LOCALIZING AN UNKNOWN TIME-VARYING NUMBER OF SPEAKERS: A BAYESIAN RANDOM FINITE SET APPROACH

Ba-Ngu Vo[†]*, Wing-Kin Ma*[†]*, and Sumeetpal Singh*[‡]

[†]Dept. Electrical & Electronic Eng., University of Melbourne, Vic., Australia
[‡]Dept. Eng., University of Cambridge, U.K.

ABSTRACT

Using time-difference-of-arrival (TDOA) measurements to perform speaker localization has received much interest recently. Motivated by the significant progress in TDOA single speaker localization, this paper presents a TDOA multi-speaker location tracking algorithm based on Bayesian particle filtering. The development is based on the random finite set framework, which provides an effective treatment to the problem of unknown time-varying number of active speakers. The proposed method can be viewed as a generalization of the existing single-speaker particle filter. Using a simulated reverberant room, we demonstrate the tracking capability of the proposed particle filter.

1. INTRODUCTION

Using speech activity to locate speakers is an important problem in microphone array processing, driven by applications such as automatic camera steering in video-conferencing. Challenges in speaker localization include room reverberation effects and multiple speaker voice activities, both of which are considered difficult signal processing problems.

Single-speaker localization techniques have recently seen significant progress [1–3]. In particular, the time-difference-of-arrival (TDOA) based localization approach has been frequently considered. In a TDOA system such as that depicted in Fig. 1, microphones are grouped into pairs and the TDOA (or the inter-sensor signal propagation delay) is measured for each pair. Under the assumption of single, direct path signal propagation, the TDOA can be measured reliably using standard methods such as the generalized cross correlation (GCC) method [4]. The TDOAs for all microphone pairs are then used to estimate the speaker location. Again, standard, simple methods are available for TDOA location estimation; see [1] and the references therein. The problem with this simple approach is that in the presence of reverberation, the GCC method can give anomalous TDOA estimates which are not formed by the direct paths. Recent research reveals two approaches for combating this problem:

I. Replace the GCC estimator by blind channel identification based TDOA estimators, which accounts for the reverberation effects by estimating the whole room impulse responses; see [1] and the references therein.

II. Apply Bayesian filtering [2, 3], which, in the target tracking context, has been shown to be effective in handling false measurements.

This work was supported in part by a research grant awarded by the Australian Research Grant Council.

In general, the Class I approach exhibits simpler structures than the Class II. However, simulation evidences have indicated that the Class II approach can provide better location estimation accuracy compared to the Class I. This is because the Bayesian treatment exploits the correlation of the speaker motion from one time window to another.

The objective of this paper is to extend the Bayesian approach to a multi-speaker scenario in which the voice activity interval for each speaker is unknown and random. This scenario poses a significant challenge in signal processing (and this is also true for the Class I approach). In this paper, we employ the theory of random finite sets (RFSs) to formulate the multiple-speaker localization problem. RFS is a rigorous mathematical discipline for dealing with random spatial patterns [5, 6] that has long been used by statisticians in many diverse applications including agriculture, geology, epidemiology [6], and more recently multi-target tracking [7–9]. Discussions regarding the differences between the RFS and other multi-target tracking techniques can be found in [10, 11]. Our previous work in multi-speaker localization [11] considers a suboptimal Bayesian RFS filter using the first-order moment approximation [8]. In this work we focus on the optimal Bayesian RFS filter. Analogous to the single-speaker work [2, 3], a particle implementation is developed for the RFS filter. In Section 4, we will use a simulated room environment to demonstrate the tracking capability of the proposed method.

