AN ONLINE MULTIPLE-SPEAKER DOA TRACKING USING THE CAPPÉ-MOULINES RECURSIVE EXPECTATION-MAXIMIZATION ALGORITHM

Koby Weisberg, Sharon Gannot

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

Ofer Schwartz

CEVA-DSP
Audio Department, Herzelia, Israel

ABSTRACT

In this paper, we present a multiple-speaker direction of arrival (DOA) tracking algorithm with a microphone array that utilizes the recursive EM (REM) algorithm proposed by Cappé and Moulines. In our model, all sources can be located in one of a predefined set of candidate DOAs. Accordingly, the received signals from all microphones are modeled as Mixture of Gaussians (MoG) vectors in which each speaker is associated with a corresponding Gaussian. The localization task is then formulated as a maximum likelihood (ML) problem, where the MoG weights and the power spectral density (PSD) of the speakers are the unknown parameters. The REM algorithm is then utilized to estimate the ML parameters in an online manner, facilitating multiple source tracking. By using Fisher-Neyman factorization, the outputs of the minimum variance distortionless response (MVDR)-beamformer (BF) are shown to be sufficient statistics for estimating the parameters of the problem at hand. With that, the terms for the E-step are significantly simplified to a scalar form. An experimental study demonstrates the benefits of the using proposed algorithm in both a simulated data-set and real recordings from the acoustic source localization and tracking (LOCATA) data-set.

***Index Terms*—** Speaker tracking, Recursive expectation-maximization, LOCATA challenge

1. INTRODUCTION

Online speaker tracking is required in many applications, including navigation, source separation and target acquisition. This task becomes challenging when multiple moving speakers are concurrently active, as well as when additive inference sources are captured by the microphone array.

In this paper we focus on the DOA estimation problem. In the audio processing community, common DOA estimators are based on the steered response power (SRP)-phase transform (PHAT) algorithm [1] or the multiple signals classification (MUSIC) algorithm [2]. However, these techniques are not optimal in the multiple-speaker case.

In [3], the expectation-maximization (EM) algorithm was utilized to estimate the DOAs of multiple static speakers with a microphone pair. Assuming a single dominant speaker in each time-frequency (TF) bin, the interaural phase differences (IPDs) from all TF bins were clustered into groups associated with a candidate speaker. The DOA of the active speakers were estimated using the groups with the highest probability. In [4], two REM versions were applied to a multichannel extension of the model in [3]: one based on Titterington recursive EM (TREM) [5] and the second on Cappé and Moulines recursive EM (CREM) [6]. The model in this approach does not directly address additive noise.

In [7, 8, 9], the phase-related feature vectors were substituted by the raw short-time Fourier transform (STFT) observations. In addition, the noise (or reverberation) was implicitly modeled and therefore improved results were obtained in noisy scenarios. The observations were modelled as a mixture of high-dimensional complex-Gaussian with zero-mean, and a spatial covariance matrix that reflects both the speech and the noise power spectral densities (PSDs). In [8], speaker localization and separation procedures for the noisy case were presented. It was shown that the PSDs of the candidate speakers can be estimated in advance (prior to the application of EM) from the output of an MVDR-BF.

In the current contribution, we extend [7, 8, 9] to address the dynamic scenario. For that, we first show that the E-step can be recast in a scalar form, rather than a vector form, which results in a lower computational burden necessary for online and real-time tracking problems. By applying the Fisher-Neyman factorization [10], it is shown that the MVDR-BF outputs can substitute the raw observation features and serve as a sufficient statistics for estimating the parameters of the problem at hand. Next, a tracking procedure is proposed by applying the CREM algorithm. Recursive equations for the DOA probabilities and the candidate speakers PSD are derived, which facilitates online DOA tracking of multiple speakers.

This proposed tracking algorithm was evaluated using both simulated room impulse responses (RIRs) and real recordings from the LOCATA data-set. Improved estimates of the speakers trajectories, compared with baseline methods, are demonstrated.

2. PROBLEM FORMULATION

In this section, the statistical model is formulated, and the associated ML is stated.

