



Speech Source Tracking Based on Distributed Particle Filter in Reverberant Environments

Ruifang Wang^{1,2} and Xiaoyu Lan¹(✉)

¹ School of Electronic and Information Engineering,
Shenyang Aerospace University, Shenyang 110136, China
rffwang0404@mail.dlut.edu.cn, lanxiaoyu1015@163.com

² School of Information and Communication Engineering,
Dalian University of Technology, Dalian 116023, China

Abstract. In reverberant and noisy environments, tracking a speech source in distributed microphone networks is a challenging problem. A speech source tracking method based on distributed particle filter (DPF) and average consensus algorithm (ACA) is proposed in distributed microphone networks. The generalized cross-correlation (GCC) function is used to approximate the time difference of arrival (TDOA) of speech signals received by two microphones at each node. Next, the multiple-hypothesis model based on multiple TDOAs is calculated as the local likelihood function of the DPF. Finally, the ACA is applied to fuse local state estimates from local particle filter (PF) to obtain a global consensus estimate of the speech source at each node. The proposed method can accurately track moving speech source in reverberant and noisy environments with distributed microphone networks, and it is robust against the node failures. Simulation results reveal the validity of the proposed method.

Keywords: Speech source tracking · Distributed particle filter · Distributed microphone networks · Average consensus

1 Introduction

The tracking of a speech source in reverberant indoor environments may help to know the speech source's position at all times, which becomes very important in many audio applications, such as audio/video conference system, source separation, beamforming and robot [1–4], and it has been an attractive research problem. Since the room reverberation brings multi-path components into speech signals received by microphones, and environmental noise can also pollute the speech signals, it could bring some challenges to accurately track moving speech source in indoor environments with microphone networks. Meanwhile, reverberation and noise will generate spurious and unreliable measurements and may lead to the tracking performance degradation for a moving speech source.

To solve this problem, Bayesian filter based speech source tracking methods have been developed, which depict the tracking problem with a state-space model and estimate the state of speech source with the state posterior [5–9]. These methods generally depended on estimated the time difference of arrival (TDOA) measurements and use both a series of past measurements and current measurements. Considering that the tracking of speech source is a nonlinear problem, an optimal approximation for the Bayesian filter via the Monte Carlo technique is applied in the speech source tracking, i.e., the particle filter (PF). A state space approach using PF was presented to track acoustic source and a general PF framework was formed in microphone networks [5]. A framework of speaker localization and tracking based on the information theory and PF was presented in [7]. In [8], a multiple talkers tracking method based on random finite set PF and time-frequency masking was discussed. A nonconcurrent multiple talkers tracking problem based upon extended Kalman particle filter (EKPF) was proposed in [9].

However, in the methods above-mentioned, their microphone networks normally are regular geometry structure, which make these methods require a central processing unit to collect all measurements for position estimate of the speech source. Thus, any failures of the central processing unit may lead to the tracking system collapse. Besides, considering the problem of node failures and lost data in the microphone networks, constructing distributed microphone networks with arbitrary layout and irregular geometry are suitable to perform speech source tracking. Then, in distributed microphone networks, the tracking methods of speech source based on distributed PF (DPF) have been discussed. In [10], a DPF algorithm was employed for the speaker tracking in the microphone pair network, in which an extended Kalman filter (EKF) is used to estimate local posterior probability for sampling particles (abbreviated as DPF-EKF). A improved distributed Gaussian PF (IDGPF) was performed to track speaker and an optimal fusion rule was employed in distributed microphone networks, in which a multiple-hypothesis model is modified as the likelihood function [11]. For non-Gaussian noise environments, a speaker tracking method based on DPF was discussed in [12]. For these methods, they employed different fusion algorithms of distributed data and different TDOA measurements.

Taking into account adverse effects of reverberation and noise, based on the DPF [13] and consensus fusion algorithm [14], a speech source tracking method in reverberant environments with distributed microphone networks is proposed in the paper. First, a dynamics model is used to describe the motion of a speech source in a room. Next, the generalized cross-correlation (GCC) estimator is applied to calculate multiple TDOAs of speech signals from each microphone pair. After that, multiple-hypothesis likelihood model is employed to compute the weights associated with particles of the local PF and the local state posterior is estimated at each node. Finally, the decentralized computation fashion of local posteriors is implemented by the average consensus algorithm and a global state estimate is obtained at each node. Especially, since the data communication does

not perform in the whole microphone network, and only occurs among the neighbor nodes, the proposed method is robust against node failures and lost data.

