Multiple-Hypothesis Extended Particle Filter for Acoustic Source Localization in Reverberant Environments

Avinoam Levy, Sharon Gannot, *Senior Member, IEEE*, and Emanuël A. P. Habets, *Member, IEEE*

***Abstract*—Particle filtering has been shown to be an effective approach to solving the problem of acoustic source localization in reverberant environments. In reverberant environment, the direct-arrival of the single source is accompanied by multiple spurious arrivals. Multiple-hypothesis model associated with these arrivals can be used to alleviate the unreliability often attributed to the acoustic source localization problem. Until recently, this multiple-hypothesis approach was only applied to bootstrap-based particle filter schemes. Recently, the extended Kalman particle filter (EPF) scheme which allows for an improved tracking capability was proposed for the localization problem. The EPF scheme utilizes a global extended Kalman filter (EKF) which strongly depends on prior knowledge of the correct hypotheses. Due to this, the extension of the multiple-hypothesis model for this scheme is not trivial. In this paper, the EPF scheme is adapted to the multiple-hypothesis model to track a single acoustic source in reverberant environments. Our work is supported by an extensive experimental study using both simulated data and data recorded in our acoustic lab. Various algorithms and array constellations were evaluated. The results demonstrate the superiority of the proposed algorithm in both tracking and switching scenarios. It is further shown that splitting the array into several sub-arrays improves the robustness of the estimated source location.**

***Index Terms*—Acoustic source localization, particle filter, optimal importance sampling.**

I. Introduction

THE physical location of acoustic sources is often required, directly or indirectly, in many applications such as camera steering for video conferencing [1], beamforming [2], and speaker separation [3]. The principal idea of all localization approaches is to exploit the spatial information of acoustic sources present in multiple microphone signals by using prior knowledge of the relative positions of the microphones.

Localization approaches are commonly split into two groups: single- and dual-step approaches. In single-step approaches the location of the source is estimated directly from the microphone signals. In this paper, we will not elaborate on these approaches. Extensive overviews can be found instead in, e.g., [4]–[7]. In dual-step approaches [8]–[11], the time difference of arrival (TDOA) is estimated for certain microphone pairs in the first step. Subsequently, the estimated TDOAs are used to perform the localization in the second step. Dual-step approaches are typically less computationally demanding than single-step approaches since the representation of the geometric space by TDOAs allows for data reduction. Several problems might be encountered in these approaches. First, TDOA errors which may be introduced in the first step are propagated to the second step resulting in an erroneous estimation of the source position. Second, the nonlinear mapping from TDOA values to location renders the derivation of a closed-form optimal estimator cumbersome. Therefore, suboptimal closed-form estimators were proposed. These suboptimal estimators exploit the mapping between the TDOA values and the geometric space by virtue of intersecting loci [12]–[16].

An important dual-step approach is the Bayesian approach, in which the source position is described as a stochastic process. It is convenient to model the dynamics of the source position using a first-order finite-state discrete-time Markov process. The variance of the process noise reflects the level of uncertainty in the source position. Under this model, sudden movements or alternating sources are manifested as high process noise. A recursive optimal Bayesian estimator for a linear state-space model with Gaussian noises (process and measurement) is given by the Kalman filter. Unfortunately, this solution is not applicable to the localization problem due to the nonlinear relation between the state (position) and the measurements (TDOA estimates) [13], [14], [16]. Furthermore, in reverberant environments, where multiple sound arrivals can occur, and in multiple speakers scenarios, the posterior probability density function (pdf) given the measurements can be a multimodal density rather than Gaussian.

Several alternatives can be found in the literature for addressing nonlinear problems and/or multimodal posterior densities. Among these alternatives, the EKF [17], the unscented Kalman filter (UKF) [18] and the Gaussian sum filter (GSF) [19] are widely used. The EKF approximates the nonlinearities only up to the first order. The iterated extended Kalman filter (IEKF) introduced by Jazwinski [20] alleviates this rough approximation by iterating the update stage of the EKF. Faubel *et al.* [21], [22] proposed improvements to the UKF which mitigate some of its inherent deficiencies. Approximated solutions to the localization problem that are based on the EKF,

Manuscript received December 24, 2009; revised June 03, 2010, October 07, 2010; accepted October 30, 2010. Date of publication November 18, 2010; date of current version May 25, 2011. This work was supported in part by the Marie Curie Intra European Fellowship within the 7th European Community Framework Program under contract number PIEF-GA-2009-237246. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tomohiro Nakatani.

A. Levy and S. Gannot are with the School of Engineering, Bar-Ilan University, Ramat-Gan, 52900, Israel (e-mail: avinoamle@gmail.com; gannot@eng.biu.ac.il).

E. A. P. Habets is with the International Audio Laboratories Erlangen, University of Erlangen, 91054 Nuremberg, Germany (e-mail: e.habets@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2093517

and UKF were proposed by Gannot and Dvorkind [23]. A solution based on the IEKF was proposed by Klee *et al.* [24]. The EKF was also applied to the speaker localization problem by incorporating both audio and visual information [25].

The *particle filter (PF)* is an alternative Bayesian approach that performs sequential Monte Carlo (SMC) estimation. In this approach, no particular model is assigned to the posterior pdf. The PF supports nonlinear and non-Gaussian state-space models and hence is suitable for solving the localization problem. An important attribute of the PF is its ability to incorporate a multiple-hypothesis model [26]. In this paper, TDOA readings were used in the localizer implementation.

Ward and Williamson [27] replaced the TDOA detector by the more robust steered beamformer (SBF). Rather than calculating the SBF output for all possible source positions (which necessitates an exhaustive grid search), the SBF is only evaluated for a relatively small number of candidate positions (the current particle set). This method is referred to as the *pseudo-likelihood* approach. In [28], a more general framework was presented and a comprehensive comparison of the pseudo-likelihood and the multiple-hypothesis approaches was carried out. Their results show the superiority of the multiple-hypothesis model combined with a TDOA detector approach.

The basic idea behind the PF is the representation of the posterior pdf as a set of random samples (also known as *particles*). One of the core building blocks of the PF is the importance sampling (IS) mechanism by which an auxiliary density function is used to generate the samples. The choice of the IS density function has a major influence on the performance and complexity of the estimator. Bootstrap-based PF approaches use the prior density as an IS density function. The prior describes the propagation of states in time and is given *a priori* by the state-space model. While this approach has practical advantages, no claims of optimality hold. Furthermore, as this IS function is measurement-independent, it does not allow the filter to adapt to diverse movement regimes. Lehmann and Williamson [29] replaced the traditional IS density with an approximated measurement-dependent IS density to allow improved tracking of multiple nonconcurrent speakers. The approximation is based on a SBF computed for low frequencies. Zhong and Hopgood [30] used an alternative approximation to the optimal measurement-dependent IS density function. This is done by incorporating an EKF into the PF to form the EPF [31]. Neither measurement-dependent IS density functions incorporate a multiple-hypothesis model into the approximation.

The PF was also applied to speaker tracking applications in related problems. When visual information is available, the PF can be used to fuse the audio and video detections, yielding a more robust location estimation [32], [33]. Talantzis *et al.* used an additional background PF initialized periodically to improve robustness against noise and reverberations. The background PF is used to release the primary PF when it is trapped in a spurious location. Zhong and Hopgood used the Rao–Blackwellized particle filter (RBPF) to track multiple, simultaneously active and time-varying speakers [34]. The multiple sources are detected and segregated using a binary time–frequency mask. The speakers, assumed to be slow-paced, are tracked by an EKF. The association of the tracked speakers to the measurements is carried out by the RBPF. The RBPF scheme allows for a reduction in particles in comparison with the conventional PF in estimation of the joint parameter space of all unknown variables. These related problems are beyond the scope of this paper.

In this paper, the problem of tracking a single active acoustic source is addressed. Two dynamic scenarios are considered, namely either one moving source or multiple nonconcurrent alternating speakers. We adapt the EPF approach [31], [30] due to its ability to incorporate measurements into the IS density. The EPF implementation takes into consideration the multiple-hypothesis model [26], characterizing the sound propagation in reverberant enclosures. Our proposed localization scheme is comprised of two steps. The first step is the feature extraction step in which the received microphone signals are mapped to *location features*, e.g., TDOA values, and beamformer output signals. In the second step, the location features are fused by the EPF scheme into a source location estimate. The first step is likely to yield outliers when the global maximum of the detector output is selected as the feature location. We propose to replace the global maximum by a set of local maxima constituting the multiple hypotheses. The EKF is employed in two levels. First, it is applied to each hypothesis to produce an unambiguous measurement vector with fewer outliers. Second, a global EKF is applied to calculate the measurement-dependent IS density. We show that this approach improves the overall performance of the localizer. In addition to the ability of the algorithm to adapt to diverse dynamic scenarios, it is shown to be more robust to reverberation and noise due to the synergy of the multiple-hypothesis model and the improved IS density function.

The structure of the paper is as follows. In Section II, localization detectors are discussed and the motivation for the proposed algorithm is presented. In Section III, PF fundamentals are given and the application of the PF approach to the localization problem is discussed. We present a novel approach to source localization based on an approximation of the optimal IS density using the EPF scheme and the multiple-hypothesis framework. In Section IV, several practical considerations concerning the implementation of the proposed algorithm are discussed. In Section V, the performance of the proposed algorithm is evaluated in different scenarios. The evaluation is conducted using both simulations and data recorded in our acoustic lab. In Section VI, the paper is summarized and concluded.

II. Detectors Survey and Motivation for the Proposed Algorithm

In this section, the first localization step, namely the feature extraction step, is discussed. In the feature extraction step, the localization detectors map the received microphone signals to a source location. In this paper, either TDOA- or beamformer-based detectors are considered. In noisy and reverberant environments, these detectors might produce erroneous readings. In this section, the inherent limitations of the detectors are discussed. Once these limitations are known, means to alleviate them in the second localization step can be derived. These means will be discussed in the rest of the paper.

Consider an M microphone array located in a 3-D space. In this paper, Cartesian coordinates are used, although spherical coordinates can be used as well. Each microphone location

is given by $\mathbf{m}_i = [m_i^{(x)}\ m_i^{(y)}\ m_i^{(z)}]^T$ for $i = 1, 2, \ldots, M$. We assume a *single active source* at each time instant k. The current active source is located at $\mathbf{s}_k = [s_x[k]\ s_y[k]\ s_z[k]]^T$. In the case of alternating sources, $\mathbf{s}_k$ can be a noncontinuous function of time. The ith microphone signal, denoted by $z_i[k]$, consists of ambient noise components, directional noise, and a desired source signal convolved with an acoustic impulse response (AIR). The envelope of the AIR can be characterized by the reverberation time T_{60} and the direct to reverberation ratio (DRR). The feature extraction step maps the received microphone signals $\mathbf{z}_k = [z_1[k], z_2[k], \ldots, z_M[k]]^T$ to the location feature vector $\mathbf{y}_k$. Given the feature vector, the localizer aims at estimating the source location $\mathbf{s}_k$. Alternative feature vectors will be discussed in the sequel.

A. TDOA-Based Detectors

The TDOA-based detectors operate on a single microphone pair and produces a TDOA estimate of the source. For two dimensions, the TDOA value can be mapped to a hyperboloid, which is the locus of all possible source positions corresponding to that TDOA. The TDOA is given by

$$\tau_\ell[k] = \frac{\|\mathbf{s}_k - \mathbf{m}_1^{(\ell)}\| - \|\mathbf{s}_k - \mathbf{m}_2^{(\ell)}\|}{c} \tag{1}$$

where τ_ℓ is the TDOA between microphone signals for time instant k and microphone pair ℓ, c is the speed of sound, and $\mathbf{m}_1^{(\ell)}\ \mathbf{m}_2^{(\ell)}$ are the microphones comprising the pair. The line that connects the two microphones is referred to as the baseline and the inter-microphone distance is the base-length. From array theory, it is known that the mapping from source position to TDOA values depends on the base-length and the angle of arrival with respect to the baseline. Increasing the base-length reduces the position estimation uncertainty for a given TDOA estimation accuracy. Furthermore, the localization accuracy depends on the orientation of the source with respect to the array. Broadside constellations are more suitable for localization than endfire constellations. Finally, the use of orthogonal pairs can further improve the performance of the location estimation algorithm. A comprehensive discussion on aspects of array design in the localization context can be found in [35], [36].

There are several approaches for TDOA estimation. In this paper, the popular generalized cross-correlation (GCC) presented by Knapp and Carter [8] is used. Specifically, the phase transform (PHAT) version of the GCC is used since it has been shown to be more robust to reverberation [37].

The performance of multiple-hypothesis model based localization schemes is dependent to a large extent on the direct-arrival being one of the local maxima at the TDOA detector output. If this condition is not met in sufficient number of location features, no subsequent processing can be performed by any particle filtering scheme to estimate the source position. In high reverberation levels, the reflections are of a diffusive nature and hence may mask the direct-arrival. The direct-arrival to reverberation ratio can be increased by exploiting the time consistency of the direct-arrival compared to the random time of arrivals of reflections. We therefore smoothed the cross-power spectral density (PSD) functions across time by an exponential weighting prior to the evaluation of the GCC-PHAT to produce a more robust detector output.

B. Beamformer-Based Detectors

The beamformer-based detector is using multiple (usually more than two) microphones for a position estimate of the source. The microphone signals are linearly combined by steering the array to a certain position. Then the power of the beamformer output signal is computed to form the steered response power (SRP). The beamformer power is expected to attain its maximum value when the array is steered towards the source position.

