Distributed Kalman filter-based speaker tracking in microphone array networks

Ye Tian, Zhe Chen, Fuliang Yin

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

Abstract

Using a microphone array network, a speaker tracking method based on distributed Kalman filter (DKF) in a noisy and reverberant environment is proposed. Firstly, the time delay of arrival (TDOA) in each microphone pair is estimated by the generalized cross-correlation (GCC) method. Next, the Langevin model is used as state equation to model the speaker’s movement, meanwhile the measurement equations with true TDOA are deduced by linearizing the TDOA model. Finally, the moving speaker’s positions are estimated by distributed Kalman filtering in a microphone array network. The proposed method is scalable. It can obtain a trajectory of the speaker’s movement smoothly with excellent tracking accuracy. Simulation results verify the effectiveness of the proposed method.

1. Introduction

Speaker localization and tracking with microphone arrays is useful in many applications, including audio/video conference system [1], smart video monitor system [2], robot, human–machine interface, far distance speech capture and recognition, etc.

The topics of speaker localization [3–5] and speaker tracking [6–11] have been studied for many years. However, traditional methods usually require dedicated devices, and need to know the positions and geometry structure of microphone arrays.

In practice, it is possible that the geometry structure of microphone arrays is irregular and the positions of them are also distributed randomly. The geometry structure and the positions of microphone arrays can be obtained by self-calibration methods [12,13]. To determine speaker’s positions in spatially irregular microphone arrays, the distributed speaker localization methods [14,15] were proposed recently. In [16], the global coherence field (GCF) method was proposed, which was defined over the space of possible sound source locations to represent the plausibility that a sound source was active at a given point. In [17,18], the GCF was extended to Oriented GCF (OGCF) which was allowed to estimate both the positions and geometry structure of microphone arrays.

In practice, it is possible that the geometry structure of microphone arrays is irregular and the positions of them are also distributed randomly. The geometry structure and the positions of microphone arrays can be obtained by self-calibration methods [12,13]. To determine speaker’s positions in spatially irregular microphone arrays, the distributed speaker localization methods [14,15] were proposed recently. In [16], the global coherence field (GCF) method was proposed, which was defined over the space of possible sound source locations to represent the plausibility that a sound source was active at a given point. In [17,18], the GCF was extended to Oriented GCF (OGCF) which was allowed to estimate both the positions and geometry structure of microphone arrays.

In this paper, the DKF theory is introduced into a distributed microphone array network and a DKF-based speaker tracking method in a noisy and reverberant environment is proposed.
Firstly, the time delay of arrival (TDOA) of the speech signals received by microphone arrays in the network is estimated by the generalized cross-correlation (GCC) method. For each node in the microphone array network, whether or not the TDOA is true is checked, and the true measurements are gathered from its neighboring nodes for speaker tracking. Then, the Langevin model [8,10] is introduced as state equation to represent the time-varying locations of a moving speaker, meanwhile the measurement equations with true TDOA are deduced by linearizing the TDOA model. Finally, the distributed Kalman filter is used to estimate the time varying speaker’s positions. Since each node in the network only communicates with its neighbors, the proposed speaker tracking method is scalable (i.e. new nodes can join in the network or any node in the network can leave freely), and is robust against lost data or link failure.

The rest of this paper is organized as follows. The DKF theory is introduced in Section 2. The DKF-based speaker tracking method is proposed in Section 3. The effectiveness of the proposed method is verified with simulations in Section 4. Finally, conclusions are drawn in Section 5.