2.1. Signal model

Consider a grid of candidate positions in the room $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_M\}$, where $M = |\mathcal{P}|$ is the number of candidates. Note that, according to this model, the number of speakers is equal to the number of grid points. As will be clarified below, the number of actual speakers is always significantly lower than M.

The speech signals, together with an additive noise, are captured by an array of N microphones. The nth microphone signal in the STFT domain is given by:

$$z_n(t,k) = \sum_{m=1}^{M} d_m(t,k) g_{m,n}(k) s_m(t,k) + v_n(t,k) \qquad (1)$$

where $t = 0, \ldots, T-1$ denotes the time index, $k = 0, \ldots, K-1$ denotes the frequency index, $g_{m,n}(k)$ denotes the direct-path transfer function from the speaker positioned at $\mathbf{p}_m$ to microphone n

(relative to the reference microphone arbitrarily chosen as microphone #1), $s_m(t,k)$ denotes the speech signal uttered by a speaker at grid point m (as received by the reference microphone), and $v_n(t,k)$ denotes the ambient noise. The indicator signal $d_m(t,k)$ indicates whether speaker m is active in the (t,k)th TF bin:

$$d_m(t,k) = \begin{cases} 1, & \text{if speaker } m \text{ is active in TF bin } (t,k) \\ 0, & \text{otherwise} \end{cases}. \quad (2)$$

Note that, according to the sparsity assumption [11], the vector $\mathbf{d}(t,k) = \text{vec}_m\{d_m(t,k)\} \in \{\mathbf{e}_1, \ldots, \mathbf{e}_M\}$, where $\mathbf{e}_m$ is a "one-hot" vector, namely equals '1' in its mth entry, and zero elsewhere. The N microphone signals can be concatenated in a vector form:

$$\mathbf{z}(t,k) = \sum_{m=1}^{M} d_m(t,k)\mathbf{g}_m(k)s_m(t,k) + \mathbf{v}(t,k),$$

where $\mathbf{z}(t,k)$, $\mathbf{g}_m(k)$ and $\mathbf{v}(t,k)$ are the respective concatenated vectors. The transfer function of the direct-arrival is given by:

$$g_{m,n}(k) = \exp\left(-\iota\frac{2\pi k}{K}\frac{\tau_{m,n}}{T_s}\right) \quad (3)$$

where T_s denotes the sampling period, and $\tau_{m,n}$ denotes the time difference of arrival (TDOA) between position $\mathbf{p}_m$ and microphone n. This TDOA can be calculated in advance from the predefined grid points and the array constellation.

2.2. Statistical model and the ML problem

Both the speech and the noise signals are modeled as zero-mean complex-Gaussian random vectors:

$$\mathbf{v}(t,k) \sim \mathcal{N}\left(\mathbf{v}(t,k), \mathbf{0}, \Phi_{\mathbf{v}}(k)\right), \quad (4)$$

$$s_m(t,k) \sim \mathcal{N}\left(s_m(t,k), 0, \phi_{s,m}(t,k)\right). \quad (5)$$

The probability density function (p.d.f.) of the indicator vector $\mathbf{d}(t,k)$ is given by:

$$f_{\mathbf{d}}(\mathbf{d}(t,k)) = \sum_{m=1}^{M} d_m(t,k)\psi_m \quad (6)$$

where ψ_m is the a priori probability of the activity of a speaker at the mth position, and $\sum_{m=1}^{M}\psi_m = 1$. Because the actual number of speakers is usually lower than the number of candidates, most of ψ_m will be close to zero [12]. Following the sparsity assumption, the observation vectors are distributed as a mixture of M zero-mean complex-Gaussians:

$$f_z(\mathbf{z}(t,k)) = \sum_{m=1}^{M} \psi_m\mathcal{N}(\mathbf{z}(t,k), \mathbf{0}, \Phi_{\mathbf{z},m}(t,k)) \quad (7)$$

where the PSD matrix of each Gaussian is given by:

$$\Phi_{\mathbf{z},m}(t,k) = \mathbf{g}_m(k)\mathbf{g}_m^H(k)\phi_{s,m}(t,k) + \Phi_{\mathbf{v}}(k), \quad (8)$$

and the speech and noise vectors are assumed to be statistically independent. Finally, by assuming TF bin independency, we obtain the p.d.f. of the entire set of observations:

