An EM Algorithm for Joint Dual-Speaker Separation and Dereverberation

Nili Cohen, Gershon Hazan, Boaz Schwartz, Sharon Gannot
Faculty of Engineering, *Bar Ilan University*, Ramat-Gan, Israel
nili.cohen@biu.ac.il, hazanshl@gmail.com, boazsh0@gmail.com, sharon.gannot@biu.ac.il

***Abstract*—The scenario of a mixture of two speakers captured by a microphone array in a noisy and reverberant environment is considered. If the problems of source separation and dereverberation are treated separately, performance degradation may result. It is well-known that the performance of blind source separation (BSS) algorithms degrades in the presence of reverberation, unless reverberation effects are properly addressed (leading to the so-called convolutive BSS algorithms). Similarly, the performance of common dereverberation algorithms will severely degrade if an interference signal is also captured by the same microphone array. The aim of the proposed method is to jointly separate and dereverberate the two speech sources, by extending the Kalman expectation-maximization for dereverberation (KEMD) algorithm, previously proposed by the authors. A statistical model is attributed to this scenario, using the convolutive transfer function (CTF) approximation, and the expectation-maximization (EM) scheme is applied to obtain a maximum likelihood (ML) estimate of the parameters. In the expectation step, the separated clean signals are extracted from the observed data by the application of a Kalman Filter, utilizing the parameters that were estimated in the previous iteration. The maximization step updates the parameters estimation according to the E-step output. Simulation results shows that the proposed method improves both the separation of the signals and their overall quality.**

***Index Terms*—Array processing, blind source separation, dereverberation, expectation-maximization, convolution in STFT**

I. Introduction

Audio (blind) source separation is a family of algorithms that aims at the extraction of a speech signal(s) from a mixture of several sound sources [1], [2]. Such algorithms can be deployed in many modern devices. e.g. hearing aids, hands-free phones, and conference call systems. BSS is also required as a preprocessing stage for automatic speech recognition (ASR) systems, commonly used in modern smart assistants.

Sound sources recorded in real environments often involve acoustic reverberation. While propagating in an acoustic enclosure, the sound wave undergoes reflections from the room facets and from various objects. These reflections are detrimental to both speech quality and, in severe cases, its intelligibility. Furthermore, reverberation increases the time dependency between speech frames, rendering source separation problems further challenging.

The BSS problem can be addressed from the perspective of simultaneous estimation of acoustic parameters and clean speech signals. Since neither the speech signals, nor the acoustical parameters are known in advance, the EM algorithm [3] can be applied and lead to a local ML estimate of the parameters, as was proposed in e.g. [4]. Other EM algorithms for BSS were proposed by using the nonnegative matrix factorization (NMF) method [5], which is a very useful tool in audio BSS applications.

In [6], an EM algorithm for both dereverberation and noise reduction is presented. The room impulse response (RIR) is modelled as an auto-regressive (AR) process in each frequency band, and the EM algorithm is used to estimate both the clean signal and the modelled system. The method was extended in [7] to simultaneously dereverberate and separate multiple speakers. Source separation based on CTF in the short-time Fourier transform (STFT) domain is presented in [8], for known mixing filters. For reverberated environment, an EM method is proposed in [9] to jointly estimate the model parameters, including the CTF coefficients of the mixing filters, and infer the sources.

A Kalman expectation-maximization (KEM) scheme for single-microphone speech enhancement in the time-domain was presented in [10], and the KEM scheme was extended in [11] for speech dereverberation in the STFT domain. In the E-step, the Kalman smoother is applied to extract the clean signal from the data, utilizing the estimated parameters. In the M-step, the parameters are updated using the output of the Kalman smoother. We refer to this algorithm as KEMD, which was further extended to the recursive, segmental, and binaural cases in [12]–[14], respectively.

In this paper, we propose the Kalman expectation-maximization for dereverberation and separation (KEMDS) algorithm, which is an extension of the KEMD to the case of two speakers. In the E-step of the proposed algorithm, a Kalman Filter is applied to jointly separate and dereverberate the two speech sources. In the M-step, the CTFs of the two speakers are updated.

II. Statistical Model

In the following, we introduce a statistical model that represents a scenario with two desired speakers, including both reverberation and ambient noise. Extension to more than two speakers is straightforward but cumbersome, and is out of the scope of this paper.

