Dorfan *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2020) 2020:9
https://doi.org/10.1186/s13636-020-00177-1

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Joint speaker localization and array calibration using expectation-maximization

Yuval Dorfan[1], Ofer Schwartz[2] and Sharon Gannot[1*†]

Abstract

Ad hoc acoustic networks comprising multiple nodes, each of which consists of several microphones, are addressed. From the ad hoc nature of the node constellation, microphone positions are unknown. Hence, typical tasks, such as localization, tracking, and beamforming, cannot be directly applied. To tackle this challenging joint multiple speaker localization and array calibration task, we propose a novel variant of the expectation-maximization (EM) algorithm. The coordinates of multiple arrays relative to an anchor array are blindly estimated using naturally uttered speech signals of multiple concurrent speakers. The speakers' locations, relative to the anchor array, are also estimated. The inter-distances of the microphones in each array, as well their orientations, are assumed known, which is a reasonable assumption for many modern mobile devices (in outdoor and in a several indoor scenarios). The well-known initialization problem of the batch EM algorithm is circumvented by an incremental procedure, also derived here. The proposed algorithm is tested by an extensive simulation study.

Keywords: Wireless acoustic sensor network, Joint calibration and localization, Expectation-maximization, Microphone array, Simultaneous speakers, W-disjoint

1 Introduction

Localization and tracking using multiple arrays of sensors are often handled under the assumption that the locations of the microphone arrays are precisely known. The recent deployment of ad hoc networks introduces a new challenge of estimating the array locations in parallel to routine tasks, such as speaker localization [1–5], noise or reverberation reduction [6–8], and speaker separation [9–13]. The solution is complex due to the amount of unknown parameters and the dependencies between them. Many scenarios do not even have a unique single solution, e.g., when the numbers of arrays or active sources are too small. In this paper, a novel expectation-maximization (EM)-based algorithm for the integrated task of speaker localization and array calibration is introduced. The new algorithm combines two tasks: direct positioning determination (DPD) and calibration for ad hoc networks.

1.1 Multiple direction of arrival estimation

The direction of arrival (DOA) estimation with known sensor positions is a well-studied problem. In [14], the steered response power (SRP)-phase transform (PHAT) algorithm is suggested, which is the generalization of the generalized cross correlation (GCC)-PHAT [15] for an array of microphones in the far-field scenario. Other known multi-channel algorithms are root multiple signal classification (MUSIC) [16, 17], minimum variance distortionless response (MVDR) [18], and audio applicable versions [19–21] These estimators were not proven to be optimal in the presence of multiple speakers. The DOA estimation [22] in the presence of various noise types can be formulated as a maximum likelihood (ML) estimation problem of deterministic parameters [23–26]. The DOA challenge in the presence of unknown noise field was dealt with in [23]. The W-disjoint orthogonality (WDO) assumption [27], commonly attributed to speech signals due to their sparseness, is often exploited for DOA estimations tasks [28].

*Correspondence: sharon.gannot@biu.ac.il
†Member, EURASIP
[1]Faculty of Engineering, Bar-Ilan University, 5290002 Ramat-Gan, Israel
Full list of author information is available at the end of the article

The problem of estimating multiple time difference of arrival (TDoA) (or DOAs) was addressed in [12, 29–31] by using the EM procedure. In [29], the task of multiple TDoA estimation is addressed considering two-microphone (binaural) case, with the WDO [27] applied, namely the dominance of a single speaker at each time-frequency (TF) bin. The authors used the EM procedure and the mixture of Gaussians (MoG) model to cluster the phase differences from each TF bin, where each cluster is associated with a TDoA value. In the E-step, a TF mask, associating each bin with a specific TDoA, was estimated. In the M-step, the probability for each TDoA was estimated, using the number of associations of TF bins.

In [30], an algorithm for estimating multiple DOAs in a reverberant environment was presented. Unlike the method presented in [29], the TF raw samples were clustered rather than their respective phase differences. The MoG model consists of explicit modeling of the reverberation properties. The resulting algorithm was able to localize multiple speakers with reverberation modeled as an additive diffuse noise with time-varying power. The reverberation power was estimated in the M-step for each speaker and for each TF bin. Note that in [30], a noiseless scenario was considered.

In the study presented in [12], the algorithm presented in [30] was extended to the problem of joint localization and separation of concurrent speakers. However, the algorithm requires a known noise power spectral density (PSD) matrix. In [31], the DOA estimation procedure presented in [12, 30] was adopted for deriving a DOA estimator for multiple sources in a noisy environment. Stationary noise was assumed with known spatial coherence but, unlike [12], the noise level was assumed unknown, and its level was estimated in the M-step.

1.2 Multiple-source cartesian localization

In this paper, when we use the term *localization*, we refer to higher-dimension problems (at least 2D). A straightforward solution to higher-dimension localization problems involves a triangulation of the 1D problems solved locally by each array of the network [32]. It has the advantage of simplicity, especially in distributed networks, where computations should be shared between nodes. There are many approaches that use triangulation of separate DOAs to solve the 2D or 3D localization problem. An example of such an approach for an acoustic ad hoc network was given in [33].

However, these solutions are not optimal, because only part of the information is utilized during the first step of the estimation. Moreover, in a small area (for example indoor scenario), a more general solution becomes a necessity since near-field conditions are often encountered. Since we do not rely on DOAs and find the locations given the measured signals, our approach is general enough to cover both near- and far-field.

A possible general solution, which directly estimates the location without any intermediate steps, is frequently referred to as DPD [34]. For acoustic localization, DPD approaches were presented in [4, 35]. In [4], a generalization of the method in [29] to the estimation of the coordinates of multiple sources, rather than only of their associated TDoAs, was presented using a grid of Cartesian coordinates that covers the room surface. The measured phase differences between microphones is then clustered to the nominal phase differences from each grid point. The probability of a speaker to be located at each grid point was estimated in the M-step. Note that in [4, 29], the spatial characteristics of the noise was not explicitly modeled and therefore not optimally treated. Some localization approaches [4, 35] use a non-realistic assumption within the context of ad hoc networks relying on perfect knowledge of array positions. This is often referred to as the *calibration* problem.

Another important challenge in ad hoc networks, tightly connected to the calibration process, is the clock synchronization. Both acoustic and non-acoustic solutions were proposed to overcome this challenge [36–45]. In the current work, we assume that the nodes are perfectly synchronized, by possibly using one of those approaches. It has been shown that current technology used by commercial personal consumer electronics, like smartphones, provides very small drift and jitters in the clock frequency that can be compensated for by these algorithms. We will hereinafter ignore the synchronization issues.

1.3 Array calibration

Finding the location of microphones is a well-covered topic in the literature. For example [46] deals with finding the location of a microphone utilizing a single loudspeaker and the room known shape. Array constellation calibration has been analyzed from a theoretical point of view for far-field [47] and near-field scenarios [48]. For acoustic arrays, a few approaches have already been proposed for calibration, some of which are only suitable for scenarios with a dedicated time for calibrations [49]. Other algorithms utilize ambient sound for finding the inter-distances of microphones [50, 51].

Calibration performed jointly with localization or tracking of sources presents a greater challenge. A family of algorithms called simultaneous localization and mapping (SLAM) for robots was described in [52–54]. In these contributions, the joint estimation of a single moving array trajectory, the positions of static sources, and the major reflectors (e.g., walls) is addressed.

Another popular problem is the estimation of static array locations jointly with tracking of moving acoustic sources [55, 56]. The problem is sometimes referred to as

simultaneous localization and tracking (SLAT) [57]. Effective solutions for array calibration in dynamic scenarios can utilize the multiple locations visited by the speakers. Such a method, based on genetic algorithm, was recently presented for a scenario where speakers move around a table in the center of the room [58]. The arrays are located on the table and the algorithm estimates the arrays' locations and tracks the speakers. The sensitivity to small movements are discussed in [59, 60].

Approaches suitable for static scenarios can also be found in the literature, e.g., [61, 62]. They rely on TDoAs between adjacent microphones. Other joint calibration approaches are described in [63–65]. Those methods currently work under a very specific set of geometrical conditions. For example, some of them require moving speakers or a minimum number of active speakers to guarantee sufficient amount of data to overcome the problem of geometrical ambiguities. In [64], the proposed algorithm automatically determine the relative three-dimensional positions of audio sensors and sources in an ad hoc network. A closed form approximate solution is derived, which is further refined by minimizing a nonlinear error function. They also account for the lack of temporal synchronization among different platforms. Recently, several approaches, suitable for the static scenario, were presented. The joint estimation problem is solved by applying various mathematical methods [66, 67].

1.4 Proposed strategy

We suggest in this paper a new EM-based speaker localization and array calibration algorithm. The microphone inter-distances in each array, as well as the orientation of each array, are assumed known in advance, as can be commonly verified in commercial devices, e.g., cellular phones. In addition, since we use omni-directional microphones, it enables usage of acoustic calibration approaches such as inter distance measurements. However, the network constellation, namely the center points of the arrays and the locations of the sources, are unknown in advance and should be jointly estimated by the algorithm.

The challenge is to solve the localization problem of multiple concurrent speakers (more than two) jointly with the calibration problem of multiple arrays without any other information or any additional calibration signals. Following [4], we use the EM and the MoG models to cluster the observed data to centroids located on a grid defined on the surface. An explicit model of the speech and noise is defined within the MoG model, as used in [12].

To address the calibration problem, we add the locations of the array centers to the estimation task. As a result, the locations of the array centers are estimated in the M-step. Maximization of the auxiliary function of the EM with respect to (w.r.t.) the array centers does not produce a closed-form expression. We utilize the simplifying assumption that the noise signals, as captured by the different arrays, are uncorrelated. This assumption enables us to avoid a multidimensional search of the array centers, i.e., a separate search for each array is obtained, and can be justified empirically if the array centers are sufficiently separated.

The initialization stage was found to be a cumbersome task, due to the large size of the parameter set. We present a new way for self-initialization, which utilizes the collected data in an incremental fashion. One of the arrays is designated as the *anchor array* and all the other elements (arrays and sources) are localized w.r.t. this anchor. First, the algorithm is applied with only the anchor array while the other arrays are disabled. Then, the other arrays in the network are sequentially added. The location of sources is kept as a soft probability map throughout the iterative procedure. Only after the last iteration, an actual localization is obtained by applying a hard threshold to the final probability map. In this paper, for simplicity, the speakers are assumed to be spatially static across time. In the case of moving speakers, a virtue of recursive EM (REM) algorithm can be utilized [4] using our EM model for the fixed speakers.

1.5 Main contributions

The main contributions of this paper are listed below:

1. The problem of joint estimation of the array center positions and multiple speaker position is addressed. The problem is statistically formulated using the probability density function (p.d.f.) of the observations. By maximizing the likelihood of the observations via the EM algorithm, the source positions are inferred.
2. Searching the array center positions is carried out separately for each array, avoiding a simultaneous multidimensional search of the entire set of possible array centers.
3. The statistical model of the multiple speech signals is based on the WDO assumption [27], which was proven to be highly efficient for speaker separation tasks.

2 Methods

We start from a mathematical description of the problem in the first subsection and then derive the new algorithm in the second subsection.

2.1 Problem formulation

We derive a batch EM solution for joint estimation of the positions of static speakers and microphone arrays. The problem formulation is divided into two parts. The first

describes the ad hoc network signals in the presence of multiple concurrent speakers and sensor noise, and the second presents the statistical model.

2.1.1 Signal model

Consider Q arrays, each of which is equipped with N microphones receiving signals from J speakers. The number of speakers is not necessarily known in advance. The measured signals are linear combinations of the incoming waveforms. Let $Z_{q,n}(t,k)$ be the signals received by the (q,n)th microphone, where $q = 1,\ldots,Q$ is the array index and $n = 1,\ldots,N$ is the microphone index within each array. Overall, there are $Q \times N$ microphones. The signals in the short-time Fourier transform (STFT) domain are given by:

$$Z_{q,n}(t,k) = \sum_{j=1}^{J} G_{q,n,j}(k) \cdot S_j(t,k) + V_{q,n}(t,k), \tag{1}$$

where $t = 0,\ldots,T-1$ and $k = 0,\ldots,K-1$ denote the time and frequency indexes respectively. $G_{q,n,j}(k)$ is the direct transfer function (DTF) associating speaker j and microphone (q,n). $S_j(t,k)$ is the speech signal uttered by speaker j and $V_{q,n}(t,k)$ is the ambient noise, namely noise signals that result from the environment. Specific spatial characteristics of the noise signals will be later discussed.

