A Bayesian Performance Bound for Time-Delay of Arrival based Acoustic Source Tracking in a Reverberant Environment

Xionghu Zhong†, Wenwu Wang°, Mohsen Naqvi⋆, Eng Siong Chng†
†School of Computer Engineering, Nanyang Technological University, Singapore, 639798.
°CVSSP, Department of Electronic Engineering, University of Surrey, UK, GU2 7XH.
⋆School of Electronic, Electrical and Systems Engineering, Loughborough University, UK, LEH 3TU.
†{xhzhong and aseschng}@ntu.edu.sg, °w.wang@surrey.ac.uk, ⋆S.M.R.Naqvi@lboro.ac.uk

***Abstract*—Acoustic source tracking in a room environment based on a number of distributed microphone pairs has been widely studied in the past. Based on the received microphone pair signals, the time-delay of arrival (TDOA) measurement is easily accessible. Bayesian tracking approaches such as extended Kalman filter (EKF) and particle filtering (PF) are subsequently applied to estimate the source position. In this paper, the Bayesian performance bound, namely posterior Cramér-Rao bound (PCRB) is derived for such a tracking scheme. Since the position estimation is indirectly related to the received signal, a two-stage approach is developed to formulate the Fisher information matrix (FIM). First, the Cramér-Rao bound (CRB) of the TDOA measurement in the noisy and reverberant environment is calculated. The CRB is then regarded as the variance of the TDOAs in the measurement function to obtain the PCRB. Also, two different TDOA measurement models are considered: 1) single TDOA corresponding to the largest peak of the generalized cross-correlation (GCC) function; and 2) multiple TDOAs from several peaks in GCC function. The later measurement model implies a higher probability of detection and heavier false alarms. The PCRB for both measurement models are derived. Simulations under different noisy and reverberant environments are organized to validate the proposed PCRB.**

I. Introduction

Estimating the position of an acoustic source in a room environment plays an important role in many speech and audio applications such as speaker diarization, hearing aids, hands-free distant speech recognition and communication, and teleconferencing systems [1]–[4]. Once the position of the source is known, it can be fed into a higher processing stage for speech acquisition of a focussed region, enhancement of a specific speech signal in the presence of competing talkers, or keeping a camera focused on the talker in a video-conferencing scenario. However, it is a challenge to provide an accurate position estimation since the received audio signal can be significantly distorted and its statistical properties can be 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.

To perform the position estimation, a number of distributed microphone pairs/arrays are usually deployed in the room environment. The time-delay of arrival (TDOA) is usually employed as the measurement due to its accessibility and robustness under noisy and reverberant conditions [3], [5]–[11]. The TDOA measurement is extracted from the microphone pair signals by using, for example, generalized cross-correlation (GCC) method [17]. Since each TDOA yields a half hyperboloid of two sheets, the TDOA measurements from distributed microphone pairs/arrays can be used to triangulate a target position. Such a triangulation is traditionally approximated by using a localization approach such as linear intersection (LI) algorithm [5]. For source that is dynamic, a state space model can be used to model the trajectory of the moving source and the uncertainties in the TDOA measurements. Correspondingly, different Bayesian approaches such as extended Kalamn filter (EKF) [7], [9] and particle filtering (PF) [8], [12], [13] have been developed to track the source. In the EKF implementation, the measurement at each microphone pair consists of a single TDOA which maximizes the GCC function, and its performance can be seriously degraded due to inaccurate TDOA estimation. The PF approach incorporates multiple peaks of the GCC function to increase the probability of detection. A bi-hypothesis model is defined for each TDOA to evaluate its likelihood of being a false alarm or a real detection. Generally, the PF approach is more robust than the LI and EKF approaches in adverse environments [16].

In a room environment, the received signal to reverberation ratio (SRR) is quadratically proportional to the inverse of the distance between the source and the microphone [14]. This results in an inhomogenous signal quality at different microphone receiver and hence the TDOA estimation are affected differently. When the source is close to the microphone pair, better TDOA measurement can be expected, and vice versa. Hence, the microphone deployment has a significant impact on the performance of tracking algorithms. In this paper, we attempt to derive a lower bound to characterize the potential tracking performance in the room environment. The posterior Cramér-Rao bound (PCRB) [15] that is developed for nonlinear state space model and dynamic source tracking problem is introduced. Since the received signal is indirectly related to the source position, a two-stage scheme is developed to formulate the Fisher information matrix (FIM) [15]. First, the traditional Cramér-Rao bound (CRB) of the TDOA estimation is obtained according to the derivations in [14]. The CRB is then regarded as a variance of the TDOA measurement in the measurement function. The PCRB is derived based on the spatial information (from the source dynamic model) and the temporal information (from the measurement model). For each microphone pair, two different TDOA measurement models are also considered: *i)* single TDOA corresponding to the largest peak of the GCC function; and *ii)* multiple TDOAs from several peaks in the GCC function. The former is overly optimistic since it assumes perfect TDOA detection, i.e., the probability of detection equals one. Multiple TDOA measurement model takes into account miss detection and false alarms. It is observed from our simulations that the PCRB based on the multiple TDOA measurement model is more practical and achievable. The developed PCRB has a broad range of applications including microphone selection and deployment in a room environment.

This work is the first attempt to derive the PCRB for the TDOA measurement based acoustic source tracking in the room environment. The rest of this paper is organized as follows. In Section II, the signal model and GCC method based TDOA estimation are introduced. Section III presents different measurement models and the tracking algorithm developed based on these models. The detailed derivation of the PCRB is given in Section IV. Simulations are organized in Section V and conclusions are drawn in Section VI.

II. Signal model and TDOA measurements

This section provides the signal model of distributed microphone pairs in a room environment. The GCC method based TDOA estimation is also addressed.

