

Tracking Multiple Speakers Using CPHD Filter

Nam Trung Pham, Weimin Huang
Institute for Infocomm Research, Singapore
{stuntp, wmhuang}@i2r.a-star.edu.sg

S. H. Ong
Department of ECE
National University of Singapore
eleongsh@nus.edu.sg

ABSTRACT

In this paper, we present an efficient method for tracking multiple speakers in a reverberant environment. The proposed method is based on the cardinalized probability hypothesis density (CPHD) filter. Because the CPHD filter can handle a large amount of clutter measurements, our method has a high reliability when tracking multiple speakers. Simulation experiments are presented to demonstrate the performance of the proposed method.

Categories and Subject Descriptors

G.3 [Probability and statistics]: Probabilistic algorithms;
J.7 [Computers in other systems]: Real time; I.2 [Artificial intelligence]: Miscellaneous

General Terms

Algorithms, Experimentation, Performance

Keywords

Random finite set, CPHD filter, TDOA, speaker tracking

1. INTRODUCTION

Speaker tracking is an important part of multimedia applications. It will allow us to determine the speaker trajectories and help us in analyzing the behavior of speakers.

There are many approaches for the single-speaker tracking problem based on search methods [2]. However, these methods may not be efficient in a reverberant acoustic environment because of multi-path effects. Recently, some approaches for speaker tracking based on the particle filter have been proposed to cope with the effects of reverberations [14]. Unfortunately, in some scenarios, many people speak simultaneously at a time. Tracking multiple speakers is challenging because of the varying number of speakers, multi-sensor data fusion, high clutter, and data association.

Methods for tracking multiple speakers should rely on a multiple-object tracking framework. There are some related

works on multiple-object tracking, such as the multiple hypothesis tracker (MHT) [8] and the joint probabilistic data association (JPDA) [1]. These methods estimate the states of multiple objects based on possible hypotheses for data association and are computationally expensive. Recently, there has been increasing research interests on using random set theory to solve multiple-object tracking. Mahler [6] proposed a probability hypothesis density (PHD) filter. This method operates on the single-object state space and avoids the combinatorial problem from data association. There are some implementations of the PHD filter, such as the particle PHD filter [11] and the Gaussian mixture PHD filter [10]. To improve the performance of the PHD filter, Mahler [5] also presented a cardinalized probability hypothesis density (CPHD) filter that is a generalization of the PHD recursion. The CPHD filter jointly propagates the posterior intensity and the posterior cardinality distribution at time steps. Vo [13] presented an implementation of the CPHD filter by using the Gaussian mixture model.

With the development of multiple-object tracking methods, several approaches to multiple-speaker tracking have been proposed [9], [7], [12], [4]. Among them, the particle PHD filter in [12] and the random finite set sequential Monte Carlo (RFS-SMC) Bayes filter [4] are based on the random finite set (RFS) approach, and have good performances.

In this paper, we propose an efficient method using the CPHD filter for tracking multiple speakers in a reverberant environment. The proposed method fuses the time delay of arrival (TDOA) measurements from microphone pairs by using asynchronous sensor fusion with the CPHD filter to obtain the positions of the speakers. Because the CPHD filter can handle a large amount of clutter measurements, our method is more reliable than methods in [9], [12] and [4].

2. CPHD FILTER APPROACH

The posterior density function in multiple-object tracking is difficult to obtain because of the large computation of data association. Fortunately, this posterior density can be approximately recovered from the first moment of this distribution, the probability hypothesis density (PHD) (or intensity function). The states of objects can be estimated by investigating peaks of PHD. To obtain the PHD at each time step, we can use the PHD filter [6] or the CPHD filter [5]. In this section, we review an implementation that is considered the closed-form of the CPHD filter [13]. First, there are some assumptions. Each object follows a linear

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.
Copyright 2007 ACM 978-1-59593-701-8/07/0009 ...\$5.00.

Gaussian model, i.e.,

$$f_{k|k-1}(x|\zeta) = N(x; F_{k-1}\zeta, Q_{k-1}), \quad (1)$$

$$g_k(z|x) = N(z; H_k x, R_k), \quad (2)$$

where F_{k-1} is a state transition matrix, Q_{k-1} is a process noise covariance, H_k is an observation matrix, and R_k is an observation noise covariance. The survival and detection probabilities are $p_{S,k}$ and $p_{D,k}$, respectively. The intensity of clutter measurements is $\kappa_k(\cdot)$ and the cardinality distribution of clutter is $p_{K,k}(\cdot)$. The intensity of the spontaneous birth RFS is

$$\gamma_k(x) = \sum_{i=1}^{J_{\gamma,k}} w_{\gamma,k}^{(i)} N(x; w_{\gamma,k}^{(i)}, P_{\gamma,k}^{(i)}) \quad (3)$$

The cardinality distribution of births is $p_{\Gamma,k}(\cdot)$. The posterior intensity at time $k-1$ is

$$v_{k-1}(x) = \sum_{i=1}^{J_{k-1}} w_{k-1}^{(i)} N(x; w_{k-1}^{(i)}, P_{k-1}^{(i)}) \quad (4)$$

The posterior cardinality distribution at time $k-1$ is $p_{k-1}(\cdot)$.

