Tracking an Unknown Time-Varying Number of Speakers using TDOA Measurements: A Random Finite Set Approach

Wing-Kin Ma, Ba-Ngu Vo, S. Singh, and A. Baddeley

***Abstract*— Speaker location estimation techniques based on time-difference-of-arrival (TDOA) measurements have attracted much attention recently. Many existing localization ideas assume that only one speaker is active at a time. In this paper, we focus on a more realistic assumption that the number of active speakers is unknown and time-varying. Such an assumption results in a more complex localization problem, and we employ the random finite set (RFS) theory to deal with that problem. The RFS concepts provide us with an effective, solid foundation where the multi-speaker locations and the number of speakers are integrated to form a single set-valued variable. By applying a sequential Monte Carlo (SMC) implementation, we develop a Bayesian RFS filter that simultaneously tracks the time-varying speaker locations and number of speakers. The tracking capability of the proposed filter is demonstrated in simulated reverberant environments.**

***Index Terms*— Random finite set, source localization, time difference of arrival, sequential Monte Carlo, Bayesian filter**

I. Introduction

Speaker localization using voice activity is an important problem in microphone array processing, driven by applications such as automatic camera steering in video-conferencing. By localization, one can consider estimating the directions of the speaker sources, or estimating the Cartesian coordinates of the sources. In this paper we are interested in the Cartesian coordinate localization, with particular emphasis on the time-difference-of-arrival (TDOA) approach. Readers who are interested in the direction finding approach (which presents a rather different signal processing framework compared to the TDOA) are referred to the literature such as [1]–[4] and the references therein.

Fig. 1 depicts a microphone placement for the TDOA approach. Essentially microphones are grouped into pairs and those pairs are distributed in the room (note that one can also choose to place those pairs in proximity to each other; see the setting in [5] for example). For each pair the TDOA is measured independently. Assuming single speaker activity and no reverberation, the TDOAs can be reliably measured using a generalized cross correlation (GCC) method [6], [7]. The measured TDOAs, which embed location information relative to the microphone pair locations, are then fused to estimate the Cartesian coordinate of the speaker. This second stage process can be done by some simple, effective algorithms such as [5], [8]–[10].

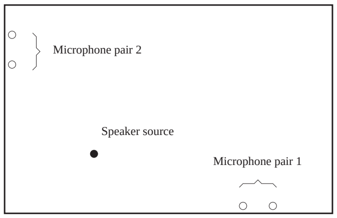

Fig. 1. TDOA microphone array system.

The TDOA single-speaker localization approach mentioned above is vulnerable to reverberation, a problem that is quite common in room environments. The rich multipaths resulting from reverberation can lead to anomalous GCC TDOA estimates. Under such circumstances it is suggested to employ the phase transform (PHAT) GCC method for reducing TDOA estimation error [7]. To further suppress the effect of anomalous TDOA measurements, blind channel identification methods have been introduced to replace the role of GCC [5], [11], [12]. An alternative that has emerged more recently is to apply the Bayesian object tracking framework [13], [14] in the source localization stage. In this approach the possibility of GCC TDOA outliers is incorporated in the problem formulation, thereby making the resultant speaker location estimator less prone to reverberation (compared to localization methods that do not assume the presence of TDOA outliers). Another salient feature of the Bayesian approach is that the speaker location is sequentially tracked with respect to (w.r.t.) time, by following a Markov speaker motion model. In other words, the Bayesian approach exploits the correlation of speaker motions w.r.t. time, which can help improve localization accuracy. The Bayesian tracking approach has been numerically demonstrated [13], [14] to be robust against the effects of reverberation. Moreover, performance comparisons of the Bayesian approach with the blind channel identification based localization approach have been examined in [14][1].

This work is supported in part by discovery grant DP0345215 awarded by the Australian Research Council.

W.-K. Ma is with the Department of Electrical Engineering, National Tsing Hua University, 101, Hsinchu,Taiwan 30013. (e-mail: wkma@ieee.org).

B.-N. Vo is with the Electrical Engineering Department the University of Melbourne, Melbourne, Vic. 3010, Australia. (email: bv@ee.unimelb.edu.au).

S. Singh is with the Department of Engineering, Cambridge University, CB2 1PZ Cambridge, U.K. (email: sss40@eng.cam.ac.uk).

A. Baddeley is with the school of Mathematics and Statistics, the University of Western Australia. (email: adrian@maths.uwa.edu.au).

[1] It is interesting to mention that conceptually, the Bayesian tracking idea can be applied to the blind channel identification methods to further improve localization accuracy. However, no such work has yet appeared in the speaker localization literature.

The objective of this paper is to deal with TDOA based localization with unknown time-varying number of speakers and with reverberation, by applying a Bayesian tracking framework based on the random finite set (RFS) formulation [15]–[18]. This multi-speaker localization problem presents a challenge in signal processing, and very recently some attempts [19], [20] have been made to tackle that problem for known number of speakers. Our attempt for locating unknown time-varying numbers of speakers is based on the multi-object tracking approach, a generalization of the single-object tracking approach. There is a variety of techniques for multi-object tracking; see the reviews in [16], [21], [22]. The RFS approach employed here is a recently emerged framework that has been found to be promising for multi-object tracking [15]–[18]. RFS is a rigorous mathematical discipline for dealing with random spatial patterns [23]–[26] that has long been used by statisticians in many diverse applications including agriculture, geology, and epidemiology; see [25] and the references therein for further details. In essence, an RFS is a finite collection of elements where not only each RFS constituent element is random, but the number of elements is also random. The RFS approach to multi-object tracking is elegant in that the multiple object states and the number of objects are integrated to a single RFS. More importantly, RFS provides a solid foundation for Bayesian multi-object tracking, that is not found in traditional multi-object tracking approaches. For further discussions of the differences between the RFS and traditional approaches, please read [15], [16], [27]. An exposition of RFS theory is rather involved particularly when it comes to the constructions of probability densities for RFSs; see [15], [18], [25] for the details. Fortunately, for most engineering applications it suffices to know how to apply several key concepts and results, which in our opinion are presently not well publicized for the generally knowledgeable readers and therefore will be demonstrated in this paper.

The summary of this paper, together with the organization, are as follows. After a background review in Section II, in Section III we propose an RFS model for the multi-speaker tracking application. Section III also lays several reasonable assumptions for the application, that turn out to greatly facilitate the RFS tracking implementation. Those assumptions include

i) At each time instant, at most one speaker source can be born.
ii) The number of simultaneously active speaker sources is small.

The assumption in ii) is particularly true in applications such as video-conferencing, in which the most frequently encountered events are no voice activity, one speaker voice activity, and one speaker interrupting another. Section IV describes the Bayes RFS filter (or tracker) and its implementation using the sequential Monte Carlo (SMC) method in [18]. The Bayes RFS SMC filter is known to be computationally expensive for large number of objects [18], [22]. For those cases it would be appropriate to consider computationally efficient approximations such as the probability hypothesis density method [16], [18], [28]. Fortunately for small number of objects it is still computationally affordable to employ the Bayes RFS SMC filter (see, for example, [22]), and it appears that the multi-speaker tracking problem considered here falls into such case. Section V deals with track association in the RFS framework. The presently available track association resolutions work by combining the RFS module with some other tracking modules [29], [30], but in this work we exploit the 'at-most-one-birth' assumption to come up with a simple track association method. The idea is to augment the speaker state vector by a discrete variable that records track association information. It is interesting to point out that such an idea has been alluded to in [31]. It is further shown that by using the proposed track association method, the respective RFS state estimation process can be greatly simplified. To demonstrate the performance of the Bayes RFS SMC filter, in Section VI we provide two sets of simulation results based on relatively realistic room simulations. In particular, in one of the simulation examples we tested the robustness of the Bayes RFS SMC filter against the effects of model mismatch.

This work is a more complete version of the conference paper [32]. In particular, [32] did not consider the track association method and the respective RFS state estimation method in Section V.

II. Background

This section provides a brief review on TDOA speaker tracking, by considering the simple case of single speaker activity and no reverberation. We should point out that the method reviewed in the following subsections is a simplified version of the TDOA single-speaker location tracking method in [13], [14], in which the reverberation problem was also addressed.

In the first subsection, some aspects regarding TDOA measurements are described. Then, in the second subsection we consider some basic concepts of Bayesian tracking.

A. TDOA Measurement

In the scenario of a single speaker without reverberation, the received signals at a microphone pair can be modeled as [7]

$$y_1(t) = s(t) + \nu_1(t), \quad y_2(t) = s(t-\tau) + \nu_2(t) \tag{1}$$

where $s(t)$ is the signal due to the source, $\nu_i(t)$, $i = 1, 2$ are background noise, and τ is the TDOA between the 1st and 2nd microphones. Let $\boldsymbol{\alpha} \in \mathbb{R}^2$ denote the (x, y) position vector of the speaker source. The TDOA τ is dependent on $\boldsymbol{\alpha}$ through the following nonlinear relation:

$$\tau = \tfrac{1}{c}(\|\boldsymbol{\alpha} - \mathbf{u}_2\| - \|\boldsymbol{\alpha} - \mathbf{u}_1\|) \tag{2}$$

where $\mathbf{u}_1$ and $\mathbf{u}_2$ are the positions of the two microphones, $\|.\|$ is the 2-norm, and c is the speed of sound (note that extension to the 3-dimensional coordinate is straightforward). The TDOA can be measured by a generalized cross-correlation (GCC) estimator [6], given as follows:

$$\hat{\tau} = \arg \max_{\tau \in [-\tau_{max}, \tau_{max}]} R_{gcc}(\tau) \tag{3}$$

$$R_{gcc}(\tau) = \int_{-\infty}^{\infty} \Phi(\omega) S_{y_1,y_2}(\omega) e^{j\omega\tau} d\omega \tag{4}$$

Here, $R_{gcc}(\tau)$ is called the GCC function, $\tau_{max} = \|\mathbf{u}_2 - \mathbf{u}_1\|/c$ is the maximum admissible TDOA value, $S_{y_1,y_2}(\omega)$ is the cross spectral density of $y_1(t)$ and $y_2(t)$, and $\Phi(\omega)$ is a weighting function; see [6], [7] for details regarding the choices of $\Phi(\omega)$. A popular choice of $\Phi(\omega)$ is the phase transform (PHAT), where $\Phi(\omega) = 1/|S_{y_1,y_2}(\omega)|$. In this work we will employ the PHAT.