A. Microphone Signal Model

Assume that a distributed microphone system consisting of L microphone pairs is deployed in a room environment. Let $\boldsymbol{p}_{\ell,i} \in \mathbb{R}^3$, $i \in \{1, 2\}$ denote the position of the ith microphone of the ℓth ($\ell = 1, \ldots, L$) microphone pair. Also, let $\mathbf{x}_t \in \mathbb{R}^3$ denote the position of the source signal at discrete time t. The discrete time signal received from a single source can be modeled as

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

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 independent between different channels, and $\star$ denotes convolution. To formulate TDOA estimates, the impulse response can be rewritten in terms of direct path and multipath components as

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

where $r_{\ell,i}(t) = \|\mathbf{x}_t - \boldsymbol{p}_{\ell,i}\|$ is the Euclidean distance between the source and the microphone, and $\tau_{\ell,i}(t) = r_{\ell,i}(t)/c$ is the direct path time delay with c denoting the speed of sound, and $g(\boldsymbol{p}_{\ell,i}, \mathbf{x}_t)$ is the reverberant part of the 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.

Although speech signal is non-stationary and its statistics change over time, it is usually assumed to be short-time quasi-stationary. Hence, the signal received at each microphone can be processed in short 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 ith microphone of the ℓth pair can then be written as $\mathbf{s}(k)$ and $\mathbf{y}_{\ell,i}(k)$, given as

$$\mathbf{s}(k) = [s(kT_0), s(kT_0 + 1), \ldots, s((k+1)T_0 - 1)], \tag{3}$$

$$\mathbf{y}_{\ell,i}(k) = [y_{\ell,i}(kT_0), \ldots, y_{\ell,i}((k+1)T_0 - 1)]. \tag{4}$$

Correspondingly, the additive noise is written as $\mathbf{n}_{\ell,i}(k)$. Further, it is assumed that in each frame the position of the source is spatially stationary. The parameters characterizing the source motion and the source signal are fixed in the kth frame, e.g., the source position, $\mathbf{x}_k$, and the corresponding room impulse response (RIR), $h(\boldsymbol{p}_{\ell,i}, \mathbf{x}_k)$.

The additive noise $\mathbf{n}_{\ell,i}(k)$ is assumed to be zero-mean, white, and Gaussian, i.e., $\mathbf{n}_{\ell,i}(k) \sim \mathcal{N}(0, \sigma^2)$. The signal to noise ratio (SNR) is thus defined as

$$\text{SNR}(k, \omega) = \frac{P_{ss}(k, \omega)}{\sigma^2}, \tag{5}$$

where $P_{ss}(k, \omega)$ is the power spectrum of the source signal. The reverberation time T_{60} is usually used to evaluate the reverberation in the room environment. However, for tracking problem, the source is dynamic and the distance between the source and the microphones can change drastically. When the source is closer to the microphones and the wall reflection coefficient is smaller, the direct path component is relatively stronger. The effect of distance and reflection coefficient can be summarized by one parameter: the signal to reverberation ratio (SRR). Given the reflection coefficient ρ and the distance r_k^ℓ between the source and the centroid of the ℓth microphone pair, the SRR can be defined as [14]

$$\text{SRR}^\ell(k, \omega) = \frac{\mathcal{A}(1 - \rho^2)}{16\pi (r_k^\ell)^2 \rho^2}, \tag{6}$$

where $\mathcal{A}$ is the whole wall reflection area. Detailed derivation of equation (6) can be found in [16]. The expression (6) indicates that the distortionness of the received signal is determined by the room dimension, reflection coefficients and the distance between the source and the microphones. It shows that in the same room environment, SRR increases quadratically with the decreasing of the distance between the source and the sensor. Better position estimation performance is thus likely to be obtained when the source is located at the close-end of the microphones, and vice versa.

B. TDOA Measurements

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 $\mathbf{y}_{\ell,1}(k)$ and $\mathbf{y}_{\ell,2}(k)$ collected at the ℓth microphone pair at the time step k, the GCC function can be approximated as [17]:

$$R_\ell(k,\tau) = \int_\Omega \Phi_\ell(k,\omega) Y_{\ell,1}(k,\omega) Y^*_{\ell,2}(k,\omega) e^{j\omega\tau} d\omega, \qquad (7)$$

where $\mathbf{y}_{\ell,i}(k) \rightleftharpoons Y_{\ell,i}(k,\omega)$ are discrete Fourier transform (DFT) pairs, and $\Phi_\ell(k,\omega) = |Y_{\ell,1}(k,\omega)Y^*_{\ell,2}(k,\omega)|^{-1}$ is the PHAT weighting term, and Ω is the frequency range over which the integration is carried out. The TDOA measurement at the ℓth microphone pair can thus be estimated by exploring the potential TDOA τ that maximizes the GCC function

$$\hat{\tau}_k^\ell = \arg \max_{\tau \in [-\tau_{\max}, \tau_{\max}]} R_\ell(k,\tau), \qquad (8)$$

where $\tau_{\max} = \|\boldsymbol{p}_{\ell,1} - \boldsymbol{p}_{\ell,2}\|/c$ is the possible maximum delay. This TDOA measurement could, for example, be directly used in (12) for 3-D position estimation.

Assume that: 1) the dimension of the room is large relative to the wavelength of $s(t)$. This is satisfied since for tracking problem, the frequency band of interest is usually $\Omega = 2\pi \times [300, 3500]$Hz; and 2) the source and the microphones are located in the interior of the room, at least half-wavelength away from the walls, and microphone separation $d \ll \min(r_{\ell,1}(t), r_{\ell,2}(t))$ so that the two microphones receive an equal amount of energy from the direct path. The Cramér-Rao bound (CRB) σ_k^ℓ for the TDOA estimates is given as [14]

$$\sigma_k^\ell = \left(2 \sum_{\omega \in \Omega} \frac{\eta_k^\ell(k,\omega)^2}{1 + 2\eta_k^\ell(k,\omega)} \omega_k^2 \right)^{-1}, \qquad (9)$$

where η_k^ℓ is the signal to noise and reverberation ratio (SNRR) defined as

$$\eta_k^\ell(k,\omega) = \frac{P_{ss}(k,\omega)\frac{1}{4\pi(r_k^\ell)^2}}{P_{ss}(k,\omega)\frac{4\rho^2}{\mathcal{A}(1-\rho^2)} + \sigma^2}. \qquad (10)$$

The detailed study of the CRB (9) under different noise and reverberant environments is given in [14]. In this work, the CRB (9) will be employed as a theoretical variance of the estimated TDOA in the measurement function.