Under these assumptions, the predicted intensity and predicted cardinality distribution to time k is given by

$$p_{k|k-1}(n) = \sum_{j=0}^n p_{\Gamma,k}(n-j) \sum_{\ell=j}^{\infty} C_j^\ell p_{k-1}(\ell) p_{S,k}^j (1-p_{S,k})^{\ell-j} \quad (5)$$

$$v_{k|k-1}(x) = v_{S,k|k-1}(x) + \gamma_k(x), \quad (6)$$

where C_j^ℓ is the binomial coefficient and

$$v_{S,k|k-1}(x) = p_{S,k} \sum_{j=1}^{J_{k-1}} w_{k-1}^{(j)} N(x; m_{S,k|k-1}^{(j)}, P_{S,k|k-1}^{(j)}), \quad (7)$$

$$m_{S,k|k-1}^{(j)} = F_{k-1} m_{k-1}^{(j)}, \quad (8)$$

$$P_{S,k|k-1}^{(j)} = Q_{k-1} + F_{k-1} P_{k-1}^{(j)} F_{k-1}^T. \quad (9)$$

Because $v_{S,k|k-1}(x)$ and $\gamma_k(x)$ are Gaussian mixtures, $v_{k|k-1}(x)$ can be expressed as a Gaussian mixture of the form

$$v_{k|k-1}(x) = \sum_{i=1}^{J_{k|k-1}} w_{k|k-1}^{(i)} N(x; m_{k|k-1}^{(i)}, P_{k|k-1}^{(i)}) \quad (10)$$

Then, the posterior intensity and cardinality distribution at time k are given by

$$p_k(n) = \frac{\Psi_k^0[w_{k|k-1}, Z_k](n) p_{k|k-1}(n)}{\langle \Psi_k^0[w_{k|k-1}, Z_k], p_{k|k-1} \rangle}, \quad (11)$$

$$v_k(x) = \frac{\langle \Psi_k^1[w_{k|k-1}, Z_k], p_{k|k-1} \rangle}{\langle \Psi_k^0[w_{k|k-1}, Z_k], p_{k|k-1} \rangle} (1-p_{D,k}) v_{k|k-1}(x) + \sum_{z \in Z_k} \sum_{j=1}^{J_{k|k-1}} w_k^{(j)}(z) N(x; m_k^{(j)}(z), P_k^{(j)}) \quad (12)$$

where

$$\Psi_k^u[w, Z](n) = \sum_{j=0}^{\min(|Z|, n)} (|Z| - j)! p_{K,k}(|Z| - j) P_{j+u}^n \times$$

$$\frac{(1-p_{D,k})^{n-(j+u)}}{\langle 1, w \rangle^{j+u}} e_j(\Lambda_k(w, Z)), \quad (13)$$

$$\Lambda_k(w, Z) = \left\{ \frac{\langle 1, \kappa_k \rangle}{\kappa_k(z)} p_{D,k} w^T q_k(z) : z \in Z \right\}, \quad (14)$$

$$w_{k|k-1} = \left[w_{k|k-1}^{(1)}, \dots, w_{k|k-1}^{(J_{k|k-1})} \right]^T, \quad (15)$$

$$q_k(z) = \left[q_k^{(1)}(z), \dots, q_k^{(J_{k|k-1})}(z) \right]^T, \quad (16)$$

$$q_k^{(j)}(z) = N\left(z; \eta_{k|k-1}^{(j)}, S_{k|k-1}^{(j)}\right), \quad (17)$$

$$\eta_{k|k-1}^{(j)} = H_k m_{k|k-1}^{(j)}, \quad (18)$$

$$S_{k|k-1}^{(j)} = H_k P_{k|k-1}^{(j)} H_k^T + R_k, \quad (19)$$

$$w_k^{(j)}(z) = p_{D,k} w_{k|k-1}^{(j)} q_k^{(j)}(z) \times \frac{\langle \Psi_k^1[w_{k|k-1}, Z_k \setminus \{z\}], p_{k|k-1} \rangle \langle 1, \kappa_k \rangle}{\langle \Psi_k^0[w_{k|k-1}, Z_k], p_{k|k-1} \rangle \kappa_k(z)}, \quad (20)$$

$$m_k^{(j)}(z) = m_{k|k-1}^{(j)} + K_k^{(j)} \left(z - \eta_{k|k-1}^{(j)} \right), \quad (21)$$

$$P_k^{(j)} = \left[I - K_k^{(j)} H_k \right] P_{k|k-1}^{(j)}, \quad (22)$$

$$K_k^{(j)} = P_{k|k-1}^{(j)} H_k^T \left[S_{k|k-1}^{(j)} \right]^{-1} \quad (23)$$