In practice the speaker position $\boldsymbol{\alpha}$ can change over time, in which case it is appropriate to estimate τ from a relatively short time frame so that $\boldsymbol{\alpha}$ is (almost) static over each frame. Thus, we replace $S_{y_1,y_2}(\omega)$ in (4) by a short-time estimate

$$\hat{S}_{y_1,y_2}(\omega;k) = Y_1(\omega;k)Y_2^*(\omega;k) \tag{5}$$

$$Y_i(\omega;k) = \int_{(k-1)T}^{kT} y_i(t)e^{-j\omega t}dt, \quad i=1,2 \tag{6}$$

where T is the time frame length, and k is the time frame index.

B. Sequential State Estimation

We consider a standard state space model [13], [14] for the single-speaker TDOA problem mentioned above. We use the notation $\boldsymbol{\alpha}_k$ to represent the speaker location at the kth time frame. By defining a state vector $\mathbf{x}_k = [\ \boldsymbol{\alpha}_k^T, \boldsymbol{\phi}_k^T\ ]^T \in \mathbb{R}^n$ where n is the state dimension and $\boldsymbol{\phi}_k$ contains some kinematic variables for the speaker motion (e.g., velocity), we model $\mathbf{x}_k$ by a dynamic process:

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{w}_k \tag{7}$$

where $\mathbf{A}$ and $\mathbf{B}$ are some pre-specified matrices, and $\mathbf{w}_k$ is a time-uncorrelated random Gaussian vector with zero mean and covariance $\mathbf{I}$. In speaker location tracking, it is popular to employ the Langevin model [13], [14] in which $\boldsymbol{\phi}_k$ consists of the (x,y) velocities. The state space equations for the Langevin model are given by

$$\boldsymbol{\alpha}_k = \boldsymbol{\alpha}_{k-1} + T\boldsymbol{\phi}_{k-1} \tag{8}$$

$$\boldsymbol{\phi}_k = e^{-\beta T}\boldsymbol{\phi}_{k-1} + \nu\sqrt{1-e^{-2\beta T}}\mathbf{w}_k \tag{9}$$

Here, β and ν are model parameters called the rate constant and the steady-state root-mean-square velocity, respectively.

Next, we consider the TDOA measurements. We denote by $z_k^{[q]}$ the TDOA measured from the qth microphone pair at time frame k. The measured TDOAs are modeled by:

$$z_k^{[q]} = \tau_q(\mathbf{C}\mathbf{x}_k) + v_k^{[q]}, \quad q=1,\ldots,Q. \tag{10}$$

Here, $\mathbf{C} = [\ \mathbf{I}\ \mathbf{0}\ ]$ so that $\mathbf{C}\mathbf{x}_k = \boldsymbol{\alpha}_k$,

$$\tau_q(\boldsymbol{\alpha}_k) = \tfrac{1}{c}(\|\boldsymbol{\alpha}_k - \mathbf{u}_{2,q}\| - \|\boldsymbol{\alpha}_k - \mathbf{u}_{1,q}\|) \tag{11}$$

is the true TDOA value, $\{\mathbf{u}_{1,q}, \mathbf{u}_{2,q}\}$ are the position vectors of the qth microphone pair, and $v_k^{[q]}$ is time-uncorrelated noise. We assume that $v_k^{[q]}$ is independent of $v_k^{[p]}$ for any $q \neq p$, and that each $v_k^{[q]}$ follows a Gaussian distribution with zero mean and variance σ_v^2.

Our goal is to estimate $\mathbf{x}_k$ over time. In the sequential Bayesian framework, we assume knowledge of the probability density function (p.d.f.) of $\mathbf{x}_k$ conditioned on $\mathbf{x}_{k-1}$, which we denote by

$$f(\mathbf{x}_k|\mathbf{x}_{k-1}), \tag{12}$$

and the p.d.f. of $z_k^{[q]}$ given $\mathbf{x}_k$, which we denote by

$$g_q(z_k^{[q]}|\mathbf{x}_k). \tag{13}$$

Eqs. (12) and (13) are called the state transition density and the likelihood function, respectively. By letting $\mathcal{N}(.;\boldsymbol{\mu},\mathbf{P})$ denotes the Gaussian density function with mean $\boldsymbol{\mu}$ and covariance $\mathbf{P}$, the expressions (12) and (13) are $f(\mathbf{x}_k|\mathbf{x}_{k-1}) = \mathcal{N}(\mathbf{x}_k;\mathbf{A}\mathbf{x}_{k-1},\mathbf{B}\mathbf{B}^T)$ and $g_q(z_k^{[q]}|\mathbf{x}_k) = \mathcal{N}(z_k^{[q]};\tau_q(\mathbf{C}\mathbf{x}_k),\sigma_v^2)$, respectively. Let $z_{1:k}^{[1:Q]}$ define the sequence containing $z_i^{[q]}$ for $i=1,\ldots,k$ and for $q=1,\ldots,Q$. The Bayesian approach considers finding the posterior p.d.f.

$$p(\mathbf{x}_k|z_{1:k}^{[1:Q]}), \tag{14}$$

which then allows us to estimate $\mathbf{x}_k$ using some optimal criterion such as the expected *a posteriori* (EAP). The posterior p.d.f. obeys the following recursion [33], [34]:

$$p(\mathbf{x}_k|z_{1:k-1}^{[1:Q]}) = \int f(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|z_{1:k-1}^{[1:Q]})d\mathbf{x}_{k-1} \tag{15}$$

$$p(\mathbf{x}_k|z_{1:k}^{[1:Q]}) = \frac{\prod_{q=1}^{Q} g_q(z_k^{[q]}|\mathbf{x}_k)p(\mathbf{x}_k|z_{1:k-1}^{[1:Q]})}{\int \prod_{q=1}^{Q} g_q(z_k^{[q]}|\mathbf{x}_k)p(\mathbf{x}_k|z_{1:k-1}^{[1:Q]})d\mathbf{x}_k} \tag{16}$$

To solve (15) and (16) exactly is not easy due to the nonlinearity of $\tau_q(.)$. Presently, in TDOA single-speaker tracking, a promising approach to approximating (15) and (16) is the sequential Monte Carlo methods [33], [34]; see [13], [14] for the details.

Our proposed RFS method follows the same paradigm as the above Bayesian tracking framework. This is illustrated in the following sections. Moreover, we should point out that in [13], [14], the above Bayesian framework has been extended to handle single-speaker tracking in the presence of reverberation. In those works, the GCC method was slightly modified to cater for the possibility of false TDOA peaks caused by reverberation. Since this work will employ the same modified GCC method (which will be described in the next section), the proposed method may be considered as a multi-speaker generalization of [13], [14].

III. RFS Formulation for Multi-Speaker Localization

This section describes our problem formulation for TDOA multi-speaker tracking in the presence of reverberation, using the random finite set (RFS) framework. In the first subsection, we outline the characteristics of our multi-speaker problem. An RFS formulation for the problem is then presented in the second subsection.

A. The Multi-Speaker Problem

The multi-speaker scenario considered here has the following characteristics: i) each speaker follows the state space motion model described in Section II-B, but his/her own voice activity interval is unknown to the system; and ii) each speaker's voice undergoes a multipath propagation, due to room reverberation. In the proposed RFS treatment, we represent the state vectors of the speakers by a single finite set, given by:

$$\mathcal{X}_k = \{\mathbf{x}_{1,k}, \ldots, \mathbf{x}_{N_k,k}\}, \tag{17}$$

where $N_k = |\mathcal{X}_k|$ ($|.|$ stands for the cardinality) is the number of active speakers at time k, and each $\mathbf{x}_{i,k}$ represents a distinct speaker state vector. We assume that $|\mathcal{X}_k| \leq N_{max}$ for some given N_{max}, and that $|\mathcal{X}_k|$ is unknown. In the next subsection we will develop a statistical finite set model for $\mathcal{X}_k$, which describes not only the state space motion mechanism, but also the appearance and disappearance events for each speaker.

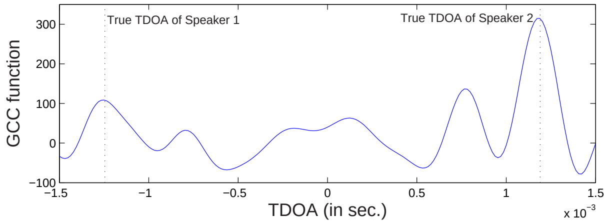

Fig. 2. GCC function response in the presence of multi-speaker activity and reverberation. The response was obtained by using recorded real speech, and by simulating a room environment where reverberation is present.

In the presence of multi-speaker, reverberation-induced multipath signal propagations, the GCC function in (4) is composed of the cross-correlations of the various paths. Hence, some of the peaks of the GCC function are expected to be contributed by the direct path components of the speaker sources. This can be seen from the illustration in Fig. 2, which shows a GCC function response in the presence of multi-speaker multipath propagation. We follow the TDOA extraction scheme in [13], [14] where multiple TDOAs are measured from one GCC function by picking more than one peak (or locally maximum point) in (4). By collecting those TDOAs to form a finite set, we have the following finite-set-valued measurement at time k for the qth microphone pair:

$$\mathcal{Z}_k^{[q]} = \left\{ z_{1,k}^{[q]}, \ldots, z_{M_k^{[q]},k}^{[q]} \right\}, \tag{18}$$

where $M_k^{[q]} = |\mathcal{Z}_k^{[q]}|$ is the number of measured TDOAs. We are now faced with the following problems: i) Given an $\mathbf{x}_{i,k}$ in $\mathcal{X}_k$ we expect that one of the measured TDOAs in $\mathcal{Z}_k^{[q]}$ is generated by $\mathbf{x}_{i,k}$, but we do not know which element in $\mathcal{Z}_k^{[q]}$ is due to $\mathbf{x}_{i,k}$. ii) It may occasionally turn out that $\mathbf{x}_{i,k}$ does not contribute a measurement to $\mathcal{Z}_k^{[q]}$. This measurement miss situation can occur when one speaker cross-correlation response masks that of another, and/or when the speech signal powers are too weak in certain time frames due to the nonstationarity of speech signals. iii) $\mathcal{Z}_k^{[q]}$ may contain false TDOAs; i.e., TDOAs that are not generated by the direct paths. Such an effect can also be seen in Fig. 2.

