TRACKING MULTIPLE SPEAKERS USING RANDOM SETS

Ba-Ngu Vo, Sumeetpal Singh and Wing Kin Ma

Department of Electrical and Electronic Engineering
The University of Melbourne, VIC 3010, Australia
E-mail:{b.vo,ssss,w.ma}@ee.mu.oz.au

ABSTRACT

Tracking multiple speakers in an acoustic environment involves jointly estimating the number of speakers and their states. This important problem in signal processing is challenging in theory as well as implementation. This paper presents a novel and fundamentally well-grounded framework for tracking multiple speakers using random finite sets. Simulations are also presented to demonstrate the performance in tracking a randomly varying number of speakers in a reverberant room.

Keywords: Multi-target Tracking, Optimal Filtering, Particle Methods, Point Processes, Random Sets, Sequential Monte Carlo.

1. INTRODUCTION

Tracking the positions of multiple speakers (or sources) in an acoustic environment has several applications in multimedia such as automatic camera steering for video conferencing, discriminating between individual speakers in multi-speaker environments, and providing steering information for micro-phone arrays [2]. Several approaches for tracking a single speaker have been proposed [1], [6], [8], [3]. Traditional approaches [1], [6] transform the received frame of data into a *localisation function* that exhibits a peak in the location due to the speaker. However, reverberation causes spurious peaks in the localisation function that may have greater magnitudes than the peak associated with the speaker. Recent developments [8], [3] exploited the fact that the peak due to the true speaker follows a dynamical model from frame to frame, whereas the spurious peaks, also known as *clutter*, exhibit no temporal consistency. [8], [3] formulated the speaker tracking as nonlinear non-Gaussian filtering problem and solved it using a sequential Monte Carlo (SMC) algorithm (particle filter), yielding better performance than traditional approaches.

The emphasis of this paper is the generalisation to tracking multiple speakers, which is far from trivial. In a multi-speaker environment, speakers can appear or disappear in a random manner and tracking involves jointly estimating the randomly time-varying number of speakers as well as their *states*. The object of interest in this case is the set of states of all speakers, whose cardinality varies with time, whereas in single-speaker tracking, the object of interest (the speaker state) has a fixed dimension. As discussed in Section 2, multi-speaker tracking is a very difficult problem both in theory and implementation and single-speaker tracking algorithms cannot be easily extended for this purpose.

Using the theory of random finite sets (RFSs), or simple point processes, in this paper we propose a novel and mathematically well-founded framework to track multiple speakers in a noisy reverberative environment. The key is to treat the collection of speakers as a single *set-valued state*, and the collection of observations as a single *set-valued observation*. With appropriate notions of probability density for RFS [4], [9], the multi-speaker tracking problem can be rigorously cast as a Bayesian *set-valued estimation/filtering* problem. Simulations show the proposed RFS or set-valued estimation framework tracks multiple speakers in a reverberant room well.

2. MOTIVATION: MULTIPLES SPEAKERS AND RANDOM FINITE SETS

This section outlines the differences between the objectives of multiple hypothesis tracking (MHT) and set valued estimation for tracking multiple speakers. The discussion below assumes a noisy reverberant room with a single microphone pair. Extension to multiple microphone pairs is given in 3.2.

2.1. Time varying Speaker and Measurement Model

Let the state of a speaker at time (or frame) k be

$$x = [x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3]^T \in E_s \left(\subset R^6\right) \tag{1}$$

where $[x_1, x_2, x_3]^T$ is the speaker location, and $[\dot{x}_1, \dot{x}_2, \dot{x}_3]^T$ is the speaker velocity. The motion of each speaker is modeled by a discrete-time Markov process described by the transition density

$$f_{k|k-1}(x|x') \tag{2}$$

A Langevin process satisfies this assumption and is used for modeling human motion in [8]. In the absence of clutter, a speaker with state x at time k generates a *true* observation $z \in E_o$ that can be modeled by the likelihood function

$$g_k(z|x). \tag{3}$$

For example, suppose that at the k-th frame the speaker is located at x and S_1, S_2 are the signals received at the microphone pair whose coordinates are $w_1, w_2 \in R^3$. In the absence of other speakers, noise and reverberation, the signals S_1 and S_2 are delayed versions of each other whose cross-correlation peaks at the time difference of arrival (TDOA)

$$\Delta(x) = c^{-1}(\|x - w_1\| - \|x - w_2\|) \in E_o := [-\Delta_{\max}, \Delta_{\max}]$$

where c is the speed of sound and $\Delta_{\max}$ is the maximum possible delay defined by the dimensions of the room and microphones locations. Thus, the TDOA observation z can be modeled by a Gaussian with mean $\Delta(x)$ and variance σ^2 which is truncated to

$[-\Delta_{\max}, \Delta_{\max}]$ [8]. We assume that each clutter point is distributed according to the density on E_o

$$c_k(z), \quad z \in E_o. \tag{4}$$

A typical choice for the clutter density is the uniform density on E_o [8].

