INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Online Blind Audio Source Separation using Recursive Expectation-Maximization

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 challenging problem of online multi-microphone blind audio source separation (BASS) in noisy environment is addressed in this paper. We present a sequential, non-iterative, algorithm based on the recursive EM (REM) framework. In the proposed algorithm, the compete-data, which constitutes the separated sources and residual noise, is estimated in the E-step by applying a multichannel Wiener filter (MCWF); and the corresponding parameters, comprised of acoustic transfer functions (ATFs) relating the sources and the microphones and power spectral densities (PSDs) of the desired sources, are sequentially estimated in the M-step. The separated speech signals are further enhanced using matched-filter beamformers. The performance of the algorithm is demonstrated in terms of the separation capabilities, the resulting speech intelligibility and the ability to track the direction of arrival (DOA) of the moving sources.
Index Terms: blind audio source separation, recursive expectation maximization, multichannel Wiener filter beamforming

1. Introduction

Blind audio source separation (BASS) is an essential component in various applications such as hearing aids, conference calls, virtual assistants and robot audition. Its goal is to separate or extract the audio signals of the interesting speakers in the acoustic scene. A comprehensive survey of state-of-the-art multichannel audio separation methods can be found in [1–3].

The expectation-maximization (EM) algorithm [4] is widely used for BASS [5], due to its ability to jointly estimate the signals and the associated parameters, e.g. the coefficients of the multichannel filters. In [6], two EM algorithms were developed based on different statistical models describing the desired speakers. In [7], a combination of EM and deep neural network (DNN) was proposed. In the E-step, a MCWF is used for extracting the desired sources. In the M-step, the power spectral density (PSD) of the reverberant signals are estimated using multiple DNNs, and the rest of the parameters are estimated via regular application of the EM iterations.

The recursive EM (REM) is a version of the batch EM in which the iterations are substituted by time-recursion. Two variants of the REM are known in the literature. The variant proposed by Titterington [8] is based on the Newton method for maximizing the likelihood function. Cappé and Moulines [9] proposed a second variant based on the maximization of a recursive accumulation of the auxiliary function. On top of their computational efficiency, these variants facilitate the estimation of time-varying parameters encountered in dynamic scenarios.

For example, in [10] both a (constrained) Titterington recursvie EM (TREM) algorithm and a Cappé-Moulines recursive EM (CREM) algorithm were developed for multi-speaker localization. In [11] a dereverberation method using the CREM algorithm was proposed. The method recursively tracks the time-varying acoustic channels and concurrently estimates the dereverberated speech signal by applying a Kalman filter. An online DOA tracking using CREM was developed in [12].

Only a few methods addressing the challenging online BASS problem, can be found in the literature. A common approach is to simultaneously apply DOA tracking and signal separation [13, 14]. In [15], a DNN approach for single-channel online BASS was proposed. In [16, 17], it was proposed to map anechoic observations in the short-time Fourier transform (STFT) domain to an embedding space. Time-varying deep attractor networks (DANets) are then applied to cluster these representations to the various sources. Variational EM (VEM) approaches were developed in [18, 19], assuming the ATFs are random processes.

Another family of online signal processing algorithms is applying a recursive least-squares (RLS) procedure for tracking the major eigenvector(s) of the spatial correlation matrix of the received signals [20]. An extension that jointly estimate the number of signals (which is equivalent to the rank of the spatial correlation matrix) and the signals' subspace can be found in [21]. In [22] the subspace tracking paradigm was adopted for speakers extraction assuming prior knowledge on the activity patterns of speakers.

In the current contribution we extend our previous work [23] in order to address dynamic scenarios. We modify the iterative EM scheme into a recursive scheme by using the CREM procedure [9]. In the E-step of the proposed algorithm a MCWF is applied to separate the sources and in the M-step, the filter parameters are updated.

2. Problem Formulation

A scenario comprising D concurrent speakers, captured by J microphones in a reveberant and noisy environment, is addressed. The problem is formulated in the STFT domain, with $k \in \{0, \dots, K-1\}$ and $t \in \{0, \dots, T-1\}$ representing the frequency index and time-frame index, respectively, with T and K the total number of time-frames and frequency bands, respectively. Let $s_d(t,k)$ denote the clean and anechoic speech signal of the d-th speaker. The observed signal, as received by the microphones array, can be modelled as

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

where $\mathbf{h}_d(t,k)$ 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 zero-mean, complex-Gaussian random vector with time-invariant covariance matrix $\mathbf{Q}(k)$.

