2020 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2020, ESPOO, FINLAND

BLIND AUDIO SOURCE SEPARATION USING TWO EXPECTATION-MAXIMIZATION ALGORITHMS

Aviad Eisenberg, Boaz Schwartz and Sharon Gannot

Faculty of Engineering, Bar-Ilan University, Ramt-Gan ,Israel

aviad.eisenberg@biu.ac.il; boazsh0@gmail.com; sharon.gannot@biu.ac.il

ABSTRACT

The problem of multi-microphone blind audio source separation in noisy environment is addressed. The estimation of the acoustic signals and the associated parameters is carried out using the expectation-maximization algorithm. Two separation algorithms are developed using either deterministic representation or stochastic Gaussian distribution for modelling the speech signals. Under the deterministic model, the speech sources are estimated in the M-step by applying in parallel multiple minimum variance distortionless response (MVDR) beamformers, while under the stochastic model, the speech signals are estimated in the E-step by applying in parallel multiple multichannel Wiener filters (MCWF). In the simulation study, we generated a large dataset of microphone signals, by convolving speech signals, with overlapping activity patterns, by measured acoustic impulse responses. It is shown that the proposed methods outperform a baseline method in terms of speech quality and intelligibility.

***Index Terms*—** Blind audio source separation, Expectation-maximization algorithm, MVDR and multichannel Wiener filter beamforming

1. INTRODUCTION

Separation of a desired speaker from multi-microphone mixtures of multiple simultaneous speakers is required in many modern applications and devices, such as virtual assistants, hearing aids and smartphones. A comprehensive survey of state-of-the-art multichannel audio separation methods can be found in [1].

The W-disjoint orthogonality (WDO) property of the speech signal in the time-frequency (TF) domain is widely-used in the context of speech processing and it forms the basis for the degenerate unmixing estimation technique (DUET) algorithm [2]. This property is also used in [3, 4] for joint direction of arrival (DOA) estimation and source separation. This property is also widely used in deep learning methods for separation [5–7].

Many blind audio source separation (BASS) methods are utilizing the expectation-maximization (EM) algorithm [8], since it enables to separately estimate signals and associated parameters. In [9] an EM algorithm using the nonnegative matrix factorization (NMF) approach for BASS was proposed. This algorithm requires a good initialization of the mixing filters. In [10], the problem of high reverberation was addressed by using the convolutive transfer function (CTF) representation of the acoustic filter, which parameters are estimated by the EM algorithm. In [11] under the sparsity assumption, an EM algorithm was proposed, to classify various features derived from the binaural input. Under the WDO assumption, two statistical models for the BASS problem were presented in [12]. From these models, and using the EM algorithm, two enhancement schemes were derived, based on either minimum variance distortionless response (MVDR) or multichannel Wiener filter (MCWF) beamformers. Note that due to the WDO property of the speech signals, the cancellation of the interfering source is only obtained by the noise suppression capabilities of the baeamformers and the application of TF masking.

In [13, 14], two statistical models are presented. The *complete data* of both EM-based algorithms are the separated source signals, as received by the microphones, together with an arbitrary portion of the additive noise. The algorithms differ in the signal modelling. In the first, the desired signals are assumed to be known, and in the second, the desired signals are assumed to be a stochastic Gaussian. These models are then used to devise an EM algorithm for estimating the DOA of multiple concurrent sources. An extension to the case of deterministic unknown signals can be found in [15].

In the current contribution, we adopt the EM framework of [13, 14]. The WDO property is only utilized in initializing the EM algorithm, to obtain a rough separation between the speech sources. Unlike these papers, which focus on the DOA estimation task, our goal in this paper is to separate and enhance the speech sources. We therefore do not make any restrictive assumptions on the acoustic transfer functions (ATFs) relating the sources and the microphones. Estimating the ATFs under the deterministic model, boils down to least squares (LS) fit between the estimated sources and the complete data, while under the stochastic model to the calculation of the dominant eigenvector of a weighted correlation matrix. In the former model, the speech is estimated by applying an MVDR beamformer, and in the latter by applying an MCWF.

2. PROBLEM FORMULATION AND STATISTICAL MODELS

2.1. signal model

We assume that D concurrent speakers are captured by J microphones in a reveberant and noisy environment. The BASS problem is formulated in the short-time Fourier transform (STFT) domain, where $k \in \{0, \ldots, K-1\}$ and $t \in \{0, \ldots, T-1\}$ represent the frequency index and time-frame index, respectively and T and K are the total number of time frames and frequency bands, respectively. Let $s_d(t,k)$ denote the clean, anechoic speech signal of the d-th speaker. By assuming that the ATFs are time-invariant, the observed signal, as received at the microphones array, can be modelled

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme, Grant Agreement No. 871245.