B. The RFS Formulation

We consider an RFS formulation that models the multi-speaker, multi-measurement problems described in the last subsection. The multi-speaker finite set $\mathcal{X}_k$ is modeled by

$$\mathcal{X}_k = \mathcal{B}_k(\mathbf{b}_k) \cup \left\{ \bigcup_{i=1,\ldots,|\mathcal{X}_{k-1}|} \mathcal{S}_k(\mathbf{x}_{i,k-1}, \mathbf{w}_{i,k}) \right\} \tag{19}$$

where $\mathcal{B}_k(\mathbf{b}_k)$ contains state vectors of speakers 'born' at time k, $\mathcal{S}_k(\mathbf{x}_{i,k-1}, \mathbf{w}_{i,k})$ is a finite set associated with the previous speaker state $\mathbf{x}_{i,k-1}$, and the vectors $\mathbf{w}_{i,k}$ and $\mathbf{b}_k$ are random variables independent of one other. For $\mathcal{S}_k$, we have the following hypotheses:

$$\mathcal{S}_k(\mathbf{x}_{i,k-1}, \mathbf{w}_{i,k}) = \begin{cases} \emptyset, & H_{death} \\ \{\mathbf{A}\mathbf{x}_{i,k-1} + \mathbf{B}\mathbf{w}_{i,k}\}, & \bar{H}_{death} \end{cases} \tag{20}$$

where H_{death} and $\bar{H}_{death}$ are respectively the death and no-death hypotheses. Note that for the no-death hypothesis, the state space process is exactly the same as that of the simple single-speaker case in (7). The hypothesis H_{death} occurs with probability P_{death}. For the birth process, we assume that at most 1 speaker is born at a time. If $|\mathcal{X}_{k-1}| = N_{max}$ then we have $\mathcal{B}_k = \emptyset$. Otherwise, the following hypotheses apply:

$$\mathcal{B}_k(\mathbf{b}_k) = \begin{cases} \emptyset, & \bar{H}_{birth} \\ \{\mathbf{b}_k\}, & H_{birth} \end{cases} \tag{21}$$

where H_{birth} and $\bar{H}_{birth}$ are respectively the birth and no-birth hypotheses, and $\mathbf{b}_k$ is an initial state vector under the birth hypothesis. We denote the probability of H_{birth} by P_{birth}. Moreover, $\mathbf{b}_k$ is assumed to follow an initial state distribution in which the (x, y) position is uniformly distributed within the room enclosure and the other kinematic variables are zero.

For the measurement model, we assume that $\mathcal{Z}_k^{[q]}$ is independent of $\mathcal{Z}_k^{[p]}$ for any $q \neq p$. Each $\mathcal{Z}_k^{[q]}$ is modeled by

$$\mathcal{Z}_k^{[q]} = \left\{ \bigcup_{i=1,\ldots,|\mathcal{X}_k|} \mathcal{T}_k^{[q]}(\mathbf{x}_{i,k}, v_{i,k}^{[q]}) \right\} \cup \mathcal{C}_k^{[q]} \tag{22}$$

where $\mathcal{C}_k^{[q]}$ is the finite set of false TDOAs, and $\mathcal{T}_k^{[q]}$ is given by

$$\mathcal{T}_k^{[q]}(\mathbf{x}_{i,k}, v_{i,k}^{[q]}) = \begin{cases} \emptyset, & H_{miss} \\ \left\{ \tau_q(\mathbf{C}\mathbf{x}_{i,k}) + v_{i,k}^{[q]} \right\}, & \bar{H}_{miss} \end{cases} \tag{23}$$

with $v_{i,k}^{[q]} \sim \mathcal{N}(0, \sigma_v^2)$. Here, $\tau_q(\mathbf{C}\mathbf{x}_k)$ is given in (11), and $\bar{H}_{miss}$ and H_{miss} are respectively the detection and miss hypotheses. The hypothesis H_{miss} happens with probability P_{miss}. For the false TDOAs, we follow the standard assumption in [13], [14] that each $c_k^{[q]} \in \mathcal{C}_k^{[q]}$ independently follows a uniform distribution over the admissible TDOA interval $[-\tau_{max}, \tau_{max}]$, where $\tau_{max} = \|\mathbf{u}_{2,q} - \mathbf{u}_{1,q}\|/c$ (For simplicity the inter-sensor distance $\|\mathbf{u}_{2,q} - \mathbf{u}_{1,q}\|$ for every microphone pair is assumed to be the same). In addition, the number of false TDOAs $|\mathcal{C}_k^{[q]}|$ is assumed to follow a Poisson distribution with an average rate of λ_c.

Some remarks are now in order:

Remark 1: The above described RFS model is applicable to any N_{max}; that is, the maximum number of simultaneously active speakers. However, the performance of the GCC TDOA measurement method in practice incurs a limitation on the choice of N_{max}. GCC benefits from its simplicity, but it is not a super resolution method in the multiple TDOA estimation context. When there are many speakers or when the TDOAs of two speakers are close, GCC may only be able to obtain a few true TDOAs that are associated with the dominant sources. Fortunately, for speech applications it is generally true that the number of simultaneously active speakers is small, such as 2 (some justification for this has been presented in Section I). Hence, in this TDOA speaker localization application, it is pertinent to focus on a small N_{max}.

Remark 2: The probability of birth P_{birth} and the probability of death P_{death} are not known in reality. In practice it is reasonable to make a guess on P_{birth} and P_{death}, and such parameter adjustments generally lead to some tradeoffs on the performance of multi-speaker tracking (presented in the next sections). Increasing P_{birth} is expected to improve the chance of identifying a newly born speaker source. Likewise, to quickly identify speaker source death is advised to increase P_{death}. However, increasing P_{birth} and/or P_{death} may also increase the chance of over-estimation and under-estimation on the number of speakers, especially in the presence of false measurements. In other words, the tuning of P_{birth} and P_{death} is a tradeoff between sensitivity and robustness.

Remark 3: Like P_{birth} and P_{death}, in practice the probability of miss P_{miss} is decided by some rough guess. Increasing P_{miss} is expected to improve the robustness against the measurement miss situations. However, increasing P_{miss} may also reduce the accuracy of speaker state estimation.

IV. Bayesian RFS Filter

With the above RFS problem formulation, we can develop a Bayesian framework for estimating $\mathcal{X}_k$; i.e., sequentially estimating both the multi-speaker locations and the number of active speakers. In the first subsection, we examine some probabilistic results that are essential to the Bayesian RFS framework. Then, the second subsection proposes an implementation for Bayesian RFS estimation using the sequential Monte Carlo (SMC) technique, also known as the particle filter.

A. Bayes Recursion for RFSs

The RFS theory provides two important tools for the multi-speaker tracking application. First, we can construct p.d.f.s for the RFS $\mathcal{X}_k$ and $\mathcal{Z}_k^{[q]}$ according to the model outlined in the previous section. In particular, we can determine a multi-speaker RFS state transition density, denoted by

$$f(\mathcal{X}_k|\mathcal{X}_{k-1}), \tag{24}$$

and RFS likelihood functions, denoted by

$$g_q(\mathcal{Z}_k^{[q]}|\mathcal{X}_k) \tag{25}$$

for $q = 1, \ldots, Q$. To construct these p.d.f.s, some involved mathematical concepts are required and the details are beyond the scope of this application paper. Readers are referred to [15], [18], [25] for complete descriptions of the RFS p.d.f. concepts. From an application viewpoint, we are more interested in the results, particularly the expressions for (24) and (25). The derivations of (24) and (25) are given in the Appendix. Second, many ideas in RFS Bayes estimation are essentially the same as those of the standard Bayes framework (c.f., Section II-B). To explain this, let $\mathcal{Z}_{1:k}^{[1:Q]}$ define a sequence consisting of the finite sets $\mathcal{Z}_i^{[q]}$ for all $i = 1, \ldots, k$ and $q = 1, \ldots, Q$. In an RFS Bayesian framework, we consider determining the posterior p.d.f. $p(\mathcal{X}_k|\mathcal{Z}_{1:k}^{[1:Q]})$ thereby estimating $\mathcal{X}_k$ over time. Moreover, $p(\mathcal{X}_k|\mathcal{Z}_{1:k}^{[1:Q]})$ has a recursive relation reminiscent of the prediction and update formulae in (15) and (16), given as follows:

$$p(\mathcal{X}_k|\mathcal{Z}_{1:k-1}^{[1:Q]}) = \int_{\mathcal{F}(\mathbb{R}^n)} f(\mathcal{X}_k|\mathcal{X}_{k-1})p(\mathcal{X}_{k-1}|\mathcal{Z}_{1:k-1}^{[1:Q]})\mu(d\mathcal{X}_{k-1}) \tag{26}$$

$$p(\mathcal{X}_k|\mathcal{Z}_{1:k}^{[1:Q]}) = \frac{\prod_{q=1}^{Q} g_q(\mathcal{Z}_k^{[q]}|\mathcal{X}_k)p(\mathcal{X}_k|\mathcal{Z}_{1:k-1}^{[1:Q]})}{\int_{\mathcal{F}(\mathbb{R}^n)} \prod_{q=1}^{Q} g_q(\mathcal{Z}_k^{[q]}|\mathcal{X}_k)p(\mathcal{X}_k|\mathcal{Z}_{1:k-1}^{[1:Q]})\mu(d\mathcal{X}_k)}. \tag{27}$$

where $\mathcal{F}(\mathbb{R}^n)$ is the class of all finite subsets of $\mathbb{R}^n$, and μ is a measure on $\mathcal{F}(\mathbb{R}^n)$; see [18], [35], [36] for the details.

The next section considers the implementation of (26) and (27) using SMC.

B. Sequential Monte Carlo Implementation

The Bayes recursion in (26) and (27) can be computed, in an approximate manner, by applying an RFS SMC method [18], [37]. In the single-speaker scenario, SMC has been shown [13], [14] to be effective in handling the nonlinearity of the TDOA function. In this multi-speaker extension where the p.d.f.s. exhibit even more complicated structures (c.f., the Appendix), the SMC implementations are particularly favorable.

The implementation employed here is the RFS bootstrap SMC method, which is a special case of the generic RFS SMC method in [18] but is particularly easy to use (note that the RFS bootstrap SMC method here is not related to the method in [37]). The RFS boostrap SMC method is briefly described as follows. We use a random measure $\{\mathcal{X}_k^{(i)}, w_k^{(i)}\}_{i=1}^{L}$ to approximate the posterior p.d.f.:

$$p(\mathcal{X}_k|\mathcal{Z}_{1:k}^{[1:Q]}) \approx \sum_{i=1}^{L} w_k^{(i)} \delta_{\mathcal{X}_k^{(i)}}(\mathcal{X}_k). \tag{28}$$

Here, $\mathcal{X}_k^{(i)}$ is the ith (finite set) particle, $w_k^{(i)}$ is the weight associated with $\mathcal{X}_k^{(i)}$, L is the number of particles applied, and $\delta_{\mathcal{X}_k^{(i)}}$ is a set-valued version of the standard Dirac delta function[2]. As an approximation to probability densities, the weights have the properties that $w_k^{(i)} \geq 0$ for all i and $\sum_{i=1}^{L} w_k^{(i)} = 1$. The particles $\{\mathcal{X}_k^{(i)}\}_{i=1}^{L}$ are randomly drawn conditioned on the time $k-1$ random measure $\{\mathcal{X}_{k-1}^{(i)}, w_{k-1}^{(i)}\}_{i=1}^{L}$. Specifically, for each $i = 1, \ldots, L$ we generate

$$\mathcal{X}_k^{(i)} \sim f(.|\mathcal{X}_{k-1}^{(i)}). \tag{29}$$

[2] A set-valued Dirac delta function $\delta_{\mathcal{Y}}(\mathcal{X})$ is a function such that given every $\mathcal{A} \subseteq \mathcal{F}(\mathbb{R}^n)$, we have $\int_{\mathcal{A}} \delta_{\mathcal{Y}}(\mathcal{X})\mu(d\mathcal{X}) = \mathbf{1}_{\mathcal{A}}(\mathcal{Y})$. Here, $\mathbf{1}_{\mathcal{A}}(\mathcal{Y})$ is an indicator function where $\mathbf{1}_{\mathcal{A}}(\mathcal{Y}) = 1$ if $\mathcal{Y} \in \mathcal{A}$, and $\mathbf{1}_{\mathcal{A}}(\mathcal{Y}) = 0$ otherwise.