2.2. Multi-speaker Model

In a multi-speaker scenario, speakers appear and disappear randomly. For the duration the speaker is speaking, it moves according to the dynamic model (2) in a reverberant room with background noise. At time k, let $M(k)$ be the number of speakers present with states $x_{k,1}, \ldots, x_{k,M(k)}$, and let

$$X_k = \{x_{k,1}, \ldots, x_{k,M(k)}\} \subset E_s. \tag{5}$$

$$Z_k = \{z_{k,1}, \ldots, z_{k,N(k)}\} \subset E_o \tag{6}$$

denotes the set of measurements received at time k where some of the $N(k)$ observations may be due to clutter. If $z \in Z_k$ is due to a speaker with state x, then it is distributed according to $g_k(\cdot|x)$ (3), which we denote by $z \sim g_k(\cdot|x)$. If z is due to clutter, then $z \sim c_k(\cdot)$ (4). The number of clutter points are assumed to be Poisson distributed with rate λ_k. Let $Z_{1:k} = \{Z_1, \ldots, Z_k\}$ be the collection of all measurements sets received until time k. *The aim of multi-speaker tracking is to extract from $Z_{1:k}$ the information concerning when a speaker appeared and disappeared, as well as the trajectory it took, for all speakers that generated measurements between time 1 to k.* This is also the aim of classical multi-target tracking and achieved by the *multiple hypothesis tracking* (MHT) filter and its variants [5].

2.3. Set Valued Estimation vs MHT based filters

The MHT approach is one way to achieve the above stated aim of multi-speaker tracking. Given $Z_{1:k}$, we may hypothesize that a certain speaker, called speaker i, was present from time 1 to time k and generated observations $\{z_{1,i_1}, z_{1,i_2}, \ldots, z_{1,i_k}\}$. $z_{1,i_m} \in Z_m$ was the observation at time m due to speaker i, $1 \leq m \leq k$. Obviously $1 \leq i_m \leq N(m)$ and it is not necessary that $i_1 = i_2 = \ldots = i_k$. Assuming that new speakers have a state distributed according to the prior density γ, we may use γ, the observations $\{z_{1,i_1}, z_{1,i_2}, \ldots, z_{1,i_k}\}$, together with the transition (4) and likelihood (3) to compute the filtered density at time k for speaker i using the standard Bayes recursion. (It could also be hypothesised that speaker i appeared at time m, disappeared at time m' ($1 \leq m \leq m' \leq k$) and generated only one observation, or two, and so on.) An *association hypothesis* is defined as an assignment of all measurements in $Z_{1:k}$ to speakers and clutter subject to the constraint that a speaker generates none *or* one measurement at a time. The set $\mathcal{H}_k$ of all association hypotheses at time k is a large set and the problem of generating $\mathcal{H}_{k+1}$ from $\mathcal{H}_k$ is intractable. The MHT filter and its variants recursively propagate $\mathcal{H}_k$ and relies heavily on hypothesis pruning for a tractable implementation [5].

In target tracking applications, the MHT filter performs well when clutter is low, targets are well separated and follow predictable trajectories. For speaker tracking, the trajectories are complicated, the speakers can be close to each other and clutter density is high due to reverberation. Thus, MHT is not expected to work well here. Furthermore, the time between measurements are short and requires a fast filtering algorithm for real-time tracking. MHT is computationally intensive and is not suited for real-time implementation.

In Section 3 of this paper we propose a practical alternative to speaker tracking based on set valued estimation. Unlike the MHT approach that generates associations, the random set approach generates estimates of the speaker *states* but *without* any association.

Advantage: It is important to note that only the estimates of the speaker states are obtained at time 1 to k with set valued estimation. There is no trajectory information, or the path individual speakers took, as was available in $\mathcal{H}_k$. However, using post MHT like processing, it is possible to extract such information. The important practical advantage here is that recursively estimating the state of all speakers present at time k is tractable and can be done quickly using SMC [9]. This provides a quick online algorithm that may be use for beam steering, assuming we are only interested in locating and following acoustic sources without discriminating between them. The other advantage is that we have decoupled the state estimation problem from the problem of association and forming tracks. In the MHT, one did the latter and then the former. So, we are free to use more computationally intensive discrete optimisation based algorithms to solve the association problem off-line while still tracking states on-line. On a more technical note, we remark that the method proposed in [8] is not a rigorous Bayesian framework as one is using the *standard* Bayes recursion for filtering when the states and observations have fixed dimensions on a problem where the observation dimension is varying.