978-1-7281-6662-9/20/$31.00

as

$$\mathbf{y}(t,k) = \sum_{d=1}^{D} \mathbf{h}_d(k) \cdot s_d(t,k) + \mathbf{n}(t,k)\,, \tag{1}$$

where $\mathbf{h}_d(k)$ is a $J \times 1$ vector of the ATFs relating the d-th source and the microphones array, and $\mathbf{n}(t,k)$ is the $J \times 1$ vector of additive noise as received by the microphone array, modelled as a statistically independent, zero-mean, complex-Gaussian random vector with time-invariant covariance matrix $\mathbf{Q}(k)$.

Following [13, 14], we rewrite the observed signal as a sum of arbitrarily-chosen D noisy components. Under this reformulation, the observed signal can be written in the following matrix form:

$$\mathbf{y}(t,k) = \sum_{d=1}^{D} \mathbf{x}_d(t,k) = [\mathbf{I}, \mathbf{I}, \ldots, \mathbf{I}] \begin{bmatrix} \mathbf{x}_1(t,k) \\ \mathbf{x}_2(t,k) \\ \vdots \\ \mathbf{x}_D(t,k) \end{bmatrix} = \mathbf{H}\mathbf{x}(t,k) \tag{2}$$

where $\mathbf{I}$ is a $J \times J$ identity matrix, $\mathbf{H}$ is a non-invertible matrix comprised of D row-concatenated identity matrices, and $\mathbf{x}(t,k)$ is the *complete data* constructed by:

$$\mathbf{x}_d(t,k) = \mathbf{h}_d(k) \cdot s_d(t,k) + \mathbf{n}_d(t,k);\ d = 1, \ldots, D\,, \tag{3}$$

where $\mathbf{n}_d(t,k)$ is the, arbitrarily-defined, d-th component of the noise $\mathbf{n}(t,k)$ such that

$$\sum_{d=1}^{D} \mathbf{n}_d(t,k) = \mathbf{n}(t,k)\,. \tag{4}$$

The arbitrary decomposition of the noise is chosen such that $\mathbf{n}_d(t,k)$ are mutually uncorrelated, zero-mean, complex-Gaussian random variables

$$\mathbf{n}_d(t,k) \sim \mathcal{N}_c\big(\mathbf{n}_d(t,k); \mathbf{0}, \mathbf{Q}_d(k)\big)\,. \tag{5}$$

The D covariance matrices share the same structure of the full noise covariance matrix, and their portions in the total noise level is determined by a set of scalar weights

$$\mathbf{Q}_d(k) = \beta_d \cdot \mathbf{Q}(k)\,. \tag{6}$$

such that $\sum_d \beta_d = 1$.

In the next two sections, we develop two statistical models for the speech signals $s_d(t,k)$.

2.2. Deterministic Model

Under the deterministic model $s_d(t,k)$, $d = 1, \ldots, D$ are referred to as a deterministic unknown parameters. As a result, the set of unknown parameters are

$$\bar{\boldsymbol{\theta}} = \{s_d(t,k), \mathbf{h}_d(k), \mathbf{Q}_d(k)\}_{d=1}^{D}\,. \tag{7}$$

Denoting $\mathbf{m}_d(t,k) = s_d(t,k) \cdot \mathbf{h}_d(k)$, the probability density function (p.d.f.) of the observation data, given the vector of parameters $\bar{\boldsymbol{\theta}}$ is complex-Gaussian,

$$f\big(\mathbf{y}(t,k)|\bar{\boldsymbol{\theta}}\big) = \mathcal{N}_c\left(\mathbf{y}(t,k); \sum_{d=1}^{D} \mathbf{m}_d(t,k), \bar{\mathbf{P}}(k)\right) \tag{8}$$

where the measurement covariance matrix is given by:

$$\bar{\mathbf{P}}(k) = \sum_{d=1}^{D} \mathbf{Q}_d(k) = \mathbf{Q}(k)\,. \tag{9}$$

2.3. Stochastic Model

As an alternative approach the speech signals are modelled as zero-mean complex-Gaussian random variables, such that

$$s_d(t,k) \sim \mathcal{N}_c(s_d(t,k); 0, \phi_d(t,k)) \tag{10}$$

where $\phi_d(t,k)$ is the power spectral density (PSD) of the d-th speaker. Under this model, the set of unknown parameters is (compare to (7)):

$$\tilde{\boldsymbol{\theta}} = \{\phi_d(t,k), \mathbf{h}_d(k)\}_{d=1}^{D}\,. \tag{11}$$

