16th International Conference on Information Fusion
Istanbul, Turkey, July 9-12, 2013

# Acoustic Source Tracking in a Reverberant Environment Using a Pairwise Synchronous Microphone Network

Xionghu Zhong†, Arash Mohammadi⋆, Wenwu Wang°, A. B. Premkumar† and Amir Asif⋆
†CEMNET, School of Computer Engineering, Nanyang Technological University, Singapore, 639798.
⋆Department of Electrical Engineering and Computer Science, York University, Toronto, Canada, M3J 1P3.
°CVSSP, Department of Electronic Engineering, University of Surrey, UK, GU2 7XH.
†{xhzhong and asannamalai}@ntu.edu.sg, ⋆{marash and asif}@cse.yorku.ca, °w.wang@surrey.ac.uk

***Abstract*—This paper considers acoustic source tracking in a room environment using a distributed microphone pair network. Existing time-delay of arrival (TDOA) based approaches usually require all received signals to be transmitted to central processor and synchronized to extract the TDOA measurements. The source positions are then obtained by using a subsequent localization or tracking approach. In this paper, we propose a distributed particle filtering (PF) approach to track the source using a microphone pair network. Each node is constructed by a microphone pair and TDOA measurements are extracted at local nodes. An extended Kalman filter based PF is developed to estimate the first order and the second order statistics of the source state. A consensus filter is then applied to fuse these local statistics between neighboring nodes to achieve a global estimation. Under such an approach, only the state statistics need to be transmitted and the received signals need only to be pairwise synchronized. Consequently, both communication and computational cost can be significantly reduced. Simulations under different reverberant environments demonstrate that the proposed approach outperforms the centralized sequential importance sampling based PF approach in single source tracking as well as in nonconcurrent multiple source tracking.**

***Index Terms*—Acoustic source tracking, reverberant environment, time-delay of arrival, extended Kalman particle filtering, consensus filter.**

## I. Introduction

Tracking an acoustic source (speaker) in a room environment plays an important role in many speech and audio applications such as diarisation, hearing aids, hands-free distant speech recognition and communication, and teleconferencing systems [1]–[4]. Once the speaker is localized and tracked, the position information can be fed into a higher processing stage for: high-quality speech acquisition; enhancement of a specific speech signal in the presence of other competing talkers; or keeping a camera focused on the talker in a video-conferencing scenario. Usually, a distributed system equipped with a number of microphone pairs/arrays is employed to localize or track the source [5]–[8]. However, it is a challenge to provide an accurate position estimation since the received audio signal can be significantly distorted and its statistical properties drastically changed due to room reverberations. The difficulties also arise from the uncertainty in the source motion and the non-stationary characteristics of the speech signal.

Time-delay of arrivals (TDOAs) have been shown to provide robust measurements in the noisy and reverberant environment for room acoustic localization and tracking [3], [5]–[11]. Usually, all received signals are transmitted to the central processor (CP) and TDOA estimates are obtained by, for example, using generalized cross-correlation (GCC) method [12]. Since each TDOA yields a half hyperboloid of two sheets, multiple TDOA measurements from distributed microphone pairs/arrays are employed to triangulate a target position. Such a triangulation is traditionally approximated either by using a linear intersection (LI) algorithm [5] or by using an extended Kalamn filter (EKF) algorithm [7], [9]. However, the performance of these algorithms can be seriously degraded due to incorrect TDOA measurements caused by reverberation and different kinds of noise. Recently, sequential importance resampling based particle filtering (SIR-PF) [8], [13] was introduced into the room acoustic source localization and tracking problem. It is able to reduce the TDOA errors and is more robust than LI and EKF approaches in adverse environments. More advanced PF algorithms have been developed for room acoustic source localization and tracking. In [14], a voice activity detection (VAD) is employed to reduce the effect of heavy false alarms due to the weak source signals in silence gaps. In [10], an EKF is incorporated into a PF to enhance sampling efficiency. In [11], a multiple-hypothesis based PF for acoustic localization is proposed.

However, transmitting all received signals to the CP and synchronizing them can be cumbersome and expensive in practice, particularly when a large number of microphones are deployed. Recently, advances in distributed wireless sensor networks in providing unprecedented capabilities for target detection and localization have motivated the deployment of distributed wireless microphone network for acoustic source localization and tracking [15]. In this paper, we propose to track an acoustic source using networked microphone pairs. Each node in the network is constructed from a microphone pair so that the local TDOA measurement is available. Based on the TDOA measurements extracted at each node, a distributed particle filtering approach is developed to track the acoustic source. Firstly, an extended Kalman filter based PF

978-605-86311-1-3 ©2013 ISIF

(EKPF) is implemented at each local node to coarsely estimate the source position. A consensus filter is then applied to fuse these local estimates between neighboring nodes to achieve a global estimation. Under such an architecture, only the state statistics need to be transmitted between neighboring nodes. Consequently, the communication cost can be significantly reduced and the received signals need only to be pairwise synchronized. It is worth mentioning that the deployment of microphone pairs here is similar to that in [8]–[10], the difference is that communications between neighboring nodes are allowed and the source position is estimated in a distributed manner in our work. Received acoustic energy measurement based microphone network is formulated in [15]. However, the acoustic energy measurement is not appropriate for room acoustic source tracking since it can be seriously violated by room reverberation.

The core contributions of this work is that a distributed PF approach has been derived for room acoustic source localization and tracking problem. The rest of this paper is organized as follows. In Section II, the network signal model and GCC method based TDOA esmation are introduced. Section III presents the distributed EKPF algorithm for room acoustic source tracking. Simulations are organized in Section IV and conclusions are drawn in Section V.

## II. Signal model and TDOA measurements

This section provides the received signal model of a microphone network in a room environment. The GCC method based TDOA measurement extraction at each local node is also addressed. It is worth mentioning that the received signal model employed here is the same as that described in [8].

