This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2013.2292361

Speaker Tracking Using Recursive EM Algorithms

Ofer Schwartz and Sharon Gannot

***Abstract*—The problem of localizing and tracking a known number of concurrent speakers in noisy and reverberant enclosures is addressed in this paper. We formulate the localization task as a maximum likelihood (ML) parameter estimation problem, and solve it by utilizing the expectation-maximization (EM) procedure.**

For the tracking scenario, we propose to adapt two recursive EM (REM) variants. The first, based on Titterington's scheme, is a Newton-based recursion. In this work we also extend Titterington's method to deal with constrained maximization, encountered in the problem at hand. The second is based on Cappé and Moulines' scheme. We discuss the similarities and dissimilarities of these two variants and show their applicability to the tracking problem by a simulated experimental study.

I. Introduction

In many scenarios, an estimation of the location of speakers is required. These scenarios may include navigation, surveillance, beamforming [1], source separation [2], target acquisition and tracking geared towards the steering of automated cameras [3].

In reverberant environments, secondary reflections, may result in biased location estimates. Multiple concurrent speakers scenarios, that require multiple estimations of several dominant directions, are more vulnerable to these secondary sound reflections. The movement of the speakers further complicates the localization problem, as the amount of data available in each location is limited, especially in multiple speaker scenarios. In these scenarios, tracking procedures, that maintain smooth trajectory of the speakers should be applied. The objective of this contribution is, therefore, to derive computationally efficient tracking algorithms for multiple sources that are not affected by the coexistence of the sources in reverberant and noisy environments.

The target of tracking (and localization) schemes can be either the coordinates of the speaker location, or the time difference of arrival (TDOA) between two signals observed by two adjacent microphones. Although this contribution is dedicated to coordinate localization, TDOA estimation methods will also be addressed, as some of the core techniques for solving both problems are common.

The mathematical relations between the coordinates of the speakers (or the respective TDOAs) and the observed signals is nonlinear and non-injective. Hence, the estimation cannot be achieved by linear estimators. Attempts to estimate the TDOA in the time-domain encounter difficulties in associating the observations with the various speakers. On the contrary, the separation between speakers' signals in the frequency-domain is achievable due to the sparseness of the speech signals. Therefore, the frequency domain will be the preferable choice for our estimation procedures.

A plethora of algorithms for speaker localization and tracking has been proposed. Some of them will be reviewed in the sequel. The simplest estimator for finding the TDOA between two observed signals is the cross-correlation method and its variants, mainly the generalized cross correlation (GCC) [4] with phase transform (PHAT) normalization. In [5], the authors presented the steered response power (SRP)-PHAT algorithm, which is the generalization of the GCC-PHAT to an array of microphones in the far-field scenarios. However, in the multiple speakers scenario, these estimators will not necessarily yield the required results. These basic techniques are widely used and can be incorporated into many algorithms. In the experimental study of the current contribution we will compare the proposed localization methods with the SRP-PHAT algorithm.

Two paradigms can be adopted in deriving tracking algorithms, namely the Bayesian and non-Bayesian families of estimators. By defining the trajectory, or the TDOAs time-sequences, as *stochastic* processes, Bayesian methods that aim at the minimum mean square error (MMSE) estimation can be utilized [6], [7], [8], [9], [10], [11]. In contrast, by defining the trajectory as a *deterministic* and time-varying parameter, the maximum likelihood estimator (MLE) can be used. Bayesian tracking algorithms are out of the scope of this contribution.

In this work we adopt the non-Bayesian framework. A deterministic system identification approach is taken in [12],[13] for estimating the room impulse response and in [14] for estimating the *relative transfer function*. The TDOA can be deduced from both estimates. Dvorkind and Gannot [14] also derive tracking procedure based on a recursive Gauss scheme.

In [15] the ML criterion was used for estimating the TDOA. In [16], the EM algorithm [17],[18] is adopted to evaluate the TDOA values of the sources. In the EM formulation in [16] the hidden data is defined as the individual source signals. As a result, the *E-step* simplifies to the MMSE estimator of each of the source signals, given the measured microphone signals at the current estimated TDOA. The *M-step* simplifies to a maximization of a steered beamformer output yielding new TDOA estimates for all speakers. The authors do not utilize the sparseness often attributed to speech signal activity patterns.

The EM framework presented above estimates the TDOAs of static sources. For moving sources the TDOA changes from frame to frame. Implementing the EM algorithm for each frame separately discards the smooth nature of the source trajectory. A REM algorithm can mitigate this problem. Two versions of the REM algorithm are described in the literature, one of which was proposed by Titterington [19] and the other by Cappé and Moulines [20]. Based on the model in [16],

Ofer Schwartz and Sharon Gannot are with the faculty of Engineering, Bar-Ilan University, Ramat-Gan, 52900, Israel (e-mail: ofer.shwartz@live.biu.ac.il, Sharon.Gannot@biu.ac.il)

Copyright (c) 2013 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2013.2292361

REM algorithms were derived to track moving sources. In [21] and in [22] the Titterington recursive EM (TREM) variant is applied and in [23] the Cappé and Moulines recursive EM (CREM) variant is applied.

In [24],[25] the EM algorithm is also adopted to estimate the direction of arrival (DOA) values. DOA candidates, taken by maximizing the SRP output at every frequency bin, are associated with the various speakers by applying (a variant) of mixture of Gaussians (MoG) clustering, based on the EM algorithm. Assuming disjoint activity of the speakers in the Fourier domain, it is possible to cluster these DOA candidates to various speakers.

In [26] a two microphone (binaural) scenario is considered and the model-based EM source separation and localization (MESSL) algorithm is derived. The phase ratios and the absolute value ratios of the microphone signals in the short-time Fourier transform (STFT) domain are associated with the speakers by utilizing the EM algorithm to perform MoG clustering. The mean of the Gaussians can be associated with any potential TDOA value for every speaker. Hypothetically, the model "duplicates" each speaker to multiple candidate locations. The number of the Gaussians in the MoG distribution is equal to the number of candidate locations. The means of these Gaussians are associated with all possible TDOA values. The E-step is responsible for determining a time-frequency mask, that is the extent to which each time-frequency bin associates with each of the Gaussians. The M-step determines the probabilities of the Gaussians by its number of associations. The final TDOA estimates can be taken as the TDOA value associated with the Gaussians with the largest probability. Another contribution [27] improves the estimation in the presence of reverberation.

The algorithm in [26] will be the starting point of our algorithm development. Our contribution is threefold: 1) generalization of [26] to the estimation of the coordinates of multiple sources, rather than only their associated TDOAs; 2) the development of REM schemes for constrained optimization problems (often encountered in tracking procedures); and 3) the development of efficient tracking algorithms for multiple speakers in reverberant environments.

The remainder of this paper is organized as follows. In Section II the problem of multiple speakers location estimation with a spatially distributed microphone constellation is defined. In Section III we summarize the EM framework developed in the context of *static* TDOA estimation [26] and generalize it to coordinate localization. This generalization will serve as the basis of the tracking algorithms developed in the subsequent sections. Sections IV and V are dedicated to the main contribution of this paper, the adaption of two versions of the REM procedure to the multiple speakers tracking scheme. A simulative experimental study, that demonstrates the tracking capabilities of the proposed algorithms, can be found in Section VI. Conclusions are outlined in Section VII.

II. Problem Formulation and the Probabilistic Model

Consider an array with M microphone pairs receiving signals from S speakers as illustrated in Fig. 1. The measured array output is a linear combination of the incoming waveforms, corrupted by additive Gaussian noise. Let z_m^i be the signals received by the ith microphone of pair m, where $i = 1, 2$ and $m = 1, \dots, M$. The signals in the STFT domain are given by

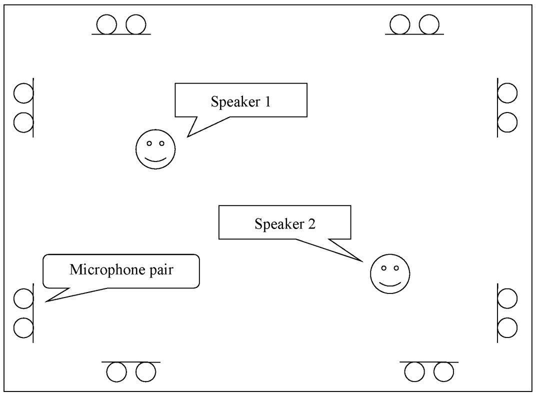


Fig. 1. Illustration of microphone and speaker constellation.

$$z_m^i(t,k) = \sum_{s=1}^{S} a_{sm}^i(t,k) \cdot v_s(t,k) + n_m^i(t,k) \tag{1}$$

where $t = 0, \dots, T-1$ denotes the time index and $k = 0, \dots, K-1$ denotes the frequency index. The speaker index is denoted $s = 1, \dots, S$. The number of speakers S assumed to be a priori known. $v_s(t,k)$ denotes the speech signal emanating from speaker s, $a_{sm}^i(t,k)$ denotes the acoustic transfer function (ATF) relating source s and microphone i in pair m, and $n_m^i(t,k)$ denotes additive noise as received by the microphones. Define

$$\mathbf{z} = \text{vec}_{m,i,t,k}\left(\{z_m^i(t,k)\}\right), \tag{2}$$

where the vec operation denotes the concatenation of all time-frequency observations received by all microphones. In low-reverberant environments the ATF can be approximated by the direct arrival alone:

$$a_{sm}^i(t,k) \backsimeq \frac{1}{||\mathbf{p}_s - \mathbf{p}_m^i||} \cdot \exp\left(-j\frac{2\pi k}{K}\frac{\tau_{sm}^i}{T_s}\right) \tag{3}$$

where T_s denotes the sampling period, $\mathbf{p}_s$ and $\mathbf{p}_m^i$ are the speaker s and microphone i in pair m locations, respectively, and τ_{sm}^i denotes the travel time between the respective speaker and microphone. $||\cdot||$ denotes the Euclidean norm. The travel time τ_{sm}^i is given by

$$\tau_{sm}^i = \frac{1}{c} \cdot \left(||\mathbf{p}_s - \mathbf{p}_m^i||\right) \tag{4}$$

where c is the sound velocity. Define the vector of the concatenated unknown source locations as:

$$\mathbf{p}^{\text{con}} = \begin{bmatrix} \mathbf{p}_1^T & \cdots & \mathbf{p}_S^T \end{bmatrix}^T.$$

Estimating these, possibly time-varying, parameters is the goal of this paper.