2 Problem Formulation and Fundamental Algorithm

2.1 Problem Formulation

In a distributed microphone network with J nodes, where a node consists of two microphones and the communication among nodes can be modeled as an undirected graph $\mathcal{G} = (\mathcal{V}, \varepsilon)$, where $\mathcal{V} = \{1, 2, \dots, J\}$ is the node set of the network and $\varepsilon \subset \{\{j, j'\} | j, j' \in \mathcal{V}\}$ is the edge set between nodes in the network. An edge $\{j, j'\} \subset \varepsilon$ indicates that node j and j' can exchange information each other. $\mathcal{M}^j = \{j' \in \mathcal{V} | (j, j') \in \varepsilon\}$ is the neighbors' set of node j .

Let the varying state $\mathbf{x}_k = [x_k, y_k, \dot{x}_k, \dot{y}_k]^T$ denote the speech source's position $[x_k, y_k]$ and velocity $[\dot{x}_k, \dot{y}_k]$ at time k in a reverberant and noisy environment, and \mathbf{y}_k denote measurement vector. The transition function f_k and measurement function h_k are used to describe the nonlinear relationships between \mathbf{x}_k and \mathbf{x}_{k-1} , between \mathbf{x}_k and \mathbf{y}_k , respectively [5, 12].

$$\mathbf{x}_k = f_k(\mathbf{x}_{k-1}) + u_k \quad (1)$$

$$\mathbf{y}_k = h_k(\mathbf{x}_k) + v_k \quad (2)$$

where u_k and v_k are the measurement and process noise at time k , respectively, both with known probability density functions.

The tracking problem of speech source is to estimate the state \mathbf{x}_k at time k based on all measurements, i.e., $\mathbf{y}_{1:k} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k\}$. The Bayesian filter for tracking problem is to recursively calculate the posterior probability density $p(\mathbf{x}_k | \mathbf{y}_{1:k})$ of \mathbf{x}_k based on the posterior probability $p(\mathbf{x}_{k-1} | \mathbf{y}_{1:k-1})$

$$p(\mathbf{x}_k | \mathbf{y}_{1:k-1}) = \int p(\mathbf{x}_k | \mathbf{x}_{k-1}) p(\mathbf{x}_{k-1} | \mathbf{y}_{1:k-1}) d\mathbf{x}_{k-1} \quad (3)$$

$$p(\mathbf{x}_k | \mathbf{y}_{1:k}) = \frac{p(\mathbf{y}_k | \mathbf{x}_k) p(\mathbf{x}_k | \mathbf{y}_{1:k-1})}{p(\mathbf{y}_k | \mathbf{y}_{1:k-1})} \quad (4)$$

where $p(\mathbf{x}_k | \mathbf{x}_{k-1})$ is the state transition density and calculated via Eq. (1); $p(\mathbf{y}_k | \mathbf{x}_k)$ is the global likelihood function and computed by Eq. (2); $p(\mathbf{y}_k | \mathbf{y}_{1:k-1}) = \int p(\mathbf{y}_k | \mathbf{x}_k) p(\mathbf{x}_k | \mathbf{y}_{1:k-1}) d\mathbf{x}_k$ is the normalized parameter [5, 6].

2.2 Particle Filter

In the tracking of speech source, taking into account that the measurement \mathbf{y}_k is nonlinear, the particle filter (PF) can obtain the optimal solution to Eqs. (3) and (4) by Monte Carlo technique. Let $\{\mathbf{X}_k^n\}_{n=1}^N$ be sampled particles and $\{w_k^n\}_{n=1}^N$ be associated weights, respectively. The PF represents the posterior probability $p(\mathbf{x}_k | \mathbf{y}_{1:k})$ using weighted particles. Using the sampling importance resampling

(SIR) filter [6], the N weighted particles are drawn from the state-transition density $p(\mathbf{X}_k^n | \mathbf{X}_{k-1}^n)$ as the proposal function, and the weight w_k^n corresponding to the n -th particle \mathbf{X}_k^n is updated as $w_k^n = p(\mathbf{y}_k | \mathbf{X}_k^n)$ [5,6].

Then the $p(\mathbf{x}_k | \mathbf{y}_{1:k})$ is written as

$$p(\mathbf{x}_k | \mathbf{y}_{1:k}) \approx \sum_{n=1}^N \tilde{w}_k^n \delta(\mathbf{x}_k - \mathbf{X}_k^n) \quad (5)$$

where \tilde{w}_k^n is the normalized weight, i.e., $\tilde{w}_k^n = w_k^n / \sum_{n=1}^N w_k^n$ and $\delta(\cdot)$ is the multi-dimensional Dirac delta function.

Based on the posterior probability, the minimum mean-square error (MMSE) estimate $\hat{\mathbf{x}}_k$ and covariance $\hat{\mathbf{P}}_k$ of \mathbf{x}_k are obtained as [6]

$$\hat{\mathbf{x}}_k = E \{ \mathbf{x}_k | \mathbf{y}_{1:k} \} = \sum_{n=1}^N \tilde{w}_k^n \mathbf{X}_k^n \quad (6)$$

$$\hat{\mathbf{P}}_k = \sum_{n=1}^N \tilde{w}_k^n (\hat{\mathbf{x}}_k - \mathbf{X}_k^n) (\hat{\mathbf{x}}_k - \mathbf{X}_k^n)^T \quad (7)$$

where $E[\bullet]$ is the mathematical expectation operation.

