Distributed Expectation-Maximization Algorithm for Speaker Localization in Reverberant Environments

Yuval Dorfan, Axel Plinge, Gershon Hazan, and Sharon Gannot

***Abstract*—Localization of acoustic sources has attracted a considerable amount of research attention in recent years. A major obstacle to achieving high localization accuracy is the presence of reverberation, the influence of which obviously increases with the number of active speakers in the room. Human hearing is capable of localizing acoustic sources even in extreme conditions.**

In this study, we propose to combine a method based on human hearing mechanisms and a modified incremental distributed expectation-maximization algorithm (IDEM).

Rather than using phase difference measurements that are modeled by a mixture of *complex-valued* Gaussians, as proposed in the original IDEM framework, we propose to use time difference of arrival (TDoA) measurements in multiple subbands and model them by a mixture of *real-valued* truncated Gaussians. Moreover, we propose to first filter the measurements in order to reduce the effect of the multi-path conditions. The proposed method is evaluated using both simulated data and real-life recordings.

***Index Terms*—Precedence effect; Onset dominance; Distributed expectation-maximization; Auditory scene analysis; Sound source localization; Spectral masking; Incremental expectation-maximization; Truncated Gaussian; multi-path; Time difference of arrival**

I. Introduction

The challenge of distributed localization in reverberant environment for concurrent speakers is addressed. A vast number of algorithms for localization of sources employing multiple concurrent sensor measurements have been developed [1], [2]. Important distinguishing aspects between methods are the sensor configuration, assumptions on the signal propagation, measurement type and probabilistic modeling.

Acoustic measurements by a spatially condensed array can only estimate the angle to the source [3]–[7]. In order to compute the sources' two- or three-dimensional coordinates, sensors have to be spatially distributed over a wider area. In spatially distributed approaches it is very common to use the term node for each local unit, which often has some computation power, sensors and communication capabilities. Such configurations, called acoustic sensor networks (ASNs), comprise of several nodes with one or more microphones. If these nodes are coupled by wireless links, communication constraints have to be taken into account.

The challenge of multiple speaker localization using generalized cross-correlation with phase transform (GCC-PHAT) is presented in [8]. The separation of concurrent speakers in time domain is limited when there are power differences and when their time difference of arrivals (TDoAs) are too close. Another way to apply GCC-PHAT for multiple sources is given in [9] for direction of arrival (DoA) estimation. A multiple speaker application with tight relations to GCC-PHAT is the blind source separation (BSS) problem [10] using a single array of microphones.

We can distinguish ASN approaches by the point in which calculations take place. Centralized solutions gather all measurements in a center point, where it is processed [11]–[13]. Hierarchical methods compute partial results in the nodes before combining them. In several hierarchal localization methods, the nodes provide DoAs estimates that are later combined by triangulation at a central processing point [14], [15].

In contrast, distributed algorithms carry out most of the calculations (or even all of them) in the nodes of the network. Some distributed approaches deal with the localization [16] or tracking [17] of a single source. In [17], Bayesian estimation based on a Kalman filter, applied to tracking, was described. Distributed localization is a relatively new field in the acoustic signal processing domain. Energy-based distributed localization approaches were described in [18]–[24].

In several algorithms, the nodes positions are assumed to be unknown [25]. Even if they are not measured beforehand, methods for automatic calibration of positions can be employed [26]–[28]. Recently, self-calibration methods that can be applied online were introduced [29], [30]. In this study, we assume perfect knowledge of the sensors' positions.

Acoustic signals often suffer from high level of reverberation that hamper localization accuracy and sometimes also from noise of various types. The main obstacles that hamper localization and DoA estimation accuracy are reverberation level [31], diffused noise [32] or sensor noise [33]–[35].

The next distinguishing criterion for localization algorithms is the assumption regarding strong unimpeded direct path signals. As many algorithms rely on a strong signal from the direct path, reverberations degrade the localization performance, especially in indoor scenarios [36]. This problem becomes more severe when the number of speakers increases, since each speaker produces additional reflections.

Several approaches are applicable without relying on the direct path from the target to the sensor [37]. In the method proposed in [38], direct-path dominance is not mandatory and various room shapes are supported. Methods that utilize the multi-path signals are limited by a set of assumptions about the room and might be more sensitive to changes in room

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

Axel Plinge is with the Department of Computer Science, Technische Universität Dortmund. He was supported by a fellowship within the FITweltweit programme of the German Academic Exchange Service (DAAD).

characteristics [39]–[44]. The set of techniques vary from un-supervised through semi-supervised to fully supervised. Our proposed approach is un-supervised, which uses multiple distributed sensor nodes and assumes that the direct path is dominant in at least part of them.

Next, we may also distinguish between algorithms according to the type of measurements used. A large number of acoustic localization and direction finding methods employ correlation between the microphone signals in order to determine the TDoA. If a node is equipped with more than a single microphone, the DoA can be estimated from TDoA values. Some approaches use full band measurements and some use subbands [45] or different frequencies.

A plethora of methods are based on the GCC-PHAT [46] or the steered response power with phase transform (SRP-PHAT) [12], [47], which are full-band TDoA estimation algorithms. Some algorithms compute subband TDoAs in order to improve the robustness and to facilitate concurrent speaker direction finding [48], [49].

Biologically inspired approaches use multiple frequency bands as well [50], [51]. The benefit of applying modeling of the sound processing in the human cochlea and mid-brain [52], [53] stems from the ability to include a number of mechanisms that render the estimation robust. The approaches summarized in [53] focus on the binaural case include generalized cross correlation (GCC) based methods. They present direction finding by subband GCC. This includes modeling of the precedence effect, which focuses on the first wavefronts from the direct path [54].

One of the consequences of the precedence effect is the neural onset dominance [55], which can be implemented in mono channel processing [56]. Methods that use phase locked spikes centered on the maxima in the certain frequency bands [57] can be advantageous [56] as compared to the more common zero crossing methods [50], [52].

More generally, the fact that only high signal to noise ratio (SNR) segments are used [58] can be exploited in tracking applications [59]. These principles were shown to add robustness in tracking with multiple microphone arrays [60] and were combined with auditory scene analysis (ASA)-oriented processing for application to distributed nodes [15].

The next distinguishing criterion between localization algorithms refers to the probabilistic modeling used. Many algorithms employ the maximum likelihood (ML) criterion, which is often implemented by applying the expectation-maximization (EM) procedure, to localize sources. This procedure is adopted for estimating the DoA [61], or the source position after triangulation [13], by utilizing the measured TDoAs.

A different approach is to model room positions by a mixture of Gaussians (MoG). This idea was first applied in [62] for binaural measurements. In [63], this has been applied for localization and tracking using ASN. The authors have defined a set of parameters and a set of hidden variables and developed a localization algorithm based on the EM and two tracking variants of the recursive expectation-maximization (REM) technique. A detailed mathematical description of the application of the EM to MoG can be found in [64]. The EM often faces a convergence problem. The random swap expectation-maximization (RSEM), based on MoG distribution [65], reduces the dependency of the EM algorithm on initial conditions.

In contrast to the methods that use a global MoG [63], [64], distributed localization versions [66], [67] define multiple local MoGs that share a set of weights, referred to as the global parameters. The hidden variables used in [66], [67] are local, enabling utilization of various network topologies. Although exhibiting good performance in a wide range of acoustic conditions, high reverberation still degrades performance. This may be attributed to the influence of multiple secondary reflections on the phase information embedded in the direct-path of the acoustic propagation [63], [66]–[68] or in the respective TDoA measurements.

We adopt the incremental distributed expectation-maximization (IDEM) methodology presented in [66], which was later adapted for BSS [69]. In the proposed method, we combine the power of machine hearing processing and that of the distributed MoG estimation. The bio-inspired pre-filtering eliminates a large percentage of the reflections. Consequently, the effect of reverberation on the localization accuracy is significantly reduced.

We chose in this contribution a different feature vector that necessitates a new probabilistic description. In the original model, a complex MoG is defined to describe the pair-wise relative phase ratio (PRP), which is a complex feature vector. Modeling the PRP as a complex-Gaussian is only an approximation, since magnitude 1 is assumed (see [67]). Here the feature vector is the real-valued TDoAs, hence a real-Gaussian can be used. Moreover, as the values of the TDoAs are confined to a physically plausible range, the truncated Gaussian can be a good model for the estimated TDoAs values. We use subband TDoA estimates, similarly to the model described in [70], and derive a modified version of the IDEM algorithm.

The new statistical model is described as a mixture of *truncated* Gaussians [71], which is a real-valued distribution consisting of random variables with a finite support. In [71], two alternative strategies were described for measurements confined to a finite range: *truncation* and *censor*. The truncation strategy classifies each value outside the allowed range as illegal, whereas the censor strategy substitutes these values with the closest value within the physical range. In our case, the direct path can only produce a TDoA within a confined range of values. We adopted the truncation strategy, since a TDoA reading that exceeds the permissible range, most probably does not contain any meaningful localization information. Unlike truncation, adopting the censor strategy may cause artifacts in certain positions in the room.

An additional modification applied to the original IDEM algorithm [66] is the ordering of the nodes. The ordering refers to the sequence of processing along the nodes of the network. Instead of a constant order, as used by the regular directed ring, we adopt here the concept of pseudo-random order, which is known to improve the convergence of the EM [65] and the robustness of the network to failures. In order to enable the new ordering we should replace the directed-ring

with a topology that contains a much reacher connectivity. For simplicity we use the fully connected topology.

The contribution of this paper is a combination of the bio-inspired pre-filtering, filter-bank approach, truncated model and the IDEM algorithm. It is shown that it produces good results for multiple speakers in reverberant rooms.

The resulting algorithm is also attractive from the network point of view, since, similarly to the method proposed in [66], it can be implemented without any central node. In addition, exact, sample-level, synchronization of the nodes is not required. In contrast to the model presented in [66], by virtue of its randomized order, it is capable of overcoming link failures.

The paper is organized as follows. In Section II, we formulate the localization problem. In Section III, the method used for solving the problem is described. This section explains the speech peak detection method, the calculation of TDoA readings, and their usage with the IDEM algorithm. Section IV is dedicated to simulations and experimental results. The summary and conclusions are given in Section V.