Copyright (c) 2013 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2013.2292361

Rather than using the measurements $\mathbf{z}$ directly we will use the pair-wise relative phase ratio (PRP) defined as:

$$\phi_m(t,k) \triangleq \frac{z_m^2(t,k)}{z_m^1(t,k)} \cdot \frac{|z_m^1(t,k)|}{|z_m^2(t,k)|}. \tag{5}$$

The observed measurement vector at time instant t and frequency bin k is constructed by augmenting the PRPs of all sensors:

$$\boldsymbol{\phi}(t,k) = \left[\begin{array}{ccc} \phi_1(t,k) & \cdots & \phi_M(t,k) \end{array} \right]^T. \tag{6}$$

The set $\boldsymbol{\phi}(t,k)$ will be designated the *observed data* in the EM formulation, defined in the sequel. The various speakers are assumed to exhibit disjoint activity in the STFT domain. Therefore, by means of a clustering process, every time-instant and frequency-bin of $\boldsymbol{\phi}(t,k)$ can be associated with a single active source.

We use MoG probability function to characterize the coordinates of the entire speakers set in the following way, adapted from [26]. The mean of each of the Gaussian is a candidate PRP associated with a specific coordinate on a predefined grid. The set of grid points is denoted $\mathcal{P}$. The total number of Gaussians is therefore $S \times |\mathcal{P}|$. Each speaker is attributed with a subset of the $|\mathcal{P}|$ Gaussians. Hence, localization boils down to selecting the most probable Gaussian per speaker.

Based on the disjoint activity of the sources, we give the observations the following probabilistic description:

$$\boldsymbol{\phi}(t,k) \sim \sum_{s,\mathbf{p}} \psi_{s\mathbf{p}} \cdot \mathcal{N}^c(\boldsymbol{\phi}(t,k); \tilde{\boldsymbol{\phi}}^k(\mathbf{p}), \Sigma_s) \tag{7}$$

where $\psi_{s,\mathbf{p}}$ is the (unknown) probability of speaker s to be in location $\mathbf{p}$ and $\mathcal{N}^c(\cdot;\cdot,\cdot)$ denotes the complex Gaussian probability[1] with covariance matrix Σ_s. The mean of each Gaussian, $\tilde{\boldsymbol{\phi}}^k(\mathbf{p}) = \left[\begin{array}{ccc} \tilde{\phi}_1^k(\mathbf{p}) & \cdots & \tilde{\phi}_M^k(\mathbf{p}) \end{array} \right]^T$, is set to the expected PRP from all candidate locations in the room $\mathbf{p} \in \mathcal{P}$ to the microphone pairs, satisfying:

$$\tilde{\phi}_m^k(\mathbf{p}) \triangleq \exp\left(-j\frac{2\pi k}{K} \frac{\cdot(||\mathbf{p}-\mathbf{p}_m^2|| - ||\mathbf{p}-\mathbf{p}_m^1||)}{c \cdot T_s} \right) \tag{8}$$

where $\mathbf{p}_m^1$,$\mathbf{p}_m^2$ are the locations of the microphones in pair m. The set $\mathbf{p} \in \mathcal{P}$ can be depicted as a dense grid of points in the enclosure. Other, more sophisticated point selection mechanisms can be applied as well.

It is reasonable to further assume that the PRP readings at the microphone pairs are independent, since they experience different sound reflection patterns. Hence, the covariance matrix can be modeled as $\Sigma_s = \text{diag}(\sigma_{1s}^2, \ldots, \sigma_{Ms}^2)$, and consequently:

$$\mathcal{N}^c(\boldsymbol{\phi}(t,k); \tilde{\boldsymbol{\phi}}^k(\mathbf{p}), \Sigma_s) = \prod_m \mathcal{N}^c(\phi_m(t,k); \tilde{\boldsymbol{\phi}}_m^k(\mathbf{p}), \sigma_{ms}^2). \tag{9}$$

Collecting all terms, the distribution of the PRP readings at microphone m is given by:

$$\mathcal{N}^c(\phi_m(t,k); \tilde{\phi}_m^k(\mathbf{p}), \sigma_{ms}^2) = \frac{1}{\pi\sigma_{ms}^2} \cdot \exp\left(-\frac{|\phi_m(t,k) - \tilde{\phi}_m^k(\mathbf{p})|^2}{\sigma_{ms}^2} \right). \tag{10}$$

In this contribution we further simplify the model and set $\sigma_{ms} = \sigma_s \forall m$. Alternative, more general models, can be used as well.

Finally, by augmenting all PRP readings for $t = 1, \ldots, T$ and $k = 0, \ldots, K-1$ in $\boldsymbol{\phi} = \text{vec}_{t,k}(\{\boldsymbol{\phi}(t,k)\})$, the probability density function (p.d.f.) of the entire observation set can be stated as:

$$f(\boldsymbol{\phi}) = \prod_{t,k} \sum_{s,\mathbf{p}} \psi_{s\mathbf{p}} \prod_m \mathcal{N}^c(\phi_m(t,k); \tilde{\phi}_m^k(\mathbf{p}), \sigma_s^2) \tag{11}$$

where we assumed that the PRP readings for all time segments and frequency bins are independent. This assumption follows directly from the disjoint activity of all speakers.

Let the unknown parameter set be

$$\boldsymbol{\theta} = \left[\boldsymbol{\psi}^T, \left(\boldsymbol{\sigma}^2\right)^T \right]^T \tag{12}$$

where, $\boldsymbol{\psi} = \text{vec}_{s\mathbf{p}}\left(\{\psi_{s\mathbf{p}}\}\right)$ and $\boldsymbol{\sigma}^2 = \text{vec}_s\left(\{\sigma_s^2\}\right)$. The MLE problem can readily be stated as:

$$\{\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\sigma}^2}\} = \underset{\{\boldsymbol{\psi},\boldsymbol{\sigma}^2\}}{\text{argmax}} \quad \log f(\boldsymbol{\phi}; \boldsymbol{\psi}, \boldsymbol{\sigma}^2). \tag{13}$$

III. The EM Algorithm for Localization

In this section the MLE of the (static) source locations is derived, utilizing the EM algorithm. We extend previous work [26] developed for TDOA estimation and set the foundations to the development of tracking procedures in Sections IV,V.

The EM algorithm requires the definition of three datasets and their probability model: the observations, the target parameters (already defined in Sec. II) and the hidden data. In our case, we define the hidden data to be the *association* of each time-frequency bin with a single source located in a particular location. The parameters to be estimated are the Gaussians' standard deviations and the probability of selecting one Gaussian. The mean of the Gaussians are pre-defined according to the distance between the prospective source positions on the grid and the known microphone locations, as explained in Sec. II.

A. The Hidden Data

The EM algorithm is a common procedure for finding the MLE in complex problems. Define the hidden data, $x(t,k,s,\mathbf{p})$, to be the indicator that the time-frequency bin (t,k) belongs to speaker s located at $\mathbf{p}$. The total number of indicators in the problem is $T \times K \times S \times |\mathcal{P}|$. The expectation of the indicator is readily equal to $\psi_{s\mathbf{p}}$, the probability of speaker s to be in location $\mathbf{p}$:

$$E\{x(t,k,s,\mathbf{p})\} = \psi_{s\mathbf{p}}. \tag{14}$$

[1] Note, that in [26] the argument of the PRP is approximated by a (real) Gaussian probability. We preferred to directly approximate the PRP by a (complex) Gaussian probability to avoid the nonlinearity involved in the argument operation. The absolute value of the PRP is confined to the unit circle. Note that distances between two PRP values are approximated by the chord of the unit circle connecting these values, rather than by the correct distance measured on the arc connecting them. For small distances both values are approximately identical. For any two PRP values, long arc-distances are mapped to long chord-distances. In our experiments both approximations yield the same performance.

Copyright (c) 2013 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2013.2292361

Let $\mathbf{x} = \text{vec}_{t,k,s,\mathbf{p}}\left(\{x(t,k,s,\mathbf{p})\}\right)$ be the set of all indicators. The probability density function of $\mathbf{x}$ is given by:

$$f(\mathbf{x};\boldsymbol{\theta}) = \prod_{t,k}\sum_{s,\mathbf{p}} \psi_{s\mathbf{p}} x(t,k,s,\mathbf{p}) \tag{15}$$

where we applied our model assumptions, that each observation can only be associated with one source, i.e, that at each time-frequency bin only a single indicator equals 1.