2.3 Time Difference of Arrival

The time difference of arrival (TDOA) measurements are calculated from the generalized cross-correlation (GCC) function $R_{k,j}(\tau)$ between two microphone signals of node j . The $R_{k,j}(\tau)$ based on the phase transform is given as [5,15]

$$R_{k,j}(\tau) = \int \frac{X_j^1(f) X_j^{2*}(f)}{|X_j^1(f) X_j^{2*}(f)|} e^{j2\pi f \tau} df \quad (8)$$

where $X_j^1(f)$ and $X_j^2(f)$ denote the frequency domain signals received by microphone pair, and superscript $*$ denotes the complex conjugation.

The TDOA measurement estimate $\hat{\tau}_k^j$ at node j corresponds to the large peak of the $R_{k,j}(\tau)$, written as

$$\hat{\tau}_k^j = \underset{\tau \in [-\tau^{j\max}, \tau^{j\max}]}{\operatorname{argmax}} (R_{k,j}(\tau)) \quad (9)$$

where $\tau^{j\max}$ is the maximal TDOA at node j .

However, in reverberant and noisy environments, only considering a TDOA estimate from the largest peak of $R_{k,j}(\tau)$ in Eq. (9) may bring ambiguous TDOA estimates, which can lead to spurious estimates of the speech source's position. Generally, taking multiple TDOA estimates from local largest peak of $R_{k,j}(\tau)$ has become popularly in speech source tracking problem [11,12]. Calculate U_k TDOA estimates to constitute local measurement of node j , i.e., $\mathbf{y}_k^j = [\hat{\tau}_{k,1}^j, \hat{\tau}_{k,2}^j, \dots, \hat{\tau}_{k,U_k}^j]^T$, where $\hat{\tau}_{k,i}^j$ ($i = 1, 2, \dots, U_k$) is taken from the i -th largest local peak of $R_{k,j}(\tau)$.

3 Speech Source Tracking Based on DPF

3.1 Speech Source Dynamical Model

The Langevin model [5] is used to be speech source dynamical model, which describe the varying speech source's motion in indoor environments. It is assumed to be independent in each Cartesian coordinate, written as

$$\mathbf{x}_k = \begin{bmatrix} \mathbf{I}_2 & aT \otimes \mathbf{I}_2 \\ 0 & a \otimes \mathbf{I}_2 \end{bmatrix} \mathbf{x}_{k-1} + \begin{bmatrix} bT \otimes \mathbf{I}_2 & 0 \\ 0 & b \otimes \mathbf{I}_2 \end{bmatrix} \mathbf{u}_{k-1} \quad (10)$$

where T is the discrete time interval, \mathbf{I}_2 is the second-order identity matrix, \otimes is the Kronecker product operation, \mathbf{u}_{k-1} is the time-uncorrelated Gaussian noise vector, $a = \exp(-\beta\Delta T)$ and $b = \bar{v}\sqrt{1-a^2}$, where β and \bar{v} are the rate constant and steady-state velocity parameter, respectively. Setting suitable values for β and \bar{v} can simulate the realistic speech source motion.

3.2 Distributed Particle Filter Based on Average Consensus Algorithm

In [13], a distributed particle filter (DPF) is presented to achieve a consensus-based calculation of posterior parameters from the local PF at each node in the distributed network. All posterior parameters are assumed as Gaussian probability density. Then the global state posterior is calculated based on local posterior parameters via distributed data fusion algorithm.

In the distributed microphone network with J nodes, the local measurements $\mathbf{y}_k^j (j = 1, 2, \dots, J)$ of the state \mathbf{x}_k form the measurement vector \mathbf{y}_k , written as

$$\mathbf{y}_k = \left[(\mathbf{y}_k^1)^T, (\mathbf{y}_k^2)^T, \dots, (\mathbf{y}_k^J)^T \right]^T \quad (11)$$

Node j first performs a local PF and calculates a local posterior $p(\mathbf{x}_k | \mathbf{y}_{1:k-1}, \mathbf{y}_k^j)$ incorporating all nodes' measurements $\mathbf{y}_{1:k-1}$ up to time $k-1$ and the local measurement \mathbf{y}_k^j . Then the Gaussian estimation of the $p(\mathbf{x}_k | \mathbf{y}_{1:k-1}, \mathbf{y}_k^j)$ is computed in term of weighted particles from Eqs. (6) and (7), which are local MMSE estimate $\hat{\mathbf{x}}_{k,j}$ and covariance estimate $\hat{\mathbf{P}}_{k,j}$, respectively.