The actual TDOA of a microphone pair is expressed in terms of the source and sensor geometry by

$$\tau_k^\ell(\mathbf{x}_k) = \tau_{\ell,1}(k) - \tau_{\ell,2}(k) = \frac{\|\mathbf{x}_k - \boldsymbol{p}_{\ell,1}\| - \|\mathbf{x}_k - \boldsymbol{p}_{\ell,2}\|}{c}. \qquad (11)$$

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

$$\hat{\mathbf{x}}_k = \arg \min_{\mathbf{x}_k} \sum_{\ell=1}^{L} \left(\hat{\tau}_k^\ell - \tau_k^\ell(\mathbf{x}_k) \right)^2. \qquad (12)$$

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. Tracking algorithm review

This section presents a brief review of the Bayesian tracking approaches for room acoustic source tracking. Different measurement models are also introduced.

A. Bayesian Tracking Framework

To formulate a Bayesian framework, the source dynamic model is defined first. In the $x-y$ plan, the source dynamics can be assumed to follow the Langevin motion model [8], [12]. Since the height of a talker is often fixed during a conversation, it is reasonable to use a random walk model to describe the height uncertainties in $z-$direction. The complete motion model in 3-D space can be addressed as

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{Q}\mathbf{v}_k, \qquad (13)$$

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

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}_2 & a\Delta T \mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & a\mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 \end{bmatrix}; \quad \mathbf{Q} = \begin{bmatrix} b\Delta T \mathbf{I}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & b\mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 \end{bmatrix}, \qquad (14)$$

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 β are the velocity parameter and the rate constant respectively. Equation (13) 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], [12], [18] are found to be adequate for room acoustic source tracking and are employed here.

Assume that the TDOA measurement set at the time step k is $\mathcal{Z}_k$. The solution based on Bayesian recursive estimation can be given as

- Predict:

$$p(\mathbf{x}_k|\mathcal{Z}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{x}_{k-1}|\mathcal{Z}_{1:k-1}) d\mathbf{x}_{k-1}; \qquad (15)$$

- Update:

$$p(\mathbf{x}_k|\mathcal{Z}_{1:k}) \propto p(\mathcal{Z}_k|\mathbf{x}_k) p(\mathbf{x}_k|\mathcal{Z}_{1:k-1}). \qquad (16)$$

where $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ is the transition probability density that can be obtained according to (13), and $p(\mathcal{Z}_k|\mathbf{x}_k)$ is the likelihood. There is no closed form solution to the recursion (15) and (16) since the TDOA measurement function is nonlinear. The extended Kalman filter (EKF) and PF approaches are widely employed to approximate this recursion [8], [12], [19].

B. EKF Implementation

Assume that L TDOAs are extracted from the GCC functions. The measurement vector can be written as

$$\mathbf{z}_k = [\hat{\tau}_k^1, \hat{\tau}_k^2, \ldots, \hat{\tau}_k^L]^T. \qquad (17)$$

Note that here we only use the largest peak in the GCC function at each microphone pair to obtain the TDOA measurement. Since the measurement function (11) is nonlinear, the EKF can be employed to approximate the posterior distribution [9]. The first-order Taylor expansion on $\tau_k^\ell(\mathbf{x}_k)$ from (11) is

$$\tau_k^\ell(\mathbf{x}_k) = \tau_k^\ell(\mathbf{x}_{k-1}) + \mathbf{C}_k^\ell \left[\mathbf{x}_k - \mathbf{x}_{k-1}\right]^T + \bar{n}_k, \tag{18}$$

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

$$\mathbf{C}_k^\ell = \frac{1}{c}\left[\frac{\mathbf{x}_k - \boldsymbol{p}_{\ell,1}}{\|\mathbf{x}_k - \boldsymbol{p}_{\ell,1}\|} - \frac{\mathbf{x}_k - \boldsymbol{p}_{\ell,2}}{\|\mathbf{x}_k - \boldsymbol{p}_{\ell,2}\|}\right]\Bigg|_{\mathbf{x}_k = \mathbf{x}_{k-1}}. \tag{19}$$

Define

$$\bar{\tau}_k^\ell = \hat{\tau}_k^\ell - \tau_k^\ell(\hat{\mathbf{x}}_{k-1}) + \mathbf{C}_k^\ell \hat{\mathbf{x}}_{k-1}, \tag{20}$$

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 (11). The nonlinear measurement is thus approximated by

$$\bar{\tau}_k^\ell \approx \mathbf{C}_k^\ell \mathbf{x}_k + \bar{n}_k. \tag{21}$$

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.

Regarding (13) as the state dynamic process, the implementation of an EKF can be written as [20]

$$\bar{\mathbf{x}}_{k|k-1} = \mathbf{A}\widehat{\mathbf{x}}_{k-1}; \tag{22a}$$
$$\overline{\mathbf{P}}_{k|k-1} = \mathbf{A}\widehat{\mathbf{P}}_{k-1}\mathbf{A}^T + \mathbf{Q}\boldsymbol{\Sigma}_k\mathbf{Q}^T; \tag{22b}$$
$$\mathbf{S}_k = \mathbf{R}_k + \mathbf{C}_k\overline{\mathbf{P}}_{k|k-1}\left(\mathbf{C}_k\right)^T; \tag{22c}$$
$$\mathbf{K}_k = \overline{\mathbf{P}}_{k|k-1}(\mathbf{C}_k)^T(\mathbf{S}_k)^{-1}; \tag{22d}$$
$$\widetilde{\mathbf{x}}_k = \bar{\mathbf{x}}_{k|k-1} + \mathbf{K}_k(\mathbf{z}_k - \boldsymbol{\tau}_k(\bar{\mathbf{x}}_{k|k-1})); \tag{22e}$$
$$\widetilde{\mathbf{P}}_k = \overline{\mathbf{P}}_{k|k-1} - \mathbf{K}_k\mathbf{C}_k\overline{\mathbf{P}}_{k|k-1}. \tag{22f}$$

where $\mathbf{C}_k = [\mathbf{C}_k^1, \ldots, \mathbf{C}_k^L]^T$ and $\boldsymbol{\tau}_k(\bar{\mathbf{x}}_{k|k-1}) = [\tau_k^1(\bar{\mathbf{x}}_{k|k-1}), \ldots, \tau_k^L(\bar{\mathbf{x}}_{k|k-1})]^T$ are obtained according to (19) and (11) respectively. After the EKF steps, the posterior distribution is given by $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathcal{Z}_k) = \mathcal{N}(\mathbf{x}_k; \widetilde{\mathbf{x}}_k, \widetilde{\mathbf{P}}_k)$ where $\widetilde{\mathbf{x}}_k$ and $\widetilde{\mathbf{P}}_k$ are the estimated state vector and the corresponding covariance matrix respectively. In next section, a PF approach which does not need to use the first-order approximation of the measurement function will be introduced.