Beamformers, generally result in high computational burden, as they require an exhaustive 3-D grid search. As shown in [27], using the pseudo-likelihood approach in bootstrap-based approaches alleviates the grid search. The search-domain of the pseudo-likelihood approach is confined to the locations determined by the distribution of the particles. Therefore, the extraction of multiple location-candidates [26] is inherently limited with the pseudo-likelihood approach.

In practice, beamforming approaches are typically implemented as follows. First, the fixed beamformer is steered to all potential source positions using grid-search. Then a few local maxima of the SRP output are designated as position candidates. The directivity of the beamformer increases the probability that one of these candidate peaks will correspond to the direct-arrival. This probability is presumably higher than the one obtained by TDOA-based detectors.

C. Motivation for the Proposed Algorithm

Originally, both TDOA and beamformer-based detectors were designed for free-space scenarios, single source and spatially white noise environments. Such conditions are rarely met in room acoustic environments. In highly reverberant and noisy environments, both alternative detectors are characterized by highly spurious response masking the direct-arrival. The direct-arrival and the reflections cannot necessarily be separated. Choosing the global maximum for the direct-arrival is an alternative often taken by traditional localization schemes. However, if an erroneous choice is made, the source position estimate will be poor. The multiple-hypothesis model can alleviate this problem (see, e.g., Vermaak and Blake [26]). In this approach, rather than choosing the maximal value, several candidates, corresponding to TDOA or beamformer detector peaks, are considered. In the proposed scheme determining the source location is based on these candidate peaks as well as other building blocks of the PF.

A major benefit of the EPF scheme [31], [30] is its improved tracking and acquisition capabilities obtained by incorporating the measurements into the propagation stage of the PF. As shown in Section III, the incorporation of the multiple-hypothesis model into the EPF scheme is not straightforward. The objective of the current contribution is therefore to propose an applicable source localization method that efficiently combines the benefits of the multiple-hypothesis model and the EPF scheme.

III. Proposed Algorithm

This section starts with a presentation of particle filtering fundamentals. A full formulation of particle filtering theory can be found in [38]–[41]. We proceed by describing the proposed multiple-hypothesis extended particle filter (MH-EPF) localizer which combines merits of the multiple-hypothesis model and the EPF scheme. The structure of the applied algorithm is determined by the type and number of detectors used in the feature extraction step. We conclude this section with a derivation of an efficient approximated variant of the MH-EPF.

A. Particle Filter Preliminaries

Let $\mathbf{x}_k \in \mathbb{R}^{n_x}$ be a state-vector representing a stochastic process to be estimated at time instants $0 \leq k < \infty$ and let $\mathbf{y}_{0:k} \in \mathbb{R}^{n_y}$ be a set of all available (causal) measurement vectors. $\mathbb{R}^n$ is the set of all n dimensional real-valued vectors. PF techniques, are Bayesian state-space estimation techniques aimed at recursively estimating the posterior pdf, $\Pr(\mathbf{x}_k|\mathbf{y}_{0:k})$, also known as the filtering distribution. Given the posterior pdf, any desired Bayesian state estimation can be derived [e.g., maximum *a posteriori* probability (MAP), and minimum mean squared error (MMSE)].

The state-space model is defined by the following system:

$$\mathbf{x}_k = f_k(\mathbf{x}_{k-1}) + \mathbf{v}_k \tag{2}$$

$$\mathbf{y}_k = h_k(\mathbf{x}_k) + \mathbf{w}_k \tag{3}$$

where $\mathbf{v}_k \in \mathbb{R}^{n_x}$ is the process noise with known pdf $\Pr(\mathbf{v}_k)$, $\mathbf{w}_k \in \mathbb{R}^{n_y}$ is the measurement noise with known pdf $\Pr(\mathbf{w}_k)$, f_k is the process function defining the time evolution of the state, and h_k is the measurement equation defining the mapping from the state-vector to the measurements vector for specific time instant. The process noise signal is assumed zero-mean. $E\{\mathbf{v}_k\mathbf{v}_{k'}^T\} = \mathbf{Q}_k\delta[k-k']$, where $\delta[\cdot]$ is the Kronecker delta, $E\{\cdot\}$ denotes expectation, $\mathbf{Q}_k$ is the process noise covariance matrix at time instant k, and k, k' are time instants. The measurement noise signal is assumed zero-mean and $E\{\mathbf{w}_k\mathbf{w}_{k'}^T\} = \mathbf{R}_k\delta[k-k']$, where $\mathbf{R}_k$ is the measurement noise covariance matrix for time instant k. The vectors $\mathbf{v}_k$ and $\mathbf{w}_{k'}$ are mutually independent for all k, k'. In our formulation the noise signals are additive. As a consequence $\Pr(\mathbf{v}_k)$ and $\Pr(\mathbf{w}_k)$ represent the prior $\Pr(\mathbf{x}_k|\mathbf{x}_{k-1})$ and the likelihood $\Pr(\mathbf{y}_k|\mathbf{x}_k)$ probability density functions, respectively.

The basic idea behind the PF is as follows. For time instant k, the posterior pdf is represented by a set of P random samples (event-space samples also known as *particles*) $\mathbf{x}_k^{(p)}$ with corresponding *importance weights* $w_k^{(p)}$, where $1 \leq p \leq P$ is the particle index. As the posterior density is unknown, the samples are obtained using the IS method in which an IS density (also known as the auxiliary or *proposal* density) is used to generate the samples. The IS density has a crucial influence on the performance and complexity of the estimation. The PF consists of two stages: the *propagation stage* and the *update stage*. In the propagation stage, the particles obtained by the previous time instant $\mathbf{x}_{k-1}^{(p)}$ are propagated to the current time instant according to the IS density $\Pr^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{y}_k)$:

$$\mathbf{x}_k^{(p)} \sim \Pr{}^{(\mathrm{IS})}\left(\mathbf{x}_k^{(p)} \middle| \mathbf{x}_{k-1}^{(p)}, \mathbf{y}_k\right). \tag{4}$$

In the update stage the particles weights are updated according to

$$w_k^{(p)} \propto w_{k-1}^{(p)} \frac{\Pr\left(\mathbf{y}_k \middle| \mathbf{x}_k^{(p)}\right) \Pr\left(\mathbf{x}_k^{(p)} \middle| \mathbf{x}_{k-1}^{(p)}\right)}{\Pr^{(\mathrm{IS})}\left(\mathbf{x}_k^{(p)} \middle| \mathbf{x}_{k-1}^{(p)}, \mathbf{y}_k\right)}. \tag{5}$$

In practice, due to the proportionality, only $\tilde{w}_k^{(p)}$, a scaled version of $w_k^{(p)}$, is available. The normalized weights are derived by setting

$$w_k^{(p)} = \frac{\tilde{w}_k^{(p)}}{\sum_{p \in P} \tilde{w}_k^{(p)}}. \tag{6}$$

The PF usually consists of a *resampling* stage [39] which prevents the *degeneration* phenomenon of the particles over time. Equations (4)–(5) together with the resampling stage are commonly referred to as the *sequential importance resampling (SIR)* algorithm.

B. State-Space Formulation

In our case the state-vector $\mathbf{x}_k$ is the source position and is given by $\mathbf{s}_k$. The measurement vector $\mathbf{y}_k$ consists of either position or TDOA readings depending on the implemented feature extraction step. The corresponding process and measurement model equations are defined as follows.

1) Process Equation: The question of what constitutes a suitable speaker trajectory model is still open. Determining trajectory models is beyond the scope of this paper and we decided to stick to well-known models in the literature. The Langevin model is widely used to model the source dynamics [26]–[28]. However, according to our experiments choosing a *random walk* model

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{v}_k \tag{7}$$

suffices to describe the speaker movement [23]. To simplify the exposition we further assume that the process noise covariance matrix is diagonal, i.e.,

$$\mathbf{Q}_k = q_0[k]\mathrm{diag}\left(\sigma_{q_x}^2[k], \sigma_{q_y}^2[k], \sigma_{q_z}^2[k]\right) \tag{8}$$

where $\mathrm{diag}(\cdot)$ is a diagonal matrix with the elements in parentheses on its main diagonal and where $\sigma_{q_i}^2[k]$ is the variance of the ith dimension. The auxiliary factor $q_0[k]$ is used to adjust the process noise variance to adapt to the source dynamics. The process noise variance reflects the level of uncertainty regarding the source movement. The propagation of the state in time in the PF scheme is determined according to the IS density variance. Hence, the current level of IS density variance has a significant effect on the estimation quality. Low values reduce the estimation error variance but hamper the tracking ability of fast maneuvering or alternating sources. In bootstrap-based PF schemes, the IS density is chosen as the prior density and hence the IS variance is the process noise variance. For this scheme, adjustment of the process noise variance to the source dynamics is therefore beneficial and can take place by various

measures, e.g., the elapsed time from the previous speech activity [24] or the level of the importance weights spread. Although these measures are shown in our experiments to be effective, they require additional mechanisms to alleviate the effects of miss-detections.

For some applications, *a priori* knowledge of the speaker dynamics can be exploited. Hence, lower variance can be assigned to dimensions with lower expected source dynamics or higher sensitivity. For example, in many applications, speakers are not expected to move in the z-dimension. A higher variance (i.e., lower sensitivity) can be used for far-field setups in which movement in the radial dimension are not reflected in the detector output. The EPF-based PF scheme adopted in this work, uses an improved measurement-dependent IS density which depends on both the previous state and current measurements and is adapted online according to the source dynamics. Therefore additional measures to control the process noise variance are not necessary and a degenerated time invariant process noise covariance matrix can be used:

$$\mathbf{Q} = q_0\mathbf{I} \tag{9}$$

where q_0 in this scheme is the initial value for the IS density variance and $\mathbf{I} = \mathrm{diag}(1,1,1)$ is the identity matrix with the same dimensions as $\mathbf{Q}$.

Position estimators usually have two main working modes, namely *tracking* and *acquisition*. In the tracking mode, the estimator has already converged and only relatively small adjustments are required; therefore, low process noise variance is required. When either a discontinuous source position change occurs or the estimator loses track, the estimator switches to the acquisition mode and higher variance is preferred. A smooth transition between modes is obtained by adapting the IS variance.

2) Measurement Equation: The measurement equation is determined according to the number and type of detectors used in the feature extraction step.

We start with beamformer-based detectors. In this case, the nonlinear measurement equation in (3) is set according to the SRP output:

$$h_k(\mathbf{x}_k) = \mathbf{x}_k. \tag{10}$$

By splitting the overall array into $L > 1$ sub-arrays, a diversity can be introduced in the measurement vector which is now the concatenation of all L position features:

$$\mathbf{y}_k = \left[\left(\mathbf{y}_k^{(1)}\right)^T, \left(\mathbf{y}_k^{(2)}\right)^T, \ldots, \left(\mathbf{y}_k^{(L)}\right)^T\right]^T \tag{11}$$

where $\mathbf{y}_k^{(\ell)} = \hat{\mathbf{s}}_k^{(\ell)};\ 1 \le \ell \le L$ is the source position detected by the ℓth sub-array. The covariance matrix is block diagonal:

$$\mathbf{R}_k = \mathrm{blk\ diag}\left(\mathbf{R}_k^{(1)}, \mathbf{R}_k^{(2)}, \ldots, \mathbf{R}_k^{(L)}\right) \tag{12}$$

where $\mathbf{R}_k^{(\ell)}$ is the measurement noise covariance matrix of the ℓth feature vector

$$\mathbf{R}_k^{(\ell)} = \mathrm{diag}\left(\sigma^2_{r_x^{(\ell)}}[k], \sigma^2_{r_y^{(\ell)}}[k], \sigma^2_{r_z^{(\ell)}}[k]\right) \tag{13}$$

and $\sigma^2_{r_i^{(\ell)}}$ is the variance of the ith dimension for the ℓth feature vector.

Next, the case of TDOA-based detector is addressed. The feature vector $\mathbf{y}_k$ is again given by (11), where in this case $\mathbf{y}_k^{(\ell)}$ is reduced to a scalar:

$$\mathbf{y}_k^{(\ell)} = \hat{\tau}_{(\ell)}[k]. \tag{14}$$

To keep the exposition general, vector notation is used for $\mathbf{y}_k^{(\ell)}$ throughout the paper. $\hat{\tau}_{(\ell)}[k]$ denotes the TDOA extracted from microphone pair $1 \le \ell \le L$, where $L \le \binom{M}{2}$ is the maximum number of available array pairs. In this case, the nonlinear measurement equation in (3) is set according to (1):

$$h_k(\mathbf{x}_k) = \begin{pmatrix} \frac{\|\mathbf{x}_k - \mathbf{m}_1^{(1)}\| - \|\mathbf{s}_k - \mathbf{m}_2^{(1)}\|}{c} \\ \frac{\|\mathbf{x}_k - \mathbf{m}_1^{(2)}\| - \|\mathbf{s}_k - \mathbf{m}_2^{(2)}\|}{c} \\ \vdots \\ \frac{\|\mathbf{x}_k - \mathbf{m}_1^{(L)}\| - \|\mathbf{s}_k - \mathbf{m}_2^{(L)}\|}{c} \end{pmatrix}. \tag{15}$$

The covariance matrix of the measurement noise vector $\mathbf{w}_k$ is assumed to be diagonal:

$$\mathbf{R}_k = \mathrm{diag}\left(\sigma^2_{r_1}[k], \sigma^2_{r_2}[k], \ldots, \sigma^2_{r_L}[k]\right) \tag{16}$$

where $\sigma^2_{r_\ell}[k]$ is the TDOA estimation variance for the ℓth pair.