Let $x[n]$ and $y[n]$ be the time-domain clean speech signals of the first and second speakers. The signals are captured by an array of J microphones, with the jth microphone signal given by

$$z_j[n] = x[n] * h_j[n] + y[n] * g_j[n] + v_j[n] \ , \qquad (1)$$

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

with $h_j[n]$ and $g_j[n]$ the RIRs between the first and second speakers and the jth microphone, respectively, $*$ denotes time-domain convolution, and $v_j[n]$ represents the respective additive noise.

The problem is formulated in the STFT domain, where $x(t,k)$ and $y(t,k)$ denote the STFT representation of $x[n]$ and $y[n]$, respectively, with $t \in \{1,\dots,T\}$ the time-frame and $k \in \{1,\dots,K\}$ the frequency-bin. Assuming the source signals are short-term stationary signals, and applying a proper STFT analysis, $x(t,k)$ and $y(t,k)$ can be modelled as independent complex-Gaussian random variables:

$$x(t,k) \sim \mathcal{N}_C\left\{0, \sigma_x^2(t,k)\right\} \ , \ y(t,k) \sim \mathcal{N}_C\left\{0, \sigma_y^2(t,k)\right\}, \tag{2}$$

where $\sigma_x^2(t,k)$ and $\sigma_y^2(t,k)$ denote the short-time power spectrum of $x[n]$ and $y[n]$, respectively, and $\mathcal{N}_C$ denotes a proper complex-Gaussian distribution.

In the STFT domain, the RIRs can be approximately modelled by a CTF model [15]; an approximation that was successfully applied for dereverberation [16] and beamforming [17]. Using the CTF model, (1) can be approximated by

$$z_j(t,k) = \mathbf{h}_j^T(k) \cdot \mathbf{x}_t(k) + \mathbf{g}_j^T(k) \cdot \mathbf{y}_t(k) + v_j(t,k), \tag{3}$$

where the CTF systems are

$$\begin{aligned} \mathbf{h}_j(k) &= [h_{j,L-1}(k), \ \dots \ , \ h_{j,0}(k)]^T, \\ \mathbf{g}_j(k) &= [g_{j,L-1}(k), \ \dots \ , \ g_{j,0}(k)]^T, \end{aligned} \tag{4}$$

and the state-vectors of the desired signals are

$$\begin{aligned} \mathbf{x}_t(k) &= [x(t-L+1,k), \ \dots \ , x(t,k)]^T, \\ \mathbf{y}_t(k) &= [y(t-L+1,k), \ \dots \ , y(t,k)]^T \end{aligned} \tag{5}$$

with L the CTF length that depends on the reverberation time. We further assume that $v_j(t,k)$ are stationary complex-Gaussian uncorrelated random processes, namely:

$$v_j(t,k) \sim \mathcal{N}_C\left\{0, \sigma_{v_j}^2(k)\right\} \tag{6}$$

and $E\{v_j(t,k)v_i^*(t,k)\} = 0$ for $j \neq i$.

III. Algorithm Derivation

We now derive an EM-based algorithm for the joint parameter estimation. The desired signals will be inferred as a byproduct.

A. Parameter Estimation Problem

Let $\mathcal{Z}$ be the set of all available measurements:

$$\mathcal{Z} = \{z_j(t,k): \ j=1,\dots,J \ , \ t=1,\dots,T \ , \ k=1,\dots,K\}.$$

As often encountered in many statistical models, maximizing the likelihood function $f(\mathcal{Z};\mathbf{\Theta})$ is intractable, and therefore necessitates the application of the EM procedure. The set of parameters comprises the following subsets:

$$\begin{aligned} \mathbf{\Theta} &\equiv \{\mathbf{\Theta}_X, \mathbf{\Theta}_Y, \mathbf{\Theta}_H, \mathbf{\Theta}_G, \mathbf{\Theta}_V\} \\ \mathbf{\Theta}_X &\equiv \left\{\sigma_x^2(t,k)\right\}_{t,k}, \ \mathbf{\Theta}_Y \equiv \left\{\sigma_y^2(t,k)\right\}_{t,k} \\ \mathbf{\Theta}_H &\equiv \left\{\mathbf{h}_j(k)\right\}_{j,k}, \ \mathbf{\Theta}_G \equiv \left\{\mathbf{g}_j(k)\right\}_{j,k}, \ \mathbf{\Theta}_V \equiv \left\{\sigma_{v_j}^2(k)\right\}_{j,k} \end{aligned} \tag{7}$$

for all $j=1,\dots,J$, $t=1,\dots,T$, and $k=1,\dots,K$. The latent data in this problem is defined as the STFT coefficients of the two clean speech signals:

$$\begin{aligned} \mathcal{X} &= \{x(t,k): \ t=1,\dots,T \ , \ k=1,\dots,K\} \\ \mathcal{Y} &= \{y(t,k): \ t=1,\dots,T \ , \ k=1,\dots,K\} \ . \end{aligned} \tag{8}$$