### A. Network Signal Model

Assume that a microphone network which consists of $L$ microphone pairs is deployed. Let $\boldsymbol{p}_{\ell,i} \in \mathbb{R}^3$, $i \in \{1, 2\}$ denote the position of the $i$th microphone of the $\ell$th, for $\ell = 1, \ldots, L$ microphone pair, and let $\mathbf{x}_t \in \mathbb{R}^3$ denote the position of the source signal at time $t$. The discrete time signal received from a single source can be modeled as

$$z_{\ell,i}(t) = s(t) \star h(\boldsymbol{p}_{\ell,i}, \mathbf{x}_t) + n_{\ell,i}(t), \tag{1}$$

where $s(t)$ is the source signal, $h(\boldsymbol{p}_{\ell,i}, \mathbf{x}_t)$ is the overall impulse response cascading the room and the microphone channel response, $n_{\ell,i}(t)$ is an additive noise process assumed to be uncorrelated with the source, and $\star$ denotes convolution. To formulate TDOA estimates, the impulse response can be rewritten in terms of direct path and multipath components as

$$\begin{aligned} z_{\ell,i}(t) &= \frac{1}{r_{\ell,i}(t)} s(t - \tau_{\ell,i}(t)) + \underbrace{s(t) \star g(\boldsymbol{p}_{\ell,i}, \mathbf{x}_t) + n_{\ell,i}(t)} \\ &= \frac{1}{r_{\ell,i}(t)} s(t - \tau_{\ell,i}(t)) \qquad + \qquad v_{\ell,i}(t), \end{aligned} \tag{2}$$

where $r_{\ell,i}(t) = \|\mathbf{x}_t - \boldsymbol{p}_{\ell,i}\|$ is the Euclidean distance between source and microphone, $\tau_{\ell,i}(t) = r_{\ell,i}(t)/c$ is the direct path time delay, $c$ is the speed of sound, and $g(\boldsymbol{p}_{\ell,i}, \mathbf{x}_t)$ is a modified impulse response which is defined as the original response minus the direct path component. The new noise term $v_{\ell,i}(t)$ contains the additive noise $n_{\ell,i}(t)$ and the reverberant signal $s(t) \star g(\boldsymbol{p}_{\ell,i}, \mathbf{x}_t)$. This model is the free-field model in that it regards reverberation as part of the noise term.

As usual, although speech is non-stationary and its statistics change over time, it is assumed that speech is quasi-stationary. Hence, the signal received at each microphone is processed in frames. Let $T_0$ and $k$ denote the length and the time index of the frame, respectively. The source signal and the signal collected at the $i$th microphone of the $\ell$th pair can then be written as $\mathbf{s}(k)$ and $\mathbf{z}_{\ell,i}(k)$. Further, it is assumed that in each frame the position of the source is spatially stationary. The parameters characterizing the source are fixed in the $k$th frame, e.g., the source position, $\mathbf{x}_k$, and the corresponding room impulse response (RIR), $h(\boldsymbol{p}_{\ell,i}, \mathbf{x}_k)$. Without loss of generality, we assume that each node of the microphone network consists of two microphones (i.e., a microphone pair). All nodes of the network are modeled as vertices of the graph $\boldsymbol{\mathcal{G}} = (\boldsymbol{\nu}, \boldsymbol{\mathcal{E}})$, namely as elements of the node set $\boldsymbol{\nu} = \{1, \ldots, L\}$. The edge set $\boldsymbol{\mathcal{E}} \subseteq \boldsymbol{\nu} \times \boldsymbol{\nu}$ represents the network's communication constraints, i.e., if node $\ell$ can send information to node $m$ then $(\ell, m) \in \boldsymbol{\mathcal{E}}$. For graph $\boldsymbol{\mathcal{G}}$, the maximum degree $\Delta_{\mathcal{G}} = \max_\ell D^{(\ell)}$, where $D^{(\ell)}$ is the number of neighboring nodes for node $\ell$. Also relevant is the Laplacian matrix $\boldsymbol{L}$ for graph $\boldsymbol{\mathcal{G}}$, defined as $L_{\ell\ell} = D^{(\ell)}$ and $L_{\ell m} = -1$ if $(\ell, m) \in \boldsymbol{\mathcal{E}}$, else $L_{\ell m} = 0$.

### B. TDOA Measurements at Local Nodes

Due to its popularity and simplicity in time-delay estimation, the phase transform based generalized cross-correlation (PHAT-GCC) method is used in this paper to extract the TDOA measurements. Given the speech frames $\mathbf{z}_{\ell,1}(k)$ and $\mathbf{z}_{\ell,2}(k)$ collected at the $\ell$th microphone pair at time step $k$, the GCC function can be approximated as [12]:

$$R_\ell(k, \tau) = \int_\Omega \Phi_\ell(k, \omega) Z_{\ell,1}(k, \omega) Z^*_{\ell,2}(k, \omega) e^{j\omega\tau} d\omega, \tag{3}$$

where $\mathbf{z}_{\ell,i}(k) \rightleftharpoons Z_{\ell,i}(k, \omega)$ are DFT pairs, $\Phi_\ell(k, \omega) = |Z_{\ell,1}(k, \omega) Z^*_{\ell,2}(k, \omega)|^{-1}$ is the PHAT weighting term, and $\Omega$ is the frequency range over which the integration is carried out. The TDOA measurement at the $\ell$th microphone pair can thus be estimated by exploring the potential TDOA $\tau$ that maximizes the GCC function

$$\hat{\tau}^\ell_k = \arg \max_{\tau \in [-\tau_{\max}, \tau_{\max}]} R_\ell(k, \tau), \tag{4}$$

where $\tau_{\max} = \|\boldsymbol{p}_{\ell,1} - \boldsymbol{p}_{\ell,2}\|/c$ is the maximum delay possible. This TDOA measurement could, for example, be directly used in (6).