C. The MH-EPF

The implementation of the PF requires the evaluation of (4) and (5). The specific structure is determined by the chosen IS density. In previous bootstrap-based work [26], [28], [42], the prior $\Pr(\mathbf{x}_k|\mathbf{x}_{k-1})$, was used as an IS density. For this choice, the propagation and update stages reduce to

$$\mathbf{x}_k^{(p)} \sim \Pr\left(\mathbf{x}_k^{(p)} \middle| \mathbf{x}_{k-1}^{(p)}\right)$$
$$w_k^{(p)} \propto w_{k-1}^{(p)} \Pr\left(\mathbf{y}_k \middle| \mathbf{x}_k^{(p)}\right).$$

Hence, only the prior and the likelihood are required to implement the PF. Although simpler, this choice restricts the tracking and adaptation ability of the estimator in highly dynamic scenarios in comparison with the optimal IS density choice.

In [30], the optimal IS density $\Pr(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{y}_k)$ is used rather than the prior density in the implementation of the PF. For this choice, the propagation and update stages are given by

$$\mathbf{x}_k^{(p)} \sim \Pr\left(\mathbf{x}_k^{(p)} \middle| \mathbf{x}_{k-1}^{(p)}, \mathbf{y}_k\right) \tag{17}$$

$$w_k^{(p)} \propto w_{k-1}^{(p)} \Pr\left(\mathbf{y}_k \middle| \mathbf{x}_{k-1}^{(p)}\right) = w_{k-1}^{(p)} \frac{\Pr\left(\mathbf{y}_k \middle| \mathbf{x}_k^{(p)}\right) \Pr\left(\mathbf{x}_k^{(p)} \middle| \mathbf{x}_{k-1}^{(p)}\right)}{\Pr\left(\mathbf{x}_k^{(p)} \middle| \mathbf{x}_{k-1}^{(p)}, \mathbf{y}_k\right)}. \tag{18}$$

Direct sampling from the optimal IS function is often not possible due to nonlinearities and hence it is approximated by an EKF to form the EPF [31]. This general approach relies on the assumption of an unambiguous measurement vector which is not valid in reverberant environments. We propose a different approach for the evaluation of the optimal IS density function. According to our approach, the EPF scheme is combined with

the multiple-hypothesis model [26] to adapt to the ambiguous nature of the measurement vector.

1) Soft and Hard Decision Frameworks: Due to the limited reliability of the feature extraction step (as discussed in Section II), we replace the original feature vector $\mathbf{y}_k$ by

$$\tilde{\mathbf{y}}_k = \left[\left(\tilde{\mathbf{y}}_k^{(1)}\right)^T, \left(\tilde{\mathbf{y}}_k^{(2)}\right)^T, \ldots, \left(\tilde{\mathbf{y}}_k^{(L)}\right)^T\right]^T \tag{19}$$

where $\tilde{\mathbf{y}}_k^{(\ell)}$ is the concatenation of all candidates for feature ℓ given by

$$\tilde{\mathbf{y}}_k^{(\ell)} \triangleq \left[\left(\mathbf{y}_k^{(\ell,1)}\right)^T, \left(\mathbf{y}_k^{(\ell,2)}\right)^T, \ldots, \left(\mathbf{y}_k^{(\ell,N_k^{(\ell)})}\right)^T\right]^T. \tag{20}$$

$\mathbf{y}_k^{(\ell,n)}, \quad 1 \le n \le N_k^{(\ell)}$ is the nth candidate of the feature ℓ. It can be either a candidate for the source position as measured by sub-array ℓ or a candidate for the TDOA as measured by the ℓth microphone pair. $N_k^{(\ell)}$ is the number of candidate peaks passing a predefined threshold (adaptively defined with respect to the global maximum). Therefore, the number of declared peaks varies with the frame and feature indices. To reduce the number of spurious peaks, $N_k^{(\ell)}$ is upper bounded by a predefined number N. The length of the concatenated feature vector $\tilde{\mathbf{y}}_k$ is $\sum_{\ell=1}^{L} N_k^{(\ell)}$.

We adopt the multiple-hypothesis model proposed by Vermaak and Blake [26]. Let $H_n;\ 1 \le n \le N_k^{(\ell)}$ denote the hypothesis that the candidate n corresponds to the direct-arrival and let H_0 denote the hypothesis that none of the candidates can be attributed to the source. The hypothesis H_0 is valid in either highly reverberant environments or non-active speech segments.

In previous work [26], [28], [42], all hypotheses for each feature were weighted using a soft-decision approach. In the proposed algorithm, an EKF is integrated into the PF structure. A method for weighting multiple-hypotheses in the EKF structure using the soft decision approach was proposed [43]. This method assumes non-maneuvering targets and necessitates a validation region used to discriminate between hypotheses. The existence of such regions cannot be assumed in switching scenarios. Furthermore, by incorporating all measurements in a global EKF, spatial relations between measurements can be exploited to improve robustness against outliers. The resulting number of hypotheses using the entire measurement vector grows exponentially with the measurement-vector length and impose high computational load had soft-decision approach been selected. For these reasons, we used the following combination of hard-decision and soft-decision approaches. The procedure is carried out in two levels. In the first level, the feature level, a *hard-decision* rule is applied to all available candidates of a given feature. The most probable hypothesis is selected according to a weight calculated using a local EKF. Denote the index of the most probable hypothesis for feature ℓ by $\tilde{n}^{\ell}$. Based on these decisions an *unambiguous* measurement vector is constructed:

$$\bar{\mathbf{y}}_k = \left[\left(\mathbf{y}_k^{(1,\tilde{n}^{(1)})}\right)^T, \left(\mathbf{y}_k^{(2,\tilde{n}^{(2)})}\right)^T, \ldots, \left(\mathbf{y}_k^{(L,\tilde{n}^{(L)})}\right)^T\right]^T. \tag{21}$$

In the second level, the feature-vector level, a global EKF is applied to the *unambiguous* measurement vector. This global EKF is used to approximate the desired IS density and to propagate the particles.

The *hard-decision* approach is sensitive to erroneous decisions in the selection of the most-probable hypotheses. The soft-decision approach allows for more flexibility by weighting all hypotheses but cannot be practically incorporated in all algorithm stages. Hence, the *soft-decision* approach is utilized only in the evaluation of the importance weights.

2) The Propagation Stage: In the propagation stage, particles are drawn from the proposal density $\Pr(\mathbf{x}_k|\mathbf{x}_{k-1}, \tilde{\mathbf{y}}_k)$. Since this procedure requires actual drawing of particles, the uncertainties embedded in the concatenated feature vector must be first resolved. Hence, the soft-decision approach is not applicable and the hard-decision approach must be adopted.

We used the maximum hypothesis importance weight criterion as the hard-decision rule give by

$$\tilde{n}^{\ell} = \underset{n\in\{1,2,\ldots,N_k^{(\ell)}\}}{\operatorname{argmax}} \left\{\Pr\left(\mathbf{y}_k^{(\ell,n)}\middle|\mathbf{x}_{k-1}^{(p)}, H_n\right)\right\} \quad \text{for} \quad \ell = 1,2,\ldots,L, \tag{22}$$

where

$$\Pr\left(\mathbf{y}_k^{(\ell,n)}\middle|\mathbf{x}_{k-1}^{(p)}, H_n\right) = \frac{\Pr\left(\mathbf{y}_k^{(\ell,n)}\middle|\mathbf{x}_k^{(p)}, H_n\right)\Pr\left(\mathbf{x}_k^{(p)}\middle|\mathbf{x}_{k-1}^{(p)}, H_n\right)}{\Pr\left(\mathbf{x}_k^{(p)}\middle|\mathbf{x}_{k-1}^{(p)}, \mathbf{y}_k^{(\ell,n)}, H_n\right)} \tag{23}$$

is calculated for each candidate n in feature ℓ. We stress that the set of particles $\mathbf{x}_k^{(p)}$ is independent of ℓ and n. It is assumed that this procedure will yield the correct hypotheses for particles with high importance weights. Particles with low importance weights might introduce errors in the hard-decision procedure but their overall contribution to the state estimation is negligible. The calculation of (23) requires the prior $\Pr(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, H_n)$, the likelihood $\Pr(\mathbf{y}_k^{(\ell,n)}|\mathbf{x}_k^{(p)}, H_n)$, and the proposal $\Pr(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{y}_k^{(\ell,n)}, H_n)$, all under hypothesis H_n.[1]

The prior is independent of the features and hence given for all hypotheses as the process noise pdf:

$$\begin{aligned}\Pr(\mathbf{x}_k|\mathbf{x}_{k-1}) &= \Pr(f(\mathbf{x}_{k-1}) + \mathbf{v}_k|\mathbf{x}_{k-1}) \\ &= \Pr(\mathbf{v}_k|\mathbf{x}_{k-1}) = \Pr(\mathbf{v}_k).\end{aligned}$$

For simplicity we assume Gaussian distribution

$$\Pr\left(\hat{\mathbf{x}}_k^{(p)}\middle|\mathbf{x}_{k-1}^{(p)}, H_n\right) = \mathcal{N}\left(\hat{\mathbf{x}}_k^{(p)}; \mathbf{x}_{k-1}^{(p)}, \mathbf{Q}_k\right). \tag{24}$$

The likelihood under hypothesis H_n is given as the measurement noise pdf:

$$\Pr\left(\mathbf{y}_k^{(\ell,n)}\middle|\hat{\mathbf{x}}_k^{(p)}, H_n\right) = \mathcal{N}\left(\mathbf{y}_k^{(\ell,n)}; h_k(\hat{\mathbf{x}}_k^{(p)}), \mathbf{R}_k^{(\ell)}\right). \tag{25}$$

The proposal density cannot be evaluated analytically in our problem. However, as proposed in [31], it can be approximated

[1]Note that the calculation involves the current particle set $\mathbf{x}_k^{(p)}$ which is not yet available. We therefore use instead the particle set $\hat{\mathbf{x}}_k^{(p)}$ obtained by the previous proposal density, i.e., $\hat{\mathbf{x}}_k^{(p)} \sim \Pr(\mathbf{x}_{k-1}^{(p)}|\mathbf{x}_{k-2}^{(p)}, \bar{\mathbf{y}}_{k-1}) = \mathcal{N}(\mathbf{x}_{k-1}^{(p)}; \boldsymbol{\eta}_{k-1}^{(p)}, \boldsymbol{\Sigma}_{k-1}^{(p)})$.

Algorithm 1: Extended Kalman filter algorithm.

Input: $\boldsymbol{\eta}_{k-1}, \mathbf{y}_k, \boldsymbol{\Sigma}_{k-1}, \mathbf{Q}_k, \mathbf{R}_k$
Output: $\boldsymbol{\eta}_k, \boldsymbol{\Sigma}_k$
begin
Propagation step:
State propagation: $\hat{\boldsymbol{\eta}}_{k|k-1} = f_k(\boldsymbol{\eta}_{k-1})$
Calculate the Jacobians:
$\mathbf{F}_k(\boldsymbol{\eta}_{k-1}) = \frac{\partial f_k(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}|_{\boldsymbol{\eta}=\boldsymbol{\eta}_{k-1}}$
$\mathbf{H}_k(\boldsymbol{\eta}_{k|k-1}) = \frac{\partial h_k(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}|_{\boldsymbol{\eta}=\boldsymbol{\eta}_{k|k-1}}$
Covariance propagation:
$\boldsymbol{\Sigma}_{k|k-1} = \mathbf{F}_k(\boldsymbol{\eta}_{k-1})\boldsymbol{\Sigma}_{k-1}(\mathbf{F}_k(\boldsymbol{\eta}_{k-1}))^T + \mathbf{Q}_k$
Update step:
Evaluate Kalman gain:
$\mathbf{K}_k = \boldsymbol{\Sigma}_{k|k-1}(\mathbf{H}_k(\boldsymbol{\eta}_{k|k-1}))^T$
$\times(\mathbf{H}_k^{\boldsymbol{\eta}_{k|k-1}}\boldsymbol{\Sigma}_{k|k-1}(\mathbf{H}_k(\boldsymbol{\eta}_{k|k-1}))^T + \mathbf{R}_k)^{-1}$
State update: $\boldsymbol{\eta}_k = \boldsymbol{\eta}_{k|k-1} + \mathbf{K}_k\left(\mathbf{y}_k - \mathbf{H}_k(\boldsymbol{\eta}_{k|k-1})\right)$
Covariance update:
$\boldsymbol{\Sigma}_k = \left(\mathbf{I} - \mathbf{K}_k(\boldsymbol{\eta}_{k|k-1})\mathbf{H}_k(\boldsymbol{\eta}_{k|k-1})\right)\boldsymbol{\Sigma}_{k|k-1}$
end

Algorithm 2: Evaluation of (23) under hypothesis H_n.