2. RFS FORMULATION FOR MULTI-SPEAKER LOCALIZATION

Before describing the RFS formulation for multi-speaker localization, it is instructive to provide a brief review on the case of single speaker and no reverberation. We define $\boldsymbol{\alpha}_k \in \mathbb{R}^2$ to be the speaker (x, y) position vector at the kth time frame. Then, define $\mathbf{x}_k = [\,\boldsymbol{\alpha}_k^T, \boldsymbol{\phi}_k^T\,]^T$ where $\boldsymbol{\phi}_k$ contain some kinematic variables for the speaker motion (e.g., velocity). A state space equation is used to model the time dependence of $\mathbf{x}_k$:

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

where $\mathbf{A}$ and $\mathbf{B}$ are some pre-specified matrices, and $\mathbf{w}_k$ is a time-uncorrelated random vector. For example, we can choose $\mathbf{x}_k = \boldsymbol{\alpha}_k$, $\mathbf{A} = \mathbf{B} = \mathbf{I}_2$ and $\mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \sigma_w^2\mathbf{I})$ for some variance σ_w^2, which will lead to the random walk model. Some more sophisticated motion models are also available; e.g., the Langevin model [2, 3]. Next, we consider the TDOA measurements. The TDOA obtained at the kth time frame from the qth microphone

pair is given by

$$z_k^{[q]} = \tau_q(\boldsymbol{\alpha}_k) + v_k^{[q]} \tag{2}$$

$$\tau_q(\boldsymbol{\alpha}_k) = \frac{1}{c}(\|\boldsymbol{\alpha}_k - \mathbf{u}_{2,q}\| - \|\boldsymbol{\alpha}_k - \mathbf{u}_{1,q}\|) \tag{3}$$

Here, $v_k^{[q]}$ is measurement error which we assume $v_k^{[q]} \sim \mathcal{N}(0, \sigma_v^2)$ for some error variance σ_v^2, $\{\mathbf{u}_{1,q}, \mathbf{u}_{2,q}\}$ are the position vectors of the qth microphone pair, and c is the speed of sound.

In the RFS case, the state vector $\mathbf{x}_k$ is extended to a finite set

$$\mathcal{X}_k = \{\mathbf{x}_{1,k}, \ldots, \mathbf{x}_{N_k,k}\} \tag{4}$$

which contains the state vectors of speakers active at time k. Here, $N_k = |\mathcal{X}_k|$ (where $|.|$ stands for the cardinality) is the number of active speakers at time k. We assume $N_k \leq N_{max}$ where N_{max} is the maximum allowable number of speakers. The RFS state space equation is given by

$$\mathcal{X}_k = \mathcal{B}_k(\mathbf{b}_k) \cup \left\{ \bigcup_{i=1,\ldots,|\mathcal{X}_{k-1}|} \mathcal{S}_k(\mathbf{x}_{i,k-1}, \mathbf{w}_{i,k}) \right\} \tag{5}$$

where $\mathcal{B}_k(\mathbf{b}_k)$ contains state vectors of speakers 'born' at time k, $\mathcal{S}_k(\mathbf{x}_{k-1}, \mathbf{w}_{i,k})$ is contributed by the speaker associated with $\mathbf{x}_{k-1}$, and the vectors $\mathbf{w}_{i,k}$ and $\mathbf{b}_k$ are random variables accountable for the random behaviors of $\mathcal{X}_k$. For $\mathcal{S}_k$, we have the following hypotheses:

$$\mathcal{S}_k(\mathbf{x}_{i,k-1}, \mathbf{w}_{i,k}) = \begin{cases} \emptyset, & H_{death} \\ \{\mathbf{A}\mathbf{x}_{i,k-1} + \mathbf{B}\mathbf{w}_{i,k}\}, & \bar{H}_{death} \end{cases} \tag{6}$$

where H_{death} and $\bar{H}_{death}$ are respectively the death and no-death hypotheses. The hypothesis H_{death} has a probability of 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:

$$\mathcal{B}_k(\mathbf{b}_k) = \begin{cases} \emptyset, & \bar{H}_{birth} \\ \{\mathbf{b}_k\}, & H_{birth} \end{cases} \tag{7}$$

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 (such as velocity) are zero.