The signal model contains the parameters of interest, namely the TDOA $\tau_{\ell,i}(k)$. The actual TDOA of a microphone pair is expressed in terms of the source and sensor geometry by

$$\tau^\ell_k(\mathbf{x}_k) = \tau_{\ell,1}(k) - \tau_{\ell,2}(k) = \frac{\|\mathbf{x}_k - \boldsymbol{p}_{\ell,1}\| - \|\mathbf{x}_k - \boldsymbol{p}_{\ell,2}\|}{c}. \tag{5}$$

Given a sequence of TDOA estimates, $\hat{\tau}_k^\ell$, the maximum likelihood (ML) criterion [3] for the location estimate is given by:

$$\hat{\mathbf{x}}_k = \arg\min_{\mathbf{x}_k} \sum_{\ell=1}^{L} \left(\hat{\tau}_k^\ell - \tau_k^\ell(\mathbf{x}_k)\right)^2 . \tag{6}$$

The evaluation of $\hat{\mathbf{x}}_k$ at each time step involves the optimization of a non-linear function and necessitates the use of numerical search methods as no closed-form solution exists.

## III. Distributed EKPF for Source Localization and tracking

This section presents our solution to the problem of tracking an acoustic source based on the local TDOA measurements. Essentially, an EKPF is employed at each local node to estimate the first and the second order statistics of the source state. These statistics are then fused by communicating between the neighboring nodes to obtain a global estimation. The PF framework at a local node is introduced first.

### A. *Particle Filtering for Local Statistics Estimation*

Since the height of a talker is often fixed during a conversation for a certain length of time, it is reasonable to consider that the source dynamics follow the Langevin model [8], [13] in 2-dimensional ($x-y$) plan, and in $z-$direction, we use a random walk model to describe the uncertainties. The complete motion in 3-D space can be modelled as

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

where $\mathbf{v}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_k)$ is a zero-mean real Gaussian process with variance $\mathbf{\Sigma}_k$, and the state vector $\mathbf{x}_k$ is extended by appending a velocity component, i.e., $\mathbf{x}_k = [x_k\, y_k\, \dot{x}_k\, \dot{y}_k\, z_k]^T$. The coefficient matrices $\mathbf{A}$ and $\mathbf{Q}$ are given by

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}_2 & a\Delta T\mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & a\mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 \end{bmatrix}; \quad \mathbf{Q} = \begin{bmatrix} b\Delta T\mathbf{I}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & b\mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 \end{bmatrix}, \tag{8}$$

where $\Delta T = T_0/f_s$ is the time interval (in seconds) between time step $k$ and $k-1$, $f_s$ denoting the sampling frequency, and $\mathbf{I}_M$ is an $M$-order identity matrix. The parameters $a$ and $b$ are the position and velocity variance constants calculated according to $a = \exp(-\beta\Delta T)$ and $b = v\sqrt{1-a^2}$, in which $v$ and $\beta$ are the velocity parameter and the rate constant respectively. Equation (7) is used to model the source dynamics in this paper. The model parameters $v = 1\text{ms}^{-1}$ and $\beta = 10\text{s}^{-1}$ used in [8], [13], [16] are found to be adequate for room acoustic source tracking and are employed here.

Assume at each local node, $n_k^\ell$ TDOAs are collected. The measurement set for the $\ell$th node can thus be written as $\mathcal{Z}_k^\ell = \{\tau_{1,k}^\ell, \ldots, \tau_{n_k^\ell,k}^\ell\}$. At each local node, the solution based on Bayesian recursive estimation can be given as

- Predict:

$$p(\mathbf{x}_k|\mathcal{Z}_{1:k-1}^\ell) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathcal{Z}_{1:k-1}^\ell)d\mathbf{x}_{k-1}; \tag{9}$$

- Update:

$$p(\mathbf{x}_k|\mathcal{Z}_{1:k}^\ell) \propto p(\mathcal{Z}_k^\ell|\mathbf{x}_k)p(\mathbf{x}_k|\mathcal{Z}_{1:k-1}^\ell). \tag{10}$$

Obtaining the closed form solution to the recursion (9) and (10) is not easy since the TDOA measurement function is nonlinear. A promising approach to approximate this recursion is using PF approach [8], [13], [17].

Assume that a set of particles $\{\mathbf{x}_{k-1}^{(i)}\}_{i=1}^N$, with corresponding importance weight $\{w_{k-1}^{(i)}\}$ are available to approximate the posterior distribution of $p(\mathbf{x}_{k-1}|\mathcal{Z}_{1:k-1}^\ell)$ at time step $k-1$. A simple way to draw the particles at the current time step, $k$ is to follow the source dynamic model (7), stated as

$$\mathbf{x}_k^{(i)} \sim p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}). \tag{11}$$

The importance weights of the particles at the current time step are then given by

$$w_k^{(i)} = w_{k-1}^{(i)} \frac{p(\mathcal{Z}_k^\ell|\mathbf{x}_k^{(i)})p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})}{q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathcal{Z}_{1:k}^\ell)}, \tag{12}$$

where $q(\cdot)$ stands for the importance function. If the source dynamic model is employed as importance function, i.e., $q(\cdot) = p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})$, the weight becomes $w_k^{(i)} = w_{k-1}^{(i)}p(\mathcal{Z}_k^\ell|\mathbf{x}_k^{(i)})$. After performing the resampling scheme, the posterior distribution of the state is thus approximated by

$$p(\mathbf{x}_k|\mathcal{Z}_k^\ell) \approx \sum_{i=1}^{N} \tilde{w}_k^{(i)}\delta_{\mathbf{x}_k^{(i)}}(\mathbf{x}_k), \tag{13}$$

where $\delta(\cdot)$ is a Dirac-delta function.

It is worth mentioning that the above described PF is implemented at each local node and is based only on the local TDOA measurements. In [8], centralized SIR-PF approach has been developed, in which the likelihood is obtained by taking the TDOAs from all microphone pairs into account, i.e., $p(\mathcal{Z}_k|\mathbf{x}_k^{(i)}) = \prod_{\ell=1}^{L} p(\mathcal{Z}_k^\ell|\mathbf{x}_k^{(i)})$, and a prior density based importance function is employed, i.e., $q(\cdot) = p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})$. It has been shown that PF tracking approach is able to provide better accuracy of source position estimation than LI based localization approach [8].

