Tree-based recursive expectation-maximization algorithm for localization of acoustic sources

Yuval Dorfan and Sharon Gannot

***Abstract*— The problem of distributed localization for ad hoc wireless acoustic sensor networks (WASNs) is addressed in this paper. WASNs are characterized by low computational resources in each node and by limited connectivity between the nodes. Novel bi-directional tree-based distributed expectation-maximization (DEM) algorithms are proposed to circumvent these inherent limitations. We show that the proposed algorithms are capable of localizing static acoustic sources in reverberant enclosures without a priori information on the number of sources. Unlike serial estimation procedures (like ring based algorithms), the new algorithms enable simultaneous computations in the nodes and exhibit greater robustness to communication failures. Specifically, the recursive distributed EM (RDEM) variant is better suited to online applications due to its recursive nature. Furthermore, the RDEM outperforms the other proposed variants in terms of convergence speed and simplicity. Performance is demonstrated by an extensive experimental study consisting of both simulated and actual environments.**

***Index Terms*— Recursive expectation-maximization; Distributed signal processing; Bi-directional tree topologies; Wireless acoustic sensor networks; Speaker localization.**

I. Introduction

The localization of multiple acoustic sources has various civil [1] and military [2] applications. Both Bayesian [3],[4],[5] and non-Bayesian [6],[7] localization approaches have been proposed in the literature. The algorithms derived in this work belong to the family of non-Bayesian estimation algorithms, more specifically to the maximum likelihood (ML) estimation family of algorithms. ML estimation procedures for localization are usually characterized by high computational complexity and the nonexistence of closed-form solutions. For these reasons either iterative expectation-maximization (EM) procedures [8] or recursive EM (REM) procedures [9] have been suggested.

The sparsity of speech signals in the short-time Fourier transform (STFT) domain is widely used in the context of speaker localization [9],[10],[11]. In [9] for example, the localization task was carried out by using spatially distributed microphone nodes (more specifically, each node was comprised of a pair of microphones).

Sometimes, due to limited computational resources in each node and the bandwidth (BW) constraint on the communication link connecting the nodes, *distributed computation* is resorted to. Distributed networks can be utilized to jointly estimate parameters [12],[13] or signals by applying beamforming techniques [14],[15].

Distributed localization algorithms for WASN that only use the amplitude or power of the received signals, but ignore phase information, can be found in [16],[17],[18]. It is well-known from the radio frequency (RF) literature, that localization schemes based on phase differences exhibit higher accuracy, robustness and lower sensitivity than schemes that are solely based on the received signal strength [19].

An EM-based method to estimate the angle of arrival for multiple sources that uses phase information extracted from stereo recordings was proposed in [20]. This method was extended to a two dimensional localization problem by an array of microphone pairs in [9]. In addition, in [9] two recursive algorithms were proposed based on the Cappé and Moulines recursive EM (CREM) [21] and the Titterington recursive EM (TREM) [22] schemes for multiple acoustic source tracking. Both [9] and [20] use *centralized computation* approaches.

A DEM scheme was presented by Nowak in [23] for clustering stochastic variables obeying a mixture of Gaussians (MoG) probability density function (p.d.f.). MoG p.d.f. is also the underlying probabilistic model in [9]. A detailed description of EM usage for MoG can be found in [24], where *latent indicators* are used for the estimation of the MoG parameters. Our derivations are based on the same latent variables. Several paradigms for distributed inference of the parameters of a MoG p.d.f., which are based on the EM framework, are summarized in [25]. This contribution does not address speech signals and are therefore not utilizing the specific attributes of speech signals in acoustic environments, e.g. sparsity and reverberation. In [26] three distributed strategies are presented for the MoG EM (incremental, consensus and diffusion).

An incremental variant of the EM, denoted incremental EM (IEM), was proposed by Neal and Hinton [27]. The REM mechanism was also addressed in this contribution. The convergence speed and accuracy of the IEM procedure was analyzed and demonstrated in [28] and [29], where it is shown through many examples that the incremental strategy enables faster convergence. In addition, the IEM has a higher probability to converge to the ML. In other words, it does not tend to converge to a local optimum. There is no mathematical proof for these properties, but they are explained intuitively and demonstrated empirically.

Based on this method we recently proposed a distributed scheme for source localization [30] which we dubbed incremental distributed expectation-maximization (IDEM). It was implemented over a *directed-ring* constellation. The IEM concept was also mentioned in [25] in the context of a local E-step computation.

Distributed computation, as opposed to centralized solu-

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

tions, requires a new perspective on the EM iterations introducing local *hidden variables* in conjunction or instead of the global hidden variables addressed above. The current work adopts this paradigm and proposes a novel DEM family of algorithms for phase-based localization of acoustic sources in reverberant environments. We propose two different algorithms based on a bi-directional tree topology (for a detailed description of network topologies the reader is referred to [31].). Both proposed algorithms employ estimation of the local hidden variables.

In the first algorithm, dubbed batch distributed EM (BDEM), a simultaneous E-step followed by a global M-step are implemented over a bi-directional tree topology. The leaves of the tree execute *partial* E-step based on a batch of local measurements, and aggregates the results towards the root of the tree in the *fusion* stage. The root of the tree applies the M-step to produce the next estimation of the parameters (the root is the only node to apply the M-step). This estimation is then broadcast through the same bi-directional tree topology in the opposite direction towards all leaves in the *diffusion* stage. The fusion and the subsequent diffusion constitute one iteration of the algorithm.

As explained above, recursive variants of the classical EM method; namely REM schemes that exhibit some advantages over the EM framework, were introduced in [21],[22] and [27]. The second algorithm proposed in this work, dubbed recursive distributed EM, uses the TREM [22] scheme for recursively estimating the global parameters. In the proposed RDEM scheme, the BDEM iterations are substituted by a recursion along the time-axis.

The RDEM is also defined on the bi-directional tree topology. It also propagates the local estimations towards the root and then (after applying the M-step) broadcast the recent parameter estimates in the opposite direction. Unlike the BDEM, the RDEM does not operate in a batch mode, but rather adapts a recursive mode that runs along the time axis, frame by frame.

In this work we show that the RDEM has several advantages over the BDEM scheme in online applications. In addition, it is demonstrated that the RDEM and BDEM outperform the classical steered response power-phase transform (SRP-PHAT) algorithm [32] in scenarios with two concurrent speakers in an acoustic room.

The remainder of this paper is organized as follows. In Section II we present the statistical model for localizing concurrent sources in a noisy and reverberant environment. The new local hidden variables are presented in Section III. Two novel DEM algorithms for distributed localization based on the bi-directional tree are proposed in Section IV. Section V deals with implementation issues related to the proposed algorithms. Section VI is dedicated to the experimental study, based on simulations as well as actual recordings carried out in our acoustic lab. Conclusions are drawn in Section VII.

II. The Statistical Model

The problem is formulated in the STFT domain with $t = 1, \dots, T$ as the time index and $k = 0, \dots, K-1$ as the frequency index. S acoustic signals are captured by M microphone pairs. The signal received by the ith microphone, $i = 1, 2$, of the mth pair, $m = 1, \dots, M$, is given by:

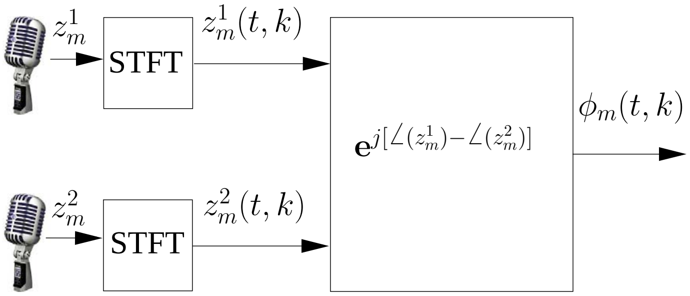


Fig. 1

PRE-PROCESSING AT THE mTH MICROPHONE PAIR TO EXTRACT THE PRP.

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

where $s = 1, \dots, S$ is the source index. $v_s(t,k)$ denotes the sth source signal, $n_m^i(t,k)$ denotes additive noise as captured by the ith microphone of the mth pair, and $a_{sm}^i(t,k)$ denotes the acoustic transfer function (ATF) from the sth source to the ith microphone of the mth node. The ATF in reverberant environments consists of a direct path (which bears the desired information for localization) and reflections (which usually degrade the localization performance). The model in (1) is commonly referred to as the multiplicative transfer function (MTF) model [33]. It is only valid if the STFT window-length is sufficiently larger than the reverberation time T_{60}. If this does not hold, the indicated relation can be viewed as an approximation.

The first stage of all localization procedures discussed below consists of pair-wise relative phase ratio (PRP) extraction, here given for the mth microphone pair:

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

The rationale for using these PRP values is explained below. First, consider a single source in a noiseless environment:

$$z_m^i(t,k) = a_{sm}^i(t,k) v_s(t,k); \; \forall m, s = 1, i = 1, 2. \qquad (3)$$

To motivate the usage of the PRP, we first assume similar amplitudes. The time difference between the received signals at the microphone pair is then given by:

$$z_m^2(t,k)/z_m^1(t,k) = e^{jw(\tau_m^2 - \tau_m^1)} \forall m. \qquad (4)$$

Based on this simple result, the PRP can be interpreted as an extension of the time difference of arrival (TDOA) to the reverberant case and (4) only approximately holds.