Next, they are propagated among the neighbor nodes in the distributed microphone network, and fused by the typical average consensus algorithm (ACA) [14] which performs distributed linear iterations make each node obtain the converging average value. The consensus iteration calculation at node j can be given as

$$\mathbf{t}_j(m+1) = \mathbf{t}_j(m) + \alpha \sum_{i \in \mathcal{M}^j} [\mathbf{t}_i(m) - \mathbf{t}_j(m)] \quad (12)$$

where α denotes the weight corresponding to edge $\{i, j\} \subset \varepsilon$ in the distributed network, and m denotes the time index of consensus iteration. The variable $\mathbf{t}_j(m+1)$ will converge to the global average value at node j after M iterations.

Finally, the global consensus posterior $p(\mathbf{x}_k | \mathbf{y}_{1:k})$ can be obtained at each node of the distributed microphone network.

The DPF based on the ACA is nearly not affected by changing topologies structure and node link failures of the distributed network since in the iteration calculations of the ACA, each node only performs the data communications among the neighbor nodes.

3.3 Multiple-Hypothesis Likelihood Model

The particles' weights of local PF at node j are considered as local likelihood functions, i.e., $p(\mathbf{y}_k^j | \mathbf{x}_k)$, which are computed by multiple-hypothesis likelihood model based on the U_k TDOAs. Due to reverberation and noise of indoor environment, in local measurement \mathbf{y}_k^j , at most one TDOA $\hat{\tau}_{k,i}^j$ corresponds to the true speech source's position. If the $\hat{\tau}_{k,i}^j$ corresponds to the true position, let $f_{k,i}^j = T$, else, let $f_{k,i}^j = F$. These hypotheses can be described as [9, 12]

$$\begin{cases} \mathcal{H}_0 = \{f_{k,i}^j = F; i = 1, 2, \dots, U_k\} \\ \mathcal{H}_i = \{f_{k,i}^j = T; f_{k,g}^j = F; g = 1, 2, \dots, U_k, i \neq g\} \end{cases} \quad (13)$$

where \mathcal{H}_0 indicates that none of TDOAs corresponds to the true speech source's position, and \mathcal{H}_i denotes that only the i -th TDOA $\hat{\tau}_{k,i}^j$ corresponds to the true position.

Assume that the hypotheses of Eq. (13) are mutually exclusive, then local likelihood function $p(\mathbf{y}_k^j | \mathbf{x}_k)$ can be given as

$$p(\mathbf{y}_k^j | \mathbf{x}_k) = \sum_{q=0}^{U_k} s_q p(\mathbf{y}_k^j | \mathbf{x}_k, \mathcal{H}_q) \quad (14)$$

where s_q is the prior probability of the hypothesis \mathcal{H}_q , and $\sum_{q=0}^{U_k} s_q = 1$.

Assume that the U_k TDOAs in local measurement \mathbf{y}_k^j are mutually independent conditioned on \mathbf{x}_k and \mathcal{H}_q , if the TDOA $\hat{\tau}_{k,i}^j$ corresponds to the true speech source's position, the likelihood function is defined as a Gaussian distribution; else, it is defined as a uniform distribution over the set of admissible TDOA $[-\tau^{j \max}, \tau^{j \max}]$, written as [9]

$$\begin{cases} p(\mathbf{y}_k^j | \mathbf{x}_k, \mathcal{H}_0) = \frac{1}{(2\tau^{j \max})^{U_k}} \\ p(\mathbf{y}_k^j | \mathbf{x}_k, \mathcal{H}_i) = \frac{1}{(2\tau^{j \max})^{U_k-1}} \mathcal{N}(\hat{\tau}_{k,i}^j; \tau_k^j(\mathbf{x}_k), \sigma^2) \end{cases} \quad (15)$$

Then, the local likelihood function $p(\mathbf{y}_k^j | \mathbf{x}_k)$ in Eq. (14) is written as

$$p(\mathbf{y}_k^j | \mathbf{x}_k) = \eta \left(\frac{s_0}{2\tau^{j \max}} + \sum_{i=1}^{U_k} s_i \mathcal{N}(\hat{\tau}_{k,i}^j; \tau_k^j(\mathbf{x}_k), \sigma^2) \right) \quad (16)$$

where $\eta = \frac{1}{(2\tau^{j \max})^{U_k-1}}$.

3.4 Speech Source Tracking Based on DPF

Based upon the above-mentioned discussions, a speech source tracking method based on the DPF and ACA is proposed in reverberant environments with distributed microphone networks (abbreviated to DPF-ACA). First, each node calculates the GCC function of speech signals received by a microphone pair and chooses multiple TDOAs as its local measurement. Based on them, predict the particles via the Langevin model and compute the local multiple-hypothesis likelihood as weights of particles for the local PF at each node. Next, estimate the local state and corresponding covariance with representation of weighted particles. Finally, fuse all local state estimates via the average consensus algorithm, and all nodes can obtain a global consensus estimate. The DPF-ACA algorithm is summarized in Algorithm 1. Furthermore, since the data communications occur only in the neighbor nodes of distributed networks, the proposed method is robust against node failures or the data lost.