Note, that under this model, the noise covariance matrix is assumed to be a priori known. The p.d.f. of the observations is given by

$$f(\mathbf{y}(t,k)|\tilde{\boldsymbol{\theta}}) = \mathcal{N}_c\big(\mathbf{y}(t,k), \mathbf{0}, \tilde{\mathbf{P}}(t,k)\big)\,. \tag{12}$$

where

$$\tilde{\mathbf{P}}(t,k) = \sum_{d=1}^{D} \phi_d(t,k)\mathbf{h}_d(k)\mathbf{h}_d^H(k) + \mathbf{Q}_d(k) \tag{13}$$

such that $\tilde{\mathbf{P}}(t,k) = \sum_{d=1}^{D} \boldsymbol{\Lambda}_d(t,k)$, with

$$\boldsymbol{\Lambda}_d(t,k) = \phi_d(t,k)\mathbf{h}_d(k)\mathbf{h}_d^H(k) + \mathbf{Q}_d(k)\,, \tag{14}$$

the covariance matrix of the d-th component of the complete data.

3. THE PROPOSED EM-BASED BASS ALGORITHMS

Two EM-based BASS algorithms will be derived now from both the deterministic and stochastic models, as presented in Sec. 2.

First we write the general formulation of the EM algorithm, and since it is identical for the two algorithms proposed, we define $\boldsymbol{\theta} \in \{\bar{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}\}$ to be either the deterministic or stochastic model parameters, respectively.

Define the *auxiliary function* as the expectation, calculated at the current parameter set $\boldsymbol{\theta}^{(\ell-1)}$:

$$U\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\ell-1)}\big) = E\big\{\log f(\mathbf{x}(t,k); \boldsymbol{\theta}) \big| \mathbf{y}(t,k); \boldsymbol{\theta}^{(\ell-1)}\big\}\,. \tag{15}$$

The EM iterates between calculating the auxiliary function (E-step), and maximizing it with respect to (w.r.t.) the parameter set $\boldsymbol{\theta}$ (M-step). As all frequency bins are assumed independent, the derivation is carried out per-frequency bin. Hence, for conciseness, the frequency index k will be omitted for the rest of the derivation.

3.1. Deterministic EM Algorithm

By (3) it follows that $\mathbf{x}_d(t,k)$ is a complex-Gaussian random vector,

$$\mathbf{x}_d(t) \sim \mathcal{N}_c\big\{\mathbf{x}_d(t); \mathbf{m}_d(t), \mathbf{Q}_d\big\}\,, \tag{16}$$

and the auxiliary function is hence given by:

$$U(\bar{\boldsymbol{\theta}}; \bar{\boldsymbol{\theta}}^{(\ell-1)}) = -\sum_{t=0}^{T-1} \sum_{d=1}^{D} \Big(\log|\mathbf{Q}_d| + \mathrm{tr}\left(\mathbf{Q}_d^{-1} \widehat{\mathbf{x}_d(t)\mathbf{x}_d^H(t)}\right) - \hat{\mathbf{x}}_d^H(t)\mathbf{Q}_d^{-1}\mathbf{m}_d(t) - \mathbf{m}_d^H(t)\mathbf{Q}_d^{-1}\hat{\mathbf{x}}_d(t) + \mathbf{m}_d^H(t)\mathbf{Q}_d^{-1}\mathbf{m}_d(t)\Big)\,. \tag{17}$$

Next, we derive the EM formula for the maximum likelihood (ML) estimation of the parameters.

3.1.1. E-Step

In the E-step the first- and second-order statistics of $\mathbf{x}_d(t)$ are estimated, using the latest value of the parameter set $\bar{\boldsymbol{\theta}}^{(\ell-1)}$:

$$\hat{\mathbf{x}}_d(t) = \mathbf{m}_d(t) + \mathbf{Q}_d \cdot \bar{\mathbf{P}}^{-1}\Big(\mathbf{y}(t) - \sum_{d=1}^{D}\mathbf{m}_d(t)\Big) \tag{18a}$$

$$\widehat{\mathbf{x}_d(t)\mathbf{x}_d^H(t)} = \mathbf{Q}_d - \mathbf{Q}_d\bar{\mathbf{P}}^{-1}\mathbf{Q}_d + \hat{\mathbf{x}}_d(t)\hat{\mathbf{x}}_d^H(t)\,. \tag{18b}$$

A detailed derivation of (18a)-(18b) can be found in [14]. These estimates are used in the M-step, as follows.