Our TDOA measurement scheme follows that of [2, 3], in which multiple TDOAs are extracted from the GCC function by multiple peak picking. Thus, the measured TDOA in (2) is extended to a set-valued observation

$$\mathcal{Z}_k^{[q]} = \left\{ z_{1,k}^{[q]}, \ldots, z_{M_k^{[q]},k}^{[q]} \right\}, \tag{8}$$

where $M_k^{[q]} = |\mathcal{Z}_k^{[q]}|$ is the number of estimated TDOAs. It is important to note that only some of the elements in $\mathcal{Z}_k^{[q]}$ follow the true TDOA measurement model in (2). Moreover, we do not know which elements in $\mathcal{Z}_k^{[q]}$ are true TDOAs. Thus, the TDOA measurement model takes the form

$$\mathcal{Z}_k^{[q]} = \left\{ \bigcup_{i=1,\ldots,|\mathcal{X}_k|} \mathcal{T}_k^{[q]}(\mathbf{x}_{i,k}, v_{i,k}^{[q]}) \right\} \cup \mathcal{C}_k^{[q]} \tag{9}$$

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

$$\mathcal{T}_k^{[q]}(\mathbf{x}_{i,k}, v_{i,k}^{[q]}) = \begin{cases} \emptyset, & H_{miss} \\ \left\{\tau_q(\mathbf{C}\mathbf{x}_{i,k}) + v_{i,k}^{[q]}\right\}, & \bar{H}_{miss} \end{cases} \tag{10}$$

with $v_{i,k}^{[q]} \sim \mathcal{N}(0, \sigma_v^2)$. Here, $\tau_q(\mathbf{C}\mathbf{x}_k)$ is given in (3), $\mathbf{C}$ is such that $\mathbf{C}\mathbf{x}_k = \boldsymbol{\alpha}_k$, and $\bar{H}_{miss}$ and H_{miss} are respectively the detection and miss hypotheses. The hypothesis H_{miss} happens with a probability of P_{miss}. For the false TDOAs, we follow the standard assumption in [2, 3] 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 pairs 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.

The above RFS formulation is applicable to any N_{max} (i.e., the maximum number of active speakers), but in this TDOA application we usually fix $N_{max} = 2$. The reasons for this are as follows: i) The GCC method, which was designed for single source problems, only has medium time resolution to distinguish TDOAs of two speakers. 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. ii) In hands-free communication and teleconferencing applications, the mostly commonly encountered events are either no voice activity, one speaker, or one speaker interrupting another.

With the above RFS problem formulation, we can develop a Bayesian framework for estimating $\mathcal{X}_k$; i.e., estimating both the multi-speaker locations and the number of active speakers. This is considered in the next section.

3. RFS BAYESIAN FILTER

Using the RFS theory, we can construct probability density functions (p.d.f.s) for the RFS multi-speaker problem formulated above. This result is useful in developing a probabilistic framework for multi-speaker localization. More importantly, the RFS theory enables us to transfer many ideas in vector-valued Bayesian estimation directly to the RFS scenario. To illustrate this, we denote the p.d.f. of $\mathcal{X}_k$ conditioned on $\mathcal{X}_{k-1}$ by

$$f(\mathcal{X}_k|\mathcal{X}_{k-1}), \tag{11}$$

and the p.d.f. of $\mathcal{Z}_k^{[q]}$ given $\mathcal{X}_k$ by

$$g_q(\mathcal{Z}_k^{[q]}|\mathcal{X}_k). \tag{12}$$

The principles for deriving (11) and (12) can be found in the literature, such as [7]. Now, 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 $q = 1, \ldots, Q$ (where Q is the total of number of microphone pairs). The posterior p.d.f.s for $\mathcal{X}_k$ has a recursive relation reminiscent of the classic prediction and update formulae [12], given as follows:

$$p(\mathcal{X}_k|\mathcal{Z}_{1:k-1}^{[1:Q]}) = \int f(\mathcal{X}_k|\mathcal{X}_{k-1}) p(\mathcal{X}_{k-1}|\mathcal{Z}_{1:k-1}^{[1:Q]}) \mu(d\mathcal{X}_{k-1}) \tag{13}$$