II. Problem formulation

Consider M microphone pairs receiving signals from J speakers. A noiseless environment is considered. The number of speakers is unknown in advance. Let $y_{m,i}(t)$ be the signals received by the ith microphone of the mth node. The signals in the time domain are given by

$$y_{m,i}(t) = \sum_{j=1}^{J} (g_{j,m,i}(t) * s_j(t) + r_{j,m,i}(t) * s_j(t)), \qquad (1)$$

where $t = 0, \ldots, T-1$ denotes the time index, j is the speaker index, and $s_j(t)$ denotes the speech signal produced by speaker j. The node index is $m = 1, \ldots, M$. At each node, we use $i = 1, 2$ as the index of each microphone.

The direct part of the room impulse response (RIR), $g_{j,m,i}(t)$, comprises an attenuation and a phase shift. The residual part of the RIR, $r_{j,m,i}(t)$, consists of all the reflections, namely the multi-path components. As the reverberation of each speaker is independent, the total power increases with the number of speakers, J.

In previously proposed localization algorithms [66], [67], only the direct-path is utilized for the estimation. These algorithms tend to work well, although they did not take the other reflections into account. In contrast, the method we propose in this paper tries to reduce the late reflections by the construction of the feature vector.

In the presence of additive noise we can add a noise reduction algorithm similarly to [34]. Although the focus is on reverberation, additive noise influence is discussed in the experimental section as well.

The network used here has a pseudo-random topology, unlike [66], i.e. the connectivity between the nodes changes between iterations. We adopted this topology for two reasons. First, changing the order of the nodes may allow faster and more accurate convergence of the IDEM algorithm [65], [72]. Second, it is less prone to communication failures, since each node has more than one link, which can be used for initiating the next sub-iteration. If a specific link has a failure, the network can overcome it by skipping a specific iteration, which contains this failure.

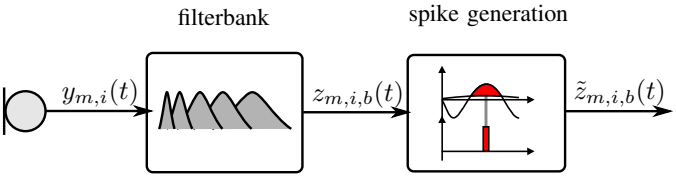


Figure 1. Signal preconditioning stage. The microphones' signals are each split into multiple bands by a gammatone filterbank and then transformed into spike trains.

III. The proposed method

The proposed distributed method comprises three major steps. In the first step, a bio-inspired model is applied to the microphone signals for reducing the effect of reverberation. In the second step, cross-correlation is applied to estimate the TDoAs. The final step, which is based on the IDEM, is derived from the ML model in order to find the number of speakers and their locations. These steps and the statistical model of the measurements are described in this section.

A. Cochlear model

The speech peak detection method applied as a preprocessing procedure for each microphone independently is described in this subsection. The efficient on-line cochlear model introduced in [56] is used to generate a sparse spike representation of the microphone signals. It consists of a filter bank modeled on the basilar membrane and a model of the cochlear nucleus generating spikes from the band filtered signals. The preconditioning is illustrated in Fig. 1. It is tailored to facilitate localization in reverberant environment from three aspects: first, only highly modulated parts of the signal are used; second, phase-locked spikes are generated; and third, echo suppression is achieved by modeling the neural saturation (cf. [55]).

A common approach in computational auditory scene analysis (CASA) is to use an infinite impulse response (IIR) filter bank to model the basilar membrane [50]. In our approach, a fast Fourier transform (FFT) filter bank is used. The filters are defined in the spectral domain using a gammatone [4] approximation [73]:

$$\hat{G}_b(f) = \left(1 + \frac{\imath(f - f_b)}{w \cdot w_b}\right)^{-4}, \qquad (2)$$

where $\imath$ is the imaginary unit and $b = 1, \ldots, B$ denotes the frequency band index. The total number of bands is $B = 16$. The bands are defined to imitate the critical bands found in human hearing [50], [51].

The nonlinear spacing of the bands with center frequencies ($f_b = f_1, \ldots, f_B$) between 0.3 kHz and 3 kHz is chosen as equidistant on the quasi-logarithmic equivalent rectangular bandwidth (ERB) scale [51], [74]. The Glasberg-Moore bandwidth [74], w_b, is widened by a factor w in order to increase frequency content.

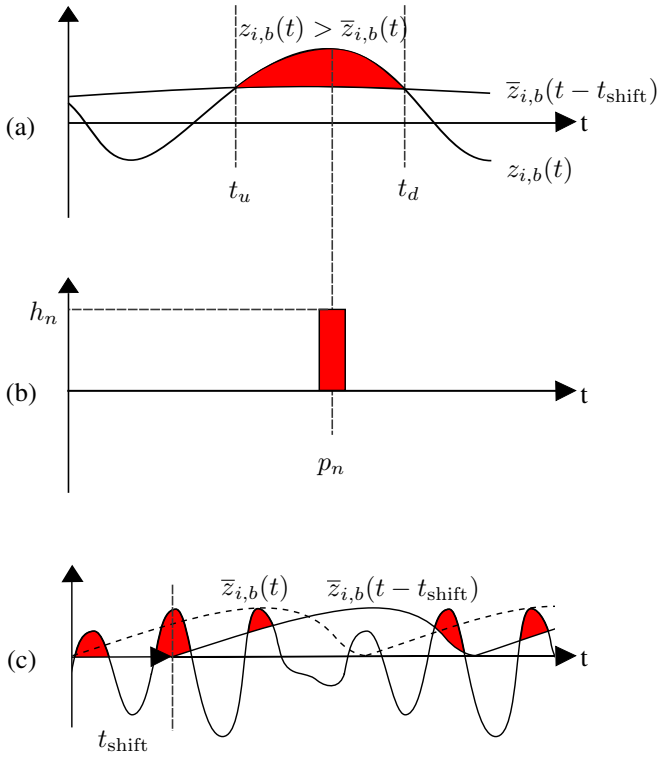


Figure 2. Spike generation procedure. (a) The band signal is compared to its short time average. (b) When the area above the average exceeds the signal by 6 dB, a spike of height h_n is generated at the maximum position p_n. (c) By shifting the average in time, the first wavefronts are enhanced.

For each microphone signal $y_{m,i}(t)$, B signals $z_{m,i,b}(t)$ are synthesized in the time domain. In each band, the signal is transformed into a sparse spike representation. The process of computing the spikes is illustrated in Fig. 2. One of the key concepts is to model the neural saturation in order to detect the modulation and imitate monaural echo suppression. This is achieved by a halfway rectification and a peak over average comparison. Thus, for each band signal $z_{m,i,b}(t)$ is halfway rectified $\max(0, z_{m,i,b}(t))$ before a moving average $\overline{z}_{m,i,b}(t)$ is computed. The averages are calculated over an intermediate interval of, e.g. 30 msec to encompass the pitch modulation. This delay is a non-issue for this off-line static case. If applied to dynamic on-line applications, the latency might be still reasonable, since indoor movements of speakers is not expected to be too fast.

Modulated intervals $U_n = [t_u, t_d]$ are found as time segments, where the signal exceeds the shifted average. Its maximum position in time, p_n is a candidate position for a spike. If the peak is 6 dB higher than the shifted average, an impulse spike is generated at p_n in the output signal $\tilde{z}_{m,i,b}$.

The pulse height h_n of the spike corresponds to the amplitude computed by square root compression:

$$h_n = \sum_{t=t_u}^{t_d} \sqrt{z_{m,i,b}(t) - \overline{z}_{m,i,b}(t - t_{\text{shift}})}, \tag{3}$$

where t_{shift} is a time shift used to enhance the first wavefronts as illustrated in Fig. 2. The pulse width is set to 20μsec to imitate the temporal resolution of human neural processing.

Algorithm 1: Cochlear model processing

initialize $\tilde{z}_{m,i,b}(t) = 0$
calculate moving average $\overline{z}_{m,i,b}(t)$
find modulated intervals $U_n = [t_u, t_d]$
  where $z_{m,i,b}(t) > \overline{z}_{m,i,b}(t - t_{\text{shift}}) \quad \forall t_u < t < t_d$
foreach U_n **do**
  $p_n = \arg\max_{t_u < t < t_d} z_{m,i,b}(t) - \overline{z}_{m,i,b}(t - t_{\text{shift}})$
  if $z_{m,i,b}(p_n) > 2\overline{z}_{m,i,b}(p_n - t_{\text{shift}})$ *(6 dB)* **then**
    Compute
    $h_n = \sum_{t=t_u}^{t_d} \sqrt{z_{m,i,b}(t) - \overline{z}_{m,i,b}(t - t_{\text{shift}})}$
    Insert h_n with width 20 μsec at p_n in $\tilde{z}_{m,i,b}$
  end
end

This nonlinear processing is summarized in Algorithm 1. The resulting signal of that processing is sparse. Although only few bands are used, it still complies with the model of sparse representation of the signal, which will be required for the EM algorithm.

The signals $\tilde{z}_{m,i,b}(t)$ of the microphones of all nodes are used for calculating the features, the multiple pairwise TDoAs (MPTs). The MPTs, denoted by $\tau_{m,b}(\mathbf{p})$, are used as feature vectors in the ML. The model of these features is discussed in the next subsection.

B. Multiple pairwise time difference of arrival

In order to calculate the MPT, $\tau_{m,b}(\mathbf{p})$ we apply the following steps at each node to $\tilde{z}_{m,i,b}(t)$: 1) Calculate the cross-correlation of the two microphones. 2) Find the peak and compare it to a threshold. 3) If it passes the threshold, calculate $\tau_{m,b}(\mathbf{p})$, the exact (interpolated) time difference of the peak.

In contrast to the PRP used in [66], the MPTs are real-valued quantities. The speakers are assumed to exhibit disjoint activity in the time-frequency domain [63], [75]–[77]. Therefore, every time-frequency band can be associated with at most a single active position (or speaker). Here we claim that the sparsity assumption is still valid although our frequency resolution is poorer. The validity of the sparseness assumption may be attributed to the bio-inspired reverberation reduction mechanism.