Algorithm 1. DPF and ACA Based Speech Source Tracking.

- 1: Calculate the GCC function $R_{k,j}(\tau)$ according to Eq.(8), $k = 1, 2, \dots, K$, $\forall j \in \mathcal{V}$, $j = 1, 2, \dots, J$, where K denotes the maximal time index;
 - 2: Choose U_k TDOAs to form local measurement \mathbf{y}_k^j ;
 - 3: Predict particles $\{\tilde{\mathbf{X}}_{k,j}^n\}_{n=1}^N$ by broadcasting $\{\mathbf{X}_{k-1,j}^n\}_{n=1}^N$ according to Eq. (10);
 - 4: Compute the local weights $\{w_{k,j}^n\}_{n=1}^N$ according to Eq. (16);
 - 5: Normalize the weights: $\tilde{w}_{k,j}^n = w_{k,j}^n / \sum_{n=1}^N w_{k,j}^n$;
 - 6: Estimate the local state $\hat{\mathbf{x}}_{k,j}$ and corresponding covariance $\hat{\mathbf{P}}_{k,j}$ according to Eqs. (6) and (7);
 - 7: Fuse $\{\hat{\mathbf{x}}_{k,j}\}_{j=1}^J$ and $\{\hat{\mathbf{P}}_{k,j}\}_{j=1}^J$ according to Eq. (12);
 - 8: Calculate the global estimate at node j : $\hat{\mathbf{x}}_{k,j} = \hat{\mathbf{x}}_{k,j}(M)$, $\hat{\mathbf{P}}_{k,j} = \hat{\mathbf{P}}_{k,j}(M)$;
 - 9: Sample particles $\{\mathbf{X}_{k,j}^n\}_{n=1}^N$ from $\mathcal{N}(\mathbf{x}_{k,j}; \hat{\mathbf{x}}_{k,j}, \hat{\mathbf{P}}_{k,j})$;
 - 10: **return** $\hat{\mathbf{x}}_{k,j}, \hat{\mathbf{P}}_{k,j}$;
-

4 Simulations and Result Discussions

4.1 Simulation Setup

In the simulation, consider that a female speech source moves in a office room, whose moving trajectory is a curve which start point is (0.9 m, 2.65 m) and end point is (4.1 m, 2.65 m). There are 12 omni-direction microphone pairs irregularly

and randomly installed in the room shown in Fig. 1. In advance, the distributed microphone network has been constructed via choosing microphones adaptively [16]. A microphone pair is considered as a node, in which the spacing distance of two microphones is set as 0.6 m. The communication graph of the distributed microphone network is shown in Fig. 2, where the line between two nodes indicates that they can exchange information each other. Each node has its neighbor nodes and can communicate with neighbor node only when their communication radius between them is less than 1.8 m. Meanwhile, the height of microphones is set as 1.5 m, which is same as that of the speech source, and a two-dimensional speech source tracking problem is focused in this paper.

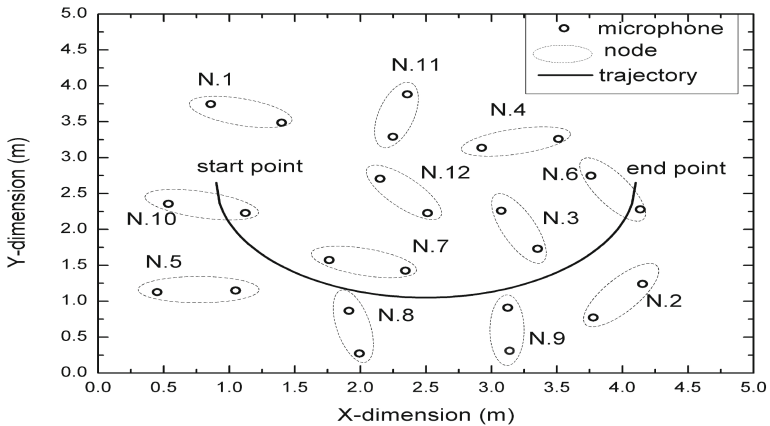


Fig. 1. Speech source trajectory and layout of the 12 microphone pairs in $X - Y$ plane.

