Two Model-Based EM Algorithms for Blind Source Separation in Noisy Environment

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

***Abstract*—The problem of blind separation of speech signals in the presence of noise using multiple microphones is addressed. Blind estimation of the acoustic parameters and the individual source signals are carried out by applying the expectation-maximization (EM) algorithm. Two models for the speech signals are used, namely an unknown deterministic signal model and a complex-Gaussian signal model. For the two alternatives, we define a statistical model and develop EM-based algorithms to jointly estimate the acoustic parameters and the speech signals. The resulting algorithms are then compared from both theoretical and performance perspectives. In both cases, the latent data (differently defined for each alternative) is estimated in the E-step, where in the M-step, the two algorithms estimate the acoustic transfer functions of each source and the noise covariance matrix. The algorithms differ in the way the clean speech signals are used in the EM scheme. When the clean signal is assumed deterministic unknown, only the a posteriori probabilities of the presence of each source are estimated in the E-step, while their time-frequency coefficients are designated as parameters, and are estimated in the M-step using the minimum variance distortionless response beamformer. If the clean speech signals are modelled as complex Gaussian signals, their power spectral densities (PSDs) are estimated in the E-step using the multichannel Wiener filter output. The proposed algorithms were tested using reverberant noisy mixtures of two speech sources in different reverberation and noise conditions.**

I. Introduction

In many applications it is required to recover one or more speech signals from a set of microphone observations, which might include competing speakers, reverberation, and ambient noise. Speech enhancement is required in human-to-human communication systems, conferencing systems, and as a pre-processing step for speech recognition systems. Throughout the past two decades, many methods for multi-microphone speech enhancement were proposed [1]. Separating the desired sources from the undesired sources may rely on spatial information, the statistical independence of the sources, or both.

Spatial information can be utilized by using array processing theory. Different criteria were proposed to estimate the desired signal. The minimum variance distortionless response (MVDR) criterion [2], [3] is based on the preservation of the desired signal while minimizing the power of the interfering sources. An adaptive implementation of this beamformer is the generalized side-lobe canceller (GSC) [4] which was reformulated in the frequency domain, and extended to deal with reverberant environments in [5]. In [6], an multichannel Wiener filter (MWF) approach was proposed to estimate the desired speech in the minimum mean square error (MMSE) sense. The MMSE criterion can be tuned to trade-off between noise reduction and speech distortion [7]. To make use of these algorithms, several sets of parameters governing the desired sources (acoustic transfer functions (ATFs) or covariance matrices) and the noise covariance matrix are required. These parameters are usually not available and should be estimated from the data.

In the general field of signal processing (not necessarily speech signal processing), when only the mixed signals are available, and the acoustic parameters are unknown, the separation problem is usually referred to as a blind source separation (BSS) problem. Several BSS methods utilize the mutual statistical independence of the signals, usually by applying independent component analysis (ICA) algorithms [8]. These methods are also commonly applied to speech separation tasks [9]. Alternatively, the sparsity of the speech sources in the time-frequency (TF) domain can be utilized. Using this property, speech components of simultaneously active speakers can be assumed non-overlapping. In [10] and the references therein, it was shown that due to the sparsity, even when there are fewer microphones than speakers, it is possible to obtain an acceptable separation performance. Under the sparsity property, one should estimate which of the sources is active in each TF bin, namely to calculate a (soft) TF mask. This estimation commonly results in an a posteriori speech presence probability (SPP).

The BSS problem can be addressed also from the perspective of the simultaneous estimation of the acoustical parameters and the clean speech signals. Since neither the speech signals, nor the acoustical parameters are known in advance, the expectation-maximization (EM) algorithm [11], which converges to a local maximum likelihood estimate of the parameters, can be used to solve this problem, as was proposed in [12]–[24]. In [12], the ATFs were approximated by delay-only systems, and the noise covariance matrix was assumed diffuse with time-varying power. In the E-step, each source is estimated using an MVDR beamformer, and in the M-step, the noise power and the delays were estimated. In [13], a blind source separation technique is proposed based on spectral masking and the MVDR beamformer. The direction of arrivals (DOAs) of the different sources are estimated using the EM algorithm, and are used to calculate the masks and the associated MVDR BFs. In [14], the phase difference between channels was exploited for source separation, using an EM algorithm as well as a sequential variant of random sample consensus (RANSAC). Another EM algorithm for acoustic source separation was proposed in [15] for two-channel re-

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

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

cordings, using the interaural phase and level differences.

In reverberant environments, the multichannel ATFs contain more information than phase and level differences, and this information is utilized in [16]–[24]. In [16], the acoustic systems are approximated in the short-time Fourier transform (STFT) domain by the multiplicative transfer function (MTF) model, and an iterative algorithm was applied to estimate these MTFs and the noise covariance matrix. In [17], the MTFs and the acoustic parameters of reverberation are estimated using an EM algorithm for joint dereverberation and noise reduction. In [18], blind source separation is carried out by using a Gaussian mixture model (GMM) and the EM algorithm. The ATF can be modelled also as an auto-regressive (AR) system in the STFT domain, as proposed in [19] in the context of joint speech separation and dereverberation. A different signal model is proposed in [20], where each source is modelled by a complex-Gaussian process with a full-rank covariance matrix. A full-rank model of the sources' covariance matrices is also used in [21], where an online algorithm was developed using the incremental EM [25] for the parameters' update. Another online version was proposed in [22], where the algorithm in [20] is extended to an online algorithm by substituting the iterative-batch M-step by an online recursive version. Two models of the ATF of a mixture are used for a binaural microphone system in [23], and an EM algorithm is applied for the source separation. A method for offline and online noise reduction that uses complex Gaussian model for speech and an EM algorithm was proposed in [24], where it is extended to an online speech enhancement scenario.

Further EM algorithms for BSS were proposed by using the nonnegative matrix factorization (NMF) approach, which is a very useful tool in audio BSS applications [26]. Using this approach, the periodogram of the received signals (usually, calculated as the absolute-squared value of the signal in the STFT domain, or a smoothed version thereof), is decomposed by the NMF to the multiplication of non-negative basis and activation functions. This representation was used in [27], for BSS of speech signals, using the EM algorithm. In [28], [29] the problem of time-varying acoustical systems was targeted, and the NMF representation was estimated using a variational EM approach.

In [30], an EM algorithm is proposed to estimate the location of multiple sources that are considered complex random processes. In the E-step, a delay and sum beamformer (DSB) steered towards the DOA of each of the signals is applied to estimate the desired signals, and in the M-step, the signals estimates are used to update the beamformer weights. In [31], this approach is extended, and the signals are modelled either as deterministic unknown, or as stochastic processes. For each of the models, an EM algorithm is derived for the joint estimation of the source signals and the DOAs.

By the sparsity property of speech signal, a multiple-hypothesis model can be attributed to a mixture of speech sources, where under each hypothesis only a single source is active. Under this model, the MMSE estimate of the separate sources is shown to be an average of the conditional expected values under each hypothesis, weighted by its respective a posteriori speaker presence probability. These a posteriori probabilities appear in several speech enhancement methods, as in [32], where a method for SPP estimation in the presence of stationary noise is presented. Assuming a multivariate complex-Gaussian model for the noise and a rank-1 covariance matrix for the speech, the a posteriori probability of speech presence is derived. An important theoretical result in [32] is that for spatially coherent noise fields, the performance of the speech activity detection can be perfect. The utilization of the a posteriori presence probability is also part of the BSS algorithms proposed in [16], [33], [34] and in the spotforming algorithms proposed in [35], [36].

In the current contribution, we develop two EM-based algorithms for BSS, using the sparsity model, and compare two models of the clean speech signals in the STFT domain; namely an unknown deterministic signal and a complex-Gaussian signal. This extends the methods in [30], [31] by applying these two general alternatives to acoustic signals, using MTF model for the ATFs and a multi-hypothesis model for the speech signals. For the two alternative models, we define the appropriate latent data set, and develop EM-based algorithms for the simultaneous estimation of the acoustic parameters and the speech signal. The resulting algorithms are compared from both theoretical and performance perspectives. In both algorithms, the E-step necessitates the evaluation of the posterior presence probability of each speech source, and then, these posterior probabilities are used for signal estimation. The contribution of an accurate posterior estimates to the total performance of the algorithm is investigated as well. In this light, this paper extends the discussion in [32] to the multiple-speaker case.

To summarize, the contribution of this paper is four-fold: 1) developing of two EM algorithms for BSS, by using two different statistical models, 2) showing that the well-known MVDR and MWF beamformers result naturally from these statistical models, 3) analyzing and demonstrating the similarities and differences between the two algorithms, and 4) incorporating in one of the methods an a posteriori probability estimator which serves as a soft mask to enhance the separation capabilities.

This paper is organized as follows. The BSS problem formulation and the two models are given in Sec. II. In Secs. III and IV, the EM algorithms are derived for both models. The differences between the models and the respective algorithms are shown in Sec. V, and some practical considerations are given in Sec VI. An experimental study for the two algorithms is presented in Sec. VII, and conclusions are drawn in Sec. VIII.

II. Problem formulation and statistical model

In this section, the problem and models are formulated. In Sec. II-A, general notation and the observation model is described, and the alternative models for the speech signals are described in Secs. II-B and II-C.

A. Notation and observation model

The following model is formulated in the STFT domain, where K and T denote the number of frequency bands and

time-frames, respectively. We assume that D speech sources are concurrently active, but based on the sparsity property, each TF bin is dominated by only one source. The discrete random variable $d(t,k)$ indicates which speaker is active in the (t,k)-th bin, where $d(t,k) \in \mathbb{n}_D$ and $\mathbb{n}_D \triangleq \{0,...,D\}$. Under the d-th hypothesis, i.e. $d(t,k) = d > 0$, only the d-th speaker is present, and under the *null hypothesis*, i.e. $d(t,k) = 0$, no speaker is active. Now, define the set of indicators,

$$\mathcal{D} = \{d(t,k) \ : \ t \in \mathbb{n}_T, \ k \in \mathbb{n}_K\} \ , \tag{1}$$

where $\mathbb{n}_T \triangleq \{1,...,T\}$, and $\mathbb{n}_K \triangleq \{1,...,K\}$. Finally, define the set of clean speech signals by

$$\mathcal{X} \triangleq \{x_d(t,k) : t \in \mathbb{n}_T, k \in \mathbb{n}_K, d \in \mathbb{n}_D\} \ , \tag{2}$$

where by definition $x_0(t,k) = 0$ is used only to simplify the notation.

Using these definitions, and in addition, assuming that the room impulse responses (RIRs) are time-invariant and shorter than the STFT frame, the observation vector is given by,

$$\mathbf{z}(t,k) = \begin{cases} \mathbf{v}(t,k) & ; \quad d(t,k) = 0 \\ \mathbf{v}(t,k) + \mathbf{h}_1(k) \cdot x_1(t,k) & ; \quad d(t,k) = 1 \\ \vdots & \vdots \\ \mathbf{v}(t,k) + \mathbf{h}_D(k) \cdot x_D(t,k) & ; \quad d(t,k) = D \end{cases} \tag{3}$$

where $\mathbf{z}(t,k)$ is a $J \times 1$ vector comprising all J microphone signals, $\mathbf{v}(t,k)$ is a stationary additive noise, and $\mathbf{h}_d(k)$ represent the relative transfer functions (RTFs) between the d-th source and a reference microphone. For more details about the RTF, the reader is referred to [5].