2. Distributed Kalman filter (DKF)

2.1. Data model and problem formulation

Consider a sensor network with N nodes labeled by an index \( i = 1, 2, \ldots, N \) spatially distributed as shown in Fig. 1. The communications between them are modeled by a graph \( \mathcal{G} = (\mathcal{V}, \mathcal{E}) \), where \( \mathcal{V} = \{1, 2, \ldots, N\} \) is the vertex set, \( \mathcal{E} \subset \{(i,j) | i,j \in \mathcal{V}\} \) is the edge set. An edge \((i,j)\) in \(\mathcal{E}\) if and only if the node \(i\) can communicate with the node \(j\). We only consider undirected communication structures, i.e. graphs in which \((i,j)\) \(\in\mathcal{E} \iff (j,i) \in \mathcal{E}\) and assume that \(\mathcal{G}\) is connected, i.e. there is a path between any two nodes. Let \(\mathcal{N}_i = \{j \in \mathcal{V} | (i,j) \in \mathcal{E}\} \cup \{i\}\) denote the set of neighbors of node \(i\) at time \(k\), where a node is also regarded as a neighbor of itself. Node \(i\) takes a measurement \(y_{i,k} \in \mathbb{R}^p\) of the common environment state \(x_i \in \mathbb{R}^p\) independently at time \(k\), where \(\mathbb{R}^p\) denotes a \(n \times 1\) real column vector space. The state-space model associated with the measurement of node \(i\) is of the form

\[
\begin{align*}
\mathbf{x}_{i,k+1} &= F_k \mathbf{x}_i + \Gamma_k w_k \\
\mathbf{y}_{i,k} &= H_k \mathbf{x}_i + v_{i,k}
\end{align*}
\]

(1)

where \(w_k\) is the process noise; \(v_{i,k}\) is the measurement noise; and \(F_k, H_k\) and \(\Gamma_k\) are the real transformation matrices at time \(k\). The initial state \(x_0\) is assumed to be zero-mean with covariance matrix \(P_0 > 0\), uncorrelated with \(w_0\) and \(v_{i,0}\).

Assume \(w_k\) and \(v_{i,k}\) are zero-mean, uncorrelated white noises with the relationship below

\[
\mathbb{E} \begin{bmatrix} w_k \\ v_{i,k} \\ W_{k} \\ V_{i,k} \end{bmatrix} = \begin{bmatrix} Q_k \delta_{k,i} & 0 \\ 0 & R_{k} \delta_{k,i} \delta_{i,j} \end{bmatrix}.
\]

(2)

where superscript \((\cdot)^T\) denotes the matrix transpose; \(\delta_{k,i}\) is the Kronecker delta function; and \(Q_k, R_{k,i}\) are assumed to be positive definite.

2.2. Distributed Kalman filtering algorithm

In [23], a distributed Kalman filter was proposed with objective for every node in the network to compute an estimate of the unknown state \(x_i\) by sharing data only with its neighbors \(\mathcal{N}_i\) at time \(k\). The configuration is shown in Fig. 1. The distributed Kalman filter, also called diffusion Kalman filter, is obtained by adding a diffusion step, which is a convex combination of neighboring estimates after a conventional Kalman filtering measurement update.

The diffusion Kalman filtering algorithm defines an \(N \times N\) matrix \(C\) with real, non-negative entries \(c_{ij}\), which satisfies

\[
i^T C = 1^T, \quad c_{ii} > 0 \quad \forall i \quad \text{and} \quad c_{ij} = 0 \quad \text{if} \quad i \neq j
\]

(3)

where \(i\) is a \(N \times 1\) column vector with unity entries; and \(c_{ij}\) is the \((i,j)\) element of matrix \(C\). The matrix \(C\) is called the diffusion matrix, since it governs the diffusion process, and plays an important role in the steady-state performance of the network.

The diffusion Kalman filtering algorithm is summarized below.

**Algorithm 1.** Diffusion Kalman filter

Start with: \(\mathbf{x}_{i,0-1} = 0, \quad P_{i,0-1} = P_0 > 0, \quad k = 0\).

At time \(k\), the following steps are calculated.