For conciseness, the frequency index k will be omitted in the rest of the derivation.

In the E-step of the EM procedure, the auxiliary function is first calculated:

$$Q\left(\mathbf{\Theta}\middle|\hat{\mathbf{\Theta}}^{(p-1)}\right) \equiv E\Big\{ \log f(\mathcal{Z},\mathcal{X},\mathcal{Y};\mathbf{\Theta})\Big|\mathcal{Z};\hat{\mathbf{\Theta}}^{(p-1)}\Big\}, \tag{9}$$

where $\hat{\mathbf{\Theta}}^{(p-1)}$ is the parameter estimate at iteration p. In the M-step, the new parameter estimate $\hat{\mathbf{\Theta}}^{(p)}$ is calculated by:

$$\hat{\mathbf{\Theta}}^{(p)} = \arg\max_{\mathbf{\Theta}} Q\left(\mathbf{\Theta}\middle|\hat{\mathbf{\Theta}}^{(p-1)}\right) \ . \tag{10}$$

Under the statistical model presented in Sec. II, the complete-data log-likelihood is given by

$$\begin{aligned} \log f(\mathcal{X},\mathcal{Y},\mathcal{Z};\mathbf{\Theta}) &= C \\ &- \frac{1}{2}\sum_{t=1}^{T}\left[\log(\sigma_x^2(t)\sigma_y^2(t)) + \frac{|x(t)|^2}{\sigma_x^2(t)} + \frac{|y(t)|^2}{\sigma_y^2(t)}\right] \\ - \frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{J}&\left[\log\sigma_{v_j}^2 + \frac{1}{\sigma_{v_j}^2}\sum_{t=1}^{T}\left|z_j(t) - \mathbf{h}_j^T\mathbf{x}_t - \mathbf{g}_j^T\mathbf{y}_t\right|^2\right] \end{aligned} \tag{11}$$

with C a constant, independent of $\mathbf{\Theta}$.

A few notes are in place. The term in the second line of (11) is the log-likelihood of the clean speech signals, and due to the independence between time frames, it can be expressed as a summation over the time index t. The term in the third line is the log-likelihood of the additive noise, and due to the independence of noise signals across microphones it decomposes to a sum over the J microphones. Calculating the expected value of $\log f(\mathcal{X},\mathcal{Y},\mathcal{Z};\mathbf{\Theta})$ involves the computation of the first- and second-order statistics of $\mathbf{x}_t$ and $\mathbf{y}_t$, derived in the next section.

B. E-Step: Kalman Filter

The *Kalman smoother* implements the conditional expectation in (9) in the Gaussian case for time-varying problems. The application of the Kalman smoother necessitates forward-backward recursions. Our preliminary examination showed that avoiding the backward path results in only a marginal performance degradation. We have therefore decided to only apply the forward recursion, namely the *Kalman filter*, that is extensively covered in the science and engineering literature.

The application of the Kalman filtering requires state-space formulation, and therefore we concatenate the state-vectors in (5) as:

$$\boldsymbol{\mu}_t = [\mathbf{x}_t^T \ , \ \mathbf{y}_t^T]^T \ . \tag{12}$$

The dynamical model is given by a simple random walk:

$$\boldsymbol{\mu}_t = \boldsymbol{\Phi}\boldsymbol{\mu}_{t-1} + \mathbf{w}_t \quad , \quad \boldsymbol{\Phi} = \left[\begin{array}{c|c} \boldsymbol{\Phi}_X & \mathbf{0} \\ \hline \mathbf{0} & \boldsymbol{\Phi}_Y \end{array}\right] \tag{13}$$

where $\boldsymbol{\Phi}_X$ and $\boldsymbol{\Phi}_Y$ are identical $L \times L$ matrices