Further denote by $p_d(k) \equiv \Pr\{d(t,k) = d\}$ the time-invariant a priori probability of the d-th hypothesis, satisfying $\sum_{d \in \mathbb{n}_D} p_d(k) = 1$, and the set of observations

$$\mathcal{Z} = \{\mathbf{z}(t,k) \ : \ t \in \mathbb{n}_T, \ k \in \mathbb{n}_K\} \ . \tag{4}$$

The noise signal is modelled as a zero-mean complex-Gaussian random variable,

$$\mathbf{v}(t,k) \sim \mathcal{N}_c\{\mathbf{v}(t,k); \mathbf{0}, \mathbf{R}_v(k)\} \ , \tag{5}$$

where $\mathbf{R}_v(k)$ is the time-invariant covariance matrix of the noise, that may comprise a superposition of several noise sources.

In Secs. III and IV we develop two algorithms to estimate $\mathcal{X}$ from $\mathcal{Z}$ using two alternative models for the speech signals. In Sec. II-B, $\mathcal{X}$ is modelled as a deterministic unknown signal, and in Sec. II-C, as a stochastic complex-Gaussian signal. For the two models, we define the speech enhancement and separation task as a hidden-data problem and develop two EM algorithms to estimate the parameters and the hidden data. The observation set for the two models is $\mathcal{Z}$, but the set of hidden data and the set of unknown parameters differ. In Sec. II-B, $\mathcal{X}$ is treated as part of the parameters set, while in Sec. II-C $\mathcal{X}$ is treated as hidden data.

We conclude this section by defining the set of unknown time-invariant parameters that are common to both models,

$$\boldsymbol{\theta} = \{p_d(k), \ \mathbf{h}_d(k), \ \mathbf{R}_v(k), \ : \ k \in \mathbb{n}_K, \ d \in \mathbb{n}_D\} \ . \tag{6}$$

The notation is summarized in Table. I.

TABLE I
DEFINITIONS - (k IS OMITTED FOR BREVITY)

Sym.	Definition	Value
$d(t)$	d-th hypothesis indicator	
$x_d(t)$	d-th source signal	
$v_j(t)$	j-th microphone noise	
$z_j(t)$	j-th microphone signal	
$\mathbf{v}(t)$	Noise vector	$[v_0(t),...,v_{J-1}(t)]^T$
$\mathbf{z}(t)$	Observation vector	$[z_0(t),...,z_{J-1}(t)]^T$
p_d	d-th a priori probability	
$\mathbf{h}_d$	d-th steering vector	$\left[h_{d,0},...,h_{d,J-1}\right]^T$
$\mathbf{R}_v$	Noise covariance	
$\mathcal{Z}$	Observation set	$\{\mathbf{z}(t) : t \in \mathbb{n}_T\}$
$\mathcal{D}$	Indicators set	$\{d(t) : t \in \mathbb{n}_T\}$
$\mathcal{X}$	Speech STFT coefficients	$\{x_d(t) : t \in \mathbb{n}_T, d \in \mathbb{n}_D\}$
$\boldsymbol{\Phi}$	Speech PSD coefficients	$\{\phi_d(t) : t \in \mathbb{n}_T, d \in \mathbb{n}_D\}$
$\boldsymbol{\theta}$	Common parameter set	$\{p_d, \mathbf{h}_d, \mathbf{R}_v : d \in \mathbb{n}_D\}$
	Parameters and hidden - I	$\overline{\boldsymbol{\theta}} = \{\boldsymbol{\theta}, \mathcal{X}\}$, $\overline{\mathcal{H}} = \mathcal{D}$
	Parameters and hidden - II	$\widetilde{\boldsymbol{\theta}} = \{\boldsymbol{\theta}, \boldsymbol{\Phi}\}$, $\widetilde{\mathcal{H}} = \{\mathcal{D}, \mathcal{X}\}$

B. Model I - Deterministic unknown speech signals

In model I, $\mathcal{X}$ is defined as a set of unknown parameters, hence the observation model (3) is closely related to a GMM. Under this model, the observation $z(t,k)$ belongs to one of $(D+1)$ Gaussians, where the mean of the d-th Gaussian is $\mathbf{h}_d(k) \cdot x_d(t,k)$, and all Gaussians have identical covariance matrix $\mathbf{R}_v(k)$.

The set of parameters is (see (6)) $\overline{\boldsymbol{\theta}} = \{\boldsymbol{\theta}, \mathcal{X}\}$, the hidden data is $\overline{\mathcal{H}} = \mathcal{D}$, namely the association of each TF bin to one of $D+1$ classes, and the probability density function (p.d.f.) of the complete data is

$$f(\mathcal{Z}, \overline{\mathcal{H}}; \overline{\boldsymbol{\theta}}) = \prod_{t,k} \sum_{d} \mathbb{1}_{t,k,d} \cdot p_d(k) \cdot f\left(\mathbf{z}(t,k) \middle| d; \overline{\boldsymbol{\theta}}\right) \ , \tag{7}$$

where $\mathbb{1}_{t,k,d}$ is an indicator random variable that equals one if $d(t,k) = d$ and zero otherwise. In a GMM, $p_d(k)$ is the a priori probability of the d-th Gaussian, which in this case, is a frequency-dependent parameter indicating the probability of a specific source activity. For example, at the harmonic frequencies of the d-th speaker, $p_d(k)$ is expected to be relatively high.

The log-likelihood of the complete data is therefore

$$\log f(\mathcal{Z}, \overline{\mathcal{H}}; \overline{\boldsymbol{\theta}}) = C + \sum_{t,k,d} \mathbb{1}_{t,k,d} \left[\log p_d(k) + \log f\left(\mathbf{z}(t,k) \middle| d; \overline{\boldsymbol{\theta}}\right)\right] \ , \tag{8}$$

where C is a constant that is independent of the parameters, and

$$f\left(\mathbf{z}(t,k) | d(t,k) = d; \overline{\boldsymbol{\theta}}\right) = \mathcal{N}_c\{\mathbf{z}(t,k); \ \mathbf{h}_d(k) \cdot x_d(t,k), \mathbf{R}_v(k)\} \ . \tag{9}$$

In the derivation of (8), we exchanged the logarithm and summation order by using the fact that $\mathbb{1}_{t,k,d}$ is nonzero only for a single hypothesis.

Under this model, the estimation of the various speech sources is obtained by the maximum likelihood (ML) estimation

of $\mathcal{X}$, which is a subset of $\overline{\boldsymbol{\theta}}$. This is achieved by applying a generalized EM (GEM) algorithm, derived in Sec. III.

C. Model II - Stochastic speech signals

As mentioned earlier, an alternative approach is proposed by modelling $\mathcal{X}$ as a set of complex Gaussian random variables. The elements of $\mathcal{X}$ are assumed statistically independent, with zero mean, and different variances

$$x_d(t,k) \sim \mathcal{N}_c\left\{x_d(t,k); 0, \phi_{t,k,d}\right\} \ , \tag{10}$$

where $\phi_{t,k,d}$ is the variance of the d-th speaker at time t and frequency k. Now, the set of unknown parameters is $\widetilde{\boldsymbol{\theta}} = \{\boldsymbol{\theta}, \Phi\}$, where

$$\Phi \triangleq \left\{\phi_{t,k,d} : t \in \mathbb{n}_T, \ k \in \mathbb{n}_K, \ d \in \mathbb{n}_D\right\} \ , \tag{11}$$

and the hidden data set comprises both the indicators set and the speech coefficients, $\widetilde{\mathcal{H}} = \{\mathcal{X}, \mathcal{D}\}$. The p.d.f. of the complete data is now (compare with (7))

$$f(\mathcal{Z}, \widetilde{\mathcal{H}}; \widetilde{\boldsymbol{\theta}}) = \prod_{t,k} \sum_{d} \mathbb{1}_{t,k,d} \cdot p_d(k) \times f\left(\mathbf{z}(t,k), x_d(t,k) \middle| d; \widetilde{\boldsymbol{\theta}}\right) \ . \tag{12}$$

Unlike the model in Sec. II-B, (12) comprises additional hidden data, hence

$$\begin{aligned} &f\left(\mathbf{z}(t,k), x_d(t,k) \middle| d; \widetilde{\boldsymbol{\theta}}\right) \\ &\quad = f\left(\mathbf{z}(t,k) \middle| x_d(t,k), d; \widetilde{\boldsymbol{\theta}}\right) \cdot f\left(x_d(t,k) \middle| d; \widetilde{\boldsymbol{\theta}}\right) \\ &\quad = \mathcal{N}_c\left\{\mathbf{z}(t,k); \ \mathbf{h}_d(k) x_d(t,k), \mathbf{R}_v(k)\right\} \\ &\qquad\qquad \times \mathcal{N}_c\left\{x_d(t,k); \ 0, \phi_{t,k,d}\right\} \ . \end{aligned} \tag{13}$$

By (3), the marginal p.d.f. of $\mathbf{z}(t,k)$ is

$$\begin{aligned} &f(\mathbf{z}(t,k)|d; \widetilde{\boldsymbol{\theta}}) \\ &\quad = \mathcal{N}_c\left\{\mathbf{z}(t,k); \mathbf{0}, \phi_{t,k,d}\mathbf{h}_d(k)\mathbf{h}_d^H(k) + \mathbf{R}_v(k)\right\} \ . \end{aligned} \tag{14}$$

Similarly to (8), the log-likelihood of the complete data is

$$\log f(\mathcal{Z}, \widetilde{\mathcal{H}}; \widetilde{\boldsymbol{\theta}}) = C + \sum_{t,k,d} \mathbb{1}_{t,k,d} \left[\log p_d(k) + \log f\left(\mathbf{z}(t,k), x_d(t,k) \middle| d; \widetilde{\boldsymbol{\theta}}\right)\right] \ . \tag{15}$$

An EM algorithm based on this model is derived in Sec. IV.

III. Algorithm I - Estimation of Deterministic Signals

In this section a GEM algorithm is derived for the model described in Sec. II-B. This algorithm requires a hypothesis testing in the E-step, which is discussed in Sec. III-A, and beamforming and parameter estimation, that are carried out in the M-step and discussed in Sec. III-B. We dub this method GEM for beamforming with posterior (GEMBFP).

For brevity, the frequency index k is omitted whenever no confusion arises, and use $\overline{(\cdot)}^{(\ell)}$ to denote an estimated value that was calculated at the ℓ-th iteration of the GEMBFP algorithm.

A. E-step: Posterior probabilities

In this section, the E-step of the GEMBFP algorithm is derived, which is shown to comprise only the computation of the a posteriori speaker presence probabilities. Since the TF coefficients of the speech signals are considered parameters of the model, they are estimated in the M-step in Sec. III-B.