Note that the DTF model accounts for near-field scenarios and hence comprises the attenuation of the direct speech wave as well as the respective inter-microphone phase. Also note that the attenuation is known to be much less reliable than the phase. Therefore, multiple arrays should be used. It is demonstrated in the Section 3 by adding arrays of sensors one by one. The DTF is given by:

$$G_{q,n,j}(k) = \frac{1}{d_{q,n,j}} \exp\left(-\iota \frac{2\pi k}{K} \frac{d_{q,n,j}}{c \cdot T_s}\right), \tag{2}$$

where c is the sound velocity and T_s denotes the sampling period. The distance from speaker j to microphone (q,n), $d_{q,n,j}$ is calculated from geometrical considerations as:

$$d_{q,n,j} = ||\mathbf{p}_j - \mathbf{p}_{q,n}||, \tag{3}$$

where $\mathbf{p}_j$ is the location of the jth speaker and $\mathbf{p}_{q,n}$ is the location of the (q,n)th microphone given by:

$$\mathbf{p}_{q,n} = \mathbf{p}_q + \mathbf{p}_n(q), \tag{4}$$

where $\mathbf{p}_q$ is the position of the center of the qth array and $\mathbf{p}_n(q)$ is the relative position of the nth microphone w.r.t. the array center. The inter-structure of the arrays and their orientation, namely $\mathbf{p}_n(q)$, are assumed to be known in advance. Note that the *orientation* of the arrays can be extracted by various means, for example, compass-based technology [68, 69]. The orientation accuracy is often reported around 5° indoor and much better for outdoor scenarios. For simplicity, we assume hereinafter that the orientation of the nodes is perfectly known to the algorithms, since joint estimation of positions and orientation is too cumbersome, at this stage.

To address reverberant environments, an additional term representing the ambient reverberation field can be added to (1). As indicated in, e.g., [30], the reverberation components can be modeled as an additive multi-dimensional Gaussian interference with spatially diffuse sound field with time-varying level, following the anechoic speech level. In such a case, the reverberation level can also be estimated by the M-step of the EM procedure. In this paper, for the sake of simplicity, the reverberation phenomenon is ignored. It means that the solution will fit indoor with low reverberation levels and outdoor scenarios that are dominant by random noise.

The N microphone signals in the qth array can be concatenated in a vector form:

$$\mathbf{z}_q(t,k) = \sum_{j=1}^{J} \mathbf{g}_{q,j}(k) S_j(t,k) + \mathbf{v}_q(t,k), \tag{5}$$

where:

$$\mathbf{z}_q(t,k) = \left[\, Z_{q,1}(t,k) \;\ldots\; Z_{q,N}(t,k) \,\right]^{\mathrm{T}} \tag{6a}$$

$$\mathbf{g}_{q,j}(k) = \left[\, G_{q,1,j}(k) \;\ldots\; G_{q,N,j}(k) \,\right]^{\mathrm{T}} \tag{6b}$$

$$\mathbf{v}_q(t,k) = \left[\, V_{q,1}(t,k) \;\ldots\; V_{q,N}(t,k) \,\right]^{\mathrm{T}}. \tag{6c}$$

The overall observation set, DTFs, and noise components can be concatenated in compound vectors:

$$\mathbf{z}(t,k) = \left[\, \mathbf{z}_1^T(t,k) \;\ldots\; \mathbf{z}_Q^T(t,k) \,\right]^{\mathrm{T}}, \tag{7a}$$

$$\mathbf{g}_j(k) = \left[\, \mathbf{g}_{1,j}^T(k) \;\ldots\; \mathbf{g}_{Q,j}^T(k) \,\right]^{\mathrm{T}}, \tag{7b}$$

$$\mathbf{v}(t,k) = \left[\, \mathbf{v}_1^T(t,k) \;\ldots\; \mathbf{v}_Q^T(t,k) \,\right]^{\mathrm{T}}, \tag{7c}$$

such that:

$$\mathbf{z}(t,k) = \sum_{j=1}^{J} \mathbf{g}_j(k) S_j(t,k) + \mathbf{v}(t,k). \tag{8}$$

The goal of this study is to jointly estimate the speaker locations $\mathbf{p}_j$ and the array center positions $\mathbf{p}_q$, in (3),(4).

2.1.2 Statistical model

We use a MoG probability function to characterize the speech signals of all potential speakers. Each speaker can be assumed to be a complex-Gaussian source emitting acoustic waveforms from location $\mathbf{p}_m$, where m is the index of the Gaussian component. Because the number of speakers and their locations are unknown in advance, we use a predefined grid as candidate source positions.

The various speakers are assumed to exhibit disjoint activity in the STFT domain (WDO assumption [27]). Therefore, by means of clustering, every TF bin of $\mathbf{z}(t,k)$ can be associated with a single active source.

Based on the disjoint activity of the sources, the observations are given the following probabilistic description:

$$\mathbf{z}(t,k) \sim \sum_{m=1}^{M} \psi_m \cdot \mathcal{N}^c\left(\mathbf{z}(t,k); \mathbf{0}, \Phi_m(t,k)\right), \tag{9}$$

where ψ_m is the (unknown) probability of a speaker present at $\mathbf{p}_m$ and M is the number of Gaussians. $\mathcal{N}^c(\cdot;\cdot,\cdot)$ denotes the complex Gaussian p.d.f.:

$$\mathcal{N}^c(\mathbf{y}; \mathbf{0}, \mathbf{\Sigma}) = \frac{1}{\pi^{(QN)} \det(\mathbf{\Sigma})} \exp\left(\mathbf{y}^{\mathrm{H}} \mathbf{\Sigma}^{-1} \mathbf{y}\right), \tag{10}$$

with $\mathbf{y}$ a zero-mean complex-Gaussian random vector and $\mathbf{\Sigma}$ its PSD matrix.

The matrix $\Phi_m(t,k)$ is the PSD of $\mathbf{z}(t,k)$, given that $\mathbf{z}(t,k)$ is associated with the speaker located at $\mathbf{p}_m$:

$$\Phi_m(t,k) = \mathbf{g}_m(k)\mathbf{g}_m^{\mathrm{H}}(k)\phi_{S,m}(t,k) + \Phi_{\mathbf{v}}(k), \tag{11}$$

where the DTF $\mathbf{g}_m(k)$ is defined in (7b).

The direct-path temporal PSD $\phi_{S,m}(t,k)$ and the noise PSD matrix $\Phi_{\mathbf{v}}(t,k)$ are defined as:

$$\phi_{S,m}(t,k) = E\left\{|S_m(t,k)|^2\right\}, \tag{12}$$

$$\Phi_{\mathbf{v}}(k) = E\left\{\mathbf{v}(t,k)\mathbf{v}^{\mathrm{H}}(t,k)\right\}. \tag{13}$$

The noise components from different arrays are often assumed to be uncorrelated [23], and thus:

$$\Phi_{\mathbf{v}}(k) = \mathrm{Blockdiag}\left[\ \Phi_{\mathbf{v}_1}(k)\ \ldots\ \Phi_{\mathbf{v}_Q}(k)\ \right], \tag{14}$$

where $\Phi_{\mathbf{v}_q}(k) = E\left\{\mathbf{v}_q(t,k)\mathbf{v}_q^{\mathrm{H}}(t,k)\right\}$. This assumption is a key assumption (as elaborated later) because it allows the estimation of the array centers to be separately executed for each array. This assumption can be well justified in the presence of a spatially white or diffuse noise field, assuming that the inter-array distances are large enough. For the case of a directional noise field, this assumption is invalid.

The PSD matrices of the noise are assumed to be time-invariant and known in advance or can be estimated during speech absence segments.

Finally, by augmenting all observations for $t = 0,\ldots,T-1$ and $k = 0,\ldots,K-1$ in $\mathbf{z} = \mathrm{vec}_{t,k}(\{\mathbf{z}(t,k)\})$, the p.d.f. of the entire observation set can be stated as:

$$f(\mathbf{z}) = \prod_{t,k} \sum_{m=1}^{M} \psi_m \cdot \mathcal{N}^c\left(\mathbf{z}(t,k); \mathbf{0}, \Phi_m(t,k)\right), \tag{15}$$

where the readings for all TF bins are assumed independent [27].

Let the unknown parameter set be $\boldsymbol{\theta} = \left[\mathbf{p}^{\mathrm{T}}, \boldsymbol{\psi}^{\mathrm{T}}, \boldsymbol{\phi}_S^{\mathrm{T}}\right]^{\mathrm{T}}$, where $\mathbf{p} = \mathrm{vec}_q(\mathbf{p}_q)$, $\boldsymbol{\psi} = \mathrm{vec}_m(\psi_m)$, and $\boldsymbol{\phi}_S = \mathrm{vec}_{m,t,k}\left(\phi_{S,m}(t,k)\right)$. It should be emphasized that, unlike the array locations, the speaker locations are indirectly estimated by the soft variables $\boldsymbol{\psi}$ that form a probability map. The number of speakers and their locations are inferred from this probability map.

The maximum likelihood estimation (MLE) problem can readily be stated as:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log f(\mathbf{z}; \boldsymbol{\theta}). \tag{16}$$

The various assumptions leading to the MLE problem statement are summarized in the following list:

1. Noise signals for different arrays are assumed uncorrelated in (14). This assumption is valid for non-coherent sources (i.e., spatially white or diffuse noise fields). This assumption will be used to simplify the optimization problem.
2. Speakers exhibit disjoint activity in each TF bin, namely $\mathbf{z}(t,k)$, is dominated by a single source in (9), as suggested in [27] and subsequent contributions.
3. Noise and speech signals are modeled by complex-Gaussian variables. This assumption is widely used in many speech processing algorithms and can be attributed to the properties of the Fourier transform of sufficiently long frames.
4. Each microphone array is calibrated, i.e., array internal geometry, $\mathbf{p}_n(q)$ is known.
5. Each microphone array orientation is also known (for example, by using a compass-based technology or a GPS).
6. The speakers are assumed static, namely their positions are fixed and do not change in time. In future research, moving speakers scenarios will be addressed using a virtue of recursive EM, inspired by [4].
7. The reverberation phenomenon is ignored. The presented algorithm is therefore better suited to scenario that are dominated by random noise, e.g., outdoor scenarios.

In the next subsection, an algorithm is derived for estimating $\boldsymbol{\theta}$. The first two components are the required parameters (array centers and source positions). The last component $\boldsymbol{\phi}_S$ is a set of nuisance parameters. Since the MLE in this case is of high complexity, it is necessary to use an iterative search algorithm. A widely used algorithm for this type of problems is the EM algorithm. We derive the basic (batch) version of the algorithm. For performance improvement and for mitigating the dependency on the algorithm initialization, we also further introduce a novel modification of that basic EM.

2.2 Localization and calibration expectation-Maximization sequence (LACES)

The MLE of $\boldsymbol{\theta}$ is developed using the EM algorithm. It uses three datasets and their probability models: the observations, the target parameters (these datasets were already defined in Section 2.1), and the hidden datasets that will be estimated by the algorithm. In our case, we set the

hidden data to comprise: (1) the speech signals $S_m(t,k)$, which are potentially emitted from each location m in the room, and (2) the *association* of each TF bin with a single source emitting from a particular location, as in [4].

The *association* of each TF bin is expressed by $x(t,k,m)$, an indicator that the bin (t,k) is associated with a speaker located at $\mathbf{p}_m$. The total number of indicators in the problem is $T \times K$. Note that, under the WDO assumption [27], each TF bin is dominated by a single speaker.

This subsection is split into five parts. In the first part, the basic EM equations are derived. The second one is dedicated to the E-step and the third to the M-step. The fourth summarizes the algorithm and its initialization process. Complexity analysis is given in the last part.