It is assumed that the speakers can be located in a final set of positions on a grid with a desired resolution. The time differences can be calculated in advance for every candidate location $\mathbf{p}$ on that grid:

$$\tilde{\tau}_{m,b}(\mathbf{p}) \triangleq \frac{||\mathbf{p} - \mathbf{p}_m^1|| - ||\mathbf{p} - \mathbf{p}_m^2||}{c}, \tag{4}$$

where $\mathbf{p}_m^1$ and $\mathbf{p}_m^2$ are the locations of the microphones, assumed to be known, $||\cdot||$ denotes the Euclidean norm, and c is the sound velocity. The set $\mathbf{p} \in \mathcal{P}$ contains the grid of points.

In the following subsection, we describe the statistical model of the MPTs. Since they have different physical properties than the PRPs, a new statistical model is presented.

C. Statistical model

We assume that the MPTs are random observations drawn from a mixture of truncated normal distributions:

$$\tau_{m,b}(t) \sim \sum_{\mathbf{p}} \psi_{\mathbf{p}} g\left(\tau_{m,b}(t); \tilde{\tau}_{m,b}(\mathbf{p}), \sigma^2, F, \tau_{\max}\right), \quad (5)$$

where $\psi_{\mathbf{p}}$ is the probability of a speaker being at position $\mathbf{p}$.

The probability density function (p.d.f.), $g(\cdot;\cdot,\cdot,\cdot,\cdot)$ is the truncated Gaussian probability [71]:

$$g\left(x;\mu,\sigma^2,F,\tau_{\max}\right) = \begin{cases} \frac{\mathcal{N}(x;\mu,\sigma^2)}{\int_{x_L}^{x_H} \mathcal{N}(x;\mu,\sigma^2)}, & x \in [x_L, x_H] \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

$$x_L = \max(\mu - F\sigma, -\tau_{\max})$$
$$x_H = \min(\mu + F\sigma, \tau_{\max})$$

where

$$\mathcal{N}\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\cdot\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\cdot\sigma^2}\right) \quad (7)$$

is the normal distribution.

The maximal time difference, $\tau_{\max}$ is used to limit the truncated supports. It is the maximal time difference that can be produced by sound traveling between the microphones at each node. We denote τ_{max} the maximal physical delay.

The factor of the truncation F is empirically determined in the experimental section. The variance σ^2 is used to control errors in estimating the MPTs due to noise and reverberation. It also supports off-grid source positions.

The truncation is applied according to the lower and higher support limits, x_L and x_H, respectively. The truncated Gaussian distribution is used here rather than the Gaussian used in [68], since it allows irrelevant measurements to get zero probability, hence better fitting the physical model. The truncation for grid positions near the end-fire is asymmetric due to the maximal physical delay.

We state now the ML estimation procedure of the localization parameters from a given set of local measurements. As in [66], the parameters to be estimated are the weights, $\psi_{\mathbf{p}}$, which are the probabilities of finding acoustic sources in each position, $\mathbf{p}$.

The other parameters are assumed to be known. The variance σ^2 was manually tuned to 6 [samples^2]. The variance could be assumed unknown, as often done, and its estimation can be easily incorporated in the EM iteration [64]. It is kept fixed, since we have not observed any significant advantage of adding it to the parameter estimation in this particular algorithm. Further variance analysis is presented in the experimental section.

The vector of the unknown parameters is defined as

$$\boldsymbol{\theta} \triangleq \text{vec}_{\mathbf{p}}(\psi_{\mathbf{p}}). \quad (8)$$

Note that $\psi_{\mathbf{p}}$ has a probability interpretation, namely $\sum_{\mathbf{p}} \psi_{\mathbf{p}} = 1$. The goal of the algorithm is to estimate these parameters. Then we will apply a threshold in order to estimate the number of active speakers (also referred to as speaker detection) and their locations. These parameters are global, meaning common to all nodes.

Augmenting all MPT readings in $\boldsymbol{\tau} = \text{vec}_{m,t,b}\left(\tau_{m,b}(t)\right)$ and following [67], the p.d.f. of the observation set can be written as

$$f(\mathbf{T} = \boldsymbol{\tau}; \boldsymbol{\theta}) = \prod_{m,t,b} \sum_{\mathbf{p}} \psi_{\mathbf{p}} g\left(\tau_{m,b}(t); \tilde{\tau}_{m,b}(\mathbf{p}), \sigma^2, F, \tau_{\max}\right). \quad (9)$$

The maximum likelihood estimator (MLE) problem can be stated as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}} [\log f(\mathbf{T} = \boldsymbol{\tau}; \boldsymbol{\theta})]. \quad (10)$$

A straightforward solution of the ML is centralized and of high complexity. This type of problems is often solved by the EM in order to reduce its complexity.

As already mentioned above, we are interested in a distributed solution of this estimation problem. The global parameters will be estimated jointly by the nodes of the network. Several algorithms can be suggested for this problem. In this contribution, we are interested mainly in the application of the IDEM algorithm [66] to the MPTs.

D. An incremental distributed expectation-maximization with pseudo-random topology

For complexity reduction of the ML, EM and REM techniques are often used. For dynamic cases, REM is used. For distributed applications, distributed expectation-maximization (DEM) algorithms, which were first suggested in [78], can be used for localization and tracking of acoustic speakers. This section focuses on an IDEM algorithm for the static case.

This subsection is divided into two parts. The first deals with the choice of the hidden variables of the DEM algorithms that can be used for the problem at hand. Here, we focus on one of these algorithms, the IDEM [66], which is presented in the second part. The IDEM processes the measurements and the hidden variables incrementally through the network.

It is shown that for updating the minimal sufficient statistics (MSS), only the recent parameter estimation is needed. This estimation is very compact and hence requires less communication bandwidth (BW).

1) Local hidden variables: The global parameters can be estimated by applying EM techniques using global hidden variables [63] or local hidden variables [66].

Synchronization for acoustic network has been dealt in [79], [80]. Rough synchronization of nodes is maintained in our case by the communication link. We also assume that the signals acquired by the microphones of the same node are fully synchronized, since the accuracy of the MPTs depends on it. A much looser assumption is made regarding the signals of different nodes [70].

For DEM algorithms, it was suggested that local hidden variables be defined [66]. The hidden variables are defined to be the indicators

$$y_m(t,b,\mathbf{p}) = \begin{cases} 1, & \mathbf{p} \text{ active for } (m,t,b) \\ 0, & \text{otherwise} \end{cases}. \quad (11)$$

For each time-frequency band (t,b) they equal to zero everywhere except to $\mathbf{p}$, the position of the active speaker, since not more than one speaker is likely to be active at each time-frequency band [75].

The total number of indicators in the problem is $T\times B\times M$. Each of these is defined over $|\mathcal{P}|$ possible values, with $|\mathcal{P}|$ the cardinality of the set of all possible grid positions. Please note that, in contrast to methods where global hidden variables are used [63], local hidden data support a case where some of the nodes measurements are physically unfeasible and hence assigned with zero probability.

Let $\mathbf{y}=\text{vec}_{t,b,m,\mathbf{p}}\left(y_m(t,b,\mathbf{p})\right)$ be the set of all local indicators. The expectation of the indicator is given by

$$E\left\{y_m(t,b,\mathbf{p})\right\}=\psi_{\mathbf{p}}. \tag{12}$$

Under the W-disjoint assumption [75], namely that each observation can be associated with only a single speaker (and hence a single position) and under the static model assumption, the probability density function of $\mathbf{y}$ is given by

$$f(\mathbf{Y}=\mathbf{y};\boldsymbol{\theta})=\prod_{m,t,b}\sum_{\mathbf{p}}\psi_{\mathbf{p}}y_m(t,b,\mathbf{p}). \tag{13}$$

Given the local hidden variables, the p.d.f. of the observations is

$$f(\mathbf{T}=\boldsymbol{\tau}|\mathbf{y};\boldsymbol{\theta})= \tag{14}$$
$$\prod_{m,t,b}\sum_{\mathbf{p}}y_m\left(t,b,\mathbf{p}\right)g\left(\tau_{m,b}(t);\tilde{\tau}_{m,b}(\mathbf{p}),\sigma^2,F,\tau_{\max}\right).$$

The p.d.f. of the complete data can be deduced from (13)-(14) and from some simplifications due to the properties of the indicator $y_m\left(t,b,\mathbf{p}\right)$:

$$\begin{aligned}&f(\mathbf{T}=\boldsymbol{\tau},\mathbf{Y}=\mathbf{y};\boldsymbol{\theta})=f(\mathbf{Y}=\mathbf{y};\boldsymbol{\theta})f(\mathbf{T}=\boldsymbol{\tau}|\mathbf{y};\boldsymbol{\theta})\\&=\prod_{m,t,b}\sum_{\mathbf{p}}\psi_{\mathbf{p}}y_m\left(t,b,\mathbf{p}\right)\cdot\\&g\left(\tau_{m,b}(t);\tilde{\tau}_{m,b}(\mathbf{p}),\sigma^2,F,\tau_{\max}\right).\end{aligned} \tag{15}$$

A family of distributed algorithms can be derived using these local hidden variables. The algorithm simulated in [66], [69] was implemented over a simple fixed directed ring. In [67], we suggested a batch distributed expectation-maximization (BDEM) over a bi-directional tree. In the subsequent part of this subsection, we derive a version of the IDEM for the case of MPTs measurements, which is implemented over a pseudo randomly-ordered topology.

2) Random order incremental distributed expectation-maximization: The gist of this algorithm is to run the EM with local computations in a pseudo-random order of executing the local *E-step* and updating the localization parameters. Each such local *E-step* is referred to as a sub-iteration.

In addition, the *M-step* is applied after each such sub-iteration locally according to the incremental equation, which updates only the local contribution to the parameters estimation. The incremental principle was first introduced in [81].

The idea is that after the initialization round, the algorithm has two types of iterations. The global iterations are defined as those introduced for every batch EM algorithm. However, for the IDEM we also define a sub-iteration index, which is

$$i=M*(\ell-1)+n(m,\ell). \tag{16}$$

The order index function, $n(m,\ell)$ assigns a unique integer (between 1 and M) to every node m, which determines the order in which the M nodes are updated in the global iteration ℓ. This sub-iteration is local.