Consider the $(\ell+1)$-th iteration, where the previous parameter estimate $\overline{\boldsymbol{\theta}}^{(\ell)}$ is available. To derive the equations for the GEM algorithm, we define the auxiliary function, using the complete data p.d.f. in (8),

$$\begin{aligned} Q\left(\overline{\boldsymbol{\theta}} \middle| \overline{\boldsymbol{\theta}}^{(\ell)}\right) &\triangleq E\left\{\log f(\mathcal{Z}, \overline{\mathcal{H}}; \overline{\boldsymbol{\theta}}) \middle| \mathcal{Z}; \overline{\boldsymbol{\theta}}^{(\ell)}\right\} \\ &= \sum_{t,k,d} \overline{w}_d^{(\ell)}(t) \cdot \left[\log p_d(k) + \log f\left(\mathbf{z}(t,k) \middle| d; \overline{\boldsymbol{\theta}}\right)\right] \ , \end{aligned} \tag{16}$$

where $\overline{w}_d^{(\ell)}(t)$ is the a posteriori probability of the d-th hypothesis,

$$\begin{aligned} \overline{w}_d^{(\ell)}(t) &\triangleq \Pr\left\{d(t,k) = d \middle| \mathcal{Z}; \overline{\boldsymbol{\theta}}^{(\ell)}\right\} \\ &= \frac{p_d^{(\ell)}(k) \cdot f\left(\mathbf{z}(t,k) \middle| d; \overline{\boldsymbol{\theta}}^{(\ell)}\right)}{\sum_{d' \in \mathbb{n}_D} p_{d'}^{(\ell)}(k) \cdot f\left(\mathbf{z}(t,k) \middle| d'; \overline{\boldsymbol{\theta}}^{(\ell)}\right)} \ , \end{aligned} \tag{17}$$

and $f(\mathbf{z}(t,k)|d; \overline{\boldsymbol{\theta}})$ is given in (9). Equations (16) and (17) are now used in the M-step, as shown in the next section.

B. M-step: Beamforming and parameters update

In the M-step, the set of parameters is updated by the maximization of (16) with respect to (w.r.t.) each of the elements of $\overline{\boldsymbol{\theta}}$. We start by substituting (9) in (16),

$$\begin{aligned} &Q\left(\overline{\boldsymbol{\theta}} \middle| \overline{\boldsymbol{\theta}}^{(\ell)}\right) \\ &\quad = \sum_{t,d} \overline{w}_d^{(\ell)}(t) \left[\log p_d - \log|\mathbf{R}_v| - \mathbf{e}_d^H(t) \mathbf{R}_v^{-1} \mathbf{e}_d(t)\right] \ , \end{aligned} \tag{18}$$

where

$$\mathbf{e}_d(t) = \begin{cases} \mathbf{z}(t) & ; \quad d = 0 \\ \mathbf{z}(t) - \mathbf{h}_d x_d(t) & ; \quad d > 0 \end{cases} .$$

It is important to note that $\mathbf{e}_d(t)$ and $\mathbf{v}(t)$ are identical under the d-th hypothesis, i.e. $d(t) = d$, but they are not equal under the other hypotheses. The M-step of the $(\ell+1)$-th iteration is carried out by solving

$$\overline{\boldsymbol{\theta}}^{(\ell+1)} = \underset{\overline{\boldsymbol{\theta}}}{\operatorname{argmax}} \, Q\left(\overline{\boldsymbol{\theta}} \middle| \overline{\boldsymbol{\theta}}^{(\ell)}\right). \tag{19}$$

We begin with calculating the derivative of (19) w.r.t. $x_d(t)$ for every $d > 0$ (recall that, by definition, $x_0(t) = 0$),

$$\begin{aligned} &\nabla_{x_d^*(t)} Q\left(\overline{\boldsymbol{\theta}} \middle| \overline{\boldsymbol{\theta}}^{(\ell)}\right) \\ &\quad = \overline{w}_d^{(\ell)}(t) \cdot \mathbf{h}_d^H \cdot \mathbf{R}_v^{-1} \cdot \left(\mathbf{z}(t) - \mathbf{h}_d \cdot x_d(t)\right) \ , \ d > 0 \ , \end{aligned}$$

and by equating the derivative to zero the updated value is obtained

$$\overline{x}_d^{(\ell+1)}(t) = \overline{\mathbf{u}}_d^H\left(\overline{\boldsymbol{\theta}}^{(\ell)}\right) \cdot \mathbf{z}(t) \ , \ d > 0 \ , \tag{20}$$

where

$$\overline{\mathbf{u}}_d\left(\overline{\boldsymbol{\theta}}\right) = \frac{\mathbf{R}_v^{-1}\mathbf{h}_d}{\mathbf{h}_d^H\mathbf{R}_v^{-1}\mathbf{h}_d} \ , \ d>0 \ . \tag{21}$$

A few notes regarding the theory behind (21) are in order. First, since $\overline{x}_d^{(\ell+1)}(t)$ depends on the parameters $\mathbf{h}_d$ and $\mathbf{R}_v$, which are also estimated in the M-step, the entire algorithm is a GEM rather than EM (see Sec. 3.2 in [37]). In practice, we substitute $\mathbf{R}_v$ by $\overline{\mathbf{R}}_v^{(\ell)}$ and $\mathbf{h}_d$ by $\overline{\mathbf{h}}_d^{(\ell)}$ in (21), which denote the estimates of $\mathbf{R}_v$ and $\mathbf{h}_d$ in the ℓ-th iteration. Second, the filter (21) fits the result in [38] (Sec. 6.2.1.2), that the ML estimator of $x_d(t)$ coincides with the MVDR output. Third, when trying to separate different signals as we propose in this paper, one might expect the algorithm to completely cancel the undesired sources, e.g. by a linear constraint minimum variance beamformer. However, since we assumed sparsity in the STFT domain, the MVDR (21) provides the optimal solution that minimizes the noise while maintaining the desired source.

Next, we maximize (16) w.r.t. p_d under the constraint $\sum_{d\in\mathbb{n}_D} p_d = 1$ by writing the Lagrangian

$$\mathcal{L}(\overline{\boldsymbol{\theta}}) = Q\big(\overline{\boldsymbol{\theta}}|\overline{\boldsymbol{\theta}}^{(\ell)}\big) + \lambda\Big(\sum_d p_d - 1\Big) \ , \ \forall d\in\mathbb{n}_d \ . \tag{22}$$

After calculating the derivative w.r.t. p_d and equating to zero we obtain

$$\overline{p}_d^{(\ell+1)} = \frac{1}{T}\sum_t \overline{w}_d^{(\ell)}(t) \ , \ \forall d\in\mathbb{n}_d \ . \tag{23}$$

The derivative of $Q\big(\overline{\boldsymbol{\theta}}|\overline{\boldsymbol{\theta}}^{(\ell)}\big)$ w.r.t. $\mathbf{h}_d^H$ for $d>0$, is

$$\nabla_{\mathbf{h}_d^H} Q\big(\overline{\boldsymbol{\theta}}|\overline{\boldsymbol{\theta}}^{(\ell)}\big) = \sum_t \overline{w}_d^{(\ell)}(t)\ x_d^*(t)\ \mathbf{R}_v^{-1}\ \left[\mathbf{z}(t) - \mathbf{h}_d\cdot x_d(t)\right] \ , \ d>0 \ ,$$

and, by equating to zero, we get $\overline{\mathbf{h}}_d$ that depends on the values of $\overline{x}_d(t)$. As was done in (20), the GEM algorithm is applied instead of the regular EM, and the substitution $x_d(t) = \overline{x}_d^{(\ell+1)}(t)$ is used to obtain

$$\overline{\mathbf{h}}_d^{(\ell+1)} = \frac{\sum_t \overline{w}_d^{(\ell)}(t)\cdot\mathbf{z}(t)\cdot\left[\overline{x}_d^{(\ell+1)}(t)\right]^*}{\sum_t \overline{w}_d^{(\ell)}(t)\cdot\left|\overline{x}_d^{(\ell+1)}(t)\right|^2} \ , \ d>0 \ , \tag{24}$$

which is the weighted least squares estimation of the linear system between $\overline{x}_d^{(\ell+1)}(t)$ and $\mathbf{z}(t)$.

Finally, by calculating the derivative of $Q\big(\overline{\boldsymbol{\theta}}|\overline{\boldsymbol{\theta}}^{(\ell)}\big)$ w.r.t. $\mathbf{R}_v$ and equating to zero, it follows that

$$\overline{\mathbf{R}}_v^{(\ell+1)} = \frac{1}{T}\sum_{t,d}\overline{w}_d^{(\ell)}(t)\cdot\overline{\mathbf{e}}_d^{(\ell+1)}(t)\left[\overline{\mathbf{e}}_d^{(\ell+1)}(t)\right]^H \ , \tag{25}$$

where

$$\overline{\mathbf{e}}_d^{(\ell+1)}(t) = \begin{cases} \mathbf{z}(t) & ; \quad d=0 \\ \mathbf{z}(t) - \overline{\mathbf{h}}_d^{(\ell+1)}\cdot\overline{x}_d^{(\ell+1)}(t) & ; \quad d>0 \end{cases} .$$

The algorithm is summarized on the left-hand side of Table II.

IV. Algorithm II - Estimation of Stochastic Signals

In this section, an EM algorithm is derived for the statistical model described in Sec. II-C. The resultant algorithm requires hypothesis testing and beamforming in the E-step as discussed in Sec. IV-A. The parameters estimation, carried out in the M-step, is discussed in Sec. IV-B. We dub this method EM for beamforming with posterior (EMBFP).

For clarity, we use the convention $\widetilde{(\cdot)}^{(\ell)}$ to denote an estimated value that was calculated at the ℓ-th iteration of the EMBFP algorithm. When applied to a parameter (e.g. $\widetilde{p}_d^{(\ell)}$), it represents an estimate at the ℓ-th iteration, and when applied to a random variable (e.g. $\widetilde{x}_d^{(\ell)}(t)$), it represents an MMSE estimate at the ℓ-th iteration.

A. E-step: Posterior and beamforming

In this section, the E-step of the EMBFP algorithm is derived. As in the GEMBFP algorithm, the a posteriori speaker presence probabilities are computed in the E-step. However, unlike the GEMBFP, the speech signals are considered stochastic signals, which leads to the estimation of their first- and second-order moments in the E-step.

To derive the equations for the EM algorithm, we define the auxiliary function using the log-likelihood of the complete data (15) as,

$$\begin{aligned} Q\big(\widetilde{\boldsymbol{\theta}}|\widetilde{\boldsymbol{\theta}}^{(\ell)}\big) &\triangleq E\left\{\log f(\mathcal{Z},\widetilde{\mathcal{H}};\widetilde{\boldsymbol{\theta}})\middle|\mathcal{Z};\widetilde{\boldsymbol{\theta}}^{(\ell)}\right\} \\ &= \sum_{t,d}\widetilde{w}_d^{(\ell)}(t)\ E\left\{\log f\left(\mathbf{z}(t),x_d(t)\middle|d;\widetilde{\boldsymbol{\theta}}\right)\middle|\mathbf{z}(t);\widetilde{\boldsymbol{\theta}}^{(\ell)}\right\} \\ &\quad + \sum_{t,d}\widetilde{w}_d^{(\ell)}(t)\ p_d \ , \end{aligned} \tag{26}$$

where the a posteriori probability is now given by

$$\begin{aligned} \widetilde{w}_d^{(\ell)}(t) &= \Pr\left\{d_t=d\middle|\mathcal{Z};\widetilde{\boldsymbol{\theta}}^{(\ell)}\right\} \\ &= \frac{p_d\ f\left(\mathbf{z}(t)\middle|d;\widetilde{\boldsymbol{\theta}}^{(\ell)}\right)}{\sum_{d'}p_{d'}\ f\left(\mathbf{z}(t)\middle|d';\widetilde{\boldsymbol{\theta}}^{(\ell)}\right)} \ . \end{aligned} \tag{27}$$

While $\widetilde{w}_d^{(\ell)}(t)$ has the same structure as $\overline{w}_d^{(\ell)}(t)$ (compare (27) and (17)), the resulting posterior probabilities differ due to the different model of the clean speech signals.