**Step 1** Incremental Update:

\[
\begin{align*}
\psi_{i,k} &= \mathbf{x}_{i,k-1} \\
P_{i,k} &= P_{i,k-1}
\end{align*}
\]

for every neighboring node \(i \in \mathcal{N}_{i,k}\), repeat:

\[
\begin{align*}
R_k &= R_k + H_k P_{i,k} H_k^T \\
\psi_{j,k} &= \mathbf{x}_{j,k} + P_{i,k} H_k^T (y_{j,k} - H_k \psi_{j,k}) \\
P_{j,k} &= P_{j,k} - P_{i,k} H_k R_k^{-1} H_k^T P_{j,k}
\end{align*}
\]

end

**Step 2** Diffusion Update:

\[
\begin{align*}
\mathbf{x}_{i,k} &= \mathbf{x}_{i,k-1} - c_{i,j} \psi_{j,k} \\
P_{i,k} &= P_{i,k-1} - c_{i,k} P_{i,k} F_k^T + F_k R_k F_k^T
\end{align*}
\]

where \(P_{i,k}\) denotes the covariance matrix of estimation error \(\tilde{x}_{i,k} = x_i - \hat{x}_{i,k}\); and ‘\(-\)’ denotes a sequential, or non-concurrent assignment. According to the algorithm above, node \(i\) receives the message \(m_{si} = (H_k R_k y_{i,k})\) from its neighbors \(\mathcal{N}_{i,k}\) for the incremental update and sends message \(m_{is} = (\psi_{i,k})\) to its neighbors \(\mathcal{N}_{i,k}\) for the diffusion update at time \(k\).

The works in [24,25] applied a weight adaptive strategy to diffusion matrix, which could be adapted to the changes in the data statistics in the diffusion update of the diffusion Kalman filtering algorithm.

Let \(n_i\) denote the degree of node \(i\) (i.e. the number of nodes connected to the node \(i\) including itself), and \(\{i_1, i_2, \ldots, i_{n_i}\}\) denote the indexes of the neighbors of node \(i\). A matrix \(S_i\) is defined as

\[
S_i = [e_{i_1}, e_{i_2}, \ldots, e_{i_{n_i}}]^T, \quad (4)
\]

where \(e_i\) denotes the \(i\)th column of an \(N \times N\) identity matrix. According to [24,25], the diffusion matrix can be obtained adaptively as follows.

![Fig. 1. At time \(k\), node \(i\) collects a measurement \(y_{i,k}\).](image-url)
Start with $b_{i0} \in \mathbb{R}^n$ for every node $i$ such that $1^T b_{i0} = 1$. Then, for $k \geq 0$, repeat the following operations

$$
\begin{align*}
\begin{cases}
    b_{i,k+1} = b_{i,k} + \mu A_i q_{i,k} - Q_{i,k} b_{i,k} \\
    c_i^{(k+1)} = S b_{i,k+1}
\end{cases}
\end{align*}
$$

(5)

where $\mu$ denotes the step size; $c_i^{(k+1)}$ denotes the $i$th column of diffusion matrix $C$ at time $k+1$; and $q_{i,k} = E\{\psi_i^T x_i\} \approx \psi_i^T x_{i,k-1}$. $Q_{i,k} = E\{\psi_i^T x_{i,k}\} \approx \psi_i^T x_{i,k-1}$. $A_i = I_n - \frac{1}{\mu} 1^T 1^T$.

3. A distributed Kalman filter based speaker tracking method

In this section, a distributed Kalman filter based speaker tracking method using a microphone array network is proposed. First, the Langevin model is used to represent the time-varying locations of a moving speaker in a room. Then, the generalized cross-correlation (GCC) method is used to estimate the TDOAs. Finally, the measurement equation is deduced to track a speaker.