$$p(\mathcal{X}_k|\mathcal{Z}_{1:k}^{[1:Q]}) = \frac{\prod_{q=1}^{Q} g_q(\mathcal{Z}_k^{[q]}|\mathcal{X}_k) p(\mathcal{X}_k|\mathcal{Z}_{1:k-1}^{[1:Q]})}{\int \prod_{q=1}^{Q} g_q(\mathcal{Z}_k^{[q]}|\mathcal{X}_k) p(\mathcal{X}_k|\mathcal{Z}_{1:k-1}^{[1:Q]}) \mu(d\mathcal{X}_k)} \tag{14}$$

$$g_q(\mathcal{Z}_k^{[q]}|\emptyset) = e^{-\lambda_c}\left(\frac{\lambda_c}{2\tau_{max}}\right)^{|\mathcal{Z}_k^{[q]}|} \tag{16}$$

$$g_q(\mathcal{Z}_k^{[q]}|\{\mathbf{x}_k\}) = g_q(\mathcal{Z}_k^{[q]}|\emptyset)\left(P_{miss} + (1-P_{miss})\sum_{z_k^{[q]}\in\mathcal{Z}_k^{[q]}}\left(\frac{2\tau_{max}}{\lambda_c}\right)\mathcal{N}(z_k^{[q]};\tau_q(\mathbf{C}\mathbf{x}_k),\sigma_v^2)\right) \tag{17}$$

$$\begin{aligned} g_q(\mathcal{Z}_k^{[q]}|\{\mathbf{x}_{1,k},\mathbf{x}_{2,k}\}) = g_q(\mathcal{Z}_k^{[q]}|\emptyset)\Bigg\{\prod_{i=1,2}\left(P_{miss} + (1-P_{miss})\sum_{z_k^{[q]}\in\mathcal{Z}_k^{[q]}}\left(\frac{2\tau_{max}}{\lambda_c}\right)\mathcal{N}(z_k^{[q]};\tau_q(\mathbf{C}\mathbf{x}_{i,k}),\sigma_v^2)\right) \\ -(1-P_{miss})^2\sum_{z_k^{[q]}\in\mathcal{Z}_k^{[q]}}\left(\frac{2\tau_{max}}{\lambda_c}\right)^2\mathcal{N}(z_k^{[q]};\tau_q(\mathbf{C}\mathbf{x}_{1,k}),\sigma_v^2)\mathcal{N}(z_k^{[q]};\tau_q(\mathbf{C}\mathbf{x}_{2,k}),\sigma_v^2)\Bigg\} \end{aligned} \tag{18}$$

where μ is the extended Lebesque measure on the space of finite subsets of the state space [9].

Given $p(\mathcal{X}_k|\mathcal{Z}_{1:k}^{[1:Q]})$, we can estimate $\mathcal{X}_k$ using a Bayes optimal criterion such as the expected *a posteriori* (EAP). This work employs the EAP approach, the details of which can be found in [9, 10]. To compute $p(\mathcal{X}_k|\mathcal{Z}_{1:k}^{[1:Q]})$, we adopt a particle filter implementation for (13) and (14). The benefits of this implementation are reminiscent of those in the single-speaker scenario; see [2, 3] for the details. The idea of RFS particle filtering was presented in [9], in which some theoretical aspects were also explored. Table 1 shows a bootstrap particle filter developed for our RFS multi-speaker problem. Essentially, the algorithm uses a set of set-valued particles $\{\mathcal{X}_k^{(i)}\}$ and weights $\{w_k^{(i)}\}$ to recursively approximate the posterior p.d.f.:

$$p(\mathcal{X}_k|\mathcal{Z}_{1:k}^{[1:Q]}) \approx \sum_{i=1}^{L} w_k^{(i)}\delta_S(\mathcal{X}_k - \mathcal{X}_k^{(i)}) \tag{15}$$

where $\delta_S(.)$ is a set-valued version of the standard Dirac delta function. An advantage of this bootstrap filter is that we do not need to evaluate the state transition density $f(\mathcal{X}_k|\mathcal{X}_{k-1})$. The particle generation at Step 1 of Table 1 can be easily done by following the state space process in (5) to (7). At the top of this page we show expressions for $g_q(\mathcal{Z}_k^{[q]}|\mathcal{X}_k)$ that are sufficient for the speaker tracking application. In (17) to (18), the notation $\mathcal{N}(z;\bar{z},\sigma_z^2)$ stands for a Gaussian density function with mean $\bar{z}$ and variance σ_z^2.