Next, the expected value in (26) is calculated by using (13). For $d=0$ we obtain

$$\begin{aligned} &E\left\{\log f\left(\mathbf{z}(t),x_d(t)\middle|d=0;\widetilde{\boldsymbol{\theta}}\right)\middle|\mathbf{z}(t);\widetilde{\boldsymbol{\theta}}^{(\ell)}\right\} \\ &\qquad = C - \log|\mathbf{R}_v| - \mathbf{z}^H(t)\ \mathbf{R}_v^{-1}\ \mathbf{z}(t) \ , \end{aligned} \tag{28}$$

TABLE II
SUMMARY OF THE TWO ALGORITHMS

	GEMBFP	(Speech is a deterministic signal)		**EMBFP**	(Speech is a stochastic process)
Eq.	**Notation**	**Value**	**Eq.**	**Notation**	**Value**
(9)	$f(\mathbf{z}(t)\|d;\overline{\boldsymbol{\theta}})$	$\mathcal{N}_c\left\{\mathbf{z}(t);\mathbf{h}_d\cdot x_d(t),\mathbf{R}_v\right\}$	(14)	$f(\mathbf{z}(t)\|d;\widetilde{\boldsymbol{\theta}})$	$\mathcal{N}_c\left\{\mathbf{z}(t);\mathbf{0},\phi_d(t)\mathbf{h}_d\mathbf{h}_d^H+\mathbf{R}_v\right\}$
(17)	$\overline{w}_d^{(\ell)}(t)$	$\frac{p_d\cdot f\left(\mathbf{z}(t)\middle\|d;\overline{\boldsymbol{\theta}}^{(\ell)}\right)}{\sum_{d'}p_{d'}\cdot f\left(\mathbf{z}(t)\middle\|d';\overline{\boldsymbol{\theta}}^{(\ell)}\right)}$	(27)	$\widetilde{w}_d^{(\ell)}(t)$	$\frac{p_d\cdot f\left(\mathbf{z}(t)\middle\|d;\widetilde{\boldsymbol{\theta}}^{(\ell)}\right)}{\sum_{d'}p_{d'}\cdot f\left(\mathbf{z}(t)\middle\|d';\widetilde{\boldsymbol{\theta}}^{(\ell)}\right)}$
(20)	$\overline{x}_d^{(\ell+1)}(t)$	$\overline{\mathbf{u}}_d^{(\ell)H}\cdot\mathbf{z}(t)$	(30)	$\widetilde{x}_d^{(\ell)}(t)$	$\widetilde{w}_d^{(\ell)}(t)\cdot\widetilde{\mathbf{u}}_d^{(\ell)H}(t)\cdot\mathbf{z}(t)$
(21)	$\overline{\mathbf{u}}_d^{(\ell)}$	$\frac{\mathbf{R}_v^{-1}\mathbf{h}_d}{\mathbf{h}_d^H\mathbf{R}_v^{-1}\mathbf{h}_d}\Big\|_{\overline{\boldsymbol{\theta}}=\overline{\boldsymbol{\theta}}^{(\ell)}}$	(32)	$\widetilde{\mathbf{u}}_d^{(\ell)}(t)$	$\frac{\mathbf{R}_v^{-1}\mathbf{h}_d}{\mathbf{h}_d^H\mathbf{R}_v^{-1}\mathbf{h}_d}\cdot\frac{\phi_d(t)}{\phi_d(t)+\zeta_d(t)}\Big\|_{\widetilde{\boldsymbol{\theta}}=\widetilde{\boldsymbol{\theta}}^{(\ell)}}$
			(33)	$\widetilde{\lvert x_d(t)\rvert^2}^{(\ell)}$	$\left\lvert\widetilde{x}_d^{(\ell)}(t)\right\rvert^2+\zeta_d(t)\cdot\frac{\phi_d(t)}{\phi_d(t)+\zeta_d(t)}\Big\|_{\widetilde{\boldsymbol{\theta}}=\widetilde{\boldsymbol{\theta}}^{(\ell)}}$
(23)	$\overline{p}_d^{(\ell+1)}$	$\frac{1}{T}\sum_t\overline{w}_d^{(\ell)}(t)$	(36)	$\widetilde{p}_d^{(\ell+1)}$	$\frac{1}{T}\sum_t\widetilde{w}_d^{(\ell)}(t)$
(24)	$\overline{\mathbf{h}}_d^{(\ell+1)}$	$\frac{\sum_t\overline{w}_d^{(\ell)}(t)\cdot\mathbf{z}(t)\cdot\overline{x}_d^{*(\ell+1)}(t)}{\sum_t\overline{w}_d^{(\ell)}(t)\cdot\left\lvert\overline{x}_d^{(\ell+1)}(t)\right\rvert^2}$	(37)	$\widetilde{\mathbf{h}}_d^{(\ell+1)}$	$\frac{\sum_t\widetilde{w}_d^{(\ell)}(t)\cdot\mathbf{z}(t)\widetilde{x}_d^{*(\ell)}(t)}{\sum_t\widetilde{w}_d^{(\ell)}(t)\ \widetilde{\lvert x_d(t)\rvert^2}^{(\ell)}}$
(25)	$\overline{\mathbf{R}}_v^{(\ell+1)}$	$\frac{1}{T}\sum_{t,d}\overline{w}_d^{(\ell)}(t)\cdot\overline{\mathbf{e}}_d^{(\ell+1)}(t)\overline{\mathbf{e}}_d^{H(\ell+1)}(t)$	(38)	$\widetilde{\mathbf{R}}_v^{(\ell+1)}$	$\frac{1}{T}\sum_{t,d}\widetilde{w}_d^{(\ell)}(t)\cdot\widetilde{\mathbf{e}_d(t)\mathbf{e}_d^H}^{(\ell)}(t)$
			(41)	$\widetilde{\phi}_d^{(\ell+1)}(t)$	$\widetilde{w}_d^{(\ell)}(t)\cdot\ \widetilde{\lvert x_d(t)\rvert^2}^{(\ell)}$

and for $d>0$,

$$
\begin{aligned}
&E\left\{\log f\left(\mathbf{z}(t),x_d(t)\middle|d>0;\widetilde{\boldsymbol{\theta}}\right)\middle|\mathbf{z}(t);\widetilde{\boldsymbol{\theta}}^{(\ell)}\right\} \qquad (29)\\
&=C-\log\phi_d(t)-\ \widetilde{\lvert x_d(t)\rvert^2}^{(\ell)}\ \phi_d^{-1}(t)-\log\lvert\mathbf{R}_v\rvert\\
&\quad-\mathbf{z}^H(t)\ \mathbf{R}_v^{-1}\ \mathbf{z}(t)+\mathbf{z}^H(t)\ \mathbf{R}_v^{-1}\ \mathbf{h}_d\ \widetilde{x}_d^{(\ell)}(t)\\
&\quad+\left[\widetilde{x}_d^{(\ell)}(t)\right]^*\ \mathbf{h}_d^H\ \mathbf{R}_v^{-1}\ \mathbf{z}(t)-\mathbf{h}_d^H\ \mathbf{R}_v^{-1}\ \mathbf{h}_d\ \ \widetilde{\lvert x_d(t)\rvert^2}^{(\ell)}\ .
\end{aligned}
$$

In the following, $\widetilde{x}_d^{(\ell)}(t)$ and $\widetilde{\lvert x_d(t)\rvert^2}^{(\ell)}$ are calculated for every $d>0$ (for $d=0$ their value is zero by definition). The first-order moments are,

$$
\begin{aligned}
\widetilde{x}_d^{(\ell)}(t)&=\widetilde{w}_d^{(\ell)}(t)\cdot E\left\{x_d(t)\middle|\mathbf{z}(t),d;\widetilde{\boldsymbol{\theta}}^{(\ell)}\right\} \qquad (30)\\
&=\widetilde{w}_d^{(\ell)}(t)\cdot\widetilde{\mathbf{u}}_d^H(t)\mathbf{z}(t)\ ,\ d>0
\end{aligned}
$$

where $\widetilde{\mathbf{u}}_d(t)$ is the MWF,

$$
\begin{aligned}
\widetilde{\mathbf{u}}_d(t)=E^{-1}&\left\{\mathbf{z}(t)\ \mathbf{z}^H(t)\middle|d;\widetilde{\boldsymbol{\theta}}^{(\ell)}\right\}\\
&\times E\left\{\mathbf{z}(t)\ x_d^*(t)\middle|d;\widetilde{\boldsymbol{\theta}}^{(\ell)}\right\}\ ,\ d>0\ . \qquad (31)
\end{aligned}
$$

The MWF can be decomposed into an MVDR beamformer and a subsequent single-channel Wiener filter (SWF) [39],

$$
\widetilde{\mathbf{u}}_d(t)=\overline{\mathbf{u}}_d\left(\boldsymbol{\theta}\right)\ \frac{\phi_d(t)}{\phi_d(t)+\zeta_d(t)}\Big|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}^{(\ell)}}\ ,\ d>0\ , \qquad (32)
$$

where $\zeta_d(t)=\left(\mathbf{h}_d^H\mathbf{R}_v^{-1}\mathbf{h}_d\right)^{-1}$ denotes the power of the residual noise at the output of the MVDR.

The second-order moments are given by

$$
\begin{aligned}
\widetilde{\lvert x_d(t)\rvert^2}^{(\ell)}&=\left\lvert\widetilde{x}_d^{(\ell)}(t)\right\rvert^2\\
&\quad+E\left\{\lvert x_d(t)-\widetilde{x}_d(t)\rvert^2\middle|d;\widetilde{\boldsymbol{\theta}}^{(\ell)}\right\}\ ,\ d>0\ , \qquad (33)
\end{aligned}
$$

where

$$
\begin{aligned}
&E\left\{\lvert x_d(t)-\widetilde{x}_d(t)\rvert^2\middle|d;\widetilde{\boldsymbol{\theta}}^{(\ell)}\right\} \qquad (34)\\
&\quad=\phi_d(t)-\phi_d(t)\ \mathbf{h}_d^H\ \widetilde{\mathbf{u}}_d(t)\big|_{\widetilde{\boldsymbol{\theta}}=\widetilde{\boldsymbol{\theta}}^{(\ell)}}\\
&\quad=\frac{\phi_d(t)}{1+{}^{\phi_d(t)}/_{\zeta_d(t)}}\Big|_{\widetilde{\boldsymbol{\theta}}=\widetilde{\boldsymbol{\theta}}^{(\ell)}}\ ,\ d>0\ ,
\end{aligned}
$$

is the error covariance after applying the MWF.