$$f(\mathbf{z};\boldsymbol{\theta}) = \prod_{t,k}\sum_{m=1}^{M} \psi_m\mathcal{N}(\mathbf{z}(t,k), \mathbf{0}, \Phi_{\mathbf{z},m}(t,k)) \quad (9)$$

where $\mathbf{z} = \text{vec}_{t,k}\{\mathbf{z}(t,k)\}$, and $\boldsymbol{\theta}$ is the set of unknown parameters, namely $\boldsymbol{\theta} = \left[\boldsymbol{\psi}^T, \boldsymbol{\phi_s}^T\right]^T$ with $\boldsymbol{\psi} = \text{vec}_m\{\psi_m\}$ and $\boldsymbol{\phi_s} = \text{vec}_{t,k,m}\{\phi_{s,m}(t,k)\}$. Note that $\Phi_{\mathbf{v}}(k)$ is assumed to be known in advance, or can be estimated with no speech activity. The ML problem can readily be stated as: $\widehat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} \log f(\mathbf{z};\boldsymbol{\theta})$.

3. MAXIMUM LIKELIHOOD ESTIMATION

In this section, we derive the maximum likelihood estimator (MLE) of the $\boldsymbol{\theta}$. In the following, we omit the frequency index k for brevity.

3.1. Fisher-Neyman factorization

According to the Fisher-Neyman factorization theorem [10], the p.d.f. of $\mathbf{z}$ (given that the dominant speaker is located at $\mathbf{p}_m$) can be factorized as:

$$\mathcal{N}(\mathbf{z}(t), \mathbf{0}, \Phi_{\mathbf{z},m}(t)) = \mathcal{N}(\hat{s}_{m,\text{MVDR}}(t), 0, \phi_{s,m} + \phi_{v,m})h\left(\mathbf{z}(t)\right), \quad (10)$$

where $\hat{s}_{m,\text{MVDR}}(t) \equiv \mathbf{w}_m^H\mathbf{z}(t)$ is an estimate of the speech using the MVDR-BF, $\mathbf{w}_m = \frac{\Phi_{\mathbf{v}}^{-1}\mathbf{g}_m}{\mathbf{g}_m^H\Phi_{\mathbf{v}}^{-1}\mathbf{g}_m}$, which constitutes a sufficient statistic for estimating the speech PSD $\phi_{s,m}(t)$ given the observations $\mathbf{z}(t)$. The $\phi_{v,m} \equiv \frac{1}{\mathbf{g}_m^H\Phi_{\mathbf{v}}^{-1}\mathbf{g}_m}$ parameter is the PSD of the residual noise at the output of the MVDR-BF. The function $h\left(\mathbf{z}(t)\right)$ is independent of $\phi_{s,m}(t)$, and is given by:

$$h\left(\mathbf{z}(t)\right) = \frac{\phi_{v,m}}{\pi^{N-1}}\exp\left(-\mathbf{z}^H(t)\Phi_{\mathbf{v}}^{-1}\mathbf{z}(t) + \frac{|\hat{s}_{m,\text{MVDR}}(t)|^2}{\phi_{v,m}}\right). \quad (11)$$

3.2. Localization using Batch EM

In this section we review the results presented in [8] using the factorization (10). We will define $d_m(t,k)$ as the hidden data. The auxiliary function of the EM algorithm is given by:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\ell-1)}) = E\left\{\log\left(f(\mathbf{z},\mathbf{d};\boldsymbol{\theta})\right)|\mathbf{z};\boldsymbol{\theta}^{(\ell-1)}\right\} \quad (12)$$

where the joint p.d.f. of the observations and the hidden data (the complete data) is given by:

$$f(\mathbf{z},\mathbf{d};\boldsymbol{\theta}) = f(\mathbf{z}|\mathbf{d};\boldsymbol{\phi}_s)f(\mathbf{d};\boldsymbol{\psi}) = \prod_{t,k}\sum_{m=1}^{M}\psi_m d_m(t)\mathcal{N}(\mathbf{z}(t), 0, \Phi_{\mathbf{z},m}(t)). \quad (13)$$