As in the study reported in [66], each node in its turn obtains the recent parameters' estimation from the previous node and calculates its local (partial) *E-step* in order to update its local hidden variables. Then, it can apply a new *M-step* and transmit the result to the next node.

In contrast to the method proposed in [66], which uses directed-ring topology, the current scheme determines a random update order in each iteration. Reliable updates in each iteration can be guaranteed if an identical pseudo-random generator is applied in all nodes.

This random order has two major advantages. The first arises from the optimization aspect. The incremental expectation-maximization (IEM) is known to produce faster convergence and more accurate estimation in the case of a pseudo random order of measurements processing. The second advantage arises from the network considerations. For example, when the update order is pseudo random, easy ways to circumvent a link failure exist. The idea is that if a single (but permanent) link failure occurs, a directed-ring topology renders useless. Changing the order after each iteration enables recovery from such a link failure. It should be noted that such a strategy of changing the order depends on the network topology, which in our case assumed to be fully-connected.

As previously mentioned, each node is responsible for its local measurements and local hidden variables, but in contrast to the BDEM method that executes the *E-step* simultaneously at all nodes for the same parameter estimation, our proposed method executes the *E-step* and the *M-step* incrementally.

The *E-step* is carried out in a distributed manner for each node on its turn (sub-index i):

$$\begin{aligned}\upsilon_m^{(i)}\left(t,b,\mathbf{p}\right)&\triangleq E\left\{y_m\left(t,b,\mathbf{p}\right)|\tau_{m,b}(t);\hat{\boldsymbol{\theta}}^{(i-1)}\right\}\\&=\frac{\hat{\psi}_{\mathbf{p}}^{(i-1)}g\left(\tau_{m,b}(t);\tilde{\tau}_{m,b}(\mathbf{p}),\sigma^2,F,\tau_{\max}\right)}{\sum_{\tilde{\mathbf{p}}}\hat{\psi}_{\tilde{\mathbf{p}}}^{(i-1)}g\left(\tau_{m,b}(t);\tilde{\tau}_{m,b}(\tilde{\mathbf{p}}),\sigma^2,F,\tau_{\max}\right)}.\end{aligned} \tag{17}$$

The algorithm is based on the IEM [81]. The authors proved that, not only the classical (batch) EM converges, but also other modifications of the EM. For example partial *E-step* can be followed by the regular *M-step*. In many cases the partial version is even better in terms of convergence speed and accuracy. The IEM algorithm can be applied in the distributed case to local hidden data.

The network topology can be used for incremental updates of the hidden variables and the parameter estimation by transmitting only the recent parameters' values.

Define the local summation of the hidden variables as

$$\bar{\upsilon}_m^{(i)}\left(\mathbf{p}\right)\triangleq\sum_{t,b}\upsilon_m^{(i)}\left(t,b,\mathbf{p}\right). \tag{18}$$

Algorithm 2: IDEM Speaker localization.

Obtain the valid local measurements $\tau_{m,b}(t)$.
initialize $\bar{v}_m^{(0)}(\mathbf{p})$ and $\psi^{(0)}(\mathbf{p})$.
for $\ell = 1$ **to** L **do**
  Randomly choose the order $n(m,\ell)$.
  for $n(m,\ell) = 1$ **to** M **do**
    Calculate the current sub-iteration index: $i = M*(\ell-1)+n(m,\ell)$.
    Calculate the current node index: $m = n^{-1}(m,\ell)$.
    E-step
    Calculate locally $v_m^{(i)}(t,b,\mathbf{p})$ (17).
    Aggregate locally $\bar{v}_m^{(i)}$ (18).
    M-step
    Calculate the global parameters, $\hat{\psi}_{\mathbf{p}}^{(i)}$ (19).
    Transmit the global parameters to the next node.
  end
end
Find $\hat{J}$, the number of speakers, and their respective locations $\mathbf{p}_j$; $j = 1,\ldots,\hat{J}$ by applying a threshold to $\psi_{\mathbf{p}}^{(LM)}$, which is the result of the last *M-step*.

The *M-step* of the BDEM was shown to be the average of all the hidden variables [67]. The *M-step* of the IDEM in this case, as for many other IEM algorithms, simplifies to an update of the local contribution of node m to the global parameters:

$$\hat{\psi}_{\mathbf{p}}^{(i)} = \hat{\psi}_{\mathbf{p}}^{(i-1)} + \frac{\bar{v}_m^{(i)}(\mathbf{p}) - \bar{v}_m^{(M*(\ell-2)+n(m,\ell-1))}(\mathbf{p})}{T \cdot B \cdot M}. \quad (19)$$

The new estimation, $\bar{v}_m^{(i)}(\mathbf{p})$ is added while the previous local contribution, $\bar{v}_m^{(M*(\ell-2)+n(m,\ell-1))}(\mathbf{p})$ is subtracted. Each node of the network uses its local memory to store this local contribution in order to use it in the next iteration. The denominator contains a normalization factor.

The initialization of the global parameters is obtained by a preliminary global iteration:

$$\hat{\psi}_{\mathbf{p}}^{(0)} = \frac{\sum_m \bar{v}_m^{(0)}(\mathbf{p})}{T \cdot B \cdot M}, \quad (20)$$

where the initial hidden data estimates, $\bar{v}_m^{(0)}(\mathbf{p})$ are calculated at each node and aggregated through the network.

After the last iteration a threshold is applied to the latest estimation of the parameters $\psi_{\mathbf{p}}^{(LM)}$. Each position above the threshold is declared as a speaker location. This way we estimate the number of speakers and their locations.

This concludes the description of the proposed algorithm. Algorithm 2 summarizes the proposed IDEM approach.

IV. Experimental Study

This section is dedicated to an experimental study of the localization algorithms. We use the proposed algorithm in various versions and two other algorithms for comparison. The first, denoted IDEM2014 is described in [66]. The second, is the well-known SRP-PHAT algorithm [47]. This is an approach that searches the position that maximizes the output of a steered delay and sum beam former (BF).

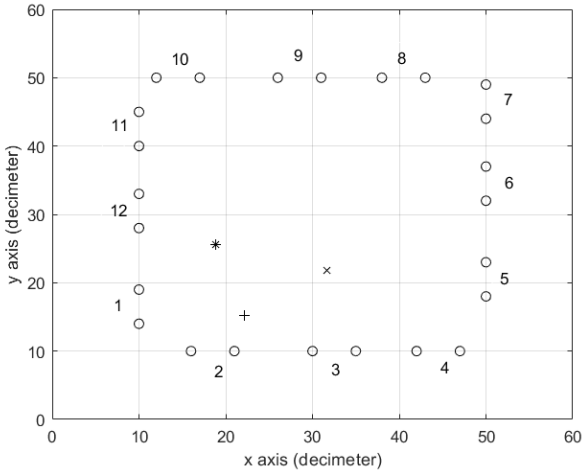


Figure 3. Room setup with twelve microphone pairs (circles) with numbers and three speakers ($*$,x,$+$). Grid of 60×60 positions.

The section is organized as follows. The first subsection presents the setup of both the simulated room and the acoustic laboratory. The parameter settings of the algorithms and complexity analysis are described in the second subsection. The third subsection presents the results of the simulation analysis. The forth subsection examines different components of the proposed algorithm. The last subsection discusses the real recording results.

A. Simulation and acoustic room setup

This subsection contains two parts: simulated room and real room descriptions.

1) Simulated room: We simulated 100 random positions of 1, 2, or 3 concurrent speakers in a reverberant room ($6 \times 6 \times 2.4$ m) with T_{60} in the range of 150 msec to 600 msec, using the image method [82]. The time duration of each localization experiment was 12sec. Note that the grid resolution, which was 10×10 cm, is sufficiently good taking into account the volume of a real speaker, which is not a point source.

The microphone pairs are placed at an inter-distance of 50 cm, which was determined to offer a good compromise between resolution and ambiguity [66]. A few simulations were also performed for 10 cm as explained below. The speakers and microphones are assumed to be located in a 2D plane at a height of 135 cm in order to examine the performance when the speakers has a typical mouth height (accuracy of height is not critical).

An example for such a room setup is shown in Fig. 3. The microphone pairs are numbered and marked with circles. The three speakers are marked with a $*$, a x and a $+$.

2) Real room: The room measures $6 \times 6 \times 2.4$ m and has a controllable reverberation time (T_{60}) in the range of 100 msec to 1100 msec. Also here the microphone pairs are placed at an inter-distance of 50 cm. An example of the room setup for one of the room recordings is shown in Fig. 4. The microphone pairs are numbered and marked with circles. The two speakers are marked with a $*$ and a $+$. For the real recordings, two

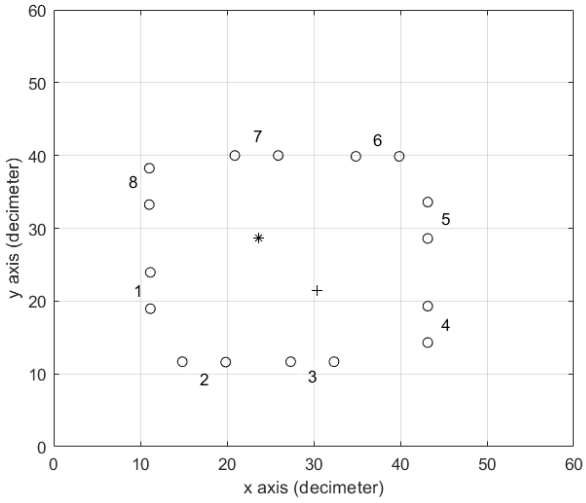


Figure 4. Room setup with eight microphone pairs (circles) with numbers and two speakers (∗,+). Grid of 60×60 positions.