2.2.1 Basic expectation-maximization steps derivation

Denote the hidden data as:

$$\mathbf{x} = \mathrm{vec}_{t,k,m}\left(\{x(t,k,m)\}\right) \tag{17}$$

$$\mathbf{s} = \mathrm{vec}_{t,k,m}\left(\{S_m(t,k)\}\right). \tag{18}$$

Following Bayes' rule, the p.d.f. of the complete dataset, $\mathbf{z}$, $\mathbf{x}$ and $\mathbf{s}$, is obtained by:

$$f(\mathbf{z},\mathbf{x},\mathbf{s};\boldsymbol{\theta}) = f(\mathbf{z}|\mathbf{x},\mathbf{s};\boldsymbol{\theta}) f(\mathbf{x}|\mathbf{s};\boldsymbol{\theta}) f(\mathbf{s};\boldsymbol{\theta}). \tag{19}$$

The conditional distribution of the observed data given the hidden data can be expressed as:

$$f(\mathbf{z}|\mathbf{x},\mathbf{s};\boldsymbol{\theta}) = \prod_{t,k}\sum_{m=1}^{M} x(t,k,m)\, f(\mathbf{z}(t,k)|x(t,k,m)=1,\mathbf{s};\boldsymbol{\theta}). \tag{20}$$

Using the assumption that the noise signals, as captured by the different arrays are uncorrelated (14), the p.d.f. of the noise signals can be decomposed to a multiplication of per-array quantities:

$$\begin{aligned} f(\mathbf{z}(t,k)|x(t,k,m)=1,\mathbf{s};\boldsymbol{\theta}) &= \mathcal{N}^c\left(\mathbf{z}(t,k) - \mathbf{g}_m(k)S_m(t,k);\mathbf{0},\Phi_{\mathbf{v}}(k)\right) \\ &= \prod_q \mathcal{N}^c\left(\mathbf{z}_q(t,k) - \mathbf{g}_{q,m}(k)S_m(t,k);\mathbf{0},\Phi_{\mathbf{v}_q}(k)\right). \end{aligned} \tag{21}$$

Since the indicator x is independent of speech signals $\mathbf{s}$, its conditional p.d.f. is given by:

$$f(\mathbf{x}|\mathbf{s};\boldsymbol{\theta}) = f(\mathbf{x};\boldsymbol{\theta}) = \prod_{t,k}\sum_{m=1}^{M} x(t,k,m)\psi_m. \tag{22}$$

The speech p.d.f. is frequently assumed to follow a complex-Gaussian distribution:

$$f(\mathbf{s};\boldsymbol{\theta}) = \prod_{t,k,m} \mathcal{N}^c\left(S_m(t,k);0,\phi_{S,m}(t,k)\right). \tag{23}$$

The p.d.f. of the complete dataset is then obtained by collecting the terms in (19)-(23):

$$\begin{aligned} f(\mathbf{x},\mathbf{z},\mathbf{s};\boldsymbol{\theta}) = &\left(\prod_{t,k}\sum_{m=1}^{M} x(t,k,m)\psi_m \right. \\ &\left. \times \prod_q \mathcal{N}^c\left(\mathbf{z}_q(t,k) - \mathbf{g}_{q,m}(k)S_m(t,k);\mathbf{0},\Phi_{\mathbf{v}_q}(k)\right)\right) \\ &\times \left(\prod_{t,k,m} \mathcal{N}^c\left(S_m(t,k);0,\phi_{S,m}(t,k)\right)\right). \end{aligned} \tag{24}$$

2.2.2 E-step

For any variable, the denotation $\widehat{(\cdot)}$ refers to $E\left\{(\cdot)|\mathbf{z};\boldsymbol{\theta}^{(\ell-1)}\right\}$. The auxiliary function in our case can be stated as:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(\ell-1)}) &\triangleq \log\widehat{f(\mathbf{z},\mathbf{x},\mathbf{s};\boldsymbol{\theta})} \\ &= Q_1(\boldsymbol{\psi}|\boldsymbol{\theta}^{(\ell-1)}) + Q_2(\mathbf{p}|\boldsymbol{\theta}^{(\ell-1)}) + Q_3(\boldsymbol{\phi}_S|\boldsymbol{\theta}^{(\ell-1)}), \end{aligned} \tag{25}$$

where:

$$Q_1(\boldsymbol{\psi}|\boldsymbol{\theta}^{(\ell-1)}) = \sum_{t,k,m} \widehat{x}(t,k,m)\log\psi_m, \tag{26a}$$

$$\begin{aligned} Q_2(\mathbf{p}|\boldsymbol{\theta}^{(\ell-1)}) = &\sum_{t,k,m,q} \\ &\widehat{x(t,k,m)\log\mathcal{N}^c\left(\mathbf{z}_q(t,k) - \mathbf{g}_{q,m}(k)S_m(t,k);\mathbf{0},\Phi_{\mathbf{v}_q}(k)\right)}, \end{aligned} \tag{26b}$$

$$Q_3(\boldsymbol{\phi}_S|\boldsymbol{\theta}^{(\ell-1)}) = \sum_{t,k,m} \widehat{\log\mathcal{N}^c\left(S_m(t,k);0,\phi_{S,m}(t,k)\right)}. \tag{26c}$$

Note that, due to the indicator properties of $x(t,k,m)$, the summation over m is carried out outside the logarithm operation.

For implementing the E-step, the sufficient statistics of the hidden variables are evaluated by the following expressions:

$$1)\quad \widehat{x}(t,k,m), \tag{27a}$$

$$2)\quad \widehat{x(t,k,m)S_m}(t,k), \tag{27b}$$

$$3)\quad \widehat{x(t,k,m)\cdot|S_m(t,k)|^2}, \tag{27c}$$

$$4)\quad \widehat{|S_m(t,k)|^2}. \tag{27d}$$

In the next list, these expressions are mathematically derived.

1. The expected associations:

$$\widehat{x}^{(\ell)}(t,k,m) \triangleq E\left\{x(t,k,m)|\mathbf{z}(t,k);\boldsymbol{\theta}^{(\ell-1)}\right\} = \frac{\psi_m^{(\ell-1)}\mathcal{N}^c\left(\mathbf{z}(t,k);\mathbf{0},\Phi_m^{(\ell-1)}(t,k)\right)}{\sum_m \psi_m^{(\ell-1)}\mathcal{N}^c\left(\mathbf{z}(t,k);\mathbf{0},\Phi_m^{(\ell-1)}(t,k)\right)}, \tag{28}$$

where:

$$\Phi_m^{(\ell-1)}(t,k) = \mathbf{g}_m^{(\ell-1)}(k)\cdot\left(\mathbf{g}_m^{(\ell-1)}(k)\right)^{\mathrm{H}}\phi_{S,m}^{(\ell-1)}(t,k) + \Phi_{\mathbf{v}}(k). \tag{29}$$

Note that the direct-path $\mathbf{g}_m^{(\ell-1)}(k)$ is calculated before each E-step according to the estimated array locations for all possible grid points. The expression for $\mathbf{g}_m^{(\ell-1)}(k)$ is given by (7b) and (2), while exchanging the source index j with the candidate location index m and using the estimated array positions $\mathbf{p}_q$ rather than its true value.

2. The next term for the E-step is the first-order statistics of the speech multiplied by the indicator, given the measurements and the parameters. Using the law of total expectation:

$$\widehat{(\cdot)} = \widehat{x}\times E\left\{(\cdot)|x=1,\mathbf{z}(t,k);\boldsymbol{\theta}^{(\ell-1)}\right\} + (1-\widehat{x})\times E\left\{(\cdot)|x=0,\mathbf{z}(t,k);\boldsymbol{\theta}^{(\ell-1)}\right\}. \tag{30}$$

Accordingly, the first-order statistics of the speech multiplied by the indicator is then given by (31). Note that the expectation of the mth speaker when the (t,k) bin is associated with the mth speaker is the multichannel Wiener filter (MCWF) (see [70, Eq. (28)]). Otherwise, the expectation is the prior of the signal, as defined in (23), namely identically zero.

$$\begin{aligned}\widehat{x(t,k,m)S_m}(t,k) &= \widehat{x}^{(\ell)}(t,k,m)\\ &\times E\left\{x(t,k,m)S_m(t,k)|x(t,k,m)=1,\mathbf{z}(t,k);\boldsymbol{\theta}^{(\ell-1)}\right\}\\ &+(1-\widehat{x}^{(\ell)}(t,k,m))\\ &\times E\left\{x(t,k,m)S_m(t,k)|x(t,k,m)=0,\mathbf{z}(t,k);\boldsymbol{\theta}^{(\ell-1)}\right\} =\\ &\widehat{x}^{(\ell)}(t,k,m)\cdot\phi_{S,m}^{(\ell-1)}(t,k)\left(\mathbf{g}_m^{(\ell-1)}(k)\right)^{\mathrm{H}}\left(\Phi_m^{(\ell-1)}(t,k)\right)^{-1}\mathbf{z}(t,k).\end{aligned} \tag{31}$$

3. The third term for the E-step is the expected speech second-order statistics multiplied by the indicator. Using the law of total expectation, the expected speech second-order statistics multiplied by the indicator is given by (32). Note that, when the (t,k)th bin is associated with the mth speaker, the expected speech second-order statistics is the squared MCWF plus the associated error co variance term (see [70, Eq. (32)]).

$$\widehat{x(t,k,m)\cdot|S_m(t,k)|^2} = \widehat{x}^{(\ell)}(t,k,m)\left[\left|\widehat{S}_m^{(\ell)}(t,k)\right|^2 + \phi_{S,m}^{(\ell-1)}(t,k) - \left(\phi_{S,m}^{(\ell-1)}(t,k)\right)^2\left(\mathbf{g}_m^{(\ell-1)}(k)\right)^{\mathrm{H}}\left(\Phi_m^{(\ell-1)}(t,k)\right)^{-1}\cdot\mathbf{g}_m^{(\ell-1)}(k)\right]. \tag{32}$$

4. The last term of the E-step is the expected speech second-order statistics. Using the law of total expectation, the expected speech second-order statistics is given by (33), which is a weighted sum (according to the estimate of the indicator) of the conditional expectation in (32) and the prior variance $\phi_{S,m}^{(\ell-1)}(t,k)$. Note that, when the (t,k)th bin is not associated with the mth speaker, the expected speech second-order statistics is simply the prior variance $\phi_{S,m}^{(\ell-1)}(t,k)$.

$$\widehat{|S_m(t,k)|^2} = \widehat{x}^{(\ell)}(t,k,m)\left[\left|\widehat{S}_m^{(\ell)}(t,k)\right|^2 + \phi_{S,m}^{(\ell-1)}(t,k) - \left(\phi_{S,m}^{(\ell-1)}(t,k)\right)^2\left(\mathbf{g}_m^{(\ell-1)}(k)\right)^{\mathrm{H}}\left(\Phi_m^{(\ell-1)}(t,k)\right)^{-1}\cdot\mathbf{g}_m^{(\ell-1)}(k)\right] + \left(1-\widehat{x}^{(\ell)}(t,k,m)\right)\left[\phi_{S,m}^{(\ell-1)}(t,k)\right]. \tag{33}$$

2.2.3 M-step

The second step of the iterative algorithm is the maximization of (25) w.r.t. the unknown deterministic parameters $\boldsymbol{\theta}$, namely the M-step:

1. Similarly to [4, Eq. (20a)], ψ_m is obtained by a constrained[1] maximization of $Q_1(\boldsymbol{\psi}|\boldsymbol{\theta}^{(\ell-1)})$ in (25):

$$\psi_m^{(\ell)} = \frac{\sum_{t,k}\widehat{x}^{(\ell)}(t,k,m)}{T\cdot K}. \tag{34}$$

2. The array locations are obtained by the maximization:

$$\mathbf{p}_1^{(\ell)},\ldots,\mathbf{p}_Q^{(\ell)} = \mathrm{argmax}_{\mathbf{p}_1,\ldots,\mathbf{p}_Q}\ Q_2(\mathbf{p}|\boldsymbol{\theta}^{(\ell-1)}). \tag{35}$$

There is no closed-form solution for the array centers, and therefore, a straightforward solution will require a tedious evaluation of the expression (35) in $|P|^Q$ points. Such a search is extremely complex. However, due to the assumption that the noise signals at different arrays are uncorrelated (14), $Q_2(\mathbf{p}|\boldsymbol{\theta}^{(\ell-1)})$ simplifies and the search can be carried out separately for each $\mathbf{p}_q^{(\ell)}$.

$$\mathbf{p}_q^{(\ell)} = \mathrm{argmax}_{\mathbf{p}_q}\sum_{t,k,m}2\mathrm{Re}\left\{\mathbf{z}_q^H(t,k)\Phi_{\mathbf{v}_q}^{-1}(k)\mathbf{g}_{q,m}(k)\widehat{x(t,k,m)S_m}(t,k)\right\} - \left(\mathbf{g}_{q,m}(k)\right)^H\Phi_{\mathbf{v}_q}^{-1}(k)\mathbf{g}_{q,m}(k)\widehat{x(t,k,m)\cdot|S_m}(t,k)|^2. \tag{36}$$

Because the search is carried out for each array separately, it requires $|P|\cdot Q$ calculations of the

[1] The sum of ψ_m equals 1. The full derivation can be found in [71, Sec. 9.2.2].

likelihood term in (35), resulting in a significant calculation saving. Note that $\mathbf{p}_q$ determines $\mathbf{g}_{q,m}(k)$, as evident from (2)-(4).

3. The variance of the speech is obtained by maximizing $Q_3(\boldsymbol{\phi}_S|\boldsymbol{\theta}^{(\ell-1)})$, resulting in:

$$\phi_{S,m}^{(\ell)}(t,k) = |\widehat{S_m(t,k)}|^2. \tag{37}$$

which is the periodogram of the speech signal, using its second-order statistics.

2.2.4 The IACES algorithm: summary

A conventional EM procedure for the problem at hand can be formalized for any number of nodes, $\tilde{Q}$ according to Algorithm 1, required L iterations.

Algorithm 1: Algorithm for EM-steps.