B. The Derivation of the EM Algorithm for Localization of Static Sources

Given the hidden data, the probability function of the observations is given by:

$$f(\boldsymbol{\phi}|\mathbf{x};\boldsymbol{\theta}) = \prod_{t,k}\sum_{s,\mathbf{p}} x(t,k,s,\mathbf{p}) \prod_m \mathcal{N}^c(\phi_m(t,k);\tilde{\phi}^k_m(\mathbf{p}),\sigma_s^2). \tag{16}$$

The p.d.f. of the complete data can be deduced from (15)-(16):

$$\begin{aligned} f(\mathbf{x},\boldsymbol{\phi};\boldsymbol{\theta}) &= f(\mathbf{x};\boldsymbol{\theta})\cdot f(\boldsymbol{\phi}|\mathbf{x};\boldsymbol{\theta}) \\ &= \prod_{t,k}\sum_{s,\mathbf{p}} \psi_{s\mathbf{p}} x(t,k,s,\mathbf{p}) \prod_m \mathcal{N}^c(\phi_m(t,k),\tilde{\phi}^k_m(\mathbf{p}),\sigma_s^2). \end{aligned} \tag{17}$$

The derivation of (17) from (15)-(16) simplifies due to the indicator properties. The EM algorithm for the problem at hand is now derived. The E-step is given by:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\ell-1)}) &\triangleq E\left\{\log\left(f(\boldsymbol{\phi},\mathbf{x};\boldsymbol{\theta})\right)|\boldsymbol{\phi};\boldsymbol{\theta}^{(\ell-1)}\right\} \\ &= \sum_{t,k,s,\mathbf{p}} E\left\{x(t,k,s,\mathbf{p})|\boldsymbol{\phi}(t,k);\boldsymbol{\theta}^{(\ell-1)}\right\} \\ &\times\left[\log\psi_{s\mathbf{p}} + \sum_m \log\mathcal{N}^c(\phi_m(t,k);\tilde{\phi}^k_m(\mathbf{p}),\sigma_s^2)\right]. \end{aligned} \tag{18}$$

For implementing the E-step it is sufficient to evaluate $\mu^{(\ell-1)}$ given by:

$$\begin{aligned} \mu^{(\ell-1)}(t,k,s,\mathbf{p}) &\triangleq E\left\{x(t,k,s,\mathbf{p})|\boldsymbol{\phi}(t,k);\boldsymbol{\theta}^{(\ell-1)}\right\} \qquad (19) \\ &= \frac{\psi^{(\ell-1)}_{s\mathbf{p}}\prod_m \mathcal{N}^c\left(\phi_m(t,k);\tilde{\phi}^k_m(\mathbf{p}),\sigma_s^{2,(\ell-1)}\right)}{\sum_{s\mathbf{p}}\psi^{(\ell-1)}_{s\mathbf{p}}\prod_m \mathcal{N}^c\left(\phi_m(t,k);\tilde{\phi}^k_m(\mathbf{p}),\sigma_s^{2,(\ell-1)}\right)}. \end{aligned}$$

Maximizing (18) with respect to the parameters $\boldsymbol{\theta}^{(\ell)}$ constitutes the M-step:

$$\psi^{(\ell)}_{s\mathbf{p}} = \frac{\sum_{t,k}\mu^{(\ell-1)}(t,k,s,\mathbf{p})}{T\cdot K} \tag{20a}$$

$$\sigma_s^{2,(\ell)} = \frac{\sum_{t,k,\mathbf{p},m}\mu^{(\ell-1)}(t,k,s,\mathbf{p})|\phi_m(t,k)-\tilde{\phi}^k_m(\mathbf{p})|^2}{M\cdot\sum_{t,k,\mathbf{p}}\mu^{(\ell-1)}(t,k,s,\mathbf{p})}. \tag{20b}$$

The resulting location estimation is determined by searching for the S (assumed to be known) most probable location of each speaker:

$$\mathbf{p}_s = \underset{\mathbf{p}}{\text{argmax}}\,\psi^{(L)}_{s,\mathbf{p}} \quad \forall s \tag{21}$$

where L is a pre-defined number of iterations. The summary of the location estimation can be found in Algorithm 1.

Algorithm 1: Static speaker localization with the EM algorithm.

```
Obtain $z^1_m(t,k)$ and $z^2_m(t,k)$.
Calculate $\phi_m(t,k)$ using (5).
set $\tilde{\phi}^k_m(\mathbf{p})$ using (8).
initialize $\psi^{(0)}_{s,\mathbf{p}}$ and $\sigma^{2,(0)}_{ms}$.
for $\ell = 1$ to $L$ do
    E-step
    Calculate $\mu^{(\ell)}(t,k,s,\mathbf{p})$ using (19).
    M-step
    Calculate $\psi^{(\ell)}_{s\mathbf{p}}$ using (20a) and $\sigma^{2,(\ell)}_s$ using (20b).
end
Find $\mathbf{p}_s$ $\forall s$ using (21).
```

The parameter $\psi^{(0)}_{s,\mathbf{p}}$ can be initialized by splitting the room area into regions, presumably containing a single speaker. Parameter $\sigma^{2,(0)}_s$ can be initialized by a uniform value for all the variances. The default value can be set to 1.

The computational complexity of the proposed localization scheme is as follows. The term $|\phi_m(t,k)-\tilde{\phi}^k_m(\mathbf{p})|^2$ can be calculated once before the iterations starts. This requires $\mathcal{O}(M\cdot|\mathcal{P}|\cdot T\cdot K)$ operations. Then, for L EM iterations we have: 1) E-step: $\mathcal{O}(S\cdot M\cdot|\mathcal{P}|\cdot T\cdot K\cdot L)$ operations; and 2) M-step: $\mathcal{O}(S\cdot|\mathcal{P}|\cdot T\cdot K\cdot L)$ operations for calculating the probability of each Gaussian and $\mathcal{O}(S\cdot M\cdot|\mathcal{P}|\cdot T\cdot K\cdot L)$ operations for the variance calculation. Note that, as shown in the experimental study in Sec. VI, in many practical scenarios the calculation of the variances can be avoided.

IV. The Derivation of TREM Algorithms for Source Tracking

The moving target scenario necessitates a tracking algorithm, that can utilize location estimates from previous frames. In this section we develop a recursive scheme, suitable for multiple speaker tracking, that uses the TREM algorithm [19]. The original TREM algorithm is summarized in the Appendix.

A. Preliminaries

The recursion for the problem at hand can be stated using (60):

$$\boldsymbol{\theta}^{(t)}_R = \boldsymbol{\theta}^{(t-1)}_R + \gamma_t\cdot\mathbf{I}^{-1}_{\mathbf{x}_t,\boldsymbol{\phi}_t;\boldsymbol{\theta}^{(t-1)}_R}\cdot\nabla_{\boldsymbol{\theta}}\log f(\boldsymbol{\phi}_t;\boldsymbol{\theta})|_{\boldsymbol{\theta}^{(t-1)}_R} \tag{22}$$

where $\boldsymbol{\theta}^{(t)}_R$ denotes recursive estimate of $\boldsymbol{\theta}$ at time step t.

$$\boldsymbol{\theta}^{(t)}_R = \left\{\left(\boldsymbol{\psi}^{(t)}_R\right)^T,\left(\boldsymbol{\sigma}^{2,(t)}_R\right)^T\right\}^T$$

with components $\boldsymbol{\psi}^{(t)}_R = \text{vec}_{s\mathbf{p}}\left(\left\{\psi^{(t)}_{s\mathbf{p},R}\right\}\right)$ and $\boldsymbol{\sigma}^{2,(t)}_R = \text{vec}_{ms}\left(\left\{\sigma^{2,(t)}_{s,R}\right\}\right)$. $\mathbf{x}_t = \text{vec}_{k,s,\mathbf{p}}\left(\{x(t,k,s,\mathbf{p})\}\right)$ is the entire set of hidden data related to time index t and $\boldsymbol{\phi}_t = \text{vec}_k\left(\{\boldsymbol{\phi}(t,k)\}\right)$ is the corresponding observed data. $\mathbf{I}_{\mathbf{x}_t,\boldsymbol{\phi}_t;\boldsymbol{\theta}^{(t-1)}_R}$ is the Fisher information matrix (FIM):

$$\mathbf{I}_{\mathbf{x}_t,\boldsymbol{\phi}_t;\boldsymbol{\theta}^{(t-1)}_R} \triangleq -E\left\{\nabla^2_{\boldsymbol{\theta}}\log(f(\mathbf{x}_t,\boldsymbol{\phi}_t;\boldsymbol{\theta}))|_{\boldsymbol{\theta}^{(t-1)}_R}\right\}. \tag{23}$$

Copyright (c) 2013 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2013.2292361

The probability density function of the complete data at time index t required for the evaluation of the FIM is given in (17). For the calculation of the gradient term in (22) we use (adapted from (11)):

$$f(\boldsymbol{\phi}_t;\boldsymbol{\theta}) = \prod_k \sum_{s,\mathbf{p}} \psi_{s\mathbf{p}} \prod_m \mathcal{N}^c(\phi_m(t,k), \tilde{\phi}_m^k(\mathbf{p}), \sigma_s^2). \quad (24)$$

In our case, $\psi_{s\mathbf{p}}$ should satisfy the constraints $\sum_{s,\mathbf{p}} \psi_{s\mathbf{p}} = 1$ and $0 < \psi_{s\mathbf{p}} < 1 \forall s, \mathbf{p}$, and hence Newton's method cannot be directly applied. In [28] a version of the generalized reduced gradient method (GRG) algorithm was suggested, in which one of the parameters is not obtained by the maximization, but rather calculated directly by applying the constraints. This method did not yield satisfactory results in our case. We therefore suggest using the constrained variant of Newton's method derived in the sequel. In this variant the Lagrangian is incorporated into Newton's iterations.