B. M-step: Parameters estimation

Similarly to Sec. III-B, the M-step is carried out via the maximization of $Q\left(\widetilde{\boldsymbol{\theta}}|\widetilde{\boldsymbol{\theta}}^{(\ell)}\right)$, i.e.,

$$
\widetilde{\boldsymbol{\theta}}^{(\ell+1)}=\underset{\widetilde{\boldsymbol{\theta}}}{\operatorname{argmax}}\,Q\left(\widetilde{\boldsymbol{\theta}}|\widetilde{\boldsymbol{\theta}}^{(\ell)}\right)\ , \qquad (35)
$$

and the a priori probabilities are calculated as in (23),

$$
\widetilde{p}_d^{(\ell+1)}=\frac{1}{T}\sum_t\widetilde{w}_d^{(\ell)}(t)\ ,\ \forall d\in\mathbb{n}_d\ . \qquad (36)
$$

Calculating the derivative of $Q\left(\widetilde{\boldsymbol{\theta}}|\widetilde{\boldsymbol{\theta}}^{(\ell)}\right)$ w.r.t. $\mathbf{h}_d$ and equating to zero results in

$$
\widetilde{\mathbf{h}}_d^{(\ell+1)}=\frac{\sum_t\widetilde{w}_d^{(\ell)}(t)\cdot\mathbf{z}(t)\cdot\widetilde{x}_d^{*(\ell)}(t)}{\sum_t\widetilde{w}_d^{(\ell)}(t)\cdot\ \widetilde{\lvert x_d(t)\rvert^2}^{(\ell)}}\ ,\ d>0\ , \qquad (37)
$$

and the same procedure for $\mathbf{R}_v$ implies

$$\widetilde{\mathbf{R}}_v^{(\ell+1)} = \frac{1}{T}\sum_{t,d} \widetilde{w}_d^{(\ell)}(t)\cdot \widetilde{\mathbf{e}_d(t)\mathbf{e}_d^H(t)}^{(\ell)} , \tag{38}$$

where

$$\widetilde{\mathbf{e}_d(t)\mathbf{e}_d^H(t)}^{(\ell)} = \mathbf{z}(t)\cdot\mathbf{z}^H(t) - 2\Re\left\{\mathbf{h}_d\cdot\widetilde{x}_d^{(\ell)}(t)\cdot\mathbf{z}^H(t)\right\} + \widetilde{|x_d(t)|^2}^{(\ell)}\cdot\widetilde{\mathbf{h}}_d^{(\ell+1)}\cdot\left[\widetilde{\mathbf{h}}_d^{(\ell+1)}\right]^H , \; d>0 , \tag{39}$$

and for $d=0$ the term is simply

$$\widetilde{\mathbf{e}_0(t)\mathbf{e}_0^H(t)}^{(\ell)} = \mathbf{z}(t)\cdot\mathbf{z}^H(t) . \tag{40}$$

Finally, the power spectral density (PSD) of the d-th speaker is estimated by

$$\widetilde{\phi}_d^{(\ell+1)}(t) = \widetilde{w}_d^{(\ell)}(t)\cdot \widetilde{|x_d(t)|^2}^{(\ell)} , \; d>0 , \tag{41}$$

which is the empirical PSD conditioned on the a posteriori probability of the d-th hypothesis. The algorithm is summarized on the right side of Table II.

V. Comparison and Discussion

The differences between the two models and the resulting algorithms are highlighted in Sec. V-A, and an interpretation of the posteriors $\overline{w}_d^{(\ell)}(t)$ and $\widetilde{w}_d^{(\ell)}(t)$ is given in Sec. V-B.

A. Comparison of the GEMBFP and EMBFP the algorithms

The difference between the beamformers $\overline{\mathbf{u}}_d$ (21) and $\widetilde{\mathbf{u}}_d(t)$ (32) is explained w.r.t. the underlying models. Under model I, $x_d(t)$ is a deterministic unknown and therefore its ML estimator is the output of the MVDR beamformer $\overline{\mathbf{u}}_d^H\cdot\mathbf{z}(t)$. However, when $x_d(t)$ is assumed a random process, the MMSE estimate is required, which is obtained by the MWF (32) that consists of an MVDR followed by an SWF.

In addition, the signal estimator in the EMBFP includes the a posteriori speaker probability (30), while in the GEMBFP the posterior is part of the statistics estimation, but not the signal estimator (20). The a posteriori speaker probability in (30) contributes to the aggressiveness of the attenuation of the noise and the interference. While the EMBFP enhances the signal by concatenating the MVDR beamformer, the SWF, and the a posteriori speaker probability, the GEMBFP solely comprises the MVDR beamformer.

In this context, we observe an important property of the EMBFP algorithms emerging when consequent iterations are executed. By (32), the value of $\phi_d(t)$ from the previous iteration is substituted in the SWF formula, which directly affects the next value of $\phi_d(t)$ (see (33) and (41)), constituting a strong positive feedback. This feedback is not necessarily bad, since noise and competing speakers are further attenuated with every iteration. However, this property affects also TF bins where the desired source PSD has medium to small values, and the desired signal's estimate becomes sparse and distorted. Note that the GEMBFP algorithm does not exhibit such behavior. Although $\overline{x}_d^{(\ell)}$ is fed back to the algorithm, it is done via (17) which is not necessarily low for small values of $\overline{x}_d^{(\ell)}$.

In order to tradeoff between the EMBFP's signal distortion and noise and interference reduction, the mathematical formulation of the problem is adjusted. We add a constraint to the original ML approach, i.e. that $\phi_d(t)$ is higher than ξ_{min}, a pre-defined minimum value. As shown in Sec. V.B in [40], the previous EM is replaced by a constrained-EM algorithm that possesses similar convergence characteristics. In practice, it is better from computational and performance aspects to set a minimum level constraint on the SWF, i.e.

$$\frac{\phi_d(t)}{\phi_d(t)+\zeta_d(t)} > \xi_{\text{min}} \quad \forall d,t , \tag{42}$$

which is mathematically equivalent to applying a constraint on $\phi_d(t)$, and will be used in the following sections.

Comparison of the algorithms' performance as well as the practical choice of ξ_{min} are discussed in Sec. VII-B.

B. Spatial interpretation of the posterior probability

In this section, we give a spatial interpretation to the posterior probability and show that $\widetilde{w}_d(t)$ obtains high value if a signal impinges the array from the direction $\mathbf{h}_d$. Although we mainly discuss $\widetilde{w}_d(t)$, the same interpretation applies to $\overline{w}_d(t)$ *mutatis mutandis*. The following analysis links the algorithms proposed in this paper to the analytic result in [32].

The posterior $\widetilde{w}_d(t)$ in (27) depends on p_d and $f\left(\mathbf{z}(t)\middle| d;\widetilde{\boldsymbol{\theta}}^{(\ell)}\right)$ in (14) for all $d\in\mathbb{n}_D$. As the spatial properties of the signal $\mathbf{z}(t)$ are manifested by its propagation model, we ignore in our analysis its temporal properties and express (14) as a function of an arbitrary time-invariant propagation vector $\mathbf{s}$,

$$f(\mathbf{s}|d;\widetilde{\boldsymbol{\theta}}) = (2\pi)^{-J}\left|\mathbf{R}_d(t)\right|^{-1}\exp\left(-\mathbf{s}^H\cdot\mathbf{R}_d^{-1}(t)\cdot\mathbf{s}\right) , \tag{43}$$

where

$$\mathbf{R}_d(t) = \phi_d(t)\cdot\mathbf{h}_d\mathbf{h}_d^H + \mathbf{R}_v . \tag{44}$$

In the following, we assume that

1) The dominant eigenvector (EV) of $\mathbf{R}_d(t)$ is parallel to the ATF of the d-th signal, and hence can be written as $\alpha\mathbf{h}_d$, where α is an arbitrary complex scalar.
2) The a priori probabilities in (27) are equal, i.e. $p_d = 1/(D+1)$, $\forall d\in\mathbb{n}_D$.

Note that Assumption 1 is satisfied if the noise is spatially white or if the signal-to-noise ratio (SNR) at the specific TF bin is high, i.e. $\phi_d(t)\gg\|\mathbf{R}_v\|_2^2$. In addition, Assumption 2 is relaxed at the end of this section.

Let $C_d=\|\mathbf{h}_d\|^2$ be the norm of the RTF of the d-th speaker, then by Assumption 1,

$$\mathbf{h}_d = \underset{\mathbf{s}}{\text{argmax}}\ \mathbf{s}^H\cdot\mathbf{R}_d(t)\cdot\mathbf{s} \ , \ \text{s.t.}\ \ \|\mathbf{s}\|^2 = C_d . \tag{45}$$

Now, since $\alpha\mathbf{h}_d$ is the most significant EV of $\mathbf{R}_d(t)$, then it is also the least significant EV of $\mathbf{R}_d^{-1}(t)$, and consequently

$$\mathbf{h}_d = \underset{\mathbf{s}}{\text{argmax}}\, f(\mathbf{s}|d;\widetilde{\boldsymbol{\theta}}) \ , \ \text{s.t.}\ \ \|\mathbf{s}\|^2 = C_d . \tag{46}$$

To conclude this discussion, it follows from (27) and Assumption 2 that the global maximum of $\widetilde{w}_d(t)$ is obtained if the input signal propagates from direction $\mathbf{h}_d$. If the null hypothesis is true, i.e. $d=0$, then $\widetilde{w}_0(t)$ is not expected to exhibit any directional preference.

Since $f(\mathbf{s}|d;\widetilde{\boldsymbol{\theta}})$ is an exponential function of $\mathbf{s}$ and a linear function of p_d, it is most likely that this result will hold even when Assumption 2 is not fully satisfied. Moreover, Assumption 1 holds for high and moderate SNR values in TF bins dominated by speaker d, even if the noise is not spatially-white.

VI. Practical aspects

In this section we discuss several practical issues that should be considered when applying the algorithms proposed in Secs. II-B and II-C to actual speech signals.

A. Parameter initialization

The initialization of parameters is an important block in any application of the EM algorithm. In this paper, we address a blind separation problem, where no information about the speakers' activity is available, i.e. we do not know which source is active in each TF bin. This makes the parameter initialization a challenging task. In the following, we propose a simple yet effective technique for initialization, which is used to initialize both the GEMBFP and the EMBFP algorithms.

In the first iteration, the clean sources are initialized according to (30)

$$\widetilde{x}_d^{(0)}(t)=\widetilde{w}_d^{(0)}(t)\cdot\left(\widetilde{\mathbf{u}}_d(t)^{(0)}\right)^H\cdot\mathbf{z}(t)\ , \tag{47}$$

where $\widetilde{\mathbf{u}}_d(t)^{(0)}$ is calculated according to (32), using the initial values of the parameters, denoted by $\widetilde{\mathbf{R}}_v^{(0)}$ and $\widetilde{\mathbf{h}}_d^{(0)}$, and assuming $\zeta_d(t)\ll\phi_d(t)$, i.e., the SWF is approximately one in this iteration. The a posteriori coefficients are initialized by

$$\widetilde{w}_d^{(0)}(t)=\frac{p_d^{(0)}\cdot f^{(0)}\left(\mathbf{z}(t)\middle|d;\widetilde{\boldsymbol{\theta}}^{(0)}\right)}{\sum_{d'\in\mathbb{n}_D}p_{d'}^{(0)}\cdot f^{(0)}\left(\mathbf{z}(t)\middle|d';\widetilde{\boldsymbol{\theta}}^{(0)}\right)}\ , \tag{48}$$

where we have used

$$\begin{aligned}&f^{(0)}\left(\mathbf{z}(t)\middle|d;\widetilde{\boldsymbol{\theta}}^{(0)}\right)\\&\quad=\mathcal{N}_c\Big(\mathbf{z}(t)\,;\mathbf{0},\widetilde{\mathbf{h}}_d^{(0)}\left[\widetilde{\mathbf{h}}_d^{(0)}\right]^H+\widetilde{\mathbf{R}}_v^{(0)}\Big)\ .\end{aligned} \tag{49}$$