E-step

Estimate $\widehat{x}(t,k,m)$ using (28), $\widehat{x(t,k,m)S_m}(t,k)$ using (31), $\widehat{x(t,k,m)\cdot|S_m(t,k)|^2}$ using (32), and $|\widehat{S_m(t,k)}|^2$ using (33).

M-step

Calculate $\psi_m^{(\ell)}$ using (34), $\phi_{S,m}^{(\ell)}(t,k)$ using (37), and $\mathbf{p}_q^{(\ell)}$ using (36) $\forall q = 2,\ldots,\tilde{Q}$.

The classical batch EM algorithm is sensitive to initialization and might converge to a local maximum instead of the global maximum likelihood [71]. Several solutions have been suggested [72] to circumvent the misconvergence phenomenon, including incremental [73], sparse [72], recursive [74], and other variants of the batch EM algorithm. Experimentally, it has been shown that the proposed algorithm might suffer from this misconvergence if a conventional initialization is applied.

In addition, because all locations of the microphones and the speakers in our model are unknown, the origin of the coordinate system should be predefined. We decided to use one of the arrays as the origin, referred to as the *anchor* node. The entire microphone/speaker constellation is then measured w.r.t. this node. Consequently, the EM algorithm should only search for $Q-1$ array center locations.

We propose the following incremental procedure that was empirically shown to converge to the MLE. First, only the anchor node is used by the algorithm. ψ_m is initialized to a uniform distribution, and $\phi_{S,m}(t,k)$ is calculated based on the anchor position. The nodes are added incrementally until all Q nodes used by the ad hoc network are included. After adding each node, EM iterations are applied with the current measurements, as captured by the $\tilde{Q}$ nodes. In general, the number of iterations can be set to $L > 1$, but empirically, we see that $L = 1$ iteration is sufficient for each node addition. The localization and calibration EM sequence (LACES) algorithm is summarized in Algorithm 2.

Algorithm 2: LACES algorithm for noisy environments.

Initialize
$\psi_m^{(0)} = \frac{1}{M}$
$\phi_{S,m}^{(0)}(t,k) = |\mathbf{z}(t,k)|^2$
$\mathbf{p} = \mathbf{p}_1$
E-step; (See Algorithm 1)
for $\tilde{Q} = 2$ **to** Q **do**
  Add node center to $\mathbf{p}$: $\mathbf{p} \leftarrow \left[\mathbf{p}^T\ \mathbf{p}_{\tilde{Q}}^T\right]^T$.
  for $\ell = 1$ **to** L **do**
    M-step (See Algorithm 1 **M-step**)
    E-step (See Algorithm 1 **E-step**)
  end
end
Find J, the number of speakers, and their positions $\mathbf{p}_j$ $\forall j \in [1, J]$ using a threshold for $\psi_m^{(L)}$.

After finalizing all iterations of the last node, the number of speakers J and their positions $\mathbf{p}_j$ $\forall j \in [1, J]$ are determined by applying a threshold to the probability map $\psi_m^{(L)}$. The threshold is applied in the way it has been suggested for iterative localization after algorithm convergence [35, 75–78]. The rationale is to keep the soft values during the EM convergence and apply the threshold only at the end.

2.2.5 Algorithm's complexity

The complexity of the proposed algorithm is high, even though we apply the calibration of each array sequentially, as described above. The complexity is a function of a few parameters. For example, it is very important to choose the correct grid resolution in the room to guarantee proper localization accuracy. However, the trade-off between accuracy and computational burden should be taken into consideration. In Table 1, the relevant param-

Table 1 Implementation parameters

Notation	Meaning
L	Number of algorithm iterations
Q	Number of nodes
T	Number of time frames
K	Number of frequency bins
M	Size of the grid

eters are listed. These parameters were already defined above during the derivation of the algorithm equations. The resources consumed by the proposed algorithm are summarized in Table 2 in terms of computational complexity, communication bandwidth (BW), and memory requirements. Due to the distributed nature of the problem at hand, these resources can be shared by the nodes, thus increasing the algorithm's efficacy. For example, we can start locally from the anchor node and then share the results with the second node and so on. The details depends on the network topology, which is beyond the scope of this paper.

3 Results and discussion

The proposed algorithm was evaluated using both simulations and real recordings. The performance of the proposed algorithm was evaluated in terms of both node calibration accuracy and concurrent speaker localization. The simulation and recording setups are described in the first subsection. The second subsection summarizes the measures used to evaluate the performance. The simulation results are given in the third subsection. The fourth subsection is about the influence of imperfections on the performance. The fifth subsection is dedicated to the evaluation of the proposed method using real-life recordings. The last subsection introduces a naïve algorithm that might be applied for the same problem. We compare the two approaches in terms of performance and their basic assumptions.

3.1 Experimental setup

For simplicity reasons only, we focus on 2D cases, namely both microphones and sources are located at the same height. The 3D cases imposes high computational complexity and will be therefore skipped in this manuscript. In addition, to avoid too strong reflections from either the floor or the ceiling of the acoustic enclosure, we have selected the height of the sources-microphone constellation in the center of the z-axis. The experimental setup for the simulation study and the real-life recordings were designed to be as similar as possible. Accordingly, the speakers were positioned to imitate a group of people sitting around a table located in the center of the room. Three to five microphone arrays were located randomly in the center area of the room to emulate mobile telecommunication devices that are located on that virtual table, each of which with a few microphones. This geometry also simulate an outdoor scenario for which the sensors are restricted to be located within a close area and the sources are located in the perimeter of this area.

Table 2 Implementation complexity table for the localization and calibration EM sequence algorithm

Criteria	LACES algorithm
Computation	$\mathcal{O}(L \cdot Q^2 \cdot T \cdot K \cdot M)$
BW	$\mathcal{O}(L \cdot Q \cdot T \cdot K \cdot M)$
Memory	$\mathcal{O}(Q \cdot T \cdot K \cdot M)$

The nodes jointly constitute an ad hoc acoustic sensor network. The nodes are rectangular with four microphones each, simulating smartphones with known dimensions and orientations. An example of such an array is shown in Fig. 1.

The sampling frequency was set to 8 kHz and the frame length of the STFT to 64 ms with an overlap of 75%. The number of frequency bins was 512. Utterances of simultaneously active male and female speakers were used (signals lengths is 1 s). The speakers were located randomly around the table. The number of speakers was five for the simulations and six for the real recordings.

The frequency band that was proven sufficient for our array sizes was $500 - 2000$ Hz. In the simulations, the speech signals were convolved with simple room impulse responses (RIRs) of an anechoic chamber. In the real-life recordings, we recorded the signals in our acoustic lab, set to a low reverberation level ($T_{60} = 120$ ms). In both cases, a synthetic additive white Gaussian noise (AWGN) was added with various signal to noise ratio (SNR) levels.

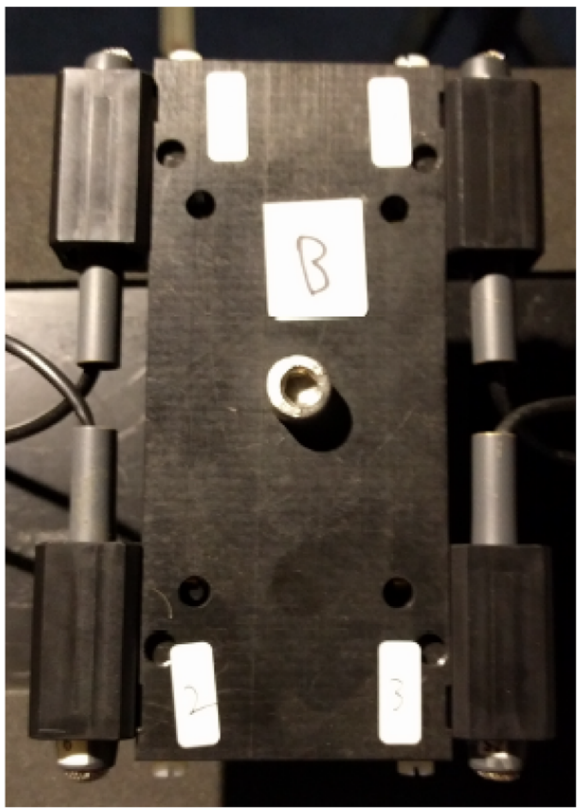


Fig. 1 Cellular phone form-factor array with four omnidirectional AKG CK32 microphones at the corners

A picture depicting the recordings setup can be found in Fig. 2. The rectangular arrays mentioned above were used in the acoustic lab together with Fostex model 6301BX loudspeakers, serving as sources. A high-quality recording system (by RME) was used to measure the T_{60} and to generate the input signals. Although the full size of the room was $6 \times 6 \times 2.4$ m, here, we focus on a smaller search area of 5×5 m with a constant height of 135 cm.

3.2 Performance measures

Calibration success rate (SR) was calculated using Monte-Carlo simulations according to the number of times the estimation of the node center was sufficiently accurate (up to 20 cm):

$$\mathrm{SR}(\%) = 100 * S_c/A_e, \tag{38}$$

where S_c is the number of successful calibrations and A_e is the total number of nodes to be calibrated. This is the only measure used for the calibration stage. If the calibration is sufficiently accurate, then the calibration error in centimeters will be very good; if the calibration fails, the results of subsequent localization stage also fails.

For the localization stage, we adapted three statistical measures used in [35, 75]. They are only calculated for the cases of successful calibration. The misdetections (MDs) are counted according to the percentage of misdetected speakers:

$$\mathrm{MD}(\%) = 100 * M_s/R_s, \tag{39}$$

where M_s is the number of misdetected sources and R_s is the total number of real sources.

The false alarm (FA) is the percentage of wrongly detected speakers:

$$\mathrm{FA}(\%) = 100 * F_s/R_s, \tag{40}$$

where F_s is the number of falsely detected sources.

Localization root mean square error (RMSE) is a measure of the estimation accuracy of all detected speakers:

$$\mathrm{RMSE} = \sqrt{\frac{1}{R_s - M_s} \sum_{s=1}^{R_s - M_s} e^2(s)}, \tag{41}$$

where s is the source index and $e(s)$ is its respective localization error in meters.

3.3 Simulations of random geometric setups

The geometric setup for the simulations is shown in Fig. 3. Three nodes with a square shape (10×10 cm) were randomly located with a random orientation in the middle of the room (each microphone is denoted by "$\circ$"). Six speakers (denoted by the "+" sign) were located away from the center to imitate a scenario with nodes in the center (on a table for indoor case) and speakers around that center. The main purpose of the simulation was to explore the performance for random geometric setups. The performance of the algorithm was tested for various levels of SNR and various sensor and source locations. The number of different setups generated was 100.

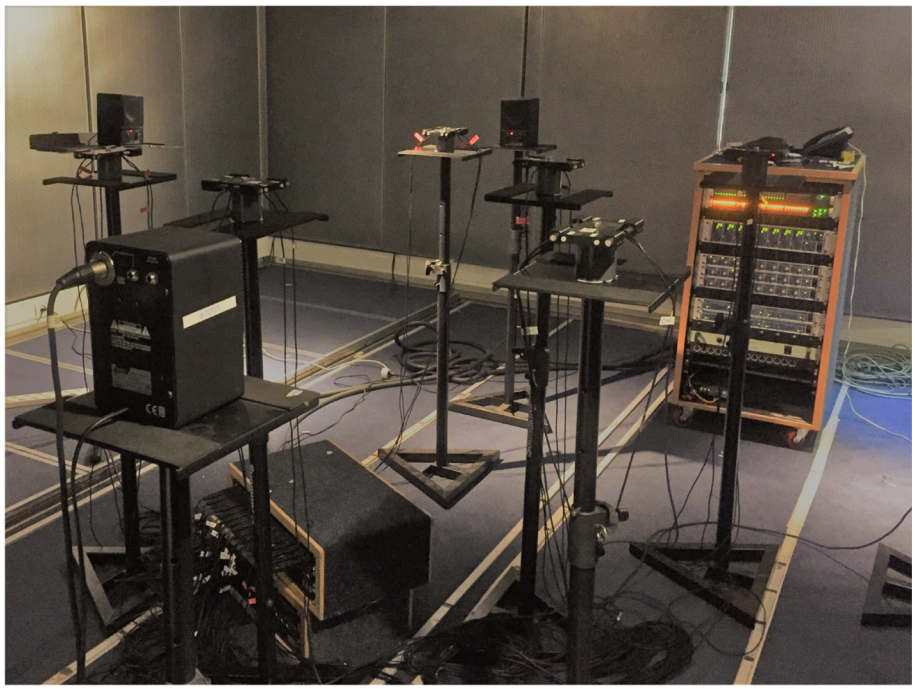


Fig. 2 Room setup example: loudspeakers, microphone arrays, and recording equipment