The E-step is then given by:

$$\hat{d}_m^{(\ell-1)}(t) = E\left\{d_m(t)|\mathbf{z}(t);\boldsymbol{\theta}^{(\ell-1)}\right\} = \frac{\psi_m^{(\ell-1)}\mathcal{N}(\mathbf{z}(t), \mathbf{0}, \Phi_{\mathbf{z},m}^{(\ell-1)}(t))}{\sum_m \psi_m^{(\ell-1)}\mathcal{N}(\mathbf{z}(t), \mathbf{0}, \Phi_{\mathbf{z},m}^{(\ell-1)}(t))}, \quad (14)$$

and the M-step by:

$$\hat{\psi}_m^{(\ell)} = \frac{\sum_{t,k}\hat{d}_m^{(\ell-1)}(t,k)}{T\cdot K} \quad (15)$$

and:

$$\hat{\phi}_{s,m}(t,k) = |\hat{s}_{m,\text{MVDR}}(t,k)|^2 - \phi_{v,m}(k). \quad (16)$$

Note that, since $\hat{\phi}_{s,m}$ is independent of the outcome of the E-step, it can be calculated prior to the application of the EM iterations.

We will now simplify the E-step term in (14) to reduce the computational complexity and to gain some insights. Using the factorization in (10 the estimate of the indicator is given by:

$$\hat{d}_m^{(\ell-1)}(t) = \frac{\psi_m^{(\ell-1)} T_m(t)}{\sum_m \psi_m^{(\ell-1)} T_m(t)}. \tag{17}$$

Following several algebraic steps[1] we obtain:

$$T_m(t) = \frac{1}{\text{SNR}_m^{\text{post}}(t)} \exp\left(\text{SNR}_m^{\text{post}}(t) - 1\right) \tag{18}$$

where $\text{SNR}_m^{\text{post}}(t) = \frac{|\hat{s}_{m,\text{MVDR}}(t)|^2}{\phi_{v,m}}$ is the posterior signal-to-noise ratio (SNR) of a signal from the mth candidate position. Note that $T_m(t)$ is the likelihood ratio test (LRT), as presented in [13, Eq. (14)], where we have substituted the a priori SNR with its instantaneous estimator $\frac{\hat{\phi}_{s,m}(t)}{\phi_{v,m}}$ using (16). The LRT tests whether $\mathbf{z}(t)$ is either associated with the m-th candidate speaker or with a noise-only candidate position. Using this interpretation of $T_m(t)$, the association of each TF bin with each speaker candidate m (as indicated in (17)) is proportional to the corresponding LRT result and the prior probability of the mth speaker, as deduced from the previous iteration, $\psi_m^{(\ell-1)}$.

4. RECURSIVE EM

In this section, we will apply the CREM algorithm, presented in [6], to the problem at hand. To allow for a smooth estimate of the speech PSD, we assume here that ϕ_s is time-independent, keeping in mind that the (smooth) time-variations of the speech PSD will be naturally obtained by the application of the CREM. In the CREM scheme, the iteration index ℓ is substituted by the time index t, and the recursive auxiliary function is based on smoothing of the instantaneous auxiliary function over time:

$$Q_R(t;\boldsymbol{\theta}) = (1-\gamma) Q_R(t;\boldsymbol{\theta}) + \gamma Q(\boldsymbol{\theta}|\boldsymbol{\theta}(t-1)) \tag{19}$$

where $Q_R(t;\boldsymbol{\theta})$ is the recursive auxiliary function, and $Q(\boldsymbol{\theta}|\boldsymbol{\theta}(t-1))$ is the instantaneous auxiliary function given only the current observations. The M-step is obtained by maximizing $Q_R(t;\boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$. Using (12) and (13) the recursion in (19) boils down to:

$$\eta_m(t) = (1-\gamma)\eta_m(t-1) + \gamma \hat{d}_m(t), \tag{20a}$$

$$\xi_m(t) = (1-\gamma)\xi_m(t-1) + \gamma \hat{d}_m(t)\,|\hat{s}_{m,\text{MVDR}}(t)|^2. \tag{20b}$$

Maximizing $Q_R(t;\boldsymbol{\theta})$ with respect to ψ_m and $\phi_{s,m}$ yields the M-step:

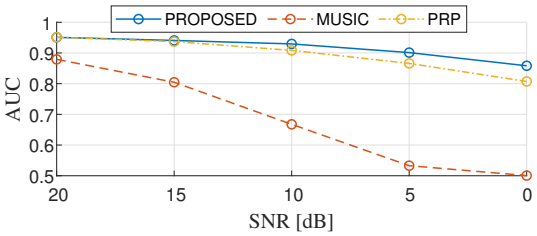
$$\hat{\psi}_m(t) = \frac{\sum_k \eta_m(t,k)}{K} \tag{21}$$

$$\hat{\phi}_{s,m}(t,k) = \frac{\xi_m(t,k)}{\eta_m(t,k)} - \phi_{v,m}(k). \tag{22}$$

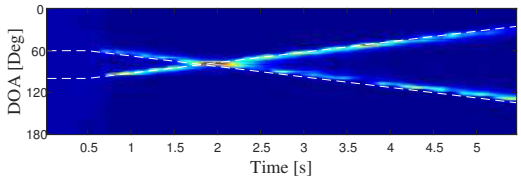
A recursive estimator of $\hat{d}_m(t,k)$ can be obtained from the CREM by substituting $\hat{\psi}_m^{(l-1)}$ with $\hat{\psi}_m(t-1)$ in (17) and by using a smoothed estimator for the a priori SNR in the LRT expression:

$$T_m(t) = \frac{1}{1+\text{SNR}_m^{\text{pri}}(t-1)} \exp\left(\frac{\text{SNR}_m^{\text{pri}}(t-1)\text{SNR}_m^{\text{post}}(t)}{1+\text{SNR}_m^{\text{pri}}(t-1)}\right) \tag{23}$$

[1] see supplementary material in `www.eng.biu.ac.il/gannot/publications/conferences-and-workshops-proceedings/`



(a) AUC vs. SNR for the proposed method and for the reference methods.



(b) An example probability map for SNR $= 25$ dB and sources velocities $\pm 15 \frac{\text{deg}}{\text{s}}$, respectively. The dashed line is the ground truth DOA. The obtained AUC ≈ 0.96.

Fig. 1: Experimental results for simulated data.

with $\text{SNR}_m^{\text{pri}}(t-1) = \frac{\hat{\phi}_{s,m}(t-1)}{\phi_{v,m}}$. Note the significant differences between (23) and (18). While the former does not take into account the smoothness of the speech PSD, and hence uses only an instantaneous SNR estimate; the latter takes the smoothness of the PSD into account through the recursively estimated a priori SNR estimate. We also note that the a priori SNR estimate obtained here by the CREM procedure is very different from the estimators presnted in [13].

4.1. Practical considerations

The original CREM uses one smoothing parameter γ. We note that in our problem, the two parameters exhibit different time behaviors: while $\boldsymbol{\psi}$, which is related to the source position, is slowly time-varying, the speech PSD $\boldsymbol{\phi_s}(t)$ is rapidly changing. Therefore, in our experiments, we used two different smoothing parameters: γ_ψ and γ_{ϕ_s}. Accordingly, for estimating $\xi(t)$, we always used $\gamma_{\phi_s} \approx 1$. For $\eta_m(t)$, we used two estimators: the first one used $\gamma_{\phi_s} \approx 1$ to obtain an estimate for $\boldsymbol{\phi_s}(t)$ in (22), and the second used $\gamma_\psi \ll 1$ to obtain an estimate of $\boldsymbol{\psi}$ in (21).

5. PERFORMANCE EVALUATION

The proposed algorithm was evaluated using two data-sets: simulated time-varying scenes generated by a signal generator[2] and real multichannel audio recordings from the LOCATA challenge [14].