C. PF Tracking Algorithm

The PF approach employs a number of particles to approximate the posterior distribution. 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 $p(\mathbf{x}_{k-1}|\mathcal{Z}_{1:k-1})$ at time step $k-1$. The particles are drawn according to the source dynamic model (13) and their importance to the estimation is evaluated by the likelihood. The state transition density is given as

$$p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}) = \mathcal{N}(\mathbf{x}_k^{(i)} \mid \mathbf{A}\mathbf{x}_{k-1}^{(i)}, \mathbf{Q}\boldsymbol{\Sigma}_k\mathbf{Q}^T), \tag{23}$$

where $\mathbf{A}$ and $\mathbf{Q}$ are defined in (14). To increase the probability of detection, a number of GCC function peaks are employed to obtain the TDOAs. Assume that n_k^ℓ TDOA measurements are available at the ℓth microphone pair, i.e., $\mathcal{Z}_k^\ell = \{\hat{\tau}_{1,k}^\ell, \ldots, \hat{\tau}_{n_k^\ell,k}^\ell\}$. The complete measurement set can be written as

$$\mathcal{Z}_k = \{\mathcal{Z}_k^1, \mathcal{Z}_k^2, \ldots, \mathcal{Z}_k^L\}. \tag{24}$$

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 [12], 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

$$\begin{aligned} &\mathcal{H}_{0,k}^\ell \triangleq \{\lambda_{q,k} = 0; q = 1, \ldots, n_k^\ell\}; \\ &\mathcal{H}_{q,k}^\ell \triangleq \{\lambda_{q,k} = 1,\, \lambda_{p,k} = 0; q \neq p = 1, \ldots, n_k^\ell\}, \end{aligned} \tag{25}$$

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 qth 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

$$p(\hat{\tau}_{q,k}^\ell|\mathbf{x}_k^{(i)}, \lambda_{q,k} = 0) = \mathcal{U}_\tau(\hat{\tau}_{q,k}^\ell) = \frac{1}{2\tau_{\max}}, \tag{26}$$

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 σ_τ^2 [12], given by

$$p(\hat{\tau}_{q,k}^\ell|\mathbf{x}_k^{(i)}, \lambda_{q,k} = 1) = \mathcal{N}(\hat{\tau}_{q,k}^\ell \mid \tau_k^\ell(\mathbf{x}_k^{(i)}), \sigma_\tau^2). \tag{27}$$

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], [12], 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 $p(\mathcal{H}_{0,k}^\ell|\mathbf{x}_k^{(i)}) = q_0$. The prior probability $p(\mathcal{H}_{q,k}^\ell|\mathbf{x}_k^{(i)})$ is equally weighted, so

$$p(\mathcal{H}_{q,k}^\ell|\mathbf{x}_k^{(i)}) = \frac{1-q_0}{n_k^\ell}; \qquad \text{for } q \in \{1, \ldots, n_k^\ell\}. \tag{28}$$

The complete likelihood over all hypotheses from the ℓth microphone pair is obtained by summing all hypotheses

$$\begin{aligned} p(\mathcal{Z}_k^\ell|\mathbf{x}_k^{(i)}) &= \sum_{q=0}^{n_k^\ell} p(\mathcal{H}_{q,k}^\ell|\mathbf{x}_k^{(i)})p(\mathcal{Z}_k^\ell|\mathbf{x}_k^{(i)}, \mathcal{H}_{q,k}^\ell) \\ &= \frac{\frac{q_0}{2\tau_{\max}} + \frac{1-q_0}{n_k^\ell}\sum_{q=1}^{n_k^\ell}\mathcal{N}(\hat{\tau}_{q,k}^\ell \mid \tau_k^\ell(\mathbf{x}_k^{(i)}), \sigma_\tau^2)}{(2\tau_{\max})^{n_k^\ell - 1}}. \end{aligned} \tag{29}$$

The particles are then weighted according to

$$w_k^{(i)} = \tilde{w}_{k-1}^{(i)} \prod_{\ell=1}^{L} p(\mathcal{Z}_k^\ell|\mathbf{x}_k^{(i)}), \tag{30}$$

where $\tilde{w}_{k-1}^{(i)}$ is the normalized weight. After resampling, the posterior distribution of the state is approximated as

$$p(\mathbf{x}_k|\mathcal{Z}_{1:k}) \approx \sum_{i=1}^{N} \tilde{w}_k^{(i)} \delta_{\mathbf{x}_k^{(i)}}(\mathbf{x}_k), \tag{31}$$

where $\delta(\cdot)$ is a Dirac-delta function with unity value if $\mathbf{x}_k = \mathbf{x}_k^{(i)}$ and 0 otherwise, and N is the number of the particles.

IV. Potential Tracking Performance Bound

The PCRB provides a lower performance bound on the mean square error (MSE) matrix for sequential Bayesian estimation of random parameters. In this section, we present the derivation of PCRB for the TDOA measurement based 3-D position estimation in the noisy and reverberant environment. Since the source position is estimated from the TDOA measurements rather than the received signal directly, a two stage approach is introduced to obtain the tracking bound.