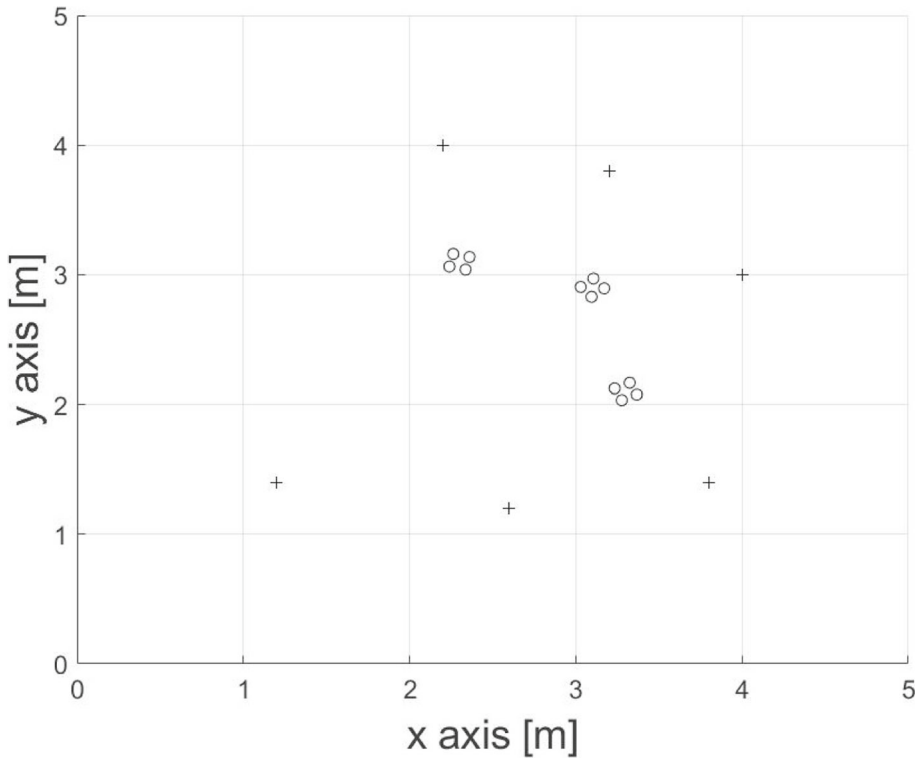


Fig. 3 Simulation room random setup example. Each microphone is denoted by the sign *o*. Six speakers are denoted by the sign +

We noticed that a single EM iteration per new node ($L = 1$) yields satisfactory results. The statistical measures for the simulation study are summarized in Table 3. In the presence of white sensor noise, as also demonstrated for the real recordings, the algorithm performance rapidly deteriorates from good results (for SNRs of 20 dB) to very bad results (around SNRs of 0 dB). Note that the localization search grid is 0.2 m $\times$0.2 m and the localization error is within the grid resolution. We noticed that some compensation for low SNR could be achieved, if we add microphones to each array as long as the noise is spatially white. However, a detailed analysis of how the number of microphones might affect the performance is beyond the scope of this contribution.

To experimentally examine the LACES convergence when arrays are added to the estimation, we plotted the intermediate results for the localization parameters, $\boldsymbol{\psi}$ in Fig. 4 for $L = 1$. The real locations of the five speakers are marked by '+'.

Table 3 Statistical measures for various SNR levels

Sensor SNR (dB)	Calib. SR (%)	MD (%)	FA (%)	Loc. RMSE [m]
0	45.5	54	-	-
10	70.5	16	1	0.16
20	71.5	6	1	0.16
40	74	6	1	0.16
60	74.5	6	1	0.16

The node calibration SR is measured in percentage (%). The source localization performance measures are MD percentage, FAs percentage, and RMSE in meters

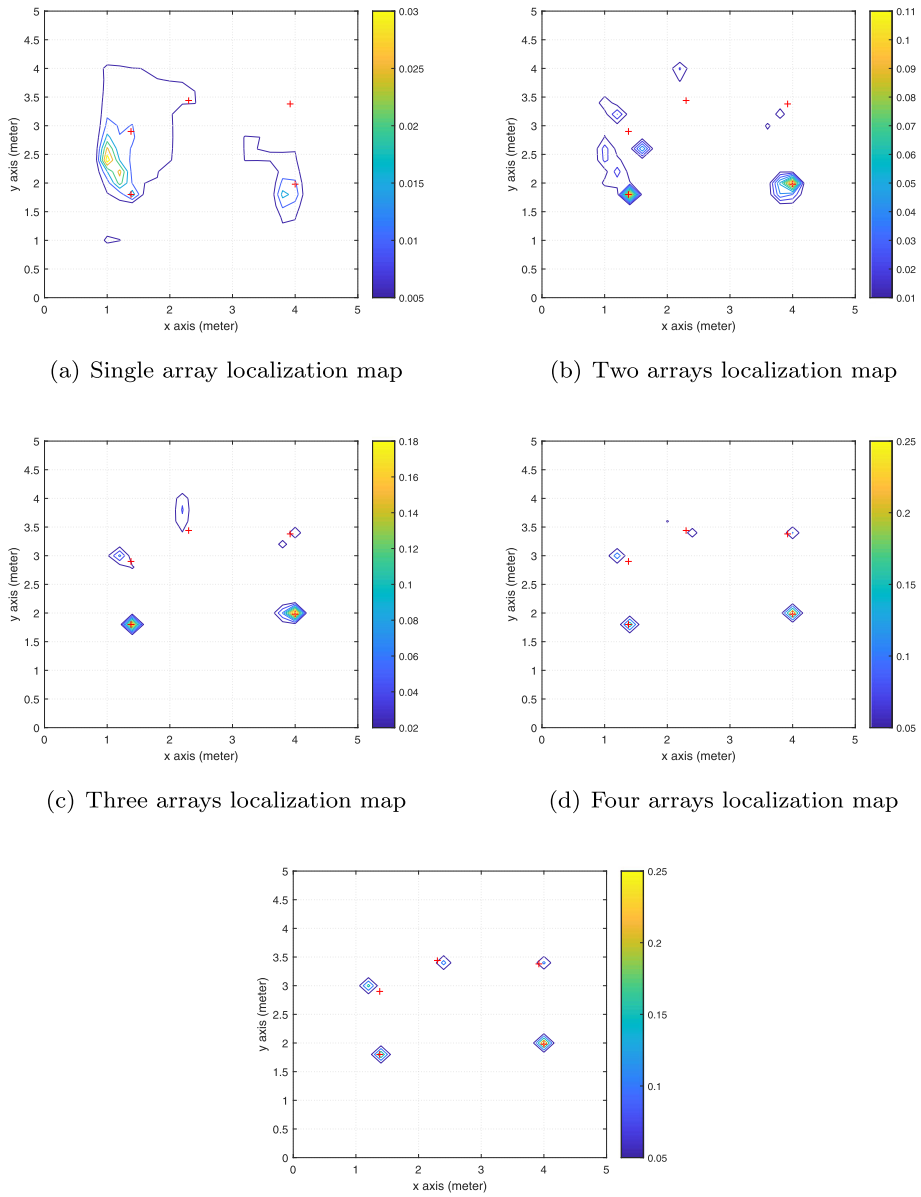
The improvement of the localization maps can be observed when additional arrays are utilized. For a single array, only a few of the speakers are detected and many errors are observed. As arrays are added, the estimation improves for all speakers. The final map can be used to infer the number of speakers and their locations.

3.4 Sensitivity to imperfections

Before discussing real recordings, it is essential to examine what is the sensitivity of the algorithm to imperfections that exist in any real system.

The first one is sensitivity to inaccurate offset values of the microphones with respect to the center of the array. We use a uniform distribution with various maximal offset. The performance of the algorithm is summarized in Table 4. In the presence of errors in microphones locations, the algorithm performance rapidly deteriorates from good results (for maximal offset of 10 mm) to very bad results (around offset of 20 mm). Seems realistic to assume internal calibration accuracy of around 1 mm, which seems to be high enough in terms of the algorithm performance.

The second analysis is sensitivity to synchronization issues between arrays. We examine the influence of clock



(a) Single array localization map

(b) Two arrays localization map

(c) Three arrays localization map

(d) Four arrays localization map

(e) Five arrays localization map

Fig. 4 Localization soft maps intermediate results (**a**–**e**). The real locations of the simulated speakers are marked by '+'. The estimation is given by colored contours. The grid resolution is 20 cm. We excluded strips of 100 cm near the walls from the search area

Table 4 Measures for various internal calibration errors

Max offset (mm)	Calib. SR (%)	MD (%)	FA(%)	Loc. RMSE [m]
0	100	0	0	0.117
10	100	0	0	0.117
20	50	33	0	0.411
30	0	100	67	-

The node calibration SR is measured in percentage (%). The source localization performance measures are MD percentage, FAs percentage, and RMSE in meters

rate differences between arrays. We use a constant frequency offset between the three arrays, which is measured compared to the anchor array in parts per million (ppm) units. One array has maximal offset as indicated in the table and the other one has half the offset. The performance of the algorithm is summarized in Table 5. In the presence of very large frequency offsets, the algorithm performance rapidly deteriorates from good results (for maximal offset of 100 ppm) to very bad results (around an offset of 1000 ppm). It means that even for very low quality of internal clocks, the performance is still satisfactory.

Table 5 Measures for various frequency offsets between arrays

Frequency offset ppm	Calib. SR (%)	MD(%)	FA (%)	Loc. RMSE [m]
0	100	0	0	0.117
100	100	0	0	0.117
250	100	33	0	0.118
1000	100	33	0	0.314
10000	0	100	83	-

The node calibration SR is measured in percentage (%). The source localization performance measures are MD percentage, FAs percentage, and RMSE in meters

The last analysis is the sensitivity to the reverberation level of the room. As stated above, we assume low reverberation levels, since we observed significant influence on the performance. The performance of the algorithm as a function of the reverberation level is summarized in Table 6. As expected, when reverberation increases, the algorithm performance rapidly deteriorates from good results ($T_{60} = 100$ ms) to very bad results ($T_{60} = 300$ ms).

3.5 Real-life recordings in low-reverberation indoor environment

The geometric setup for the real recordings taken at BIU acoustic lab is depicted in Fig. 5. Three arrays with a rectangular shape (8.2 × 14.7 cm) were located in the middle, each of which consists four microphones. Each microphone in the scheme is denoted by the symbol "○." Six speakers, denoted by the symbol "+," were located around the center in a meeting room setup. The real recordings are characterized by low reverberation level ($T_{60} = 120$ ms). We tested this array constellation with various levels of sensor AWGN. The analysis of the real recordings is therefore focused on the influence of the SNR level, rather than the reverberation level, on the calibration and localization accuracy. These acoustic conditions can also represent outdoor environments, which are usually characterized by a small number of reflections . We analyze a single scenario in this subsection with signals of the same length used above in the simulated subsection. Table 7 summarizes the results for various SNR conditions. In the Calibration SR column, we designate the number of correctly calibrated arrays out of 2 arrays (the third array is the anchor array). MD is calculated for 6 speakers.

It can be seen that for any SNR higher than 14 dB, the performance is very good: the calibration was good for the nodes, the number of MDs was zero, there were no FAs, and the localization RMSE was 0.1 m. For an SNR of 10 dB, there is some degradation in the localization results, but the calibration is still good. The algorithm fails for all SNR levels equal to or below 3 dB.

Table 6 Measures for various reverberation levels

T_{60} ms	Calib. SR (%)	MD (%)	FA (%)	Loc. RMSE [m]
100	100	0	0	0.117
120	100	0	17	0.126
200	50	50	33	0.517
300	0	100	50	-

The node calibration SR is measured in percentage (%). The source localization performance measures are MD percentage, FAs percentage, and RMSE in meters

3.6 Naïve algorithm

In this subsection, as a comparison to the proposed method, we introduce a naïve geometrical technique for estimating both the array centers $\mathbf{p}_q$ for $q = 2,\ldots,Q$ (assuming the reference array position $\mathbf{p}_1$ is known) and the speakers' positions $\mathbf{p}_j$ for $j = 1,\ldots,J$, with J the number of speakers.

Two simplifying assumptions are first made: (1) the number of speakers J is known in advance and (2) the speakers' activity patterns are non-overlapping and the time-segments in which they are active are known as well. Note that the LACES algorithm does not require these simplifying assumptions, that are rarely met in real-life scenarios.

The naïve algorithm uses two datasets: (1) $\tau_{q,j}$—the TDoA between each array centroid and the reference array centroid w.r.t. each speaker; neglecting the TDoAs between the microphones within each array, the TDoA is estimated by maximizing the cross-correlation between each possible pair of signals (one from each array and one from the reference array) and average all the obtained TDoAs—and (2) $\vartheta_{q,j}$—the DOA of each speaker w.r.t. each array. The DOA is estimated by maximizing the SRP steered to all possible DOAs. Note that the orientation of the arrays are known (same as for the LACES algorithm), and hence, the independently estimated DOAs are all referring to the same coordinate system.