5.1. Algorithm settings and baseline methods

The parameters used in the implementation of our algorithm are as follows: 1) signals re-sampled to 16 kHz; 2) STFT frame-length 64 ms with no overlap; 3) frequency band used for localization $1-6$ KHz; 4) smoothing parameters $\gamma_\psi = 0.1$, $\gamma_{\phi_s} = 0.8$; 5) grid of possible azimuth angle between -90^0 and 90^0, with resolution 2° and 5° for the simulated data and LOCATA data-set, respectively; and 6) the probabilities were uniformly initialized to

[2] `www.audiolabs-erlangen.de/fau/professor/habets/software/signal-generator`

$\psi_m(0) = \frac{1}{M}, \forall m$. The noise PSD matrix was estimated using speech absence segment at the beginning of the recording, annotated manually for the LOCATA data-set.

The proposed method provides a probability map as a function of time and not directly the DOA estimates. For estimating the actual trajectory of the speakers, one should use a peak-selection method. To circumvent the effects of the peak-selection algorithm, we have chosen to calculate instead the receiver operating characteristic (ROC) curve for each frame and to use the area under the curve (AUC) as a measure. For calculating the ROC curve, all detections in the range around the true DOA, specifically $\text{DOA}_{\text{gt}} \pm 3^\circ$, are considered *true positive*. The final score is obtained by time-averaging of the per-frame AUC, excluding noise-only frames. For baseline methods, we used both the MUSIC algorithm [2], as provided by the challenge, and the PRP-REM algorithm [4] with the same smoothing parameter, and with fixed variance for all the Gaussians, $\sigma = 0.1$. For a fair comparison, the MUSIC results were similarly smoothed and normalized to obtain a pseudo-distribution.

5.2. Evaluation using simulated data

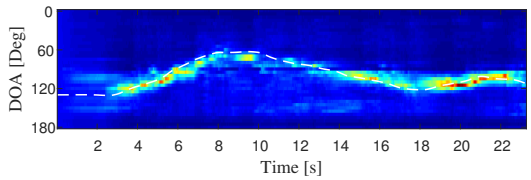
In the simulated scenario, clean anechoic speech signals were drawn from the TIMIT database [15], where speech utterances of the same speaker were concatenated to obtain a 5 s long speech signal. The speakers were randomly selected from 26 different speakers. To simulate moving sources, we used the signal generator, as mentioned above. The room dimensions were set to $6 \times 6 \times 6.1$ m with reverberation time $T_{60} \sim 200$ ms. The signals were captured by an eight-microphone linear array with inter-distances of $[3, 3, 3, 8, 3, 3, 3]$ cm from one another, together with an additive spatially-white noise with various SNR values.

Thirty Monte-Carlo trials, simulating two moving sources scenarios, were examined. In each scenario, the initial DOAs of the speakers were set to 60° and 100°, respectively. The sources moved from their initial positions in a circle with a radius of 1 m around the array center and with angular velocity randomly selected from a uniform distribution in the range $[-15 : 15]$ $\frac{\text{deg}}{\text{s}}$ to obtain random trajectories. We first examined the influence of γ_{ϕ_s} on the obtained localization score. We have noticed that the scores are insensitive to the smoothing parameter value in the range $0.6 < \gamma_{\phi_s} < 0.9$. We have therefore selected $\gamma_{\phi_s} = 0.8$ for all experiments.

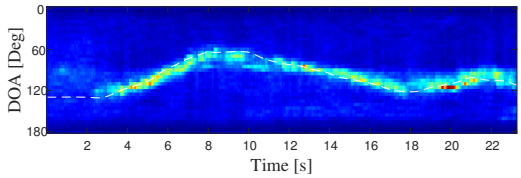
The results of the simulation study are depicted in Fig. 1. It is evident from Fig. 1(a) that the proposed algorithm outperforms the PRP-REM algorithm [4] by approximately 5% for 0 dB SNR, and that their performance converges as the SNR level increases. It is also demonstrated that the proposed method significantly outperforms the MUSIC algorithm. Moreover, we note that the proposed method is computationally more efficient than the PRP-REM, and that it additionally provides the speech PSD estimate that may be useful for further processing, e.g. in separation tasks [8]. In Fig. 1(b) we depict the probability map $\hat{\psi}_m$ of one the trials, clearly demonstrating the tracking capabilities of the proposed method.