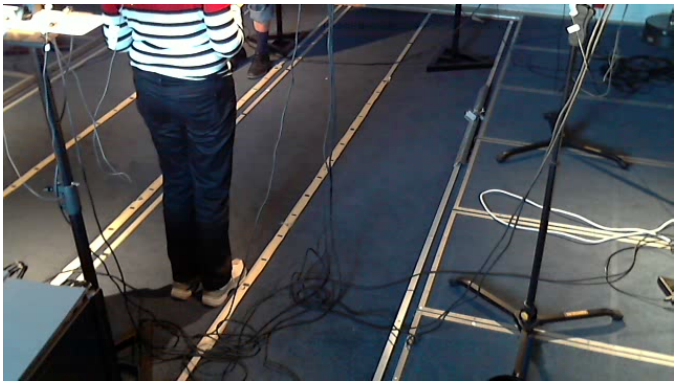


Figure 5. First scenario: low reverberation level($T_{60} = 200$ msec); two speakers with a large distance between them.

reverberation setups were tested. We recorded two concurrent speakers using eight microphone pairs. The speakers uttered English sentences while standing in the room. The sampling frequency was set to 48 KHz.

The first speakers were recorded with a low reverberation level ($T_{60} = 200$ msec) and a relatively large distance between them. A snapshot from this recording is depicted in Fig. 5.

In order to test a considerably more challenging scenario, the other speakers were recorded in the same room with a much higher reverberation time, $T_{60} = 930$ msec, and with a smaller distance between them. A snapshot from that recording is depicted in Fig. 6. The changes in the reflection nature of the floor can be seen in the pictures. Other facets of the room (ceiling and walls) were changed from absorbing to reflective as well.

B. Parameter setting and complexity of the proposed algorithm

A few practical issues must be addressed regarding the proposed algorithm. They are listed below and refer to all experimental results.

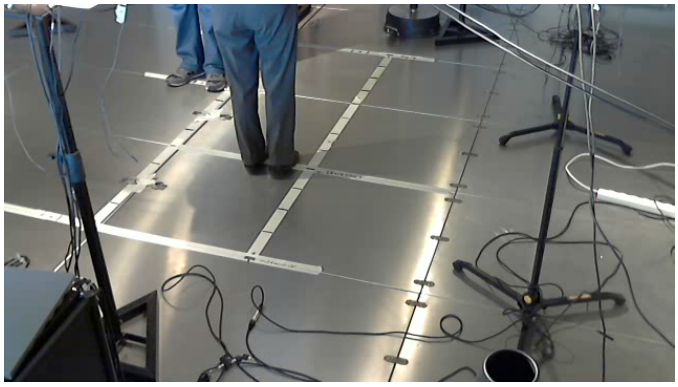


Figure 6. Second scenario: high reverberation level ($T_{60} = 930$ msec); two speakers standing close to each other.

1) Tuning onset algorithm: The cochlear model provided robust features in previous applications [56]. Some of its parameters were determined experimentally; for example, $w = 6.0$, which is used to scale the bandwidth of the filters. As mentioned above, the number of bands used was $B = 16$. We experimentally chose it, observing significant degradation for lower values in the case of multiple concurrent speakers. This may be attributed to the invalidity of the sparseness (w-disjoint orthogonality) for the lower resolution.

2) Beam width of a microphone pair: A typical problem, which arises using microphone pairs, is the reduced resolution in the planar domain in the end-fire directions. The case dealt here is 2D with both pairs and speakers on the same plane. It means that broadside measurements have a better resolution. As a large number of distributed nodes were used, many nodes would provide accurate MPTs measurements corresponding to near broadside positions.

One of the advantages of using local hidden variables rather than global hidden variables is that the consensus might be achieved even if only part of the nodes reliably measure a specific speaker.

3) Calculation precision: When summing small numbers and numbers close to 1, precision problems should be addressed. For this reason, a natural log is applied to some of the equations to alleviate dynamic range problems that may arise even for double precision calculations [83], [84]. It is first applied to the Gaussian equation (6):

$$\log\left(\mathcal{N}\left(x;\mu,\sigma^2\right)\right) = -\frac{1}{2}\log(2\pi\cdot\sigma^2) - \frac{(x-\mu)^2}{2\cdot\sigma^2} \qquad (21)$$

The same operation has been applied to the EM equations (17)-(20). For example, the log version of equation (17) is:

$$\begin{aligned}&\log \upsilon_m^{(i)}\left(t,b,\mathbf{p}\right) = \\ &\log(\hat{\psi}_{\mathbf{p}}^{(i-1)}) + \log\left(g\left(\tau_{m,b}(t);\tilde{\tau}_{m,b}(\tilde{\mathbf{p}}),\sigma^2,F,\tau_{\max}\right)\right) \\ &-\log\left(\sum_{\tilde{\mathbf{p}}}\hat{\psi}_{\tilde{\mathbf{p}}}^{(i-1)}g\left(\tau_{m,b}(t);\tilde{\tau}_{m,b}(\tilde{\tilde{\mathbf{p}}}),\sigma^2,F,\tau_{\max}\right)\right),\end{aligned} \qquad (22)$$

Table I
THE PROPOSED ALGORITHM IS ANALYZED WITH RESPECT TO COMPUTATION, COMMUNICATION BANDWIDTH AND MEMORY.

Criteria	Rough estimation
Computation	$\mathcal{O}(L \cdot T \cdot B \cdot \lvert P \rvert)$
BW	$\mathcal{O}(L \cdot \lvert P \rvert)$
Memory	$\mathcal{O}(T \cdot B + \lvert P \rvert)$

where log of sum of numbers is calculated by applying the following equation:

$$\log(\sum_{i=1}^{I}(e^{v_i})) = v_{max} + \log \sum_{i=1}^{I} e^{v_i - v_{max}}. \quad (23)$$

For probabilities calculations e_i^v are between 0 and 1. It means that v_i are negative real numbers, with v_{max} their maximum.

In addition, multiplication and division operations are substituted by summation and subtraction operations and small numbers can be stored together with much larger ones (within the same vector).

4) Truncated Gaussian: Another computational issue is the Gaussians truncation, which is applied according to a $\pm F\sigma$ and the physical valid range, where $F = 3$ was empirically tuned. It is notable again that this truncation is applied around each grid point. In addition to what mentioned above, the truncation has a computational advantage, since extremely low probability values are neglected.

5) Complexity analysis: Complexity analysis is briefly summarized. When we deal with DEM algorithms, we have to examine several aspects. Some of them can be taken from the complexity analysis in [67]:

- We are interested in the number of calculations. Some of them are done once, before the iterations starts. For example the MPTs calculation. Others are applied every iteration. Since all calculations are applied in the nodes and they are all identical, we will analyze the complexity per node. The most significant operations are multiplications and additions, when assuming usage of lookup table implementation of operations like log.
- Communication BW is addressed from a single node perspective, since the process is serial. At each sub-iteration the map of probabilities should be transmitted from one node to the next. As the algorithm converges, this map becomes sparser, since the majority of locations are not active.
- Memory usage per node relates to the local measurements and the latest contribution to the global parameters that should be subtracted in the next iteration.

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

C. Simulation results

We carried out 100 Monte Carlo trials and calculated three statistical measures: 1) the miss-detection (MD) rate, defined as the percentage of speakers that were miss-detected out of the total number of actual speakers; 2) the false alarm (FA) rate, defined as the percentage of falsely-detected speakers normalized by the total number of real speakers; and 3) the root mean square error (RMSE), defined as the accuracy of localization for all successfully detected speakers.

Table II
LOCALIZATION STATISTICS FOR 100 MONTE CARLO TRIALS (INTER DISTANCE 50 CM). THE FIRST COLUMN INDICATES THE REVERBERATION LEVEL. THE SECOND ONE CONTAINS THE NUMBER OF SIMULATED SPEAKERS THAT WERE RANDOMLY LOCATED.

T_{60}	Sim. speakers	Algorithm	MD[%]	FA[%]	RMSE[cm]
600	**1**	**Proposed**	0.0	6.0	4
600	**1**	**IDEM2014**	10.0	63.0	99
600	**1**	**SRP-PHAT**	95.0	1.0	206
400	1	Proposed	0.0	4.0	4
400	1	IDEM2014	6.0	73.0	99
400	1	SRP-PHAT	95.0	1.0	157
200	1	Proposed	0.0	3.0	4
200	1	IDEM2014	41.0	14.0	175
200	1	SRP-PHAT	91.0	1.0	153
600	2	Proposed	31.5	16.5	12
600	2	IDEM2014	88.5	3.0	116
600	2	SRP-PHAT	99.0	1.0	101
400	**2**	**Proposed**	8.0	9.0	4
400	**2**	**IDEM2014**	35.5	54.0	73
400	**2**	**SRP-PHAT**	97.0	0.5	138
200	2	Proposed	0.0	8.5	4
200	2	IDEM2014	16.0	67.0	82
200	2	SRP-PHAT	75.0	8.0	149
400	3	Proposed	21.3	34.3	10
400	3	IDEM2014	56.7	22.7	82
400	3	SRP-PHAT	96.3	1.3	103
200	**3**	**Proposed**	2.3	4.0	4
200	**3**	**IDEM2014**	41.0	35.3	97
200	**3**	**SRP-PHAT**	97.0	0.3	126

Detection, false- and miss-detections are defined with respect to the threshold applied. The statistical analysis matches each ground truth speaker position to the closest candidate (if any). The RMSE measure is calculated from all the matched positions. If a ground truth speaker does not have a matching detection, it is counted as a miss-detection. We count extra detections as false alarms.

We present here results of various reverberation levels for different speakers locations after Monte Carlo (MC) averaging. In addition, we identified for each number of speakers the level of reverberation for which the proposed algorithm performance is still good and marked them in **bold**. Table II summarizes the measures for all simulated cases.

Inspecting the results in Table II, it can be deduced that as the reverberation level increases from $T_{60} = 200$ msec to $T_{60} = 600$ msec, the number of speakers allowing reliable localization by the proposed algorithm decreases from three to one, respectively. This can be attributed to the increased density of reflections due to the co-existence of multiple concurrent speakers in the same environment. In any case, the proposed algorithm significantly outperforms the reference algorithms [47], [66] in terms of detection, as well as RMSE. The obtained RMSE values are better than the designated grid

Table III
COMPARISON WITH AND WITHOUT TRUNCATION (INTER DISTANCE 50 CM): 100 MONTE CARLO TRIALS OF TWO RANDOMLY LOCATED SIMULATED SPEAKERS.