By the notion of importance sampling [33], [34] and by applying the time $k-1$ counterpart of (28) to the Bayes recursion in (26) and (27), one can obtain

$$p(\mathcal{X}_k|\mathcal{Z}_{1:k}^{[1:Q]}) \approx \sum_{i=1}^{L} \frac{\prod_{q=1}^{Q} g_q(\mathcal{Z}_k^{[q]}|\mathcal{X}_k^{(i)}) w_{k-1}^{(i)}}{\sum_{j=1}^{L}\prod_{q=1}^{Q} g_q(\mathcal{Z}_k^{[q]}|\mathcal{X}_k^{(j)}) w_{k-1}^{(j)}} \delta_{\mathcal{X}_k^{(i)}}(\mathcal{X}_k). \tag{30}$$

Table I summarizes the RFS bootstrap SMC filter. It should be noted that in Table I there is an additional step called resampling. This step is used to reduce the degeneracy problem commonly encountered in SMC approximations; see [33], [34] for more details. The essential ingredients for the bootstrap SMC filter are the particle generation at Step 1 of Table I, and the expressions for $g_q(\mathcal{Z}_k|\mathcal{X}_k)$. In Table II we show the particle generation algorithm. The general expression for $g_q(\mathcal{Z}_k|\mathcal{X}_k)$ is shown in the Appendix. Some useful equations of $g_q(\mathcal{Z}_k|\mathcal{X}_k)$ for $N_{max}=2$ are shown at the top of the next page.

TABLE I
RFS BOOTSTRAP SMC FILTER FOR MULTI-SPEAKER TRACKING.

```
**RFS Bootstrap SMC Filter**
**Given** a particle size $L$.
**for** $k = 1, 2, \ldots$
    **Step 1. Sampling:**
    For $i = 1, \ldots, L$, generate $\mathcal{X}_k^{(i)} \sim f(.|\mathcal{X}_{k-1}^{(i)})$ and compute
        $w_k^{(i)} = \prod_{q=1}^{Q} g_q(\mathcal{Z}_k^{[q]}|\mathcal{X}_k^{(i)}) w_{k-1}^{(i)}$.
    Then, apply normalization $w_k^{(i)} := w_k^{(i)}/(\sum_{\ell=1}^{L} w_k^{(\ell)})$ for all $i$.
    **Step 2. Resampling:**
    Apply a resampling algorithm [33],[34] on $\{\mathcal{X}_k^{(i)}, w_k^{(i)}\}_{i=1}^{L}$ to obtain a resampled set $\{\tilde{\mathcal{X}}_k^{(i)}, \tilde{w}_k^{(i)}\}_{i=1}^{L}$.
    Then, update $\{\mathcal{X}_k^{(i)}, w_k^{(i)}\}_{i=1}^{L} := \{\tilde{\mathcal{X}}_k^{(i)}, \tilde{w}_k^{(i)}\}_{i=1}^{L}$.
**end**
```

We should point out that Eqs. (31), (32), (33) represent the likelihood functions for no speaker, one speaker, and two speakers, respectively. Also, recall that the parameter λ_c in (31) to (33) represents the average number of false TDOA measurement. Some further remarks are now in order:

Remark 4: An asymptotic convergence property for the RFS SMC filter, such as the above described bootstrap filter has been considered in [18]. Specifically, it has been proven that for sufficiently large L, the mean square approximation error of the RFS SMC filter is inversely proportional to L^α for some constant $0 < \alpha \leq 1$. This implies that the RFS bootstrap SMC filter is an accurate approximation for large L.

Remark 5: If we choose $N_{max} = 1$, $P_{death} = 0$, and $P_{birth} = 1$, the RFS SMC filter reduces to a form very similar to the single-speaker SMC filter in [13], [14].

Remark 6: The computational complexity of the RFS bootstrap SMC filter is linearly dependent on the particle size L. Moreover, for each particle, the complexity depends on the evaluation of the likelihood function $g_q(\mathcal{Z}_k|\mathcal{X}_k)$. It can be seen in the Appendix that the computations of $g_q(\mathcal{Z}_k|\mathcal{X}_k)$ are exponential in $|\mathcal{X}_k|$. In this application where $|\mathcal{X}_k|$ are small (see the argument in Section I and *Remark 1*), this computational issue is insignificant.

TABLE II
PARTICLE GENERATION.

```
**Particle Generation Algorithm for** $\mathcal{X}_k \sim f(.|\mathcal{X}_{k-1})$
Set $\mathcal{X}_k = \emptyset$.
**Step 1. Source Death and Survival:**
**for** each $\mathbf{x}_{k-1} \in \mathcal{X}_{k-1}$
    Draw a random number $u$ uniformly distributed over $[0, 1)$.
    **if** $u > P_{death}$
        draw a random vector $\mathbf{w}_k$ according to the state space model assumed;
        compute $\mathbf{x}_k := \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{w}_k$; and
        set $\mathcal{X}_k := \mathcal{X}_k \cup \{\mathbf{x}_k\}$.
    **end**
**end**
**Step 2. Source Birth**
**if** $|\mathcal{X}_{k-1}^{(i)}| < N_{max}$
    Draw a random number $u$ uniformly distributed over $[0, 1)$.
    **if** $u \leq P_{birth}$
        draw an initial state $\mathbf{b}_k$ according to the initial state distribution assumed; and
        set $\mathcal{X}_k := \mathcal{X}_k \cup \{\mathbf{b}_k\}$.
    **end**
**end**
```

V. REFINEMENT OF THE BAYES RFS FILTER

In this section we present some additional ideas that can further enhance the effectiveness of the proposed RFS multi-speaker tracking method. The first subsection describes the track association problem arising in the RFS framework. A simple method, called *track labeling*, is proposed to handle that problem. Then, in the second subsection, we propose a state estimation scheme that takes advantage of track label information to simplify the estimation process.

A. Track Association using Track Labeling

A problem with the RFS state formulation described in the previous sections is that it gives no information on the track association between $\mathcal{X}_k$ and $\mathcal{X}_{k-1}$. That is, given an element $\mathbf{x}_{k-1} \in \mathcal{X}_{k-1}$, we do not know which element in $\mathcal{X}_k$ is originated from $\mathbf{x}_{k-1}$. It follows that a Bayesian RFS filter based on this model will not provide such information. For the general RFS multi-object tracking scenario in which target birth can be quite complex[3], handling track association is non-trivial; see, for example, [29], [30]. In this multi-speaker tracking problem where at most one speaker source is allowed to be born at one time, track association can be quite easily handled by considering the following idea.

[3] In a general RFS multi-object tracking framework, multiple targets can be born at one time. In addition, one target can split to form two or more .

$$g_q(\mathcal{Z}_k^{[q]}|\emptyset) = e^{-\lambda_c}\left(\frac{\lambda_c}{2\tau_{max}}\right)^{|\mathcal{Z}_k^{[q]}|} \tag{31}$$

$$g_q(\mathcal{Z}_k^{[q]}|\{\mathbf{x}_k\}) = g_q(\mathcal{Z}_k^{[q]}|\emptyset)\left(P_{miss} + (1-P_{miss})\sum_{z_k^{[q]}\in\mathcal{Z}_k^{[q]}}\left(\frac{2\tau_{max}}{\lambda_c}\right)g_q(z_k^{[q]}|\mathbf{x}_k)\right) \tag{32}$$

$$\begin{aligned} g_q(\mathcal{Z}_k^{[q]}|\{\mathbf{x}_{1,k},\mathbf{x}_{2,k}\}) = g_q(\mathcal{Z}_k^{[q]}|\emptyset)\Bigg\{\prod_{i=1,2}\left(P_{miss} + (1-P_{miss})\sum_{z_k^{[q]}\in\mathcal{Z}_k^{[q]}}\left(\frac{2\tau_{max}}{\lambda_c}\right)g_q(z_k^{[q]}|\mathbf{x}_{i,k})\right) \\ -(1-P_{miss})^2\sum_{z_k^{[q]}\in\mathcal{Z}_k^{[q]}}\left(\frac{2\tau_{max}}{\lambda_c}\right)^2 g_q(z_k^{[q]}|\mathbf{x}_{1,k})g_q(z_k^{[q]}|\mathbf{x}_{2,k})\Bigg\} \end{aligned} \tag{33}$$

To avoid notational inconsistency, let us re-define the state vector used in Sections II and III to be $\boldsymbol{\xi}_k = [\ \boldsymbol{\alpha}_k^T, \boldsymbol{\phi}_k^T\ ]^T \in \mathbb{R}^n$, in place of $\mathbf{x}_k$. We define a new state vector

$$\mathbf{x}_k = [\ \boldsymbol{\xi}_k^T, \gamma_k\ ]^T \tag{34}$$

where we augment the state vector by a variable γ_k to indicate the track identity of the speaker state. The variable γ_k is set to the birth time of the speaker source. Since no two speakers share the same birth time, γ_k will provide adequate information for resolving track association when $\mathbf{x}_k$ in (34) is used in the RFS framework. We call γ_k a *track label* of a speaker source. Moreover, we refer to a state vector as the *track-*ℓ speaker state if its track label γ_k takes the value ℓ.