B. Newton's method with equality constraints

We will derive now a procedure capable of recursive constrained optimization. Our method will be based on a combination of Newton's method and the method of Lagrange multipliers. We wish to solve the following optimization problem:

$$\mathbf{x}_{\text{opt}} = \operatorname*{argmax}_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{a}^T\mathbf{x} = b. \quad (25)$$

Defining the Lagrangian $\mathcal{L}(\mathbf{x}) = f(\mathbf{x}) + \lambda(a^T\mathbf{x} - b)$ the equivalent optimization problem can be defined:

$$\mathbf{x}_{\text{opt}} = \operatorname*{argmax}_{[\mathbf{x}^T \lambda]^T} \mathcal{L}(\mathbf{x}). \quad (26)$$

Following [29], the application of one Newton step is given by:

$$\begin{bmatrix}\mathbf{x}^{(t)}\\ \lambda^{(t)}\end{bmatrix} = \begin{bmatrix}\mathbf{x}^{(t-1)}\\ \lambda^{(t-1)}\end{bmatrix} - \gamma_t \cdot \begin{bmatrix} H(\mathbf{x}^{(t-1)}) & \mathbf{a}\\ \mathbf{a}^T & 0\end{bmatrix}^{-1} \cdot \begin{bmatrix}\nabla f(\mathbf{x}^{(t-1)}) + \lambda^{(t-1)}\mathbf{a}\\ \mathbf{a}^T\mathbf{x}^{(t-1)} - b\end{bmatrix} \quad (27)$$

where $H(\mathbf{x}^{(t-1)})$ is the Hessian[2] of $\mathbf{x}$ calculated at the current value of the optimized vector $\mathbf{x}^{(t-1)}$. The matrix $\begin{bmatrix} H(\mathbf{x}^{(t-1)}) & \mathbf{a}\\ \mathbf{a}^T & 0\end{bmatrix}$ can be inverted by the block matrix inversion formula [30]:

$$\left[\begin{array}{c|c} A_{11} & A_{12}\\ \hline A_{21} & A_{22}\end{array}\right]^{-1} = \left[\begin{array}{c|c} C_1^{-1} & -A_{11}^{-1}A_{12}C_2^{-1}\\ \hline -C_2^{-1}A_{21}A_{11}^{-1} & C_2^{-1}\end{array}\right] \quad (28)$$

with the obvious definition of $A_{11}, A_{12}, A_{21}, A_{22}$ and where

$$\begin{aligned} C_1 &= A_{11} - A_{12}A_{22}^{-1}A_{21}\\ C_2 &= A_{22} - A_{21}A_{11}^{-1}A_{12}. \end{aligned} \quad (29)$$

Since, in our case $A_{22} = 0$, C_1 cannot be calculated. Fortunately, C_1^{-1} may be calculated directly by applying the Woodbury matrix inversion identity [30] given by:

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}. \quad (30)$$

When the Woodbury identity is applied to calculating C_1^{-1} it results in:

$$\begin{aligned} C_1^{-1} &= (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} \quad (31)\\ &= A_{11}^{-1} + A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1}. \end{aligned}$$

Collecting all terms results in:

$$C_1^{-1} = H^{-1}(\mathbf{x}^{(t-1)}) - \frac{H^{-1}(\mathbf{x}^{(t-1)})\mathbf{a}\mathbf{a}^T H^{-1}(\mathbf{x}^{(t-1)})}{\mathbf{a}^T H^{-1}(\mathbf{x}^{(t-1)})\mathbf{a}} \quad (32a)$$

$$C_2^{-1} = -\frac{1}{\mathbf{a}^T H^{-1}(\mathbf{x}^{(t-1)})\mathbf{a}}. \quad (32b)$$

Since we are only interested in recursive optimization algorithm for $\mathbf{x}^{(t)}$, the Lagrange multiplier λ should be cancelled out from the adaptation. Taking the first row of (27) and using the above matrix definitions results in:

$$\begin{aligned} \mathbf{x}^{(t)} = \mathbf{x}^{(t-1)} - \gamma_t \cdot \Big[& C_1^{-1}\left(\nabla f(\mathbf{x}^{(t-1)}) + \lambda\mathbf{a}\right)\\ & + A_{11}^{-1}A_{12}C_2^{-1}\left(\mathbf{a}^T\mathbf{x}^{(t-1)} - b\right)\Big]. \quad (33) \end{aligned}$$

Using the explicit matrices:

$$\begin{aligned} \mathbf{x}^{(t)} = \mathbf{x}^{(t-1)} - \gamma_t \cdot \Big[& H^{-1}(\mathbf{x}^{(t-1)})\\ & - \frac{H^{-1}(\mathbf{x}^{(t-1)})\mathbf{a}\mathbf{a}^T H^{-1}(\mathbf{x}^{(t-1)})}{\mathbf{a}^T H^{-1}(\mathbf{x}^{(t-1)})\mathbf{a}}\Big]\left(\nabla f(\mathbf{x}^{(t-1)}) + \lambda\mathbf{a}\right)\\ & - \gamma_t \cdot \frac{H^{-1}(\mathbf{x}^{(t-1)})\mathbf{a}}{\mathbf{a}^T H^{-1}(\mathbf{x}^{(t-1)})\mathbf{a}}\left(\mathbf{a}^T\mathbf{x}^{(t-1)} - b\right) \end{aligned}$$

and following straightforward algebraic manipulations finally yields:

$$\begin{aligned} \mathbf{x}^{(t)} = \mathbf{x}^{(t-1)} & - \gamma_t \cdot H^{-1}(\mathbf{x}^{(t-1)})\nabla f(\mathbf{x}^{(t-1)})\\ & - \gamma_t \cdot \frac{H^{-1}(\mathbf{x}^{(t-1)})\mathbf{a}}{\mathbf{a}^T H^{-1}(\mathbf{x}^{(t-1)})\mathbf{a}}\\ & \times \left(\mathbf{a}^T\mathbf{x}^{(t-1)} - \mathbf{a}^T H^{-1}(\mathbf{x}^{(t-1)})\nabla f(\mathbf{x}^{(t-1)}) - b\right). \quad (34) \end{aligned}$$

It is interesting to examine the special case of setting $\gamma_t = 1$. Defining the *unconstrained* Newton step:

$$\mathbf{x}_N^{(t)} \triangleq \mathbf{x}^{(t-1)} - H^{-1}(\mathbf{x}^{(t-1)})\nabla f(\mathbf{x}^{(t-1)}),$$

results in:

$$\mathbf{x}^{(t)} = \mathbf{x}_N^{(t)} - \frac{H^{-1}(\mathbf{x}^{(t-1)})\mathbf{a}}{\mathbf{a}^T H^{-1}(\mathbf{x}^{(t-1)})\mathbf{a}}\left(\mathbf{a}^T\mathbf{x}_N^{(t)} - b\right). \quad (35)$$

This recursion can be interpreted as an unconstrained Newton step with a *correcting term*, responsible for satisfying the constraint.

C. Constraint TREM Algorithm

Linear constraints[3] $a^T\boldsymbol{\theta} = b$ can be imposed on the maximization of the TREM algorithm by utilizing the result[4] in Section IV-B.

[2] Note that the Hessian is calculated without considering the constraint.

[3] Note that there is no need to add the inequality constraints, regarding positive probabilities, as they are inherently satisfied.

[4] Since the FIM is an approximation of the Hessian, it is calculated without considering the constraint as well. This approximated Hessian is not claimed to be the actual FIM that considers the constraint.

Copyright (c) 2013 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2013.2292361

Following (34), the recursive estimation of the parameters $\boldsymbol{\theta}$ with the linear constraint set $\mathbf{a}^T\boldsymbol{\theta} = b$ is given by:

$$\begin{aligned}\boldsymbol{\theta}_R^{(t)} &= \boldsymbol{\theta}_R^{(t-1)} + \\ &\gamma_t \mathbf{I}^{-1}_{\mathbf{x}_t,\boldsymbol{\phi}_t;\boldsymbol{\theta}_R^{(t-1)}} \nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\phi}_t;\boldsymbol{\theta})|_{\boldsymbol{\theta}_R^{(t-1)}} - \gamma_t \frac{\mathbf{I}^{-1}_{\mathbf{x}_t,\boldsymbol{\phi}_t;\boldsymbol{\theta}_R^{(t-1)}}\mathbf{a}}{\mathbf{a}^T \mathbf{I}^{-1}_{\mathbf{x}_t,\boldsymbol{\phi}_t;\boldsymbol{\theta}_R^{(t-1)}}\mathbf{a}} \\ &\times (\mathbf{a}^T\boldsymbol{\theta}_R^{(t-1)} + \mathbf{a}^T\mathbf{I}^{-1}_{\mathbf{x}_t,\boldsymbol{\phi}_t;\boldsymbol{\theta}_R^{(t-1)}} \nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\phi}_t;\boldsymbol{\theta})|_{\boldsymbol{\theta}_R^{(t-1)}} - b).\end{aligned} \tag{36}$$

where in our case $\mathbf{a} = [\mathbf{1}^T_{S\cdot|\mathcal{P}|} \quad \mathbf{0}^T_S]^T$ and $b = 1$.

The terms required for the calculation of the FIM are:

$$-E\left\{\frac{\partial^2}{\partial\psi^2_{s\mathbf{p}}} \log f(\mathbf{x}_t,\boldsymbol{\phi}_t;\boldsymbol{\theta})|_{\boldsymbol{\theta}_R^{(t-1)}}\right\} = \frac{K}{\psi^{(t-1)}_{s\mathbf{p},R}} \tag{37a}$$

$$-E\left\{\frac{\partial^2}{\partial(\sigma^2_s)^2} \log f(\mathbf{x}_t,\boldsymbol{\phi}_t;\boldsymbol{\theta})|_{\boldsymbol{\theta}_R^{(t-1)}}\right\} = \frac{K\cdot M}{\sigma^{4,(t-1)}_{s,R}} \sum_{\mathbf{p}} \psi^{(t-1)}_{s\mathbf{p},R}. \tag{37b}$$

All the cross-derivatives comprising the FIM are equal to zero; therefore, the FIM turns out to be diagonal. The FIM inversion can be performed by inverting each element of the diagonal. Therefore, the recursive equation system can be decoupled into multiple scalar equations for each $\psi_{s\mathbf{p}} \forall s,\mathbf{p}$ and the other for $\sigma^2_s \; \forall s$.