5.3. Evaluation on LOCATA data-set

The data for the LOCATA challenge [14] were recorded in a room of size $7.1 \times 9.8 \times 3$ m with a reverberation time $T60 \sim 0.55s$. We tested our algorithm on Task #3, which is a recording of a single moving speaker, and Task #4, which is a recording of two moving speakers. We used the data recorded by the linear array (DICIT). We used the first recording (Recording #1) of each task. As a reference method, an implementation of the MUSIC algorithm was provided, as well as ground-truth location of the speakers. We evaluate our algorithm on the azimuth estimation only. The results of the LOCATA test are shown for the single source tracking task in Fig. 2 and for the two source tracking task in Fig. 3. The proposed method clearly outperforms MUSIC in both tasks, as can be deduced from the inspection of the probability maps and from the score values. The differences are more pronounced in the two speakers case, for which the MUSIC algorithm performs poorly.

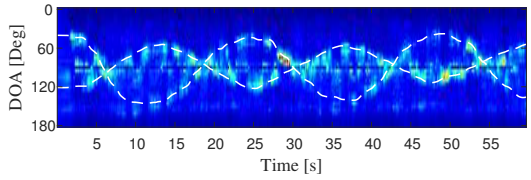


(a) Proposed method.

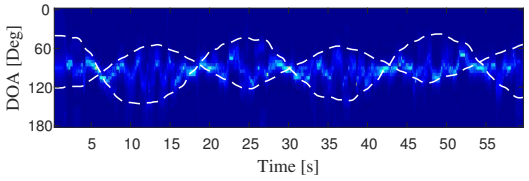


(b) PRP-REM algorithm [4]

Fig. 2: Probability maps for the LOCATA challenge (Task #3 - single moving speaker). The dashed line is the ground truth azimuth, as provided with the LOCATA database. AUC ≈ 0.95 for both methods.



(a) Proposed method.



(b) MUSIC.

Fig. 3: Probability maps for the LOCATA challenge (Task #4 - two moving speakers). The dashed line is the ground truth azimuth, as provided with the LOCATA database. AUC$= 0.82, 0.69$ for the proposed method and for the MUSIC algorithm, respectively.

6. CONCLUSIONS

A computationally efficient tracking algorithm, based on the CREM procedure, was proposed. An estimate of the speech presence probability in each candidate DOA is calculated from the respective MVDR-BF output in the E-step. In the M-step, both the direction and the speech PSD are recursively estimated. An experimental study demonstrates the advantage of the proposed algorithm compared to baseline methods, on both simulated data and recorded data.

7. REFERENCES

[1] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein, *Microphone arrays : signal processing techniques and applications*, chapter Robust Localization in Reverberant Rooms, pp. 157–180, Springer Verlag, 2001.

[2] Ralph Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[3] Michael I Mandel, Ron J Weiss, and Daniel PW 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.

[4] Ofer Schwartz and Sharon Gannot, "Speaker tracking using recursive em algorithms," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 2, pp. 392–402, 2014.

[5] Michael D Titterington, "Recursive parameter estimation using incomplete data," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 257–267, 1984.

[6] Olivier Cappé and Eric 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.

[7] Ofer Schwartz, Yuval Dorfan, Emanuël AP Habets, and Sharon Gannot, "Multi-speaker doa estimation in reverberation conditions using expectation-maximization," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016.

[8] Yuval Dorfan, Ofer Schwartz, Boaz Schwartz, Emanuël AP Habets, and Sharon Gannot, "Multiple DOA estimation and blind source separation using estimation-maximization," in *IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, 2016.

[9] Ofer Schwartz, Yuval Dorfan, Maja Taseska, Emanuël AP Habets, and Sharon Gannot, "DOA estimation in noisy environment with unknown noise power using the EM algorithm," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 86–90.

[10] Steven M Kay, *Fundamentals of statistical signal processing: Practical algorithm development: Estimation Theory*, vol. 1, Pearson Education, 2013.

[11] Ozgur Yilmaz and Scott Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[12] Xiaofei Li, Laurent Girin, Radu Horaud, Sharon Gannot, Xiaofei Li, Laurent Girin, Radu Horaud, and Sharon Gannot, "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1997–2012, 2017.

[13] Tao Yu and John HL Hansen, "A speech presence microphone array beamformer using model based speech presence probability estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 213–216.

[14] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Sheffield, UK, July 2018.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.