The positions of the speakers and arrays should match the TDoA readings between the arrays. Accordingly, the TDoA between the qth array centroid and the reference array centroid (namely, array #1) is given by $\frac{\|\mathbf{p}_j-\mathbf{p}_q\|-\|\mathbf{p}_j-\mathbf{p}_1\|}{c}F_s$, with c the sound velocity and F_s the sampling frequency. Using the observed TDoAs $\tau_{q,j}$, the following cost function should be minimized to obtain an estimate of the positions of the arrays' centroids $\mathbf{p}_q$; $q = 2,\ldots,Q$ and the speakers' positions $\mathbf{p}_j$; $j = 1,\ldots,J$:

$$\sum_{q=2}^{Q}\sum_{j=1}^{J}\left|\frac{\|\mathbf{p}_j-\mathbf{p}_q\|-\|\mathbf{p}_j-\mathbf{p}_1\|}{c}F_s-\tau_{q,j}\right|^2. \tag{42}$$

As this cost function in (42) includes both the arrays' and speakers' positions, the search for a global minimum is a cumbersome task.

The positions of the sources and the arrays should also satisfy the relations imposed the DOAs $\vartheta_{q,j}$ between the arrays and the sources. Considering only the horizontal plain, the following relation must hold:

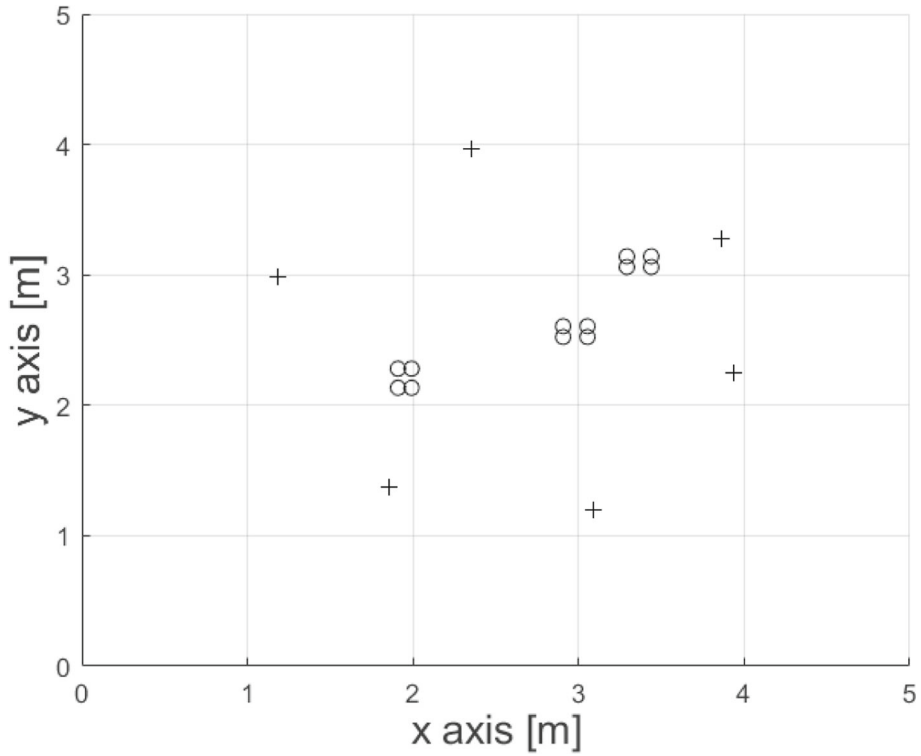


Fig. 5 Recordings room setup. Each microphone is denoted by the sign *o*. Six speakers are denoted by the sign +

$$\left[\sin(\vartheta_{q,j}), \quad -\cos(\vartheta_{q,j})\right]^T \left(\mathbf{p}_j - \mathbf{p}_q\right) = 0. \tag{43}$$

Note that this relation has an inherent ambiguity. If a specific $\bar{\vartheta}_{q,j}$ satisfies (43), then also $\bar{\vartheta}_{q,j} + \pi$ satisfies the same equation.

Concatenating the above relations for all arrays $q = 1, \ldots, Q$ yields:

$$\mathbf{A}\mathbf{p}_j = (\mathbf{A} \circ \mathbf{B}) \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{44}$$

where $\mathbf{A}$ and $\mathbf{B}$ are $Q \times 2$ matrices defined by $\mathbf{A}_{q,1:2} = \left[\sin(\vartheta_{q,j}), \quad -\cos(\vartheta_{q,j})\right]$ and $\mathbf{B}_{q,1:2} = \mathbf{p}_q^T$. The symbol $\circ$ denotes the Hadamard product (element-wise product). Equation 44 is an over-determined set of equations for $\mathbf{p}_j$, provided that $Q \geq 2$, and hence can be solved by applying the least squares procedure. The position of the jth speaker $\mathbf{p}_j$, as a function of the arrays' positions is then given by:

Table 7 Measures for room recordings in various SNR conditions

SNR (dB)	Calib. SR	MDs	FAs	Loc. RMSE [m]
$\geq$ 14	2/2	0/6	0	0.1
10	2/2	1/6	0	0.1
$\leq$ 3	0/2	N/A	N/A	N/A

The node calibration SR is given as a ratio. The source localization performance measures are MD ratio, FA ratio, and RMSE in meters

$$\widehat{\mathbf{p}}_j = \left(\mathbf{A}^T\mathbf{A}\right)^{-1} \mathbf{A}^T (\mathbf{A} \circ \mathbf{B}) \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{45}$$

Substituting (45) into the cost function in (42), the array positions $\mathbf{p}_q; \ q = 2, \ldots, Q$ can now be estimated independently of the speakers' positions, thus alleviating the computational burden:

$$\widehat{\mathbf{p}}_2, \ldots, \widehat{\mathbf{p}}_Q = \text{argmin}_{\mathbf{p}_2,\ldots,\mathbf{p}_Q} \sum_{q=2}^{Q} \sum_{j=1}^{J} \left| \frac{\|\widehat{\mathbf{p}}_j - \mathbf{p}_q\| - \|\widehat{\mathbf{p}}_j - \mathbf{p}_1\|}{c} F_s - \tau_{q,j} \right|^2. \tag{46}$$

The cost function in (46) still requires a $(Q-1)$-dimensional search. To further reduce the complexity, we propose to sequentially minimize the cost function for the sub-network of size $\tilde{Q}$, with $\tilde{Q} = 2, \ldots, Q$. At each step, only the position of the newly added array $\mathbf{p}_{\tilde{Q}}$ is estimated, while all previous arrays' positions $\mathbf{p}_2, \ldots, \mathbf{p}_{(\tilde{Q}-1)}$, that were estimated in the previous algorithmic steps, are kept unaltered:

$$\hat{\mathbf{p}}_{\tilde{Q}} = \text{argmin}_{\mathbf{p}_{\tilde{Q}}} \sum_{q=2}^{\tilde{Q}} \sum_{j=1}^{J} \left| \frac{\|\widehat{\mathbf{p}}_j - \mathbf{p}_q\| - \|\widehat{\mathbf{p}}_j - \mathbf{p}_1\|}{c} F_s - \tau_{q,j} \right|^2. \tag{47}$$

The 1-dimensional minimization can now be carried out by a simple grid search. We chose an area of 5×5 m sur-

rounding the reference array with a resolution of 0.05 m, to obtain a similar search domain as for the LACES algorithm. The number of candidate positions is denoted M and is approximately equal 10,000 in this case. The naïve geometrical technique is summarized in Algorithm 3.

Algorithm 3: Naïve geometrical technique for joint speakers' and arrays' positions estimation.

Input: Obtain an estimate of the TDoA $\tau_{q,j}$ between each array centroid and the reference array centroid w.r.t. each speaker
Input: Obtain an estimate of the DOAs $\vartheta_{q,j}$ of all speakers w.r.t. all arrays
for $\tilde{Q} = 2$ **to** Q **do**
  for *Each of the M candidate positions of the* $\tilde{Q}$*-th array* **do**
    Minimize the cost-function (47) w.r.t. $\mathbf{p}_{\tilde{Q}}$
    Correct the ambiguity problem for $\tilde{Q} = 2$, if necessary
  end
end
Output: Array positions $\hat{\mathbf{p}}_2, \ldots, \hat{\mathbf{p}}_Q$
Output: Determine the speakers' positions $\hat{\mathbf{p}}_1, \ldots, \hat{\mathbf{p}}_J$ using (45) and the estimated array positions

To exemplify the procedure, the case of six speakers and three arrays, as depicted in Fig. 6, is examined. The reference array is located at $[2.4, 2.6]$ m and its position is not estimated by the algorithm. In the first stage, only the position of the second array, located at $[1.8, 3]$ m, is estimated. The obtained cost function (47) is depicted in Fig. 7a. Two distinct minima, in $[1.8, 2.95]$ and $[3, 2.25]$, can be observed. This is attributed to the symmetric behavior of the cost function (47) w.r.t. the reference array, namely for $\mathbf{p}_1 + \mathbf{p}_2$ and $\mathbf{p}_1 - \mathbf{p}_2$, as evident from (43). Therefore, an additional disambiguity stage was applied to determine the second array position. For that, we calculated two alternative DOA estimates from the two optional array positions (either $[1.8, 2.95]$ or $[3, 2.25]$) towards the estimated position of an arbitrarily chosen speaker $\bar{j}$, using (45). The two values were compared to the observed DOA $\vartheta_{2,\bar{j}}$. Since $[1.8, 2.95]$ better fits the observed DOA than the alternative candidate $[3, 2.25]$, it was finally chosen as the position of the second array.

Next, the position of the third array ($[3, 2.2]$) was estimated using the known position of the first array and the already estimated position of the second array. The obtained cost function, which does not suffer from the above ambiguity, is depicted in Fig. 7b, and its minimum is obtained in $[3.0, 2.2]$. The final estimated positions of the arrays and the speakers versus the oracle positions are

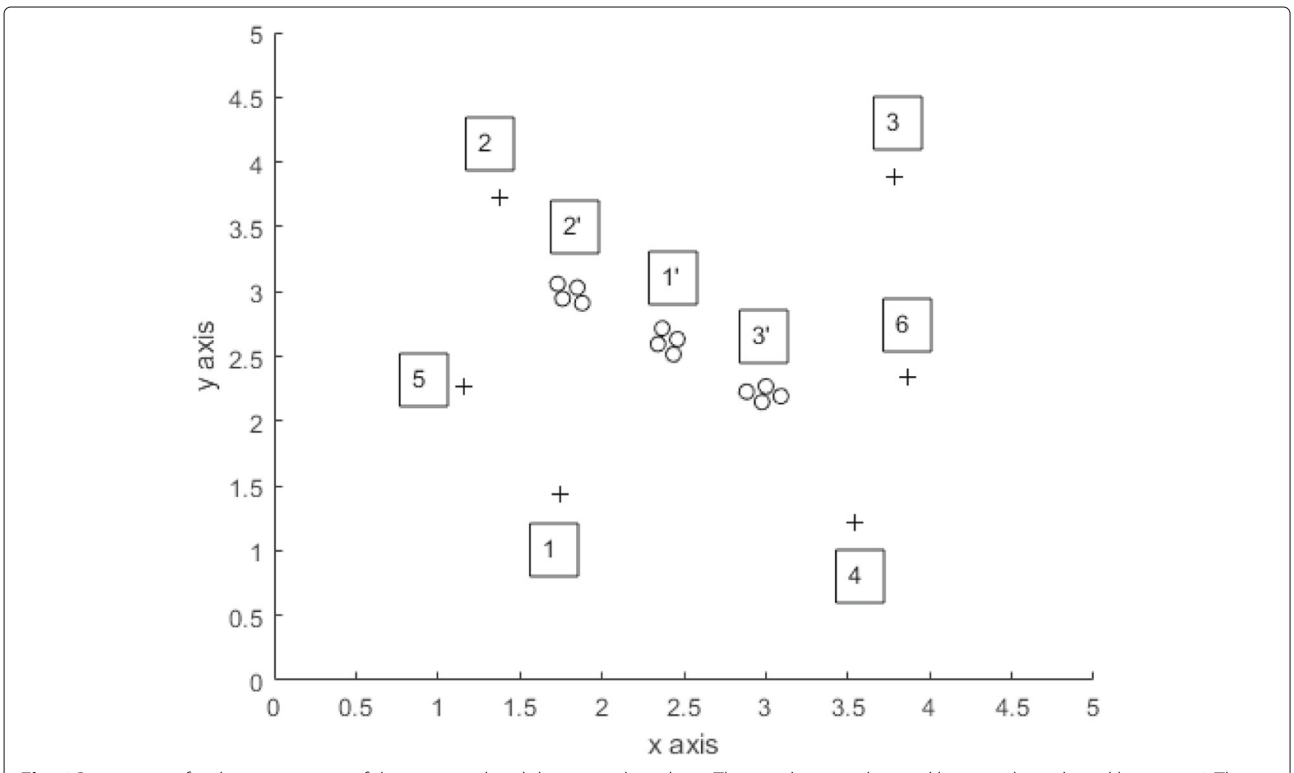


Fig. 6 Room setup for the comparison of the proposed and the naïve algorithms. The speakers are denoted by + and numbered by $1, \ldots, 6$. The microphones are denoted by *o*. The arrays are numbered by tagged numbers $1', 2', 3'$