Now, the initial values $\widetilde{\mathbf{R}}_v^{(0)}$ and $\widetilde{\mathbf{h}}_d^{(0)}$ are required. The noise covariance is simply initialized by the identity matrix,

$$\widetilde{\mathbf{R}}_v^{(0)}=\mathbf{I}\ . \tag{50}$$

For the initialization of the RTFs $\widetilde{\mathbf{h}}_d^{(0)}$, we estimate the long-term covariance matrix of the input

$$\widehat{\mathbf{R}}=\frac{1}{T}\sum_{t\in\mathbb{n}_T}\mathbf{z}(t)\mathbf{z}^H(t)\ . \tag{51}$$

Then, we calculate the EV decomposition of $\widehat{\mathbf{R}}$, and denote by $\{\lambda_1,...,\lambda_J\}$ eigenvalues in descending order, and by $\{\mathbf{q}_1,...,\mathbf{q}_J\}$ the respective EVs. Since $\widehat{\mathbf{R}}$ is calculated from the entire mixture signals, it comprises average spatial information from all sources. Initializing the RTFs with the D leading eigenvectors, i.e. $\overline{\mathbf{h}}_d^{(0)}=\mathbf{q}_d$ for all $d=1,...,D$, and substituting in (49) results in

$$\begin{aligned}&f^{(0)}\left(\mathbf{z}(t)\middle|d;\widetilde{\boldsymbol{\theta}}^{(0)}\right)\\&\qquad\propto\exp\left(-\mathbf{z}^H(t)\left[\mathbf{q}_d\mathbf{q}_d^H+\mathbf{I}\right]^{-1}\mathbf{z}(t)\right)\ .\end{aligned} \tag{52}$$

This choice, however, may result in a strong bias, as demonstrated in the sequel. Using the Woodbury identity, it can be easily verified that $\left[\mathbf{q}_d\mathbf{q}_d^H+\mathbf{I}\right]^{-1}=\left[\mathbf{I}-\frac{1}{2}\mathbf{q}_d\mathbf{q}_d^H\right]$, and the argument of the exponent in (52) becomes,

$$\begin{aligned}&-\mathbf{z}^H(t)\left[\mathbf{q}_d\mathbf{q}_d^H+\mathbf{I}\right]^{-1}\mathbf{z}(t)\\&\qquad=-\mathbf{z}(t)^H\mathbf{z}(t)+\frac{1}{2}\mathbf{q}_d^H\left[\mathbf{z}(t)\mathbf{z}^H(t)\right]\mathbf{q}_d\ .\end{aligned} \tag{53}$$

Since $\mathbf{q}_d$, $d=1,...,D$ are the EVs of $\widehat{\mathbf{R}}$, the time-average of the term $\mathbf{q}_d^H\left[\mathbf{z}(t)\mathbf{z}^H(t)\right]\mathbf{q}_d$ is the respective eigenvalue λ_d. Since $\lambda_1,...,\lambda_D$ are in descending order, it follows that $f^{(0)}\left(\mathbf{z}(t)\middle|d;\widetilde{\boldsymbol{\theta}}^{(0)}\right)$, $d=1,...,D$ are, on average, organized in descending order as well, and due to the exponential function can exhibit very large differences in their values. This may imply a strong bias in the initialization procedure towards the first eigenvector and consequently, the ability of the proposed algorithm to separate the signals will significantly deteriorate. Preliminary tests verifies this observation.

To circumvent these bias effects, we propose an alternative initialization. Rather than using the eigenvectors of $\widehat{\mathbf{R}}$, we use a set of vectors $\left\{\widetilde{\mathbf{h}}_1^{(0)},...,\widetilde{\mathbf{h}}_D^{(0)}\right\}$, satisfying the following conditions,

$$\begin{aligned}\left[\widetilde{\mathbf{h}}_d^{(0)}\right]^H\widetilde{\mathbf{h}}_{d'}^{(0)}&=0\quad,\quad\forall d\neq d'\\\left[\widetilde{\mathbf{h}}_d^{(0)}\right]^H\cdot\widehat{\mathbf{R}}\cdot\widetilde{\mathbf{h}}_d^{(0)}&=1\quad,\quad 1\le d\le D.\end{aligned} \tag{54}$$

Note that these vectors are not necessarily EVs of $\widehat{\mathbf{R}}$, but can however be directly obtained from the first D EVs, as will be later demonstrated. The underlying intuition is that 1) the orthogonality property assures that the vectors are distinctive, and 2) the equal *energy* distribution circumvents the bias discussed above. A general formula for obtaining $\left\{\widetilde{\mathbf{h}}_1^{(0)},...,\widetilde{\mathbf{h}}_D^{(0)}\right\}$ from $\{\mathbf{q}_1,...,\mathbf{q}_D\}$ is beyond the scope of this paper, and here we consider the case of $D=2$ as an example. In this case, the initial RTFs are chosen as

$$\widetilde{\mathbf{h}}_1^{(0)}=\frac{\mathbf{q}_1+\mathbf{q}_2}{\sqrt{\lambda_1+\lambda_2}}\ ,\ \widetilde{\mathbf{h}}_2^{(0)}=\frac{\mathbf{q}_1-\mathbf{q}_2}{\sqrt{\lambda_1+\lambda_2}}\ . \tag{55}$$

It can be easily verified that the constraint set in (54) is satisfied. Finally, the M-step and the subsequent iterations are carried out according to the respective procedure for each algorithm.

B. Robust and efficient computation of the posterior

Express (14) as

$$f(\mathbf{z}(t)|d;\widetilde{\boldsymbol{\theta}})\sim|\mathbf{R}_d(t)|^{-1}\cdot\exp(-\xi_{t,d})\ , \tag{56}$$

where $\xi_{t,d} = \mathbf{z}^H(t) \cdot \mathbf{R}_d^{-1}(t) \cdot \mathbf{z}(t)$, and rewrite (27) as

$$\widetilde{w}_d^{(\ell)}(t) = \frac{p_d \cdot |\mathbf{R}_d(t)|^{-1} \cdot \exp(-\xi_{t,d})}{\sum_{d'} p_{d'} \cdot |\mathbf{R}_{d'}(t)|^{-1} \cdot \exp(-\xi_{t,d'})} \ . \quad (57)$$

For large values of $\xi_{t,d}$, the term $\exp(-\xi_{t,d})$ may exceed the precision limitation. We therefore multiply the numerator and denominator of (57) by $\exp(\xi_{t,\max})$ where $\xi_{t,\max} = \max_d \{\xi_{t,d}\}$,

$$\widetilde{w}_d^{(\ell)}(t) = \frac{p_d \cdot |\mathbf{R}_d(t)|^{-1} \cdot \exp(-\xi_{t,d} + \xi_{t,\max})}{\sum_{d'} p_{d'} \cdot |\mathbf{R}_{d'}(t)|^{-1} \cdot \exp(-\xi_{t,d'} + \xi_{t,\max})} \ , \quad (58)$$

hence significantly reducing the argument in the exponent. Identical procedure can be applied to calculate $\overline{w}_d^{(\ell)}(t)$.

Another practical issue is the inversion of $\mathbf{R}_d(t)$, which is computationally expensive. By the Woodbury identity [41],

$$\mathbf{R}_d^{-1}(t) = \mathbf{R}_v^{-1} - \frac{\mathbf{R}_v^{-1}\mathbf{h}_d\mathbf{h}_d^H\mathbf{R}_v^{-1}}{\phi_d^{-1}(t) + \mathbf{h}_d^H\mathbf{R}_v^{-1}\mathbf{h}_d} \ , \quad (59)$$

where $\mathbf{R}_v^{-1}$, $\mathbf{R}_v^{-1}\mathbf{h}_d\mathbf{h}_d^H\mathbf{R}_v^{-1}$, and $\mathbf{h}_d^H\mathbf{R}_v^{-1}\mathbf{h}_d$ are calculated only once every iteration, reducing the computation to a single (real) division at each time frame. Note that the calculation of $\overline{w}_d^{(\ell)}(t)$ requires only one matrix inversion per iteration, because the covariance matrix in (9) is $\mathbf{R}_v$, which is time invariant.

C. Permutation ambiguity

The models and algorithms described in Secs. II-III apply an identical procedure in each frequency band. However, when the output signals are reconstructed, one needs to validate that all separated sources are aligned across the frequency range and no permutation occurs. In this work, we resolved the frequency alignment problem based on correlation between the signals at the different frequency bands, as proposed in [42] and [43].

VII. Performance Evaluation

In this section we describe the experiments carried out to evaluate the proposed algorithms. In Sec. VII-A we describe the setup, the reference algorithms, and the quality measures. In Sec. VII-B, we present the performance of the proposed algorithms w.r.t. the reverberation level, noise level, and compare them to each other. A comparison between the proposed algorithms and a reference method is given in Sec. VII-C.

A. Experimental setup

1) Signals and RIRs: The algorithms developed in this paper were evaluated by signals with two concurrently active speakers ($D = 2$) and diffuse noise. Clean anechoic speech signals were randomly drawn from the Wall Street Journal (WSJ) corpus, each of length 30 seconds. For RIRs we used the database that was presented in [44], which was recorded in the Speech & Acoustic Lab of the Faculty of Engineering at Bar-Ilan University[1], with controllable reverberation time. The room dimensions are $6 \times 6 \times 2.4$ m (length $\times$ width $\times$ height),

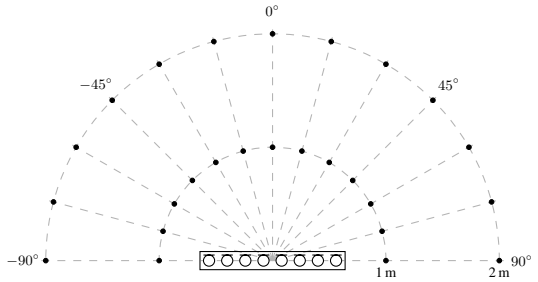


Fig. 1. Microphone-array setup.

and the tested reverberation times (T_{60}) are 0.16, 0.36, and 0.61 s. The RIRs were captured by an eight-microphone linear array with inter-distances of $\{3, 3, 3, 8, 3, 3, 3\}$ cm from one another, as depicted in Fig. 1. For the additive noise, we used a random signal with speech-like spectrum, and applied the method described in [45] to produce diffuse noise field. The speakers were positioned at different locations in the room, where the distances from the array are 1 or 2 m semi-circle with a grid of 15 degrees. The total number of signals used is 135, which resulted in a total of 1.13 hours of audio.

2) Tested algorithms: In Sec. VII-B we evaluate and compare the performance of the algorithms proposed in Secs. III and IV.

We compared the proposed algorithms with the method presented in [27], which is based on the NMF of the speech spectra. Since the algorithm in [27] was originally developed for a two-channels setup, a multi-channel extension was used for the experiment in Sec. VII-C.[2] The NMF algorithm was initialized by using the separated source signals, where each source is corrupted by the other sources with signal-to-interference ratio (SIR) of 10 dB (see [27]).