3.1. Speaker’s dynamic model

The Langevin model can represent the time-varying locations of a speaker moving in a room [8,10]. In this model, the speaker’s movement in each Cartesian coordinate is assumed to be independent. The state at time $k$ is defined as $x_k = (x_k, y_k, z_k)^T$ which includes the speaker’s position $(x_k, y_k)$ and velocity $(\dot{x}_k, \dot{y}_k)$. The speaker’s motion can be modeled by a dynamic process

$$
\begin{align*}
x_{k+1} &= F x_k + \Gamma_1 w_k, \quad (6)
\end{align*}
$$

where $F$ and $\Gamma_1$ are the real transformation matrices; and $w_k$ is a time-uncorrelated Gaussian white noise. In the $x$-coordinate, this motion is described as

$$
\begin{align*}
\begin{cases}
x_{k+1} &= \dot{x}_k + \Delta T \dot{x}_k \\
x_{k+1} &= a_s \dot{x}_k + b_s w_k,
\end{cases}
\end{align*}
$$

(7)

where $\Delta T$ is the discrete time interval, and

$$
\begin{align*}
a_s = \exp(-\beta_s \Delta T), \quad b_s = \nu_s \sqrt{1 - a_s^2},
\end{align*}
$$

(8)

where $\beta_s$ and $\nu_s$ are the rate constant and the steady-state root-mean-square velocity, respectively. The dynamics and parameters for the other Cartesian dimensions are identical.

3.2. TDOA estimates

For the sake of simplicity, we assume that each node in the network contains a pair of microphones. For speaker tracking, we are interested in the TDOA estimates that take the positions of peaks in the GCC function [26] between the signals received by microphone pairs. Let $X_i(f)$ and $X_2(f)$ denote the frequency domain signals received by two microphones. The cross-spectral density of the signals $X_i(f)$ and $X_2(f)$ is denoted by

$$
S_{X_iX_2}(f) = E\{X_i(f) X_2(f)^\dagger\},
$$

(9)

where $X_2(f)$ is the complex conjugate of $X_2(f)$. The TDOA can be computed by a GCC estimator as follows:

$$
\hat{\tau} = \arg \max_{\tau \in [-\tau_{\text{max}}, \tau_{\text{max}}]} R_{X_iX_2}(\tau),
$$

\(10\)

$$
R_{X_iX_2}(\tau) = \int_{-\infty}^{+\infty} \phi(f) S_{X_iX_2}(f) e^{2\pi i f \tau} df,
$$

(11)

where $R_{X_iX_2}(\tau)$ is the generalized cross-correlation (GCC) function; $\tau_{\text{max}}$ is the maximal TDOA; and $\phi(f)$ is the weighting function.

The phase transform (PHAT) function [7,26], which can be expressed as

$$
\phi(f) = \frac{1}{|S_{X_iX_2}(f)|}
$$

(12)

has been proved to be an effective GCC weight for TDOA estimation in reverberant environments.

3.3. Measurement equation

Let $\tau_{ik}$ denote the estimated TDOA at the microphone pair $(p_{i1}, p_{i2})$ at node $i$ and time $k$. The estimated TDOA is modeled by

$$
\hat{\tau}_{ik} = \tau_{ik} + \nu_{ik}, \quad i = 1, 2, \ldots, N
$$

(13)

where $\tau_{ik} = (|s_{ik} - p_{i1}|| - |s_{ik} - p_{i2}|)/c$ is the TDOA from the speaker to the $i$th microphone pair; $s_{ik} = (x_k, y_k)^T$ is the speaker’s position at time $k$; $c$ is the sound speed; and $\nu_{ik}$ is time-uncorrelated noise. The relationship between $s_{ik} = (x_k, y_k)^T$ and the state of a speaker $x_k = (x_k, y_k, z_k)^T$ can be described as $s_{ik} = B x_k$, where $B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

Due to room reverberation, an erroneous TDOA may be estimated. The probability of a true TDOA can be defined as

$$
P = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(\tau_{ik} - \tau_{ik-1})^2}{2\sigma^2_{ik-1}}\right\}
$$

(14)

where $\sigma_{ik-1}$ can be seen as the standard deviation of the noise $n_{ik}$ approximately; and

$$
\tau_{ik-1} = \frac{1}{c} \left\{ |s_{ik-1} - p_{i1}| - |s_{ik-1} - p_{i2}| \right\},
$$

(15)

where $s_{ik-1} = B x_{k-1}$. $x_{k-1} = F x_{k-1}$.