3. RANDOM SET FORMULATION AND ALGORITHMS

The multi-speaker state and the observation at time k are naturally represented as finite sets $X_k \subset E_s$ (5) and $Z_k \subset E_o$ (6). Uncertainty in a multi-speaker system is characterised by modeling multi-speaker state and observation as random finite sets (RFS) Ξ_k and Σ_k respectively. The formal definition of RFS and the notion of probability density for RFS can be found in [4], [9].

Given a realisation X_{k-1} (see (5)) of RFS Ξ_{k-1}, the multi-speaker state at time k can be modeled as the RFS

$$\Xi_k = S_k(X_{k-1}) \cup \Gamma_k \tag{7}$$

where $S_k(X_{k-1})$ denotes the RFS of speakers who continue to time k, and Γ_k is the RFS of new speakers at time k. The statistical behaviour of the RFS Ξ_k is characterised by the Markov transition density $f_{k|k-1}(X_k|X_{k-1})$. Note that the same notation $f_{k|k-1}$ is used for the individual speaker transition density and the multi-speaker transition density. However, the meaning is clear depending on wether the arguments of $f_{k|k-1}$ are sets or vectors. Let b_k denote the probability density of Γ_k, the RFS of new speakers, under suitable independence assumptions, it was shown in [4] that

$$f_{k|k-1}(Y|X) = \sum_{W \subseteq X} s_{k|k-1}(W|X) b_k(Y - W) \tag{8}$$

$$s_{k|k-1}(W|X) = p_S^{|W|}(1 - p_S)^{|X|-|W|} \times \sum_{1 \leq i_1 \neq \cdots \neq i_{|W|} \leq |X|} \prod_{j=1}^{|W|} f_{k|k-1}(y_j|x_{i_j}) \tag{9}$$

where p_S denotes the probability that the speaker continues to time k.

Similarly, given a realisation X_k of Ξ_k, the multi-speaker observation can be modeled by the RFS

$$\Sigma_k = \Theta_k(X_k) \cup C_k \tag{10}$$

where $\Theta_k(X_k)$ denotes the RFS of measurements generated by X_k, and C_k denotes the RFS of clutter. The statistical behaviour of the RFS Σ_k is described by the multi-speaker likelihood [4]

$$g_k(Z|X) = \sum_{W \subseteq X} h_k(W|X) l_k(Y - W) \tag{11}$$

$$h_k(W|X) = p_D^{|W|}(1-p_D)^{|X|-|W|} \times \sum_{1 \le i_1 \ne \cdots \ne i_{|W|} \le |X|} \prod_{j=1}^{|W|} g_k(z_j|x_{i_j}) \tag{12}$$

where p_D denotes the probability of detection and $l_k(\cdot)$ denote the probability density of the RFS C_k. In the case of state dependent clutter $l_k(\cdot|X)$ is used instead. Again, the same notation g_k is used for the individual speaker likelihood and the multi-speaker likelihood, but the meaning is clear depending on whether the arguments of g_k are sets or vectors.

Let $p_{k|k}(X_k|Z_{0:k})$ denote the multi-speaker posterior density. Then, the optimal multi-object Bayes filter is given by the recursion

$$p_{k|k-1}(X_k|Z_{0:k-1}) = \int f_{k|k-1}(X_k|X) p_{k-1|k-1}(X|Z_{0:k-1}) \mu_s(dX) \tag{13}$$

$$p_{k|k}(X_k|Z_{0:k}) = \frac{g_k(Z_k|X_k) p_{k|k-1}(X_k|Z_{0:k-1})}{\int g_k(Z_k|X) p_{k|k-1}(X|Z_{0:k-1}) \mu_s(dX)}. \tag{14}$$

where μ_s is a dominating measure [9]. Note (13), (14) has the same form as the standard fixed dimension Bayesian recursion except that arguments are now set-valued and the integrals are now *set integrals*.