P_{j+u}^n is the permutation coefficient and $e_j(\cdot)$ is elementary symmetric function [13].

3. TDOA MEASUREMENT FOR MULTIPLE SPEAKER TRACKING

The well-known method, generalized cross correlation function (GCC) [3], has been applied for estimating the TDOA measurements for a single speaker. Ma [4] extended the GCC method to collect measurements for multiple-speaker tracking. The technique is described briefly as follows. The GCC function is obtained as the following equation

$$\hat{R}(\tau) = \int_{-\infty}^{+\infty} \psi_{12}(w) Y_1(w) Y_2^*(w) e^{jw\tau} dw \quad (24)$$

where $Y_1(w), Y_2(w)$ are the Fourier transform of the signal $y_1(t), y_2(t)$ from microphones 1 and 2, respectively, τ is the time delay of arrival, and $\psi_{12}(w) = \frac{1}{|Y_1(w)Y_2^*(w)|}$. In the presence of multiple speakers, there are multi-path signal propagations and the GCC function in (24) 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 speaker sources. By collecting some local maximum peaks in GCC function, we have a set of measurements for multiple-speaker tracking.

4. CPHD FILTER FOR MULTIPLE SPEAKER TRACKING

The state space model for multiple-speaker tracking is defined as follows. Each moving speaker follows a dynamical model equation

$$x_k = Ax_{k-1} + w_{k-1} \quad (25)$$

where $A = [I]$ and w_{k-1} is an uncorrelated noise. We assume $w_{k-1} \sim N([0; 0], \text{diag}([0.01; 0.01]))$. This means the average distance from the previous time $k-1$ to k of a speaker is

about 10 cm. Given a speaker x_k , the TDOA measurement z_k^q is measured from the q -th microphone pair at time k . The measurement equation is

$$z_k^q = \frac{\|x_k - p_{2,q}\| - \|x_k - p_{1,q}\|}{c} + v_k^q, q = 1, \dots, Q \quad (26)$$

where $p_{i,q}$ is the position of microphone i of pair q , c is the speed of sound, and v_k^q is uncorrelated noise. In this context, we assume $v_k^q \sim N(0; 4 \times 10^{-9})$. This means the average time delay noise is for a delay 1 sample.

At each time k , let Z_k^i be the TDOA measurements that are collected at the microphone pairs i . The method to collect TDOA measurements is described in section 3. Assuming that we have $Q = 4$ microphone pairs, the RFS of measurements at time k is modelled by

$$Z_k = [Z_k^1; Z_k^2; \dots; Z_k^Q] \quad (27)$$

The idea of fusing TDOA measurements from Q microphone pairs cameras to obtain the speaker positions is to use the CPHD filter sequentially at each microphone pair. The algorithm is described as follows:

- Step 1: From $v_{k-1}(x)$ and $p_{k-1}(n)$, the prediction equations (5) and (6) are used to obtain the predicted PHD $v_{k|k-1}^1(x)$ and predicted cardinality distribution $p_{k|k-1}^1(n)$ at microphone pair 1. Because the measurement equation (26) is not linear, the unscented transform is employed in the prediction step (more details are given in [13]). Then, the TDOA measurements from the first microphone pair, Z_k^1 , is used to update the $v_{k|k-1}^1(x)$ and $p_{k|k-1}^1(n)$ to $v_k^1(x)$ and $p_k^1(n)$ by the updating step in the CPHD filter (equations (11), (12)).
- Step 2: Set $i = 2$
- Step 3: At the microphone pair i , set $v_{k|k-1}^i(x) = v_{k-1}^{i-1}(x)$ and $p_{k|k-1}^i(n) = p_{k-1}^{i-1}(n)$. Then, the updating step of the CPHD filter is used to update $v_{k|k-1}^i(x)$ and $p_{k|k-1}^i(n)$ with observations in Z_k^i . After that, $v_k^i(x)$ and $p_k^i(n)$ are obtained.
- Step 4: Set $i = i + 1$. If $i \leq Q$ then the step 3 is repeated. Otherwise, we have $v_k^Q(x)$ and $p_k^Q(n)$. The posterior intensity of the system is $v_k(x) = v_k^Q(x)$ and the posterior cardinality distribution is $p_k(n) = p_k^Q(n)$. Hence, the number of speakers is estimated by

$$\hat{N} = \arg \max_n p_k(n) \quad (28)$$

We choose means of \hat{N} Gaussian components that have the largest weights to represent the speaker position estimations.