Input: $\mathbf{x}_{k-1}^{(p)}, \mathbf{y}_k^{(\ell,n)}, \boldsymbol{\eta}_{k-1}^{(p)}, \boldsymbol{\Sigma}_{k-1}^{(p)}$
Output: $\Pr\left(\mathbf{y}_k^{(\ell,n)}|\mathbf{x}_{k-1}^{(p)}\right)$
begin
Draw $\hat{\mathbf{x}}_k^{(p)}$ according to the previous proposal density:
$\hat{\mathbf{x}}_k^{(p)} \sim \Pr\left(\mathbf{x}_{k-1}^{(p)}|\mathbf{x}_{k-2}^{(p)}, \bar{\mathbf{y}}_{k-1}\right) =$
$\mathcal{N}\left(\mathbf{x}_{k-1}^{(p)}; \boldsymbol{\eta}_{k-1}^{(p)}, \boldsymbol{\Sigma}_{k-1}^{(p)}\right)$
Calculate prior:
$\Pr\left(\hat{\mathbf{x}}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}\right) = \mathcal{N}\left(\hat{\mathbf{x}}_k^{(p)}; \mathbf{x}_{k-1}^{(p)}, \mathbf{Q}_k\right)$
Calculate likelihood for hypothesis n:
$\Pr\left(\mathbf{y}_k^{(\ell,n)}|\hat{\mathbf{x}}_k^{(p)}, H_n\right) = \mathcal{N}\left(\mathbf{y}_k^{(\ell,n)}; h_k(\hat{\mathbf{x}}_k^{(p)}), \mathbf{R}_k^{(\ell)}\right)$
Estimate local EKF:
$\left\{\boldsymbol{\mu}_k^{(p,\ell,n)}, \boldsymbol{\Sigma}_k^{(p,\ell,n)}\right\} = \text{EKF}(\mathbf{x}_{k-1}^{(p)}, \mathbf{y}_k^{(\ell,n)}, \boldsymbol{\Sigma}_{k-1}^{(p)})$
Calculate proposal for hypothesis n:
$\Pr\left(\hat{\mathbf{x}}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{y}_k^{(\ell,n)}, H_n\right) =$
$\mathcal{N}\left(\hat{\mathbf{x}}_k^{(p)}; \boldsymbol{\mu}_k^{(p,\ell,n)}, \boldsymbol{\Sigma}_k^{(p,\ell,n)}\right)$
Calculate: $\Pr\left(\mathbf{y}_k^{(\ell,n)}|\mathbf{x}_{k-1}^{(p)}, H_n\right) =$
$\frac{\Pr\left(\mathbf{y}_k^{(\ell,n)}|\hat{\mathbf{x}}_k^{(p)}, H_n\right)\Pr\left(\hat{\mathbf{x}}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}\right)}{\Pr\left(\hat{\mathbf{x}}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{y}_k^{(\ell,n)}, H_n\right)}$
end

for each hypothesis by a Gaussian with the mean and covariance calculated by an EKF:

$$\left\{\boldsymbol{\mu}_k^{(p,\ell,n)}, \boldsymbol{\Sigma}_k^{(p,\ell,n)}\right\} = \text{EKF}\left(\mathbf{x}_{k-1}^{(p)}, \mathbf{y}_k^{(\ell,n)}, \boldsymbol{\Sigma}_{k-1}^{(p)}\right) \quad (26)$$

where $\boldsymbol{\Sigma}_{k-1}^{(p)}$ is the covariance matrix of the proposal density from the previous time instant, and $\boldsymbol{\mu}_k^{(p,\ell,n)}$ and $\boldsymbol{\Sigma}_k^{(p,\ell,n)}$ are the mean and covariance matrix of the current approximated proposal density for hypothesis n in feature ℓ, respectively. Note that the previous proposal density is independent of indices n and ℓ since a decision was already made in the previous step. The EKF procedure for time instant k is depicted schematically in Algorithm 1. The proposal density is then calculated by

$$\Pr\left(\hat{\mathbf{x}}_k^{(p)}\middle|\mathbf{x}_{k-1}^{(p)}, \mathbf{y}_k^{(\ell,n)}, H_n\right) = \mathcal{N}\left(\hat{\mathbf{x}}_k^{(p)}; \boldsymbol{\mu}_k^{(p,\ell,n)}, \boldsymbol{\Sigma}_k^{(p,\ell,n)}\right). \quad (27)$$

The evaluation of (23) for each hypothesis n is summarized in Algorithm 2. After hypothesis $\tilde{n}^{(\ell)}$ is chosen according to (22)

Algorithm 3: Propagation step.

Input: $\mathbf{x}_{k-1}^{(p)}, \mathbf{y}_k^{(\ell,n)}, \boldsymbol{\eta}_{k-1}^{(p)}, \boldsymbol{\Sigma}_{k-1}^{(p)}$
Output: $\mathbf{x}_k^{(p)}, \boldsymbol{\eta}_k^{(p)}, \boldsymbol{\Sigma}_k^{(p)}$
begin
for $\ell = 1 : L$ **do**
for $n = 1 : N_k^{(\ell)}$ **do**
Calculate (23) (Using Algorithm 2)
end
Calculate the most probable candidate: $\tilde{n}^{\ell} =$
$\text{argmax}_{n\in\{1,2,\ldots,N_k^{(\ell)}\}}\left\{\Pr\left(\mathbf{y}_k^{(\ell,n)}|\mathbf{x}_{k-1}^{(p)}, H_n\right)\right\}$
end
Construct the unambiguous measurement vector:
$\bar{\mathbf{y}}_k = \left[\mathbf{y}_k^{(1,\tilde{n}^{(1)})}, \mathbf{y}_k^{(2,\tilde{n}^{(2)})}, \ldots, \mathbf{y}_k^{(L,\tilde{n}^{(L)})}\right]$
Estimate global EKF:
$\left[\boldsymbol{\eta}_k^{(p)}, \boldsymbol{\Sigma}_k^{(p)}\right] = \text{EKF}\left[\mathbf{x}_{k-1}^{(p)}, \bar{\mathbf{y}}_k, \boldsymbol{\Sigma}_{k-1}^{(p)}\right]$
Draw $\mathbf{x}_k^{(p)}$ by the proposal:
$\mathbf{x}_k^{(p)} \sim \Pr\left(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \bar{\mathbf{y}}_k\right) = \mathcal{N}\left(\mathbf{x}_k^{(p)}; \boldsymbol{\eta}_k^{(p)}, \boldsymbol{\Sigma}_k^{(p)}\right)$
end

and (23) for each feature, the unambiguous measurement vector $\bar{\mathbf{y}}_k$ can be constructed.

The propagation stage can be concluded now with the application of a global EKF:

$$\left[\boldsymbol{\eta}_k^{(p)}, \boldsymbol{\Sigma}_k^{(p)}\right] = \text{EKF}\left[\mathbf{x}_{k-1}^{(p)}, \bar{\mathbf{y}}_k, \boldsymbol{\Sigma}_{k-1}^{(p)}\right] \quad (28)$$

where $\boldsymbol{\eta}_k^{(p)}$ and $\boldsymbol{\Sigma}_k^{(p)}$ are the mean and covariance matrix of the approximated proposal density for the unambiguous measurement vector, respectively. The particles $\mathbf{x}_k^{(p)};\ 1 \le p \le P$ are now drawn by propagating $\mathbf{x}_{k-1}^{(p)};\ 1 \le p \le P$ according to $\Pr(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \bar{\mathbf{y}}_k) = \mathcal{N}(\mathbf{x}_k^{(p)}; \boldsymbol{\eta}_k^{(p)}, \boldsymbol{\Sigma}_k^{(p)})$. The propagation step for particle p is summarized in Algorithm 3.

3) The Update Stage: The update stage is evaluated according to (18). The update factor of the importance weight for feature ℓ is derived by the soft-decision approach:

$$\Pr\left(\tilde{\mathbf{y}}_k^{(\ell)}\middle|\mathbf{x}_{k-1}^{(p)}\right) = \sum_{n=0}^{N_k^{(\ell)}} \Pr\left(\tilde{\mathbf{y}}_k^{(\ell)}\middle|\mathbf{x}_{k-1}^{(p)}, H_n\right)\Pr\left(H_n\middle|\mathbf{x}_{k-1}^{(p)}\right) \quad (29)$$

where $\forall n \neq 0$

$$\Pr\left(\tilde{\mathbf{y}}_k^{(\ell)}\middle|\mathbf{x}_{k-1}^{(p)}, H_n\right) = \prod_{m=1}^{N_\ell}\Pr\left(\mathbf{y}_k^{(\ell,m)}\middle|\mathbf{x}_{k-1}^{(p)}, H_n\right).$$

It is assumed that the candidate peaks are independent. The distribution of $\mathbf{y}_k^{(\ell,m)}|\mathbf{x}_{k-1}^{(p)}, H_n;\ \forall m \neq n$ is assumed uniform over the entire feature support and hence

$$\Pr\left(\tilde{\mathbf{y}}_k^{(\ell)}\middle|\mathbf{x}_{k-1}^{(p)}, H_n\right) \sim \Pr\left(\mathbf{y}_k^{(\ell,n)}\middle|\mathbf{x}_{k-1}^{(p)}, H_n\right);\ n \neq 0$$

is proportional to the term (23) derived in the propagation step. It is further assumed that the probability $\Pr(\tilde{\mathbf{y}}_k^{(\ell)}|\mathbf{x}_{k-1}^{(p)}, H_0)$ is uniform.

To conclude the calculation of the update factor the probability of the hypotheses given the state-vector $\Pr(H_n|\mathbf{x}_{k-1}^{(p)})$ is now derived. The probability for hypothesis H_0 can be determined based on voice activity detector (VAD) [44]:

$$\Pr\left(H_0\middle|\mathbf{x}_{k-1}^{(p)}\right) = \begin{cases} p_0^{\text{low}}, & \text{speech present} \\ p_0^{\text{high}}, & \text{speech absent} \end{cases} \quad (30)$$

Algorithm 4: Update stage.

Input: $\Pr\left(\mathbf{y}_k^{(\ell,n)}|\mathbf{x}_{k-1}^{(p)}\right);\ 1 \le n \le N_k^{(\ell)}$
Output: $\tilde{w}_k^{(p)}$
begin
  for $\ell = 1 : L$ **do**
    Calculate the ℓth feature update factor:
    $\Pr\left(\tilde{\mathbf{y}}_k^{(\ell)}|\mathbf{x}_{k-1}^{(p)}\right) =$
    $\sum_{n=0}^{N_k^{(\ell)}} \Pr\left(\tilde{\mathbf{y}}_k^{(\ell)}|\mathbf{x}_{k-1}^{(p)}, H_n\right) \Pr\left(H_n|\mathbf{x}_{k-1}^{(p)}\right)$
  end
  Update the particle importance weight:
  $\tilde{w}_k^{(p)} = w_{k-1}^{(p)} \prod_{\ell=1}^{L} \Pr\left(\tilde{\mathbf{y}}_k^{(\ell)}|\mathbf{x}_{k-1}^{(p)}\right)$
end

where $0 \le p_0^{\text{low}} \ll p_0^{\text{high}} \le 1$. The VAD scheme in this mechanism can be implemented by several alternatives such as the one proposed in [45]. The probability of the remaining hypotheses $H_n;\ 1 \le n \le N_k^{(\ell)}$ is assigned according to $\sum_{i=0}^{N_k^{(\ell)}} \Pr(H_i|\mathbf{x}_{k-1}^{(p)}) = 1$. Several paradigms can be adopted in assigning the probability of the particular hypotheses. Higher probabilities can be assigned to more reliable candidates. In this work no *a priori* information is assumed, therefore equal probabilities are assigned. If $\Pr(H_0|\mathbf{x}_{k-1}^{(p)}) = p_0^{\text{high}}$, the pdf (29) will tend to be uniform as dictated by the H_0 hypothesis. We assume continues speech in this work and hence no VAD is incorporated and $\Pr(H_0|\mathbf{x}_{k-1}^{(p)})$ is set to p_0^{low}.

Finally, assuming feature independence, the general update factor $\Pr(\tilde{\mathbf{y}}_k|\mathbf{x}_{k-1}^{(p)})$ is calculated according to

$$\Pr\left(\tilde{\mathbf{y}}_k \middle| \mathbf{x}_{k-1}^{(p)}\right) = \prod_{\ell=1}^{L} \Pr\left(\tilde{\mathbf{y}}_k^{(\ell)} \middle| \mathbf{x}_{k-1}^{(p)}\right). \tag{31}$$

The update factor is used to calculate the (non-normalized) importance weights:

$$\tilde{w}_k^{(p)} = w_{k-1}^{(p)} \Pr\left(\tilde{\mathbf{y}}_k \middle| \mathbf{x}_{k-1}^{(p)}\right).$$

The update step for particle p is summarized in Algorithm 4.

In highly reverberant environments or high ambient noise, some (or all) detectors may introduce false features. The proposed derivation increases the robustness of the update factor against these scenarios. This can be attributed to the following mechanism. For every faulty detector, the term (23) vanishes and hence (29) tends to be governed by $\Pr(H_0|\mathbf{x}_k^{(p)})$. The implied non-informative uniform distribution renders the contribution of these detectors negligible.

4) Feature Weighting: The main objective for the incorporation of the more involved EKF is to allow for better adaptation and tracking performance. By applying a global EKF on the unambiguous feature vector, the inter-relations between location features are exploited and increased robustness is obtained. The probability that the direct-arrival will be excluded from the list of candidate peaks increases in highly reverberant environments. As a result, the unambiguous measurement vector $\bar{\mathbf{y}}_k$ may include outliers. A high outlier rate will result in an erroneous proposal estimation, rendering the propagation stage useless. It is therefore proposed to weight the components of the unambiguous measurement vector according to the level of confidence assigned to each component. We suggest using the update factor (29) as a confidence measure. The weighting is applied by setting the measurement covariance matrix $\bar{\mathbf{R}}_k$ inversely proportional to the update factor:

$$\bar{\mathbf{R}}_k^{(\ell)} \propto \frac{1}{\Pr\left(\tilde{\mathbf{y}}_k^{(\ell)} \middle| \mathbf{x}_{k-1}^{(p)}\right)} \tag{32}$$

where $\bar{\mathbf{R}}_k^{(\ell)};\ (1 \le \ell \le L)$ is the covariance matrix of the ℓth component of the unambiguous measurement vector.