The EM formulation necessitates a definition of a *complete data* set from which the observations can be obtained by a non-

http://dx.doi.org/10.21437/Interspeech.2021-662

invertible transformation. Following [24, 25], we rewrite the observed signal as a sum of D noisy components. Under this formulation, 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 comprising D row-concatenated identity matrices, and $\mathbf{x}(t,k)$ is the complete data with the following components:

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

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 inherit the spatial structure of the full noise covariance matrix, and their contributions to the total noise level is determined by a set of scalar weights β_d, satisfying $\sum_d \beta_d = 1$:

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

The speech signals are modelled as zero-mean complex-Gaussian random variables, such that

$$s_d(t,k) \sim \mathcal{N}_c(s_d(t,k); 0, \phi_d(t,k)) \tag{7}$$

where $\phi_d(t,k)$ is the PSD of the d-th speaker.

Under this model, the set of unknown parameters is:

$$\boldsymbol{\theta} = \{\phi_d(t,k), \mathbf{h}_d(t,k)\}_{d=1}^{D}\ . \tag{8}$$

Note, that the noise covariance matrix is assumed to be a priori known. The probability density function (p.d.f.) of the observations is given by

$$f(\mathbf{y}(t,k)|\boldsymbol{\theta}) = \mathcal{N}_c\big(\mathbf{y}(t,k); \mathbf{0}, \mathbf{P}(t,k)\big)\ , \tag{9}$$

where

$$\mathbf{P}(t,k) = \sum_{d=1}^{D} \mathbf{\Lambda}_d(t,k) \tag{10}$$

with

$$\mathbf{\Lambda}_d(t,k) = \phi_d(t,k)\mathbf{h}_d(t,k)\mathbf{h}_d^H(k) + \mathbf{Q}_d(k)\,, \tag{11}$$

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

As all frequency bins are assumed independent, the derivation is carried out per-frequency bin. Henceforth, the frequency index k is be omitted for conciseness.

3. Iterative EM Algorithm

In this section we present a brief summary of the iterative EM algorithm that we proposed in [23]. This batch EM algorithm assumes time-invariant room impulse responses (RIRs), namely that the speakers (and the microphone array) are static. In addition, we assume that the entire data is available for processing at each time-step. In the E-step of the iterative EM procedure, the auxiliary function is calculated:

$$U\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\ell-1)}\big) = E\big\{\log f(\mathcal{X}; \boldsymbol{\theta}) \big| \mathcal{Y}; \boldsymbol{\theta}^{(\ell-1)}\big\}\,, \tag{12}$$

where $\mathcal{Y}$ and $\mathcal{X}$ are the observations and the complete data sets, defined for all time-frequency (TF) bins, respectively. In the M-step, the parameter set is estimated by maximizing the auxiliary function w.r.t. the parameters.

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, namely $\mathbf{Q}_d = \beta_d \cdot \sigma^2 \cdot \mathbf{I}$. Second, the ATFs are normalized, i.e. $\|\mathbf{h}_d\|^2 = 1$, to circumvent gain ambiguity issues. The parameter estimation procedure is detailed in [23] and is omitted here for brevity.

4. Recursive EM Algorithm

In this section we propose an online version of the BASS framework that is capable of tracking time-varying ATFs relating moving sources and the microphone array. The derivation is utilizing the REM framework proposed by Cappé and Moulines [9]. In the derivation we use the two simplifying assumptions discussed above.

4.1. Cappé and Moulines auxiliary function

The aggregated auxiliary function $Q[\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}(t)]$ at time t can be recursively estimated, as follows [9]:

$$\begin{aligned} Q[\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}(t)] = Q[\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}(t-1)] + \\ \gamma_t(q[\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}(t)] - Q[\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}(t-1)]) \end{aligned} \tag{13}$$

with

$$q[\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}(t)] = E\left\{\log f(\mathbf{x}(t,k); \boldsymbol{\theta}) | \mathbf{y}(t,k); \hat{\boldsymbol{\theta}}(t)\right\} \tag{14}$$

the instantaneous auxiliary function at time t, and γ_t a time-varying weight. We next set γ_t to a fixed value $1-\beta$, with β a forgetting factor, and rewrite (13) as

$$Q[\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}(t)] = (1-\beta)\sum_{\tau=1}^{t} \beta^{t-\tau} q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}(\tau)). \tag{15}$$