5. EXPERIMENTAL RESULTS

We simulated an acoustic room to test the performance of the method. The dimensions of the room are $3\text{m} \times 3\text{m} \times 2.5\text{m}$. The reverberation time of the room impulse responses is about $T_{60} = 0.15\text{s}$. The speech signal to noise ratio is about 20dB. The time frame length for measuring TDOA is 256ms. The probability of survival and detection are $p_{S,k} = 0.95$ and $p_{D,k} = 0.7$. The maximum number of Gaussian

components $J_{max} = 30$. We assume the birth intensity of a speaker is

$$\gamma_k(x) = 0.01 \sum_{i=1}^9 N(x; m_\gamma^i, P_\gamma)$$

where $P_\gamma = \text{diag}([0.1; 0.1])$. We divide the room into 9 parts, and then consider the center points of these parts as the means of Gaussian components. With this birth intensity, the proposed method can detect new speakers at any place in the room.

Figure 1 shows the multiple-speaker tracking performance of the particle filter in [9]. Because the simulated acoustic room is a reverberant room, the steer-beamforming method to detect measurements in [9] is not efficient due to multipath effects. The tracking performance is not reliable when two persons speak simultaneously in this data set. Figure

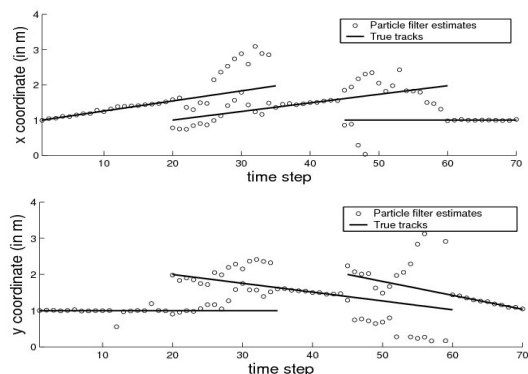


Figure 1: Position (x, y) of speakers by the method in [9]

2 shows the multiple-speaker tracking performance of our method. This performance is better than the method in [9]. Our method can give reliable estimations even when two people speak simultaneously. This is because the CPHD filter can handle a large amount of clutters in the TDOA measurement set.

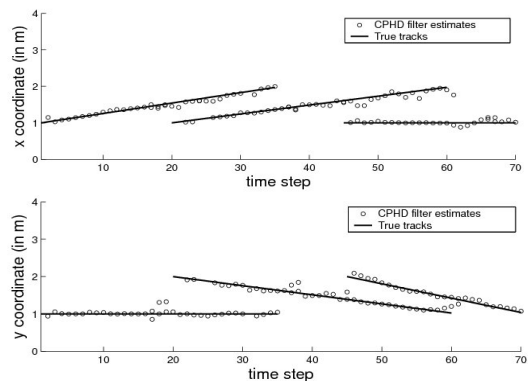


Figure 2: Position (x, y) of speakers by CPHD

To measure average performance, we used the performance measurement from [4]. It includes the probability of correct speaker number, expected absolute error on the number of

speaker and conditional mean distance error by the Wasserstein distance. We tested the performance with 500 trials. Figures 3, 4 and 5 show the performances of our method, the particle PHD filter [12] and the RFS-SMC Bayes filter [4]. Our method is more reliable than others. This is because state estimates by Gaussian mixtures are better than state estimates by clustering or extracting means of particles. Moreover the complexity computation of the CPHD filter is $O(|Z|^2 \log^2 |Z|)$ while the complexity computation of the RFS-SMC Bayes filter is exponentially growing when the number of speakers or measurements increases. The main error in our method occurs due to TDOA measurements that are not reliable in time steps when more than one person speak simultaneously.