Algorithm 5: MH-EPF algorithm.

$\left[\mathbf{x}_k^{(p)}, w_k^{(p)}, \mathbf{\Sigma}_k^{(p)}\right] =$
MH-EPF $\left[\left\{\mathbf{x}_{k-1}^{(p)}, w_{k-1}^{(p)}\right\}, \mathbf{\Sigma}_{k-1}^{(p)}, \bar{\mathbf{y}}_k\right]$
begin
  for $p = 1 : P$ **do**
    Propagation step:
    Draw $\mathbf{x}_k^{(p)}$ according to: $\Pr\left(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \tilde{\mathbf{y}}_k\right)$
    (Algorithm 3)
    Update step:
    Update the un-normalized importance weights
    according to: $\tilde{w}_k^{(p)} \propto w_{k-1}^{(p)} \Pr\left(\tilde{\mathbf{y}}_k|\mathbf{x}_{k-1}^{(p)}\right)$
    (Algorithm 4)
  end
  for $p = 1 : P$ **do**
    Normalize the importance weights according to:
    $w_k^{(p)} = \frac{\tilde{w}_k^{(p)}}{\sum_{p=1}^{P} \tilde{w}_k^{(p)}}$;
  end
  Resample step [39]:
  $\left[\left\{\mathbf{x}_k^{(p)}, \mathbf{\Sigma}_k^{(p)}, \frac{1}{P}\right\}\right] =$
  RESAMPLING $\left[\left\{\mathbf{x}_k^{(p)}, \mathbf{\Sigma}_k^{(p)}, w_k^{(p)}\right\}\right]$
end

We denote the localizer based on the improved importance sampling as MH-EPF (following the EPF notation used in [31]). The proposed algorithm is summarized in Algorithm 5. Note, that since the application of the EPF requires $\mathbf{\Sigma}_k^{(p)}$, the resampled particles inherit the covariance matrices associated with their originating particles.

D. The Approximated Multiple-Hypothesis Extended Particle Filter (aMH-EPF) Variant

The proposed algorithm requires the calculation of $P \times \sum_{\ell=1}^{L} N_k^{(\ell)}$ local EKFs and P global EKFs for each time instant. Hence, a high computational burden, compared to the traditional IS-based methods which do not use the EKF, is imposed. The proposed MH-EPF algorithm can be simplified by substituting the local EKF by a less computationally demanding block. The following practical compromises can be adopted.

The hypothesis weighting can be calculated by substituting $\Pr(\tilde{\mathbf{y}}_k^{(\ell,n)}|\mathbf{x}_{k-1}^{(p)})$ with the likelihood density $\Pr(\tilde{\mathbf{y}}_k^{(\ell,n)}|\mathbf{x}_k^{(p)})$ in (22) and (29). As a result, the unambiguous measurement vector is constructed according to the modified hard-decision rule

$$\tilde{n}^{(\ell)} = \underset{n \in \{1,2,\ldots,N_k^{(\ell)}\}}{\operatorname{argmax}} \left\{\Pr\left(\mathbf{y}_k^{(\ell,n)} \middle| \mathbf{x}_k^{(p)}\right)\right\} \quad \text{for} \quad \ell = 1, 2, \ldots, L, \tag{33}$$

rather than (22). The approximated update stage is simplified to

$$w_k^{(p)} \propto w_{k-1}^{(p)} \Pr\left(\tilde{\mathbf{y}}_k \middle| \mathbf{x}_k^{(p)}\right).$$

The subsequent steps of the propagation stage are unaltered. It should be noted that the proposal is not replaced by the prior, since the measurements are still available in the application of the global EKF. The approximated approach is denoted as aMH-EPF.

IV. Practical Considerations

In this section, guidelines for designing the array constellation are given and the excess computational burden with respect to the bootstrap-based localizer is evaluated.

It is well known that when no interferences are present, only three sensors are sufficient for estimating the source position [15]. When the received source signals are distorted by ambient noise and reverberation, more microphones are required for reliable estimation. In the context of multiple-input multiple-output (MIMO) channels for wireless communication, the tradeoff between *diversity gain* and *multiplexing gain* is well studied [46]. We adopt these terms in the context of localization where we refer to *multiplexing gain* as *processing gain*. The additional spatial information obtained by the excess microphones can either be used to increase the diversity gain or to increase the processing gain. The processing gain is increased by using the excess microphones to improve a single location representation. Beamformers are able to achieve high processing gain. The diversity gain is increased when the excess microphones are used to form multiple location representations. The diversity scheme is only effective when the acoustic scene as observed by each detector is different. TDOA-based detectors can be used for maximizing the diversity gain. Any combination of diversity and processing gains can be obtained by introducing sub-arrays.

Beamformers, generally result in a high computational burden, as they require an exhaustive three-dimensional grid search. The imposed complexity restricts the use of beamformers in real-time applications, even when suboptimal techniques [47], [48], [27] are used. As discussed in Section II, Ward and Williamson [27] showed that the grid search can be alleviated using the pseudo-likelihood approach. The proposed algorithms, which are based on the multiple-hypothesis model, necessitate the extraction of multiple location-candidates for each detector. Inherently, the pseudo-likelihood approach confines its search-domain to the locations determined by the distribution of the particles. Hence, it is not likely that this search mechanism will yield adequate location-candidates for the application of our scheme.

In most cases, the use of TDOA-based detectors is less computationally demanding as it involves a one-dimensional search (e.g., finding the peak of the cross-correlation function). In reverberant and noisy conditions, the low processing gain of TDOA-based detectors can be compensated for by the subsequent PF step. Note that the length of the feature vector that results from the TDOA-based detector is usually larger than the vector that results from the beamformer detector. However, the excess computational cost imposed on the MH-EPF is only marginal. In the remainder of the paper, for the reasons detailed above, the TDOA-based detector is used. In the sequel, design rules for improving the effectiveness of this scheme are detailed.

The effectiveness of a certain TDOA detector in suppressing interferences is highly dependent on the room-source-array constellation. There is a tradeoff in determining the base-length. On the one hand, under the free-space model, increasing the base-length improves the detector location accuracy [35]. On the other hand, in a reverberant environment, increasing the base-length increases the dissimilarity between the AIRs rendering the free-space model less adequate. The base-length is empirically set to the maximum distance that still maintains the direct-arrival as one of the candidate peaks with high probability. Due to the limited reliability of these detectors, the direct-arrival is less likely to be the global maximum in comparison with the more robust beamformer detectors. The role of the PF step is to resolve the uncertainties by exploiting the spatial consistency between the direct-arrival in the different features and inconsistencies between strong reflections. It is anticipated that strong reflections as received by remotely located pairs will be spatially inconsistent. The direct-arrival, if detected, is always consistent. In adverse reverberation conditions, the direct-arrival may be absent from the candidate peaks, rendering the subsequent PF step useless.

The excess computational burden of the proposed methods with respect to the bootstrap-based localizer is now evaluated. The implementation of the MH-EPF necessitates for each particle an additional of approximately $\sum_{\ell=1}^{L}(3+5(n_x^3+n_x^2))N^{(\ell)}$ multiplications in the calculation of the hypothesis importance weight and approximately $5n_x^2L+L^2n_x+L^3+n_x^3+n_x^2$ multiplications in the global EKF application. The aMH-EPF variant spares the excess computational burden attributed to the calculation of the hypothesis importance weight.

V. Performance Evaluation

Performance evaluation of the proposed algorithm was carried out by a set of experiments in a simulated and in an actual acoustic environment. In Section V-A, we evaluated the performance of the GCC-PHAT detector for several constellations using data recorded in our acoustic lab. The beamformer-based detectors, which require a grid-search were not evaluated. In Section V-B the proposed algorithm was evaluated for tracking and switching scenarios using simulated data. In Section V-C, the proposed algorithm is evaluated for real recordings in different sensor constellations and environments.

A. Evaluation of the GCC-PHAT Detector

To evaluate the GCC-PHAT detector, a set of data collected in our acoustic lab was used. The effects of the environmental conditions, namely reverberation and noise signals, as well as of the different room-array-source constellations were evaluated and several detector attributes were characterized. The suitability of the detector for localization is quantified using two performance measures representing the level of correspondence between the direct-arrival to the global maximum and the local maxima, respectively. These performance measures will be used in the following sections to simulate the effects of the environment conditions on the detector outputs.

1) Experimental Setup: Recordings were performed in an acoustic room located at Bar Ilan University, Israel, measuring

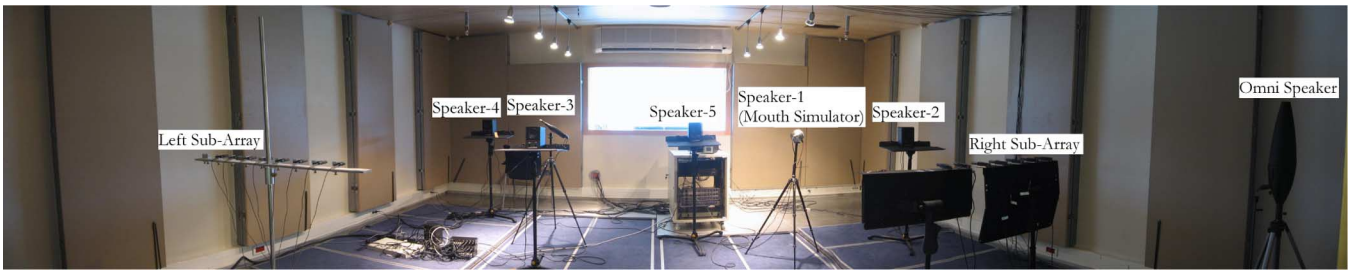

Fig. 1. Recording room setup.

6 m × 6 m × 2.4 m. The room reverberation time can be controlled by using 60 double-sided panels tiled over all six room facets. Each panel can be independently set to its absorbent side or to its reflective side. By arranging the panels in different combinations, an adjustable reverberation time in the range of $T_{60} = 0.1 - 1$ s can be obtained. The reverberation time was measured using maximum length sequence (MLS) played from a type 4295 Brüel & Kjaer omni-speaker and analyzed by the WinMLS2004 software, a product of Morset Sound Development©. A NIST clean speech recording [49] was played from a type 4227 Brüel & Kjaer mouth-simulator. It consisted of five seconds of continuous speech spoken by a male. The room setup is depicted in Fig. 1. The following scenarios were tested.

- Four room setups with a T_{60} of 0.17, 0.3, 0.4, and 0.6 s.
- Two linear 10-cm equispaced sub-arrays each with six microphones. The arrays were positioned at [1.5, 1.5, 1.5] m and [4.2, 1.5, 1.5] m.
- One source spoken by a male, positioned at [3.6, 1.8, 1.5] m.

In all experiments, the sampling frequency is set to 16 kHz and the frame length to 128 samples. To eliminate the effects of the VAD scheme on performance only continuous speech was used in all the experiments. This evaluation was also carried out for different source positions and speakers (both male and female). The results were comparable with the following results and hence are omitted.

High localization performance can only be obtained if the direct-arrival of the source position is one of the detector candidate peaks. We use the following performance measures. The first performance measure is the percentage of detections for which the global maximum of the detector output corresponds to the source direct-arrival. The second performance measure is the percentage of detections for which one of the local maxima candidates extracted from the detector output corresponds to the direct-arrival. The first performance measure is stricter and should be high to enable good performance of traditional localization techniques which consider only the global maximum of each detector. As will be shown, the proposed algorithms yield reasonable performance as long as the source direct-arrival is one of the candidate local maxima (and not necessarily the global maximum). To evaluate the two performance measures the "ground-truth" TDOA corresponding to the true direct-arrival is required. Measuring this TDOA by physical means (e.g., ruler, laser distance measure) is a cumbersome task. We therefore used the following procedure: 1) white noise signal, 5 s long, was played from the desired speaker position and the air-conditioner was switched off; 2) the room was set to the lowest reverberation level ($T_{60} \approx 0.1$ s); 3) the global maximum of the detector was declared as the ground-truth TDOA.

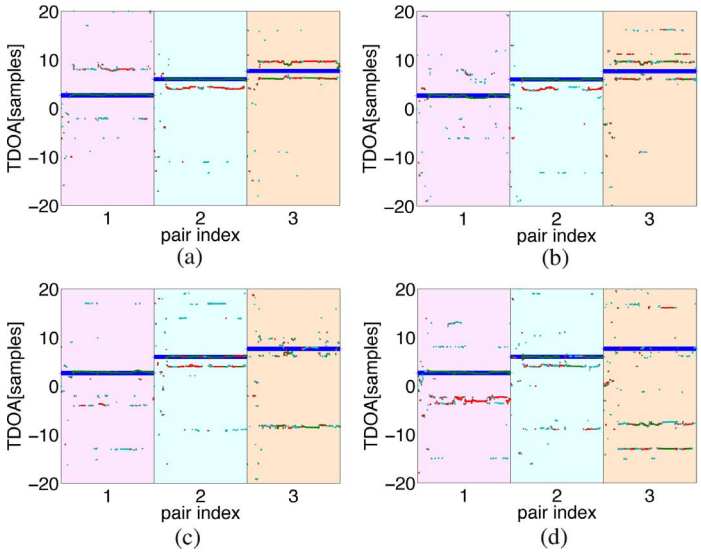

Fig. 2. Detector output versus expected TDOA for different base-lengths for the left sub-array. Each interval (indicated by a different background color) corresponds to a different base-length of 10, 20, 30 cm. The thick blue line denotes the true TDOA value, green line denotes the detector's first candidate (global maximum), red line denotes the detector's second candidate, cyan line denotes the detector's third candidate. (a) $T60 = 0.17$ s. (b) $T60 = 0.3$ s. (c) $T60 = 0.4$ s. (d) $T60 = 0.6$ s.