A. Posterior Cramér-Rao Bound

Let $\widehat{\mathbf{x}}_k$ denote an unbiased estimator of the state vector $\mathbf{x}_k$ and $\mathbf{Y}_k^\ell = (\mathbf{y}_{\ell,1}(k), \mathbf{y}_{\ell,2}(k))$ be the signal received at the ℓth microphone pair. Let $\mathbf{Y}_k = [\mathbf{Y}_k^1, \dots, \mathbf{Y}_k^L]$ denote all received signal at the time step k and $\mathbf{Y}_{1:k} = [\mathbf{Y}_1, \dots, \mathbf{Y}_k]$ represent the complete data set from time step 1 to k. The PCRB (lower bound) on the estimation error has the form [15]

$$\mathbb{E}\left\{(\widehat{\mathbf{x}}_k - \mathbf{x}_k)(\widehat{\mathbf{x}}_k - \mathbf{x}_k)^T\right\} \geq \mathbf{J}_{\mathbf{x}_k,\mathbf{x}_k}^{-1}. \tag{32}$$

The Fisher information matrix (FIM) $\mathbf{J}_{\mathbf{x}_k,\mathbf{x}_k}$ is given by

$$\mathbf{J}_{\mathbf{x}_k,\mathbf{x}_k} = \mathbb{E}\{-\mathbf{\Delta}_{\mathbf{x}_k}^{\mathbf{x}_k} \ln p(\mathbf{x}_k|\mathbf{Y}_{1:k})\}, \tag{33}$$

where $\Delta_{\mathbf{x}_k}^{\mathbf{x}_k}$ is the second order partial derivative. The notations of gradient and second order derivative are given by

$$\nabla_{\mathbf{x}_k} = \frac{\partial}{\partial \mathbf{x}_k} = \left[\frac{\partial}{\partial \mathbf{x}_k^1}, \dots, \frac{\partial}{\partial \mathbf{x}_k^M}\right]^T; \tag{34}$$

$$\Delta_{\mathbf{x}_k}^{\mathbf{x}_k} = \nabla_{\mathbf{x}_k} \nabla_{\mathbf{x}_k}^T. \tag{35}$$

where $\mathbf{x}_k^m$ denotes the mth element in vector $\mathbf{x}_k$. Here we assume that the second order derivatives and expectations in (33) exist. Given the posterior distribution $p(\mathbf{x}_k|\mathbf{Y}_{1:k})$, the FIM $\mathbf{J}_{\mathbf{x}_k,\mathbf{x}_k}$ can be calculated recursively as [15]

$$\mathbf{J}_{\mathbf{x}_k,\mathbf{x}_k} = \mathbf{D}_k^{22} - \mathbf{D}_k^{21}(\mathbf{J}_{\mathbf{x}_{k-1},\mathbf{x}_{k-1}} + \mathbf{D}_k^{11})^{-1}\mathbf{D}_k^{12}, \tag{36}$$

where $\mathbf{D}_k^{11}$, $\mathbf{D}_k^{12}$, and $\mathbf{D}_k^{21}$ are given by

$$\mathbf{D}_k^{11} = \mathbb{E}\{-\mathbf{\Delta}_{\mathbf{x}_{k-1}}^{\mathbf{x}_{k-1}} \ln p(\mathbf{x}_k|\mathbf{x}_{k-1})\} = \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}; \tag{37}$$

$$\mathbf{D}_k^{12} = \mathbb{E}\{-\mathbf{\Delta}_{\mathbf{x}_{k-1}}^{\mathbf{x}_k} \ln p(\mathbf{x}_k|\mathbf{x}_{k-1})\} = -\mathbf{A}^T\mathbf{Q}^{-1}; \tag{38}$$

$$\mathbf{D}_k^{21} = \mathbb{E}\{-\mathbf{\Delta}_{\mathbf{x}_k}^{\mathbf{x}_{k-1}} \ln p(\mathbf{x}_k|\mathbf{x}_{k-1})\} = \{\mathbf{D}_k^{12}\}^T; \tag{39}$$

$$\begin{aligned}\mathbf{D}_k^{22} &= \mathbb{E}\{-\mathbf{\Delta}_{\mathbf{x}_k}^{\mathbf{x}_k} \ln p(\mathbf{x}_k|\mathbf{x}_{k-1})\} + \mathbb{E}\{-\mathbf{\Delta}_{\mathbf{x}_k}^{\mathbf{x}_k} \ln p(\mathbf{Y}_k|\mathbf{x}_k)\} \\ &= \mathbf{Q}^{-1} + \mathbb{E}\{-\mathbf{\Delta}_{\mathbf{x}_k}^{\mathbf{x}_k} \ln p(\mathbf{Y}_k|\mathbf{x}_k)\}.\end{aligned} \tag{40}$$

Define

$$\widetilde{\mathbf{D}}_k^{22} = \mathbb{E}\{-\mathbf{\Delta}_{\mathbf{x}_k}^{\mathbf{x}_k} \ln p(\mathbf{Y}_k|\mathbf{x}_k)\}. \tag{41}$$

The task here is how to evaluate $\widetilde{\mathbf{D}}_k^{22}$. The state $\mathbf{x}_k$ is indirectly related to the received signal $\mathbf{Y}_k$. That is, the TDOAs are estimated from the signal set $\mathbf{Y}_k$ first. These TDOAs are then regarded as the measurement to track the source. Hence, the accuracy of position estimation is highly dependent on the accuracy of the TDOA measurements. Due to such an indirect relationship, the potential performance bound derived here is based on a two-stage calculation.

Consider only one TDOA at each microphone pair. The measurement model at the ℓth pair can be addressed as

$$\mathbf{z}_k = \boldsymbol{\tau}_k(\mathbf{x}_k) + \mathbf{w}_k, \tag{42}$$

where $\mathbf{w}_k$ is a zero-mean Gaussian process which models the uncertainties in the TDOA measurements. The CBR (9) provides a theoretical lower bound for the variance of the TDOA measurements. Here, we use the CRB σ_k^ℓ as the variance for each TDOA measurement $\hat{\tau}_k^\ell$ in (42), i.e., $\hat{\tau}_k^\ell \sim \mathcal{N}(\tau_k^\ell(\mathbf{x}_k), \sigma_k^\ell)$. The covariance matrix for $\mathbf{w}_k$ can thus be expressed as

$$\mathbf{R}_k = \mathrm{diag}(\sigma_k^1, \sigma_k^2, \dots, \sigma_k^L), \tag{43}$$

where diag(Ξ) denotes a diagonal matrix with diagonal elements Ξ and off diagonal elements zero.