A schematic block diagram of the pre-processing stage is depicted in Fig. 1. STFT is applied to each microphone signal and the PRPs are subsequently calculated for each time-frequency bin separately. These PRPs are induced by the

TDOA between the microphone-pair signals as a response to an acoustic source located in $\mathbf{p} \in \mathcal{P}$:

$$\tau_m(\mathbf{p}) \triangleq \frac{||\mathbf{p} - \mathbf{p}_m^2|| - ||\mathbf{p} - \mathbf{p}_m^1||}{c}, \tag{5}$$

where $\mathbf{p}_m^1$ and $\mathbf{p}_m^2$ are the locations of the microphones in pair m, $||\cdot||$ denotes the Euclidian norm and c is the sound velocity. $\mathcal{P}$ is a set of all possible source locations in the enclosure[1]. In this work we selected a regular grid of possible locations with a desired resolution. Note that any PRP can be associated with multiple source locations. The locus of all these locations is a one-sheet hyperboloid.

The various speakers are assumed to exhibit sparse activity in the STFT domain [9],[10] and [11]. This assumption is often referred to as the W-disjoint orthogonality of the speech signal [34].

Under this assumption, and an upper bound on the number of concurrent speakers, each time-frequency bin can be associated with a single active speaker (and therefore, in our case, also with a single active position). The following *deterministic* nominal set of PRPs, $\tilde{\phi}_m^k(\mathbf{p})$, associated with a possible room position $\mathbf{p}$ on the predefined grid, can be calculated in advance:

$$\tilde{\phi}_m^k(\mathbf{p}) \triangleq \exp\left(-j\frac{2\pi k \tau_m(\mathbf{p})}{KT_s}\right);\ \forall \mathbf{p} \in \mathcal{P}, \tag{6}$$

where T_s denotes the sampling time.

We can now express the PRP per node and per time-frequency bin in the following statistical model:

$$\phi_m(t,k) \sim \sum_{\mathbf{p}\in\mathcal{P}} \psi_{\mathbf{p}} \mathcal{N}^c\left(\phi_m(t,k); \tilde{\phi}_m^k(\mathbf{p}), \sigma^2\right), \tag{7}$$

where $\psi_{\mathbf{p}}$ is the probability that the speaker emitting in time-frequency bin (t,k) is located at position $\mathbf{p}$. Similarly to [9], every speaker can be located in any (fixed) position within this set of speaker positions in the room. However, unlike [9] that uses a separate distribution function for each speaker, we simplify the model and use a single joint distribution function for all speakers. Using this simplification it is not necessary to determine the number of speakers in advance.

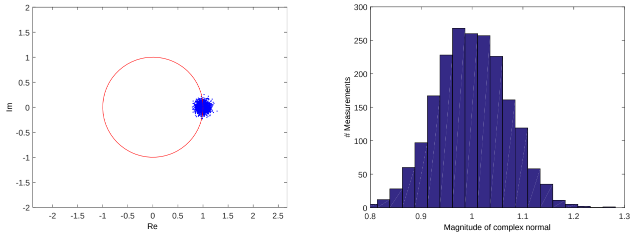
$\mathcal{N}^c(\cdot;\cdot,\cdot)$ denotes the complex-Gaussian probability with variance σ^2:

$$\mathcal{N}^c\left(\phi_m(t,k); \tilde{\phi}_m^k(\mathbf{p}), \sigma^2\right) = \frac{1}{\pi\sigma^2}\exp\left(-\frac{|\phi_m(t,k) - \tilde{\phi}_m^k(\mathbf{p})|^2}{\sigma^2}\right). \tag{8}$$

Being a probability function the following holds:

$$\sum_{\mathbf{p}\in\mathcal{P}} \psi_{\mathbf{p}} = 1, 0 < \psi_{\mathbf{p}} < 1; \forall \mathbf{p} \in \mathcal{P}. \tag{9}$$

The complex-Gaussian distribution, which cannot model the PRP with absolute value equals 1, is used here as in [9],[30], although inaccurate. In [20] the real-Gaussian is used for phase distribution, although it is inaccurate either. The von Mises distribution is sometimes used for TDOA estimation [35],[36] since it better fits the periodicity of phase measurements. To alleviate the model inaccuracies, we set the variances σ^2 (assumed equal for all complex-Gaussians in the mixture) to a small (known) value, increasing the probability that the absolute value of the PRP is sufficiently close to 1 (in our experiments we have set $\sigma^2 = 0.01$). The variance captures the level of the noise and the reverberation effects, and has therefore an important impact on the performance of the proposed algorithms. In a future study mechanisms for setting the variance value will be explored. In the current contribution, the value was empirically chosen. Scatter plot of Complex-Gaussian samples and their amplitude histogram are given in Fig. 2(a) and Fig. 2(b), respectively.



(a) Scatter plot. (b) Amplitude histogram.

Fig. 2

COMPLEX-GAUSSIAN DISTRIBUTION (VARIANCE SET TO 0.01).

Augmenting the PRP readings for all time-frequency bins, $\boldsymbol{\phi}_m = \text{vec}_{t,k}\left(\phi_m(t,k)\right)$, applying the W-disjoint property above and the assumption that the perturbations of the PRPs are independent, the p.d.f. of the observation set for each node m can be stated as [24]:

$$f(\boldsymbol{\Phi}_m = \boldsymbol{\phi}_m; \boldsymbol{\psi}) = \prod_{t,k}\sum_{\mathbf{p}\in\mathcal{P}} \psi_{\mathbf{p}} \mathcal{N}^c\left(\phi_m(t,k); \tilde{\phi}_m^k(\mathbf{p}), \sigma^2\right), \tag{10}$$

where we define the set of all probabilities of all possible grid positions in a vectorial notation:

$$\boldsymbol{\psi} = \text{vec}_{\mathbf{p}}\left(\psi_{\mathbf{p}}\right). \tag{11}$$

Note, that in [9], a separate vector of position probabilities is defined per speaker, and therefore should exhibit only one significant peak. Here, $\boldsymbol{\psi}$ should be understood as an aggregation of the probabilities over all sources, and hence might have several peaks, corresponding to the number of active sources in the measurement set. Besides model simplification, aggregating the probabilities provides an automatic mechanism for determining the number of concurrently active sources and to circumvent the requirement for setting this number in advance.

Sound localization from a single pair of microphones is widely covered in the binaural hearing literature. Some of these contributions assume the availability of a training set [37],[38]. Partial localization (azimuth and elevation angle) is dealt in [39]. Another approach [40] assumes some a priori knowledge of the surroundings and a single dynamic pair of microphones. The dynamics of the sensors enables

[1] Evidently, the speech sources can be located anywhere in the enclosure. Confining the possible locations to a finite set of room coordinates is equivalent to spatial sampling that reduces the localization resolution.

the localization of the source. In this work we consider the actual location of multiple sources; i.e., information that cannot be reliably inferred from a single microphone-pair, without limiting assumptions. We therefore concatenate all microphone-pair readings ($\boldsymbol{\phi} = \text{vec}_m\left(\boldsymbol{\phi}_m\right)$) to describe the probabilistic model of the source locations:

$$f(\boldsymbol{\Phi} = \boldsymbol{\phi}; \boldsymbol{\psi}) = \prod_m f(\boldsymbol{\Phi}_m = \boldsymbol{\phi}_m; \boldsymbol{\psi}) = \prod_{m,t,k} \sum_{\mathbf{p}\in\mathcal{P}} \psi_{\mathbf{p}} \mathcal{N}^c\left(\phi_m(t,k); \tilde{\phi}^k_m(\mathbf{p}), \sigma^2\right), \qquad (12)$$

where we assume that all microphone-pair readings are independent. This assumption can be partially justified by the different reflection patterns the signal undergoes before being captured by the microphone-pair, and more importantly, enables the derivation of the distributed algorithms proposed in this contribution. Note that the weights of the complex-Gaussians, namely $\psi_{\mathbf{p}}$, are common to all nodes, since they relate to the probability of obtaining a certain acoustic direct path from a specific location in the room, which is a global common parameter.

In the same way as described in [24], the maximum likelihood estimator (MLE) of the speakers' location (global parameter) can be obtained by maximizing the expression from equation (12) w.r.t. to $\boldsymbol{\psi}$ (note that σ^2 and $\tilde{\phi}^k_m(\mathbf{p})$ are a priori known):

$$\hat{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi}}{\text{argmax}}[\log f(\boldsymbol{\Phi} = \boldsymbol{\phi}; \boldsymbol{\psi}) \quad \text{s.t.} \sum_{\mathbf{p}\in\mathcal{P}} \psi_{\mathbf{p}} = 1; 0 < \psi_{\mathbf{p}} < 1;]. \qquad (13)$$

An example of a typical estimator of $\boldsymbol{\psi}$ as a function of all possible room positions is depicted in Fig. 3 for a two-dimensional case[2] with a resolution of 10 × 10cm. In the figure, the probability of having a source at each location in the room is represented by the z-axis with color (and height) that are proportional to its value (in natural log units). The final estimation of the number of sources and their locations can be deduced from this map by applying a proper threshold.

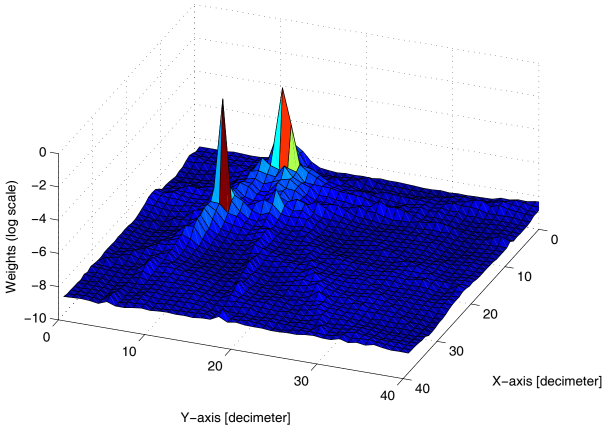


Fig. 3
MAP OF $\hat{\psi}_{\mathbf{p}}$ AS A FUNCTION OF THE ROOM'S FLOOR COORDINATES WITH A RESOLUTION OF 10 × 10CM. BY APPLYING A THRESHOLD TO THIS MAP IT CAN BE DEDUCED THAT THERE ARE TWO ACTIVE SOURCES IN THE ROOM.