2) Evaluation: Fig. 2 illustrates the performance of the multi-candidate TDOA detector for several base-lengths and reverberation times. Each subplot corresponds to a different reverberation time T_{60} =0.17, 0.3, 0.4, 0.6 s and is divided into three 5-s (or 625 frames) intervals. Each interval (indicated by different background color) corresponds to a different base-length of 10, 20, and 30 cm resulting in from pairing microphone 1 and microphones 2, 3, and 4 of the left sub-array, respectively. For each interval, the first three local maxima of the detector output (TDOA candidates) are depicted by the green, red, and cyan lines, respectively. The true TDOA value is depicted by the thick blue line. As can be seen, for a given reverberation time, the detector performance is degraded as the base-length increases. As argued in Section IV, it is beneficial to incorporate in the localization scheme microphone pairs forming a long base-length distance provided that the direct-arrival is maintained as one of the candidate peaks with high probability. For the shortest base-length, 10 cm, the detector global maximum almost always ($\approx 90\%$) corresponds to the true TDOA value for all reverberation times. The correspondence between the global maximum and

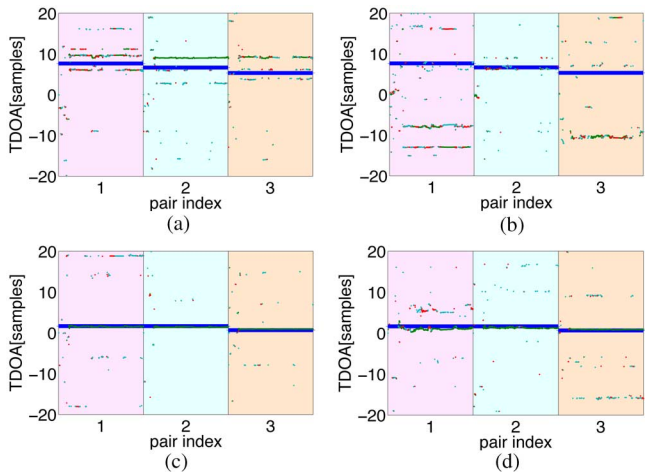

Fig. 3. Detector output versus expected TDOA for different base positions. (a) Left sub-array $T60 = 0.3$ s. (b) Left sub-array $T60 = 0.6$ s. (c) Right sub-array $T60 = 0.3$ s. (d) Right sub-array $T60 = 0.6$ s.

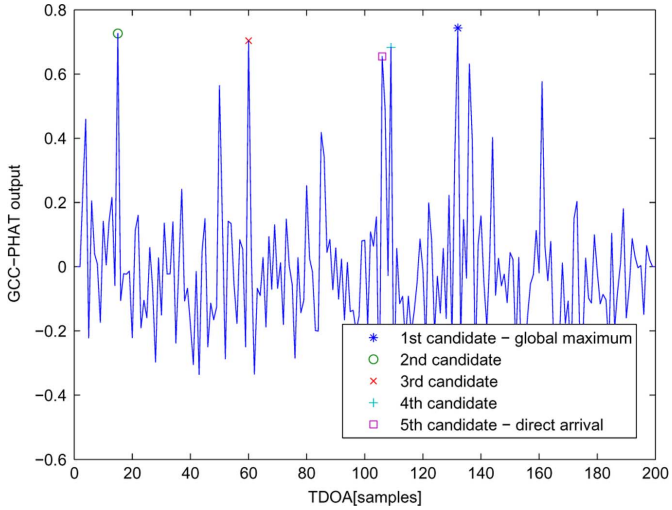

Fig. 4. GCC-PHAT output. The five strongest local maxima are indicated.

the true TDOA weakens as the base-length increases for the higher reverberation times ($\approx$ 60% for base-length of 20 cm and $<$ 5% for base-length of 30 cm). It should be noted that the correspondence between one of the candidates and the true TDOA still remains strong ($\approx$ 90%) for base-lengths of up to 20 cm. For base-lengths of 30 cm and longer the TDOA detector is useless. In high reverberation, due to the diffusive nature of the sound field, the probability of detecting the direct-arrival as one of the candidate peaks dramatically decreases. As discussed in Section II-A, to increase detector robustness, the cross-PSD functions used for the evaluation of the GCC-PHAT are smoothed across time by an exponential weighting. The forgetting factor was set to 0.95. In switching events the cross-PSD functions consist of a weighted mean of the two sources and hence the resulting GCC-PHAT output may have number of competing peaks which are observed as random values across the TDOA range. The resulting TDOAs which correspond to the direct-arrival (and also to strong time consistent reflections) may appear suddenly after the smoothed cross-PSD have stabilized.

The performance of the multi-candidate TDOA detector for different source-pair-room constellations was evaluated in a similar way. To realize different constellations, all six 30 cm pairs of the two sub-arrays were used. The results are depicted in Fig. 3. For the left sub-array, none of the pairs gave rise to reasonable detection for any of the reverberation times. All three pairs of the right sub-array yielded good detection results for all tested reverberation times. We therefore conclude that the base-length and reverberation time are inadequate to characterize the detector's performance, and that the constellation is of major importance as well. Note that in the right sub-array the source is positioned slightly to the right of the normal to the first pair and slightly to the left of the normal to the second pair. Hence, the corresponding TDOA values are approximately 0 s. Due to measurement quantization errors, the measured expected TDOA values coincide.

The output of the GCC-PHAT detector for a certain time frame is depicted in Fig. 4. This plot corresponds to a reverberation level of $T_{60} = 0.6$ s. The five strongest local maxima are also indicated. Due to the strong level of reverberation the direct-arrival peak is only the fifth local maximum. This result justifies the multiple-hypothesis methodology.

B. Evaluation of MH-EPF and aMH-EPF Using Simulated Environments

To evaluate and compare the tracking and adaptation performance of the proposed algorithms, a set of experiments were performed. These experiments are presented in this section.

1) Experimental Setup: The following simulated setup was used. The room dimensions were set to $6 \times 6 \times 3$ m. Twelve omni-directional microphones were arranged in two linear sub-arrays with orthogonal baselines. Each sub-array is comprised of six equally spaced microphones with an inter-microphone distance of 15 cm. The center of mass of the two sub-arrays was set to [1.5, 1.2, 1.5] m and [4.2, 1.2, 1.5] m, respectively.

Two approaches were used to conduct the experiments. In the first approach, the microphone signals were generated by convolving clean continuous NIST recordings [49] with a simulated AIR. The AIRs were simulated by the image method [50] efficiently implemented as described in [51].

In the second approach, the detection stage was bypassed. The location features were directly simulated according to a predefined source trajectory. As concluded from Section V-A, multiple peaks were generated to emulate the reverberation effects. Specifically, the detector outputs were simulated as multiple peaks with random amplitudes and delays. The delay of the peak corresponding to the direct-arrival was always correct. Its amplitude was uniformly distributed in a range corresponding to the reverberation conditions. As a result, the direct-arrival peak level could be lower than the level of the other candidates. Therefore, the direct-arrival could be excluded from the candidate peaks.

In both simulation approaches, only intra sub-array features were used. Hence, the measurement vector is comprised of 30 TDOA values (all available 15 microphone pairs for each sub-array).

In each localization task two modes can be distinguished. The first is the *acquisition* mode in which the localizer tries to "lock" onto the source position. The second is the *tracking* mode in

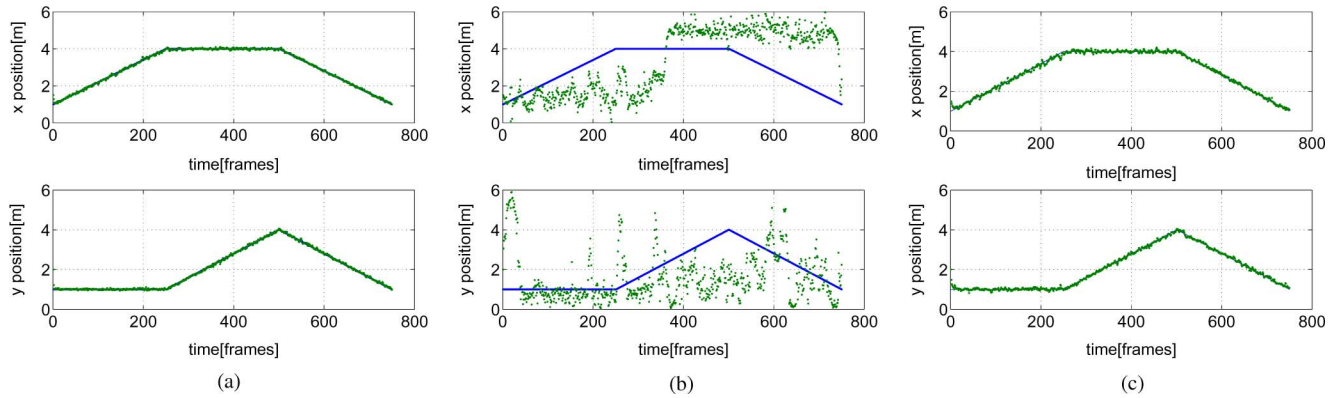

Fig. 5. Assessment of the incorporation of the multiple-hypothesis in the EPF scheme. (a) MH-EPF with 1 hypothesis in anechoic environment. (b) MH-EPF with 1 hypothesis in reverberant environment. (c) MH-EPF with five hypotheses in reverberant environment.

which the localizer adapts to the already acquired source movement. Therefore, two figures-of-merit are defined for evaluating the localizer performance. The acquisition time is used to evaluate the recovery time of the algorithm from abrupt source position changes. It is calculated as the time elapsed from the position change until 80% of the difference from the new position was reached. The root mean squared error (RMSE) is used to evaluate the localization accuracy and is computed over time intervals in which the localization estimator is in steady-state:

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} ||\hat{\mathbf{s}}_i - \mathbf{s}_i||^2} \tag{34}$$

where $\hat{\mathbf{s}}_i$ is the estimator output, $\mathbf{s}_i$ is the true source position at frame i, and $\mathcal{I}$ is the group of frames in which the estimator is in steady-state. In each of the following experiments the steady-state intervals are chosen differently based on a subjective criterion.

2) Tracking Scenario: For this evaluation we used a tracking scenario of a single fast maneuvering source in two dimensions. The source trajectory is split into three stages. First, the source is moving only at the positive x-direction at a constant speed of 1.5 ms^{-1}. Second, the source is moving only at the positive y-direction at the same speed. In the last stage the source returns to the original position by moving in both directions simultaneously. In this scenario we used emulated TDOAs as discussed in Section V-B1.

To assess the advantages of using the multiple-hypothesis model in the EPF implementation, the number of candidate peaks was either set to 1 (i.e., only the global maximum is selected) or 5. Two reverberation conditions were emulated. The first scenario represents an anechoic environment in which the true TDOA is always the global maximum of the feature vector. The second scenario represents a reverberant environment as follows. For approximately 10% of the frames of each pair, the direct-arrival corresponds to the global maximum, for approximately 75% of the frames of each pair the direct-arrival is a local rather than global maximum, and for the remaining frames, the direct-arrival is not one of the candidate peaks. As a result, in each frame, in approximately 20% of the pairs, the direct-arrival peak was absent from the list of candidate peaks. One realization of the localization results of the MH-EPF with 1 hypothesis (global maximum) and with five hypotheses for two reverberation conditions are depicted in Fig. 5. The degradation while only using the global maximum is clearly demonstrated in the reverberant condition. The EKF in the IS approximation tries to impose spatial constraints on the TDOA readings. Since the global maximum might represent an outlier, these constraints becomes invalid, rendering the algorithm more sensitive to reverberation. Using the multiple-hypothesis model alleviates this effect. The MH-EPF is less sensitive to reverberant conditions and only minor degradation of the RMSE from 0.07 m to 0.1 m is observed. The RMSE was calculated by averaging 10 realizations of tracking traces after the localizer converged. In this case, the first 100 time frames are not considered in the calculation.

Next we compare the performance of bootstrap-based localizer [26], and the proposed algorithms for an anechoic environment. The acquisition and tracking modes of the localizer impose contradictory requirements on the IS variance. In the implementation of the bootstrap-based localizer, the IS density is the prior and thus the IS variance is the process noise variance. For this evaluation we used $\mathbf{Q} = 0.005\,\mathbf{I}$ for the bootstrap-based localizer, which is the minimum process noise variance that still allows for tracking. In the implementation of the proposed algorithms, the process noise variance is only the initial value of the IS variance. We set the process noise variance of the proposed algorithms at $\mathbf{Q} = 0.005\,\mathbf{I}$ in all experiments. Note that the proposed algorithms are capable of adapting their IS density in accordance with the source dynamics. The localization results for the MH-EPF, the aMH-EPF and the bootstrap-based localizers in anechoic environments are depicted in Fig. 6. The RMSE averaged over ten realizations of the bootstrap-based localizer is 0.05 m. The RMSE of the MH-EPF and the aMH-EPF is 0.07 m. Therefore, the additional computational burden of the MH-EPF is unjustifiable.