The terms required for the calculation of the gradient are:

$$\frac{\partial}{\partial\psi_{s\mathbf{p}}} \log f(\boldsymbol{\phi}_t;\boldsymbol{\theta})|_{\boldsymbol{\theta}_R^{(t-1)}} = \sum_k \frac{\mu(t,k,s,\mathbf{p})}{\psi^{(t)}_{s\mathbf{p},R}} \tag{38a}$$

$$\begin{aligned}\frac{\partial}{\partial(\sigma^2_s)} \log f(\boldsymbol{\phi}_t;\boldsymbol{\theta})|_{\boldsymbol{\theta}_R^{(t-1)}} =& \frac{1}{\sigma^{4,(t-1)}_{s,R}} \sum_{k,\mathbf{p}} \mu(t,k,s,\mathbf{p}) \\ &\times \Big[\sum_m |\phi_m(t,k) - \tilde{\phi}^k_m(\mathbf{p})|^2 - M\cdot\sigma^{2,(t-1)}_{s,R}\Big]\end{aligned} \tag{38b}$$

where $\mu(t,k,s,\mathbf{p})$ is defined as:

$$\mu(t,k,s,\mathbf{p}) \triangleq \frac{\psi^{(t-1)}_{s\mathbf{p},R} \prod_m \mathcal{N}^c\left(\phi_m(t,k), \tilde{\phi}^k_m(\mathbf{p}), \sigma^{2,(t-1)}_{s,R}\right)}{\sum_{s\mathbf{p}} \psi^{(t-1)}_{s\mathbf{p},R} \prod_m \mathcal{N}^c\left(\phi_m(t,k), \tilde{\phi}^k_m(\mathbf{p}), \sigma^{2,(t-1)}_{s,R}\right)}. \tag{39}$$

The calculation of TREM involves the multiplication of the FIM and the gradient. Using (37a) and (38a) and due to the similarity to (20a) we have :

$$\begin{aligned}&\left(-E\left\{\frac{d^2}{d\psi^2_{s\mathbf{p}}} \log f(\mathbf{x}_t,\boldsymbol{\phi}_t;\boldsymbol{\theta})|_{\boldsymbol{\theta}_R^{(t-1)}}\right\}\right)^{-1} \\ &\times \frac{\partial}{\partial\psi_{s\mathbf{p}}} \log f(\boldsymbol{\phi}_t;\boldsymbol{\theta})|_{\boldsymbol{\theta}_R^{(t-1)}} \\ &= \sum_k \frac{\mu(t,k,s,\mathbf{p})}{K} \triangleq \psi^{(t)}_{s\mathbf{p}}.\end{aligned} \tag{40}$$

Using (37a) for the FIM and (40) in (36), a recursive estimation for $\psi_{s\mathbf{p}}$ results in:

$$\begin{aligned}\boldsymbol{\psi}_R^{(t)} = \boldsymbol{\psi}_R^{(t-1)} + \gamma_t\boldsymbol{\psi}^{(t)} - \gamma_t \frac{\boldsymbol{\psi}_R^{(t-1)}}{\mathbf{1}^T_{S\cdot|\mathcal{P}|}\boldsymbol{\psi}_R^{(t-1)}} \\ \times (\mathbf{1}^T_{S\cdot|\mathcal{P}|}\boldsymbol{\psi}_R^{(t-1)} + \mathbf{1}^T_{S\cdot|\mathcal{P}|}\boldsymbol{\psi}^{(t)} - 1)\end{aligned} \tag{41}$$

Algorithm 2: Speaker tracking using TREM.

```
set $\tilde{\phi}^k_m(\mathbf{p})$ using (8)
initialize $\psi^{(0)}_{s\mathbf{p},R}$ and $\sigma^{2,(0)}_{s,R}$
for $t = 1$ to $T$ do
    Obtain $z^1_m(t,k), z^2_m(t,k) \; \forall k, m$
    Calculate $\phi_m(t,k)$ using (5)
    Calculate $\psi^{(t)}_{s\mathbf{p},R}$ using (42)
    Calculate $\sigma^{2,(t)}_{s,R}$ using (43)
end
```

Algorithm 3: The CREM algorithm.

```
initialize $\boldsymbol{\theta}^{(0)}_R$
for $t = 1$ to $T$ do
    E-step
        $\bar{\boldsymbol{\eta}}(\boldsymbol{\phi}_t,\mathbf{x}_t) \triangleq E\left\{\boldsymbol{\eta}(\boldsymbol{\phi}_t,\mathbf{x}_t)|\boldsymbol{\phi}_t;\boldsymbol{\theta}^{(t-1)}_R\right\}$
        $\boldsymbol{\eta}^R(\boldsymbol{\phi}_t,\mathbf{x}_t) = \boldsymbol{\eta}^R(\boldsymbol{\phi}_{t-1},\mathbf{x}_{t-1}) \quad +$
        $\gamma_t\left(\bar{\boldsymbol{\eta}}(\boldsymbol{\phi}_t,\mathbf{x}_t) - \boldsymbol{\eta}^R(\boldsymbol{\phi}_{t-1},\mathbf{x}_{t-1})\right)$
    M-step
        $\boldsymbol{\theta}^{(t)}_R = \mathrm{argmax}_{\boldsymbol{\theta}}\langle\boldsymbol{\eta}^R(\boldsymbol{\phi}_t,\mathbf{x}_t),\xi(\boldsymbol{\theta})\rangle$
end
```

where $\boldsymbol{\psi}^{(t)} = \mathrm{vec}_{s\mathbf{p}}\left(\left\{\psi^{(t)}_{s\mathbf{p}}\right\}\right)$. Noting that $\mathbf{1}^T_{S\cdot|\mathcal{P}|}\boldsymbol{\psi}^{(t)} = 1$, the recursion can be further simplified to:

$$\boldsymbol{\psi}_R^{(t)} = \boldsymbol{\psi}_R^{(t-1)} + \gamma_t(\boldsymbol{\psi}^{(t)} - \boldsymbol{\psi}_R^{(t-1)}) \tag{42}$$

To recursively estimate σ^2_s, the *unconstrained* TREM procedure (22) can be applied. Using (37b) and (38b) in (36) we get:

$$\begin{aligned}\sigma^{2,(t)}_{s,R} = \sigma^{2,(t-1)}_{s,R} + \gamma_t \frac{1}{K\cdot\sum_{\mathbf{p}}\psi^{(t-1)}_{s\mathbf{p},R}} \sum_{k,\mathbf{p}} \mu(t,k,s,\mathbf{p}) \\ \times \left[\frac{1}{M}\sum_m |\phi_m(t,k) - \tilde{\phi}^k_m(\mathbf{p})|^2 - \cdot\sigma^{2,(t-1)}_{ms,R}\right].\end{aligned} \tag{43}$$

The TREM algorithm is summarized in Algorithm 2.

V. Tracking using CREM Algorithm

In this section we develop an alternative tracking procedure based on the CREM algorithm [20]. The CREM version is based on the time-smoothing of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\ell)})$ obtained through the EM iterations.

In our case, the complete data probability function is an exponential p.d.f., namely:

$$\log f(\boldsymbol{\phi}_t,\mathbf{x}_t;\boldsymbol{\theta}) = \langle\boldsymbol{\eta}(\boldsymbol{\phi}_t,\mathbf{x}_t),\boldsymbol{\xi}(\boldsymbol{\theta})\rangle \tag{44}$$

where $\boldsymbol{\eta}(\cdot)$ and $\boldsymbol{\xi}(\cdot)$ are vectors of functions and $\langle\cdot,\cdot\rangle$ denotes the scalar product. The CREM algorithm for any exponential p.d.f. is summarized in Algorithm 3.

We now turn to the derivation of the speaker tracking procedure based on the CREM. From (17) the log-likelihood

Copyright (c) 2013 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2013.2292361

of the complete data is given by:

$$\log f(\boldsymbol{\phi}_t, \mathbf{x}_t; \boldsymbol{\theta}) = \sum_{k,s,\mathbf{p}} x(t,k,s,\mathbf{p}) \times \left[\log(\psi_{s\mathbf{p}}) - \sum_m \log(\pi\sigma_s^2)\right] - \sum_{k,s,\mathbf{p},m} x(t,k,s,\mathbf{p})|\phi_m(t,k) - \tilde{\phi}^k(\mathbf{p})|^2\sigma_s^{-2}. \quad (45)$$

$\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ can now be identified in (45):

$$\boldsymbol{\eta}(\boldsymbol{\phi}_t, \mathbf{x}_t) = \begin{bmatrix} \text{vec}_{k,s,\mathbf{p}}\left(\{x(t,k,s,\mathbf{p})\}\right) \\ \text{vec}_{k,s,\mathbf{p},m}\left(\left\{x(t,k,s,\mathbf{p})|\phi_m(t,k) - \tilde{\phi}_m^k(\mathbf{p})|^2\right\}\right) \end{bmatrix} \quad (46a)$$

$$\boldsymbol{\xi}(\boldsymbol{\theta}) = \begin{bmatrix} \text{vec}_{k,s,\mathbf{p}}\left(\left\{\log(\psi_{s\mathbf{p}}) - \sum_m \log(\pi\sigma_s^2)\right\}\right) \\ \text{vec}_{k,s,\mathbf{p},m}\left(\left\{-\sigma_s^{-2}\right\}\right) \end{bmatrix}. \quad (46b)$$