3.1. The PHD Filter

The Bayesian propagation equations (13), (14) involves the evaluation of multiple set integrals. In [9] a generic particle filter has been proposed to implement it. A cheaper alternative to propagating $p_{k|k}(X_k|Z_{0:k})$ is the Probability Hypothesis Density (PHD) filter, which only propagates the first moment of $p_{k|k}(X_k|Z_{0:k})$ [4]. The PHD (also known as intensity or 1st moment) $D_\Xi : E_s \to R_+$ of a RFS Ξ is defined by

$$D_\Xi(x) \equiv \mathbf{E}\left[\sum_{w \in \Xi} \delta_w(x)\right] = \int \sum_{w \in X} \delta_w(x) P_\Xi(dX). \tag{15}$$

The PHD measure of a region $S \subseteq E$, i.e. $\int_S D_\Xi(x)dx$, gives the expected number of elements of Ξ that are in S. The peaks of the PHD provide good estimates for the elements of Ξ.

Let D_k, denote the PHD of the RFS $\Xi_k|Z_{0:k}$ (which is distributed according to the posterior $p_{k|k}$). D_k satisfies a recursion similar to the Bayes recursion, i.e., D_k is obtained from D_{k-1} and the new measurement set Z_k. Assuming that the predicted RFS $\Xi_k|Z_{0:k-1}$ is Poisson, it was shown in [4] that

$$D_k = \left(\Psi_k^{Z_k} \circ \Phi_{k|k-1}\right)(D_{k-1}). \tag{16}$$

The PHD prediction and update operators, denoted by $\Phi_{k|k-1}$ and $\Psi_k^{Z_k}$, are given respectively by

$$(\Phi_{k|k-1}\alpha)(x) = p_S \langle f_{k|k-1}(x|\cdot), \alpha \rangle + \gamma_k(x), \tag{17}$$

$$(\Psi_k^{Z_k}\alpha)(x) = \left[\upsilon + \sum_{z \in Z_k} \frac{p_D g_k(z|x)}{\lambda_k c_k(z) + p_D \langle g_k(z|\cdot), \alpha \rangle}\right] \alpha(x), \tag{18}$$

where α is a function on E_s, γ_k denotes the PHD of the RFS Γ_k of new speakers, $\upsilon = 1 - p_D$, c_k denotes the density of individual clutter point, λ_k denotes the average number of Poisson clutter points and $\langle f, g \rangle = \int_E f(\zeta) g(\zeta) d\zeta$.

Note that we are estimating the speaker states at time k with the PHD D_k, which is a function on E_s, and is cheaper to propagate than the multi-target posterior. A sequential Monte Carlo implementation of the PHD filter is given in [9].

3.2. Multiple microphone pairs

For L pairs of microphones, let $Z_k^{(i)}$ denote the observation set of the i-th microphone pair. The Bayes recursion (13), (14) still applies if we define $Z_k \equiv (Z_k^{(1)}, ..., Z_k^{(L)})$ and $g_k(Z_k|X)$ to be the joint likelihood $g_k(Z_k^{(1)}, ..., Z_k^{(L)}|X)$. Moreover, assuming conditional independence of the observations between the microphone pairs,

$$g_k(Z_k|X) = \prod_{i=1}^{L} g_k^{(i)}(Z_k^{(i)}|X). \tag{19}$$

where $g_k^{(i)}(Z_k^{(i)}|X)$ denotes the likelihood of the i-th pair. The PHD prediction equation (17) still applies, but unfortunately the PHD update (18) becomes very complex, even when $g_k(Z_k|X)$ decouples as in (19). However, the update can be approximated by

$$\Psi_k^{Z_k} \simeq \Psi_k^{Z_k^{(L)}} \circ \cdots \circ \Psi_k^{Z_k^{(1)}}. \tag{20}$$

Note that the operators $\Psi_k^{Z_k^{(i)}}$, $i = 1, ..., L$ do not commute and thus the filter output depends on the order of the updates. As a rule of thumb, update with the most 'reliable' pair of microphones should be done first.

4. SIMULATIONS

To demonstrate the applicability of the proposed framework, consider an unknown and time varying number of speakers in a $3m \times 3m \times 2.5m$ room with 2 microphone pairs placed at $([1, 0, 1]^T, [2, 0, 1]^T)$ and $([0, 1, 1]^T, [0, 2, 1]^T)$ respectively. Each speaker moves around in the room according to a Langevin model as in [8]. Speakers can appear or disappear in the scene at any time. In (9), $p_S = 0.95$. The appearance of new speakers at time k is modeled by a Poisson RFS with uniform intensity (over the room) and rate = 0.1. A probability of detection of 1 is used. The effect of reverberation is modeled by Poisson clutter with uniform intensity and rate = 5 at each microphone pair. The TDOA observations at each microphone pairs that are fed to the PHD filter. Figures 1, 2, 3 and 4 shows the PHD of positions at various times. From these Figures observe the close proximity between the peaks of the PHD and the true positions of the speakers.