According to our problem formulation, the recursive auxiliary function is given by:

$$\begin{aligned} Q[\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}(t)] = -(1-\beta)\sum_{\tau=1}^{t} \beta^{t-\tau} \\ \times \sum_{k,d} \Big[J \cdot \log \sigma^2 + \log\big(\frac{\phi_d(\tau)}{\sigma^2}||\mathbf{h}_d(\tau)||^2 + 1\big) \\ + \frac{1}{\sigma^2}\mathrm{trace}(\widehat{\mathbf{x}_d(\tau)\mathbf{x}_d^H(\tau)}) - \frac{1}{\sigma^2}\frac{\frac{\phi_d(t)}{\sigma^2}\mathbf{h}_d^H(\tau)\widehat{\mathbf{x}_d(\tau)\mathbf{x}_d^H(\tau)}\mathbf{h}_d(\tau)}{\frac{\phi_d(\tau)}{\sigma^2}||\mathbf{h}_d(\tau)||^2 + 1}\Big]. \end{aligned} \tag{16}$$

According to [25], the first- and second-order statistics of $\mathbf{x}_d(t)$ can be estimated using the MCWF:

$$\hat{\mathbf{x}}_d(t) = \mathbf{\Lambda}_d(t)\mathbf{P}^{-1}(t)\mathbf{y}(t) \tag{17a}$$

$$\widehat{\mathbf{x}_d(t)\mathbf{x}_d^H(t)} = \mathbf{\Lambda}_d(t) - \mathbf{\Lambda}_d(t)\mathbf{P}(t)^{-1}\mathbf{\Lambda}_d(t) + \hat{\mathbf{x}}_d(t)\hat{\mathbf{x}}_d^H(t)\,. \tag{17b}$$

4.2. Recursive parameter estimation

The estimates of the ATFs and the PSDs are obtained by maximizing (16) w.r.t. $\phi_d(t), \mathbf{h}_d(t),\ d = 1, \ldots, D$.

PSD estimation: The estimation of $\phi_d(t),\ d = 1, \ldots, D$ requires special attention. Normally, the PSD of a speech signal is time-varying and independent of its value in other time-frames. Hence, maximizing (16) w.r.t. $\phi_d(t)$ would result in non-smooth and high-variance estimates. To circumvent this issue, we propose to first assume that the PSDs are time-invariant, namely $\phi_d(t) = \phi_d$. Then maximization of (16) w.r.t. ϕ_d results in a time-varying estimate:

$$\hat{\phi}_d(t) = \frac{\sum_{\tau=1}^{t} \beta_\phi^{t-\tau}\left(\mathbf{h}_d^H(\tau)\widehat{\mathbf{x}_d(\tau)\mathbf{x}_d^H(\tau)}\mathbf{h}_d(\tau) - \sigma^2\right)}{\sum_{\tau=1}^{t}\beta_\phi^{t-\tau}}. \tag{18}$$

Now, using $\sum_{\tau=1}^{t}\beta_\phi^{t-\tau} = \frac{1-\beta_\phi^t}{1-\beta_\phi}$, with $0 < \beta_\phi < 1$ the forgetting factor for the PSD estimation, and assuming $t \gg 1$, results in the following time-varying PSD estimate:

$$\hat{\phi}_d(t) = \beta_\phi\hat{\phi}_d(t-1) + (1-\beta_\phi)\left[\mathbf{h}_d^H(t)\widehat{\mathbf{x}_d(t)\mathbf{x}_d^H(t)}\mathbf{h}_d(t) - \sigma^2\right]. \tag{19}$$

Finally, to circumvent numerical issues, the update is applied only if $\mathbf{h}_d^H(t)\widehat{\mathbf{x}_d(t)\mathbf{x}_d^H(t)}\mathbf{h}_d(t) \geq \xi_{min}$.