Finally, we discuss the ability of the statistical model to alleviate the influence of reverberation on the localization accuracy. Although reverberation is not explicitly modelled, we claim that the MoG model implicitly takes the reverberation effect into account. The nominal PRPs, defined as the PRP induced on a microphone pair by a source located on a grid point, models only the direct-path of the sound propagation. This nominal value merely serves as a centroid of a complex-Gaussian. The variance of this complex-Gaussians allows for small deviations of the PRPs from their nominal value. Moreover, according to the MoG model, each source has a non-zero probability to be located in *any* grid point within the enclosure, thus allowing for large deviations from its true position due to strong reflections. Combining multiple PRP reading from several microphone pairs has a tendency to emphasize (i.e. increase the probability of) the true source positions, due to the incoherent nature of the reflections. Based on these arguments, we anticipate that the algorithms derived in this paper will be able to cope with low and medium reverberation levels. For higher level of reverberation levels, other mechanisms should be considered.

III. THE HIDDEN DATA

A straightforward maximization of the likelihood function in (12) for localizing concurrent acoustic sources is a cumbersome task. We therefore propose to maximize the likelihood by applying the EM procedure. More specifically, we derive and apply distributed versions of the EM for the problem at hand. The term DEM was coined by Nowak [23].

This section discusses the definitions of the hidden variables from the node (local) perspective. The definitions are used in the derivation of the DEM variants in this work. Global hidden variables in the context of EM-based speaker localization were proposed in [9]. Rather than centralized global hidden variables, we propose to define a local version thereof. As explained above, contrary to [9] we do not assume any prior information on the number of sources, and therefore define an aggregated set of position parameters, rather than defining a separate set of parameters for each source.

The hidden variable, denoted $y_m(t,k,\mathbf{p})$, is defined as the local mth indicator associating any speaker to a certain position $\mathbf{p} \in \mathcal{P}$ in time-frequency bin (t,k). The expectation of an indicator is readily given by:

$$E\left\{y_m(t,k,\mathbf{p})\right\} = \psi_{\mathbf{p}}. \qquad (14)$$

A vectorial version of the model proposed in [41] can be defined for the 2-D (or 3-D) positioning problem. Rather than the single-node local indicator $y_m(t,k,\mathbf{p})$, a global vector of all local indicators is used. Let $\mathbf{y}(t,k,\mathbf{p}) = \text{vec}_m\left(y_m(t,k,\mathbf{p})\right)$ be a set of all local indicators associated with a certain

[2] The methods derived in this work are also applicable to the three-dimensional case. However, to simplify the exposition we only present results for the two-dimensional case.

time-frequency bin. The local components of this vector are assumed independent identically distributed (i.i.d.). This assumption can be justified by the different reflection patterns of the sound waves, as measured in the different nodes. The expectation of this vector is therefore:

$$E\left\{\mathbf{y}(t,k,\mathbf{p})\right\} = \psi_{\mathbf{p}}\cdot\mathbf{1}, \tag{15}$$

where $\mathbf{1}$ is a vector of all ones of length M. We can further define a concatenated vector $\mathbf{y}(\mathbf{p}) = \text{vec}_{t,k}\left(\mathbf{y}(t,k,\mathbf{p})\right)$ to be the set of all indicators in the problem. Under the independency assumptions, stated above, in time, frequency and node, the probability density function of $\mathbf{y}$ is given by:

$$f(\mathbf{Y}=\mathbf{y};\boldsymbol{\psi}) = \prod_{t,k,m}\sum_{\mathbf{p}\in\mathcal{P}}\psi_{\mathbf{p}} y_m(t,k,\mathbf{p}). \tag{16}$$

The total number of local indicators in $\mathbf{y}$ is $M\times T\times K$. Their support set is $\mathbf{p}\in\mathcal{P}$, the set of all discrete positions on the grid. $|\mathcal{P}|$ stands for the cardinality of the support set. The p.d.f. of the observations is given by:

$$\begin{aligned} f(\boldsymbol{\Phi}=\boldsymbol{\phi}|\mathbf{Y}=\mathbf{y};\boldsymbol{\psi}) &= \prod_{m} f(\boldsymbol{\Phi}_m=\boldsymbol{\phi}_m|\mathbf{Y}_m=\mathbf{y}_m;\boldsymbol{\psi}) \\ &= \prod_{m,t,k}\sum_{\mathbf{p}\in\mathcal{P}} y_m\left(t,k,\mathbf{p}\right)\mathcal{N}^c\left(\phi_m(t,k);\tilde{\phi}^k_m(\mathbf{p}),\sigma^2\right). \end{aligned} \tag{17}$$

The p.d.f. of the *complete data* can be deduced from (16)-(17) and some simplifications utilizing the indicator properties:

$$\begin{aligned} f(\boldsymbol{\Phi}=\boldsymbol{\phi},\mathbf{Y}=\mathbf{y};\boldsymbol{\psi}) &= f(\mathbf{Y}=\mathbf{y};\boldsymbol{\psi})f(\boldsymbol{\Phi}=\boldsymbol{\phi}|\mathbf{Y}=\mathbf{y};\boldsymbol{\psi}) \\ &= \left(\prod_{m,t,k}\sum_{\mathbf{p}\in\mathcal{P}}\psi_{\mathbf{p}} y_m(t,k,\mathbf{p})\right) \\ &\times\left(\prod_{m,t,k}\sum_{\mathbf{p}\in\mathcal{P}} y_m(t,k,\mathbf{p})\mathcal{N}^c\left(\phi_m(t,k);\tilde{\phi}^k_m(\mathbf{p}),\sigma^2\right)\right) \\ &= \prod_{m,t,k}\sum_{\mathbf{p}\in\mathcal{P}}\psi_{\mathbf{p}} y_m(t,k,\mathbf{p})\mathcal{N}^c\left(\phi_m(t,k),\tilde{\phi}^k_m(\mathbf{p}),\sigma^2\right). \end{aligned} \tag{18}$$

IV. Tree-Based Distributed Expectation-Maximization

A family of distributed algorithms can be derived using the p.d.f. in (18). We propose two EM-based algorithms, namely the BDEM and the RDEM algorithms. They are both applied in a distributed manner using a network of microphones organized in a bi-directional tree topology. The algorithms differ in the way they process the data. While the BDEM algorithm processes the samples after all of them have been acquired (as a batch), the RDEM updates the estimation along the time-axis (online processing).

We will first derive the EM iterations common to both algorithms and will then present the BDEM and the RDEM algorithms.

A. EM Iterations

The EM algorithm for the problem at hand can now be derived. Let, (ℓ) be the iteration index. From (18) the E-step can be stated as:

$$\begin{aligned} &Q\left(\boldsymbol{\psi}|\hat{\boldsymbol{\psi}}^{(\ell-1)}\right)\triangleq \\ &\quad E\left\{\log\left(f(\boldsymbol{\Phi}=\boldsymbol{\phi},\boldsymbol{Y}=\mathbf{y};\boldsymbol{\psi})\right)|\boldsymbol{\phi};\hat{\boldsymbol{\psi}}^{(\ell-1)}\right\} = \\ &\quad\sum_{m,t,k}\log\Bigg(\sum_{\mathbf{p}\in\mathcal{P}} E\{y_m(t,k,\mathbf{p})|\phi_m(t,k);\hat{\boldsymbol{\psi}}^{(\ell-1)}\} \\ &\qquad\qquad \psi_{\mathbf{p}}\mathcal{N}^c(\phi_m(t,k);\tilde{\phi}^k_m(\mathbf{p}),\sigma^2)\Bigg) = \\ &\quad\sum_{m,t,k,\mathbf{p}\in\mathcal{P}} E\left\{y_m(t,k,\mathbf{p})|\phi_m(t,k);\hat{\boldsymbol{\psi}}^{(\ell-1)}\right\} \\ &\qquad\cdot\left[\log\psi_{\mathbf{p}}+\log\mathcal{N}^c(\phi_m(t,k);\tilde{\phi}^k_m(\mathbf{p}),\sigma^2)\right], \end{aligned} \tag{19}$$

where we have used the indicator property allowing to exchange the log and $\sum$ operators. In addition, the expression $E\left\{y_m(t,k,\mathbf{p})|\phi_m(t,k);\hat{\boldsymbol{\psi}}^{(\ell-1)}\right\}$ is a node specific entity that depends solely on a single observed random variable, due to the independency in the time, frequency and node domains. This node specific expectation is given by [24]:

$$\begin{aligned} \upsilon^{(\ell)}_m(t,k,\mathbf{p}) &\triangleq E\left\{y_m(t,k,\mathbf{p})|\phi_m(t,k);\hat{\boldsymbol{\psi}}^{(\ell-1)}\right\} = \\ &\frac{\hat{\psi}^{(\ell-1)}_{\mathbf{p}}\mathcal{N}^c\left(\phi_m(t,k);\tilde{\phi}^k_m(\mathbf{p}),\sigma^2\right)}{\sum_{\tilde{\mathbf{p}}\in\mathcal{P}}\hat{\psi}^{(\ell-1)}_{\tilde{\mathbf{p}}}\mathcal{N}^c\left(\phi_m(t,k);\tilde{\phi}^k_m(\tilde{\mathbf{p}}),\sigma^2\right)}. \end{aligned} \tag{20}$$

Applying a constrained maximization to (19), the required position probabilities are obtained:

$$\hat{\psi}^{(\ell)}_{\mathbf{p}} = \frac{\sum_{m,t,k}\upsilon^{(\ell)}_m(t,k,\mathbf{p})}{M\cdot T\cdot K}. \tag{21}$$

Note, that in the proposed algorithm, the variances of the complex-Gaussians are assumed to be known and their means can be calculated in advanced from the known grid positions. The only parameter set to be estimated in our problem is therefore $\psi_{\mathbf{p}}$.