To calculate $\widetilde{\mathbf{D}}_k^{22}$, we need to take the second derivative of the likelihood $p(\mathbf{z}_k|\boldsymbol{\tau}_k(\mathbf{x}_k))$ modeled by the zero-mean Gaussian process. The FIM can be formulated by following the standard derivation of Bangs' formula [21], written as

$$\begin{aligned}\widetilde{\mathbf{D}}_k^{22}(i,j) = &\left(\frac{\partial \boldsymbol{\tau}_k(\mathbf{x}_k)}{\partial \mathbf{x}_k^i}\right)^T \mathbf{R}_k^{-1} \left(\frac{\partial \boldsymbol{\tau}_k(\mathbf{x}_k)}{\partial \mathbf{x}_k^j}\right) \\ &+ \frac{1}{2}\mathrm{tr}\left[\mathbf{R}_k^{-1}\left(\frac{\partial \mathbf{R}_k}{\partial \mathbf{x}_k^i}\right)^T \mathbf{R}_k^{-1}\frac{\partial \mathbf{R}_k}{\partial \mathbf{x}_k^j}\right],\end{aligned} \tag{44}$$

where $\frac{\partial(\cdot)}{\partial \mathbf{x}_k^i}$ denotes the partial derivative with respect to the ith element of $\mathbf{x}_k$ and $\frac{\partial \boldsymbol{\tau}_k(\mathbf{x}_k)}{\partial \mathbf{x}_k^i}$ can be obtained from (19). The partial derivative of each element in $\mathbf{R}_k$ with respect to $\mathbf{x}_k^i$ can be obtained as

$$\frac{\partial \sigma_k^\ell}{\partial \mathbf{x}_k^i} = 8\left(\sigma_k^\ell\right)^2 \sum_{\omega_k \in \Omega} \frac{(\eta_k^\ell)^2(1+\eta_k^\ell)(\mathbf{x}_k^i - \boldsymbol{p}_\ell^i)}{(1+2\eta_k^\ell)^2(r_k^\ell)^2}\omega_k^2. \tag{45}$$

Substituting (44) into (40) and then all $\mathbf{D}_k^{j_1 j_2}$ ($j_1, j_2 \in \{1,2\}$) into (36), the PCRB can be calculated in a straightforward manner. Note that by inverting $\widetilde{\mathbf{D}}_k^\ell$, the CRB of $\mathbf{x}_k$ that considers only the temporal information from current measurements can be obtained.

B. Multiple Hypothesis Model

The single TDOA measurement model (42) is overly optimistic since it regards that the TDOAs from all microphone pairs are perfectly detected, i.e., the probability of detection is one. However, false alarms can appear and also miss detection can happen due to the noise and reverberation. To increase the probability of detection, multiple TDOAs from a microphone pair are usually collected to produce a measurement set, as addressed by the measurement model (24). In multiple TDOA measurement model, each TDOA is either a real detection generated by source signal or a false alarm due to noise and reverberation. Hence, two hypotheses are defined in (26) and (27) respectively. Following the two-stage calculation scheme in Section IV.A, each TDOA measurement is assumed to be following a zero-mean Gaussian distribution with variance σ_k^ℓ. The likelihood for the ℓth microphone pair is given in (29) by ignoring the particle index, rewritten here as

$$p(\mathcal{Z}_k^\ell|\mathbf{x}_k) = \frac{\frac{q_0}{2\tau_{\max}} + \frac{1-q_0}{n_k^\ell}\sum_{q=1}^{n_k^\ell}\mathcal{N}(\hat{\tau}_{q,k}^\ell \,|\, \tau_k^\ell(\mathbf{x}_k), \sigma_k^\ell)}{(2\tau_{\max})^{n_k^\ell - 1}}. \quad (46)$$

The formulation of the PCRB remains the same as that in single TDOA model scenario except the calculation of $\widetilde{\mathbf{D}}_k^{22}$ in (40). In multiple TDOA measurement model, $\widetilde{\mathbf{D}}_k^{22}$ is the second derivative of the logarithm of the total likelihood $p(\mathcal{Z}_k|\mathbf{x}_k) = \prod_{\ell=1}^L p(\mathcal{Z}_k^\ell|\mathbf{x}_k)$, which can be given as

$$\widetilde{\mathbf{D}}_k^{22} = \mathbb{E}\left\{-\mathbf{\Delta}_{\mathbf{x}_k}^{\mathbf{x}_k}\left(\sum_{\ell=1}^{L}\ln p(\mathcal{Z}_k^\ell|\mathbf{x}_k)\right)\right\}. \quad (47)$$

Define $c_1 = q_0/(2\tau_{\max})^{n_k^\ell}$, $c_2 = (1-q_0)/(n_k^\ell(2\tau_{\max})^{n_k^\ell})$ and

$$\Lambda_k^\ell \triangleq \sum_{q=1}^{n_k^\ell}\mathcal{N}(\hat{\tau}_{q,k}^\ell \,|\, \tau_k^\ell(\mathbf{x}_k), \sigma_k^\ell). \quad (48)$$

We have $p(\mathcal{Z}_k^\ell|\mathbf{x}_k) = c_1 + c_2\Lambda_k^\ell$. The second order derivative of the ℓth item in the summation of (47) can be calculated as

$$\mathbb{E}\left\{-\mathbf{\Delta}_{\mathbf{x}_k}^{\mathbf{x}_k}\ln p(\mathcal{Z}_k^\ell|\mathbf{x}_k)\right\} = \mathbb{E}\Bigg\{ - c_2(c_1 + c_2\Lambda_k^\ell)^{-1}\Delta_{\mathbf{x}_k}^{\mathbf{x}_k}\Lambda_k^\ell + \frac{(c_2)^2}{(c_1 + c_2\Lambda_k^\ell)^2}\left(\frac{\partial\Lambda_k^\ell}{\partial\mathbf{x}_k}\right)^T\frac{\partial\Lambda_k^\ell}{\partial\mathbf{x}_k}\Bigg\}, \quad (49)$$