T_{60}	Sim. speakers	Algorithm	MD[%]	FA[%]	RMSE[cm]
150	2	Proposed	0.0	5.0	4
150	2	No-Trunc	0.0	100.0	4

Table IV
VARIANCE INFLUENCE(INTER DISTANCE 50 CM): 100 MONTE CARLO TRIALS WITH TWO RANDOMLY LOCATED SIMULATED SPEAKERS

T_{60}	Sim. speakers	σ^2[samp^2]	MD[%]	FA[%]	RMSE[cm]
400	2	16	27.0	3.0	4
400	**2**	**6**	**8.0**	**9.0**	**4**
400	2	1	1.0	50.0	4
400	2	0.01	0.0	100.0	5

resolution. Note that when either MD or FA rates are very high (above 50 percents), the exact statistics is less relevant, since it indicates that controlling the detection threshold did not yield any acceptable FA-MD combination.

D. Experimental insights

We carried out a few additional Monte Carlo simulations to gain insights about the proposed algorithm. This subsection contains five different simulation tables.

1) With or without truncation: The first comparison examines the contribution of the truncation. Working with the regular Gaussian distribution causes artifacts and hence we can produce reasonable results only for low reverberation levels. The comparison is presented for $T_{60} = 150$ msec. We kept the same setup, including inter distance of 50 cm. The results are presented in Table III.

Inspecting the results, it can be deduced that truncation is an essential part of the algorithm. Even for low reverberation level the non-truncated version produces lots of artifacts causing high FA rate. The proposed algorithm detects the same locations with a much lower FA rate.

2) The influence of setting the variance: The second analysis explores the influence of setting the variance on the performance of the proposed algorithm.

We use inter distance of 50 cm and the case of two speakers mentioned above with $T_{60} = 400$ msec. The results are presented in Table IV.

Inspecting the results, it can be deduced that setting the variance to 6 samples^2 provides a good compromise between FAs and MDs.

3) Sensor noise influence: The third analysis explores the sensor noise influence on the performance. The same setup mentioned above is used, but this time sensor noise is added. The results are presented in Table V.

Inspecting the results, it can be deduced that for this reverberation level SNR of 20 dB degrades the performance significantly.

Table V
SENSOR NOISE INFLUENCE(INTER DISTANCE 50 CM, $T_{60} = 400$ MSEC): 100 MONTE CARLO TRIALS WITH TWO RANDOMLY LOCATED SIMULATED SPEAKERS

SNR[dB]	Sim. speakers	MD[%]	FA[%]	RMSE[cm]
0	2	100.0	0.0	–
20	2	41.0	1.5	5
30	2	13.0	9.0	4
40	2	11.0	11.0	4
Noiseless	2	8.0	9.0	4

Table VI
NUMBER OF NODES INFLUENCE(INTER DISTANCE 50 CM, $T_{60} = 400$ MSEC): 100 MONTE CARLO TRIALS WITH TWO RANDOMLY LOCATED SIMULATED SPEAKERS

Nodes	Sim. speakers	MD[%]	FA[%]	RMSE[cm]
6	2	6.5	72.5	6
8	2	12.4	28.1	6
10	2	12.0	16.0	4
11	2	9.5	12.5	4
12	2	8.0	9.0	4

4) Number of nodes influence: The fourth analysis explores the number of nodes influence on the performance. The same setup mentioned above is used, but this time only part of the nodes is used. The results are presented in Table VI.

As expected, adding nodes improves the localization results. Significant degradation is encountered when the number of nodes is 8 or less.

5) Subband GCC-PHAT for TDoAs estimation: The last comparison is of the proposed algorithm and a classical TDoAs estimation, GCC-PHAT adapted to the multi-dimensional localization challenge. The preprocessing of the proposed algorithm applied at each node of the network is replaced with the subband GCC-PHAT estimator. The MPTs are now calculated by the GCC-PHAT. The rest of the algorithm is kept similar.

We tuned the setup according to the limitations of this estimator. For example the inter-distance of 50 cm is too large for this estimator, yielding aliasing effects. Therefore, we tuned it to 10 cm. In addition, the GCC-PHAT cannot deal with reverberation so well, hence a much lower level is simulated ($T_{60} = 150$ msec). After those modifications we compared the two versions using the same setup. The results are presented in Table VII.

Inspecting the results, it can be deduced that even in those conditions the GCC-PHAT has inaccurate localization results and high FA rates. The proposed algorithm has very good results even though the inter-distance is only 10 cm. We can observe very small degradation in localization accuracy in the case of three speakers, but the detection rate and FA rate are kept low.

E. Acoustic room results

To further evaluate the performance of the algorithm, we analyzed real recordings as well. The sensor SNR in those cases is around 40 dB. The first examined scenario had a large

Table VII
SUBBAND GCC-PHAT FOR TDOA ESTIMATION (INTER DISTANCE 10 CM): 100 MONTE CARLO TRIALS WITH TWO OR THREE RANDOMLY LOCATED SIMULATED SPEAKERS

T_{60}	Sim. speakers	Algorithm	MD[%]	FA[%]	RMSE[cm]
150	2	Proposed	0.0	5.0	4
150	2	GCC-PHAT	0.5	88.5	80
150	3	Proposed	2.3	0.0	6
150	3	GCC-PHAT	8.3	69.7	78

distance between the speakers and a low reverberation level. The second case is considerably more challenging, since the speakers are closer to each other and the reverberation level is much higher.

In order to examine the results we plot a map, which is the set of $\psi_{\mathbf{p}}^{(LM)}$ values for the grid of positions in the room. The lines are contours of equivalent levels.

The results of the localization algorithms for the first case with a low reverberation level (200 msec) and large distance between the speakers (225 cm) are shown in Fig. 7.

It can be observed that the proposed algorithm (shown in Fig. 7(a)) produces a very accurate map of the speakers' positions with very low levels of artifacts. It can be easily estimated that there are two speakers. Their locations are identified with high accuracy (taking into account the volume of a speaker's body). The first reference algorithm, IDEM2014 (shown in Fig. 7(b)) produces a map with a few spurious peaks and large uncertainty ellipsoid even for this simpler scenario. This means that, in addition to the two real speakers, we might get quite a few FAs. The second reference algorithm, SRP-PHAT (shown in Fig. 7(c)) is also not good even for this simple scenario.

The result of the localization algorithms for the second case with a high reverberation level (930 msec) and a small distance between the speakers (100 cm) is depicted in Fig. 8.

Even for this challenging case, the proposed algorithm (shown in Fig. 8(a)) is able to produce meaningful results. The detection of the speakers can still be obtained and location estimation is very accurate. The results for IDEM2014 (shown in Fig. 8(b)) are poor, as in addition to the two real speakers, we get many FAs. The SRP-PHAT (shown in Fig. 8(c)) demonstrates unacceptable performance in this challenging case.

V. CONCLUSIONS

In this paper we have introduced a new localization algorithm using a bio-inspired cochlear model to achieve robustness against reverberation and handle concurrent speakers. The new algorithm uses a mixture of truncated Gaussian probabilistic model instead of the regular MoG usually used in the EM and the DEM algorithms.

The proposed algorithm is a distributed solution, which uses the IEM principle. Each node locally applies most of the calculations (such as the spikes detection) and has its own hidden variables, sharing only static localization parameters.

As the reverberation effect increases with the number of concurrent speakers, it becomes more challenging to detect the number of speakers and estimate their locations.

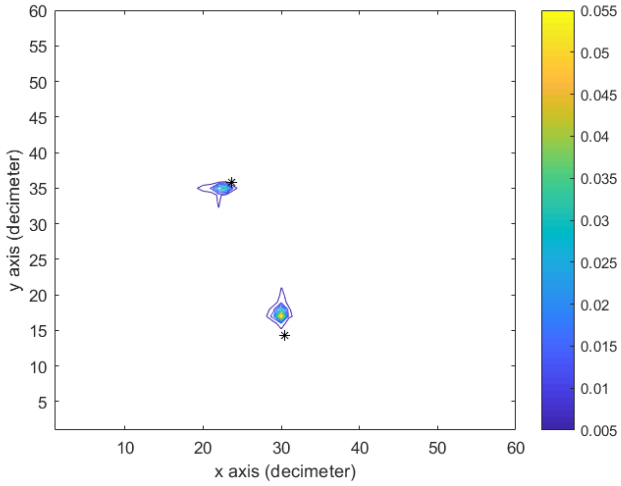
Simulations with up to three speakers were carried out in several acoustic scenarios. Real recordings analysis was carried out with two speakers in low and high reverberation levels. The proposed algorithm outperforms IDEM2014 [66] and SRP-PHAT [47] for both the simulated signals and the real-life recordings.

ACKNOWLEDGMENTS

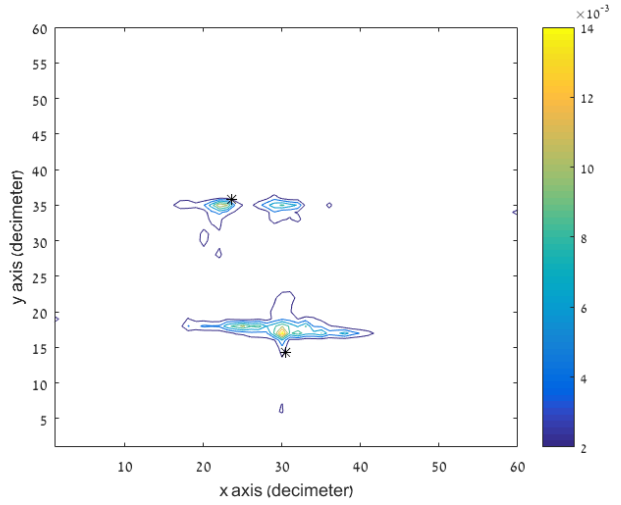
We would like to thank our four speakers recorded in the acoustic room. First, Myra and Leslie Olswang for the recording campaign. Second, Miya and Mannie Dorfan for the recording campaign and for the professional advice.