To incorporate track labels into the previously developed RFS framework, we only need minor modifications on the state space equations. For the birth hypothesis process in (21), the birth state vector is modified as

$$\mathbf{b}_k = [\ \boldsymbol{\xi}_{init,k}^T, k\ ]^T \tag{35}$$

where $\boldsymbol{\xi}_{init,k} \in \mathbb{R}^n$ is the random initial state vector described in Section III-B. For the survival process in (20), we have

$$\mathbf{x}_{i,k} = \begin{bmatrix}\mathbf{A} & 0\\ 0 & 1\end{bmatrix}\mathbf{x}_{i,k-1} + \begin{bmatrix}\mathbf{B}\\ 0\end{bmatrix}\mathbf{w}_{i,k} \tag{36}$$

in which the track label at time $k-1$ is directly carried forward to time k.

Track labeling not only helps identify speaker tracks, it also simplifies the state estimation process as shown in the following subsection.

B. State Estimation Incorporating Track Labels

Our algorithm development in the last section has focused on the particle posterior density approximation:

$$p(\mathcal{X}_k|\mathcal{Z}_{1:k}^{[1:Q]}) \approx \sum_{i=1}^{L} w_k^{(i)}\delta_{\mathcal{X}_k^{(i)}}(\mathcal{X}_k) \tag{37}$$

for some weights $w_k^{(i)}$ and for some (set-valued) particles $\mathcal{X}_k^{(i)}$. This subsection describes our proposed method for estimating $\mathcal{X}_k$ from (37), in which the track labeling idea in the previous subsection is exploited to simplify the estimation process.

In the current RFS framework (i.e., without track labeling), a number of Bayesian estimation criteria have been proposed [15], [27]. Here we are interested in the *intensity measure* [16], [18]. [4] The following quantity

$$N_k(\mathcal{S}) \triangleq \mathrm{E}\{|\mathcal{X}_k \cap \mathcal{S}|\,|\mathcal{Z}_{1:k}^{[1:Q]}\} \tag{38}$$

$$= \int |\mathcal{X}_k \cap \mathcal{S}| p(\mathcal{X}_k|\mathcal{Z}_{1:k}^{[1:Q]})\mu(d\mathcal{X}_k) \tag{39}$$

defined for any set $\mathcal{S} \subseteq \mathbb{R}^n$, is called an intensity measure of $\mathcal{X}_k$ conditioned on $\mathcal{Z}_{1:k}^{[1:Q]}$. The intensity measure is the first-order moment of $\mathcal{X}_k$. Physically, $N_k(\mathcal{S})$ describes the expected number of state vectors lying in $\mathcal{S}$; e.g., $N_k(\mathbb{R}^n)$ is the expected total number of speaker sources at time k, given the measurements $\mathcal{Z}_{1:k}^{[1:Q]}$. From (37), $N_k(\mathcal{S})$ can be approximated by

$$N_k(\mathcal{S}) \approx \sum_{i=1}^{L} w_k^{(i)}|\mathcal{X}_k^{(i)} \cap \mathcal{S}| \tag{40}$$

Roughly speaking, an intensity-measure-based state estimation method [18], [38], [39] consists of two steps: i) Obtain an estimate of the number of speakers $\hat{N}_k = \lceil N_k(\mathbb{R}^n)\rfloor$ where $\lceil . \rfloor$ is the rounding operation. ii) Determine a number of sets $\mathcal{S}_{i,k}$ for $i = 1, \ldots, \hat{N}_k$, such that the intensity $N_k(\mathcal{S}_{i,k})$ shows good response for each i whilst $\mathcal{S}_{i,k} \cap \mathcal{S}_{j,k} = \emptyset$ for $i \neq j$ and $\sum_{i=1}^{\hat{N}_k} N_k(\mathcal{S}_{i,k}) = N_k(\mathbb{R}^n)$. iii) For each i, determine the center of $\mathcal{S}_{i,k}$, denoted by $\hat{\mathbf{x}}_{i,k}$. The centers $\{\hat{\mathbf{x}}_{i,k}\}_{i=1}^{\hat{N}_k}$ are then taken as the state estimates. The challenge of this approach lies in Step ii), where some clustering algorithm is usually used to numerically determine those sets. Since clustering is a nonlinear nonconvex optimization problem, poor data fitting could occur.

The state estimation process can become simpler when track label information is available. Recall that a state vector with track labeling is in the form of $\mathbf{x}_k = [\ \boldsymbol{\xi}_k^T, \gamma_k\ ]^T \in \mathbb{R}^n \times \mathbb{Z}$. Hence, we can define the intensity measure for the track-ℓ

[4] The density of the intensity measure is called the probability hypothesis density (PHD). It is worth mentioning that PHD is an important concept in RFS multi-object tracking; see [16], [18] for the details.

speaker state:

$$N_k(\mathcal{A};\ell) \triangleq N_k(\mathcal{A}\times\{\ell\}) \tag{41}$$

$$= \int \sum_{\substack{[\ \boldsymbol{\xi}_k^T,\gamma_k\ ]^T\in\mathcal{X}_k \\ \gamma_k=\ell}} |\{\boldsymbol{\xi}_k\}\cap\mathcal{A}|\,p(\mathcal{X}_k|\mathcal{Z}_{1:k}^{[1:Q]})\mu(d\mathcal{X}_k) \tag{42}$$

for any $\mathcal{A}\subseteq\mathbb{R}^n$, and for $\ell\le k$. This track-label-dependent intensity measure allows us to perform state estimation on a speaker-by-speaker basis. First, we note that

Interpretation 1 The quantity $N_k(\mathbb{R}^n;\ell)$ is the expected number of times that the track-ℓ speaker source is present at time k, given the measurements $\mathcal{Z}_{1:k}^{[1:Q]}$.

In other words, we can detect the track-ℓ source by testing whether $N_k(\mathbb{R}^n;\ell)$ is above certain threshold, say 0.5.

Interpretation 2 The vector

$$\hat{\boldsymbol{\xi}}_k(\ell) = \frac{1}{N_k(\mathbb{R}^n;\ell)}\int_{\mathbb{R}^n}\boldsymbol{\xi}_k N_k(d\boldsymbol{\xi}_k;\ell) \tag{43}$$

is the expected state vector of the track-ℓ source at time k, conditioned on the hypothesis that the track-ℓ speaker source is present at time k, and conditioned on $\mathcal{Z}_{1:k}^{[1:Q]}$.

It is interesting to note that (43) is reminiscent of the expected *a posteriori* (EAP) estimate in the single-object tracking scenario.

Based on Interpretations 1 and 2, we propose an RFS state estimation procedure in Table III.

TABLE III

STATE ESTIMATION ALGORITHM WITH TRACK LABELING.

```
RFS state estimation algorithm
Given a random measure $\{\mathcal{X}_k^{(i)}, w_k^{(i)}\}_{i=1}^L$ at time $k$.
Set $\hat{\mathcal{X}}_k = \emptyset$.
Step 1.  Extract the track label set
        $\mathcal{I}_k = \bigcup_{i=1}^L \bigcup_{[\ \boldsymbol{\xi}_k^T,\gamma_k\ ]^T\in\mathcal{X}_k^{(i)}} \{\gamma_k\}$
Step 2.
for each $\ell\in\mathcal{I}_k$
    Obtain a particle approximation to $N_k(\mathbb{R}^n;\ell)$, denoted by
    $\hat{N}_k(\ell)$, by summing the weights associated with the track-
    $\ell$ speaker source:
        $\hat{N}_k(\ell) = \sum_{i=1}^L w_k^{(i)} \sum_{[\ \boldsymbol{\xi}_k^T,\gamma_k\ ]^T\in\mathcal{X}_k^{(i)}} \mathbf{1}\{\gamma_k=\ell\}$
    If $\hat{N}_k(\ell)\ge 0.5$, compute a particle approximation to
    $\boldsymbol{\xi}_k(\ell)$, denoted by $\hat{\boldsymbol{\xi}}_k(\ell)$, by making a particle weighted
    average
        $\hat{\boldsymbol{\xi}}_k(\ell) = \frac{1}{\hat{N}_k(\ell)} \sum_{i=1}^L w_k^{(i)} \sum_{[\ \boldsymbol{\xi}_k^T,\gamma_k\ ]^T\in\mathcal{X}_k^{(i)}} \mathbf{1}\{\gamma_k=\ell\}\boldsymbol{\xi}_k,$
    and then update $\hat{\mathcal{X}}_k := \hat{\mathcal{X}}_k \cup \{[\ \hat{\boldsymbol{\xi}}_k^T(\ell), \ell\ ]^T\}$.
end
```

VI. SIMULATION RESULTS

Two room simulation examples are used to test the tracking performance of the proposed multi-speaker RFS SMC filter.

A. Example 1

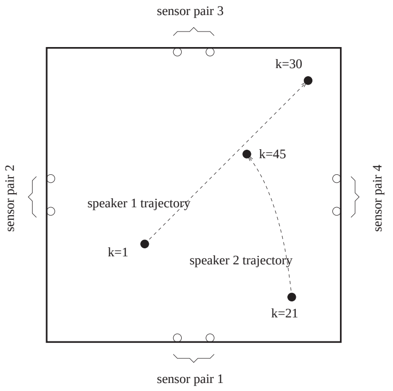

Fig. 3. Geometric settings for the room simulation in Example 1.

Fig. 3 illustrates the room settings for this example. The dimensions of the enclosure are 3m × 3m × 2.5m. We employ four microphone pairs, each of which has an inter-sensor spacing of 0.5m (which corresponds to $\tau_{max}=1.5$ms). Fig. 3 also shows the trajectories and birth/death times of the speaker sources. The speaker sources are all female. The acoustic image method [40] was used to simulate the room impulse responses. The reverberation time of the room impulse responses is about $T_{60}=0.15$s (see the literature such as [7], [14] for the definition of T_{60}). The speech-signal-to-noise ratio is about 20dB. The time frame length for measuring TDOAs is 128ms, and the time frames are non-overlapping. Fig. 4 plots the measured TDOAs against the time frame index (we only displayed the measured TDOAs for two of the microphone pairs due to page limitation). We can see that the measured data is not very informative: For each time frame the largest GCC peak does not always represent one of the true TDOAs. Moreover, in the presence of two active speakers (from time 20 to 30), the accuracy of the measured TDOAs tend to deteriorate due to mutual interference between the two speech signals.