It is worthwhile mentioning that if we set $N_{max} = 1$, $P_{death} = 0$, and $P_{birth} = 1$, the resulting RFS particle filter reduces to a form very similar to the single-speaker particle filter in [2, 3].

4. SIMULATION RESULTS

We use a simulated reverberant room to test the tracking performance of the proposed RFS particle filter. The room is illustrated in Fig. 1. The dimensions of the enclosure are 3m × 3m × 2.5m. We employ four microphone pairs, each of which has an inter-sensor spacing of $\tau_{max} = 0.5$m. Fig. 1 also shows the trajectories and birth/death times of the speaker sources. The speaker sources are all female. The acoustic image method [13] 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 [3] 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. 2 plots the measured TDOAs against the time frame index. We can see that the condition of the measured data is not so good: 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.

Table 1. RFS bootstrap particle filter for multi-speaker tracking.

Given a particle size L.
for $k = 1, 2, \ldots$

Step 1. Sampling: Generate $\breve{\mathcal{X}}_k^{(i)} \sim f(.|\mathcal{X}_{k-1}^{(i)})$ independently for $i = 1, \ldots, L$. Compute

$$\breve{w}_k^{(i)} = \prod_{q=1}^{Q} g_q(\mathcal{Z}_k^{[q]}|\breve{\mathcal{X}}_k^{(i)})w_{k-1}^{(i)} \tag{16}$$

Normalization: $\breve{w}_k^{(i)} := \breve{w}_k^{(i)}/(\sum_{\ell=1}^{L}\breve{w}_k^{(\ell)})$ for all i.

Step 2. Resampling: Apply a resampling algorithm [12] on $\{\breve{w}_k^{(i)}, \breve{\mathcal{X}}_k^{(i)}\}_{i=1}^{L}$ to obtain a resampled set $\{w_k^{(i)}, \mathcal{X}_k^{(i)}\}_{i=1}^{L}$.

end

The parameter settings for the RFS particle filter are as follows. The state space model is the Langevin model used in [2, 3], with the same parameters. The standard deviation of the TDOA measurement error is $\sigma_v = 125\mu s$ (which is also the sampling period). The other parameters are $P_{birth} = 0.2$, $P_{death} = 0.01$, $P_{miss} = 0.25$, $\lambda_c = 3$, and $L = 500$. Fig. 3 illustrates the tracking performance of the multi-speaker RFS particle filter. For comparison, we also show the performance of the existing single-speaker particle filter [2, 3] in the same figures. Clearly, the RFS particle filter is capable of identifying the locations and activity intervals of the two speakers. The single-speaker particle filter gives a reasonable single-speaker tracking performance. From time step 30 to 43, the single-speaker particle filter exhibits a transient convergence where the location estimate moves from the

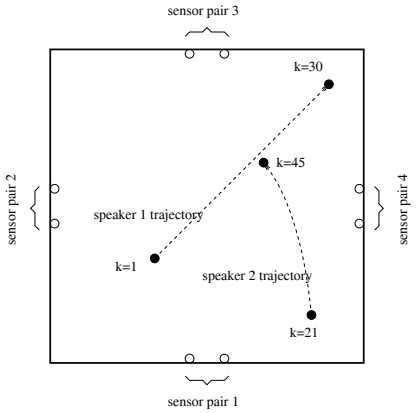

Fig. 1. Geometric settings for the room simulation.