The EM algorithm iterates between the E-step in (20) and the M-step in (21) until convergence.

B. Tree-Based Batch Distributed EM Algorithm

The EM iterations (20)-(21) can be centrally applied. A similar procedure, for a slightly different signal model, can be found in [9]. In this work we are aiming at distributed versions of the EM iterations above. The algorithms derived in the sequel are based on bi-directional tree topology, as exemplified in Fig. 4. Node '0' is designated as the root of the tree and all other nodes as its leaves. Tree topologies are less sensitive to communication link failures than e.g. directed-ring topologies [30]. Failure handling is beyond the scope of this contribution.

The key point in developing the distributed versions of the EM algorithm is the availability of local estimates of the hidden variable in each node, as evident from (20).

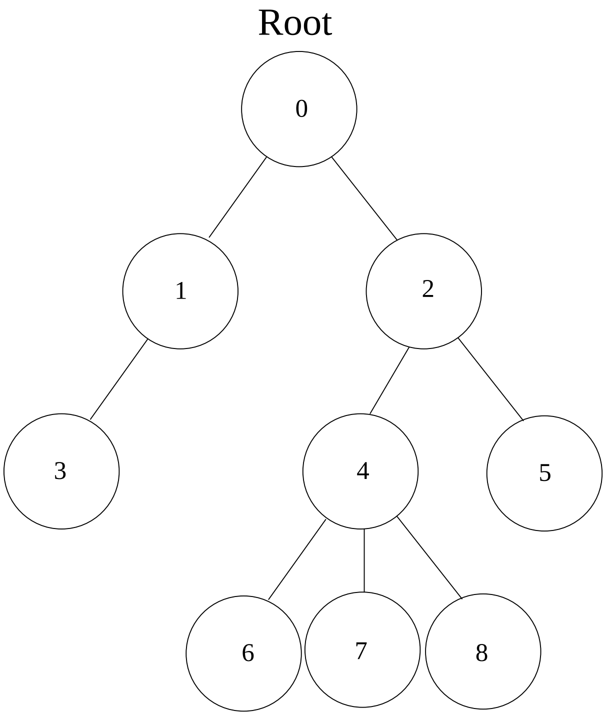


Fig. 4
A BI-DIRECTIONAL TREE NETWORK TOPOLOGY USED FOR THE PROPOSED DEM ALGORITHMS. THE EDGES REPRESENT TWO-WAY COMMUNICATION LINKS. NODE '0' IS DESIGNATED THE ROOT OF THE TREE.

Define the *average* local estimates:

$$\mu^{(\ell)}(t,k,\mathbf{p}) \triangleq \frac{1}{M}\sum_m \upsilon_m^{(\ell)}(t,k,\mathbf{p}). \tag{22}$$

that can be interpreted as a *global* estimate of the *local* hidden variables. This global entity reduces the communication volume between the nodes.

In the *fusion* stage of the proposed BDEM algorithm, these local estimates are aggregated from the leaves towards the root of the tree. The average $\mu^{(\ell)}(t,k,\mathbf{p})$, available only at the root, constitutes the E-step of the algorithm, yielding the most recent estimate of the hidden variables $y_m(t,k,\mathbf{p})$. The root is then ready to apply the M-step, yielding the current parameter estimate:

$$\hat{\psi}_{\mathbf{p}}^{(\ell)} = \frac{\sum_{t,k}\mu^{(\ell)}(t,k,\mathbf{p})}{T\cdot K}. \tag{23}$$

This estimate is subsequently *diffused* to all the leaves of the network. Note, that only the root applies the M-step, while the leaves are only responsible for partial update of the E-step and for bi-directional data transmission. All leaves can simultaneously update their contribution to the E-step, based on the current parameter estimate.

To further reduce the communication volume, each node (leaf) can average all time-frequency bins before transmission to the next node:

$$\bar{\upsilon}_m^{(\ell)}(\mathbf{p}) \triangleq \frac{\sum_{t,k}\upsilon_m^{(\ell)}(t,k,\mathbf{p})}{TK}. \tag{24}$$

Algorithm 1: Acoustic source localization with the BDEM algorithm.

```
Acquire $z_m^1(t,k)$ and $z_m^2(t,k)$; $\forall m$.
Calculate $\phi_m(t,k)$; $\forall m$ using (2).
set $\tilde{\phi}_m^k(\mathbf{p})$ using (6).
initialize $\hat{\psi}_{\mathbf{p}}^{(0)}$.
for $\ell = 1$ to $L$ do
    E-step
    $\forall m = 1:M$ calculate simultaneously and locally
    $\bar{\upsilon}_m^{(\ell)}(\mathbf{p})$ using (24).
    M-step
    Fuse local results from the leaves back to the root
    and estimate the parameters, $\hat{\psi}_{\mathbf{p}}^{(\ell)}$ using (25).
    Diffuse the result from the root of the bi-directional
    tree to all leaves.
end
By applying a threshold to $\hat{\psi}_{\mathbf{p}}^{(L)}$, the final estimation of
the number of sources $S$, and their respective locations
$\mathbf{p_s}$; $s = 1,\ldots,S$ can be obtained.
```

The M-step, applied in the root, then simplifies to:

$$\hat{\psi}_{\mathbf{p}}^{(\ell)} = \bar{\mu}^{(\ell)}(\mathbf{p}) \triangleq \frac{\sum_m \bar{\upsilon}_m^{(\ell)}(\mathbf{p})}{M}. \tag{25}$$

The algorithm iterates until convergence or until a number of pre-defined iterations has been reached. Algorithm 1 summarizes the proposed BDEM algorithm.

C. Tree-based Recursive Distributed EM Algorithm

A recursive version of the BDEM, denoted RDEM, is derived below. Although this paper only addresses static scenarios, using a recursive online algorithm can still be beneficial in several aspects, namely reducing latency and computational load. As demonstrated in Section VI, the RDEM also exhibits improved performance.

Recursive EM versions have been derived in [21] and [22]. In this paper we adopt the TREM version [22]. The basic adaptation scheme for the parameter of interest, $\boldsymbol{\psi}$ is given by:

$$\begin{aligned}\hat{\boldsymbol{\psi}}_R^{(t)} &= \hat{\boldsymbol{\psi}}_R^{(t-1)} + \\ &\gamma_t \boldsymbol{I}^{-1}_{\boldsymbol{y}_t,\boldsymbol{\phi}_t;\hat{\boldsymbol{\psi}}_R^{(t-1)}} \nabla_{\boldsymbol{\psi}} \log f(\boldsymbol{\Phi}_t = \boldsymbol{\phi}_t;\boldsymbol{\psi})\big|_{\hat{\boldsymbol{\psi}}_R^{(t-1)}},\end{aligned} \tag{26}$$

where subscript t in the notation $\mathbf{v}_t$ stands for all components of a certain vector, $\mathbf{v}$ of the current time frame. The parameter γ_t is often denoted the smoothing parameter of the recursion. Its value is chosen in the range $(0,1)$ and it has a major influence on the performance of the algorithm. Choosing too low value will slow down the converge rate of the algorithms. Overly large value will result in noisy position estimate.

The Fisher information matrix (FIM) is defined as:

$$\begin{aligned}&\boldsymbol{I}_{\boldsymbol{y}_t,\boldsymbol{\phi}_t;\hat{\boldsymbol{\psi}}_R^{(t-1)}} \triangleq \\ &\quad -E\left\{\nabla^2_{\boldsymbol{\psi}} \log f(\boldsymbol{Y}_t = \boldsymbol{y}_t, \boldsymbol{\Phi}_t = \boldsymbol{\phi}_t;\boldsymbol{\psi})\big|_{\hat{\boldsymbol{\psi}}_R^{(t-1)}}\right\}.\end{aligned} \tag{27}$$

We stress that the expectation operator, $E\{\cdot\}$ is a conditional expectation using the current parameter estimate $\hat{\boldsymbol{\psi}}_R^{(t)}$. Since the set of parameters $\psi_{\mathbf{p}}$ should be interpreted as a p.d.f., it satisfies the constraints from equation (9).

We therefore use the constrained version of TREM proposed in [9]:

$$\begin{aligned}\hat{\boldsymbol{\psi}}_R^{(t)} &= \hat{\boldsymbol{\psi}}_R^{(t-1)} + \\ &\gamma_t \boldsymbol{I}^{-1}_{\boldsymbol{y}_t,\boldsymbol{\phi}_t;\hat{\boldsymbol{\psi}}_R^{(t-1)}} \nabla_{\boldsymbol{\psi}} \log f(\boldsymbol{\Phi}_t = \boldsymbol{\phi}_t;\boldsymbol{\psi})|_{\hat{\boldsymbol{\psi}}_R^{(t-1)}} \\ &- \gamma_t \frac{\boldsymbol{I}^{-1}_{\boldsymbol{y}_t,\boldsymbol{\phi}_t;\hat{\boldsymbol{\psi}}_R^{(t-1)}}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{I}^{-1}_{\boldsymbol{y}_t,\boldsymbol{\phi}_t;\hat{\boldsymbol{\psi}}_R^{(t-1)}}\boldsymbol{a}}[\boldsymbol{a}^T\hat{\boldsymbol{\psi}}_R^{(t-1)} + \\ &\boldsymbol{a}^T\boldsymbol{I}^{-1}_{\boldsymbol{y}_t,\boldsymbol{\phi}_t;\hat{\boldsymbol{\psi}}_R^{(t-1)}} \nabla_{\boldsymbol{\psi}} \log f(\boldsymbol{\Phi}_t = \boldsymbol{\phi}_t;\boldsymbol{\psi})|_{\hat{\boldsymbol{\psi}}_R^{(t-1)}} - b],\end{aligned} \tag{28}$$

where in this case, due to the constraints in (9) we have:

$$\boldsymbol{a} = \mathbf{1}^T, b = 1. \tag{29}$$

To calculate the FIM, the expectation of the second derivative of the log-likelihood is required. The joint p.d.f. of the *instantaneous* measurements and the hidden variables (compare to (18)) is given by:

$$\begin{aligned}f(\boldsymbol{Y}_t = \boldsymbol{y}_t, \boldsymbol{\Phi}_t = \boldsymbol{\phi}_t;\boldsymbol{\psi}) &= \prod_{m,k}\sum_{\mathbf{p}\in\mathcal{P}} \psi_{\boldsymbol{p}} y_m(t,k,\boldsymbol{p}) \\ &\times \mathcal{N}^C\left(\phi_m(t,k), \tilde{\phi}_m^k(\boldsymbol{p}), \sigma^2\right).\end{aligned} \tag{30}$$