3.1.2. M-Step

In the M-Step, the parameters are updated by maximizing (17),

$$\hat{\bar{\boldsymbol{\theta}}} = \arg\max_{\bar{\boldsymbol{\theta}}} U(\bar{\boldsymbol{\theta}}; \bar{\boldsymbol{\theta}}^{(\ell-1)})\,. \tag{19}$$

The maximization is carried out by calculating the derivatives w.r.t. all parameters,[1]

$$\frac{\partial}{\partial s_d^*(t)}U(\bar{\boldsymbol{\theta}}; \bar{\boldsymbol{\theta}}^{(\ell-1)}) = \frac{\partial}{\partial s_d^*}\Big(-\mathbf{m}_d^H(t)\mathbf{Q}_d^{-1}\hat{\mathbf{x}}_d(t) + \mathbf{m}_d^H(t)\mathbf{Q}_d^{-1}\mathbf{m}_d(t)\Big) = -\mathbf{h}_d^H\mathbf{Q}_d^{-1}\hat{\mathbf{x}}_d(t) + s_d(t)\mathbf{h}_d^H\mathbf{Q}_d^{-1}\mathbf{h}_d\,. \tag{20}$$

Equating (20) to zero results in:

$$\hat{s}_d(t) = \frac{\hat{\mathbf{h}}_d^H\mathbf{Q}_d^{-1}}{\hat{\mathbf{h}}_d^H\mathbf{Q}_d^{-1}\hat{\mathbf{h}}_d}\hat{\mathbf{x}}_d(t)\,. \tag{21}$$

Interestingly, the resulting estimator is the MVDR beamformer directed towards the source $s_d(t)$ and minimizing the noise with covariance $\mathbf{Q}_d$. Note, that due to the EM procedure, each of the sources is treated separately, without considering the other sources, and only part of the noise $\mathbf{Q}_d$, the part that is associated with the signal $s_d(t)$, is minimized. A similar result was obtained in [15] for a spatially-white noise and simplified acoustic system, comprising only the direct path of the sound propagation.

The steering vector of the MVDR $\hat{\mathbf{h}}_d$ will be estimated in the sequel, by maximizing (15) w.r.t. $\mathbf{h}_d^H$:

$$\frac{\partial}{\partial \mathbf{h}_d^H}U(\bar{\boldsymbol{\theta}}; \bar{\boldsymbol{\theta}}^{(\ell-1)}) = \frac{\partial}{\partial \mathbf{h}_d^H}\sum_{t=0}^{T-1}\Big(-\mathbf{m}_d^H(t)\mathbf{Q}_d^{-1}\hat{\mathbf{x}}_d(t) + \mathbf{m}_d^H(t)\mathbf{Q}_d^{-1}\mathbf{m}_d(t)\Big) = \sum_{t=0}^{T-1}\left(-s_d^*(t)\mathbf{Q}_d^{-1}\hat{\mathbf{x}}_d(t) + s_d^*(t)\mathbf{Q}_d^{-1}s_d(t)\mathbf{h}_d\right). \tag{22}$$

By equating (22) to zero, an estimator for $\mathbf{h}_d$ is obtained:

$$\hat{\mathbf{h}}_d = \frac{\sum_{t=0}^{T-1}\hat{s}_d^*(t)\hat{\mathbf{x}}_d(t)}{\sum_t|\hat{s}_d(t)|^2}\,, \tag{23}$$

which is the LS fit between the d-th speaker estimate $\hat{s}_d(t)$ and the d-th component of the complete data $\hat{\mathbf{x}}_d(t)$.

[1] Derivatives are calculated using the Matrix Cookbook
`www.math.uwaterloo.ca/hwolkowi/matrixcookbook.pdf`

Algorithm 1: The EM algorithm: Deterministic model.

Initialize
for $\ell = 1$ **to** L **do**
  E-step
    First-order statistics: $\hat{\mathbf{x}}_d(t,k)$ (18a)
    Second-order statistics: $\widehat{\mathbf{x}_d(t,k)\mathbf{x}_d^H(t,k)}$ (18b)
  M-step
    Estimate $\hat{s}_d(t,k)$ using MVDR beamformer (21)
    Estimate $\hat{\mathbf{h}}_d(k)$ using LS fit (23)
    Estimate $\hat{\mathbf{Q}}_d(k)$ by residual error averaging (24)
end

Finally, by maximizing (15) w.r.t. $\mathbf{Q}_d(k)$ and equating to zero we obtain:

$$\hat{\mathbf{Q}}_d = \frac{1}{T}\sum_{t=0}^{T-1}\Big(\widehat{\mathbf{x}_d(t)\mathbf{x}_d^H(t)} - \mathbf{m}_d(t)\hat{\mathbf{x}}_d^H(t) - \hat{\mathbf{x}}_d(t)\mathbf{m}_d^H(t) + \mathbf{m}_d(t)\mathbf{m}_d^H(t)\Big)\,, \tag{24}$$

which can be easily recognized as the averaged residual error between $\mathbf{m}_d(t)$ and $\hat{\mathbf{x}}_d(t)$. The EM algorithm for the deterministic model is summarized in Algorithm 1.