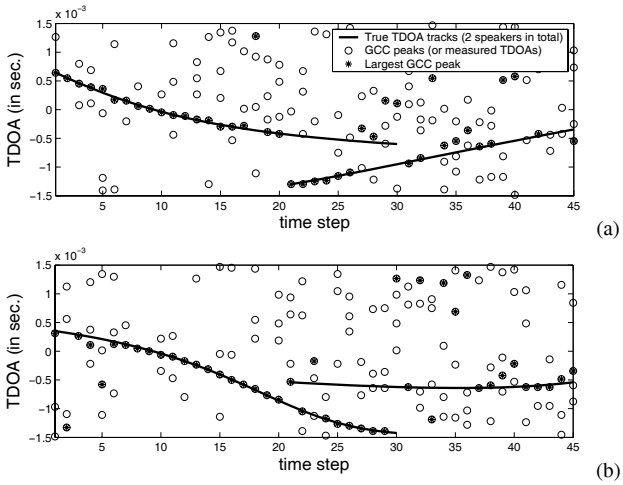

Fig. 2. Measured TDOAs at (a) sensor pair 1, and (b) sensor pair 3.

'dead' source to the new source. The RFS particle filter, however, does not have such a performance limitation.

5. CONCLUSION

Using the RFS theory and the particle filter implementation concept, 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.

6. REFERENCES

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

[2] 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.

[3] 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 Processing*, vol. 11, no. 6, pp. 826–836, 2003.

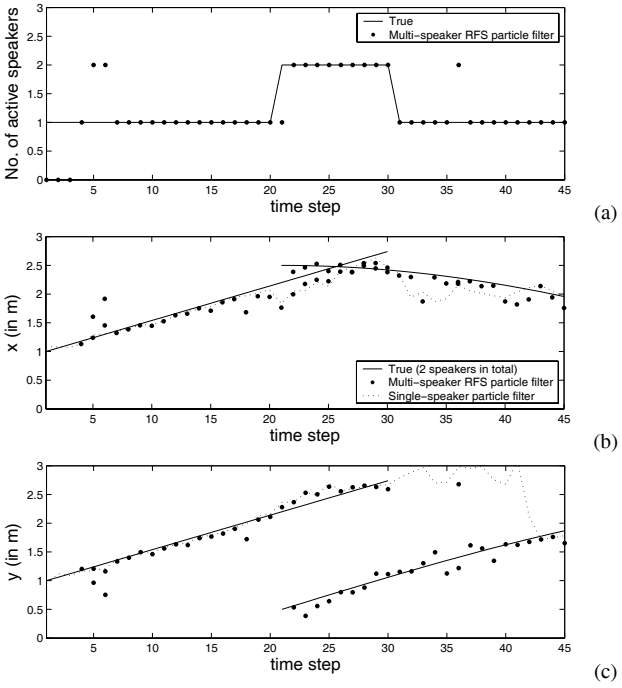

Fig. 3. (a) RFS particle filter estimates of the number of active speakers. (b)–(c) Position estimates of the RFS particle filter and the conventional single-speaker particle filter.

[4] C.H. Knapp and G.C. 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.

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

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

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

[8] R. Mahler, "Multi-target Bayes filtering via first-order multi-target moments," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1152–1178, 2003.

[9] B.-N. Vo, S. Singh, and A. Doucet, "Sequential Monte Carlo methods for Bayesian multi-target filtering with random finite sets," accepted for publication in *IEEE Trans. Aerosp. Electron. Syst.*, 2004.

[10] B.-N. Vo and W.-K. Ma, "Joint detection and tracking of multiple maneuvering targets in clutters using random finite sets," to appear in *Intl. Conf. Control, Automation, Robotics and Vision*, December 2004, also: `http://www.ee.mu.oz.au/staff/bv/publications.html`.

[11] B.-N. Vo, S. Singh, and W.-K. Ma, "Tracking multiple speakers with random sets," in *Proc. 2003 IEEE Intl. Conf. Acoust., Speech, Signal Processing*, May 2003.

[12] M.S. 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.

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