ATF estimation: The ATFs can be inferred by calculating the derivative of (16) w.r.t. $\mathbf{h}_d,\ d = 1, \ldots, D$ and setting the result to zero. This yields the following expression:

$$\lambda\mathbf{h}_d(t) = \left[\beta_h\mathbf{R}_d(t-1) + (1-\beta_h)\frac{1}{\sigma^2}\frac{\hat{\phi}_d(t)}{\hat{\phi}_d(t)+\sigma^2}\widehat{\mathbf{x}_d(t)\mathbf{x}_d^H(t)}\right]\mathbf{h}_d(t)\,, \tag{20}$$

with $0 < \beta_h < 1$ the forgetting factor ATFs estimation. We identify $\mathbf{R}_d(t)$ as the recursively aggregated second-order statistics matrix of the d-th source weighted by the corresponding single channel Wiener filter:

$$\mathbf{R}_d(t) = (1-\beta_h)\frac{1}{\sigma^2}\sum_{\tau=1}^{t}\beta_h^{t-\tau}\frac{\hat{\phi}_d(\tau)}{\hat{\phi}_d(\tau)+\sigma^2}\widehat{\mathbf{x}_d(\tau)\mathbf{x}_d^H(\tau)}. \tag{21}$$

The estimation procedure of $\mathbf{h}_d(t)$ boils down to solving the eigenvalue decomposition (EVD) problem and selecting the eigenvector with the largest eigenvalue of the matrix $\mathbf{R}_d(t)$. The normalization by σ^2 does not have any influence on the eigenvector.[1] Unlike the iterative EM scheme [23], this matrix is recursively calculated, thus enabling online estimation.

[1] If the simplifying assumption that the noise is spatially-white does not hold, the EVD should be substituted by a generalized EVD.

4.3. Efficient implementation

While the application of the EVD imposes high computational load, finding the most dominant eigenvector of a covariance matrix can be executed much more efficiently. One of the well-known procedures for estimating the dominant eigenvectors is the projection approximation subspace tracking (PAST) algorithm [20, 21].

The PAST procedure for estimating the dominant eigenvector is given by:

$$w_d(t) = \hat{\mathbf{h}}_d^H(t-1)\hat{\mathbf{x}}_d(t) \tag{22a}$$

$$l_d(t) = \beta_h l_d(t-1) + |w_d(t)|^2 \tag{22b}$$

$$\hat{\mathbf{h}}_d(t) = \hat{\mathbf{h}}_d(t-1) + \frac{1}{l_d(t)}\Big(\hat{\mathbf{x}}_d(t) - \hat{\mathbf{h}}_d(t-1)w_d(t)\Big)w_d^*(t) \tag{22c}$$

To gain more insight on the update procedure, we substitute (22a) into (22c), resulting in:

$$\begin{aligned}\hat{\mathbf{h}}_d(t) &= \hat{\mathbf{h}}_d(t-1) + \\ &\frac{1}{l_d(t)}\Big(I - \hat{\mathbf{h}}_d(t-1)\hat{\mathbf{h}}_d^H(t-1)\Big)\hat{\mathbf{x}}_d(t)\hat{\mathbf{x}}_d^H(t)\hat{\mathbf{h}}_d(t-1) \\ &= \hat{\mathbf{h}}_d(t-1) + \frac{w_d^*(t)}{l_d(t)}\mathbf{N}_{\hat{\mathbf{h}}_d(t-1)}\hat{\mathbf{x}}_d(t)\end{aligned} \tag{23}$$

where $\mathbf{N}_{\hat{\mathbf{h}}_d(t-1)}$ is the projection matrix to the orthogonal space of $\hat{\mathbf{h}}_d(t-1)$. In the transition from the second to the third line of (23), we used the normalization constraint $\|\hat{\mathbf{h}}_d(t-1)\| = 1$. The update rule progresses in a direction orthogonal to the current estimate with step-size determined by the instantaneous projection of the current estimate of the ATF on the current estimate of the signal, normalized by the time-averaged value of this projection.

5. Practical Considerations

Several practical measures are taken to facilitate proper application of the proposed method, as described below.

Source activity detection: Permutation ambiguity, especially if encountered independently for each TF, results in severe degradation in the separation performance. In dynamic scenarios, involving both time-varying ATFs and intricate activity patterns of the speakers, this phenomenon may occur frequently. To circumvent this permutation problem and to allow smooth tracking of each ATF, we propose a two-step procedure.

In the first step, each TF bin is classified to either speech-active or noise-dominant classes. This is obtained by adopting the DNN-based speech presence probability (SPP) estimator presented in [26]. Only TF bins with SPP larger than a predefined threshold are classified as speech-active and the other TF bins are discarded from further processing.