The parameter settings for the RFS SMC filter are as follows. The state space model is the Langevin model [cf., Eqs. (8) and (9)], with the model parameters $\beta = 10\text{s}^{-1}$ and $\nu = 1\text{ms}^{-1}$. The standard deviation of the TDOA measurement error is $\sigma_v = 125\mu s$ (which is also the sampling period). The other parameters are $N_{max}=2$, $P_{birth}=0.05$, $P_{death}=0.01$, $P_{miss}=0.25$, $\lambda_c=3$, and $L=500$. Fig. 5 illustrates the tracking performance of the multi-speaker RFS SMC filter. The figures show that the RFS SMC filter is able to determine the two speakers' locations and their respective activity intervals. Recall that in the legend of Fig. 5(b), the

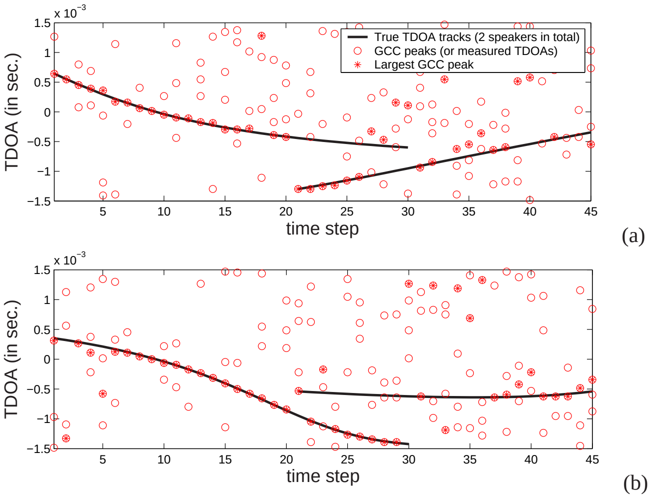

Fig. 4. Measured TDOAs at (a) sensor pair 1, and (b) sensor pair 3.

term ‘track label’ also represents the birth time of the estimated speaker track. From Figs. 5(b)–(c) we can see that the RFS SMC filter produces two tracks with track labels 21 and 22, but these two tracks actually correspond to the same speaker. This is because the RFS SMC filter can have estimation error on the birth time variables. For the readers’ interest, Figs. 5(b)–(c) also show the performance of the existing single-speaker SMC filter [13], [14].

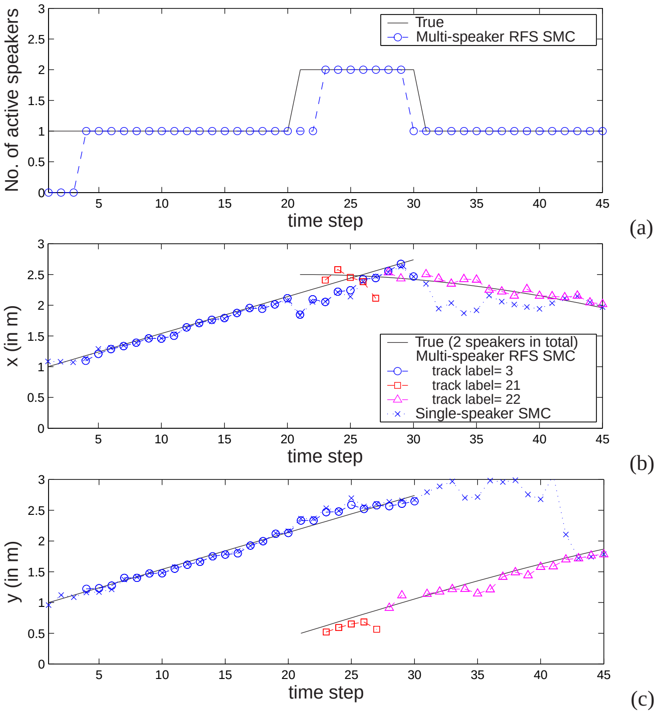

Fig. 5. Location tracking performance in Example 1. (a) RFS SMC filter estimates of the number of active speakers. (b)–(c) Position estimates of the RFS SMC filter and the conventional single-speaker SMC filter.

B. Example 2

This example considers a situation where some model assumptions are not well satisfied. In other words, we are interested in testing the robustness of the proposed method against model mismatch. The room setting is shown in Fig. 6, where the room dimensions are 5m × 3.5m × 2.5m. The reverberation time is about $T_{60} = 0.35$s, and the rest of the simulation parameters are the same as those in the previous example. In this example all the speakers are stationary, which violates the assumption that the speakers are moving following the Langevin motion model. Another model mismatch is with the measured TDOAs, which are illustrated in Fig. 7. In Fig. 7(b) we observe that from time 11 to 60, there is a false TDOA that persistently appears with a value of about 1×10^{-3} second. One can also find a few other persistent false TDOA tracks in the figures. Those false TDOAs are caused by room reverberation. Since the speaker positions are fixed, so do those reverberation-induced false TDOAs. This phenomenon violates the assumption that false TDOAs are time uncorrelated.

In this example we increase the number of microphone pairs to 8. The rationale is that the effect of model mismatch might be reduced when more sensors are available. Fig. 8 shows the localization results of the proposed multi-speaker RFS SMC filter. The figures indicate that inaccurate position estimates do happen sometimes; e.g., the track from time 59 to 61 with track label 11. But it is also seen from the figures that the proposed

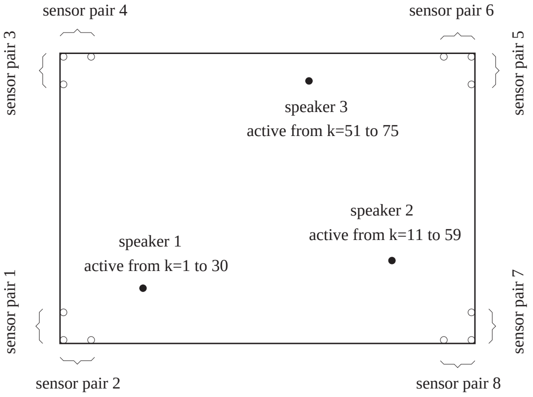

Fig. 6. Geometric settings for the room simulation in Example 2.

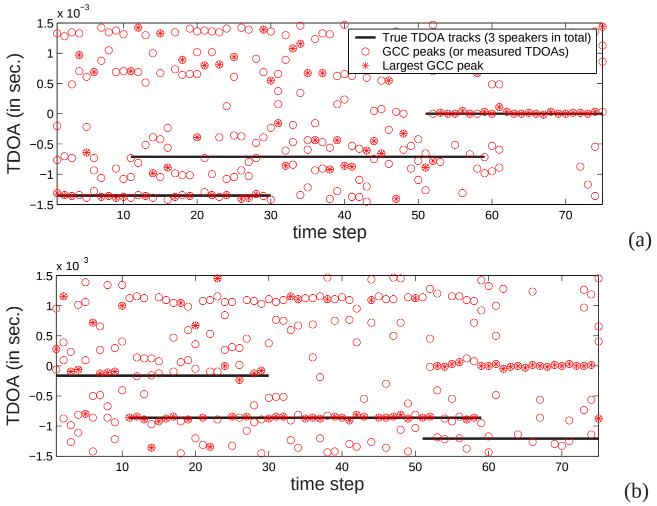

Fig. 7. Measured TDOAs at (a) sensor pair 3, and (b) sensor pair 7.

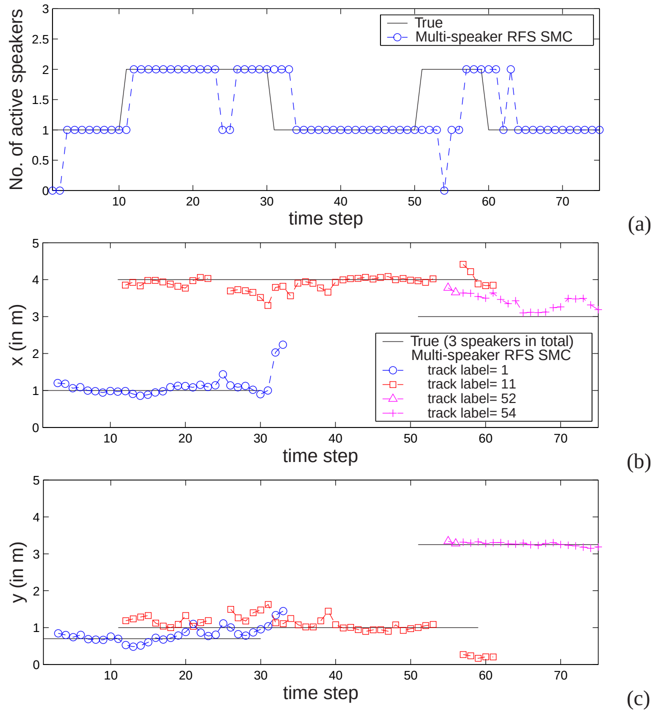

Fig. 8. Location tracking performance of the multi-speaker RFS SMC filter in Example 2. (a) Estimates of the number of active speakers. (b)–(c) Position estimates.

method provides reasonable tracking performance on average.

C. Average Performance

The above two examples show the localization performance for one trial. In this subsection we are interested in the localization performance averaged over many trials. To do so it is important to consider measures for comparing the differences between the true finite set state $\mathcal{X}_k$ and the estimated state set $\hat{\mathcal{X}}_k$. First, it is useful to evaluate the probability of correct speaker number estimation:

$$P[\ |\hat{\mathcal{X}}_k| = |\mathcal{X}_k|\ ] \tag{44}$$

Second, we are concerned with the location errors for the state vectors in $\hat{\mathcal{X}}_k$. When the speaker number estimate is incorrect such that $|\mathcal{X}_k| \neq |\hat{\mathcal{X}}_k|$, defining a localization error is a problem on its own; see [41]. Now, let us suppose that $|\mathcal{X}_k| = |\hat{\mathcal{X}}_k| = n$, and that $\mathcal{X}_k = \{\mathbf{x}_{1,k}, \ldots, \mathbf{x}_{n,k}\}$, $\hat{\mathcal{X}}_k = \{\hat{\mathbf{x}}_{1,k}, \ldots, \hat{\mathbf{x}}_{n,k}\}$. We consider the following multi-speaker distance error:

$$d(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \min_{\substack{j_i \in \{1,\ldots,n\}, i=1,\ldots,n \\ j_i \neq j_k, \forall i \neq k}} \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{C}\mathbf{x}_{i,k} - \mathbf{C}\hat{\mathbf{x}}_{j_i,k}\|^2} \tag{45}$$

where $\mathbf{C} = [\mathbf{I}\ \mathbf{0}]$ is such that given a state $\mathbf{x}_k$, $\mathbf{C}\mathbf{x}_k$ outputs the (x, y) position of that state. The idea of the minimization in (45) is to find a proper assignment between elements in $\mathcal{X}_k$ and $\hat{\mathcal{X}}_k$. Moreover, we should mention that theoretically, (45) is a special case of the *Wasserstein* distance [41]. With (45), we can measure a *conditional mean distance error*, given by

$$\mathrm{E}_{\hat{\mathcal{X}}_k}\{d(\mathcal{X}_k, \hat{\mathcal{X}}_k)|\text{correct speaker number estimate}\} \tag{46}$$

The performance measures (44) and (46) were evaluated for Examples 1 and 2 with $1,000$ trials. The results for Examples 1 and 2 are shown in Figs. 9 and 10, respectively. The figures illustrate that at the time instants where source birth/death occurs, the RFS method yields a transient behavior: At those birth/death time instants, the probability of correct speaker number estimation decreases and the conditional mean distance error increases. Then, the localization performance improves gradually with time.