### B. *Optimal Sampling Based on EKF*

It is desired that the particles are drawn according to the posterior distribution, i.e., $\mathbf{x}_k^{(i)} \sim p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathcal{Z}_{1:k}^\ell)$. As such both current measurements and previous state estimation are taken into account and the particles can be drawn at a more relevant area. Following [10], the EKF is employed in this paper to approximate the posterior distribution. The first-order Taylor expansion on $\tau_k^\ell(\mathbf{x}_k)$ from (5) is [9]

$$\tau_k^\ell(\mathbf{x}_k) = \tau_k^\ell(\mathbf{x}_{k-1}) + \mathbf{C}_k^\ell\left[\mathbf{x}_k - \mathbf{x}_{k-1}\right]^T + \bar{n}_k, \tag{14}$$

where $\bar{n}_k = O_{\mathbf{x}}(\mathbf{x}_k)$ is the higher order error of the time delay expansion, and $\mathbf{C}_k^\ell$ is the coefficient vector of Taylor expansion

$$\mathbf{C}_k^\ell = \frac{1}{c}\left[\frac{\mathbf{x}_k - \boldsymbol{p}_{\ell,1}}{\|\mathbf{x}_k - \boldsymbol{p}_{\ell,1}\|} - \frac{\mathbf{x}_k - \boldsymbol{p}_{\ell,2}}{\|\mathbf{x}_k - \boldsymbol{p}_{\ell,2}\|}\right]\Bigg|_{\mathbf{x}_k = \mathbf{x}_{k-1}}. \tag{15}$$

Define

$$\bar{\tau}_k^\ell = \hat{\tau}_k^\ell - \tau_k^\ell(\hat{\mathbf{x}}_{k-1}) + \mathbf{C}_k^\ell \hat{\mathbf{x}}_{k-1}, \quad (16)$$

where $\hat{\tau}_k^\ell$ is the TDOA measurement extracted from the largest peak of the GCC function, $\tau_k^\ell(\hat{\mathbf{x}}_{k-1})$ is calculated from (5). The nonlinear measurement is thus approximated by

$$\bar{\tau}_k^\ell \approx \mathbf{C}_k^\ell \mathbf{x}_k + \bar{n}_k. \quad (17)$$

Hence, the modified measurement $\bar{\tau}_k^\ell$ is a linear function of the state $\mathbf{x}_k$ and a standard KF can be applied. Note that the measurement here only contains the TDOA extracted from the largest peaks in the GCC function.

Assume that at the previous time step, the first order and the second order statistics of a global estimation $\widehat{\mathbf{x}}_{k-1}^\ell$ and $\widehat{\mathbf{P}}_{k-1}^\ell$ are achieved at the $\ell$th local node. Regarding (7) as the state process, the implementation of an EKF can be written as [18]

$$\overline{\mathbf{x}}_{k|k-1}^\ell = \mathbf{A}\widehat{\mathbf{x}}_{k-1}^\ell; \quad (18a)$$

$$\overline{\mathbf{P}}_{k|k-1}^\ell = \mathbf{A}\widehat{\mathbf{P}}_{k-1}^\ell \mathbf{A}^T + \mathbf{Q}\mathbf{\Sigma}_k \mathbf{Q}^T; \quad (18b)$$

$$\mathbf{S}_k^\ell = \mathbf{R}_k + \mathbf{C}_k^\ell \overline{\mathbf{P}}_{k|k-1}^\ell \left(\mathbf{C}_k^\ell\right)^T; \quad (18c)$$

$$\mathbf{K}_k^\ell = \overline{\mathbf{P}}_{k|k-1}^\ell (\mathbf{C}_k^\ell)^T (\mathbf{S}_k^\ell)^{-1}; \quad (18d)$$

$$\widetilde{\mathbf{x}}_k^\ell = \overline{\mathbf{x}}_{k|k-1}^\ell + \mathbf{K}_k^\ell(\hat{\tau}_k^\ell - \tau_k^\ell(\overline{\mathbf{x}}_{k|k-1}^\ell)); \quad (18e)$$

$$\widetilde{\mathbf{P}}_k^\ell = \overline{\mathbf{P}}_{k|k-1}^\ell - \mathbf{K}_k^\ell \mathbf{C}_k^\ell \overline{\mathbf{P}}_{k|k-1}^\ell. \quad (18f)$$

After the EKF steps, the posterior distribution is given by $p(\mathbf{x}_k^\ell | \mathbf{x}_{k-1}^\ell, \hat{\tau}_k^\ell) = \mathcal{N}(\mathbf{x}_k; \widetilde{\mathbf{x}}_k^\ell, \widetilde{\mathbf{P}}_k^\ell)$. This posterior distribution is used as an importance function. The particles are then drawn according to

$$\mathbf{x}_k^{\ell,(i)} \sim \mathcal{N}(\mathbf{x}_k^{\ell,(i)}; \widetilde{\mathbf{x}}_k^\ell, \widetilde{\mathbf{P}}_k^\ell). \quad (19)$$

The detailed expression of the transition density $p(\mathbf{x}_k^{(i)} | \mathbf{x}_{k-1}^{(i)})$ and likelihood $p(\mathcal{Z}_k^\ell | \mathbf{x}_k^{(i)})$ in (12) will be given in Section III-D. After resampling, the local statistics are estimated by

$$\overline{\mathbf{x}}_k^\ell = \sum_{i=1}^{N} w_k^{\ell,(i)} \mathbf{x}_k^{\ell,(i)}; \quad (20)$$

$$\overline{\mathbf{P}}_k^\ell = \sum_{i=1}^{L} w_k^{\ell,(i)} (\mathbf{x}_k^{\ell,(i)} - \overline{\mathbf{x}}_k^\ell)(\mathbf{x}_k^{\ell,(i)} - \overline{\mathbf{x}}_k^\ell)^T. \quad (21)$$

In next section, a consensus filter will be introduced to fuse these local statistics to obtain a global estimation of the source state and covariance matrix.

### *C. Global Estimation Based on Consensus Filter*

Based on Equations (20) and (21), node $\ell$ computes the MMSE estimate $\overline{\mathbf{x}}_k^\ell$ and its corresponding error covariance $\overline{\mathbf{P}}_k^\ell$ (local statistics) of the state variables. However, the local particles and their associated weights are based only on the local measurements $\mathcal{Z}_k^\ell$. Hence, such an estimation can result in inconsistent state estimates $\mathbb{E}(\mathbf{x}_k | \mathcal{Z}_k^\ell)$ across the network.