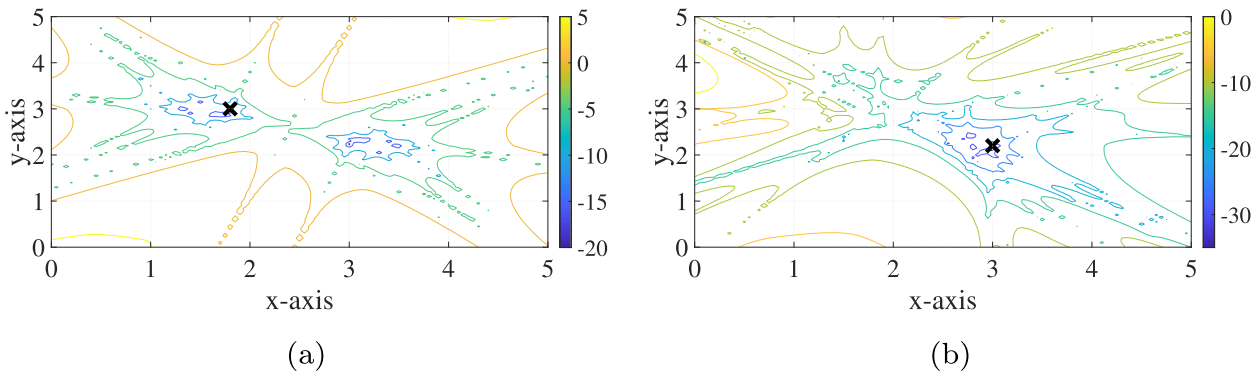


Fig. 7 Contour of the cost function in (47). The oracle position is denoted by 'x'

depicted in Fig. 8. The averaged estimation error of the speakers is 5.5 cm. The estimation error in localizing the second array is 0.05 cm, while the position of the third array is accurately estimated.

The same geometrical setup was used to evaluate the LACES algorithm, but with a more realistic scenario, where the number of speakers unknown and their activity overlapping. The position of the array centroids were accurately found (namely, negligible estimation error) and the average estimation error in localizing the speakers is 11.7 cm. The final localization map is depicted in Fig. 9. The obtained speakers positions are marked with a heat map and the real locations marked with black +.

4 Conclusions

A major challenge for ad hoc networks is to jointly localize sources and calibrate the positions of the arrays (or nodes) of the network. A novel joint calibration and localization algorithm, suitable for noisy environments, was derived using the EM algorithm. One of the nodes is used as an anchor node. The calibration, i.e., the estimation of the node positions, as well as the speakers' localization are applied relatively to the position of this anchor node.

To alleviate the initialization challenge of the batch EM, an incremental procedure was proposed that sequentially adds the nodes rather than trying to concurrently solve the entire full-dimension problem. The new algorithm, dubbed LACES algorithm, was experimentally studied using both an intensive simulated study and real recordings. It was also compared with a naïve algorithm based on geometrical considerations. While exhibiting high localization accuracy for both the nodes and the speakers in the case of non-overlapping speakers and known number of speakers, the naïve algorithm renders useless in realistic scenarios for which these simplifying assumptions do not hold. The proposed LACES algorithm maintains

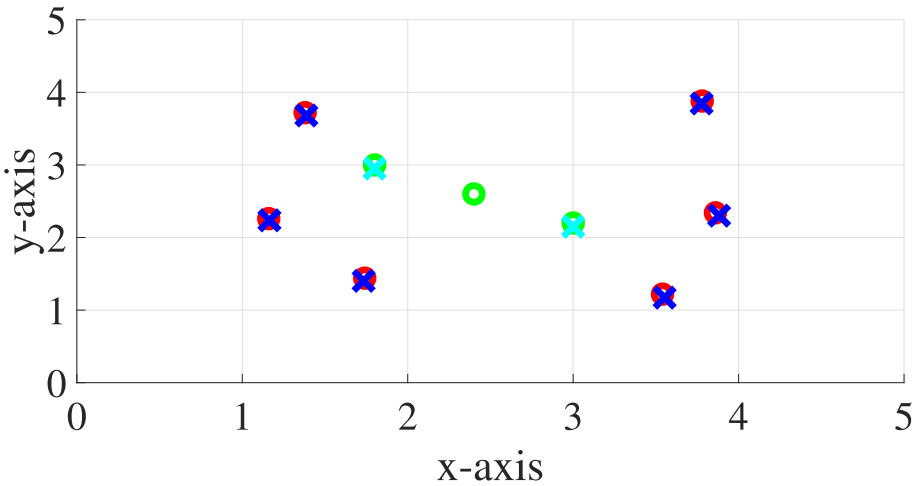


Fig. 8 The final estimated positions of the arrays and the speakers vs. the oracle positions.The oracle positions of the arrays are denoted by green circles and the estimated positions are denoted by cyan "x"s. The oracle positions of the speakers are denoted by red circles and the estimated positions are denoted by blue "x"s

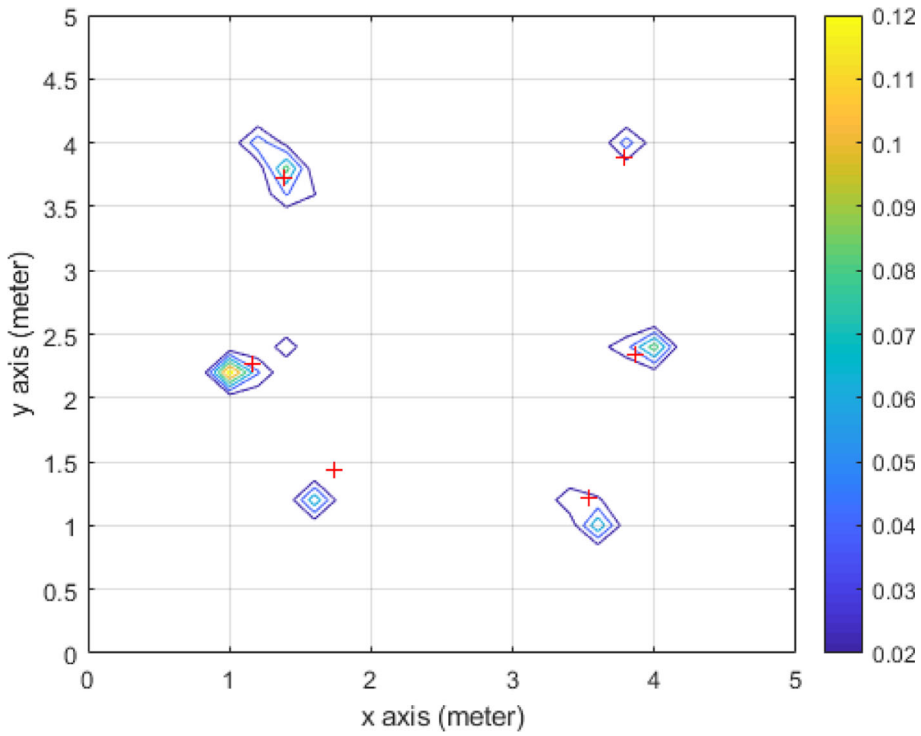


Fig. 9 The localization heat map LACES result. The final speaker positions are also added for evaluation

high localization and calibration accuracy even in these challenging scenarios.

Abbreviations

AWGN: Additive white Gaussian noise; BW: Bandwidth; DOA: Direction of arrival; DPD: Direct positioning determination; DTF: Direct transfer function; EM: Expectation-maximization; FA: False-alarm; GCC: Generalized cross correlation; LACES: Localization and calibration EM sequence; MD: Misdetection; ML: Maximum likelihood; MLE: Maximum likelihood estimation; MoG: Mixture of Gaussians; MUSIC: Multiple signal classification; MVDR: Minimum variance distortionless response; RMSE: Root mean square error; p.d.f.: Probability density function; PHAT: Phase transform; ppm: Parts per million; PSD: Power spectral density; REM: Recursive EM; RIR: Room impulse response; SLAM: Simultaneous localization and mapping; SLAT: Simultaneous localization and tracking; SNR: Signal to noise ratio; SR: Success rate; SRP: Steered response power; STFT: Short-time Fourier transform; TDoA: Time difference of arrival; TF: Time-frequency; w.r.t.: With respect to; MCWF: Multichannel Wiener filter; WDO: W-disjoint orthogonality

Acknowledgements

We would like to thank Mr. Pini Tandeitnik for his professional assistance during the acoustic room setup and the recordings.

Authors' contributions

Model development: YD, OS, and SG . Experimental testing: YD. Writing paper: YD, OS ,and SG. The authors read and approved the final manuscript.

Funding

N/A

Availability of data and materials

N/A

Consent for publication

All authors agree to the publication in this journal.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Faculty of Engineering, Bar-Ilan University, 5290002 Ramat-Gan, Israel. [2]Audio department, CEVA DSP, 4612001 Herzliya, Israel.

Received: 8 November 2019 Accepted: 17 May 2020
Published online: 09 June 2020

References

1. G. Lathoud, J.-M. Odobez, D. Gatica-Perez, in *International Workshop on Machine Learning for Multimodal Interaction*. AV16.3: an audio-visual corpus for speaker localization and tracking (Springer, 2004), pp. 182–195. https://doi.org/10.1007/978-3-540-30568-2_16
2. T. Yamada, S. Nakamura, K. Shikano, in *Fourth IEEE International Conference on Spoken Language, vol. 3*. Robust speech recognition with speaker localization by a microphone array, (1996), pp. 1317–1320. https://doi.org/10.1109/icslp.1996.607855
3. S. Doclo, M. Moonen, Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments. EURASIP J. Appl. Sig. Process. **2003**, 1110–1124 (2003)
4. O. Schwartz, S. Gannot, Speaker tracking using recursive EM algorithms. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(2), 392–402 (2014)
5. N. Madhu, R. Martin, in *Proceedings of the International Workshop on Acoustic Echo Cancellation and Noise Control (IWAENC)*. A scalable framework for multiple speaker localization and tracking, (2008)
6. E. A. P. Habets, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) vol. 4*. Multi-channel speech dereverberation based on a statistical model of late reverberation, (2005), pp. 173–176. https://doi.org/10.1109/icassp.2005.1415973
7. A. Kuklasinski, S. Doclo, S. H. Jensen, J. Jensen, in *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*. Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids, (2014), pp. 61–65