3.2. Stochastic EM Algorithm

Under the stochastic model, using (3), (10) and (14), it follows that $\mathbf{x}_d(t,k)$ is a complex-Gaussian random vector,

$$\mathbf{x}_d(t) \sim \mathcal{N}_c\big\{\mathbf{x}_d(t); \mathbf{0}, \boldsymbol{\Lambda}_d(t)\big\}\,, \tag{25}$$

with the following log-p.d.f. :

$$\log f(\mathbf{x}(t); \tilde{\boldsymbol{\theta}}) = -\sum_{t=0}^{T-1}\sum_{d=1}^{D}\Big(\log|\boldsymbol{\Lambda}_d(t)| + \mathbf{x}_d^H(t)\boldsymbol{\Lambda}_d^{-1}(t)\mathbf{x}_d(t)\Big) = -\sum_{t=0}^{T-1}\sum_{d=1}^{D}\Big(\log|\boldsymbol{\Lambda}_d(t)| + \mathrm{Tr}\left(\boldsymbol{\Lambda}_d^{-1}(t)\mathbf{x}_d(t)\mathbf{x}_d^H(t)\right)\Big)\,. \tag{26}$$

To alleviate the computational complexity involved in the calculation of the auxiliary function, we make two simplifying assumptions. First, the noise is assumed to be spatially-white. Based on (6), the component-wise covariance matrices are give by:

$$\mathbf{Q}_d = \beta_d \cdot \sigma_d^2 \cdot \mathbf{I}\,. \tag{27}$$

Furthermore, we assume the availability of speech-free segments from which the noise level σ_d^2 can be estimated, thus circumventing its estimation by the EM algorithm.[2]

The second assumption states that the ATFs are normalized, namely

$$\|\mathbf{h}_d\|^2 = 1\,. \tag{28}$$

While this may seem a restrictive assumption, it is actually mandatory to normalize the ATFs due to the inherent gain ambiguity problem of the model.

[2] If the noise in spatially non-white, we can still whiten the measurements, provided that the noise covariance matrix can be estimated in advance. This topic is left for future study.

Under these assumptions, it can be shown that

$$\log|\mathbf{\Lambda}_d(t)| = J\log(\sigma_d^2) + \log\left(\frac{\phi_d(t)}{\sigma_d^2}+1\right) \tag{29a}$$

$$\mathbf{\Lambda}_d^{-1}(t) = \frac{1}{\sigma_d^2}\left(\mathbf{I} - \frac{\mathbf{h}_d\mathbf{h}_d^H\frac{\phi_d(t)}{\sigma_d^2}}{\frac{\phi_d(t)}{\sigma_d^2}+1}\right). \tag{29b}$$

The auxiliary function for the stochastic model can now be stated :

$$U(\tilde{\boldsymbol{\theta}};\tilde{\boldsymbol{\theta}}^{(\ell-1)}) = -\sum_{t=0}^{T-1}\sum_{d=1}^{D}\left(J\log(\sigma_d^2) + \log\left(\frac{\phi_d(t)}{\sigma_d^2}+1\right) + \frac{1}{\sigma_d^2}\mathrm{Tr}(\widehat{\mathbf{x}_d(t)\mathbf{x}_d^H(t)}) - \frac{1}{\sigma_d^2}\frac{\mathrm{h}_d^H\widehat{\mathbf{x}_d(t)\mathbf{x}_d^H(t)}\mathrm{h}_d\frac{\phi_d(t)}{\sigma_d^2}}{\frac{\phi_d(t)}{\sigma_d^2}+1}\right) \tag{30}$$

3.2.1. E-Step

In the E-step the first- and second-order statistics of the components of the complete-data are estimated, using the current value of the parameter set $\tilde{\boldsymbol{\theta}}^{(\ell-1)}$:

$$\hat{\mathbf{x}}_d(t) = \mathbf{\Lambda}_d(t)\tilde{\mathbf{P}}^{-1}(t)\mathbf{y}(t) \tag{31a}$$

$$\widehat{\mathbf{x}_d(t)\mathbf{x}_d^H(t)} = \mathbf{\Lambda}_d(t) - \mathbf{\Lambda}_d(t)\tilde{\mathbf{P}}^{-1}\mathbf{\Lambda}_d(t) + \hat{\mathbf{x}}_d(t)\hat{\mathbf{x}}_d^H(t). \tag{31b}$$

The resulting estimator (and its associated variance) is the MCWF for estimating the d-th component of the complete data $\hat{\mathbf{x}}_d(t)$ given the measurements $\mathbf{y}(t)$, which is expected due to the (complex) Gaussian p.d.f. of both signals.