After obtaining the local statistics, a fusion step is used to compute a consistent set of values for the global statistics $\widehat{\mathbf{x}}_k$ and $\widehat{\mathbf{P}}_k$ at time $k$. To combine local statistics $\{\overline{\mathbf{x}}_k^\ell, \overline{\mathbf{P}}_k^\ell\}_{\ell=1}^L$ into a common set of global statistics $\widehat{\mathbf{x}}_k$ and $\widehat{\mathbf{P}}_k$ across the network, based on the Chong-Mori-Chang track-fusion theorem [19], the following rules are derived

$$\left[\widehat{\mathbf{P}}_k\right]^{-1} = \left[\overline{\mathbf{P}}_{k|k-1}^\ell\right]^{-1} + \underbrace{\sum_{\ell=1}^{L} \left\{ \left[\overline{\mathbf{P}}_k^\ell\right]^{-1} - \left[\overline{\mathbf{P}}_{k|k-1}^\ell\right]^{-1} \right\}}_{\boldsymbol{P}_c(\infty)}; \quad (22)$$

$$\widehat{\mathbf{x}}_k = \left[\widehat{\mathbf{P}}_k\right]^{-1} \left[ \left[\overline{\mathbf{P}}_{k|k-1}^\ell\right]^{-1} \overline{\mathbf{x}}_{k|k-1}^\ell + \underbrace{\sum_{j=1}^{L} \left\{ \left[\overline{\mathbf{P}}_k^\ell\right]^{-1} \overline{\mathbf{x}}_k^\ell - \left[\overline{\mathbf{P}}_{k|k-1}^\ell\right]^{-1} \overline{\mathbf{x}}_{k|k-1}^\ell \right\} \right]}_{\mathbf{x}_c(\infty)}. \quad (23)$$

In Eqs. (22) and (23), $\{\mathbf{x}_c(\infty), \boldsymbol{P}_c(\infty)\}$ are obtained by iterating the following average consensus equations where $\epsilon \in (0, 1/\Delta_{\mathcal{G}})$

$$\mathbf{x}_c^\ell(t+1) = \mathbf{x}_c^\ell(t) + \epsilon \sum_{j \in \aleph^{(l)}} (\mathbf{x}_c^j(t) - \mathbf{x}_c^\ell(t)); \quad (24)$$

$$\mathbf{P}_c^\ell(t+1) = \mathbf{P}_c^\ell(t) + \epsilon \sum_{j \in \aleph^{(l)}} (\mathbf{P}_c^j(t) - \mathbf{P}_c^\ell(t)) \quad (25)$$

till they converge to $\{\mathbf{x}_c(\infty), \mathbf{P}_c(\infty)\}$. The initial conditions are

$$\mathbf{P}_c^\ell(t=0) = \left[\overline{\mathbf{P}}_k^\ell\right]^{-1} - \left[\overline{\mathbf{P}}_{k|k-1}^\ell\right]^{-1}; \quad (26)$$

$$\mathbf{x}_c^\ell(t=0) = \left[\overline{\mathbf{P}}_k^\ell\right]^{-1} \overline{\mathbf{x}}_k^\ell - \left[\overline{\mathbf{P}}_{k|k-1}^\ell\right]^{-1} \overline{\mathbf{x}}_{k|k-1}^\ell. \quad (27)$$

Note that the consensus approach in (24)-(25) is a distributed algorithm where each node communicates only with its neighboring nodes. Convergence of the consensus algorithms is extensively studied. For example, in [20], it has shown that achieving consensus in a finite number of iterations is possible even for time-invariant topologies. In this paper, we consider the case where it is possible to communicate sufficiently fast so that consensus is reached between two successive observations which is a common practice in consensus-based distributed implementation of the particle filter. See [21] for extension to the scenario where the consensus is not reached within two consecutive time iterations. It is worth mentioning that the states are unobservable based on the TDOA measurement at each local node. However, the local nodes are able to communicate and fuse the estimates/tracks from other nodes to obtain the global estimates at the information fusion stage.

### *D. Tracking Algorithm*

The remaining issue to complete the tracking algorithm is formulating the transition density and the likelihood. After the EKF, the particles are drawn according to (19). The state transition density is given as

$$p(\mathbf{x}_k^{\ell,(i)} | \mathbf{x}_{k-1}^{\ell,(i)}) = \mathcal{N}(\mathbf{x}_k^{\ell,(i)} | \mathbf{A}\mathbf{x}_{k-1}^{\ell,(i)}, \mathbf{Q}\mathbf{\Sigma}_k \mathbf{Q}^T + \mathbf{A}\widetilde{\mathbf{P}}_k^\ell \mathbf{A}^T), \quad (28)$$

where $\mathbf{A}$ and $\mathbf{Q}$ are defined in (8). For each TDOA measurement set $\mathcal{Z}_k^\ell$ collected from a distributed microphone pair, at most one TDOA is directly generated by the source, and the other peaks are generated by clutters. Following [13], a variable $\{\lambda_{p,k}\}_{p=1}^{n_k^\ell}$ is defined to indicate the association between each TDOA measurement and its source. Two categories of hypotheses can thus be summarized for all the measurements obtained from a microphone pair

$$\begin{aligned}\mathcal{H}_{0,k}^{\ell} &\triangleq \{\lambda_{q,k}=0; q=1,\ldots,n_k^{\ell}\};\\ \mathcal{H}_{q,k}^{\ell} &\triangleq \{\lambda_{q,k}=1,\ \lambda_{p,k}=0; q\neq p=1,\ldots,n_k^{\ell}\},\end{aligned} \tag{29}$$

where $\mathcal{H}_{0,k}^{\ell}$ denotes that none of the measurements are generated by the source, and $\mathcal{H}_{q,k}^{\ell}$ represents that the $q$th TDOA measurement $\hat{\tau}_{q,k}^{\ell}$ is generated by the source, and all other TDOAs are generated by clutters.