$$\boldsymbol{\Phi}_X \equiv \boldsymbol{\Phi}_Y \equiv \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix} ,$$

and the innovation process is given by

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

Note that multiplication by $\boldsymbol{\Phi}_X$ and $\boldsymbol{\Phi}_Y$ corresponds to time-shift of the state-vector $\boldsymbol{\mu}_{t-1}$, neglecting the correlation between adjacent frames. The validity of this assumption depends on the percentage of overlap between STFT frames, as will be demonstrated in Sec. V. The observed signal is given by:

$$\mathbf{z}_t = \ \mathbf{Q}\boldsymbol{\mu}_t + \mathbf{v}_t \ , \quad \mathbf{Q} = \begin{bmatrix} \mathbf{H} & \mathbf{G} \end{bmatrix}^T , \tag{14}$$

where the observation matrices are

$$\mathbf{H} \equiv [\mathbf{h}_1,\ \ldots\ ,\ \mathbf{h}_J]^T \quad , \quad \mathbf{G} \equiv [\mathbf{g}_1,\ \ldots\ ,\ \mathbf{g}_J]^T ,$$

with $\mathbf{h}_j$ and $\mathbf{g}_j$ defined in (4), and the measurement and noise vectors are given by

$$\mathbf{z}_t \equiv [z_1(t),\ \ldots\ ,\ z_J(t)]^T \quad , \quad \mathbf{v}_t \quad \equiv [v_1(t),\ \ldots\ ,\ v_J(t)]^T \ .$$

Finally, the second-order statistics of $\mathbf{w}_t$ and $\mathbf{v}_t$ is given by

$$\mathbf{F}_t \equiv E\left\{\mathbf{w}_t\mathbf{w}_t^H\right\} = \begin{bmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \cdots & \sigma_x^2(t) & 0 & \cdots & \vdots \\ \vdots & \cdots & 0 & 0 & \cdots & \vdots \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & \sigma_y^2(t) \end{bmatrix} ,$$

$$\mathbf{B} \equiv E\left\{\mathbf{v}_t\mathbf{v}_t^H\right\} = \begin{bmatrix} \sigma_{v_1}^2 & \cdots & \cdots & 0 \\ 0 & \sigma_{v_2}^2 & & \\ & & \ddots & \\ 0 & \cdots & \cdots & \sigma_{v_J}^2 \end{bmatrix} ,$$

where $(\cdot)^H$ indicates the Hermitian operator. The corresponding Kalman filtering procedure is summarized in Algorithm 1.

The outcome of the Kalman filter is the state-vector estimator, $\hat{\boldsymbol{\mu}}_{t|t}$ and its respective error covariance matrix $\mathbf{P}_{t|t}$. Let $\mathbf{P}_{t|t}$ be partitioned as follows,

$$\mathbf{P}_{t|t} = \left[\begin{array}{c|c} [\mathbf{P}_{xx}]_{t|t} & [\mathbf{P}_{xy}]_{t|t} \\ \hline [\mathbf{P}_{xy}]_{t|t}^H & [\mathbf{P}_{yy}]_{t|t} \end{array}\right] .$$

In the M-step below, the following first- and second-order statistics terms, resulting from the application of the Kalman filter, are used [10]:

$$\hat{\boldsymbol{\mu}}_t = E\left\{\boldsymbol{\mu}_t \middle| \mathcal{Z}; \boldsymbol{\Theta}^{(p-1)}\right\} = \hat{\boldsymbol{\mu}}_{t|t} \ , \tag{15a}$$

$$\widehat{\boldsymbol{\mu}_t\boldsymbol{\mu}_t^H} = E\left\{\boldsymbol{\mu}_t\boldsymbol{\mu}_t^H \middle| \mathcal{Z}; \boldsymbol{\Theta}^{(p-1)}\right\} = \hat{\boldsymbol{\mu}}_{t|t}\hat{\boldsymbol{\mu}}_{t|t}^H + \mathbf{P}_{t|t} \ . \tag{15b}$$

C. M-Step: Parameter Estimation

The maximization of the auxiliary function (10) is obtained by setting the partial derivatives with respect to the various parameters to zero. The clean signals spectra at the pth iteration are given by:

$$\left[\sigma_x^2(t)\right]^{(p)} = \widehat{|x(t)|^2}, \qquad \left[\sigma_y^2(t)\right]^{(p)} = \widehat{|y(t)|^2}, \tag{16}$$

where $\widehat{|x(t)|^2}$ and $\widehat{|y(t)|^2}$ can be obtained from the (L)-th and $(2L)$-th diagonal elements of the second-order statistics term in (15b), respectively. Note that these estimates are simply given by the periodogram due to the assumption that the speech frames are uncorrelated and Gaussian.