Applying the log operation and the indicator properties yields:

$$\begin{aligned}\log f(\boldsymbol{Y}_t = \boldsymbol{y}_t, \boldsymbol{\Phi}_t = \boldsymbol{\phi}_t;\boldsymbol{\psi}) &= \sum_{m,k,\mathbf{p}\in\mathcal{P}} y_m(t,k,\boldsymbol{p})\times \\ &\left(\log(\psi_{\boldsymbol{p}}) + \log\left(\mathcal{N}^C\left(\phi_m(t,k), \tilde{\phi}_m^k(\boldsymbol{p}), \sigma^2\right)\right)\right).\end{aligned} \tag{31}$$

Evaluating the second derivative of (31) at the parameter estimate at the previous time instant $(t-1)$ and using definition (11) yield:

$$\begin{aligned}-\frac{\partial^2}{\partial\psi_{\boldsymbol{p}}^2} \log f(\boldsymbol{Y}_t = \boldsymbol{y}_t, \boldsymbol{\Phi}_t = \boldsymbol{\phi}_t;\boldsymbol{\psi})|_{\hat{\psi}_R^{(t-1)}} \\ = \frac{\sum_{k,m} y_m(t,k,\boldsymbol{p})}{\left(\hat{\psi}_{\boldsymbol{p},R}^{(t-1)}\right)^2}.\end{aligned} \tag{32}$$

Taking the expectation and utilizing the local indicator property (15) result in:

$$\begin{aligned}E\left\{-\frac{\partial^2}{\partial\psi_{\boldsymbol{p}}^2} \log f(\boldsymbol{Y}_t = \boldsymbol{y}_t, \boldsymbol{\Phi}_t = \boldsymbol{\phi}_t;\boldsymbol{\psi})|_{\hat{\psi}_R^{(t-1)}}\right\} \\ = \frac{K\cdot M\cdot\hat{\psi}_{\boldsymbol{p},R}^{(t-1)}}{\left(\hat{\psi}_{\boldsymbol{p},R}^{(t-1)}\right)^2} = \frac{K\cdot M}{\hat{\psi}_{\boldsymbol{p},R}^{(t-1)}}.\end{aligned} \tag{33}$$

The p.d.f. of the current observation is given by (compare to (10) and (12)):

$$\begin{aligned}f(\boldsymbol{\Phi}_t = \boldsymbol{\phi}_t;\boldsymbol{\psi}) &= \\ \prod_{m,k}\sum_{\mathbf{p}\in\mathcal{P}} &\psi_{\mathbf{p}}\mathcal{N}^c\left(\phi_m(t,k);\tilde{\phi}_m^k(\mathbf{p}),\sigma^2\right).\end{aligned} \tag{34}$$

Taking the logarithm of (34) and calculating the derivative yield:

$$\begin{aligned}\frac{\partial}{\partial\psi_{\boldsymbol{p}}}\log f(\boldsymbol{\Phi_t} = \boldsymbol{\phi_t};\boldsymbol{\psi}) &= \\ \sum_{m,k}\frac{\mathcal{N}^C\left(\phi_m(t,k),\tilde{\phi}_m^k(\boldsymbol{p}),\sigma^2\right)}{\sum_{\tilde{\mathbf{p}}\in\mathcal{P}}\psi_{\tilde{\mathbf{p}}}\mathcal{N}^C\left(\phi_m(t,k),\tilde{\phi}_m^k(\tilde{\mathbf{p}}),\sigma^2\right)}&.\end{aligned} \tag{35}$$

Evaluating (35) at the parameter estimate at the previous time and expressing the result in terms of local terms yield:

$$\begin{aligned}\frac{\partial}{\partial\psi_{\boldsymbol{p}}}\log f(\boldsymbol{\Phi_t} = \boldsymbol{\phi_t};\boldsymbol{\psi})|_{\hat{\psi}_R^{(t-1)}} &= \\ \sum_{m,k}\frac{\upsilon_m^{(t)}(k,\boldsymbol{p})}{\hat{\psi}_{\boldsymbol{p},R}^{(t-1)}}&,\end{aligned} \tag{36}$$

where $\upsilon_m^{(t)}(k,\mathbf{p})$, the local hidden variables estimates, are calculated using the estimated parameters at time instant $(t-1)$:

$$\begin{aligned}\upsilon_m^{(t)}(k,\mathbf{p}) &\triangleq \\ \frac{\hat{\psi}_{\mathbf{p},R}^{(t-1)}\mathcal{N}^c\left(\phi_m(t,k);\tilde{\phi}_m^k(\mathbf{p}),\sigma^2\right)}{\sum_{\tilde{\mathbf{p}}\in\mathcal{P}}\hat{\psi}_{\tilde{\mathbf{p}},R}^{(t-1)}\mathcal{N}^c\left(\phi_m(t,k);\tilde{\phi}_m^k(\tilde{\mathbf{p}}),\sigma^2\right)}&.\end{aligned} \tag{37}$$

Define the frequency average of the estimated local hidden variables as:

$$\bar{\upsilon}_m^{(t)}(\boldsymbol{p}) \triangleq \frac{1}{K}\sum_k \upsilon_m^{(t)}(k,\boldsymbol{p}). \tag{38}$$

Averaging over all frequencies at each node enables a significant reduction of the communication BW.

Now, multiplying the inverse of the FIM (33) and the log-likelihood gradient (36) we can define the *current* parameter estimate (*before* recursion):

$$\begin{aligned}\hat{\psi}_{\boldsymbol{p}}^{(t)} &\triangleq \left(E\left\{-\frac{\partial^2}{\partial\psi_{\boldsymbol{p}}^2}\log f(\boldsymbol{Y}_t = \boldsymbol{y}_t, \boldsymbol{\Phi}_t = \boldsymbol{\phi}_t;\boldsymbol{\psi})|_{\hat{\psi}_{\boldsymbol{p},R}^{(t-1)}}\right\}\right)^{-1} \\ &\times\frac{\partial}{\partial\psi_{\boldsymbol{p}}}\log f(\boldsymbol{\phi_t};\boldsymbol{\psi})|_{\hat{\psi}_{\boldsymbol{p},R}^{(t-1)}} = \frac{1}{M}\sum_m\bar{\upsilon}_m^{(t)}(\boldsymbol{p}) \triangleq \bar{\mu}^{(t)}(\mathbf{p}).\end{aligned} \tag{39}$$

Interestingly, this current estimates identifies with the frequency-averaged global estimate of the hidden variables as defined in (25).

Using (29) and (39) in (28) the recursive distributed estimation procedure simplifies to:

$$\hat{\boldsymbol{\psi}}_R^{(t)} = \hat{\boldsymbol{\psi}}_R^{(t-1)} + \gamma_t\left(\hat{\boldsymbol{\psi}}^{(t)} - \hat{\boldsymbol{\psi}}_R^{(t-1)}\right). \tag{40}$$

This simple recursive procedure is only applied at the root.

The parameters of the algorithm $\hat{\boldsymbol{\psi}}_R^{(t)}$ are uniformly initialized. Based on the recent parameter estimate (using the initialization at $t=0$) a local E-step is executed at every node of the bi-directional tree in the current time frame, yielding $\bar{\upsilon}_m^{(t)}(\mathbf{p})$.

This operation may, at the first glance, appear identical to the E-step of the BDEM algorithm. However, as an off-line algorithm, the BDEM calculates the E-step for all time frames

Algorithm 2: Acoustic source localization with the RDEM algorithm.

```
set $\tilde{\phi}_m^k(\boldsymbol{p})$ using (6).
initialize $\hat{\boldsymbol{\psi}}_R^{(0)}$.
for $t = 1$ to $T$ do
    Obtain $z_m^1(t,k)$ and $z_m^2(t,k)$; $\forall m$.
    Calculate $\phi_m(t,k)$; $\forall m$ using (2).
    Calculate simultaneously and locally $\bar{\upsilon}_m^{(t)}(\boldsymbol{p})$
    using (38) $\forall m = 1, \ldots, M$.
    Use the bi-directional tree to aggregate local results
    from leaves (fusion).
    Calculate at the root: $\hat{\boldsymbol{\psi}}_R^{(t)}$ using (40).
    Transmit $\hat{\boldsymbol{\psi}}_R^{(t)}$ from the root to all nodes for the next
    time frame (diffusion).
end
Find $S$, the number of sources, and their respective
locations $\mathbf{p_s}$; $s = 1, \ldots, S$ by applying a threshold to
$\hat{\boldsymbol{\psi}}_R^{(T)}$, which is the final result of the algorithm.
```

together. Its recursive version, the RDEM, processes each time frame separately immediately after being acquired.

The local hidden variables are averaged along the frequency axis locally and $\bar{\upsilon}_m^{(t)}(\boldsymbol{p})$ is aggregated through the bi-directional tree towards the root. The results are used by the root for the calculation of the current estimate (39) and the recursion equation (40). The root then transmits (diffuses) the updated parameters to all nodes through the tree before the next time frame is processed. This process is summarized in Algorithm 2.