REFERENCES

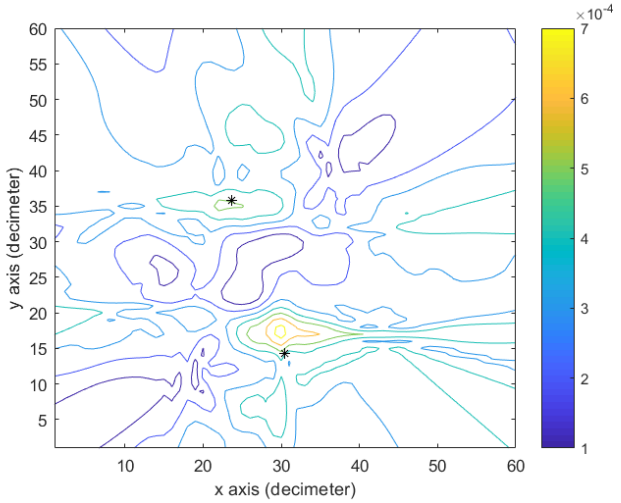
[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, vol. 8 of *Digital Signal Processing*, Springer Berlin Heidelberg, Jan. 2001.
[2] R. Martin, U. Heute, and C. Antweiler, *Advances in Digital Speech Transmission*, Wiley, 1 edition, 2008.
[3] Q. Shen, W. Liu, W. Cui, S. Wu, Y. D. Zhang, and M. G. Amin, "Low-complexity direction-of-arrival estimation based on wideband co-prime arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1445–1456, 2015.
[4] W. Xue, W. Liu, et al., "Direction of arrival estimation based on subband weighting for noisy conditions.," in *INTERSPEECH*, 2012, pp. 142–145.
[5] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.
[6] S. Tervo and A. Politis, "Direction of arrival estimation of reflections from room impulse responses using a spherical microphone array," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 10, pp. 1539–1551, 2015.
[7] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, Oct. 2017.
[8] B. Kwon, Y. Park, and Y.-s. Park, "Analysis of the GCC-PHAT technique for multiple sources," in *International Conference on Control Automation and Systems (ICCAS), 2010*, 2010, pp. 2070–2073.
[9] M. Swartling, N. Grbic, and I. Claesson, "Direction of arrival estimation for multiple speakers using time-frequency orthogonal signal separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.
[10] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.
[11] P. Pertilä, T. Korhonen, and A. Visa, "Measurement combination for acoustic source localization in a room environment," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2008, pp. 1–14, 2008.
[12] Y. Oualil, F. Faubel, and D. Klakow, "A fast cumulative steered response power for multiple speaker detection and localization," in *European Signal Processing Conference (EUSIPCO)*, Marrakech, Morocco, 2013.
[13] M. Taseska, G. Lamani, and E. A. P. Habets, "Online clustering of narrowband position estimates with application to multi-speaker detection and tracking," in *International Conference on Machine learning and Signal Processing*, 2015.
[14] A. Griffin and A. Mouchtaris, "Localizing Multiple Audio Sources from DOA Estimates in a Wireless Acoustic Sensor Network," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.
[15] A. Plinge and G. A. Fink, "Multi-speaker tracking using multiple distributed microphone arrays," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014.



(a) Proposed algorithm



(b) Original IDEM



(c) SRP-PHAT

Figure 7. Low reverberations level (T_{60} = 200 msec), two speakers, recording length 12 sec. Real positions are marked by asterisks.

[16] A. Canclini, P. Bestagini, F. Antonacci, M. Compagnoni, A. Sarti, and S. Tubaro, "A robust and low-complexity source localization algorithm for asynchronous distributed microphone networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 10, pp. 1563–1575, 2015.

[17] Y. Tian, Z. Chen, and F. Yin, "Distributed IMM-unscented Kalman filter for speaker tracking in microphone array networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 10, pp. 1637–1647, 2015.

[18] D. Li and Y. H. Hu, "Energy-based collaborative source localization using acoustic microsensor array," *EURASIP Journal on Applied Signal Processing*, pp. 321–337, 2003.

[19] D. Blatt and A. Hero, "Energy-based sensor network source localization via projection onto convex sets," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3614–3619, 2006.

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

[21] M. Chen, Z. Liu, L.-w. He, P. Chou, and Z. Zhang, "Energy-based position estimation of microphones and speakers for ad hoc microphone arrays," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2007, pp. 22 –25.

[22] C. Meesookho, U. Mitra, and S. Narayanan, "On energy-based acoustic source localization for sensor networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 365 –377, Jan. 2008.

[23] 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.

[24] W. Meng, W. Xiao, and L. Xie, "An efficient EM algorithm for energy-based multisource localization in wireless sensor networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 3, pp. 1017–1027, 2011.

[25] J. Teng, H. Snoussi, C. Richard, and R. Zhou, "Distributed variational filtering for simultaneous sensor localization and target tracking in wireless sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 5, pp. 2305–2318, June 2012.

[26] M. H. Hennecke and G. A. Fink, "Towards Acoustic Self-Localization of Ad Hoc Smartphone Arrays," in *Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, Edinburgh, UK, May 2011, pp. 127–132.

[27] P. Pertilä, M. Mieskolainen, and M. S. Hämäläinen, "Closed-form self-localization of asynchronous microphone arrays," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2011, pp. 139–144.

[28] A. Plinge, F. Jacob, R. Haeb-Umbach, and G. A. Fink, "Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms," *IEEE Signal Processing Magazine*, vol. 33, no. 4, pp. 14–29, July 2016.

[29] P. Pertilä, M. Mieskolainen, and M. S. Hämäläinen, "Passive Self-Localization of Microphones using Ambient Sounds," in *European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, Aug. 2012, pp. 1314–1318.

[30] A. Plinge, G. A. Fink, and S. Gannot, "Passive online geometry calibration of acoustic sensor networks," *IEEE Signal Processing Letters*, vol. 2017, no. 3, pp. 324–328, Mar. 2017.

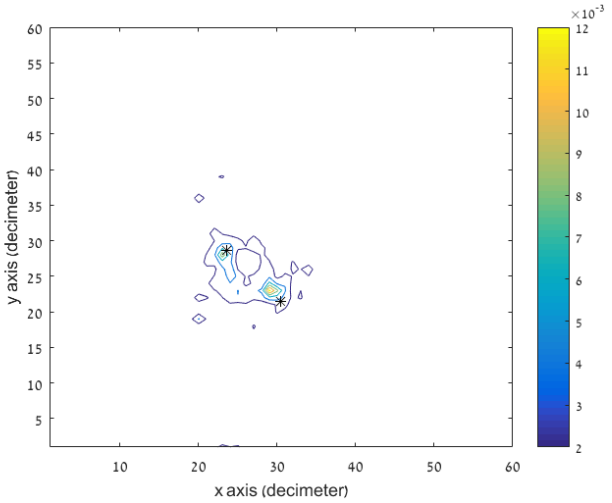
[31] O. Schwartz, Y. Dorfan, E. Habets, and S. Gannot, "Multiple DOA estimation in reverberant conditions using EM," in *International Workshop for Acoustic Echo Cancellation and Noise Control (IWAENC), Xi'an, China*, 2016.

[32] Y. Dorfan, O. Schwartz, B. Schwartz, E. A. Habets, and S. Gannot, "Multiple DOA estimation and blind source separation using estimation-maximization," in *IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, 2016.

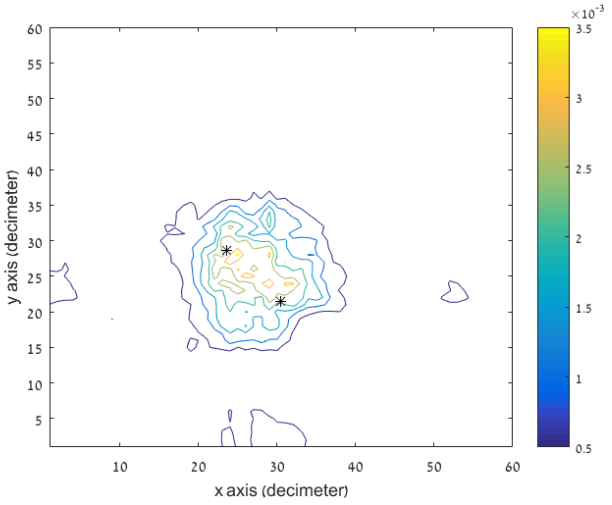
[33] A. Hassani, A. Bertrand, and M. Moonen, "Distributed node-specific direction-of-arrival estimation in wireless acoustic sensor networks," in *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*, 2013, pp. 1–5.

[34] A. Hassani, A. Bertrand, and M. Moonen, "Cooperative integrated noise reduction and node-specific direction-of-arrival estimation in a fully connected wireless acoustic sensor network," *Signal Processing*, vol. 107, pp. 68–81, 2015.

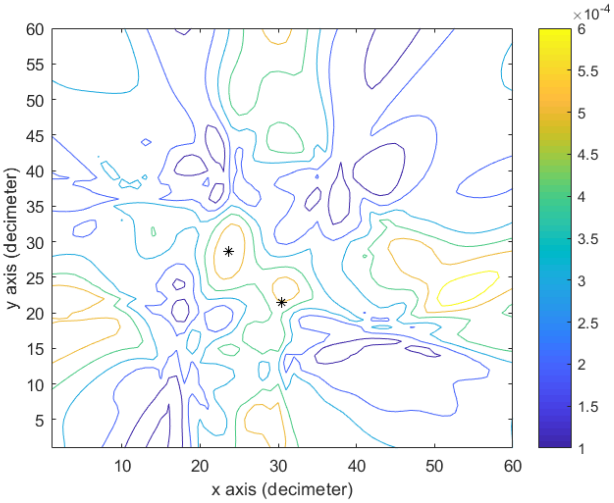
[35] O. Schwartz, Y. Dorfan, M. Taseska, E. A. Habets, and S. Gannot, "DOA estimation in noisy environment with unknown noise power using the



(a) Proposed algorithm



(b) Original IDEM



(c) SRP-PHAT

Figure 8. Maximal reverberations level ($T_{60} = 930$ msec), two speakers, recording length 12 sec. Real positions are marked by asterisks.

EM algorithm," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 86–90.