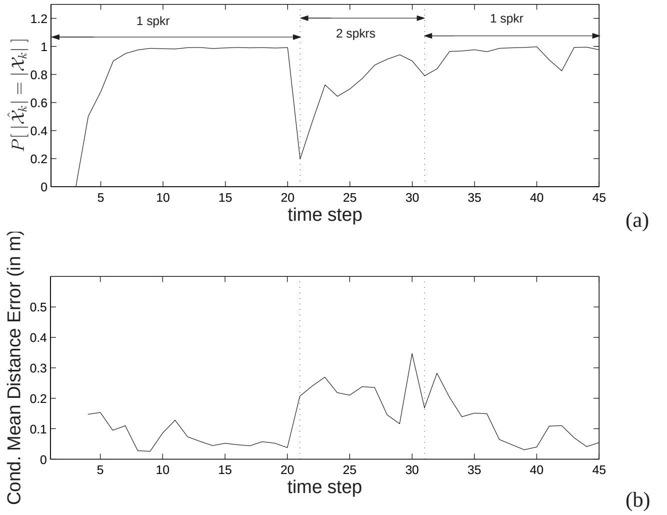

Fig. 9. Average location tracking performance of the multi-speaker RFS SMC filter in Example 1.

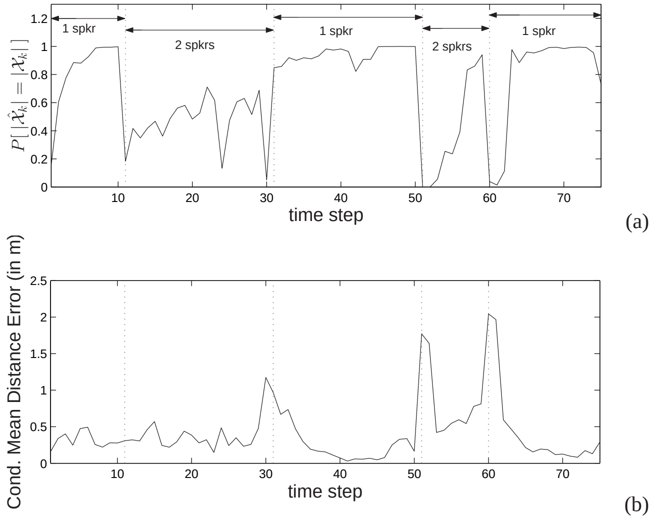

Fig. 10. Average location tracking performance of the multi-speaker RFS SMC filter in Example 2.

VII. Conclusion and Discussion

Using the RFS theory and the SMC implementation technique, we have developed a TDOA multi-speaker location tracking algorithm that can handle unknown, time-varying number of active speakers. We have used simulations to show that the proposed algorithm can correctly determine not only the speaker locations, but also the voice activity interval for each speaker.

The proposed RFS algorithm is suitable for many speech applications where the number of active speakers is usually small. As a technical challenge, it will be worthwhile to examine the case of large number of speakers. This direction leads to several open questions. First, our method (as well as the other methods in [13], [14], [20]) has been relying on the GCC TDOA measurement scheme, which has a modest resolution that generally cannot handle a large number of active speakers. To deal with the case of large number of speakers, it appears that we need to employ some more sophisticated microphone array structures and signal processing methods, such as those in the direction-of-estimation (DOA) estimation context [1]. Second, the RFS multi-speaker tracking principle is applicable to any number of speakers. However, the RFS Bayesian filter becomes more expensive to implement as the number of speaker increases. In those cases it might be appropriate to apply approximations such as the first-order moment method [16], [28]. Third, it will be interesting to extend the present method to deal with more complicated situations, such as when multiple source births are allowed at one time instant.

APPENDIX

The purpose of this section is to illustrate, in a concise manner, the derivations of the set-valued state transition density $f(\mathcal{X}_k|\mathcal{X}_{k-1})$ and the set-valued likelihood $g_q(\mathcal{Z}_k|\mathcal{X}_k)$. The principles of the derivations essentially follow those described in [15]. Readers are referred to [15] for further details. The following lemma will be frequently used:

Lemma 1 *[15] Consider*

$$\mathcal{C} = \mathcal{A} \cup \mathcal{B} \tag{47}$$

where $\mathcal{A}$ and $\mathcal{B}$ are two independent RFSs. Then, the p.d.f. of $\mathcal{C}$ is

$$p(\mathcal{C}) = \sum_{\tilde{\mathcal{C}} \subseteq \mathcal{C}} p(\mathcal{A} = \tilde{\mathcal{C}}) p(\mathcal{B} = \mathcal{C} - \tilde{\mathcal{C}}) \tag{48}$$

A. The State Transition Density

Consider the finite set state structure in (19). By applying Lemma 1 to (19), the state transition density is given by:

$$f(\mathcal{X}_k|\mathcal{X}_{k-1}) = \sum_{\tilde{\mathcal{X}}_k \subseteq \mathcal{X}_k} f_b(\tilde{\mathcal{X}}_k|\mathcal{X}_{k-1}) f_s(\mathcal{X}_k - \tilde{\mathcal{X}}_k|\mathcal{X}_{k-1}) \tag{49}$$

where

$$f_b(\mathcal{X}_k|\mathcal{X}_{k-1}) \triangleq p(\mathcal{B}_k(\mathbf{b}_k) = \mathcal{X}_k|\mathcal{X}_{k-1}) \tag{50}$$

is the p.d.f. for the birth states, and

$$f_s(\mathcal{X}_k|\mathcal{X}_{k-1}) \triangleq p\left(\bigcup_{i=1,\ldots,|\mathcal{X}_{k-1}|} \mathcal{S}_k(\mathbf{x}_{i,k-1}, \mathbf{w}_{i,k}) = \mathcal{X}_k \middle| \mathcal{X}_{k-1} \right) \tag{51}$$

is the transition density for the previous states.

The expression for the birth state p.d.f. is as follows. For $|\mathcal{X}_{k-1}| = N_{max}$ where no speaker birth is allowed, we have that

$$f_b(\mathcal{X}_k|\mathcal{X}_{k-1}) = \begin{cases} 1, & \mathcal{X}_k = \emptyset \\ 0, & \text{otherwise} \end{cases} \tag{52}$$

As for the case of $|\mathcal{X}_{k-1}| < N_{max}$, it can be shown from (21) that

$$f_b(\mathcal{X}_k|\mathcal{X}_{k-1}) = \begin{cases} 1 - P_{birth}, & \mathcal{X}_k = \emptyset \\ P_{birth}\beta(\mathbf{x}_k), & \mathcal{X}_k = \{\mathbf{x}_k\} \\ 0, & \text{otherwise} \end{cases} \tag{53}$$

where $\beta(\mathbf{x}_k) \triangleq p(\mathbf{b}_k = \mathbf{x}_k)$ is the initial state distribution.

To construct the density $f_s(\mathcal{X}_k|\mathcal{X}_{k-1})$, it is instructive to consider a one-speaker set-valued state transition density

$$\begin{aligned} f_{s,i}(\mathcal{X}_k|\mathcal{X}_{k-1}) &\triangleq p(\mathcal{S}_k(\mathbf{x}_{i,k-1}, \mathbf{w}_{i,k}) = \mathcal{X}_k|\mathcal{X}_{k-1}) \\ &= p(\mathcal{S}_k(\mathbf{x}_{i,k-1}, \mathbf{w}_{i,k}) = \mathcal{X}_k|\mathbf{x}_{i,k-1}) \end{aligned} \tag{54}$$

where $\mathbf{x}_{i,k-1}$ is an element in $\mathcal{X}_k$ with $\mathbf{x}_{i,k-1} \neq \mathbf{x}_{j,k-1}$ for $i \neq j$. From (20), it is shown that

$$f_{s,i}(\mathcal{X}_k|\mathcal{X}_{k-1}) = \begin{cases} P_{death}, & \mathcal{X}_k = \emptyset \\ (1 - P_{death}) f(\mathbf{x}_k|\mathbf{x}_{i,k-1}), & \mathcal{X}_k = \{\mathbf{x}_k\} \\ 0, & \text{otherwise} \end{cases} \tag{55}$$

where $f(\mathbf{x}_k|\mathbf{x}_{k-1})$ is the single-speaker, vector-valued p.d.f. considered in Section II-B. Let

$$\mathcal{X}_k = \{\mathbf{x}_{1,k}, \ldots, \mathbf{x}_{m,k}\}, \quad \mathcal{X}_{k-1} = \{\mathbf{x}_{1,k}, \ldots, \mathbf{x}_{n,k-1}\}$$

with $m \leq n$. By applying Lemma 1 to (51) repeatedly and by exploiting (55), it can be shown that

$$\begin{aligned} f_s(\mathcal{X}_k|\mathcal{X}_{k-1}) =& P_{death}^{n-m}(1 - P_{death})^m \times \\ & \sum_{1 \leq i_1 \neq i_m \leq n} \prod_{j=1}^{m} f(\mathbf{x}_{j,k}|\mathbf{x}_{i_j,k-1}) \end{aligned} \tag{56}$$

where the summation term in the above equation means that

$$\sum_{1 \leq i_1 \neq \ldots \neq i_m \leq n} = \sum_{i_1=1}^{n} \sum_{\substack{i_2=1 \\ i_2 \neq i_1}}^{n} \cdots \sum_{\substack{i_m=1 \\ i_m \neq i_{m-1} \neq \ldots \neq i_1}}^{n} \tag{57}$$