In the second step, each speech-active TF is associated with a dominant speaker by projecting the observation vector on the current estimate of each ATF. Hence, the d-th ATF will only be updated if the TF is indeed associated with the d-th speaker. We therefore refine (23) by incorporating an activity indicator $\mathbb{1}_d(t)$ into the update rule:

$$\hat{\mathbf{h}}_d(t) = \hat{\mathbf{h}}_d(t-1) + \mathbb{1}_d(t)\frac{w_d^*(t)}{l_d(t)}\mathbf{N}_{\hat{\mathbf{h}}_d(t-1)}\hat{\mathbf{x}}_d(t)\,. \tag{24}$$

The indicator is evaluated (per frequency bin) according to:

$$\mathbb{1}_d(t) = \begin{cases} 1 & \text{SPP}(t) \geq \alpha_i \text{ and} \\ & \hat{\mathbf{h}}_d^H(t-1)\mathbf{y}(t) \geq \hat{\mathbf{h}}_{d'}^H(t-1)\mathbf{y}(t) + \alpha_r, \\ & \forall d' \in \{1,\ldots,D\}\backslash d \\ 0 & \text{otherwise} \end{cases}$$

where $0 \leq \alpha_i \leq 1$ is a speech activity threshold and α_r a margin. In addition, the PSD estimate is modified to $\bar{\phi}_d(t)$

$$\bar{\phi}_d(t) = \mathbb{1}_d(t)\hat{\phi}_d(t) + \bar{\mathbb{1}}_d(t)\xi_{min}, \tag{25}$$

with $\bar{\mathbb{1}}_d(t)$ the indicator of the complement subset. This modification facilitates *warm start* of the estimator when the activity of a speaker resumes.

Initialization of the REM algorithms: The EM algorithm is notorious for its sensitivity to initialization. We propose the following initialization procedure. We first assume that the sources are static or moving slowly in the first T_0 seconds of the utterance. During this period we apply the batch EM algorithm summarized in Sec. 3 (as elaborated in [23]) to obtain an initial value of the PSDs of all sources and their respective ATFs.

Post-processing: Since the estimates $\hat{\mathbf{x}}_d(t)$, $d = 1,\ldots,D$ of the complete data partially include by construction a noise component, a post-processing stage is required. We adopt the procedure proposed in [23] and apply a minimum variance distortionless response (MVDR) beamformer to $\hat{\mathbf{x}}_d$ using the estimated ATFs. Under the simplifying assumptions presented in Sec. 3, the MVDR simplifies to a matched-filter beanformer:

$$\hat{s}_d(t,k) = \hat{\mathbf{h}}_d^H(t,k)\hat{\mathbf{x}}_d(t,k)\,. \tag{26}$$

6. Simulation Results

Evaluation setup: The proposed algorithm was evaluated and compared with the iterative EM algorithm [23] using the following simulation setup. The speech signals were randomly drawn from 26 speakers from the TIMIT database [27]. Speech utterances of the same speaker were concatenated to obtain a 15 s long speech signal. A scenario with two moving speakers was simulated using a signal generator[2] based on the image method [28]. In the tested scenario, the speakers moved along an arc with 2 m radius around the microphone array. Additional trajectories, not reported here, were also tested, demonstrating consistent results. At the first $T_0 = 3$ sec the speakers were static and at the rest of the utterance, the velocity of the two speakers was set to 10 degrees per second (equivalent to 0.35 m/sec). An angular distance 50° between the two speakers was maintained throughout the entire utterance. The room dimensions were set to $6\times6\times3$ m and the reverberation time to $T_{60} = 160, 200$ ms. The signals were captured by an eight microphone linear array with inter-distances of $[3,3,3,8,3,3,3]$ cm, together with an additive spatially-white babble noise using [29] with signal-to-noise ratio (SNR) value of 10 dB. The sampling frequency of the signals was 16 kHz. The STFT frame-size was 32 ms with 75% overlap. The values of the various parameters chosen for this experimental study are depicted in Table 1.

Performance measures and results: Separation results were evaluated using the signal-to-interference ratio (SIR) measure from the BSSeval toolbox [30]. The speech intelligibility was evaluated using the short-time objective intelligibility (STOI) measure [31]. We compared the proposed recursive method with the baseline batch EM method [23], which ignores the dynamics of the problems and assumes the sources to be static. The performance measures, depicted in Table 2, are averaged over 25 scenarios and the two reverberation levels. As can be expected, the batch EM algorithm cannot handle the moving-speakers scenario while the proposed method significantly improves the speech quality and intelligibility. Sound examples are available in our website.[3]

Table 1: *Parameters.*

Parameter	β_h	β_ϕ	α_i	α_r	ξ_{min}	T_0
Value	0.6	0.4	0.5	0.05	$0.5\sigma^2$	3 sec.