The term $\bar{\boldsymbol{\eta}}(\boldsymbol{\phi}_t, \mathbf{x}_t)$ required for the E-step can be simplified by only calculating the term:

$$\mu(t,k,s,\mathbf{p}) = E\left\{x(t,k,s,\mathbf{p})|\boldsymbol{\phi}_t; \boldsymbol{\theta}_R^{(t-1)}\right\} \quad (47)$$

which is readily given by (39). Using the above definition, $\bar{\boldsymbol{\eta}}(\boldsymbol{\phi}_t, \mathbf{x}_t)$ can be written as:

$$\bar{\boldsymbol{\eta}}(\boldsymbol{\phi}_t, \mathbf{x}_t) = \begin{bmatrix} \text{vec}_{k,s,\mathbf{p}}\left(\{\mu(t,k,s,\mathbf{p})\}\right) \\ \text{vec}_{k,s,\mathbf{p},m}\left(\left\{\mu(t,k,s,\mathbf{p})|\phi_m(t,k) - \tilde{\phi}_m^k(\mathbf{p})|^2\right\}\right) \end{bmatrix} \quad (48)$$

and the recursive term $\boldsymbol{\eta}^R(\boldsymbol{\phi}_t, \mathbf{x}_t)$ can be calculated by defining:

$$\boldsymbol{\eta}^R(\boldsymbol{\phi}_t, \mathbf{x}_t) \triangleq \begin{bmatrix} \text{vec}_{k,s,\mathbf{p}}\left(\{\eta_1^R(t,k,s,\mathbf{p})\}\right) \\ \text{vec}_{k,s,\mathbf{p},m}\left(\{\eta_2^R(t,k,s,\mathbf{p},m)\}\right) \end{bmatrix} \quad (49)$$

where

$$\eta_1^R(t,k,s,\mathbf{p}) = \eta_1^R(t-1,k,s,\mathbf{p}) + \gamma_t\left[\mu(t,k,s,\mathbf{p}) - \eta_1^R(t-1,k,s,\mathbf{p})\right] \quad (50a)$$

$$\eta_2^R(t,k,s,\mathbf{p},m) = \eta_2^R(t-1,k,s,\mathbf{p},m) + \gamma_t\Big[\mu(t,k,s,\mathbf{p})|\phi_m(t,k) - \tilde{\phi}_m^k(\mathbf{p})|^2 - \eta_2^R(t-1,k,s,\mathbf{p},m)\Big]. \quad (50b)$$

The maximization step in the CREM is similar to the maximization step in the batch EM. Hence, the M-step $\boldsymbol{\theta}_R^{(t)} = \text{argmax}_{\boldsymbol{\theta}}\langle\boldsymbol{\eta}^R(\boldsymbol{\phi}_t, \mathbf{x}_t), \xi(\boldsymbol{\theta})\rangle$ yields:

$$\psi_{s\mathbf{p},R}^{(t)} = \frac{\sum_k \eta_1^R(t,k,s,\mathbf{p})}{K} \quad (51a)$$

$$\sigma_{s,R}^{2,(t)} = \frac{\sum_{k,\mathbf{p},m}\eta_2^R(t,k,s,\mathbf{p},m)}{M\cdot\sum_{k,\mathbf{p}}\eta_1^R(t,k,s,\mathbf{p})}. \quad (51b)$$

The M-step and the E-step can be merged into one recursion:

$$\boldsymbol{\psi}_R^{(t)} = \boldsymbol{\psi}_R^{(t-1)} + \gamma_t(\boldsymbol{\psi}^{(t)} - \boldsymbol{\psi}_R^{(t-1)}) \quad (52)$$

Algorithm 4: Speaker tracking using CREM.

```
set $\tilde{\phi}_m^k(\mathbf{p})$ using (8)
initialize $\psi_{s\mathbf{p},R}^{(0)}$ and $\sigma_{s,R}^{2,(0)}$
for $t = 1$ to $T$ do
    Obtain $z_m^1(t,k), z_m^2(t,k)\ \forall k, m$
    Calculate $\phi_m(t,k)$ using (5)
    Calculate $\psi_{s\mathbf{p},R}^{(t)}$ using (52)
    Calculate $\sigma_{s,R}^{2,(t)}$ using (53)
end
```

where $\boldsymbol{\psi}^{(t)} = \text{vec}_{s\mathbf{p}}\left(\left\{\psi_{s\mathbf{p}}^{(t)}\right\}\right)$, the latter was defined in (40), and

$$\sigma_{s,R}^{2,(t)} = \sigma_{s,R}^{2,(t-1)}\frac{\sum_{\mathbf{p}}\psi_{s\mathbf{p},R}^{(t-1)}}{\sum_{\mathbf{p}}\psi_{s\mathbf{p},R}^{(t)}} + \gamma_t\left(\frac{\sum_{k,\mathbf{p},m}\mu(t,k,s,\mathbf{p})|\phi_m(t,k) - \tilde{\phi}_m^k(\mathbf{p})|^2}{K\cdot M\cdot\sum_{\mathbf{p}}\psi_{s\mathbf{p},R}^{(t)}} - \sigma_{s,R}^{2,(t-1)}\frac{\sum_{\mathbf{p}}\psi_{s\mathbf{p},R}^{(t-1)}}{\sum_{\mathbf{p}}\psi_{s\mathbf{p},R}^{(t)}}\right). \quad (53)$$

Interestingly, the recursion for $\boldsymbol{\psi}_R^{(t)}$ is identical for the CREM and TREM variants. However, the recursion for $\sigma_{s,R}^{2,(t)}$ differs. It is not clear which of the versions for estimating $\sigma_{s,R}^{2,(t)}$ is advantageous. Whereas the CREM exhibits a faster convergence rate, the TREM may produce smoother contour estimates for slowly moving speakers.

Both tracking procedures (based on either the TREM or CREM algorithms) have the same computational complexity. For the recursive algorithms the E-step and M-step are combined and no iterations are required. Hence, for each time step t, $\mathcal{O}(S\cdot M\cdot|\mathcal{P}|\cdot K)$ operations are required.

The CREM tracking procedure for our case is summarized in Algorithm 4.

VI. Experimental Study

This section is dedicated to an experimental study of the proposed localization and tracking algorithms. We start with a simulated experimental study of the localization algorithm derived in Section III-B. We then turn to an extensive simulated experimental study of the tracking algorithms developed in Sections IV and V.

A. Localization of Static Speakers

1) Setup: A two speakers scenario was tested. The sentences uttered by different speakers were drawn from the TIMIT [31] database. The signals were analyzed by a STFT with $K = 1024$ frequency bins. Only bins corresponding to a range of $500 - 1500$Hz were considered. We simulated a room with dimensions $6\times 6\times 6.1$ m and with a reverberation time set to either $T_{60} = 0.4$ s or $T_{60} = 0.7$ s. Twelve pairs of microphones with an inter-microphone distance of $d = 0.2$ m were placed around the room 1 m from the walls.

Copyright (c) 2013 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2013.2292361

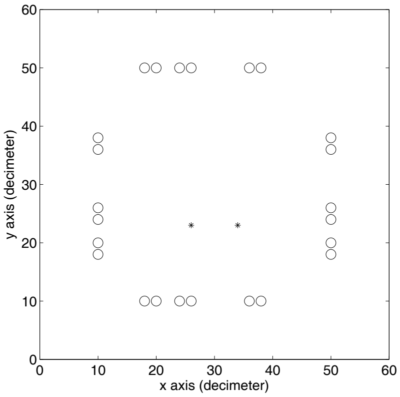


Fig. 2. Room setup with twelve pairs of microphones. 'o' denotes microphones and '*' denotes speakers.

The microphones constellation and the speakers locations are illustrated in Fig. 2.

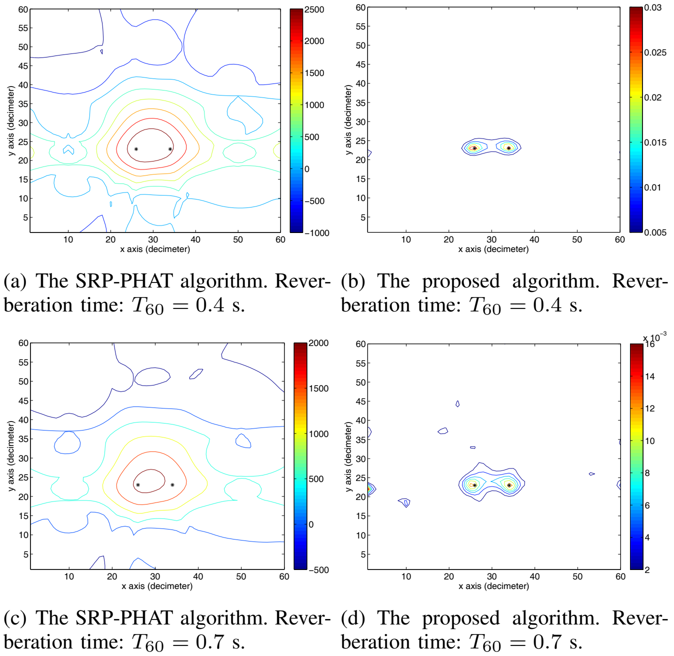
Since, in our simulations, the speakers and the microphone are placed at the same height (1 m from the floor), we have used 2-D Gaussians in our MoG model in accordance with the two-dimensional search grid. $\mathcal{P}$, the set of possible source locations, was chosen to be a grid of points, with a resolution of 10 cm. Overall, $\mathcal{P}$ consisted of a total of 60×60 possible locations. The speech signals received by the microphones were contaminated by temporally and spatially white noise with signal to noise ratio (SNR) of 30dB. We evaluated the algorithm described in Section III and compared its performance with the performance of the SRP-PHAT algorithm [5].

$\boldsymbol{\theta}^{(0)}$ was initialized with a Uniform distribution as follows. $\psi_{1\mathbf{p}}^{(0)} = \frac{1}{60\times 60}$ for all points on the left-hand side of the room and $\psi_{2\mathbf{p}}^{(0)} = \frac{1}{60\times 60}$ for all points on the right-hand side of the room. Both variances were identically initialized to a value of 1. The simulation setup is summarized in Table I.

2) Results: The two location of the speakers estimates are depicted in Fig. 3 for the two values of T_{60}. The true positions of the speakers are marked with black asterisks. The estimations of $\psi_{1\mathbf{p}}^{(10)}$ and $\psi_{2\mathbf{p}}^{(10)}$ are merged into one figure. To demonstrate the resolution of the proposed algorithm the speakers were positioned in a rather close proximity. It can be seen from the figures that the proposed algorithm is capable of correctly estimating the locations of the speakers, with sharp and distinguishable peaks at the correct source locations. The SRP-PHAT algorithm, however, exhibits a wide summit surrounding both correct locations, making them indistinguishable.

