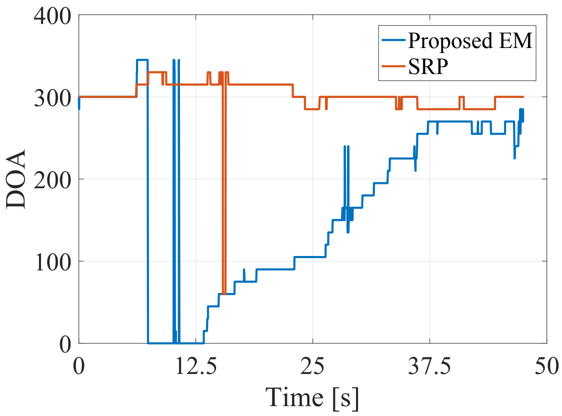


Fig. 2: Example of DOA tracking.

B. Noise PSD Matrix Estimation with Known DOA

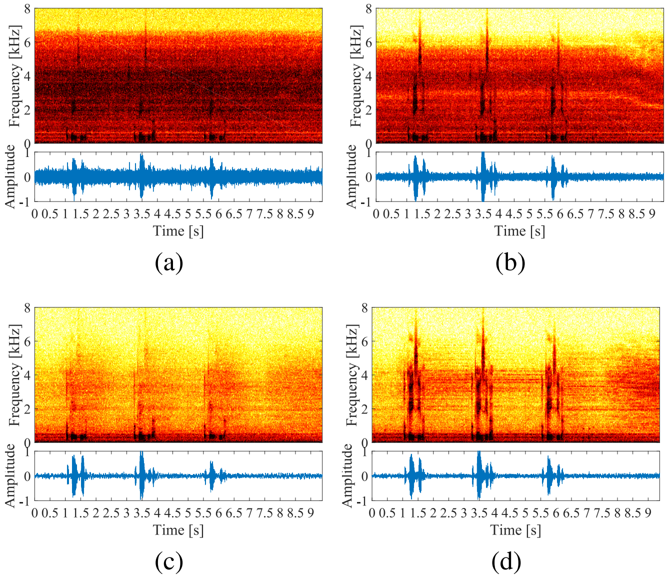
In this section, the results of the proposed algorithm are shown for a known speaker DOA and an unknown noise PSD matrix (that is, only ϕ_X and $\mathbf{\Phi_v}$ were estimated in the M-step). The case of a static noise source was not tested because the noise PSD matrix can be easily estimated using speech absence segments detected by VAD. The proposed noise PSD estimation was tested using a dynamic noise source, in which noise-only segments are useless because the noise PSD is highly time-variable. The smoothing parameters α_η and $\alpha_\mathbf{Z}$ were set to 0.1 and 0.95, respectively. For comparison, the minimum power distortionless response (MPDR) and the delay and sum (DS) BFs were calculated using the known DOA.

To circumvent self-attenuation, the MPDR BF and $\mathbf{Z}_R(m)$ in (16) were updated only when low speech was assumed. To detect low speech, simple directional VAD was used using the steered response to signal ratio (SRSR),

$$\text{SRSR} \equiv \frac{|\mathbf{g}^H\mathbf{y}|^2}{\mathbf{y}^H\mathbf{y}}.$$

The MPDR BF and $\mathbf{Z}_R(m)$ were updated only when $\text{SRSR} < \eta_1$, where η_1 denoted a threshold (set to 0.8 in our implementation).

[1] For more details about the chip, please visit https://www.ceva-dsp.com/product/ceva-teaklite-4/



(a)

(b)

(c)

(d)

Fig. 3: Example sonograms for input SNR of 10 dB. (a) Observed signal. (b) DAS output. (c) MPDR output. (d) Proposed EM-based MVDR.

The observed signal was recorded by a CEVA platform installed at BIU acoustic lab. The speaker was positioned at 0^o w.r.t. the platform, and the noise source was moved in a circle around the platform. Sonograms of the various signals are depicted in Figure 3. The audio examples are available online: `http://www.eng.biu.ac.il/gannot/speech-enhancement`.

The DS is characterized by low speech distortion, but exhibits limited noise reduction performance. Our version of MPDR reduces more noise (w.r.t. the DS) but severely degrades the speech signal. The proposed EM-based MVDR reduces more noise since the time-varying noise PSD matrix can be estimated even during speech segments.

V. Conclusions

In this paper, an online algorithm for speech enhancement based on the REM algorithm was presented. The PSD of the speech, the DOA, and the PSD matrix of the noise were estimated by the M-step of the REM algorithm. The hidden data was defined to be the speech, as received by the first microphone. The parameter-based estimation of the speech was obtained in the E-step as the MCWF. The algorithm was experimentally tested for two cases: 1) known noise PSD matrix and unknown DOA for both static and dynamic speakers 2) known DOA and unknown and varying noise PSD matrix caused by a moving noise source. In terms of the objective performance measures, the proposed algorithm outperforms baseline multichannel denoising algorithms for the considered scenarios. This conclusion can also be qualitatively deduced by inspecting the speech sonograms and listening to a few sound samples.

VI. Acknowledgments

The authors would like to thank Ofer Shahen-Tov and Aviv David from CEVA DSP for their technical assistance.

References

[1] S. Affes and Y. Grenier, “A signal subspace tracking algorithm for microphone array processing of speech,” *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 425–437, 1997.

[2] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[3] O. Schwartz, S. Gannot, and E. A. 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, 2015.

[4] ——, “An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1495–1510, 2016.

[5] J. DiBiase, H. Silverman, and M. Brandstein, “Post-filtering techniques,” in *Microphone Arrays*. Berlin Heidelberg: Springer, 2001, pp. 157–180.

[6] 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.

[7] M. Feder, A. Oppenheim, and E. Weinstein, “Methods for noise cancellation based on the em algorithm,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP’87*, vol. 12. IEEE, 1987, pp. 201–204.

[8] M. Feder, A. V. Oppenheim, and E. Weinstein, “Maximum likelihood noise cancellation using the em algorithm,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 2, pp. 204–216, 1989.

[9] E. Weinstein, A. V. Oppenheim, and M. Feder, “Signal enhancement using single and multi-sensor measurements,” Research Laboratory of Electronics, Massachusetts Institute of Technology (MIT), Tech. Rep., 1990.

[10] 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.

[11] D. Titterington, “Recursive parameter estimation using incomplete data,” *J. Roy. Statist. Soc. Ser. B*, vol. 46, pp. 257–267, 1984.

[12] O. Cappé and E. Moulines, “On-line expectationmaximization algorithm for latent data models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 3, pp. 593–613, 2009.

[13] O. Schwartz and S. Gannot, “Speaker tracking using recursive em algorithms,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 2, pp. 392–402, 2014.

[14] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, “Speech distortion weighted multichannel wiener filtering techniques for noise reduction,” in *Speech enhancement*. Springer, 2005, pp. 199–228.

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

[16] 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.

[17] 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.

[18] 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.