[36] F. Antonacci, D. Lonoce, M. Motta, A. Sarti, and S. Tubaro, "Efficient source localization and tracking in reverberant environments using microphone arrays," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, vol. 4, pp. 1061–1064.

[37] W. Li, Y. Jia, J. Du, and J. Zhang, "Distributed multiple model estimation for simultaneous localization and tracking with NLOS mitigation," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 6, 2013.

[38] O. Oçal, I. Dokmanic, and M. Vetterli, "Source localization and tracking in non-convex rooms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1429–1433.

[39] F. Ribeiro, D. Ba, C. Zhang, and D. Florêncio, "Turning enemies into friends: Using reflections to improve sound source localization," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2010, pp. 731–736.

[40] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised source localization on multiple-manifolds with distributed microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1477 – 1491, 2017.

[41] 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.

[42] Y. Dorfan, E. Christine, S. Gannot, and P. A. Naylor, "Speaker localization with moving microphone arrays," in *the 24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1003–1007.

[43] C. Busso, S. Hernanz, C.-W. Chu, S.-i. Kwon, S. Lee, P. G. Georgiou, I. Cohen, and S. Narayanan, "Smart room: participant and speaker localization and identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[44] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Speaker tracking on multiple-manifolds with distributed microphones," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 59–67.

[45] T. Wolff, M. Buck, and G. Schmidt, "A subband based acoustic source localization system for reverberant environments," in *ITG Conference on Voice Communication (SprachKommunikation)*. VDE, 2008.

[46] A. Brutti, M. Omologo, and P. Svaizer, "Localization of multiple speakers based on a two step acoustic map analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4349–4352.

[47] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[48] A. D. Firoozabadi and H. R. Abutalebi, "Combination of nested microphone array and subband processing for multiple simultaneous speaker localization," in *the Sixth IEEE International Symposium on Telecommunications (IST)*, 2012, pp. 907–912.

[49] A. D. Firoozabadi and H. R. Abutalebi, "Localization of multiple simultaneous speakers by combining the information from different subbands," in *21st Iranian Conference on Electrical Engineering (ICEE)*, 2013.

[50] D. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: principles, algorithms, and applications*, Wiley, 2006.

[51] R. F. Lyon, *Human and Machine Hearing: Extracting Meaning from Sound*, Cambridge University Press, June 2017.

[52] R. F. Lyon, "A computational model of binaural localization and separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1983, vol. 8, pp. 1148–1151.

[53] T. May, S. Par, and A. Kohlrausch, "Binaural localization and detection of speakers in complex acoustic scenes," in *The Technology of Binaural Listening*, J. Blauert, Ed., pp. 397–425. Springer, Berlin, Heidelberg, 2013.

[54] K. J. Palomäki, G. J. Brown, and D. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Communication*, vol. 43, no. 4, pp. 361–378, 2004.

[55] M. Bürck and J. L. van Hemmen, "Modeling the cochlear nucleus: A site for monaural echo suppression?," *Journal Acoustic Society of America*, vol. 122, pp. 2226–2235, 2007.

[56] A. Plinge, M. H. Hennecke, and G. A. Fink, "Robust neuro-fuzzy speaker localization using a circular microphone array," in *International Workshop on Acoustic Echo and Noise Control (IWAENC), Tel Aviv, Israel*, 2010.

[57] B. Grothe, "New roles for synaptic inhibtion in sound localisation," *Nature*, vol. 4, no. 7, pp. 540–550, 2003.
[58] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.
[59] A. Plinge, D. Hauschildt, M. H. Hennecke, and G. A. Fink, "Multiple speaker tracking using a microphone array by combining auditory processing and a Gaussian mixture cardinalized probability hypothesis density filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 2476–2479.
[60] A. Plinge and G. A. Fink, "Online multi-speaker tracking using multiple microphone arrays informed by auditory scene analysis," in *European Signal Processing Conference (EUSIPCO)*, 2013, pp. 1–5.
[61] A. Plinge, M. H. Hennecke, and G. A. Fink, "Reverberation-Robust Online Multi-Speaker Tracking by using a Microphone Array and CASA Processing," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, 2012.
[62] M. I. Mandel, R. J. Weiss, and D. P. 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.
[63] 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.
[64] C. M. Bishop, *Pattern recognition and machine learning*, Springer New York, 2006.
[65] Q. Zhao, V. Hautamäki, I. Kärkkäinen, and P. Fränti, "Random swap EM algorithm for Gaussian mixture models," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2120–2126, Dec. 2012.
[66] Y. Dorfan, G. Hazan, and S. Gannot, "Multiple acoustic sources localization using distributed expectation-maximization algorithm," in *4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, May 2014, pp. 72–76,.
[67] Y. Dorfan and S. Gannot, "Tree-based recursive expectation-maximization algorithm for localization of acoustic sources," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 10, pp. 1692–1703, 2015.
[68] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound: Sources in reverberant environments," *Advances in neural information processing systems*, pp. 953–960, 2007.
[69] Y. Dorfan, D. Cherkassky, and S. Gannot, "Speaker localization and separation using incremental distributed expectation-maximization," in *European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1256–1260.
[70] A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic source localization with distributed asynchronous microphone networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 439–443, 2013.
[71] G. Lee and C. Scott, "EM algorithms for multivariate Gaussian mixture models with truncated and censored data," *Computational Statistics & Data Analysis*, vol. 56, no. 9, pp. 2816–2829, 2012.
[72] W. Jank, "The EM algorithm, its randomized implementation and global optimization: Some challenges and opportunities for operations research," *Perspectives in operations research*, pp. 367–392, 2006.
[73] M. Unoki and M. Akagi, "A method of signal extraction from noisy signal based on auditory scene analysis," *Speech Communication*, vol. 27, no. 3, pp. 261–279, 1999.
[74] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1–2, pp. 103–138, August 1990.
[75] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.
[76] N. Madhu and J. Wouters, "Localisation-based, situation-adaptive mask generation for source separation," in *4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, Mar. 2010, pp. 1–6.
[77] F. Nesta and M. Omologo, "Enhanced multidimensional spatial functions for unambiguous localization of multiple sparse acoustic sources," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 213–216.
[78] 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.
[79] D. Cherkassky and S. Gannot, "Blind synchronization in wireless acoustic sensor networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 651–661, 2017.
[80] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *18th IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*. IEEE, 2011, pp. 1–6.
[81] R. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*. 1998, pp. 355–368, Kluwer Academic Publishers.
[82] 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.
[83] A. Nádas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 10, pp. 1495–1503, 1989.
[84] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE transactions on speech and audio processing*, vol. 10, no. 6, pp. 341–351, 2002.

Yuval Drofan (M'13) received the B.Sc. degree (summa cum laude) from Ben-Gurion University, Beer Sheva, Israel, in 1998, the M.Sc. degree (magna cum laude) from the Technion Israel Institute of Technology, Haifa, Israel, in 2000 both in electrical engineering. He got the MBA degree from the Interdisciplinary Center, Herzeliya, Israel, in 2007. His Ph.D. studies were in Bar Ilan University, Ramat Gan, Israel in electrical engineering till 2017.

He 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, distributed speaker tracking and bioinformatics. Currently he holds a postdoctoral research position at the Department of Biological Engineering, MIT, Cambridge, MA, USA.

Axel Plinge (M'11) received a diploma degree with distinction in computer science in 2010 from TU Dortmund University, Germany.

From 2000 he worked in different areas of psychophysical research from hearing to color vision and depth perception at the Leibniz Research Centre for Working Environment and Human Factors, Dortmund, Germany. There he participated in EC research projects and is co-inventor of novel methods in speech technology for persons with impaired hearing.

In 2012 he joined the pattern recognition group of the department of computer science at TU Dortmund University, Dortmund, Germany to pursue his PhD. There he published multiple papers on acoustic sensor networks geometry calibration, speaker tracking, and sound classification.

In 2017 he joined the AudioLabs at the Fraunhofer IIS Institute in Erlangen, Germany. He is currently working on virtual reality and IoT audio applications.

Gershon Hazan received the BS.c and the MS.c degrees from Bar Ilan University, Ramat Gan, Israel, in 1964 and 1967 respectively, both in physics. He got the Ph.D degree from Weizmann Institute of Science, Rehovot, Israel, in 1973 in biophysics. In 1973, he held a Postdoctoral position at the medical physics group, Hebrew University, Jerusalem, Israel, sponsored by the Katsir Grant. He spent a few years at the Biological Institute, Nes Ziona, Israel, a few years in Cancer Research at the Beilinson Hospital Petach-Tikva, Israel, and a few years in the Security Services of Israel, where he became a project manager in speech processing. In 2005 he retired, and joined the Biophysics Department at Bar Ilan University, Ramat Gan, Israel. In the last few years he is with the speech processing group at the faculty of engineering, Bar Ilan University.

Sharon Gannot (S'92-M'01-SM'06)) received the 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, Tel Aviv, Israel, in 1995 and 2000, respectively, all in electrical engineering. In 2001, he held a Postdoctoral position at the Department of Electrical Engineering, KULeuven, 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 a Full Professor at the Faculty of Engineering, Bar-Ilan University, Ramat Gan, Israel, where he is heading the Speech and Signal Processing laboratory and the Signal Processing Track. His 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. 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 the ELSEVIER Speech Communication and Signal Processing journals. He has served as an Associate Editor of the IEEE Transactions on Audio, Speech, and Language Processing in 2009–2013, and an area chair for the same journal 2013-2017. Currently, he serves as a moderator for arXiv in the field of audio and speech processing. He also serves as a reviewer of many IEEE journals and conferences. He is a member of the Audio and Acoustic Signal Processing technical committee of the IEEE since January 2010. Since January 2017, he serves as the committee chair. 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. He has served as the General Co-chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in October 2013. He was selected (with colleagues) to present tutorial sessions in ICASSP 2012, EUSIPCO 2012, ICASSP 2013, and EUSIPCO 2013 and a keynote speaker for IWAENC 2012 and LVA/ICA 2017. He received the Bar-Ilan University outstanding lecturer award for 2010 and 2014. Prof. Gannot is also a co-recipient of eight best paper awards.