3.2.2. M-Step

In the M-step, the auxiliary function should be maximized. However, as the norm of the ATFs are constrained, we should apply a constrained maximization. Define the Lagrangian:

$$L(\tilde{\boldsymbol{\theta}},\lambda) = U(\tilde{\boldsymbol{\theta}};\tilde{\boldsymbol{\theta}}^{(\ell-1)}) + \lambda(\|\mathbf{h}_d\|^2 - 1) \tag{32}$$

with λ the Lagrange multiplier. Calculating the derivative of (32) w.r.t. $\mathbf{h}_d^H$, equating to zero and rearranging terms yield:

$$\frac{1}{T}\cdot\sum_{t=0}^{T-1}\frac{\phi_d(t)}{\phi_d(t)+\sigma_d^2}\widehat{\mathbf{x}_d(t)\mathbf{x}_d^H(t)}\cdot\mathbf{h}_d = \lambda\cdot\mathbf{h}_d. \tag{33}$$

It can be deduced that the estimate of the ATF $\hat{\mathbf{h}}_d$ is parallel to the eigenvector corresponding to the largest eigenvalue of

$$\frac{1}{T}\cdot\sum_{t=0}^{T-1}\frac{\phi_d(t)}{\phi_d(t)+\sigma_d^2}\widehat{\mathbf{x}_d(t)\mathbf{x}_d^H(t)}. \tag{34}$$

Calculating the derivative of (32) w.r.t. $\phi_d(t)$ yields and estimate of the PSD of the d-th speech source:

$$\hat{\phi}_d(t) = \mathbf{h}_d^H\widehat{\mathbf{x}_d(t)\mathbf{x}_d^H(t)}\mathbf{h}_d - \sigma_d^2 \tag{35}$$

To circumvent numerical issues, we further limit the minimum value of the PSD:

$$\hat{\phi}_d(t) \geq \xi_{min}. \tag{36}$$

The estimator in the stochastic case is summarized in Algorithm 2.

Algorithm 2: The EM algorithm: Stochastic model.

Initialize
  Estimate the noise PSD from speech-absent frames
for $\ell = 1$ **to** L **do**
  E-step
    First-order statistics: $\hat{\mathbf{x}}_d(t,k)$ (31a)
    Second-order statistics: $\widehat{\mathbf{x}_d(t,k)\mathbf{x}_d^H(t,k)}$ (31b)
  M-step
    Estimate $\hat{\mathbf{h}}_d(k)$ from the "largest" eigenvector of (34)
    Estimate $\hat{\phi}_d(t,k)$ from (35),(36)
  Post-filter
    Estimate the clean speech signal $\hat{s}_d(t,k)$
end

4. PRACTICAL CONSIDERATIONS

4.1. Initialization of the EM parameters

The EM algorithm is notorious for its convergence problems. In this section we will present initialization procedures for the two presented methods that may circumvent the tendency of the EM algorithm to be trapped in local maximum.

4.1.1. Initialization of the Deterministic Algorithm

We propose to initialize the desired speech signal $s_d(t)$ using the DUET method [2] that takes advantage of the WDO property of the speech signal in the STFT domain. Although the performance of the DUET algorithm in noisy and reverberant environments is limited, it can still provide a good initial separation of the speech sources.

In the proposed method the ATFs $\mathbf{h}_d(k)$ are estimated by a LS fit between $s_d(t,k)$ and $\mathbf{x}_d(t,k)$. As both are unavailable prior to the application of the iterative procedure, we propose to substitute these signals with the estimated separated signals from the DUET initialization and the measured signals $\mathbf{y}(t,k)$. The noise covariance is simply initialized as an identity matrix, $\mathbf{Q}(k) = \mathbf{I}$.

4.1.2. Initialization of the Stochastic Algorithm

For the stochastic algorithm we should initialize the parameters $\mathbf{h}_d(k)$ and $\phi_d(t,k)$. The procedure for initializing the ATFs is similar to the initialization of the deterministic algorithm, namely first the separated speech signal $s_d(t,k)$ are initialized by applying the DUET algorithm [2] and then the ATFs are estimated by LS fitting of these signals and the microphone signals. The PSDs of all sources are simply initialized by $\phi_d(t,k) = 1$.