Finally, we compare the performance of bootstrap-based localizer, and the proposed algorithms for reverberant environment. To evaluate the influence of the process noise variance on the bootstrap-based localizer, two alternative values are used. The first alternative allows for minimum RMSE as determined in the previous experiment. In the second alternative we used $\mathbf{Q} = 0.1\,\mathbf{I}$ which is the process noise variance that yields an acquisition time that is comparable to the proposed algorithms.

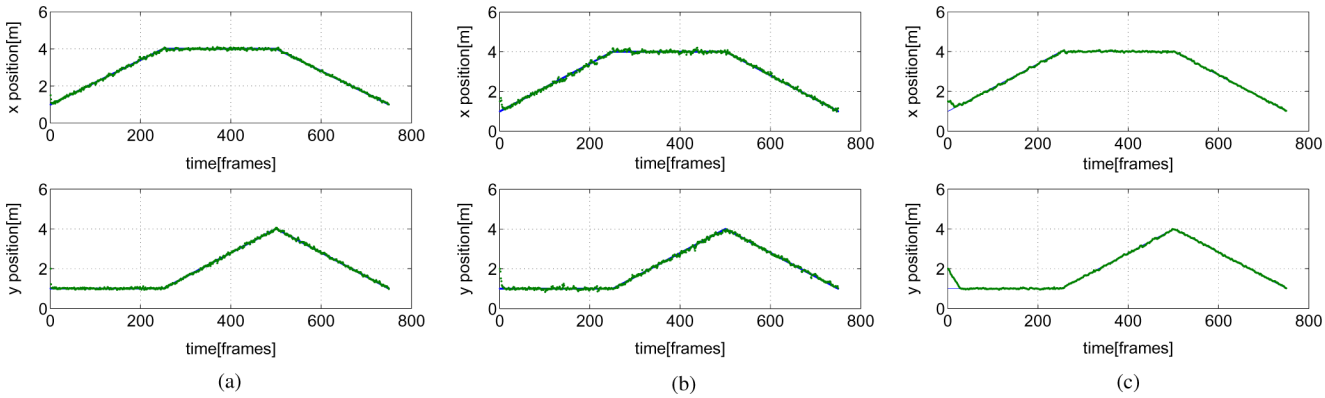

Fig. 6. Tracking performance results in anechoic environment. (a) MH-EPF. (b) aMH-EPF. (c) Bootstrap-based localizer.

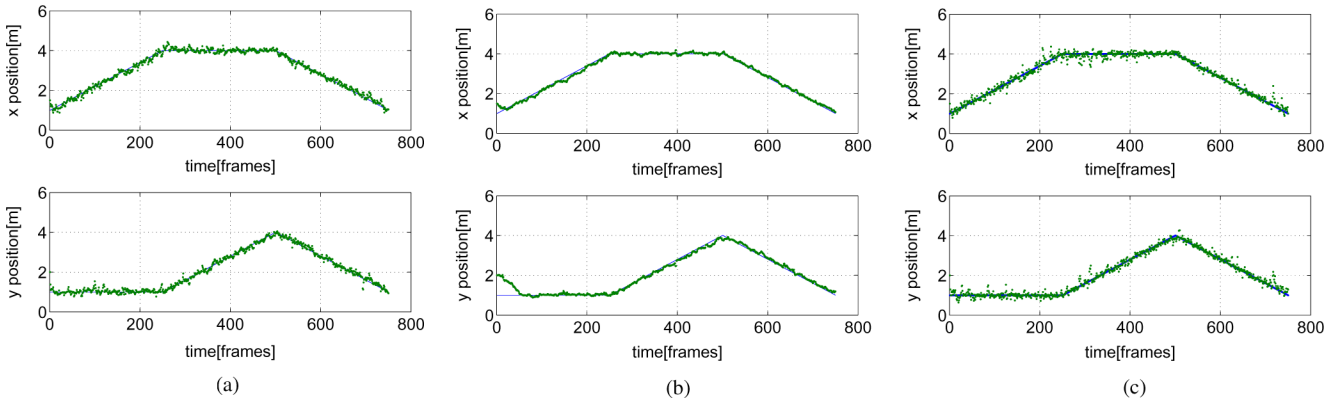

Fig. 7. Tracking performance results of aMH-EPF and bootstrap-based localizer in reverberant conditions. (a) aMH-EPF. (b) Bootstrap-based localizer with $\mathbf{Q} = 0.005\,\mathbf{I}$. (c) Bootstrap-based localizer with $\mathbf{Q} = 0.1\,\mathbf{I}$.

The localization results for the aMH-EPF and the bootstrap-based localizers (two alternative process noise variance) in reverberant environment are depicted in Fig. 7. The RMSE averaged over ten realizations of the bootstrap-based localizer for $\mathbf{Q} = 0.005\,\mathbf{I}$ and $\mathbf{Q} = 0.1\,\mathbf{I}$ are 0.09 m and 0.20 m, respectively. The RMSE of the aMH-EPF in reverberant condition is 0.13 m which is slightly larger than the RMSE obtained by the MH-EPF and the bootstrap-based localizer. It is clearly evident that the acquisition time of the aMH-EPF is comparable to bootstrap-based localizer with $\mathbf{Q} = 0.1\,\mathbf{I}$ while maintaining lower RMSE. The flexibility of the proposed methods obtained by the adaptive IS density and the fact that it is not tailored to a specific scenario, allows for a better tradeoff between acquisition time and RMSE. Hence, it is a good choice for intricate scenarios such as fast maneuvering speakers.

3) Switching Scenario: In this section, the tradeoff between acquisition and tracking for the proposed method was evaluated. We evaluated the aMH-EPF in comparison with bootstrap-based PF scheme. The evaluation was carried out using the following switching scenario. This scenario consists of three different source positions and three switching events. The source positions are [2, 2, 1.5] m, [5, 2, 1.5] m, [5, 5, 1.5] m. The switching are performed first in the x-direction only, then in the y-direction only and finally in both directions simultaneously (return to the first source position). For simplicity, the z-coordinate of all source positions is fixed and therefore the sub-arrays aperture in this axis is degenerated. In each position, continuous 2 s long speech is uttered. We used the first simulation approach detailed in Section V-B1; namely, a simulated AIR [50], [51]. Three noiseless different environmental conditions characterized by T_{60} =0.2, 0.4, 0.6 s were used. The localization accuracy is measured by calculating the RMSE over a number of consecutive frames for each source position, approximately 100 frames after the slowest localizer has converged. The number of consecutive frames used for the RMSE measurement is different for each source position and varies between 100–250 frames. The acquisition time is measured as defined in Section V-B1. Note that the acquisition time is comprised of the delays of both the feature extraction and the PF scheme blocks. To increase the measurement accuracy, ten-fold higher resolution is obtained by interpolating the trace. For a fair comparison between the two localization methods, the process noise variance of the bootstrap-based method was set to obtain comparable RMSE results to the aMH-EPF. The value was empirically found to be $\mathbf{Q} = 0.001\,\mathbf{I}$. Recall that the aMH-EPF is capable of adapting its IS density variance. Under this setup, the algorithms are solely compared by their respective acquisition time.

The performance of the bootstrap-based localizer and the aMH-EPF methods for one realization and three reverberation levels is depicted in Fig. 8. The mean of the RMSE and the acquisition time, averaged over 50 realizations, are given in Table I.

As depicted in Table I, the RMSE values of both localizers are indeed comparable. Note, that for the designated T_{60} values and the chosen array constellation, the RMSE results are not very sensitive to the reverberation level. This can be explained

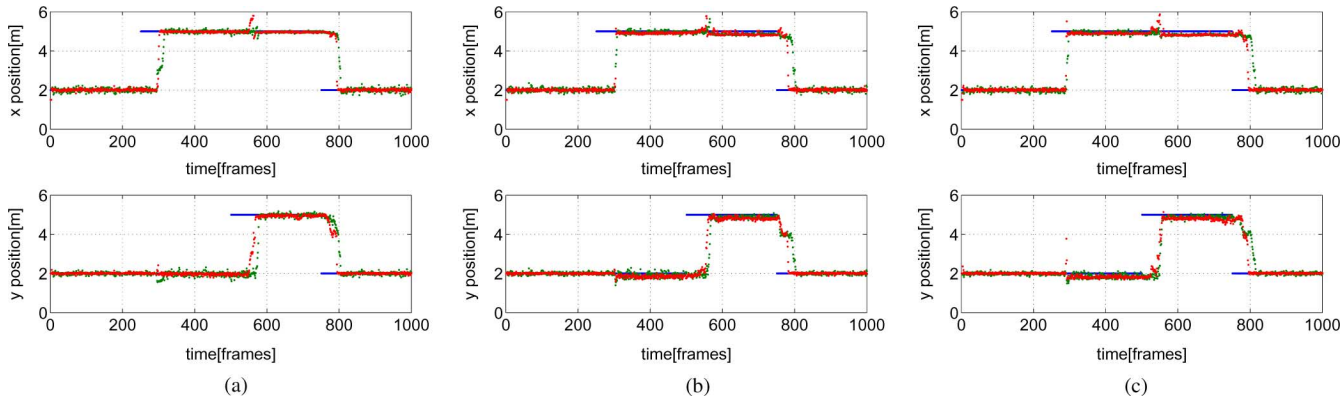

Fig. 8. Acquisition performance results for bootstrap-based method and aMH-EPF. Blue line depicts the true source position, green dotted line the bootstrap-based localizer output and red dotted line the aMH-EPF localizer output. (a) $T_{60} \sim 0.2$ s. (b) $T_{60} \sim 0.4$ s. (c) $T_{60} \sim 0.6$ s.

TABLE I
MEAN RMSE AND ACQUISITION TIME RESULTS FOR BOOTSTRAP-BASED AND AMH-EPF LOCALIZERS

		mean RMSE[cm]			mean acquisition time[ms]		
reverberation		position 1	position 2	position 3	switch $1 \rightarrow 2$	switch $2 \rightarrow 3$	switch $3 \rightarrow 1$
$T_{60} \sim 0.2$ s	bootstrap	10.8	7.8	6.4	714.4	784.8	633.6
	aMH-EPF	10.7	6.7	6.9	659.2	720.0	419.2
$T_{60} \sim 0.4$ s	bootstrap	11.1	7.8	7.1	732.8	802.4	704.8
	aMH-EPF	10.9	6.9	6.7	699.2	721.6	474.4
$T_{60} \sim 0.6$ s	bootstrap	10.9	7.5	8.7	748.8	801.6	746.4
	aMH-EPF	11.3	7.1	8.9	704.0	748.0	483.2

by the results in Section V-A, where it was shown that several speaker-array-room constellations are less affected by reverberation. Since the RMSE results are mostly affected by the correspondence of the direct-arrival to the candidate peaks, only minor differences between reverberation level are expected after the localization process has reached its steady-state. As evident from the table, the acquisition time of the MH-EPF is faster at all reverberation levels. Note that only minor differences exist between the acquisition time of the two methods for the first two switching events (the aMH-EPF is slightly faster than the bootstrap-based method). For the third switching event, the difference is more pronounced. The acquisition time of the aMH-EPF is approximately 250 ms shorter than that of the bootstrap-based method. This behavior can be attributed to the more complex switching event involving transitions in both coordinates.

The simulation performed in this section was repeated for various number of speaker positions, environmental conditions, and speaker-array-room constellations, exhibiting similar trends.

C. Evaluation of Localization Using Real Recordings

In this section, real recordings were used to evaluate the alternative localization schemes using the setup detailed in Section V-A1. Three reverberation levels, T_{60} =0.17, 0.3, 0.4 s, were used for the switching scenario. An air conditioner was activated for the higher reverberation time. Clean speech recordings [49] were played from a type 4227 Brüel & Kjaer mouth-simulator and four Fostex 6301B personal monitor loudspeakers. It consists of five seconds of non-overlapping segments of five alternating sources spoken by three males and two females. The localization algorithm necessitates the positions of the microphones. This calibration was performed using the procedure described in Section V-A1 in two stages. First, the inter-microphone spacings for each sub-array were determined. Second, the distance between the sub-arrays was determined.

1) Array Constellation: In this section, the effect of various array constellations on TDOA-based localization scheme is evaluated. We used only the aMH-EPF algorithm for this evaluation. Four array constellations were tested: left sub-array only, right sub-array only, two sub-arrays with intra-microphone pairs only and two sub-arrays with all available microphone pairs. The measured reverberation level was $T_{60} = 0.3$ s. The localization results for the x-coordinate are given in Fig. 9. Due to the array orientation, the results for the y-coordinate are inadequate. These data suggest the following conclusions. 1) Not all sub-array locations are appropriate for all source positions. For instance, the positions of speakers #4 and #5 are better estimated by the left sub-array while the position of speaker #1 is better estimated in the right sub-array. 2) Introducing spatial diversity is a good practice. In our evaluation the spatial diversity is obtained by using the two spatially apart sub-arrays with short intra-pair distances. This can be attributed to the inconsistency of the received reflections in the different sub-arrays. 3) Including long baseline pairs deteriorates the localization performance. This can be attributed to the inability of long baseline GCC-based detectors to provide reliable TDOA readings in a reverberant environment as explained in Section II. The third constellation is advantageous in this scenario since on the one hand it increases the diversity gain and on the other hand it does not use microphone pairs with a long baseline.