Table 2: *Performance measures of the proposed REM algorithm and the baseline batch EM [23].*

	Input	EM	REM
SIR [dB]	0.1	5.6	10.7
STOI [%]	67.1	67.3	87

In order to assess the quality of the estimated ATFs, we extracted the DOAs of the sources using the following procedure. First, the relative transfer functions (RTFs) are constructed by normalizing all estimated ATFs by the ATF associated with the reference microphone. Then, by transforming the obtained RTFs back to the time-domain and picking their major peaks, the time difference of arrivals (TDOAs) can be inferred. From the TDOAs it is straightforward to calculate the sources' DOAs using the given array-source constellation. The tracking results are presented in Fig. 1, demonstrating the ability of the proposed algorithm to accurately track the two DOA trajectories.

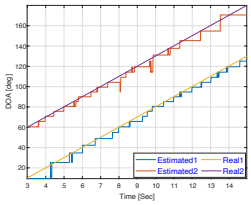

Figure 1: *DOA estimation of moving speakers.*

7. Conclusion

In this paper, an online EM algorithm for BASS was presented. In the E-step, the instantaneous first- and second-order statistics of the complete-data are calculated. In the M-step, the PSDs of the signals and the ATFs between the sources and the microphones are inferred. The ATFs are calculated using EVD, reminiscent of the estimation procedure in the batch EM algorithm, but with recursively estimated correlation matrix. Alternatively, we proposed to utilize the computationally-efficient PAST procedure. Both alternatives provided similar separation results. A post-filtering stage, implemented as a matched-filter beamformer, is then applied to estimate the clean speech signals. The algorithm was evaluated in simulated dynamic scenarios, and demonstrated significant improvements in terms of separation capabilities and speech intelligibility.

8. Acknowledgements

The project received funding from the EU Horizon 2020 Research and Innovation Programme, Grant Agreement #871245.

[2] https://github.com/ehabets/Signal-Generator

[3] https://www.eng.biu.ac.il/gannot/speech-enhancement/

9. 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] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

[3] S. Makino, Ed., *Audio source separation*, ser. Signals and Communication Technology. Springer International Publishing, 2018.

[4] 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.

[5] 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.

[6] 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.

[7] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[8] D. M. Titterington, "Recursive parameter estimation using incomplete data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 46, no. 2, pp. 257–267, 1984.

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

[10] O. Schwartz and S. Gannot, "Speaker tracking using recursive em algorithms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 392–402, 2013.

[11] B. Schwartz, S. Gannot, and E. A. 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, 2014.

[12] K. Weisberg, S. Gannot, and O. Schwartz, "An online multiple-speaker doa tracking using the Cappé-Moulines recursive expectation-maximization algorithm," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 656–660.

[13] M. Taseska and E. A. Habets, "Blind source separation of moving sources using sparsity-based source detection and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 657–670, 2017.

[14] T. Higuchi, N. Takamune, T. Nakamura, and H. Kameoka, "Underdetermined blind separation and tracking of moving sources based on DOA-HMM," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3191–3195.

[15] C. Han, Y. Luo, and N. Mesgarani, "Online deep attractor network for real-time single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 361–365.

[16] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 246–250.

[17] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[18] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "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.

[19] Y. Laufer and S. Gannot, "A bayesian hierarchical model for speech enhancement with time-varying audio channel," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 225–239, 2018.

[20] B. Yang, "Projection approximation subspace tracking," *IEEE Transactions on Signal processing*, vol. 43, no. 1, pp. 95–107, 1995.

[21] ——, "An extension of the PASTd algorithm to both rank and subspace tracking," *IEEE Signal Processing Letters*, vol. 2, no. 9, pp. 179–182, 1995.

[22] S. Markovich-Golan, S. Gannot, and I. Cohen, "Subspace tracking of multiple sources and its application to speakers extraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 201–204.

[23] A. Eisenberg, B. Schwartz, and S. Gannot, "Blind audio source separation using two expectation-maximization algorithms," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2020.

[24] 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.

[25] ——, "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.

[26] S. E. Chazan, J. Goldberger, and S. Gannot, "A hybrid approach for speech enhancement using mog model and neural network phoneme classifier," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2516–2530, 2016.

[27] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus cd-rom. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[28] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[29] E. A. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2911–2917, 2008.

[30] 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.

[31] 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.