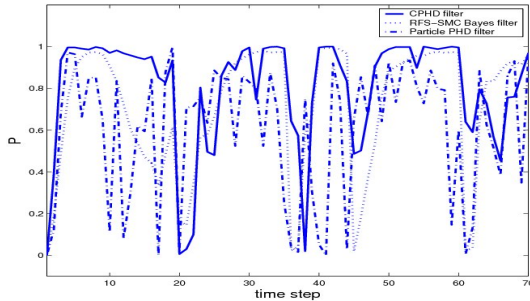


Figure 3: Probability of correct speaker number

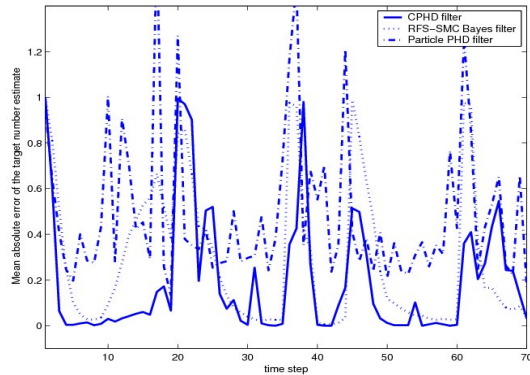


Figure 4: Absolute error on the number of speaker

6. CONCLUSIONS

In this paper, we developed an efficient method for tracking multiple speakers in a reverberant environment. The proposed method fused TDOA measurements from microphone pairs in the CPHD filter. Using synthetic audio data, we demonstrated that our method is more efficient than some methods in a reverberant acoustic environment. Moreover, this method can be applied in other multiple-sensor multiple-object tracking applications.

7. ACKNOWLEDGEMENT

The authors would like to thank Prof. Ba Ngu Vo at Melbourne University for his helps and fruitful discussions. This work is partially supported by EU project ASTRALS (FP6-IST-0028097).

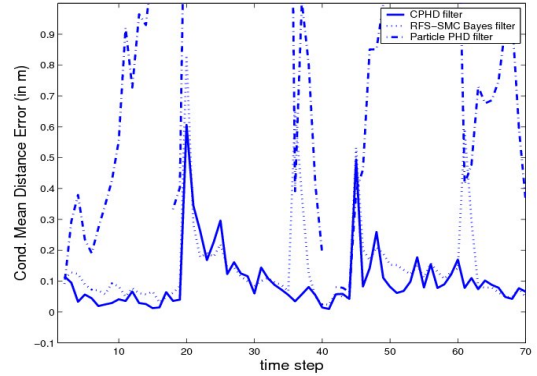


Figure 5: Conditional mean distance error of multi-speaker tracking

8. REFERENCES

- [1] Y. Bar-Shalom and T. E. Fortmann. *Tracking and data association*. Academic Press, San Diego, 1988.
- [2] J. C. Chen, K. Yao, and R. E. Hudson. Source localization and beamforming. *IEEE Signal Processing Mag.*, 19, 2002.
- [3] C. Knapp and G. Carter. The generalized correlation method of estimation of time delay. *IEEE Trans. Acoust., Speech, Signal Processing*, 24(4), 1976.
- [4] W. K. Ma, B. N. Vo, S. Singh, and A. Baddeley. Tracking an unknown time-varying number of speakers using TDOA measurements: a random finite set approach. *IEEE Trans. Signal Processing*, 54(9), 2006.
- [5] R. Mahler. PHD filters of higher order in target number. *IEEE Trans. AES (accepted)*.
- [6] R. Mahler. Multi-target Bayes filtering via first-order multi-target moments. *IEEE Trans. on Aerospace and Electronic Systems*, 39(4), 2003.
- [7] I. Potamitis, H. Chen, and G. Tremoulis. Tracking of multiple moving speakers with multiple microphone arrays. *IEEE Trans. Speech Audio Processing*, 12(5), 2002.
- [8] D. Reid. An algorithm for tracking multiple targets. *IEEE Trans. Automatic Control*, 24(6), 1979.
- [9] J. M. Valin, F. Michaud, and J. Rouat. Robust 3D localization and tracking of sound sources using beamforming and particle filtering. In *ICASSP*, 2006.
- [10] B. N. Vo and W. K. Ma. The Gaussian mixture probability hypothesis density filter. *IEEE Trans. Signal Processing*, 54(11), 2006.
- [11] B. N. Vo, S. Singh, and A. Doucet. Sequential Monte Carlo methods for Bayesian multi-target filtering with random finite sets. *IEEE Trans. Aerospace and Electronic Systems*, 41(4), 2005.
- [12] B. N. Vo, S. Singh, and W. K. Ma. Tracking multiple speakers with random sets. In *ICASSP*, 2004.
- [13] B. T. Vo, B. N. Vo, and A. Cantoni. Analytic implementations of the cardinalized probability hypothesis density filter. *IEEE Trans. Signal Processing*, 55(7), 2007.
- [14] D. B. Ward, E. A. Lehmann, and R. C. Williamson. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Trans. on Speech and Audio Processing*, 11(6), 2003.