In the simulation, the speech source is a female speech with a length of nearly 8 s with sampling frequency is 16 kHz. The speech signals are split as 120 frames and the discrete time interval T is 64 ms. The speech signals captured by each microphone are created by the well-known image method [17], which can simulate the indoor environment acoustics under different background noise and reverberations. The parameters T_{60} and signal noise ratio (SNR) are used to simulate different reverberations and different background noise, respectively. The configuration of simulation parameter is as follows. For the Langevin model, the parameter settings are $\bar{v} = 1\text{ms}^{-1}$ and $\beta = 10\text{ s}^{-1}$. For the PF, the sampling number of particles is $N = 500$ and the initial prior of the speech source state is considered as a Gaussian distribution, with mean vector $\mu_0 = [1.0, 2.6, 0.01, 0.01]^T$ and covariance $\Sigma_0 = \text{diag}([0.05, 0.05, 0.0025, 0.0025])$ set randomly. For the TDOA estimates, the number of the TDOA candidates is $U_k = 4$. For the multi-hypothesis likelihood, the standard deviation of the TDOA error is $\sigma = 50\mu\text{s}$ and the prior of \mathcal{H}_0 is $s_0 = 0.25$. For the average consensus algorithm, consensus iterations is $M = 25$.

Root Mean Square Error (RMSE) of the speech source's position has been employed in tracking performance evaluation widely and the average of RMSEs (ARMSE) over M_c Monte Carlo simulations is given as

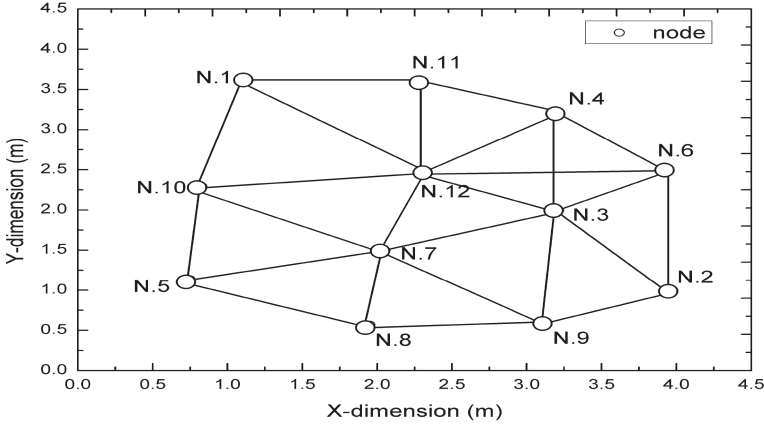


Fig. 2. Communication graph of the distributed microphone network with 12 nodes.

$$\text{ARMSE} = \frac{1}{M_c} \sum_{m=1}^{M_c} \sqrt{\frac{1}{K} \sum_{k=1}^K \|\mathbf{s}_{\mathbf{x}_k} - \mathbf{s}_{\hat{\mathbf{x}}_{m,k}}\|^2} \tag{17}$$

where m is the cycle index of the Monte Carlo simulation running, $\mathbf{s}_{\mathbf{x}_k}$ denotes the true position of the speech source, and $\mathbf{s}_{\hat{\mathbf{x}}_{m,k}}$ is the speech source’s position estimate of the m -th Monte Carlo simulation running.

4.2 Result Discussions

To evaluate the validation of the proposed method (DPF-ACA), some simulation experiments under different SNRs and different T_{60} values are conducted, comparing with the existing speech source tracking methods, i.e., [5] (abbreviated to PF), [10] (abbreviated to DPF-EKF), and [11] (abbreviated to IDGPF). Based on the same simulation setup, all methods are evaluated in form of the ARMSE results according to Eq. (17) averaged over Monte Carlo simulations, where times of Monte Carlo simulations is $M_c = 70$.

Effect of Reverberation Time T_{60} . Figure 3 indicates that the tracking results of all methods under the different T_{60} values, i.e., $T_{60} = \{100, 150, \dots, 600\}$ ms when the SNR is 10 dB.

With the rise of reverberation time, it can be observed from Fig. 3 that tracking performance of all methods becomes worse and worse. Specially, the IDGPF method has larger errors under different T_{60} values. It means heavier reverberation will bring bad TDOA estimations at each node in the distributed microphone network. Due to taking only one TDOA from the peak of GCC function for sampling particles in the DPF-EKF method, it has poor tracking accuracy when $T_{60} > 200$ ms. We can find that the tracking performance of the PF is the best when T_{60} changes from 100 ms to 600 ms. However, its central processing fashion requires that the central processing unit can not have any

failures. It can be clearly seen that the proposed method always has lower values of ARMSE and better tracking accuracy when reverberation time T_{60} becomes heavier and heavier, which indicates the proposed method is robust against the changes of environmental reverberation.