B. Speakers Tracking

1) Setup: In this section we demonstrate the ability of the algorithms derived in Sec. IV and Sec. V to track multiple moving speakers. Two microphone signals 4 s long were segmented into frames of 64 msec. The reverberation time was set to $T_{60} = 0.7$ s. The variance of all Gaussians was set to a fixed value of 1, hence only the recursive estimation procedure for $\psi_{s\mathbf{p}}^{(t)}$ was activated. The recursive estimation procedure for $\psi_{s\mathbf{p}}^{(t)}$ is identical for both the TREM and CREM algorithms, as derived from (52) and (42).



(a) The SRP-PHAT algorithm. Reverberation time: $T_{60} = 0.4$ s.

(b) The proposed algorithm. Reverberation time: $T_{60} = 0.4$ s.

(c) The SRP-PHAT algorithm. Reverberation time: $T_{60} = 0.7$ s.

(d) The proposed algorithm. Reverberation time: $T_{60} = 0.7$ s.

Fig. 3. Two speaker location estimate. SRP-PHAT versus the proposed localization algorithm.

We have conducted two sets of experiments. In the first experiment set each of the two sources was synthesized to move along a straight trajectory, both with a velocity of 1.25 m/s. The starting points for the two speakers were set to $[1.5, 0.5, 1.0]$ m and $[4.5, 0.5, 1.0]$ m, respectively. $\psi_{1\mathbf{p}}^{(0)}$ and $\psi_{2\mathbf{p}}^{(0)}$ were initialized in the same manner as in the localization experiment. The step-size γ was set to three different values: $0.1, 0.5, 1.0$. Note, that for $\gamma = 1.0$ the algorithm degenerates to the instantaneous estimate (52), hence to fast but noisy tracking. If γ is set to a very low number the algorithm exhibits long memory and tends to yield a fixed estimate close to the initialization point.

In the second set of experiments, each source was synthesized to move along a half circle trajectory, both with a velocity of 0.8 m/s. The starting points for the two speakers were set to $[2.8, 4.5, 1]$ m and $[3.2, 1.5, 1]$ m, respectively. The speakers were set to move counterclockwise. This choice of trajectory kept the speakers well-separated, avoiding the algorithm from collapsing to a single trajectory.

2) Results: The tracking results for the straight line scenario are shown in Fig. 4. In the figures the true trajectories are depicted by thin blue lines. In addition, contours corresponding to 99% of the highest peak of the p.d.f are also depicted. We show the entire set of estimated locations concurrently. It is evident that the algorithm tracks the true trajectory of each speaker with slight bias towards the competing speaker. The best results were obtained by setting $\gamma = 0.1$. The performance of the algorithm improves with lower reverberation levels.

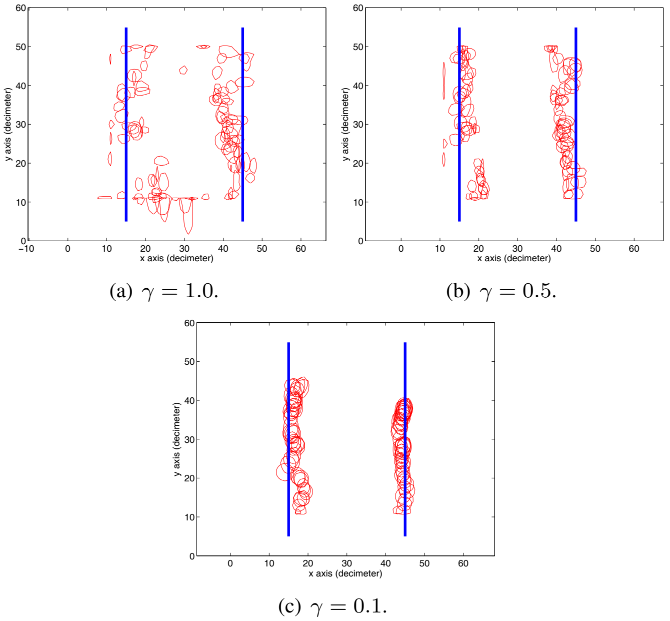
Similar trends are evident for the more complex scenario, for which the speakers were synthesized to move counterclock-

Copyright (c) 2013 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2013.2292361

TABLE I
THE SIMULATION SETUP.

f_s	T	Bins	k	T_{60}
16000Hz	10	1024	$32:96$	$0.4, 0.7$ s
Room [m]	$\|\|\mathbf{p}_m^1 - \mathbf{p}_m^2\|\|$ $[m]$	Distance from walls [m]	$\mathbf{p}_1$ [m]	$\mathbf{p}_2$ [m]
$6 \times 6 \times 6.1$	0.2	1	$[2.6, 2.3, 1]$	$[3.4, 2.3, 1]$

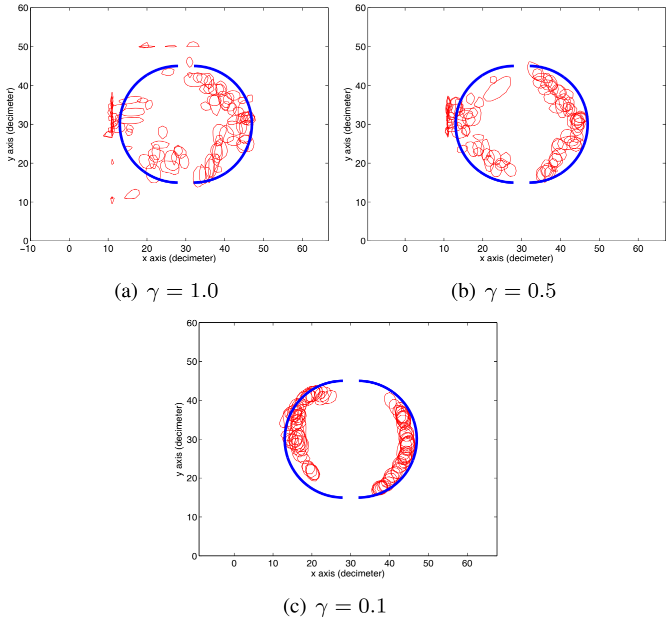


(a) $\gamma = 1.0$.

(b) $\gamma = 0.5$.

(c) $\gamma = 0.1$.

Fig. 4. Two speaker tracking by the proposed algorithm with $T_{60} = 0.7$. True trajectories of the speakers are depicted by blue lines. Red closed contours present the upper 1% values of $\psi_{s\mathbf{p}}^{(t)}$.



(a) $\gamma = 1.0$

(b) $\gamma = 0.5$

(c) $\gamma = 0.1$

Fig. 5. Two speaker tracking by the proposed algorithm with $T_{60} = 0.7$ s. True trajectories of the speakers are depicted by blue lines. Red closed contours present the upper 1% values of $\psi_{s\mathbf{p}}^{(t)}$.

wise along a half circle trajectory. Again the parameter value $\gamma = 0.1$ yields the best results. It is evident that the algorithms tracks the two speakers and the estimate is slightly biased towards closer trajectories.

3) Practical Consideration: Additional adjustments to improve the tracking ability of time-varying parameters were incorporated in the algorithm framework. We observed a degeneracy phenomenon. The values of the parameters $\psi_{s\mathbf{p}}^{(t)}$ might aggregate at a specific location estimate, disabling the algorithm from further tracking of the speakers. To mitigate this problem we perturbed the current probability values at the end of each time instant in preparation for the next stage. This method is conceptually similar to the *propagation stage* of the particle filter (PF).

The algorithm searches for the most probable track of both speakers. In some scenarios, especially if the speakers are close to each other or if one the speakers is stronger, we have encountered a practical problem in which the algorithm output collapses to a single speaker track. To circumvent this phenomenon we have taken the following measure. In each time instant, the estimated probabilities $\psi_{s\mathbf{p}}^{(t)}$ for speaker s are set to zero for all room locations that are closer to the other tracked speakers. As a consequence, the search area of the speaker location at the subsequent time instant is confined to relatively close locations hence avoiding tracks degeneration.

VII. Conclusion

In this work we considered the problem of localizing and tracking multiple speakers in reverberant environment. We first extend the MESSL algorithm [26] to deal with localization rather than TDOA estimation.

The main contribution of this work is the derivation of two tracking algorithms. The algorithms are based on two REM variants, namely the TREM and the CREM. In this work we also extended the TREM to deal with a constrained maximization, encountered in the MoG formulation of the problem at hand. The recursive, constrained MLE is obtained by incorporating the Lagrange multiplier method into the Newton recursion. In fact, the two algorithms differ only in the estimation of the variances of the MoG and become identical in the estimation procedure for the probabilities of the Gaussians. An experimental study demonstrates the capability of the derived algorithms to track multiple speakers in highly reverberant environment.

Appendix

In this appendix a recursive version of the EM algorithm, attributed to Titterington [19] is presented.