where $\frac{\partial\Lambda_k^\ell}{\partial\mathbf{x}_k}$ and $\Delta_{\mathbf{x}_k}^{\mathbf{x}_k}\Lambda_k^\ell$ are respectively the first order and the second order derivatives of Λ_k^ℓ, given as

$$\mathbb{E}\left\{\frac{\partial\Lambda_k^\ell}{\partial\mathbf{x}_k}\right\} = 0, \quad (50)$$

$$\mathbb{E}\Big\{\Delta_{\mathbf{x}_k}^{\mathbf{x}_k}\Lambda_k^\ell\Big\} = \sum_{q=1}^{n_k^\ell}u_k^q\left\{\left(\frac{\partial\tau_k^\ell(\mathbf{x}_k)}{\partial\mathbf{x}_k^i}\right)^T(\sigma_k^\ell)^{-1}\left(\frac{\partial\tau_k^\ell(\mathbf{x}_k)}{\partial\mathbf{x}_k^j}\right) + \frac{1}{2}\mathrm{tr}\left[(\sigma_k^\ell)^{-1}\left(\frac{\partial\sigma_k^\ell}{\partial\mathbf{x}_k^i}\right)^T(\sigma_k^\ell)^{-1}\frac{\partial\sigma_k^\ell}{\partial\mathbf{x}_k^j}\right]\right\}, \quad (51)$$

where $u_k^q = \mathcal{N}(\hat{\tau}_{q,k}^\ell \,|\, \tau_k^\ell(\mathbf{x}_k), \sigma_k^\ell)$.

Substituting the $\widetilde{\mathbf{D}}_k^{22}$ in (47) into (36) and taking the inverse operation, the PCRLB for multiple TDOA measurement model can be obtained. Since multiple TDOAs are incorporated at each microphone pair, the FIM is different from that based on the single TDOA measurement model by taking both false alarms and miss detection into account.

V. Simulations

In this section, the performance of the developed PCRLB under the single TDOA measurement model (MM1) and the multiple TDOA measurement model (MM2) is studied. The EKF and PF algorithms are also implemented to validate the proposed PCRB. A number of microphone pairs are deployed in the room environment to formulate a distributed microphone pair network. The room dimension is $9 \times 5 \times 3\text{m}^3$ with background noise yielding an SNR level of 20dB. The positions of microphone pairs in the $x-y$ plane are randomly drawn but assumed to be fixed and known during the tracking process. The separation of each microphone pair is 50cm. The height of all microphones is set to be 0.5m below the ceiling. The whole experimental setup is depicted in Fig. 1. Different noisy environments are simulated by adding white Gaussian noise (WGN) with different energy level into the received signal. Various reverberation time T_{60}s are used to describe different reverberant environments. The RIR at each microphone is generated by using the image method [22].

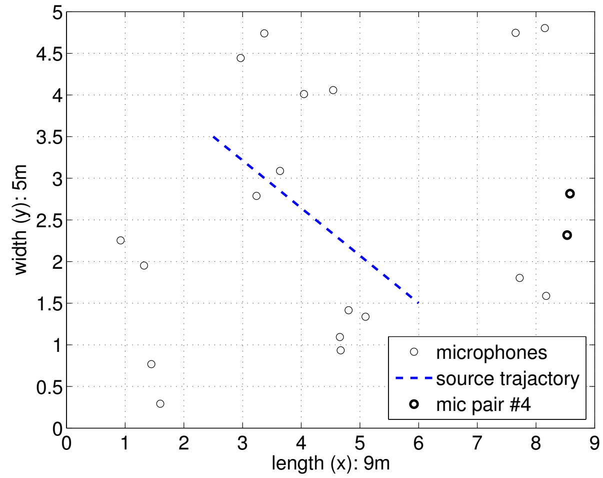
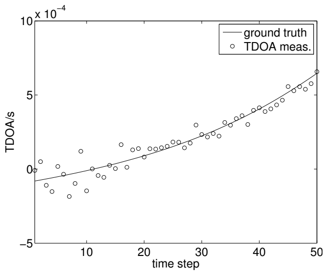


Fig. 1. Source trajectory and the configuration of microphone pairs.

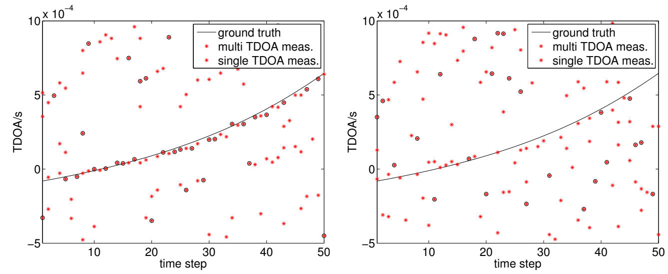
Given the system dynamic model (13), the tracking performance is determined by the variance of the TDOA estimates in (9). In our first study, single TDOA measurement is generated at each time step according to the measurement model (42). Under such a scenario, $q_0 = 0$ and MM1 is the same as MM2. The variance of each TDOA measurement is given by (9). Fig. 3 shows the tracking results and the performance bound under $T_{60} = 0.3$s (case #1) and $T_{60} = 0.6$s (case #2). Fig. 2 shows the TDOA estimates by using the PHAT-GCC method for the fourth microphone pair (indicated by dark circles in Fig. 1). As the reverberation time T_{60} increases, the variance of TDOA becomes larger. Also, since the source moves towards the fourth microphone pair, the SRR becomes lower at the last few time steps. Accordingly, the variance of TDOA decreases. Fig. 3 shows the tracking performance and the developed PCRB under two different reverberation cases.

Since a single TDOA is simulated for each microphone pair, the PCRB based on the multiple TDOA measurement model is coincident with that based on the single TDOA measurement model. Hence, only one PCRB is presented here. It can be observed that as the reverberation time T_{60}



(a) (b)

Fig. 2. Simulated TDOA measurements from microphone pair $\#4$ under (a) $T_{60} = 0.3$s and (b) $T_{60} = 0.6$s.



(a) (b)

Fig. 4. TDOA measurements generated by a Gaussian source signal from microphone pair $\#4$ under (a) $T_{60} = 0.3$s and (b) $T_{60} = 0.6$s.