For the estimation of the CTF and noise parameters, we will use the following time-averaged, second-order terms,

$$\mathbf{r}_{z_j z_j} \equiv \frac{1}{T}\sum_{t=1}^{T} |z_j(t)|^2 \ , \qquad \mathbf{r}_{\mu z_j}^{(p-1)} \equiv \frac{1}{T}\sum_{t=1}^{T} \hat{\boldsymbol{\mu}}_t z_j^*(t),$$

$$\mathbf{R}_{\mu\mu}^{(p-1)} \equiv \frac{1}{T}\sum_{t=1}^{T} \widehat{\boldsymbol{\mu}_t\boldsymbol{\mu}_t^H} \ .$$

The vector of CTF coefficients for the jth microphone at the pth iteration is given by:

$$\mathbf{q}_j{}^{(p)} \equiv \begin{bmatrix} \mathbf{h}_j{}^* \\ \mathbf{g}_j{}^* \end{bmatrix}^{(p)} = \left[\mathbf{R}_{\mu\mu}^{(p-1)}\right]^{-1} \mathbf{r}_{\mu z_j}^{(p-1)} \ , \tag{17}$$

which is the least-squares (LS) fit between the estimated state-vector and the measurement. The noise spectra coefficients are

Algorithm 1: The Kalman Filter.

for $t = 1$ **to** T **do**
- **Predict:**
- $\hat{\boldsymbol{\mu}}_{t|t-1} = \boldsymbol{\Phi} \cdot \hat{\boldsymbol{\mu}}_{t-1|t-1}$
- $\mathbf{P}_{t|t-1} = \boldsymbol{\Phi} \cdot \mathbf{P}_{t-1|t-1} \cdot \boldsymbol{\Phi}^T + \mathbf{F}_t$
- **Update:**
- $\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{Q}^H \left[\mathbf{Q}\mathbf{P}_{t|t-1}\mathbf{Q}^H + \mathbf{B}\right]^{-1}$
- $\mathbf{e}_t = \mathbf{z}_t - \mathbf{Q}\hat{\boldsymbol{\mu}}_{t|t-1}$
- $\hat{\boldsymbol{\mu}}_{t|t} = \hat{\boldsymbol{\mu}}_{t|t-1} + \mathbf{K}_t \cdot \mathbf{e}_t$
- $\mathbf{P}_{t|t} = \left[\mathbf{I} - \mathbf{K}_t\mathbf{Q}\right]\mathbf{P}_{t|t-1}$

end

the residual energy of this LS fit, given by

$$(\sigma_{v_j}^2)^{(p)} = \frac{1}{T}\sum_{t=1}^{T}\overbrace{\left|z_j(t) - \left(\mathbf{q}_j{}^T\right)^{(p)}\boldsymbol{\mu}_t\right|^2} = \mathbf{r}_{z_j z_j} - 2\Re\left[\left(\mathbf{q}_j{}^H\right)^{(p)}\mathbf{r}_{\mu z_j}^{(p-1)}\right] + \left(\mathbf{q}_j{}^H\right)^{(p)}\mathbf{R}_{\mu\mu}^{(p-1)}\mathbf{q}_j{}^{(p)} \ . \quad (18)$$

The proposed algorithm is dubbed KEMDS.

IV. Initialization of the EM Iterations

The parameter initialization strategy is a crucial component in any application of the EM algorithm, and it is important in directing the EM to the desired local maximum. In the current contribution, we do not assume any prior information on the clean sources signals, their activity patterns, or their locations, which makes the initialization task more challenging.

We propose to initialize the clean speech variances $\sigma_x^2(t,k)$ and $\sigma_y^2(t,k)$ using the output of degenerate unmixing estimation technique (DUET) method [18]. This method separates the sources under the assumption that the sources are W-disjoint orthogonal, that is, that the supports of the signals in the STFT domain are disjoint sets. In the presence of reverberation, the separation performance of the DUET algorithm is rather limited. Nevertheless, we found that the DUET method can provide a good initialization for the clean speech signals variances.