8. O. Schwartz, S. Gannot, E. A. P. Habets, Multi-microphone speech dereverberation and noise reduction using relative early transfer functions. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(2), 240–251 (2015)
9. D. P. Morgan, E. B. George, L. T. Lee, S. M. Kay, Cochannel speaker separation by harmonic enhancement and suppression. IEEE Trans. Speech Audio Process. **5**(5), 407–424 (1997)
10. A. M. Reddy, B. Raj, Soft mask methods for single-channel speaker separation. IEEE Trans. Audio Speech Lang. Process. **15**(6), 1766–1776 (2007)
11. B. Raj, P. Smaragdis, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (ICASSP)*. Latent variable decomposition of spectrograms for single channel speaker separation, (2005), pp. 17–20. https://doi.org/10.1109/aspaa.2005.1540157
12. Y. Dorfan, O. Schwartz, B. Schwartz, E. A. P. Habets, S. Gannot, in *International Conference on the Science of Electrical Engineering (ICSEE)*. Multiple DOA estimation and blind source separation using expectation-maximization algorithm, (Eilat, Israel, 2016). https://doi.org/10.1109/icsee.2016.7806066
13. O. Schwartz, S. Braun, S. Gannot, E. A. P. Habets, in *International Conference on Latent Variable Analysis and Signal Separation*. Source separation, dereverberation and noise reduction using LCMV beamformer and postfilter (Springer, 2017), pp. 182–191. https://doi.org/10.1007/978-3-319-53547-0_18
14. J. H. DiBiase, H. F. Silverman, M. S. Brandstein, Robust localization in reverberant rooms. Microphone arrays: signal processing techniques and applications, 157–180
15. C. Knapp, G. Carter, The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. **24**(4), 320–327 (1976)
16. R. Schmidt, Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986)
17. V. Vasylyshyn, Removing the outliers in root-music via pseudo-noise resampling and conventional beamformer. Sig. Process. **93**(12), 3423–3429 (2013)
18. D. Rahamim, J. Tabrikian, R. Shavit, Source localization using vector sensor array in a multipath environment. IEEE Trans. Signal Process. **52**(11), 3096–3103 (2004)
19. A. Herzog, E. A. Habets, in *2019 27th European Signal Processing Conference (EUSIPCO)*. On the relation between doa-vector eigenbeam esprit and subspace pseudointensity-vector (IEEE, 2019), pp. 1–5. https://doi.org/10.23919/eusipco.2019.8902715
20. A. Herzog, E. A. Habets, Eigenbeam-ESPRIT for DOA-vector estimation. IEEE Sig. Process. Lett. **26**(4), 572–576 (2019)
21. H. Teutsch, W. Kellermann, in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference On, vol. 3*. EB-ESPRIT: 2D localization of multiple wideband acoustic sources using eigen-beams (IEEE, 2005), p. 89. https://doi.org/10.1109/icassp.2005.1415653
22. R. Wang, Z. Chen, F. Yin, DOA-based three-dimensional node geometry calibration in acoustic sensor networks and its Cramér–Rao bound and sensitivity analysis. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(9), 1455–1468 (2019)
23. S. A. Vorobyov, A. B. Gershman, K. M. Wong, Maximum likelihood direction-of-arrival estimation in unknown noise fields using sparse sensor arrays. IEEE Trans. Signal Process. **53**(1), 34–43 (2005)
24. H. Ye, R. D. DeGroat, Maximum likelihood DOA estimation and asymptotic Cramér-Rao bounds for additive unknown colored noise. IEEE Trans. Signal Process. **43**(4), 938–949 (1995)
25. K. Yao, J. C. Chen, R. E. Hudson, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 3*. Maximum-likelihood acoustic source localization: experimental results, (2002), pp. 2949–2952. https://doi.org/10.1109/icassp.2002.1005305
26. H. Wang, C.-E. Chen, A. Ali, S. Asgari, R. E. Hudson, K. Yao, D. Estrin, C. Taylor, in *Proc. of SPIE, Advanced Signal Processing Algorithms, Architectures, and Implementations*. Acoustic sensor networks for woodpecker localization, (2005). https://doi.org/10.1117/12.617983
27. O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Sig. Process. **52**(7), 1830–1847 (2004)
28. S. Araki, H. Sawada, R. Mukai, S. Makino, DOA estimation for multiple sparse sources with arbitrarily arranged multiple sensors. J. Sig. Process. Syst. **63**(3), 265–275 (2011)
29. M. I. Mandel, D. P. W. Ellis, T. Jebara, An EM algorithm for localizing multiple sound sources in reverberant environments. Adv. Neural Inf. Process. Syst. **19**, 953–960 (2007)
30. O. Schwartz, Y. Dorfan, E. A. P. Habets, S. Gannot, in *International Workshop for Acoustic Echo Cancellation and Noise Control (IWAENC)*. Multiple DOA estimation in reverberant conditions using EM, (Xi'an, China, 2016)
31. O. Schwartz, Y. Dorfan, M. Taseska, E. A. P. Habets, S. Gannot, in *Hands-free speech communications and microphone arrays (HSCMA)*. DOA estimation in noisy environment with unknown noise power using the EM algorithm, (2017), pp. 86–90. https://doi.org/10.1109/hscma.2017.7895567
32. J. C. Chen, K. Yao, R. E. Hudson, Source localization and beamforming. IEEE Sig. Process. Mag. **19**(2), 30–39 (2002)
33. A. Griffin, A. Alexandridis, D. Pavlidi, Y. Mastorakis, A. Mouchtaris, Localizing multiple audio sources in a wireless acoustic sensor network. Sig. Process. **107**, 54–67 (2015)
34. A. J. Weiss, A. Amar, Direct position determination of multiple radio signals. EURASIP J. Adv. Signal Process. **2005**(1), 37–49 (2005)
35. Y. Dorfan, S. Gannot, Tree-based recursive expectation-maximization algorithm for localization of acoustic sources. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(10), 1692–1703 (2015)
36. Y.-C. Wu, Q. Chaudhari, E. Serpedin, Clock synchronization of wireless sensor networks. IEEE Sig. Process. Mag. **28**(1), 124–138 (2011)
37. L. Schenato, F. Fiorentin, Average timesync: a consensus-based protocol for time synchronization in wireless sensor networks1. IFAC Proc. Vol. **42**(20), 30–35 (2009)
38. Q. M. Chaudhari, E. Serpedin, K. Qaraqe, On maximum likelihood estimation of clock offset and skew in networks with exponential delays. IEEE Trans. Sig. Process. **56**(4), 1685–1697 (2008)
39. W. Su, I. F. Akyildiz, Time-diffusion synchronization protocol for wireless sensor networks. IEEE/ACM Trans. Netw. **13**(2), 384–397 (2005)
40. S. Wehr, I. Kozintsev, R. Lienhart, W. Kellermann, in *IEEE Sixth International Symposium on Multimedia Software Engineering*. Synchronization of acoustic sensors for distributed ad-hoc audio networks and its use for blind source separation, (2004), pp. 18–25. https://doi.org/10.1109/mmse.2004.79
41. S. Markovich-Golan, S. Gannot, I. Cohen, in *Internation Workshop on Acoustic Signal Enhancement (IWAENC)*. Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming, (2012)
42. S. Miyabe, N. Ono, S. Makino, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain, (2013), pp. 674–678
43. Y. Zeng, R. C. Hendriks, N. D. Gaubitch, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. On clock synchronization for multi-microphone speech processing in wireless acoustic sensor networks, (2015), pp. 231–235. https://doi.org/10.1109/icassp.2015.7177966
44. L. Wang, S. Doclo, Correlation maximization-based sampling rate offset estimation for distributed microphone arrays. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP). **24**(3), 571–582 (2016)
45. D. Cherkassky, S. Gannot, Blind synchronization in wireless acoustic sensor networks. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP). **25**(3), 651–661 (2017)
46. R. Parhizkar, I. Dokmanić, M. Vetterli, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Single-channel indoor microphone localization (IEEE, 2014), pp. 1434–1438. https://doi.org/10.1109/icassp.2014.6853834
47. Y. Rockah, P. Schultheiss, Array shape calibration using sources in unknown locations–part I: Far-field sources. IEEE Trans. Acoust. Speech Sig. Process. **35**(3), 286–299 (1987)
48. Y. Rockah, P. Schultheiss, Array shape calibration using sources in unknown locations–part II: Near-field sources and estimator implementation. IEEE Trans. Acoust. Speech Signal Process. **35**(6), 724–735 (1987)
49. R. L. Moses, D. Krishnamurthy, R. M. Patterson, A self-localization method for wireless sensor networks. EURASIP J. Adv. Signal Process. **2003**(4), 348–358 (2003)
50. S. Zhayida, F. Andersson, Y. Kuang, K. Åström, in *The 22nd European Signal Processing Conference (EUSIPCO)*. An automatic system for microphone self-localization using ambient sound, (2014), pp. 954–958

51. P. Pertilä, M. Mieskolainen, M. S. Hämäläinen, in *The 20th European Signal Processing Conference (EUSIPCO)*. Passive self-localization of microphones using ambient sounds, (2012), pp. 1314–1318
52. H. Durrant-Whyte, T. Bailey, Simultaneous localization and mapping: part I. IEEE Robot. Autom. Mag. **13**(2), 99–110 (2006)
53. T. Bailey, H. Durrant-Whyte, Simultaneous localization and mapping (SLAM): part II. IEEE Robot. Autom. Mag. **13**(3), 108–117 (2006)
54. C. Evers, P. A. Naylor, Acoustic slam. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(9), 1484–1498 (2018)
55. N. Kantas, S. S. Singh, A. Doucet, Distributed maximum likelihood for simultaneous self-localization and tracking in sensor networks. IEEE Trans. Signal Process. **60**(10), 5038–5047 (2012)
56. M. Syldatk, F. Gustafsson, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Simultaneous tracking and sparse calibration in ground sensor networks using evidence approximation, (2013), pp. 3108–3112. https://doi.org/10.1109/icassp.2013.6638230
57. C. Taylor, A. Rahimi, J. Bachrach, H. Shrobe, A. Grue, in *The 5th ACM International Conference on Information Processing in Sensor Networks*. Simultaneous localization, calibration, and tracking in an ad hoc sensor network, (2006), pp. 27–33. https://doi.org/10.1145/1127777.1127785
58. A. Plinge, G. A. Fink, S. Gannot, Passive online geometry calibration of acoustic sensor networks. IEEE Sig. Process. Lett. **24**(3), 324–328 (2017)
59. J. C. Chen, R. E. Hudson, K. Yao, Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field. IEEE Trans. Sig. Process. **50**(8), 1843–1854 (2002)
60. R. Lefort, G. Real, A. Drémeau, Direct regressions for underwater acoustic source localization in fluctuating oceans. Appl. Acoust. **116**, 303–310 (2017). https://doi.org/10.1016/j.apacoust.2016.10.005
61. L. Wang, T.-K. Hon, J. D. Reiss, A. Cavallaro, Self-localization of ad-hoc arrays using time difference of arrivals. IEEE Trans. Sig. Process. **64**(4), 1018–1033 (2016)
62. M. Pollefeys, D. Nister, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Direct computation of sound and microphone locations from time-difference-of-arrival data, (2008), pp. 2445–2448. https://doi.org/10.1109/icassp.2008.4518142
63. V. C. Raykar, I. Kozintsev, R. Lienhart, in *The Eleventh ACM International Conference on Multimedia*. Position calibration of audio sensors and actuators in a distributed computing platform, (2003), pp. 572–581. https://doi.org/10.1145/957013.957133
64. V. C. Raykar, I. V. Kozintsev, R. Lienhart, Position calibration of microphones and loudspeakers in distributed computing platforms. IEEE Trans. Speech Audio Process. **13**(1), 70–83 (2005)
65. A. Plinge, F. Jacob, R. Haeb-Umbach, G. A. Fink, Acoustic microphone geometry calibration: an overview and experimental evaluation of state-of-the-art algorithms. IEEE Sig. Process. Mag. **33**(4), 14–29 (2016)
66. D. Salvati, C. Drioli, G. L. Foresti, Sound source and microphone localization from acoustic impulse responses. IEEE Sig. Process. Lett. **23**(10), 1459–1463 (2016)
67. S. Woźniak, K. Kowalczyk, Passive joint localization and synchronization of distributed microphone arrays. IEEE Sig. Process. Lett. **26**(2), 292–296 (2018)
68. T.-L. Chou, L.-J. ChanLin, Augmented reality smartphone environment orientation application: a case study of the Fu-Jen University mobile campus touring system. Procedia-Soc. Behav. Sci. **46**, 410–416 (2012)
69. D. Nield, All the sensors in your smartphone, and how they work. Dostopno na:. gizmodo (2017). https://fieldguidecom/all-the-sensors-in-your-smartphone-and-how-theywork-1797121002
70. O. Schwartz, S. Gannot, E. A. P. Habets, An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(9), 1495–1510 (2016)
71. C. M. Bishop, *Pattern recognition and machine learning*. (Springer, New York, US, 2006)
72. R. M. Neal, G. E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants. Learn Graph Models. **89**, 355–368 (1998)
73. S.-K. Ng, G. J. McLachlan, On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures. Stat. Comput. **13**(1), 45–55 (2003)
74. L. Frenkel, M. Feder, Recursive expectation-maximization (EM) algorithms for time-varying parameters with applications to multiple target tracking. IEEE Trans. Signal Process. **47**(2), 306–320 (1999)
75. Y. Dorfan, G. Hazan, S. Gannot, in *The 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. Multiple acoustic sources localization using distributed expectation-maximization algorithm, (2014), pp. 72–76. https://doi.org/10.1109/hscma.2014.6843254
76. Y. Dorfan, D. Cherkassky, S. Gannot, in *The 23rd European Signal Processing Conference (EUSIPCO)*. Speaker localization and separation using incremental distributed expectation-maximization, (2015), pp. 1256–1260. https://doi.org/10.1109/eusipco.2015.7362585
77. Y. Dorfan, C. Evers, S. Gannot, P. A. Naylor, in *The 24th European Signal Processing Conference (EUSIPCO)*. Speaker localization with moving microphone arrays, (2016), pp. 1003–1007. https://doi.org/10.1109/eusipco.2016.7760399
78. Y. Dorfan, A. Plinge, G. Hazan, S. Gannot, Distributed expectation-maximization algorithm for speaker localization in reverberant environments. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP). **26**(3), 682–695 (2018)

Publisher's Note