The root of the bi-directional tree has two roles in this process. Its first role is to broadcast the latest global parameter estimation, $\hat{\boldsymbol{\psi}}_R^{(t)}$ to all nodes efficiently. The second role of the root (in the opposite direction) is to aggregate local results that will enable the calculation of the next recursion step.

In the following sections we evaluate the BDEM and the RDEM algorithms in terms of implementation and performance.

V. Implementation Analysis

This section deals with various implementation issues concerning the DEM algorithms. The major factors that influence the efficiency of the implementation are: L the number of iterations, M the number of nodes, T the number of time frames, K the number of frequency bins and $|P|$ the cardinality of the set of grid points. Here, we selected a regular grid of possible positions with a desired resolution, but other schemes are applicable as well. The grid and its resolution have an influence on the computational complexity.

A. Computational Complexity

The computational complexity is calculated by counting the number of basic mathematical operations. Multiplications, divisions, additions and subtractions are equally weighted.

Before calculating the complexity of each algorithm, a few general statements should be made. The following three steps are applied in the inner loops of the DEM algorithms: the M-step and the E-step. For the latter step, we identify an operation common to all algorithms. The complexity of the E-step, involving the calculation of the hidden variables $\upsilon_m(t,k,\mathbf{p})$, is $\mathcal{O}(|P|)$ per time-frequency bin and per node. The other operations, which are specific to the various algorithms, are listed below.

1) BDEM Complexity: Define L_{BDEM} the number of iterations for the BDEM algorithm. At each iteration the M-step requires $\mathcal{O}(M \cdot |P|)$. In addition, at each iteration per time-frequency bin and per node the E-step requires $\mathcal{O}(|P|)$ operations. Hence, the total number of operations is given by:

$$\begin{aligned} \text{CMP}_{\text{BDEM}} &= \mathcal{O}\left(L_{\text{BDEM}} \cdot (M \cdot |P| + T \cdot K \cdot M \cdot |P|)\right) \\ &= \mathcal{O}(L_{\text{BDEM}} \cdot M \cdot T \cdot K \cdot |P|). \end{aligned} \quad (41)$$

2) RDEM Complexity: The RDEM runs T time frames and requires no iterations. At each time frame the M-step requires $\mathcal{O}(M \cdot |P|)$ operations. In addition, at each time frame per node the E-step requires $\mathcal{O}(K \cdot |P|)$ operations. Hence, the total number of operations is given by:

$$\begin{aligned} \text{CMP}_{\text{RDEM}} = \mathcal{O}\left(T \cdot (M \cdot |P| + M \cdot K \cdot |P|)\right) = \\ \mathcal{O}(T \cdot M \cdot K \cdot |P|) < \text{CMP}_{\text{BDEM}}. \end{aligned} \quad (42)$$

B. Latency

In online systems latency is a critical issue. There are a few time constants that need to be defined to analyze the latency of the proposed algorithms.

The first constant, T_{B} is the block length (in seconds) required for reliable localization. The BDEM algorithm sometimes need a batch of a few seconds to estimate the locations. The second constant, T_{Global} is the latency caused by the global calculations applied by the algorithm. The third constant is T_{Local}, the latency resulting from the local calculations applied. For RDEM we denote global and local latencies as T_{Global}^R, which is smaller than T_{Global}, and T_{Local}^R, which is usually smaller than T_{Local}, since only the current time frame is processed.

1) BDEM Latency: The BDEM latency is given by:

$$\text{LTC}_{\text{BDEM}} = \mathcal{O}\left(T_B + L_{\text{BDEM}} \cdot (T_{\text{Global}} + T_{\text{Local}})\right). \quad (43)$$

2) RDEM Latency: One of the major reasons for using the RDEM is its reduced latency. The RDEM latency is smaller, since it does not require sample aggregation and since it applies no iterations. It is given by:

$$\text{LTC}_{\text{RDEM}} = \mathcal{O}(T_{\text{Global}}^R + T_{\text{Local}}^R) < \text{LTC}_{\text{BDEM}}. \quad (44)$$

C. Communication Bandwidth

The communication BW is another major issue, especially when BW or power are constrained.

1) BDEM Communication BW: The communication BW for the BDEM algorithm is small due to averaging along time and frequency. It is given by:

$$\text{BW}_{\text{BDEM}} = \mathcal{O}(L_{\text{BDEM}} \cdot M \cdot |P|). \quad (45)$$

2) RDEM Communication BW: The communication BW for the RDEM algorithm is small, since no iterations are applied and all frequencies are aggregated in the node before transmission:

$$\text{BW}_{\text{RDEM}} = \mathcal{O}(M \cdot T \cdot |P|). \tag{46}$$

D. Memory Requirements

The BDEM has no memory requirements:

$$\text{MEM}_{\text{BDEM}} = 0. \tag{47}$$

The RDEM only requires the most recent estimation of the parameters:

$$\text{MEM}_{\text{RDEM}} = \mathcal{O}(|P|). \tag{48}$$

E. Summary

Table I summarizes the computational complexity, latency, communication BW and memory requirements of the proposed algorithms.

When comparing the RDEM to the BDEM, significant improvements can be observed. The RDEM requires lower computational complexity and has lower latency. They both occupy a narrow BW. The RDEM memory requirements are small, but larger than those of the BDEM (which does not impose any storage requirements).

VI. Experimental Study

This section reports an experimental study of the two proposed DEM localization algorithms. As a reference algorithm we used a modified version of the SRP-PHAT [42].

In order to evaluate performance, we use both simulation and real-life recordings of concurrent sources. For simplicity, we limited the localization problem to the two-dimensional case. It should be noted that the algorithms can be applied to the three-dimensional case as well.

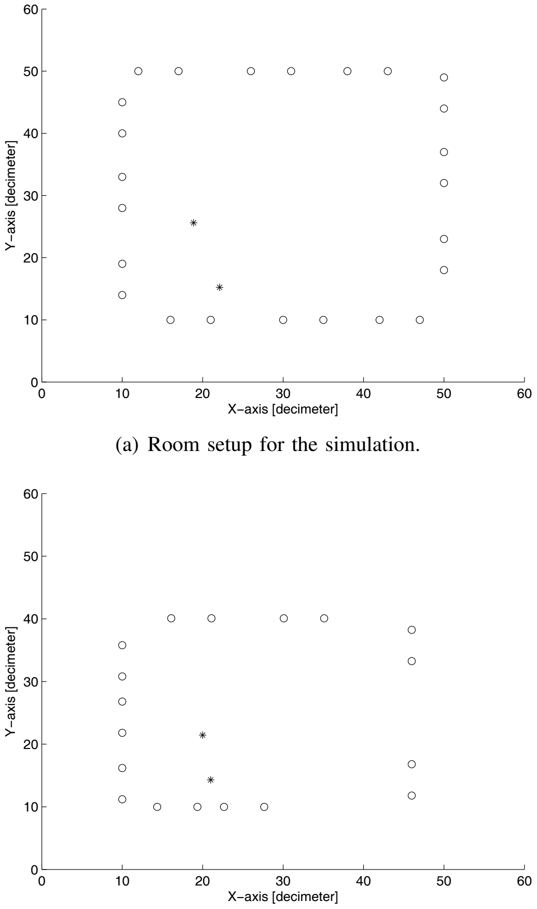
A. Practical Considerations

There are a few practical considerations to be addressed regarding the proposed algorithms.

1) Sensor Positions: As in most of the localization approaches, we also assume perfect knowledge of the sensors' positions in the room. As mentioned above, we assume a 2-D set-up purely for simplicity reasons. Therefore, the elevation value is ignored. However, the algorithms tested can be easily applied to 3-D cases as well. In the general case a 3-D location can be estimated.

2) Node Synchronization: We assume perfect synchronization between the nodes. In practice, it is most likely to have synchronization since cell phones and other equipment are synchronized through the network. In cases where there are clock differences, synchronization methods such as [43] can be adapted.

3) Microphone Inter-Distance in each Node: We used a microphone inter-distance of 50cm, which is a good compromise between resolution and ambiguity.



(a) Room setup for the simulation.

(b) Room setup for the recording analysis.

Fig. 5

Room setups for the experimental section. Microphone pairs are marked by pairs of circles and sources are marked by $*$.

4) Calculation Precision: For all calculations we used natural log operations, since they convert multiplications and divisions into additions and subtractions, while maintaining high precision. In other words, the probability $\hat{\psi}_{\mathbf{p}}^{(\ell)}$ is replaced by $\log\left(\hat{\psi}_{\mathbf{p}}^{(\ell)}\right)$. In a similar way, $\log\left(\upsilon_m^{(\ell)}(t,k,\mathbf{p})\right)$ is used rather than its original counterpart.

5) Number of iterations for BDEM: We used 20 iterations for the BDEM algorithm. If resources are limited, the number of iterations can be reduced to around 10 iterations.