If the measurement is generated by a clutter, such that $\lambda_{q,k}=0$, the likelihood is assumed to be uniform within the admissible TDOA range, given as

$$p(\hat{\tau}_{q,k}^{\ell}|\mathbf{x}_k^{\ell,(i)},\lambda_{q,k}=0)=\mathcal{U}_{\tau}(\hat{\tau}_{q,k}^{\ell})=\frac{1}{2\tau_{\max}}, \tag{30}$$

where $\tau=[-\tau_{\max},\,\tau_{\max}]$ denotes the possible TDOA range. If the measurement is generated by a real source, the likelihood is modelled as the true TDOA corrupted by white Gaussian noise with variance $\sigma_{\tau}^2$ [13], given by

$$p(\hat{\tau}_{q,k}^{\ell}|\mathbf{x}_k^{\ell,(i)},\lambda_{q,k}=1)=\mathcal{N}(\hat{\tau}_{q,k}^{\ell}\,|\,\tau_k^{\ell}(\mathbf{x}_k^{\ell,(i)}),\sigma_{\tau}^2). \tag{31}$$

Of course, it is unknown whether each TDOA estimate is generated by the target or clutter. The correct hypothesis $\mathcal{H}_{q,k}^{\ell}$ is thus unknown *a priori*. In [8], [13], all the collected TDOA estimates are deemed with equal importance. Assume that the prior probability for $\mathcal{H}_{0,k}^{\ell}$ is $q_0$, so $p(\mathcal{H}_{0,k}^{\ell}|\mathbf{x}_k^{\ell,(i)})=q_0$. The prior probability $p(\mathcal{H}_{q,k}^{\ell}|\mathbf{x}_k^{\ell,(i)})$ is equally weighted, so

$$p(\mathcal{H}_{q,k}^{\ell}|\mathbf{x}_k^{\ell,(i)})=\frac{1-q_0}{n_k^{\ell}}; \qquad \text{for } q\in\{1,\ldots,n_k^{\ell}\}. \tag{32}$$

The complete likelihood over all hypotheses from the $\ell$th microphone pair is obtained by summing all hypotheses:

$$\begin{aligned}p(\mathcal{Z}_k^{\ell}|\mathbf{x}_k^{\ell,(i)}) &= \sum_{q=0}^{n_k^{\ell}} p(\mathcal{H}_{q,k}^{\ell}|\mathbf{x}_k^{\ell,(i)})p(\mathcal{Z}_k^{\ell}|\mathbf{x}_k^{\ell,(i)},\mathcal{H}_{q,k})\\ &= \frac{\frac{q_0}{2\tau_{\max}}+\frac{1-q_0}{n_k^{\ell}}\sum_{q=1}^{n_k^{\ell}}\mathcal{N}(\hat{\tau}_{q,k}^{\ell}\,|\,\tau_k^{\ell}(\mathbf{x}_k^{\ell,(i)}),\sigma_{\tau}^2)}{(2\tau_{\max})^{n_k^{\ell}-1}}.\end{aligned} \tag{33}$$

This likelihood depicts all possible hypotheses generated by the measurement set, and performs well in room acoustic source localization and tracking scenarios [8], [13].

Substituting the likelihood (33), the transition density (28) and the importance function (19) into (12), the weight of particles can be calculated in a straightforward manner. After resampling, the first and the second order statistics are estimated according to (20) and (21) respectively. These local statistics are then fed into the consensus filter to obtain the global estimation.

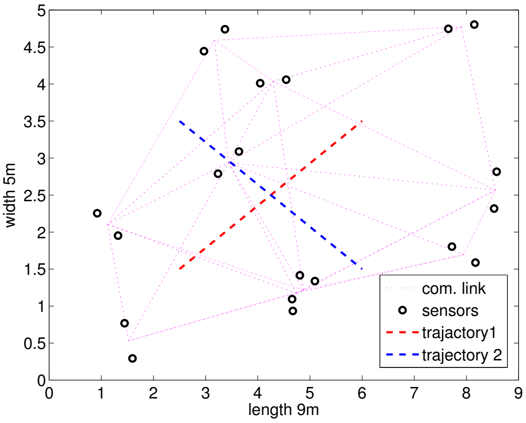

Fig. 1. Plot of source trajectories and an example of sensor network configuration. Two motion cases are considered: 1) single source indicated as source trajectory 1; and 2) one source motion as source trajectory 1 and then followed by the other indicated as source trajectory 2.

## IV. Simulations

In this section, the performance of the proposed distributed tracking algorithm is demonstrated. A number of microphone pairs are deployed to formulate a distributed microphone network. The room dimension is $9\times5\times3\text{m}^3$ with background noise yielding a SNR level of 30dB. Each microphone pair are randomly located in the $x-y$ plane with a separation of 50cm. The height of the microphones is set to 2.5m, which is 0.5m below the ceiling. Two typical motion trajectories are considered: motion as a line or "switch-source". For the line trajectory case, there is only one source moving in the room; the motion trajectory is marked as trajectory 1 in Fig. 1. The switch-speaker case involves a source change at the time center of the whole voice period; source 1 is active first and then source 2 follows. Such a case often happens in a conversation where nonconcurrent multiple speakers exist. The trajectories of two sources are also indicated as trajectory 1 and trajectory 2 respectively in Fig. 1. For the first motion case, one speaker is moving from $(2.5, 1.5, 1.5)$m to $(6.0, 3.5, 1.5)$m. For the "switch-speaker" case, the other speaker is active from $(2.5, 3.5, 1.5)$m to $(6.0, 1.5, 1.5)$m after the first one. Such motions result in a velocity of $\pm0.5$m/s roughly. The source signal is taken from the TIMIT database [22] with a sampling frequency of 1.6kHz. The background noise is simulated by adding white Gaussian noise into the received signal. Various reverberation time $T_{60}$s are employed to describe different reverberant environments. The RIR at each microphone is generated using the image method [23]. The whole experimental setup is depicted in Fig. 1.