If $P$ is larger than $\beta$, the TDOA $\tau_{ik}$ is true, and can be used for speaker tracking, where $\beta$ is the decision threshold.

The distance difference from the speaker to the ith microphone pair is calculated by

$$
f(s) = ||s - p_{i1}|| - ||s - p_{i2}||.
$$

(16)

To overcome the nonlinearity of the TDOA model in (13) and obtain the measurement equation in (1), (16) need to be linearized. The first order Taylor series expansion of (16) is given by

$$
f(s) = f(s_0) + \nabla f(s_0)^T (s - s_0),
$$

(17)

where $\nabla$ denotes the gradient operator. Let

$$
\begin{align*}
\begin{cases}
    s_{0} = s_{ik-1} \\
    H_{ik} = \nabla f(s_{ik-1})^T B \\
    y_{ik} = c \tau_{ik} - f(s_{ik-1}) + \nabla f(s_{ik-1})^T s_{ik-1},
\end{cases}
\end{align*}
$$

(18)

where $s_{ik-1} = B x_{k-1}$. The measurement equation

$$
y_{ik} = H_{ik} x_k + \nu_{ik}
$$

(19)

is obtained.

Based on the discussions above, the DKF-based speaker tracking method can be summarized as follows.

Let $I$ denote the maximal time index and $N$ denote the number of nodes in the microphone array network, respectively.
Input: $\mathbf{x}_{0:i-1} = \mathbf{x}_0$, $P_{0:i-1} = P_0 > 0$, $k = 0$, $i = 1, \ldots, N$, and speech signals received by microphones at each node in the microphone array network.

Output: $\mathbf{x}_{k:k}$, $k = 1, \ldots, l$, $i = 1, \ldots, N$.

For $k = 1$ to $l$, the following steps are computed:

**Step 1** For $i = 1$ to $N$, each node receives the speech signals, calculates $S_{i,i} (f)$ and estimates the TDOA $\tau_{ik}$ between the two received signals with the GCC method.

$$\tau_{ik} = \arg \max_{\tau \in \{- \tau_{min}, \tau_{max}\}} R_{i,i}(\tau)$$

$$R_{i,i}(\tau) = \int_{-\infty}^{\infty} \phi(f) S_{i,i}(f) e^{j2\pi f \tau} df$$

**Step 2** For $i = 1$ to $N$, for each node, whether or not TDOA $\tau_{ik}$ is true is checked; and when TDOA $\tau_{ik}$ is true, the measurement matrix $H_{ik}$ and the measurement value $y_{ik}$ are calculated.

(a) Check whether $\tau_{ik}$ is true or not.

$$P = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{\left( \tau_{ik} - \tau_{ik-1} \right)^2}{2\sigma_{\tau_{ik}}^2} \right\}$$

If $P > \beta$, the TDOA $\tau_{ik}$ is true, where $\beta$ is the decision threshold.

(b) When the TDOA $\tau_{ik}$ is true, the measurement matrix $H_{ik}$ and the measurement value $y_{ik}$ are calculated by

$$H_{ik} = \nabla f(s_{ik}^*) \delta \theta.$$  

$$y_{ik} = c_{ik}^* - f(s_{ik}^*) + \nabla f(s_{ik}^*) s_{ik-1}.$$  

**Step 3** For $i = 1$ to $N$, speaker tracking is implemented based on the DKF algorithm.

(a) $\mathbf{w}_{ik} \leftarrow \mathbf{w}_{ik-1}$

$P_{ik} \leftarrow P_{ik-1}$

for every neighboring node $l \in N_{ik}$, repeat:

$$R_{ik} = R_{ik} + H_{ik}P_{ik}H_{ik}^T$$

$$\mathbf{y}_{il} \leftarrow y_{il} + P_{ik}H_{ik}^T[y_{il} - R_{ik}\mathbf{y}_{ik}]$$