2) Comparisons of Particle Filter Schemes: Based on the conclusions of the previous section, we continued with the constellation consisting of two sub-arrays with intra-microphone

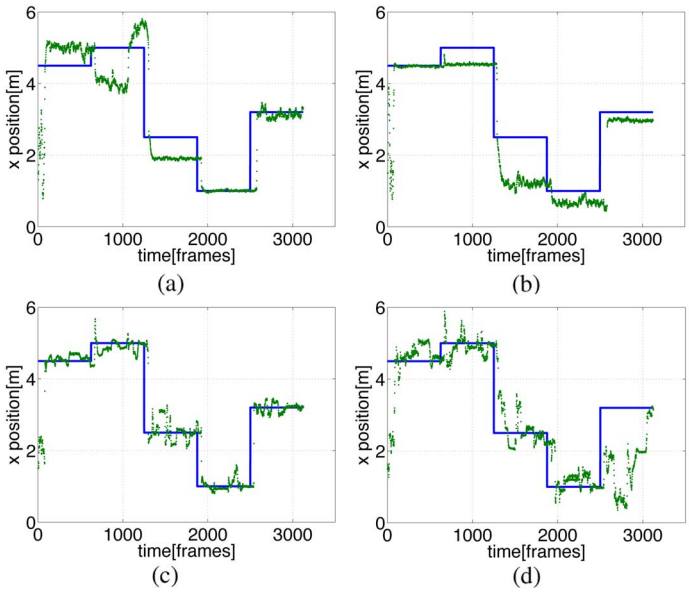

Fig. 9. Localization performance for aMH-EPF for different array constellations. Blue line—the true position, green dotted line—$T_{60} = 0.3$ s. (a) Left sub-array. (b) Right sub-array. (c) Two sub-arrays—intra-pairs only. (d) Two sub-arrays—all pairs.

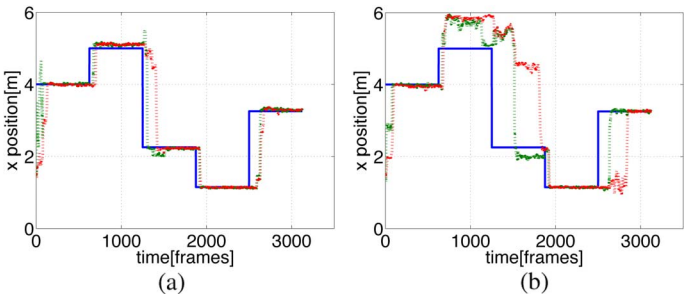

Fig. 10. Localization performance for two sub-arrays with intra-microphone pairs only. Blue line—the true position, green dotted line—aMH-EPF, red dotted line—bootstrap-based localizer. (a) $T_{60} = 0.17$ s. (b) $T_{60} = 0.4$ s.

pairs only. The test scenario of the previous section was repeated for the bootstrap-based localizer and aMH-EPF algorithms. The results for the tested algorithms are depicted in Fig. 10. In consistence with Section V-B3, the aMH-EPF estimates are more accurate and exhibits faster acquisition time than the bootstrap-based method in both reverberation levels. For high reverberation, both methods can be locked on spurious peaks. To enable better performance in high reverberation, an appropriate constellation should be used, as illustrated in Section V-C1.

VI. Summary

In this paper, a new localization method, based on an approximation of the optimal IS density was proposed. By combining the EPF with the multiple-hypothesis model, the proposed PF scheme is able to adapt its IS density (and hence its propagation step) according to the source dynamics while maintaining robustness to reverberation.

As observed in the performance evaluation, in addition to the reverberation level, the distance between microphones and the room-source-array constellation significantly affects the performance of localization schemes. By using remotely positioned small sub-arrays, a diversity can be introduced in the location features. The diversity introduced by the localization scheme is exploited in the PF to suppress the undesired reverberation effects. Hence, multiple-hypothesis based localization schemes outperform traditional localization schemes, that consider only the global maximum of the detector. Proper array design, enables the use of the non-robust but flexible GCC detectors. These detectors are easily incorporated into the proposed scheme forming a location estimator with improved tracking and acquisition performance.

References

[1] Y. Huang, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Istanbul, Turkey, Jun. 2000, vol. 2, pp. 909–912.

[2] T. Nishiura, R. Nishioka, T. Yamada, S. Nakamura, and K. Shikano, "Multiple beamforming with source localization based on CSP analysis," *Syst. Comput. Jpn.*, vol. 34, no. 5, pp. 69–80, 2003.

[3] K. H. Knuth, "Bayesian source separation and localization," in *SPIE98 Proc.: Bayesian Inference for Inverse Problems*, 1998, pp. 147–158.

[4] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.

[5] R. Roy and T. Kailath, "ESPRIT—Estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[6] W. Bangs and P. Schultheis, "Space-time processing for optimal parameter estimation," in *Signal Process.*, J. Griffiths, P. Stocklin, and C. Van Schooneveld, Eds., 1973, pp. 577–590.

[7] G. Carter, "Variance bounds for passively locating an acoustic source with a symmetric line array," *J. Acoust. Soc. Amer.*, vol. 62, pp. 922–926, Oct. 1977.

[8] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

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

[10] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, 2000.

[11] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environment," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 11, pp. 1110–1124, 2003.

[12] Y. T. Chan and K. C. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 1905–1915, Aug. 1994.

[13] H. C. Schau and A. Z. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, no. 12, pp. 1223–1225, Dec. 1987.

[14] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 12, pp. 1661–1669, Dec. 1994.

[15] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A closed form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 1, pp. 45–50, Jan. 1997.

[16] Y. A. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, "Real-time passive source localization: A practical linear correction least-squares approach," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 943–956, Aug. 2001.

[17] S. F. Schmidt, "Computational techniques in Kalman filtering," *NATO Advisory Group for Aerospace Research and Development*, 1970.

[18] S. Julier and J. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. IEEE*, vol. 92, no. 3, pp. 401–422, Mar. 2004.

[19] H. Sorenson and D. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, vol. 7, no. 4, pp. 465–479, 1971.

[20] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. New York: Academic, 1970.

[21] F. Faubel and D. Klakow, "A transformation-based derivation of the Kalman filter and an extensive unscented transform," in *Proc. IEEE/SP 15th Workshop Statist. Signal Process.*, Aug. 2009, pp. 161–164.

[22] F. Faubel, J. McDonough, and D. Klakow, "The split and merge unscented Gaussian mixture filter," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 786–789, Sep. 2009.

[23] S. Gannot and T. G. Dvorkind, "Microphone array speaker localizers using spatial-temporal information," *EURASIP J. Appl. Signal Process.*, vol. 2006, 2006, article ID 59625.
[24] U. Klee, T. Gehrig, and J. McDonough, "Kalman filters for time delay of arrival-based source localization," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–15, 2005, Article 12378.
[25] T. Gehrig, K. Nickel, H. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoustics*, Oct. 2005, pp. 118–121.
[26] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'01)*, May 2001, pp. 3021–3024.
[27] D. B. Ward and R. C. Williamson, "Particle filter beamforming for acoustic source localization in a reverberant environment," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'02)*, May 2002, pp. 1777–1780.
[28] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 826–836, Nov. 2003.
[29] E. A. Lehmann and R. C. Williamson, "Particle filter design using importance sampling for acoustic source localisation and tracking in reverberant environments," *EURASIP J. Appl. Signal Process.*, pp. 168–168, 2006.
[30] X. Zhong and J. Hopgood, "Nonconcurrent multiple speakers tracking based on extended Kalman particle filter," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'08)*, Apr. 2008, pp. 293–296.
[31] E. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Proc. IEEE Adaptive Syst. for Signal Process., Commun., Control Symp. AS-SPCC*, 2000, pp. 153–158.
[32] F. Talantzis, A. Pnevmatikakis, and A. Constantinides, "Audio-visual active speaker tracking in cluttered indoors environments," *IEEE Trans. Syst., Man, Cybern., B: Cybern.* , vol. 38, no. 3, pp. 799–807, Jun. 2008.
[33] D. Zotkin, R. Duraiswami, and L. Davis, "Multimodal 3-D tracking and event detection via the particle filter," in *Proc. IEEE Workshop Detection and Recognition of Events in Video* , 2001, pp. 20–27.
[34] X. Zhong and J. R. Hopgood, "Time-frequency masking based multiple acoustic sources tracking applying Rao-Blackwellised Monte-Carlo data association," in *Proc. IEEE/SP 15th Workshop Statist. Signal Process. SSP'09*, Sep. 2009, pp. 253–256.
[35] S. Kay, *Fundamentals of Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1993, vol. 1, Estimation Theory.
[36] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part IV: Optimum Array Processing*. New York: Wiley, 2002.
[37] J. H. Dibiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown Univ., Providence, RI, May 2000.
[38] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statist.Comput.*, vol. 10, no. 3, pp. 197–208, 2000.
[39] N. J. Gordon, D. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEE Proc.-F*, vol. 140, no. 2, pp. 107–113, 1993.
[40] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond The Kalman Filter: Particle Filters for Tracking Applications*. Norwood, MA: Artech House, 2004.
[41] A. Doucet, J. F. G. Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag, 2001.
[42] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Experimental comparison of particle filtering algorithms for acoustic source localization in a reverberant room," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP 03)*, May 2003, vol. 5, pp. 177–180.
[43] Y. Bar-Shalom and E. Tse, "Tracking in a cluttered environment with probabilistic data association," *Automatica*, vol. 11, pp. 451–460, 1975.
[44] E. A. Lehmann and A. M. Johansson, "Particle filter with integrated voice activity detection for acoustic source tracking," *EURASIP J. Adv. Signal Process.*, pp. 22–28, 2007.
[45] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.
[46] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory,* , vol. 49, no. 5, pp. 1073–1096, May 2003.
[47] M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-31, no. 5, pp. 1210–1217, Oct. 1983.
[48] M. F. Berger and H. F. Silverman, "Microphone array optimization by stochastic region contraction," *IEEE Trans. Signal Process.*, vol. 39, no. 11, pp. 2337–2386, Nov. 1991.
[49] "Wall Street Journal-based continuous speech recognition (CSR) corpus phase II (WSJ1)," *Linguistic Data Consortium*, Apr. 1994.
[50] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.
[51] E. A. P. Habets, "Room impulse response (RIR) generator," 2009 [Online]. Available: http://home.tiscali.nl/ehabets/rir_generator.html

Avinoam Levy received the B.Sc. degree in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 1999. He is currently pursuing the M.Sc. degree in electrical engineering at Bar-Ilan University, Ramat-Gan, Israel.

From 1999 to 2005, he served as an Algorithms Engineer and System Engineer in the Israel Defense Forces. From 2005 to 2006, he was an Algorithms Engineer with ELTA, Israel. From 2006 to 2008, he was a DSP Algorithm Engineer with Radlive, Israel. His research interests include parameter estimation, statistical signal processing, and speech processing using either single- or multi-microphone arrays.

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 post-doctoral position in the Department of Electrical Engineering (SISTA), K.U.Leuven, Leuven, Belgium. From 2002 to 2003, he held a research and teaching position at the Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa. Currently, he is an Associate Professor in the School of Engineering, Bar-Ilan University, Ramat-Gan, Israel.

Dr. Gannot was the recipient of Bar-Ilan University Outstanding Lecturer Award for the year 2010. He is an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, and a member of the IEEE Audio and Acoustic Signal Processing Technical Committee. He is also Associate Editor of the *EURASIP Journal on Advances in Signal Processing*, an Editor of two special issues on Multi-microphone Speech Processing of the same journal, a guest editor of the *ELSEVIER Speech Communication Journal* and a reviewer of many IEEE journals and conferences. He has been a member of the Technical and Steering Committee of the International Workshop on Acoustic Echo and Noise Control (IWAENC) since 2005 and the general cochair of IWAENC 2010 held in Tel-Aviv, Israel. His research interests include parameter estimation, statistical signal processing, array processing, and speech processing.

Emanuël A. P. Habets (S'02–M'07) was born in 1976. He received the B.Sc. degree in electrical engineering from the Hogeschool Limburg, Limburg, The Netherlands, in 1999 and the M.Sc. and Ph.D. degrees in electrical engineering from the Technische Universiteit Eindhoven, Eindhoven, The Netherlands, in 2002 and 2007, respectively.

From March 2007 to February 2009, he was a Postdoctoral Researcher at the Technion—Israel Institute of Technology, Haifa, and at the Bar-Ilan University, Ramat-Gan, Israel. From February 2009 to November 2010, he was a Member of the Research Staff in the Communication and Signal Processing Group, Imperial College London, London, U.K.. In November 2010, he joined the International Audio Laboratories Erlangen at the University of Erlangen-Nuremberg, Nuremburg, Germany as an Associate Professor. His research interests are in the areas of speech and audio signal processing, and he has worked in particular on speech dereverberation, microphone array processing, echo cancellation and suppression, system identification and equalization, and localization and tracking of stationary and moving acoustic sources.

Dr. Habets was a member of the organization committee of the 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC) and is a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing. In 2009, he was awarded a Marie Curie Intra-European Fellowship for Career Development.