4.2. Post-processing Stage for the Stochastic Algorithm

The stochastic algorithm, as opposed to the deterministic algorithm, is not providing an estimate of the separated speech sources $s_d(t,k),\ d = 1,\ldots,D$. Instead, the outcome of this algorithm are the vectors $\mathbf{x}_d(t,k)$, each of which comprised of the contribution of the d-th source to all microphones and a part of the original noise signal. Hence, this algorithm only separates the sources but does not reduce the noise. To obtain the required separated and denoised signals we propose to apply the following matched-filter beamformer,

utilizing the two simplifying assumptions above:

$$\hat{\tilde{s}}_d(t,k) = \hat{\mathbf{h}}_d^H(k)\mathbf{x}_d(t,k)\,. \tag{37}$$

While no claims of optimally of this heuristic post-filtering hold, it can still serve as a plausible procedure for spatially-white noise reduction and distortion correction. Note that since

$$\mathbf{h}_d(k)s_d(t,k) = \frac{\mathbf{h}_d(k)}{\|\mathbf{h}_d(k)\|}\cdot(\|\mathbf{h}_d(k)\|s_d(t,k)) \tag{38}$$

the outcome of stochastic algorithm will be $\tilde{s}_d(t,k) = \|\mathbf{h}_d(k)\|s_d(t,k)$ rather than its an-echoic counterpart.

5. SIMULATION RESULTS

The proposed algorithm variants were evaluated and compared to the baseline DUET algorithm [2] using the following simulation setup. The microphones signals were generated by convolving two (partially) overlapping sources ($D = 2$) with real room impulse responses (RIRs) drawn from a publicly-available database [16] recorded in our lab, with dimensions $6 \times 6 \times 2.4$ m. The reverberation level can be controlled by flipping dedicated panels covering the room facets. We have used two of the reverberation levels in the database, namely $T_{60} = \{0.36, 0.61\}$ s. Two equi-power speech signals were acquired by an eight-microphone linear array with inter-distances of [3-3-3-8-3-3-3] cm. The distance between the sound sources and the microphone array was 1 m for all experiments. We simulated two angular distances between the sources, 120° and 60°, respectively. A pseudo-diffused noise signal (generated by four loudspeakers facing the room corners) was added to the microphone signals with signal-to-noise ratio (SNR) levels of 10 or 20 dB. The anechoic speech signals, 30 sec long, were drawn from the Wall Street Journal (WSJ) corpus [17] with gender balance. The total number of experiments was 600, where in each experiment each source was randomly drawn from the database, summing up to 3 hours of acoustic recordings. The sampling frequency of the speech signals was 16 kHz. The STFT frame size was 64 ms with 50% overlap. The parameters β_d and ξ_{min} were set as 0.5 and $0.5 \cdot \sigma_d^2$, respectively.

Separation results and distortion levels were evaluated using the BSS eval toolbox [18]. The intelligibility of the speech signal was evaluated with the short-time objective intelligibility (STOI) measure [19]. All measures are reported as a function of three parameters: the input SNR, reverberation level and speakers overlap percentage.

The performance measures for the two reverberation levels are depicted in Table 1, for the two SNR levels in Table 2 and as a function of the overlap percentage (either 50%, 80% or 100%) in Table 3. It is evident that the DUET algorithm (that serves also as the initialization stage for the proposed methods) outperforms both proposed variants in terms of signal-to-interference ratio (SIR) measures. The stochastic algorithms is slightly better than the deterministic algorithm in this measure. However, for the distortion measures (signal-to-distortion ratio (SDR) and signal-to-artifact ratio (SAR)), the proposed algorithms clearly outperform the DUET algorithm, with the deterministic algorithm slightly better than the stochastic algorithm. In terms of STOI, the deterministic algorithm achieves the best results with significant improvement w.r.t. the input signal and the DUET output, indicating low distortion. The STOI scores of the DUET algorithm is significantly worse than STOI scores of the mixed and noisy input signal. The quality of the proposed algorithm variants is also demonstrated by assessing the sonograms in Fig. 1.

Table 1. Performance measures for different reverberation level. Other parameters averaged.

	T_{60} (s)	Input	DUET	Deter.	Stoch.
SIR [dB]	0.36	0.4	13.4	10.9	12.2
	0.61	0.3	13	10.1	11.1
SDR [dB]	0.36	-2.4	-1.8	5.4	4.5
	0.61	-2.9	-2.1	4	3.5
SAR [dB]	0.36	4.7	-1.4	7.5	5.8
	0.61	3.4	-2	5.8	5
STOI [%]	0.36	68.9	47.6	92	89
	0.61	64.1	38.9	87.2	83.4

Table 2. Performance measures for different SNR levels. Other parameters averaged.