B. Simulation Results

To evaluate the localization performance of the algorithms, we simulated the following scenario. Twelve pairs of omnidirectional microphones were located around a room. The dimensions of the simulated room were $6\times6\times4$m, with a low reverberation level, $T_{60} = 200$msec. Two sources randomly located in the room, were simulated using short speech files and an efficient implementation [44] of the image method [45]. An example of the speaker-microphone constellation is depicted in Fig. 5(a). In the simulations we have used the following values of the parameters. The sampling rate is 16KHz. The

Criteria	BDEM	RDEM
Computation	$\mathcal{O}(L_{\text{BDEM}} \cdot M \cdot T \cdot K \cdot \lvert P \rvert)$	$\mathcal{O}(M \cdot T \cdot K \cdot \lvert P \rvert)$
Latency	$\mathcal{O}(T_{\text{B}}) + \mathcal{O}(L_{\text{BDEM}}(T_{\text{Global}} + T_{\text{Local}}))$	$\mathcal{O}(T^R_{\text{Global}}) + \mathcal{O}(T^R_{\text{Local}})$
BW	$\mathcal{O}(L_{\text{BDEM}} \cdot M \cdot \lvert P \rvert)$	$\mathcal{O}(T \cdot M \cdot \lvert P \rvert)$
Memory	0	$\mathcal{O}(\lvert P \rvert)$

TABLE I
IMPLEMENTATION FEATURE TABLE. BDEM AND RDEM ARE COMPARED WITH RESPECT TO COMPUTATION, DELAY, COMMUNICATION BW AND MEMORY.

number of samples per frame is 1024 (64mSec) with 75% overlap. The number of spectrogram windows aggregated for localization is 245 frames, a bit less than 4 seconds. Note, that this amount of data is usually available in static localization tasks. The threshold applied to the SRP-PHAT maps is 0.0016. For the BDEM and RDEM we applied to the probability maps, a threshold which is the maximum between a fixed threshold (0.0014) and the maximal value of the map multiplied by a factor (0.55).

To compare the performance of the algorithms we followed the procedure described in [30]. We executed 100 Monte-Carlo trials and calculated three statistical measures: 1) The miss detection (MD) rate, defined as the percent of sources that were miss-detected out of the total number of sources; 2) The false alarm (FA) rate, defined as the percent of falsely-detected sources normalized by the total number of sources; and 3) The mean square error (MSE), defined as the accuracy of localization for all successfully detected sources. Note that the accuracy of the location estimation was limited by the grid resolution which was 10×10cm. Table II summarizes the measures for all algorithms. The reference algorithm SRP-PHAT [42] exhibits higher MD rate and FA rates than the proposed algorithms. The MSE of all algorithms was low with respect to the grid resolution.

Algorithm	MD[%]	FA[%]	MSE[cm]
SRP-PHAT	7.5	11.5	4
BDEM	6.5	9.0	4
RDEM	6.5	8.5	4

TABLE II
LOCALIZATION STATISTICS FOR 100 MONTE-CARLO TRIALS WITH TWO RANDOMLY LOCATED ACOUSTIC SOURCES.

C. Analysis of Actual Recordings

The algorithms were also tested using real recordings of two simultaneous sources and nine synchronized microphone pairs. Real-life recordings are important to validate localization algorithms, since some physical phenomena encountered in real-life scenarios (e.g. sources' volume, directivity of the emitted speech and the speaker orientation) cannot always be accurately simulated. For example, the spatial volume of the sources is complicated to simulate.

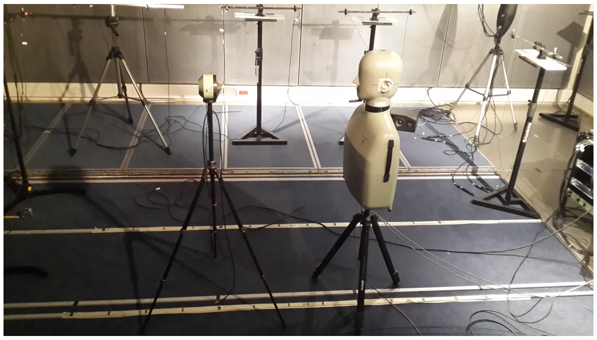


Fig. 6
EXPERIMENTAL SETUP IN THE SPEECH AND ACOUSTIC LAB OF THE ENGINEERING FACULTY AT BAR-ILAN UNIVERSITY. TWO SOURCES 61.5CM FROM EACH OTHER AND $T_{60} = 150$MSEC.

The recordings were carried out in the speech and acoustic lab of Bar-Ilan University. This is a $6 \times 6 \times 2.4$m room that has a reverberation time controlled by 60 interchangeable panels covering the room facets.

To simulate real human sources, we used a mouth simulator (B&K, type 4227) and a head and torso simulator (HATS) mannequin (B&K, type 4128C-002) to emulate head and torso shadowing effects. The measurement equipment also included a RME Hammerfall DSP Digiface sound-card and a RME Octamic (for Microphone Pre-Amplification and digitization (A/D)). AKG type CK-32 omnidirectional microphones were used. All measurements were carried out with a sampling frequency of 48KHz and a resolution of 24-bits. The multi-microphone signals were acquired using Matlab$^{\copyright}$. An example of the room layout is depicted in Fig. 5(b). Two different reverberation levels were tested by changing the panel configuration; namely $T_{60} = 150$msec (low) and $T_{60} = 450$msec (medium). For reverberation levels higher than $T_{60} = 450$msec, the localization results were not accurate enough. A picture of the room setup, with the two sources facing each other 61.5cm apart and a low reverberation level, is depicted in Fig. 6. In the following figures we depict the results of the various algorithms. In all figures, only the area encircled by the microphones is shown. Figure 7 depicts the localization probability maps ($\hat{\psi}_{\mathbf{p}}$) of the proposed algorithms and the output of the SRP-PHAT algorithm for the low reverberation level and with a source inter-distance of 61.5cm. The SRP-PHAT demonstrates poor resolution and exhibits a wide peak resulting in undistinguishable sources. The BDEM detected only one source. The RDEM algorithm detected the two sources.

Figure 8 depicts a two source localization (inter-distance of 71.5cm) for the medium reverberation level. The SRP-PHAT only detected one of the sources. The BDEM and RDEM algorithms detected both sources.

VII. Conclusion

In this paper we presented a family of DEM algorithms for multiple concurrent source localizations in reverberant environments, with an unknown number of sources.

Two novel algorithms that are members of this family were presented, namely the BDEM and the RDEM algorithms. They are both implemented over a bi-directional tree-based topology. They both use the same local hidden variables. We analyzed their implementation from a *distributed computation* point of view, by evaluating computational complexity, latency, communication BW and memory requirements. We also compared the two algorithms in terms of localization performance using simulations and real-life recordings. As a reference, we used a centralized algorithm, namely the SRP-PHAT.

The RDEM outperformed the other algorithms both in terms of performance and implementation. Moreover, it emerged as better suited for online applications, since samples are processed along the time axis (unlike batch processing) and no iterations are required.

Acknowledgments

We would like to thank Dr. Carmit Hazay for her advice on distributed topologies and their analysis. Special thanks to Dr. Gershon Hazan for his assistance in the lab recordings and especially for his advice and assistance in developing the algorithms.

References

[1] J. Kotus, K. Lopatka, and A. Czyzewski, “Detection and localization of selected acoustic events in acoustic field for smart surveillance applications,” *Multimedia Tools and Applications*, vol. 68, no. 1, pp. 5–21, Jan. 2014.

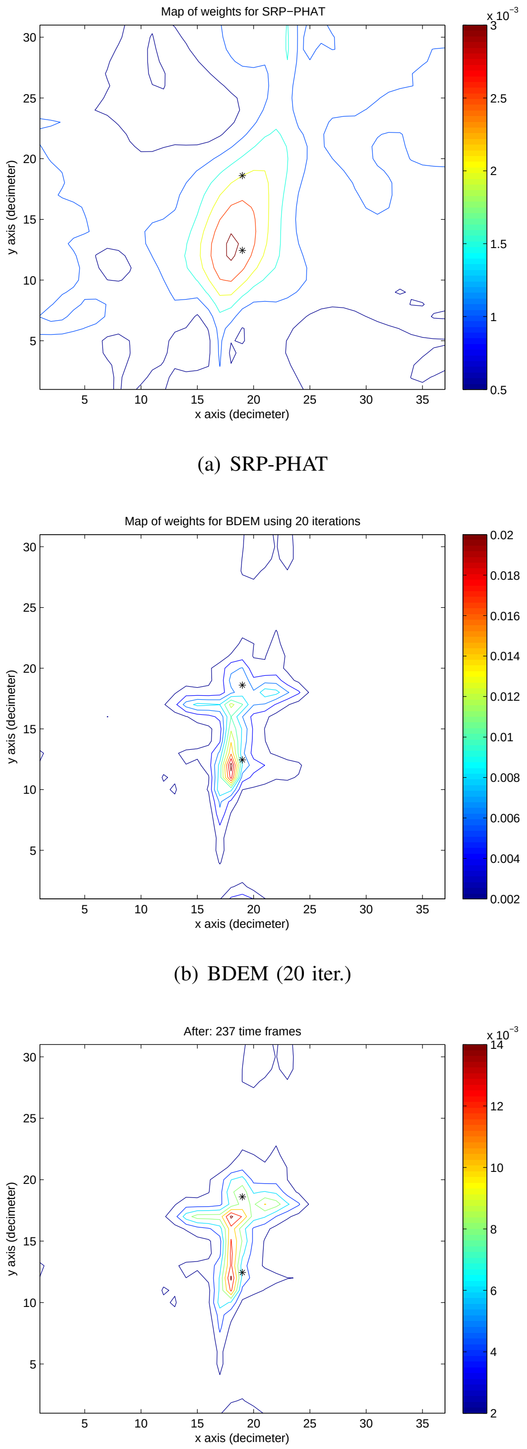
[2] J. Sallai, W. Hedgecock, P. Volgyesi, A. Nadas, G. Balogh, and A. Ledeczi, “Weapon classification and shooter localization using distributed multichannel acoustic sensors,” *Journal of Systems Architecture*, vol. 57, no. 10, pp. 869–885, Nov. 2011.

[3] S. E. Dosso and J. Dettmer, “Efficient Bayesian multi-source localization using a graphics processing unit,” *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, p. 070095, Jun. 2013.

[4] S. Prasad, “Asymptotics of Bayesian error probability and source super-localization in three dimensions,” *Optics Express*, vol. 22, no. 13, pp. 16 008–16 028, Jun. 2014.