To compare the position estimation performance, two centralized tracking approach, centralized SIR-PF (CPF) algorithm in [8] and centralized EKPF (CEKPF) in [10] are also implemented. The parameters for proposed DEKPF are set as: $L=500$, $\mathbf{\Sigma}_k=\text{diag}(\mathbf{I}_4, 0.1)$, $\sigma_{\tau}=1.25\times10^{-4}$ and $q_0=0.2$. This parameter setup is found empirically to be adequate for

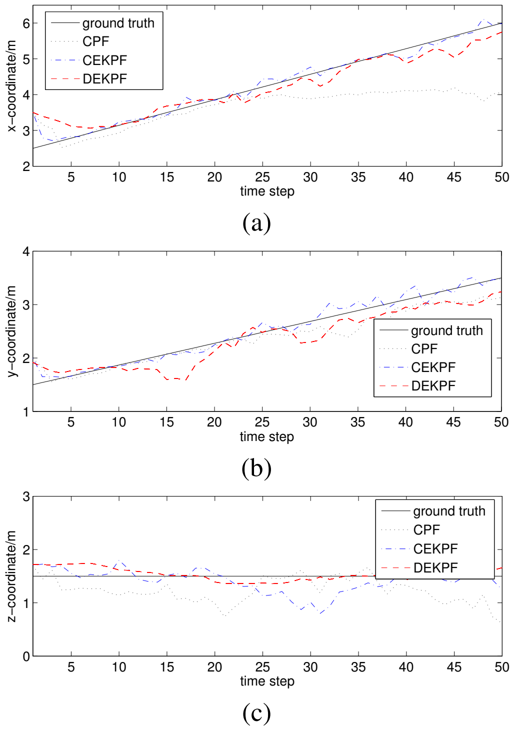

Fig. 2. Single source estimation results under $T_{60} = 250$ms for (a) $x-$; (b) $y-$; and (c) $z-$ coordinate.

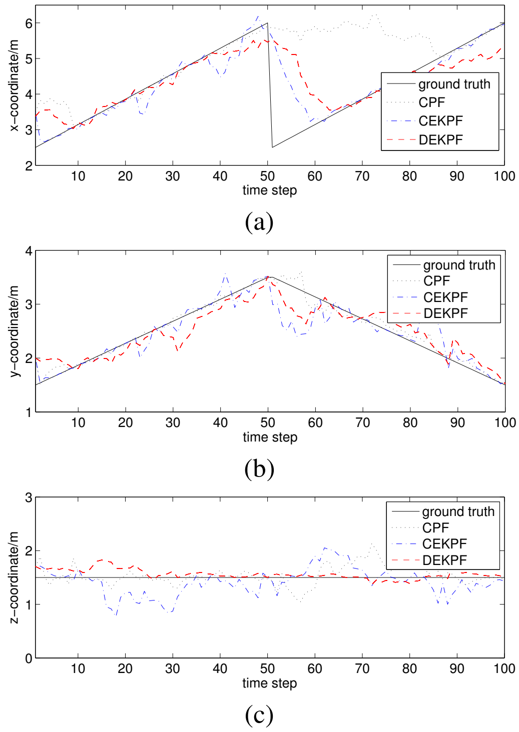

Fig. 4. Nonconcurrent multiple source estimation results under $T_{60} = 250$ms for (a) $x-$; (b) $y-$; and (c) $z-$ coordinate.

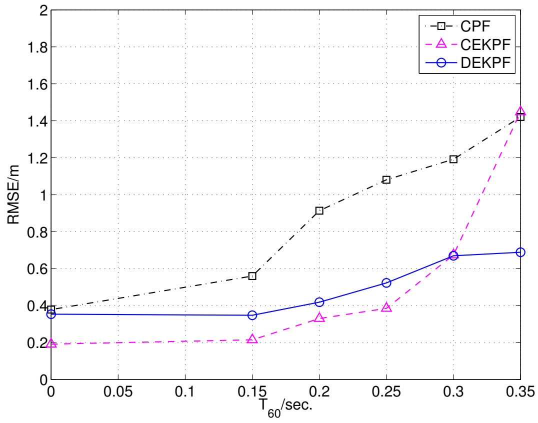

Fig. 3. RMSE over 100 MC runs for single acoustic source tracking.

all following simulations. The source velocities are initialized around the ground truth. The positions are initialized around the center of the room. The root mean square error (RMSE) is employed to evaluate the performance of position estimation. It is worth pointing out that for the proposed DEKPF approach, RMSE is calculated by averaging all RMSE at local nodes.

Figure 2 presents the tracking result of a single implementation under a single source scenario. The source motion follows source trajectory 1, as shown in Fig. 1. The reverberation time $T_{60}$ is 0.25s. It shows that the proposed DEKPF tracking approach is able to track the source trajectory accurately. All three components ($x-$, $y-$ and $z-$ coordinates) are estimated accurately. It performs better than CPF approach. Although CEKPF presents the best tracking result in $x-$ and $y-$ coordinates, its estimation in $z-$ coordinate is degraded. To fully study the performance, the proposed approach is implemented under different simulated reverberant environments. Various reverberation time $T_{60} = [0, 0.15, 0.2, 0.25, 0.3, 0.35]$s are employed to generate different reverberant environments. Fig. 3 shows the RMSE over 100 MC runs of each algorithm. Both the centralized and distributed EKPF approaches perform better than the centralized SIR-PF since an optimal importance function is approximated. The particles are sampled according to the current TDOA measurements as well as the previous position estimates. Hence, such algorithms are capable of keeping their lock on the source trajectory. Also, the performance of the centralized EKPF is better than that of the distributed one when the reverberation time is relatively small. However, when reverberation becomes stronger, the centralized EKPF degrades very rapidly; see $T_{60} = 0.35$s for example. This is because the TDOA measurements under such a reverberant environment become unreliable and particles cannot be drawn accurately by using EKF. The proposed distributed EKPF performs more consistently under different reverberant environments than the CEKPF does since it is less affected by inaccurate TDOA measurements due to the reverberations.