$$P_{il} \leftarrow P_{il} - P_{ik}H_{ik}^T H_{ik}P_{ik}$$

end

(b) $b_{ik} = b_{ik-1} + \mu [y_{ik-1} - Q_{ik-1}b_{ik-1}]$

(c) $c_{ik} = S b_{ik}$

$$\mathbf{x}_{ik+1:k} \leftarrow \sum_{k=1}^{N_x} c_{ik} \mathbf{y}_{ik}$$

$$P_{ik+1:k} \leftarrow P_{ik} - F P_{ik} F^T + \Gamma S b_{ik}$$

$$F = I + \frac{\Gamma}{\sigma_{\tau_{ik}}}$$

$$\Gamma = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$$

From the proposed method, a new nodes can easily join in the network and any node in the network can leave freely. For each node in the network, whether the TDOA $\tau_{ik}$ is true or not is checked; and the true measurements $y_{ik}$ is gathered from its neighboring nodes to track a moving speaker. Hence, spurious TDOA can be detected and erroneous estimates can be avoid.

**4. Simulations and result discussions**

In this section, we design two trajectory and use simulated audio samples to evaluate the proposed speaker tracking method. The GCF [16], which is a distributed speaker localization method, and the iterative extended Kalman filter (IEKF) based speaker tracking method [11] are implemented to compare with the proposed method. In this paper, a two dimensional tracking problem is considered, where the height of the speech source is set to be the same as the height of all the microphones. It should be stated, however, that the methods can easily be extended to perform 3D tracking.

**4.1. Simulation setup**

Image method [27] is used for simulating room acoustics for a set of reverberation times $T_{60}$ ranging from 50 ms to 250 ms with a step size 50 ms. A loose male speech is used as the source signal that is digitized by 16 bit and 32 kHz sampling frequency.

As shown in Fig. 2, the dimensions of the room is $4 \text{ m} \times 3 \text{ m} \times 2.5 \text{ m}$; the first speaker’s trajectory is a semi-circle from (2.5, 1.5) to (2.5, 2.49); the second speaker's trajectory is along a line from (0.75, 0.5) to (1.5, 0.5), and a line to (1.5, 1.31). The source signal is analyzed in 32 ms frame length. The data received at microphones is obtained by convolving these frames of source signal with the corresponding room impulse responses. Random Gaussian noise is added to each microphone at an SNR level from $-5 \text{ dB}$ to $20 \text{ dB}$ with a step size 5 dB.

The parameter configuration for the DKF-based speaker tracking method is as follows. The state space model is the Langevin model, with the parameters $\beta_0 = 10 \text{s}^{-1}$, $\Delta T = 32 \text{ms}$, and $\beta_1 = 1 \text{m/s}$. The standard deviation of TDOA measurement error is $\sigma_{\tau} = 0.001 \text{s}$. The sound speed is generally $c = 343 \text{m/s}$. $N = 12$ pairs of omni-directional microphones distributed randomly are used to form networks as shown in Fig. 3. The X, Y axis positions of the microphones are shown in Fig. 4. All the microphones are at the constant height of 1.5 m, each pair of which has an inter-microphone spacing of 0.5 m. To check whether or not the estimated TDOAs is true, the decision threshold is generally $\beta = 0.379$. The step size for calculating the diffusion matrix is set as $\mu = 0.01$.

**4.2. Simulation results**

The IEKF-based speaker tracking method in [11] uses a regular microphone array instead of a microphone array network. For comparison, the IEKF-based speaker tracking method is implemented using all the microphone arrays with the geometry structure as the proposed method in a centralized manner. The global coherence field (GCF) [16] technique is a distributed speaker localization method, using the same network as the proposed method. The results obtained from the GCF and IEKF-based methods are compared with the proposed speaker tracking method.
To evaluate the proposed speaker tracking method, the Root Mean Square Error (RMSE) is used, and is defined as

$$\text{RMSE} = \sqrt{\frac{1}{M} \left( \sum_{k=1}^{M} \| \mathbf{s}_k - \hat{\mathbf{s}}_k \|^2 \right)},$$

where $M$ is the number of the frames in the audio sample, $\mathbf{s}_k$ is the true speaker's position, while $\hat{\mathbf{s}}_k$ is the estimated speaker's position at time $k$. The RMSE gives an indication about how much the speaker's location estimate deviates from the true source position. A high RMSE value always reflects an inaccurate tracking result.