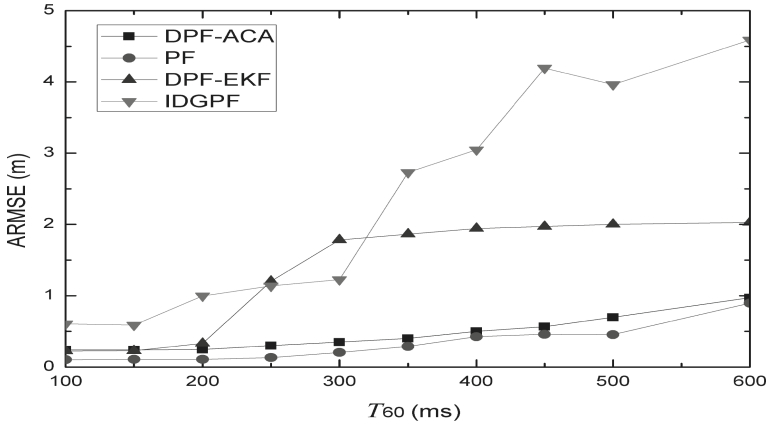


Fig. 3. ARMSE results versus different T_{60} in the environment with SNR = 10 dB.

Effect of Signal Noise Ratio (SNRs). Figure 4 illustrates that the tracking results of all methods under different SNRs, i.e., SNR = {0, 5, ..., 35} dB, when the reverberation time $T_{60} = 100$ ms.

With the increases of SNR, we can observe from Fig. 4 that the ARSME values of all methods change smaller gradually and their tracking accuracies become higher and higher. It can be seen that when SNR < 10 dB, the IDGPF method and DPF-EKF method bring serious degradation of tracking performance, which means background noise has an important influence in their tracking performances. Although the PF has the best tracking accuracy in Fig. 4, it is limit to the central data processing manner. Furthermore, as a distributed tracking algorithm, when the SNR increases from 0 dB to 35 dB, ARMSE values of the DPF-ACA are on the decline, which shows more stable and better tracking performance. It implies that the proposed method is a valid tracking method for speech source under lower SNRs environments, especially.

Effect of Node Failures. Node failures in microphone networks indicate they can not exchange data with their own neighbor nodes, which can cause the decline of the valid node’s number and the change of communication graph in Fig. 2. Tracking trajectories of all tracking methods are displayed in Fig. 5 under the environment with $T_{60} = 100$ ms and SNR = 15 dB, when there are three fault nodes in the distributed microphone network, i.e., N.1 node, N.5 node, and N.11 node shown in Fig. 2.

When three nodes can not communicate with their neighbor nodes in the distributed microphone network, the number of valid nodes of network in Fig. 2

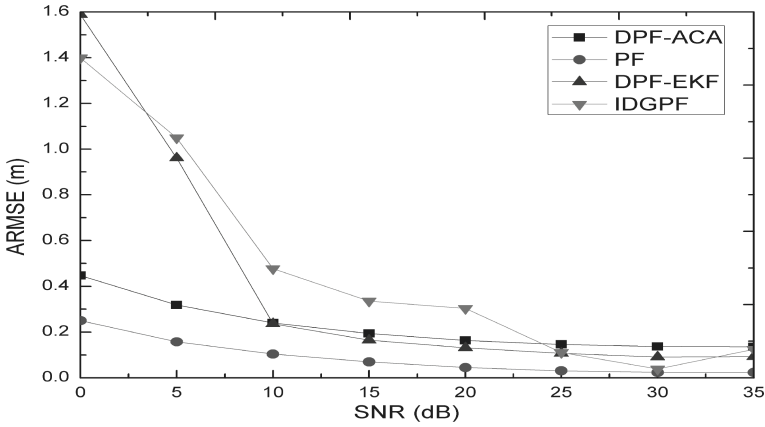


Fig. 4. ARMSE results versus different SNRs in the environment with $T_{60} = 100$ ms.

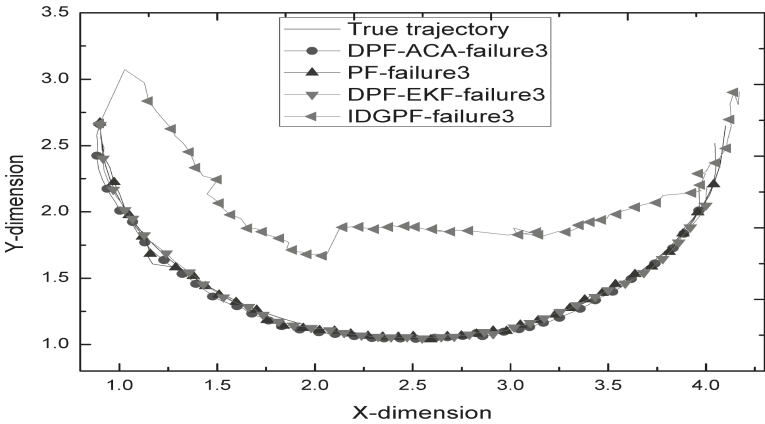


Fig. 5. Speech source tracking results in the network with three fault nodes.

changes to 9 and there are only 9 local measurements, which could affect the tracking performance of all tracking methods of speech source. However, it can be observed that the DPF-ACA, PF, and DPF-EKF methods can successfully track the moving speech source with smaller ARMSE values, which indicates the node failures have little impact them. For the IDGPF method, it generates larger tracking errors in tracking true trajectory of the speech source. Meanwhile, since the DPF-ACA only executes local data exchange, the tracking performance of the proposed method as a distributed tracking algorithm of the speech source, is nearly unaffected by node faults and it is scalable in distributed microphone networks.