Figure 4 presents the tracking result of a single implementation under nonconcurrent multiple source scenario. The source motion follows trajectory 1 and then the other follows

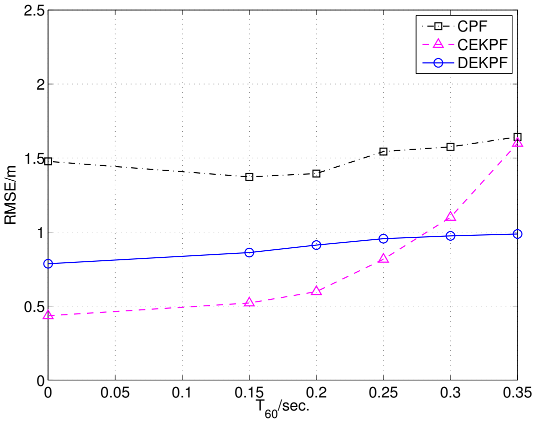

Fig. 5. RMSE over 100 MC runs for nonconcurrent multiple acoustic source tracking.

trajectory 2, as shown in Fig. 1. The reverberation time $T_{60}$ is set to 0.25s. Since the source switches, there is a sharp change on the source position. It is desired that the tracking algorithm lock on to the new source as fast as possible. The results show that the CEKPF performs the best in following up with the new source. The results of the proposed DEKPF is slightly worse than that of CEKPF but significantly better than that of CPF. The main reason is that both CEKPF and DEKPF introduced an EKF step to draw the particles. Hence, the optimal importance function is approximated and the particles are drawn at a more relevant area. The RMSE over 100 MC runs is shown in Fig. 5. The advantage of the proposed DEKPF over CPF is even more obvious. Under all different reverberant environments, the average tracking accuracy is at least 0.5m better than that of CPF approach. This is because the DEKPF approach is able to lock to the sharp position change between different speakers quickly. Although CEKPF performs better than DEKPF under low reverberant environment, the proposed DEKPF is more consistent when $T_{60}$ becomes larger.

Under both tracking scenarios, the tracking accuracy of the proposed DEKPF is favorably comparable with that of CEKPF. However, it is worth pointing out that the DEKPF needs only to pairwise synchronize the received signals and to exchange the first order and the second order statistics between the neighboring nodes. Hence, it is more efficient than CEKPF in terms of both computation complexity and communication cost. The detailed analysis of the computational complexity and communication cost will be conducted in the journal version of this paper.

## V. Conclusions

A distributed PF tracking approach for room acoustic source tracking using a microphone pair network is studied in this paper. At each local node, TDOA measurements are extracted and an EKPF is introduced to estimate the local statistics. A consensus filter is then applied to achieve a global estimation. The proposed DEKPF is able to estimate the source position accurately in the 3-D space. It performs significantly better than centralized PF approach in [8] due to the incorporation of an optimal importance sampling. Also the proposed DEKPF needs only to be pairwise synchronized and to transmit the fist order and the second order statistics between neighboring nodes. Hence, both the computation and communication cost can be reduced. The simulations under different reverberant environments show that the performance of the proposed approach is favorably comparable to that of the CEKPF tracking approach. In our future work, real deployment of wireless microphone pair network and tracking multiple simultaneously active sources will be considered.

## References

[1] J. H. Dibiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," PhD thesis, Brown University, Providence, U.S.A., 2000.

[2] Y. Huang, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, vol. 2, Jun. 5–9, 2000, pp. II909–II912.

[3] M. Brandstein and D. Ward, *Microphone Arrays. Signal Process. Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

[4] F. Talantzis, A. Pnevmatikakis, and A. Constantinides, "Audio-visual active speaker tracking in cluttered indoors environments," *IEEE Trans. Syst., Man, Cybern., B: Cybern.*, vol. 38, no. 3, pp. 799–807, 2008.

[5] M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," vol. 5, no. 1, pp. 45–50, Jan. 1997.

[6] M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer, Speech, and Language*, vol. 11, pp. 91–126, Apr. 1997.

[7] S. Gannot and T. G. Dvorkind, "Microphone array speaker localizers using spatial-temporal information," *EURASIP Journal on Applied Signal Process.*, vol. 2006, pp. 1–17, 2006.

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

[9] U. Klee, Tobias, and J. McDonough, "Kalman filters for time delay of arrival-based source localization," *EURASIP J. Applied Signal Process.*, vol. 2006, pp. 1–15, 2006.

[10] X. Zhong and J. R. Hopgood, "Nonconcurrent multiple speakers tracking based on extended kalman particle filter," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 293–296.

[11] A. Levy, S. Gannot, and E. A. P. Habets, "Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 19, no. 6, pp. 1540–1555, Aug. 2011.

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

[13] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, vol. 5, May 2001, pp. 3021–3024.

[14] E. A. Lehmann and A. M. Johansson, "Particle filter with integrated voice activity detection for acoustic source tracking," *EURASIP Journal on Advances in Signal Process.*, vol. 2007, pp. 1–11, 2007.

[15] X. Sheng and Y. H. Hu, "Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 53, no. 1, pp. 44–53, Jan. 2005.

[16] X. Zhong and J. R. Hopgood, "Time-frequency masking based multiple acoustic sources tracking applying rao-blackwellised monte carlo data association," in *Proc. IEEE 15th Workshop on Statistical Signal Process.*, Aug. 2009, pp. 253–256.

[17] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," vol. 50, no. 2, pp. 174–188, Feb. 2002.

[18] D. Simon, *Optimal State Estimation*. John Wiley and Sons, 2006.

[19] C. Chong, S. Mori, and K. Chang, *Multitarget-Multisensor Tracking: Advanced Applications*. Artech House, 1990, ch. Distributed Multitarget Multisensor Tracking.

[20] A. Dimakis, S. Kar, J. Moura, M. Rabbat, and A. Scaglione, “Gossip algorithms for distributed signal processing,” *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1847 –1864, nov. 2010.

[21] A. Mohammadi and A. Asif, “Distributed particle filter implementation with intermittent/irregular consensus convergence,” *IEEE Transactions on Signal Processing*, In Press, 2013.

[22] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, Philadelphia, 1993.

[23] J. B. Allen and D. Berkley, “Image method for efficiently simulating small-room acoust.” *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Jul. 1979.