[5] 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*, vol. 19, no. 6, pp. 1540–1555, Aug. 2011.

[6] M. Angjelichinoski, D. Denkovski, V. Atanasovski, and L. Gavrilovska, “SPEAR: Source position estimation for anchor position uncertainty reduction,” *IEEE Communications Letters*, vol. 18, no. 4, pp. 560–563, Apr. 2014.



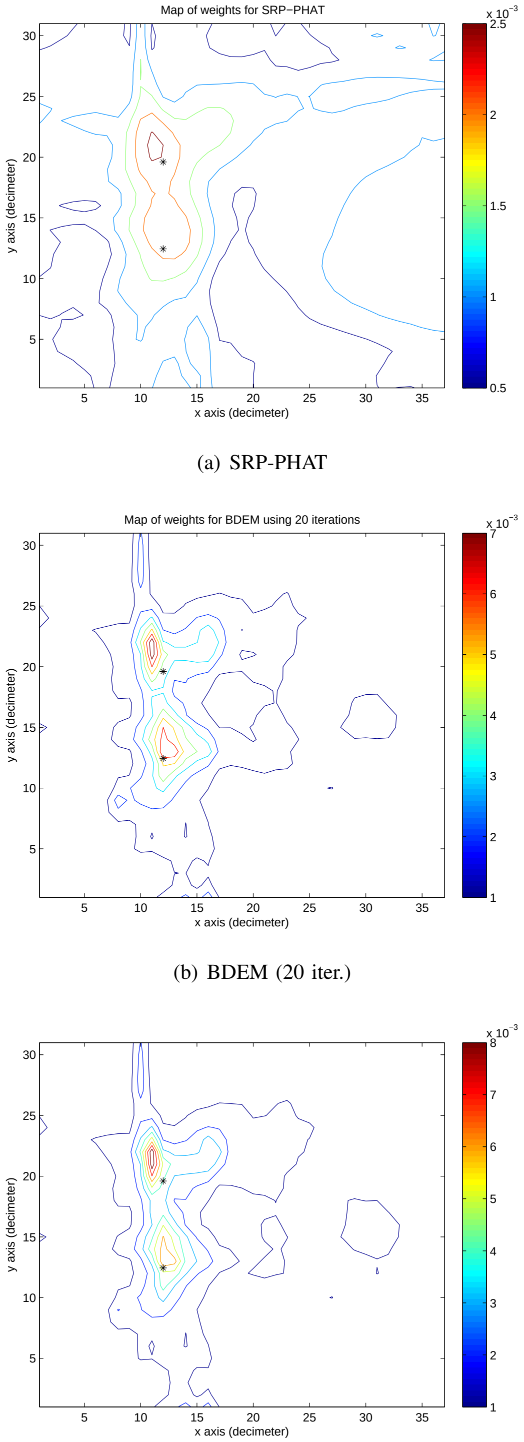
(a) SRP-PHAT

(b) BDEM (20 iter.)

(c) RDEM

Fig. 7

Map of one experimental trial for $T_{60} = 150$mSec and a source inter-distance of 61.5cm. The poor resolution of SRP-PHAT is depicted in (a). The BDEM detected only a single source, as depicted in (b). The result achieved by RDEM, the detection of both sources, is shown in (c).



(a) SRP-PHAT

(b) BDEM (20 iter.)

(c) RDEM

Fig. 8

MAP OF EXPERIMENTAL TRIAL WITH $T_{60} = 450$MSEC AND A SOURCE INTER-DISTANCE OF 71.5CM. THE SRP-PHAT ALGORITHM ONLY DETECTED ONE OF THE SOURCES, AS DEPICTED IN (A). THE BDEM AND RDEM ALGORITHMS WERE CAPABLE OF DETECTING BOTH SOURCES AS SHOWN IN (B)-(C).

[7] T. Routtenberg and J. Tabrikian, "Non-Bayesian periodic Cramér-Rao bound," *IEEE Transactions on Signal Processing*, vol. 61, no. 4, pp. 1019–1032, Feb. 2013.

[8] S. S. Iyengar, K. G. Boroojeni, and N. Balakrishnan, "Expectation maximization for acoustic source localization," in *Mathematical Theories of Distributed Sensor Networks*. Springer New York, Jan. 2014, pp. 37–54.

[9] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 392–402, 2014.

[10] N. Madhu and J. Wouters, "Localisation-based, situation-adaptive mask generation for source separation," in *Proceedings of the 4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, Mar. 2010, pp. 1–6.

[11] F. Nesta and M. Omologo, "Enhanced multidimensional spatial functions for unambiguous localization of multiple sparse acoustic sources," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 213–216.

[12] S. Haykin and K. J. R. Liu, *Handbook on Array Processing and Sensor Networks*. John Wiley & Sons, Feb. 2010.

[13] W. Yang, G. Chen, X. Wang, and L. Shi, "Stochastic sensor activation for distributed state estimation over a sensor network," *Automatica*, vol. 50, no. 8, pp. 2070–2076, Aug. 2014.

[14] A. Bertrand and M. Moonen, "Distributed signal estimation in sensor networks where nodes have different interests," *Signal Processing*, vol. 92, no. 7, pp. 1679–1690, Jul. 2012.

[15] S. Markovich-Golan and S. Gannot, "Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 343–356, 2013.

[16] D. Li and Y. H. Hu, "Energy-based collaborative source localization using acoustic microsensor array," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 4, pp. 321–337, Mar. 2003.

[17] D. Ampeliotis and K. Berberidis, "Low complexity multiple acoustic source localization in sensor networks based on energy measurements," *Signal Processing*, vol. 90, no. 4, pp. 1300–1312, Apr. 2010.

[18] Z. Liu, Z. Zhang, L.-W. He, and P. Chou, "Energy-based sound source localization and gain normalization for ad hoc microphone arrays," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Apr. 2007, pp. 761–764.

[19] C. Hekimian-Williams, B. Grant, X. Liu, Z. Zhang, and P. Kumar, "Accurate localization of RFID tags using phase difference," in *Proceedings of the IEEE International Conference on RFID*, Apr. 2010, pp. 89–96.

[20] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound: Sources in reverberant environments," *Proceedings of the 19th conference on advances in neural information processing systems*, pp. 953–960, 2007.

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

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

[23] R. Nowak, "Distributed EM algorithms for density estimation and clustering in sensor networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2245–2253, 2003.

[24] C. M. Bishop, *Pattern recognition and machine learning*. Springer, New York, 2006.

[25] D. Gu, "Distributed EM algorithm for gaussian mixtures in sensor networks," *IEEE Transactions on Neural Networks*, vol. 19, no. 7, pp. 1154–1166, 2008.

[26] Y. Weng, W. Xiao, and L. Xie, "Diffusion-based EM algorithm for distributed estimation of Gaussian mixtures in wireless sensor networks," *Sensors*, vol. 11, no. 6, pp. 6297–6316, 2011.

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

[28] P. Liang and D. Klein, "Online EM for unsupervised models," in *Proceedings of Human Language Technologies*, ser. NAACL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 611–619.

[29] M.-A. Sato, "Fast learning of on-line EM algorithm," *Rapport Technique, ATR Human Information Processing Research Laboratories*, 1999.

[30] Y. Dorfan, G. Hazan, and S. Gannot, "Multiple acoustic sources localization using distributed expectation-maximization algorithm," in *Proceedings of the 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, May 2014, pp. 72–76.

[31] N. A. Lynch, *Distributed Algorithms*. Morgan Kaufmann, Apr. 1996.

[32] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, ser. Digital Signal Processing. Springer, Berlin Heidelberg, Jan. 2001, pp. 157–180.

[33] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007.

[34] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, May 2002, pp. 29–532.

[35] I. Marković and I. Petrović, "Speaker localization and tracking with a microphone array on a mobile robot using von mises distribution and particle filtering," *Robotics and Autonomous Systems*, vol. 58, no. 11, pp. 1185–1196, 2010.

[36] F. Jacob, J. Schmalenstroeer, and R. Haeb-Umbach, "DOA-based microphone array postion self-calibration using circular statistics," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 116–120.

[37] A. Deleforge and R. Horaud, "The cocktail party robot: Sound source separation and localisation with an active binaural head," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 431–438.

[38] B. Laufer, R. Talmon, and S. Gannot, "Relative transfer function modeling for supervised source localization," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[39] A. Kulaib, M. Al-Mualla, and D. Vernon, "2D binaural sound localization for urban search and rescue robotic," in *Proceedings of the 12th Conference on Climbing and Walking Robots*, Istanbul, Turkey, Sep. 2009, pp. 9–11.

[40] Y.-C. Lu, M. Cooke, and H. Christensen, "Active binaural distance estimation for dynamic sources." in *INTERSPEECH*, 2007, pp. 574–577.

[41] M. Mandel, R. Weiss, and D. 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.

[42] H. Do, H. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Apr. 2007, pp. 121–124.

[43] D. Cherkassky and S. Gannot, "Blind synchronization in wireless sensor networks with application to speech enhancement," in *Proceedings of the Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 183–187.

[44] E. A. P. Habets, "Room impulse response (RIR) generator," http://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator, Sep. 2010.

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

Yuval Dorfan received his B.Sc. degree (summa cum laude) from Ben-Gurion University, Beer-Sheva, Israel in 1998, his M.Sc. degree (magna cum laude) from the Technion Israel Institute of Technology, Haifa, Israel in 2000 both in Electrical Engineering and his MBA degree from the Interdisciplinary Center, Herzliya, Israel in 2007. He is currently pursuing the Ph.D. degree at the Engineering Faculty in Bar-Ilan University.

Mr. Dorfan has twenty years of research and development experience in signal processing and communications industries, and holds several patents. His research interests include distributed sensor networks, distributed acoustic source localization, blind source separation and distributed speaker 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 and 2014.

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 several 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.