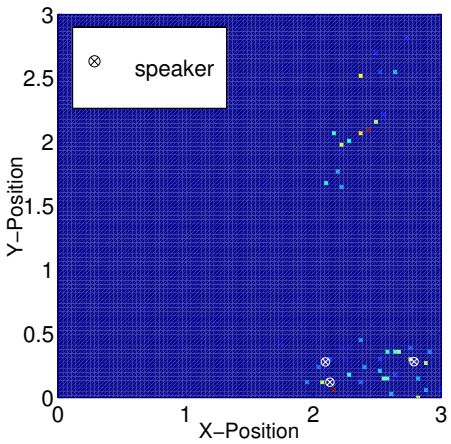

Fig. 1. PHD at k=1

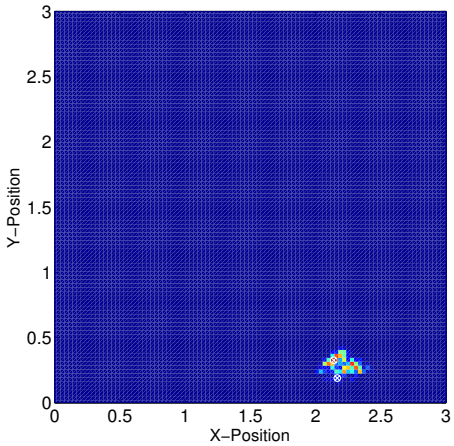

Fig. 2. PHD at k=6

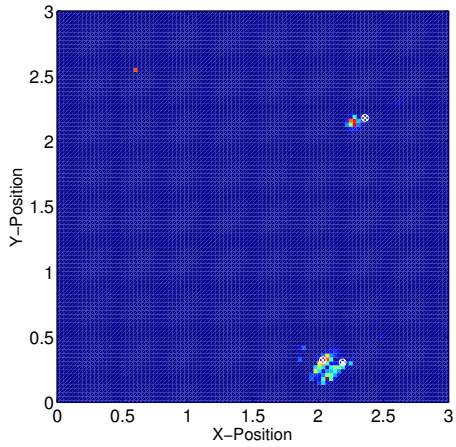

Fig. 3. PHD at k=11

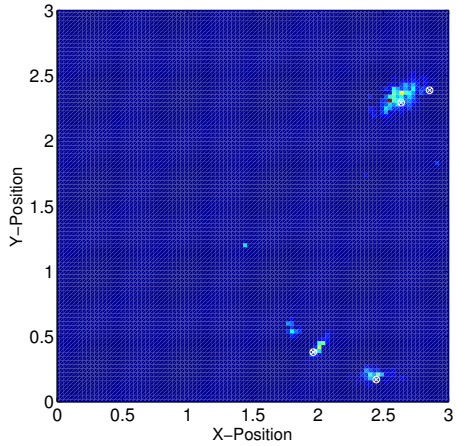

Fig. 4. PHD at k=16

5. CONCLUSION

Tracking multiple speakers is a challenging problem both in theory and implementation. While traditional tracking approaches are hypothesis based, this paper advocates a set-valued estimation approach. Modeling the multiple speakers as a dynamic random finite set enabled the tracking problem to be cast as a set-valued Bayesian estimation problem. Moreover, we have demonstrated that the set-valued estimation approach results in tracking algorithms that can be implemented on-line. Simulations show good tracking performance and motivates a real scenario study.

6. REFERENCES

[1] M. Brandstein and H. Silverman, "A practical methodology for speech source localisation with microphone arrays," *Comp., Speech & Language*, vol. 11, no. 2, pp. 91-126, 1997.

[2] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag Berlin, 2001.

[3] E. Lehmann, D. Ward and R. Williamson, "Experimental comparison of particle filtering algorithms for acoustic source localization in a reverberant room," *Proc. IEEE ICASSP*, vol. 5, pp. 177-180, Hong Kong, 2003.

[4] R. Mahler, "Approximate multisensor-multitarget joint detection, tracking and identification using a first order multitarget moment statistic," *IEEE Trans. AES*, to appear.

[5] D. Reid, "An algorithm for tracking multiple targets," *IEEE Trans. Automatic Control*, Vol AC-24, No. 6, 1979.

[6] H. Silverman and E. Kirtman, "A two-stage algorithm for determining talker location from linear microphone array data," *Comp., Speech & Language*, vol. 6, pp. 129-152, 1992.

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

[8] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," *Proc. IEEE ICASSP*, Salt Lake city, UT USA 2001.

[9] B. Vo, S. Singh and A. Doucet, "A sequential Monte Carlo Implementation of the PHD filter for Multi-target tracking," *Proc. ISF FUSION*, Cairns, Australia 2003.