B. The Likelihood Functions

The ideas behind deriving the likelihood functions are similar to those in the previous subsection. By applying Lemma 1 to the measurement model in (22), the likelihood function for the TDOAs of the qth microphone pair is shown to be

$$g_q(\mathcal{Z}_k^{[q]}|\mathcal{X}_k) = \sum_{\tilde{\mathcal{Z}}_k^{[q]} \subseteq \mathcal{Z}_k^{[q]}} g_{true,q}(\tilde{\mathcal{Z}}_k^{[q]}|\mathcal{X}_k) c_q(\mathcal{Z}_k^{[q]} - \tilde{\mathcal{Z}}_k^{[q]}) \tag{58}$$

where

$$g_{true,q}(\mathcal{Z}_k^{[q]}|\mathcal{X}_k) \triangleq p\left(\bigcup_{i=1,\ldots,|\mathcal{X}_k|} \mathcal{T}_k^{[q]}(\mathbf{x}_{i,k}, v_{i,k}^{[q]}) = \mathcal{Z}_k^{[q]} \middle| \mathcal{X}_k \right) \tag{59}$$

is the likelihood function of the true TDOAs, and

$$c_q(\mathcal{Z}_k^{[q]}) \triangleq p(\mathcal{C}_k^{[q]} = \mathcal{Z}_k^{[q]}) \tag{60}$$

is the p.d.f. of the false TDOAs. It can be shown, in a way similar to that for the state transition density in (55) to (56), that for

$$\mathcal{Z}_k^{[q]} = \{z_{1,k}^{[q]}, \ldots, z_{m,k}^{[q]}\}, \quad \mathcal{X}_k = \{\mathbf{x}_{1,k}, \ldots, \mathbf{x}_{n,k}\}$$

with $n \le m$, the true TDOA likelihood function is given by

$$g_{true,q}(\mathcal{Z}_k^{[q]}|\mathcal{X}_k) = P_{miss}^{n-m}(1-P_{miss})^m \times \sum_{1\le i_1 \neq \ldots \neq i_m \le n} \prod_{j=1}^{m} g_q(z_{j,k}^{[q]}|\mathbf{x}_{i_j,k}). \qquad (61)$$

where $g_q(z_k^{[q]}|\mathbf{x}_k) = \mathcal{N}(z_k^{[q]}; \tau_q(\mathbf{C}\mathbf{x}_k), \sigma_v^2)$ is the single-speaker likelihood function described in Section II-B. As for the false TDOA p.d.f., it is shown that

$$c_q(\{z_{1,k}^{[q]}, \ldots, z_{m,k}^{[q]}\}) = P_{|\mathcal{Z}_k^{[q]}|}(m) \left(m! \prod_{i=1}^{m} \kappa(z_{i,k}^{[q]}) \right) \qquad (62)$$

where $P_{|\mathcal{Z}_k^{[q]}|}(m) = P[|\mathcal{Z}_k^{[q]}| = m]$ is the probability of the number of false TDOAs, and $\kappa(z)$ is a uniform density with an interval $[-\tau_{max}, \tau_{max}]$. Under the assumption that the number of false TDOAs is Poisson distributed with an average rate λ_c, we have $P_{|\mathcal{Z}_k^{[q]}|}(m) = e^{-\lambda_c}\lambda_c^m/m!$ and Eq. (62) can be re-expressed as

$$c_q(\mathcal{Z}_k^{[q]}) = e^{-\lambda_c} \prod_{z_k^{[q]} \in \mathcal{Z}_k^{[q]}} \lambda_c \kappa(z_k^{[q]}) \qquad (63)$$

Substituting (63) into (58), the likelihood function can be simplified as

$$g_q(\mathcal{Z}_k^{[q]}|\mathcal{X}_k) = c_q(\mathcal{Z}_k^{[q]}) \sum_{\tilde{\mathcal{Z}}_k^{[q]} \subseteq \mathcal{Z}_k^{[q]}} \frac{g_{true,q}(\tilde{\mathcal{Z}}_k^{[q]}|\mathcal{X}_k)}{\prod_{\tilde{z}_k^{[q]} \in \tilde{\mathcal{Z}}_k^{[q]}} \lambda_c \kappa(\tilde{z}_k^{[q]})} \qquad (64)$$

REFERENCES

[1] H. Krim and M. Viberg, “Two decades of array signal processing research: the parametric approach,” *IEEE Signal Processing Mag.*, vol. 13, no. 4, pp. 67–94, 1996.

[2] H. Wang and M. Kaveh, “Coherent signal-subspace processing for the detection and estimation of angles of arrivals of multiple wide-band sources,” *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 33, no. 4, pp. 823–831, 1985.

[3] B. Friedlander and A. Weiss, “Direction finding for wide-band signals using an interpolated array,” *IEEE Trans. Signal Processing*, vol. 41, no. 4, pp. 1618–1634, 1993.

[4] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, “Speech enhancement based on the subspace method,” *IEEE Trans. Speech Audio Processing*, vol. 8, no. 5, pp. 497–507, 2000.

[5] Y. Huang, J. Benesty, and G. Elko, “Microphone arrays for video camera steering,” in *Acoustic Signal Processing for Telecommunications,* S.L. Gay and J. Benesty, Eds., Kluwer Academic, 2000.

[6] C. Knapp and G. Carter, “The generalized correlation method of estimation of time delay,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, 1976.

[7] T. Gustafsson, B. Rao, and M. Trivedi, “Source localization in reverberant environments: modeling and statistical analysis,” *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 791–803, 2003.

[8] Y. Chan and K. Ho, “A simple and efficient estimator for hyperbolic location,” *IEEE Trans. Signal Processing*, vol. 42, no. 8, pp. 1905–1915, 1994.

[9] M. Brandstein, J. Adcock, and H. Silverman, “A closed-form location estimator for use with room environment microphone arrays,” *IEEE Trans. Speech Audio Processing*, vol. 5, no. 1, pp. 45–50, 1997.

[10] Y. Huang, J. Benesty, G. Elko, and R. Mersereau, “Real-time passive source localization: an unbiased linear-correction least-squares approach,” *IEEE Trans. Speech Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

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

[12] Y. Huang and J. Benesty, “A class of frequency-domain adaptive approaches to blind multi-channel identification,” *IEEE Trans. Signal Processing*, vol. 51, no. 1, pp. 11–24, 2003.

[13] J. Vermaak and A. Blake, “Nonlinear filtering for speaker tracking in noisy and reverberant environments,” in *Proc. 2001 IEEE Intl. Conf. Acoust., Speech, Signal Processing*, May 2001.

[14] D. Ward, E. Lehmann, and R. Williamson, “Particle filtering algorithms for tracking an acoustic source in a reverberant environment,” *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 826–836, 2003.

[15] R. Mahler, *An introduction to Multisource-Multitarget Statistics and Applications*. Lockheed Martin Technical Monograph, 2000.

[16] ——, “Multi-target Bayes filtering via first-order multi-target moments,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1152–1178, 2003.

[17] B.-N. Vo, S. Singh, and A. Doucet, “Sequential Monte Carlo implementation of the PHD filter for multi-target tracking,” in *Proc. Int. Conf. Information Fusion,* Cairns, Australia, pp. 792–799, 2003, also: http://www.ee.unimelb.edu.au/staff/bv/publications.html.

[18] ——, “Sequential Monte Carlo methods for Bayesian multi-target filtering with random finite sets,” *IEEE Trans. Aerospace and Electronic Systems*, vol. 41, no. 4, pp. 1224–1245, 2005, also: http://www.ee.unimelb.edu.au/staff/bv/publications.html.

[19] D. Zotkin and R. Duraiswami, “Accelerated speech source localization via hierarchical search of steer response power,” *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 499–508, 2004.

[20] I. Potamitis, H. Chen, and G. Tremoulis, “Tracking of multiple moving speakers with multiple microphone arrays,” *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 520–529, 2003.

[21] C. Hue, J.-P. L. Cadre, and P. Pérez, “Sequential Monte Carlo methods for multiple target tracking and data fusion,” *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 309–325, 2002.

[22] M. Vihola, *Random Sets for Multitarget Tracking and Data Association*. Licentiate thesis, Dept. Inform. Tech. & Inst. Math., Tampere Univ. Technology, Finland, Aug. 2004.

[23] G. Mathéron, *Random Sets and Integral Geometry*. J. Wiley, 1975.

[24] D. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes*. Springer-Verlag, 1988.

[25] D. Stoyan, D. Kendall, and J. Mecke, *Stochastic Geometry and its Applications*. John Wiley & Sons, 1995.

[26] M. N. van Lieshout, *Markov Point Processes and their Applications*. Imperial College Press, 2000.

[27] I. Goodman, R. Mahler, and H. Nguyen, *Mathematics of Data Fusion*. Kluwer Academic, 1997.

[28] B.-N. Vo, S. Singh, and W.-K. Ma, “Tracking multiple speakers with random sets,” in *Proc. 2004 IEEE Intl. Conf. Acoust., Speech, Signal Processing,* May 2004.

[29] L. Lin, T. Kirubarajan, and Y. Bar-Shalom, “Data association combined with the probability hypothesis density filter for multitarget tracking,” in *Proc. SPIE Conf. on Signal and Data Processing of Small Targets*, Orlando, FL, April 2004.

[30] K. Panta, B.-N. Vo, S. Singh, and A. Doucet, “Probability hypothesis density filter versus multiple hypothesis tracking,” in *Proc. SPIE’2004, Florida*, 2004.

[31] R. Mahler, ““statistics 101” for multisensor, multitarget data fusion,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 19, no. 1, pp. 53–64, Jan. 2004.

[32] B.-N. Vo, W.-K. Ma, and S. Singh, “Localizing an unknown time-varying number of speakers: A Bayesian random finite set approach,” in *Proc. 2005 Int. Conf. Acoust., Speech and Signal Processing*, 2005.

[33] A. Doucet, J. de Freitas, and N. Gordon, “An introduction to sequential Monte Carlo methods,” in *Sequential Monte Carlo Methods in Practice,* A. Doucet, J.F.G. de Freitas, and N.J. Gordon, Eds., Springer-Verlag, 2001.

[34] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, “A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking,” *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[35] C. Geyer, “Likelihood inference for spatial point processes,” in *Stochastic Geometry Likelihood and Computation,* Barndorff-Nielsen *et. al* eds., pp. 79–140, 1999.

[36] J. Moller, "Markov chain Monte Carlo and spatial point processes," in *Stochastic Geometry Likelihood and Computation*, Barndorff-Nielsen *et. al* eds., pp. 141–172, 1999.

[37] H. Sidenbladh and S.-L. Wirkander, "Tracking random sets of vehicles in terrain," in *IEEE Workshop on Multi-Object Tracking*, 2003.

[38] T. Zajic and R. Mahler, "A particle-systems implementation of the PHD multitarget tracking filter," in *Proc. SPIE*, vol. 5096, pp. 291–299, 2003.

[39] B.-N. Vo and W.-K. Ma, "Joint detection and tracking of multiple maneuvering targets in clutters using random finite sets," *Intl. Conf. Control, Automation, Robotics and Vision*, pp. 1485–1490, December 2004, also: http://www.ee.unimelb.edu.au/staff/bv/publications.html.

[40] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, 1979.

[41] J. Hoffman and R. Mahler, "Multitarget miss distance via optimal assignment," *IEEE Trans. Syst. Man Cybernetics*, vol. 34, no. 3, pp. 327–336, 2004.