5 Conclusions

In the paper, a speech source tracking method based on the DPF and ACA in distributed microphone networks is proposed. Each node first performs the local PF to obtain local state posterior. Next, taking into account the environmental noise and reverberation, the GCC-PHAT function is used to estimate multiple TDOAs which are employed to calculate multiple-hypothesis model as weights of particles for the local PF. Finally, the local state estimates are fused via average consensus algorithm to acquire the global consensus estimate at each node in the distributed microphone networks. Simulation experiments with existing speech source tracking methods indicate that the proposed method has better tracking performance in environments of lower SNRs and heavier reverberations. Besides, owing to only executing communications among neighbor nodes, the proposed method is almost unaffected by node failures.

Acknowledgement. This work was supported by National Science Foundation for Young Scientists of China (Grant No.61801308).

References

1. Spexard, T.P., Hanheide, M., Sagerer, G.: Human-oriented interaction with an anthropomorphic robot. *IEEE Trans. Robotics* **23**(5), 852–862 (2007)
2. Chen, B.W., Chen, C.Y., Wang, J.F.: Smart homecare surveillance system: behavior identification based on state-transition support vector machines and sound directivity pattern analysis. *IEEE Trans. Syst., Man, Cybern. A Syst.* **43**(6), 1279–1289 (2013)
3. Kapralos, B., Jenkin, M.R.M., Evangelos, M.: Audiovisual localization of multiple speakers in a video teleconferencing setting. *Int. J. Imaging Syst. Technol.* **13**(1), 95–105 (2003)
4. Nakadai, K., Nakajima, H., Murase, M., et al.: Robust tracking of multiple sound sources by spatial integration of room and robot microphone arrays. In: *International Conference on Acoustic, Speech, Signal Process*, Toulouse, France, pp. IV-929–IV-932 (2006)
5. Ward, D.B., Lehmann, E.A., Williamson, R.C.: Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Trans. Speech Audio Process.* **11**(6), 826–836 (2003)
6. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.* **50**(2), 174–188 (2002)
7. Talantzis, F.: An acoustic source localization and tracking framework using particle filtering and information theory. *IEEE Trans. Audio Speech Lang. Process.* **18**(7), 1806–1817 (2010)
8. Zhong, X., Hopgood, J.R.: A time-frequency masking based random finite set particle filtering method for multiple acoustic source detection and tracking. *IEEE Trans. Audio Speech Lang. Process.* **23**(12), 2356–2370 (2015)
9. Zhong, X., Hopgood, J.R.: Particle filtering for TDOA based acoustic source tracking: nonconcurrent multiple talkers. *Signal Process.* **96**(5), 382–394 (2014)

10. Zhong, X., Mohammadi, A., Wang, W., Premkumar, A.B., Asif, A.: Acoustic source tracking in a reverberant environment using a pairwise synchronous microphone network. In: 16th International Conference on Information Fusion, Istanbul, Turkey, pp. 953–960 (2013)
11. Zhang, Q., Chen, Z., Yin, F.: Speaker tracking based on distributed particle filter in distributed microphone networks. *IEEE Trans. Syst. Man Cybern. Syst.* **47**(9), 2433–2443 (2017)
12. Wang, R., Chen, Z., Yin, F., Zhang, Q.: Distributed particle filter based speaker tracking in distributed microphone networks under non-Gaussian noise environments. *Digital Signal Process.* **63**, 112–122 (2017)
13. Hlinka, O., Hlawatsch, F., Djuric, P.M.: Distributed particle filtering in agent networks: a survey, classification, and comparison. *IEEE Signal Process. Mag.* **30**(1), 61–81 (2013)
14. Lin, X. and Boyd, S.: Fast linear iterations for distributed averaging. In: 42nd International Conference on Decision and Control, Maui, USA, pp. 4997–5002 (2004)
15. Knapp, C., Carter, G.C.: The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech Signal Process.* **24**(4), 320–327 (1976)
16. Mohammadi, A., Asif, A.: Consensus-based distributed dynamic sensor selection in decentralized sensor networks using the posterior Cramér-Rao lower bound. *Signal Process.* **108**, 558–575 (2015)
17. Lehmann, E.A., Johansson, A.M. and Nordholm, S.: Reverberation-time prediction method for room impulse responses simulated with the image-source model. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, USA, pp. 159–162 (2007)