In addition, an oracle version of the GEMBFP was applied for the experiment described in Sec. VII-B. Explicitly, we carried out the following steps, dubbed as oracle beamformer (BF) with posterior (OBFP):

- Estimate $\mathbf{h}_d$ from $\{x_d(t) : t \in \mathbb{n}_T\}$ and $\{\mathbf{z}(t) : t \in \mathbb{n}_T\}$, using a least-squares identification.
- Estimate $\mathbf{R}_v$ directly from $\{\mathbf{v}(t) : t \in \mathbb{n}_T\}$.
- Calculate $\overline{\mathbf{u}}_d$ by (21).
- Apply (17) to calculate $\overline{w}_d^{(\ell)}(t)$.
- Calculate $\overline{x}_d(t)$ according to (20) using $\overline{\mathbf{u}}_d^H$ and $\overline{w}_d^{(\ell)}(t)$.

The STFT analysis window length is a crucial parameter for the correctness of the statistical model in Sec. II. Specifically, in the multichannel signal model (3) we assume that the analysis window is longer than the length of the RIRs. However, according to our experience it is sufficient to use analysis window of length $T_{60}/2$.

3) Computational load: The algorithms were implemented in MATLAB, on an Intel Core i7-3770 CPU at 3.4 GHz with four cores, and using 8 GB of RAM. Since most of the computation is done independently for each frequency band,

[1]The RIRs database is available at http://www.eng.biu.ac.il/gannot/downloads

[2]We thank Dr. D. Kounades-Bastian for his kind help in developing and sharing this version.

computation can be parallelized, with every core executing the computation for different frequency bands. Every iteration of the EMBFP requires 4.88 seconds of computation to process 10 seconds of eight-channel signal sampled at 16 kHz (and similar computation is required for the GEMBFP algorithm). In this experimental study, we used 10 iterations for each algorithm, thus 48.8 seconds were required to process 10 seconds of signal.

4) Quality measures: As in [16], [46], [47], we computed the reverberant-signal to noise ratio (RSNR), SIR, and signal-to-distortion ratio (SDR) for the input and output signals. The three quality measures (QMs) were calculated by the same formula, i.e.

$$\mathrm{QM}\left\{x_d,\eta\right\} \triangleq \frac{\sum_{t,k}\left|x_d(t,k)\right|^2}{\sum_{t,k}\left|\eta(t,k)\right|^2}, \tag{60}$$

where only η changes from one measure to another. The input RSNR and SIR were calculated by

$$\mathrm{RSNR}_d^{\mathrm{in}} = \mathrm{QM}\left\{x_d, v\right\} \quad , \quad \mathrm{SIR}_{d,d'}^{\mathrm{in}} = \mathrm{QM}\left\{x_d, x_{d'}\right\}, \tag{61}$$

where $d, d' \in \{1,2\}$, and $\mathrm{SDR}_d^{\mathrm{in}} = \infty$ by definition. At the output, we applied the estimated filter (obtained blindly from the mixture) to the separated signals as follows. For example, when evaluating theEMBFP, we applied (30) to the clean speech and interference signals separately, i.e.

$$\widetilde{s}_{d,v}(t,k) = \left[\widetilde{w}_d^{(\ell)}(t)\cdot\widetilde{\mathbf{u}}_d(t)^{(\ell)}\right]^H \cdot \mathbf{v}(t,k) \ , \tag{62}$$

$$\widetilde{s}_{d,d'}(t,k) = \left[\widetilde{w}_d^{(\ell)}(t)\cdot\widetilde{\mathbf{u}}_d(t)^{(\ell)}\right]^H \cdot \left[\mathbf{h}_{d'}(k)\cdot x_{d'}(t,k)\right], \tag{63}$$

$$\widetilde{s}_{d,d}(t,k) = \left[\widetilde{w}_d^{(\ell)}(t)\cdot\widetilde{\mathbf{u}}_d(t)^{(\ell)}\right]^H \cdot \left[\mathbf{h}_{d}(k)\cdot x_{d}(t,k)\right] \ , \tag{64}$$

therefore the RSNR, SIR, and SDR at the output are given by

$$\mathrm{RSNR}_d^{\mathrm{out}} = \mathrm{QM}\left\{\widetilde{s}_{d,d}, \widetilde{s}_{d,v}\right\} \ , \tag{65}$$

$$\mathrm{SIR}_d^{\mathrm{out}} = \mathrm{QM}\left\{\widetilde{s}_{d,d}, \widetilde{s}_{d,d'}\right\} \ , \tag{66}$$

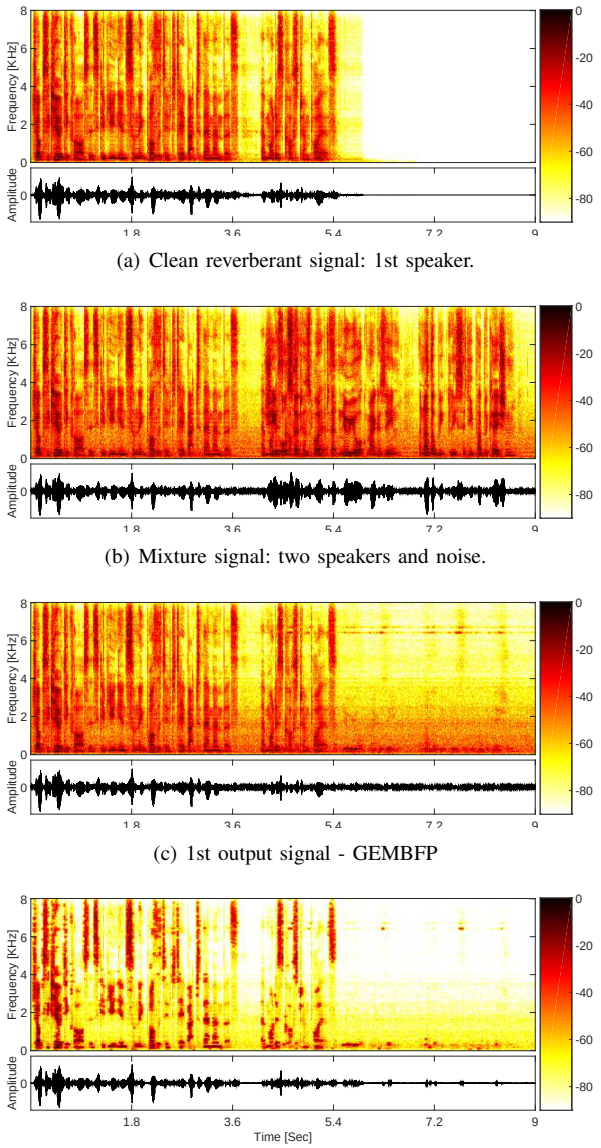
$$\mathrm{SDR}_d^{\mathrm{out}} = \mathrm{QM}\left\{\widetilde{s}_{d,d}, (x_d(t,k) - \widetilde{s}_{d,d})\right\} \ . \tag{67}$$

The same calculation is carried out for the GEMBFP output signals.

B. Algorithms' performance

In this section we investigate the performance of the proposed algorithms for the experimental setup described in Sec. VII-A. We start by comparing the performance of the two proposed algorithms, and then examine the QMs as a function of a single setup parameter (e.g. T_{60}, and RSNR) while all other parameters remain fixed.

1) Comparison between the proposed algorithms: As an example, sonograms are given in Fig. 2, where the clean, mixture, and output signals are depicted for both algorithms. As can be seen, the EMBFP reduce more noise and interference, in the price of higher signal distortion. This result confirms the discussion in Sec. V-A, attributing this distortion to the strong feedback loop of the speech PSD in the EMBFP algorithm. This point is further clarified in Fig. 3, where the QMs are presented for different values of ξ_{min}. It can be seen that as



(a) Clean reverberant signal: 1st speaker.

(b) Mixture signal: two speakers and noise.

(c) 1st output signal - GEMBFP

(d) 1st output signal - EMBFP

Fig. 2. Sonograms and waveforms for the clean (reverberant) signal of the 1-st speaker (a), noisy mixture of two speakers (b), and output signals of the GEMBFP algorithm - which considers the speech signal to be deterministic and unknown (c) and the EMBFP algorithm - which considers the speech signal to be stochastic process (d). SIR=0, RSNR=20 dB, eight microphones, 10 EM iterations. $\xi_{\mathrm{min}} = 0.3$

the value of ξ_{min} increases, the RSNR decreases due to the lower attenuation of noise, and the SDR increases due to the lower distortion of the desired source. The SIR decreases very slowly with ξ_{min}, suggesting that the SWF focuses mainly on noise. For $\xi_{\mathrm{min}} = 0.87$ the RSNR and SIR of both algorithms are approximately equal, while the EMBFP provides a lower SDR compared to the GEMBFP. In the rest of the experimental study, we used $\xi_{\mathrm{min}} = 0.3$.

2) QM vs. T_{60}: In Table III, a comparison between the GEMBFP, EMBFP, and the OBFP algorithms as a function of the reverberation level is depicted. As expected, the OBFP algorithm achieves better performance than the GEMBFP and the EMBFP algorithms, with the EMBFP obtaining lower SDR

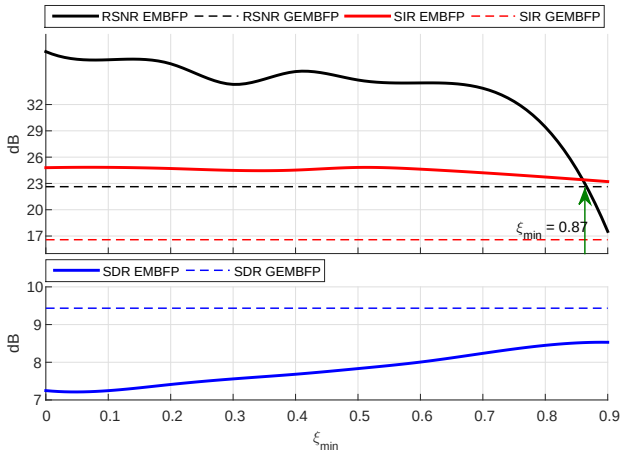


Fig. 3. QMs at the output for different values of ξ_{min}. Experiments were executed with $T_{60} = 0.36$ s, and input RSNR $= 30$ dB.

TABLE III
COMPARISON BETWEEN THE PROPOSED METHOD AND THE OBFP BEAMFORMER FOR VARIOUS REVERBERATION LEVELS.

Measure	T_{60} (s)	Input	OBFP	GEMBFP	EMBFP
SIR	0.16	0.0	18.4	12.0	**15.8**
	0.36	0.0	18.9	10.5	**14.4**
	0.61	0.0	17.0	11.0	**15.3**
	All	0.0	18.1	11.2	**15.2**
SDR	0.16	—	11.5	**8.4**	6.4
	0.36	—	10.3	**7.4**	6.2
	0.61	—	10.1	**7.0**	3.5
	All	—	10.7	**7.6**	5.4
RSNR	0.16	19.5	29.7	14.2	**26.6**
	0.36	19.5	29.4	15.7	**23.9**
	0.61	21.6	30.4	15.9	**23.7**
	All	20.2	29.8	15.3	**24.8**

values, and higher SIR and RSNR values, as discussed in Sec. V-A. In general, the longer the reverberation time, the more difficult the separation, and the performance degrades for the proposed and oracle algorithms.

Regarding the noise reduction performance, the EMBFP improves the RSNR scores by 5 dB, while the GEMBFP degrades them by 5 dB, as compared to the input. We will elaborate on this issue in Sec. VII-B4.

3) QM vs. WDO: In this paper, we assumed that speech signals do not overlap in the STFT domain, as was done in numerous speech BSS algorithms e.g. [12], [16]. However, this assumption is not fully satisfied in all practical scenarios, as will be shown in the following. We examine the proposed algorithms w.r.t. the validity of the assumption, following the W-disjoint orthogonality (WDO) definition in [10]. To this end,

TABLE IV
QM VERSUS WDO.