An initial value of the acoustic systems of the two speakers, namely $\mathbf{H}$ and $\mathbf{G}$, should also be set. To this end, we further utilize the DUET output, by applying a LS fit between its outputs and the noisy reverberated mixed signal. The length L of the CTFs should be set in accordance with the reverberation time and the STFT parameters [19]. However, preliminary examination showed that as L increases, the estimation error increases as well. Therefore, shorter filters length was used, with the additional purpose of reduced computational load. Furthermore, we have noticed that better convergence is achieved if the initial estimate of the CTF coefficients are forced to decay. This is obtained by multiplying the LS estimate with an exponentially decaying series.

V. Simulation Results

A. Setup

The KEMDS algorithm was evaluated using simulated mixtures of two concurrently active speakers in a reverberant environment. Clean signals from the two genders were drawn from the TIMIT database [20] and concatenated to form utterances of length 8 sec, where the sampling rate was 16 kHz. To construct the microphone signals, each sentence was convolved with time-invariant RIRs downloaded from the open database recorded at the acoustic lab at Bar-Ilan University [21]. The reverberation level of the $(6 \times 6 \times 2.4)$ m acoustic Lab can be controlled by flipping 60 dedicated panels covering the lab facets. The RIRs were captured by an eight-microphone linear array with inter-distances of $\{3, 3, 3, 8, 3, 3, 3\}$ cm. We have selected room setup with reverberation time $T_{60} = 610$ ms to also demonstrate the dereverberation effects. The speakers' positions were arbitrarily selected from 13 different available angles on a semi-circle with a radius of 2 m with a resolution of 15°. Finally, the reverberant signals were contaminated with spatially white noise to obtain reverberated-signal to noise ratio (RSNR) value of 20 or 30 dB.

TABLE I
Comparison between the proposed method and the DUET method for $T_{60} = 610$ ms and RSNR $= 30$ dB. The input SIR values of the y- and x-speakers were 1.43 and -0.11 dB, respectively.

Measure	Speaker	DUET	KEMDS	Improvement
SDR	x	-0.59	1.26	1.86
	y	-0.51	1.24	1.75
SIR	x	4.2	5.24	1.04
	y	2.38	6.41	1.03
SAR	x	3.01	5.14	2.12
	y	2.14	4.3	2.15

The STFT analysis utilized a 32 ms Hamming window, with 50% overlap between consecutive time-frames. Note that higher percentage of overlap will result in a significant dependency between adjacent frames. This renders the statistical model of Sec. II inaccurate, and leads to performance degradation, as well as higher computational complexity. Considering this overlap, the reverberation time, and the consideration in Sec. IV, the CTF length L was set to 10 frames. The exponential decay constant, which was used for the CTF coefficients attenuation (see Sec. IV), was set to 0.2. Three EM iterations were executed in all experiments.

B. Separation Results

For the evaluation of the signal separation task, we compared the proposed method and the baseline DUET method, which was also used for the initialization of the proposed EM procedure. We computed three quality measures: signal to distortion ratio (SDR), signal to interference ratio (SIR) and signal to artifacts ratio (SAR) [22] for both methods and for each speaker. The average results are summarized in Table I. It can be seen that the proposed method improves the signal quality and separation level for each speaker.

C. Dereverberation Results

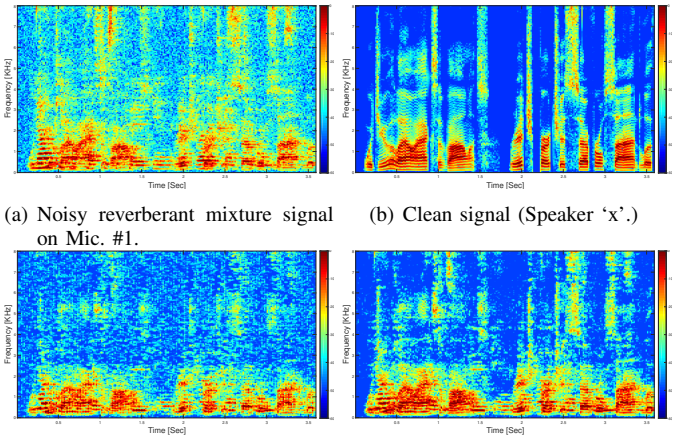
The Dereverberation performance was evaluated using the signal to reverberant ratio (SRR). An estimator for SRR is derived from the power ratio between the early sound and the reverberation tail. The SRR estimator for a tested speaker $s[n]$ is calculated by