Copyright (c) 2013 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2013.2292361

Lange [32] suggests utilizing a gradient search method to perform the maximization in the M-step. Let $f(\boldsymbol{x})$ be a function of a vector of variables $\boldsymbol{x}$. Maximization of $f(\boldsymbol{x})$ can be obtained by Newton's method:

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - H^{-1}(\boldsymbol{x}_{t-1}) \cdot \nabla_{\boldsymbol{x}} f(\boldsymbol{x})|_{\boldsymbol{x}_{t-1}} \tag{54}$$

where $H(\boldsymbol{x})$ is the Hessian of the function $f(\boldsymbol{x})$, i.e. $H(\boldsymbol{x}_{t-1}) = \nabla^2_{\boldsymbol{x}} f(\boldsymbol{x})|_{\boldsymbol{x}_{t-1}}$. This method can be adopted to accelerate the maximization of the M-step in the EM algorithm:

$$\begin{aligned}\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} - &\left[\nabla^2_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)})|_{\boldsymbol{\theta}^{(t-1)}}\right]^{-1} \\ &\times \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)})|_{\boldsymbol{\theta}^{(t-1)}}.\end{aligned} \tag{55}$$

Titterington [19] proposes a fully recursive scheme by considering only the current hidden and measured data at each iteration step, namely $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$ is redefined as:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)}) \triangleq E\left\{\log\left(y(t), \boldsymbol{x}(t); \boldsymbol{\theta}\right) |\boldsymbol{y}(t); \boldsymbol{\theta}^{(t-1)}\right\} \tag{56}$$

where $\boldsymbol{x}(t)$ is the entire set of hidden data related to time index t and $\boldsymbol{y}(t)$ is the corresponding observed data. Explicitly, the Newton's iterations can be written as:

$$\begin{aligned}&\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} \\ &\quad - \left[E\left\{\nabla^2_{\boldsymbol{\theta}} \log f(\boldsymbol{y}(t), \boldsymbol{x}(t); \boldsymbol{\theta})|_{\boldsymbol{\theta}^{(t-1)}} |\boldsymbol{y}(t); \boldsymbol{\theta}^{(t-1)}\right\}\right]^{-1} \\ &\quad \times E\left\{\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{y}(t), \boldsymbol{x}(t); \boldsymbol{\theta})|_{\boldsymbol{\theta}^{(t-1)}} |\boldsymbol{y}(t); \boldsymbol{\theta}^{(t-1)}\right\}\end{aligned} \tag{57}$$

where we exchanged the derivative and expectation operations. This expression may be simplified, by using the Fisher Identity [17]:

$$E\left\{\nabla_{\theta} \log f(\boldsymbol{y}(t), \boldsymbol{x}(t); \boldsymbol{\theta})|\boldsymbol{y}(t); \boldsymbol{\theta}\right\} = \nabla_{\theta} \log f(\boldsymbol{y}(t); \boldsymbol{\theta}). \tag{58}$$

The inversion of the Hessian matrix is a cumbersome task. Moreover, it is not guaranteed that the expected Hessian is an invertible matrix. To mitigate these problems Titterington further suggests to approximate the expected Hessian by the FIM of both the observed and the hidden data. Note, that by this approximation the conditional expectation in the original expression is replaced by the ordinary expectation operation. The FIM is a positive definite matrix and is therefore invertible. Hence, by replacing

$$\begin{aligned}&-E\left\{\nabla^2_{\boldsymbol{\theta}} \log f(\boldsymbol{y}(t), \boldsymbol{x}(t); \boldsymbol{\theta})|_{\boldsymbol{\theta}^{(t-1)}} |\boldsymbol{y}(t); \boldsymbol{\theta}^{(t-1)}\right\} \\ &\quad \to -E\{\nabla^2_{\theta} \log f(\boldsymbol{y}(t), \boldsymbol{x}(t); \boldsymbol{\theta})|_{\boldsymbol{\theta}^{(t-1)}}; \boldsymbol{\theta}^{(t-1)}\} \\ &\quad = \mathbf{I}_{\boldsymbol{y}(t), \boldsymbol{x}(t); \boldsymbol{\theta}^{(t-1)}}\end{aligned} \tag{59}$$

and using (58) we finally get:

$$\boldsymbol{\theta}_R^{(t)} = \boldsymbol{\theta}_R^{(t-1)} + \gamma_t \mathbf{I}^{-1}_{\boldsymbol{y}(t), \boldsymbol{x}(t); \boldsymbol{\theta}^{(t-1)}} \cdot \nabla_{\theta} \log f(\boldsymbol{y}(t); \boldsymbol{\theta})|_{\boldsymbol{\theta}_R^{(t-1)}} \tag{60}$$

where $\boldsymbol{\theta}_R^{(t)}$ is the recursive parameter vector. The variable γ_t introduces further convergence control to the algorithm. A convergence proof of the TREM algorithm is provided in [33].

References

[1] T. Nishiura, R. Nishioka, T. Yamada, S. Nakamura, and K. Shikano, "Multiple beamforming with source localization based on CSP analysis," *Systems and Computers in Japan*, vol. 34, no. 5, pp. 69–80, 2003.

[2] K. Knuth, "Bayesian source separation and localization," in *Proceedings of SPIE*, vol. 3459, 1998, pp. 147–158.

[3] Y. Huang, J. Benesty, and G. Elko, "Passive acoustic source localization for video camera steering," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2000, pp. 909–912.

[4] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

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

[6] U. Klee, T. Gehrig, and J. McDonough, "Kalman filters for time delay of arrival-based source localization," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 167–167, 2006.

[7] S. Gannot and T. Dvorkind, "Microphone array speaker localizers using spatial-temporal information," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 174–174, 2006.

[8] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, 2001, pp. 3021–3024.

[9] X. Zhong and J. Hopgood, "Nonconcurrent multiple speakers tracking based on extended kalman particle filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 293–296.

[10] A. Levy, S. Gannot, and E. Habets, "Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments," *IEEE Transactions on Audio, Speech, and Language Processing*, no. 99, pp. 1540–1555, 2009.

[11] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 728–739, 2008.

[12] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, pp. 384–391, 2000.

[13] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 1110–1124, 2003.

[14] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177 – 204, 2005.

[15] K. Yao, J. Chen, and R. Hudson, "Maximum-likelihood acoustic source localization: experimental results," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, 2002, pp. III–2949.

[16] 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, apr 1985, pp. 1762–1765.

[17] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[18] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *International Computer Science Institute*, vol. 4, p. 126, 1998.

[19] D. Titterington, "Recursive parameter estimation using incomplete data," *J. Roy. Statist. Soc. Ser. B*, vol. 46, pp. 257–267, 1984.

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

[21] L. Frenkel and M. Feder, "Recursive expectation-maximization (EM) algorithms for time-varying parameters with applications to multiple target tracking," *IEEE Transactions on Signal Processing*, vol. 47, no. 2, pp. 306–320, 1999.

[22] O. Cappé, M. Charbit, and E. Moulines, "Recursive EM algorithm with applications to DOA estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[23] P. Chung, J. Böhme, and A. Hero, "Tracking of multiple moving sources using recursive EM algorithm," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 50–60, 2005.

Copyright (c) 2013 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2013.2292361

[24] N. Madhu and R. Martin, "Acoustic source localization with microphone arrays," *Advances in Digital Speech Transmission*, pp. 135–170, 2008.

[25] ——, "A scalable framework for multiple speaker localization and tracking," in *International Workshop for Acoustic Echo Cancellation and Noise Control (IWAENC), Seattle, WA*, 2008.

[26] M. Mandel, D. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," *Advances in Neural Information Processing Systems*, vol. 19, p. 953, 2007.

[27] M. Mandel and D. Ellis, "The ideal interaural parameter mask : A bound on binaural separation systems," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 85–88.

[28] D. Kazakos, "Recursive estimation of prior probabilities using a mixture," *IEEE Transactions on Information Theory*, vol. 23, no. 2, pp. 203–211, 1977.

[29] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge Univ Press, 2004.

[30] K. Petersen and M. Pedersen, "The matrix cookbook," *Technical University of Denmark*, pp. 45–46, 2008.

[31] J. Garofolo, *Darpa TIMIT: Acoustic-phonetic Continuous Speech Corps CD-ROM*. US Department of Commerce, National Institute of Standards and Technology, 1993.

[32] K. Lange, "A gradient algorithm locally equivalent to the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 425–437, 1995.

[33] C. Wu, "On the convergence properties of the EM algorithm," *The Annals of Statistics*, vol. 11, no. 1, pp. 95–103, 1983.

Ofer Schwartz received his B.Sc. (Cum Laude) and M.Sc. degrees in electrical engineering from Bar-Ilan University, Israel in 2010 and 2013 respectively. His research interests include statistical signal processing and in particular speech enhancement using microphone arrays and speaker localization and tracking.

Sharon Gannot (S'92-M'01-SM'06) received his B.Sc. degree (summa cum laude) from the Technion Israel Institute of Technology, Haifa, Israel in 1986 and the M.Sc. (cum laude) and Ph.D. degrees from Tel-Aviv University, Israel in 1995 and 2000 respectively, all in electrical engineering. In 2001 he held a post-doctoral position at the department of Electrical Engineering (ESAT-SISTA) at K.U.Leuven, Belgium. From 2002 to 2003 he held a research and teaching position at the Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. Currently, he is an Associate Professor at the Faculty of Engineering, Bar-Ilan University, Israel, where he is heading the Speech and Signal Processing laboratory. Prof. Gannot is the recipient of Bar-Ilan University outstanding lecturer award for 2010.

Prof. Gannot has served as an Associate Editor of the EURASIP Journal of Advances in Signal Processing in 2003-2012, and as an Editor of two special issues on Multi-microphone Speech Processing of the same journal. He has also served as a guest editor of ELSEVIER Speech Communication and Signal Processing journals. Prof. Gannot has served as an Associate Editor of IEEE Transactions on Speech, Audio and Language Processing in 2009-2013. Currently, he is a Senior Area Chair of the same journal. He also serves as a reviewer of many IEEE journals and conferences. Prof. Gannot is a member of the Audio and Acoustic Signal Processing (AASP) technical committee of the IEEE since Jan., 2010. He is also a member of the Technical and Steering committee of the International Workshop on Acoustic Signal Enhancement (IWAENC) since 2005 and was the general co-chair of IWAENC held at Tel-Aviv, Israel in August 2010. Prof. Gannot has served as the general co-chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in October 2013. Prof. Gannot was selected (with colleagues) to present a tutorial sessions in ICASSP 2012, EUSIPCO 2012, ICASSP 2013 and EUSIPCO 2013. Prof. Gannot research interests include multi-microphone speech processing and specifically distributed algorithms for ad hoc microphone arrays for noise reduction and speaker separation; dereverberation; single microphone speech enhancement and speaker localization and tracking.

Copyright (c) 2013 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.