Measure	WDO	Input	OBFP	GEMBFP	EMBFP
SIR	[0.99,1]	0.0	17.5	10.4	**12.3**
	[0.95,0.99]	0.0	17.8	11.8	**16.3**
	[0.89,0.95]	0.0	19.0	11.3	**16.9**
	All	0.0	18.1	11.2	**15.2**
SDR	[0.99,1]	—	10.1	**7.6**	5.5
	[0.95,0.99]	—	9.9	**7.4**	4.4
	[0.89,0.95]	—	11.9	**7.9**	6.3
	All	—	10.7	**7.6**	5.4
RSNR	[0.99,1]	19.5	29.6	15.3	**24.3**
	[0.95,0.99]	21.7	30.2	16.7	**25.7**
	[0.89,0.95]	19.5	29.7	13.9	**24.3**
	All	20.2	29.8	15.3	**24.8**

we calculate a binary mask from the clean signals by

$$m_d(t,k) = \begin{cases} 1 & ; \quad |x_d(t,k)|^2 > \beta \cdot |x_{d'}(t,k)|^2 \\ 0 & ; \quad \text{else} \end{cases}, \qquad (68)$$

for $d, d' \in \{1,2\}$ and $d \neq d'$, and β set to 100. Then, the WDO of the d-th speaker was calculated

$$\text{WDO}_d = \frac{1}{B_d} \cdot \sum_{t,k} |m_d(t,k) \cdot x_d(t,k)|^2 - \frac{1}{B_d} \cdot \sum_{t,k} |m_d(t,k) \cdot x_{d'}(t,k)|^2, \qquad (69)$$

where $B_d = \sum_{t,k} |x_d(t,k)|^2$ is the total power of the d-th source, and finally WDO $= \frac{1}{2} \cdot (\text{WDO}_1 + \text{WDO}_2)$. Note that this procedure is identical to (16) in [10]. The results of the experiment were split according to their WDO value, and the summary is given in Table IV. The range of WDO values in Table IV is $[0.89, 1]$, where lower values indicate a high degree of overlap, and higher values indicate a higher sparsity in the STFT domain. The WDO values in the experiment Table IV are equally distributed, i.e. third of the signals has WDO between 0.99 and 1, third between 0.95 and 0.99, etc. In this experimental study, the WDO value is monotonically decreasing when T_{60} is increasing, since reverberation causes more overlap between the different speakers. However, this influence is only a secondary factor, since the WDO levels are influenced more by the specific speech signals than by T_{60}. It can be deduced that the proposed algorithms exhibit low sensitivity to the level of WDO among the tested values.

4) QM vs. Input RSNR: Next we analyze the experiments w.r.t. the input RSNR levels, where the average results for each level are depicted in Table V. The relation between the RSNR at the input and the SIR and SDR values at the output is not clear, and anyway, only a marginal difference can be observed between the different input noise levels. While the output RSNR is approximately 5 dB higher than the input

TABLE V
QM VERSUS RSNR.

Measure	RSNR	Input	OBFP	GEMBFP	EMBFP
SIR	10	0.0	18.6	10.8	**14.6**
	20	0.0	18.3	10.0	**15.2**
	30	0.0	17.5	12.7	**15.6**
	All	0.0	18.1	11.2	**15.2**
SDR	10	—	12.2	**8.0**	5.4
	20	—	10.7	**7.2**	4.5
	30	—	9.1	**7.7**	6.3
	All	—	10.7	**7.6**	5.4
RSNR	10	10.5	25.2	10.2	**15.7**
	20	19.8	28.8	14.1	**24.2**
	30	30.3	35.5	21.6	**34.4**
	All	20.2	29.8	15.3	**24.8**

RSNR for the EMBFP, the noise reduction degrades (actually, the noise is amplified) for the GEMBFP as the input RSNR increases. This may be attributed to the low input SIR (0 dB in all the signals) directing the beamformer towards interference suppression rather than noise reduction.

C. Comparison to a reference algorithm

In this section, we compare the proposed methods to the method proposed in [27] that is based on NMF. Unlike the proposed method, it is impossible to apply the method in [27] to $\mathbf{v}(t,k)$ or $\mathbf{h}_d(k)\cdot x_d(t)$ as described in Sec. VII-A4. Therefore, we use instead the signal to error ratio (SER) that is defined as follows (see (60)),

$$\text{SER} = \text{QM}\left\{x_d, x_d - \widehat{x}_d\right\} \ , \tag{70}$$

where $\widehat{x}_d$ is the output signal, i.e., $\widetilde{x}_d$ for the EMBFP, $\overline{x}_d$ for the GEMBFP, or the output of the NMF algorithm. For the input SER, we use $\widehat{x}_d = z_1$. The SER values for the input signal, the proposed algorithms, and the algorithm [27] as a function of the T_{60}, RSNR and WDO, are given in Table VI.

TABLE VI
SER VS. WDO, VS. T_{60}, AND VS. RSNR.

WDO	Input	OBFP	GEMBFP	EMBFP	NMF
[0.99,1]	0.2	6.5	2.4	**3.5**	1.4
[0.95,0.99]	0.1	6.4	2.8	**3.3**	1.6
[0.89,0.95]	0.2	8.8	2.8	**4.4**	3.3
All	0.1	7.3	2.7	**3.7**	2.1
T_{60}					
0.160	0.2	7.9	3.0	**4.1**	1.8
0.360	0.2	6.7	2.4	**3.5**	2.0
0.610	0.1	7.1	2.7	**3.6**	2.6
All	0.1	7.3	2.7	**3.7**	2.1
RSNR					
10	0.4	9.2	3.0	**4.4**	3.8
20	0.0	7.5	2.1	**3.5**	2.1
30	0.0	5.1	3.0	**3.3**	0.5
All	0.1	7.3	2.7	**3.7**	2.1

VIII. CONCLUSION

Two EM-based algorithms for simultaneous speech separation and noise reduction were presented, where both the acoustic parameters and the enhanced signals are estimated. We started by assuming two alternative clean signal models, i.e., either stochastic or deterministic unknown, and applied the EM scheme for an ML estimation of the parameters. While a deterministic model for the signal results in an MVDR beamformer, a stochastic model results in an MWF as well as an a posteriori speaker probability. The algorithm that uses the MVDR is denoted GEMBFP, since it is based on the generalized-EM scheme, while the one that uses the MWF is denoted EMBFP algorithm.

The algorithms were tested in various RSNR, WDO, and reverberation levels. For these scenarios, the proposed algorithms were compared to an oracle reference filter, that rely on a non-blind estimation of the parameters, and to an NMF-based algorithm proposed in [27]. The performance of the proposed algorithms are lower than the oracle algorithm's but higher than NMF algorithm. Unlike the GEMBFP, the EMBFP comprises SWF and an a posteriori speaker probability which enable better noise reduction and interference suppression at the cost of signal distortion. It was further shown that longer reverberation time degrades the performance, while the input RSNR and the levels of overlap between the competing speakers has a minor effect on the output quality.

ACKNOWLEDGMENTS

The authors would like to thank Yaron Laufer from the Faculty of Engineering, Bar Ilan University, for his helpful comments, and Dr. D. Kounades-Bastian from INRIA, Grenoble Rhône-Alpes, for his kind help in developing and sharing the reference method displayed in Sec. VII-C.

REFERENCES

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.
[2] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.
[3] I. Frost, O.L., "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.
[4] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[5] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614 –1626, Aug. 2001.

[6] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2230– 2244, Sep. 2002.

[7] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[8] A. Hyvrinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2004, vol. 46.

[9] K. Kumatani, J. McDonough, B. Rauch, D. Klakow, P. N. Garner, and W. Li, "Beamforming with a maximum negentropy criterion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 994–1008, 2009.

[10] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[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] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007, pp. 147–150.

[13] Y. Dorfan, O. Schwartz, B. Schwartz, E. A. Habets, and S. Gannot, "Multiple doa estimation and blind source separation using estimation-maximization," in *Science of Electrical Engineering (ICSEE), IEEE International Conference on the*. IEEE, 2016, pp. 1–5.

[14] J. Traa and P. Smaragdis, "Multichannel source separation and tracking with ransac and directional statistics," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 2233–2243, 2014.

[15] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[16] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1913–1928, 2013.

[17] O. Schwartz, S. Gannot, and E. A. P. Habets, "An Expectation-Maximization algorithm for multi-microphone speech dereverberation and noise reduction with coherence matrix estimation," *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 8, pp. 1393–1407, Aug. 2016.

[18] F. Gu, H. Zhang, W. Wang, and S. Wang, "An expectation-maximization algorithm for blind separation of noisy mixtures using gaussian mixture model," *Circuits, Systems, and Signal Processing*, pp. 1–30, 2016.

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

[20] N. 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, Sep. 2010.

[21] M. Togami, "Online speech source separation based on maximum likelihood of local Gaussian modeling," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 213–216.

[22] L. S. Simon and E. Vincent, "A general framework for online audio source separation," in *International Conference on Latent Variable Analysis and Signal Separation*, Tel Aviv, Israel, Mar. 2012, pp. 397–404.

[23] A. Alinaghi, P. J. Jackson, Q. Liu, and W. Wang, "Joint mixing vector and binaural model based stereo source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 9, pp. 1434–1448, 2014.

[24] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5210–5214.

[25] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*. Springer, 1998, pp. 355–368.

[26] C. Fevotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[27] A. Ozerov and C. Fvotte, "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.

[28] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of moving sound sources," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2015.

[29] ——, "A variational EM algorithm for the separation of time-varying convolutive audio mixtures," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1408–1423, 2016.

[30] M. Feder and E. Weinstein, "Optimal multiple source location estimation via the EM algorithm," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 1985, pp. 1762–1765.

[31] ——, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 4, pp. 477–489, Apr. 1988.

[32] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1072–1077, 2010.

[33] M. Taseska and E. A. P. Habets, "Informed spatial filtering for sound extraction using distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1195–1207, 2014.

[34] S. Araki and T. Nakatani, "Hybrid approach for multichannel source separation combining time-frequency mask with multi-channel Wiener filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 225–228.

[35] M. Taseska and E. A. P. Habets, "Spotforming using distributed microphone arrays," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[36] ——, "Spotforming: spatial filtering with distributed arrays for position-selective sound acquisition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1291–1304, 2016.

[37] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley and Sons, Feb. 2008.

[38] H. L. Van Trees, *Optimum array processing: part IV of detection, estimation, and modulation*. Wiley, New York, 2002.

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

[40] B. Schwartz, S. Gannot, and E. 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.

[41] M. A. Woodbury, "Inverting Modified Matrices," *Memorandum Rept. 42, Statistical Research Group, Princeton University, Princeton, NJ*, 1950.

[42] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Speech Enhancement*. Springer, 2005, pp. 299–327.

[43] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[44] F. Heese, M. Schafer, P. Vary, E. Hadad, S. M. Golan, and S. Gannot, "Comparison of supervised and semi-supervised beamformers using real audio recordings," in *IEEE 27th Convention of Electrical Electronics Engineers in Israel (IEEEI)*, Nov. 2012, pp. 1–5.

[45] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.

[46] E. Vincent, R. Gribonval, and C. Fvotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[47] S. Markovich-Golan, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.