Fig. 5 shows the localization results for the GCF method and the tracking results for the DKF-based and IEKF-based methods for reverberation times of $\text{RT}_{60} = 50$ ms and $\text{RT}_{60} = 250$ ms with a background noise level of SNR = 20 dB. In Fig. 5, the DKF-based and GCF methods are implemented in the second network as shown in Fig. 3. It can be observed from Fig. 5 that owing to the fact that the spurious TDOAs can be detected the proposed method can suppress greatly the effect of anomalous measurements and obtain high tracking performance. Because of not considering the fact that whether or not the TDOAs is true, the GCF and IEKF-based methods are not robust against room reverberations. Only depending on the current measurement, the GCF method even fails to localize a
speaker under a large room reverberation time and generates spurious sources, like position point P.

As a function of the RT60 value with a background noise level of SNR = 20 dB, the RMSE for the proposed speaker tracking method is obtained by comparing the true speaker positions and is shown in Fig. 6. As a function of the SNR value with reverberation times of RT60 = 50 ms and RT60 = 250 ms, respectively, the RMSE is shown in Fig. 7. The RMSEs plotted in Figs. 6 and 7 correspond to the average RMS errors, resulting from 100 independent experiments.

As shown in Fig. 6, since the false TDOA estimates are picked out, the tracking performance of the proposed method degrades moderately as reverberation time is increased. It can be observed from Figs. 6 and 7 that the proposed speaker tracking method performs better in densely linked network than in sparsely linked network.

As a function of the SNR value with reverberation times of RT60 = 50 ms and RT60 = 250 ms, respectively, the RMSEs for DKF-based, the IEKF-based and GCF methods are obtained by comparing the true speaker’s positions as shown in Fig. 8. As a function of the RT60 value with a background noise level of SNR = 20 dB, the RMSEs are obtained as shown in Fig. 9. In Figs. 8 and 9, the DKF-based and GCF methods are implemented in the second network as shown in Fig. 3. The RMSEs plotted in Figs. 8 and 9 correspond to the average RMS errors, resulting from 100 independent experiments.

As shown in Figs. 8 and 9, the presence of outliers in the TDOAs estimated by the GCC method results in deterioration of the overall localization and tracking accuracy of the GCF and IEKF-based methods. Picking out obvious error measurements and avoiding the error position estimates, the proposed speaker tracking method gets a distinct improvement in tracking performance compared with the IEKF-based method. And taking into account not only the current measurement but also a series of past
measurements, the proposed speaker tracking method avoids the spurious speaker sources generated under heavy noisy and reverberant environments, and is more accurate than the GCF speaker localization method.

5. Conclusions

In this paper, a distributed Kalman filter based speaker tracking method using a microphone array network is proposed. Results obtained from the GCF and IEKF-based methods have also been investigated and used as reference for an overall comparison of proposed method's tracking ability. Based on a series of past measurements rather than the current observation only, the proposed method can obtain robustly a smooth trajectory of the speaker’s movement. Since for each node in the network, whether or not these TDOAs are true is checked, and the true measurements are gathered from its neighboring nodes, the proposed speaker tracking method can get a distinct improvement in tracking performance compared with the GCF and IEKF-based methods. Simulation results demonstrate that the proposed method is more robust against room reverberation and background noise.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Nos. 61172107, 61172110, 60772161), Dalian Municipal Science and Technology Fund Scheme of China (No. 2008J23H025) and Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 200801410015) and the Fundamental Research Funds for the Central Universities of China (No. DUT13LAB06).

References