$$\mathrm{SRR}\{s\} = 10\log_{10}\frac{\sum_n |s[n] * a_{\text{early}}[n]|^2}{\sum_n |s[n] * a_{\text{late}}[n]|^2} \ , \quad (19)$$

where a_{early} and a_{late} are the early and late reflections of the corresponding RIR at both the input and output of the proposed algorithm. A comparison between the SRR at the input and the output of the KEMDS algorithm is given in Table II, where it can be seen that the proposed algorithm increases the SRR values, indicating reduced reverberation.

TABLE II
SRR VALUES FOR KEMDS AT $T_{60} = 610$ MS AND RSNR $= 30$ DB

Speaker	Input SRR	Output SRR	Improvement
x	-4.02	-1.10	2.91
y	-4.66	-1.63	3.02



(a) Noisy reverberant mixture signal on Mic. #1.

(b) Clean signal (Speaker 'x'.)

(c) DUET output for speaker 'x'.

(d) KEMDS Output for speaker 'x'.

Fig. 1. spectrograms for $T_{60} = 610$ ms and RSNR=20 dB.

D. Subjective Evaluation

The performance of the proposed algorithm was also tested by the assessment of spectrograms and by informal listening tests. spectrograms of the signal as received by the rightest microphone, the clean signal of speaker 'x', the output signal for signal 'x' of the DUET algorithm and of the proposed algorithm, respectively, are depicted in Figure 1. The spectrograms for the second speaker exhibit similar trends and are omitted due to space constraints. It is clearly depicted that the proposed method outperforms the DUET algorithm in both tasks.

Sound samples can be found in the lab website.[1]

VI. CONCLUSIONS

An EM-based algorithm for dual-speaker, multi-microphone speech separation and dereverberation was presented, where both the acoustic parameters and the enhanced signals are estimated. An estimate of the separated, denoised, and dereverberated speech signal is obtained (as a byproduct of the algorithm) at the E-step by applying the Kalman filter. The iterative procedure converges in a low number of iterations. The entire algorithm is applied in the STFT domain, enabling an efficient parallel implementation. Simulation results show better performance compared to the DUET method, with respect to both the signal quality and the separation capabilities. A significant reduction of the reverberation level is also demonstrated. These improvements are validated by both objective measures and by the subjective assessment of speech spectrograms and sound samples.

[1] www.eng.biu.ac.il/gannot/speech-enhancement/

REFERENCES

[1] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*. Wiley, Sep. 2018.

[2] S. Makino, *Audio Source Separation*. Springer, 2018.

[3] 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.

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

[5] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Tran. on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

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

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

[8] X. Li, L. Girin, S. Gannot, and R. Horaud, "Multichannel speech separation and enhancement using the convolutive transfer function," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, 2019.

[9] X. Li, L. Girin, and R. Horaud, "An EM algorithm for audio source separation based on the convolutive transfer function," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 56–60.

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

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

[12] ——, "Online Speech Dereverberation Using Kalman Filter and EM Algorithm," *IEEE/ACM Tran. on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 394–406, Feb. 2015.

[13] ——, "LPC-based speech dereverberation using Kalman-EM algorithm," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Antibes – Juan les Pins, France, Sep. 2014.

[14] ——, "An online dereverberation algorithm for hearing aids with binaural cues preservation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, Oct. 2015.

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

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

[17] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Tran. on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1420–1434, Sep. 2009.

[18] A. Jourjine, S. Rickard, and Ö. Yılmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2000, pp. 2985–2988.

[19] Y. Avargel, "System identification in the short-time fourier transform domain," Ph.D. dissertation, Technion - Israel Institute of Technology, 2008.

[20] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD, USA*, 1988.

[21] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *International Workshop on Acoustic Signal Enhancement 2014 (IWAENC 2014)*, Antibes-Juan les Pins, France, Sep. 2014.

[22] R. G. E. Vincent and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE/ACM Tran. on Acoustics, Speech, and Signal Processing*, vol. 14, no. 4, pp. 1462—-1469, Jul. 2006.