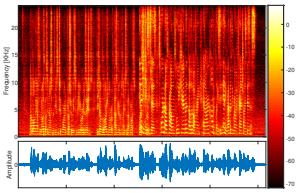
	Input SNR	Input	DUET	Deter.	Stoch.
SIR [dB]	10	0.3	13	10.3	11.6
	20	0.3	12.1	10.7	11.9
SDR [dB]	10	-3	-2.5	4.1	3.5
	20	-2.2	-1.8	5.6	4.6
SAR [dB]	10	3	-2.2	5.7	4.8
	20	5.4	-1.4	7.7	6
STOI [%]	10	61.2	38.6	86.9	83
	20	72	49	92.3	88.6

6. CONCLUSIONS

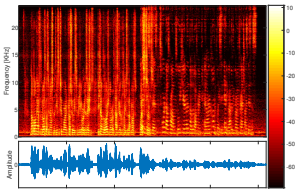
Two multi-microphone EM-based algorithms for BASS, derived from a deterministic and a stochastic models of the speech signals, were presented. In both algorithms, the E-step consists of the calculation of the first- and second-order statistics of the complete-data, and in the M-step, the acoustic parameters of the signals are estimated. In the algorithm derived from the deterministic model, the output signal is the outcome of the E-step, while in the algorithm derived from the stochastic model, an additional step is required to reduce the additive noise. We have heuristically chosen a matched-filter beamformer for this purpose. The algorithms were derived and implemented in the STFT domain, independently for each frequency bin. Different cases were tested, covering various levels of input SNR, overlap between competing speakers, and reverberation level. In all the tested cases, the two proposed algorithm variants performed similarly. Comparing these algorithms to the DUET algorithm, a significant improvement of the signal quality was achieved, in terms of SDR and STOI measures, while only mildly sacrificing competing speaker suppression, as recorded by the SIR measure.

7. REFERENCES

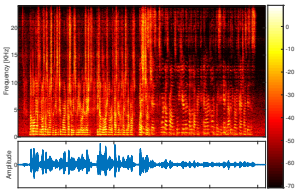
[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Au-*



(a) Mixture signal.



(b) 1st signal: Stoch. algorithm



(c) 1st signal: Deter. algorithm

Fig. 1. Sonograms and waveforms for the noisy mixture of two speakers (a), and the output signal for the 1st speaker of the stochastic and deterministic algorithms, (b) and (c), respectively. SNR=20 dB, eight microphones, 8 EM iterations.

Table 3. Performance measures as a function of overlap percentage between speakers. Other parameters averaged.

	Overlap[%]	Input	DUET	Deter.	Stoch.
SIR [dB]	50	0.3	14	10.8	12
	80	0.3	13	11	11.6
	100	0.3	12.6	10.4	11.5
SDR [dB]	50	-2.6	-1.9	5	4.3
	80	-2.4	-2.1	4.9	4
	100	-2.4	-2.2	5	3.9
SAR [dB]	50	4.1	-1.6	6.8	5.5
	80	4	-1.8	6.7	5.4
	100	4	-1.7	6.5	5.4
STOI [%]	50	76.5	47.4	92.1	90
	80	65.9	43	89.3	85.9
	100	57.4	39.2	87	82.4

dio, Speech, and Language Processing, vol. 25, no. 4, pp. 692–730, 2017.

[2] 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.

[3] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2006.

[4] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007, pp. 139–142.

[5] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

[6] Z. Wang, J. L. Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[7] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger, and S. Gannot, "Multi-microphone speaker separation based on deep DOA estimation," in *The 27th European Signal Processing Conference (EUSIPCO)*, 2019.

[8] 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)*, vol. 39, no. 1, pp. 1–22, 1977.

[9] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2009.

[10] 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.

[11] 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.

[12] B. Schwartz, S. Gannot, and E. A. Habets, "Two model-based em algorithms for blind source separation in noisy environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2209–2222, 2017.

[13] 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)*, vol. 10, 1985, pp. 1762–1765.

[14] ——, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 4, pp. 477–489, 1988.

[15] ——, "Optimal multiple source location via the EM algorithm," Woods Hole Oceanographic Institution, MA, USA, Tech. Rep. 1986-07, 1986. [Online]. Available: https://hdl.handle.net/1912/7915

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

[17] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[18] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.