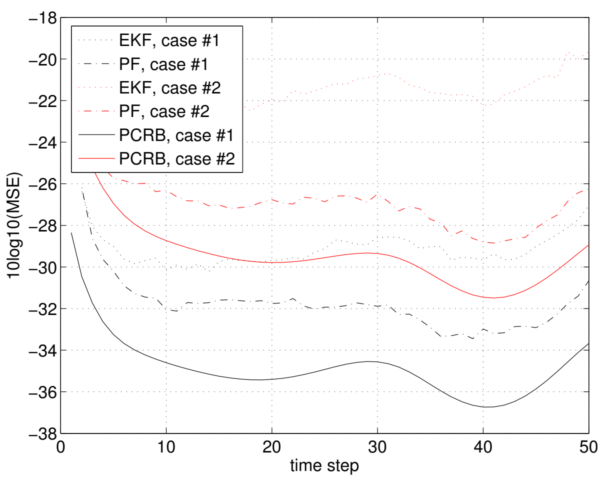


Fig. 3. PCRB and MSE of the tracking algorithms under simulated TDOAs.

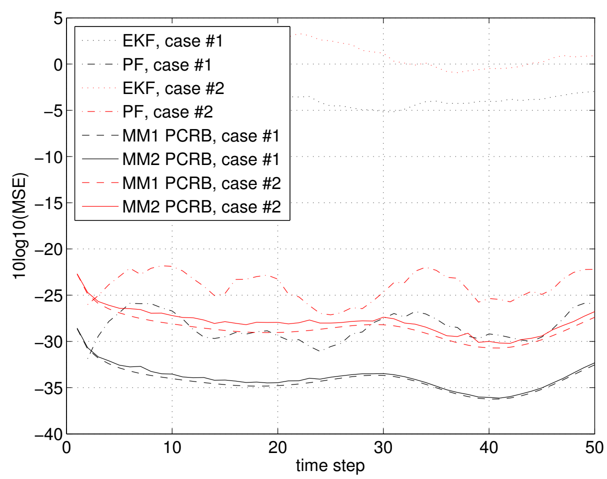
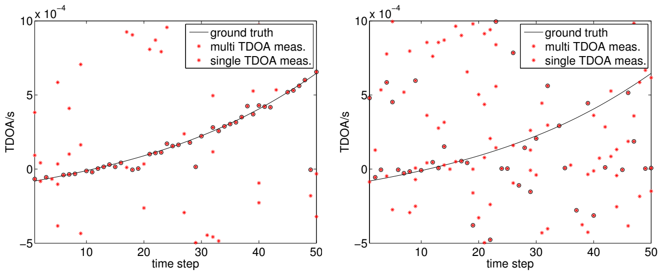


Fig. 5. PCRB and MSE of the tracking algorithms for WGN source signal.

decreases, the PCRLB becomes lower. This indicates that better tracking performance can be expected. Also, for such a microphone deployment, better performance can be achieved at time steps $15-20$ and $35-45$. The mean square error (MSE) of the tracking approaches closely follows the PCRB. This observation further validates the effectiveness of the derived bound. In addition, this simulation shows that PF performs better than EKF in the noisy and reverberant environment, particularly when the reverberation is strong. The EKF can be significantly degraded due to inaccurate TDOA measurements.

In our second study, a WGN signal is used as the source signal so that the source signal is stationary and has a flat spectrum. The SNR is thus the same over the interested frequency band. The maximum number of TDOAs collected at each microphone pair is 4. The parameter for the tracking algorithm is $q_0 = 0.2$ and $\sigma_\tau = 1.5 \times 10^{-5}$. Two above mentioned reverberant environments, case $\#1$ and case $\#2$ are considered. The SNR is fixed to 20dB. Fig. 4 shows the measurements generated by a Gaussian source signal from the microphone pair $\#4$. Since for both cases, the reverberation is strong, the TDOA estimation is degraded significantly. Only a few TDOA estimates from the largest peaks of the GCC function are correct. When multiple peaks are selected, more accurate TDOA can be obtained from those secondary largest peaks. The tracking performance and the derived PCRB are shown in Fig. 5. The single TDOA measurement model (MM1) based PCRB is a theoretical bound since it is derived from the state space model. Basically, it assumes that the TDOAs are always the correct detections and ignores the false alarms and the miss detection. Hence, it is overly optimistic and not achievable. The PCRB derived from the multiple TDOA measurement model (MM2) takes false TDOA measurement and miss detection probability into account. It is higher than the PCRB based on the single TDOA measurement model. For the tracking performance, it is observed that the EKF cannot track the source since the TDOA measurements are degraded seriously and large miss detection and false alarms appear when a single TDOA is used. The PF algorithm, on the other hand, is able to incorporate multiple TDOAs to increase the probability of detection and also considers false alarms and miss detection. It thus performs better than EKF.

The PCRB for real speech signal is also studied. The speech signals are taken from the WSJCAM0 corpus [23] with a sampling frequency of 16kHz. Estimating the TDOAs



(a) (b)

Fig. 6. TDOA measurements generated by a speech signal from microphone pair $\#4$ under (a) $T_{60} = 0$s and (b) $T_{60} = 0.3$s.

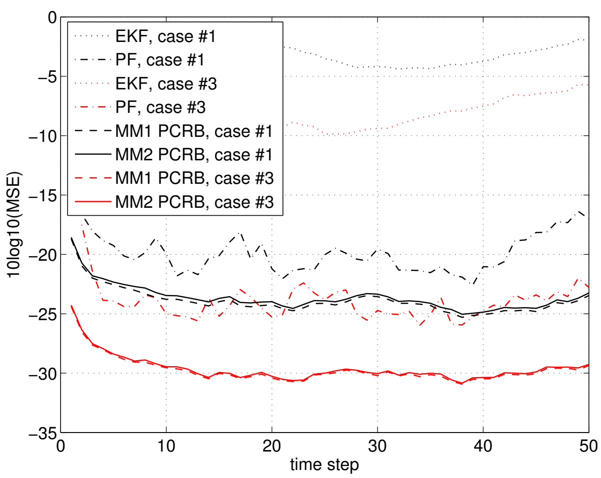


Fig. 7. PCRB and MSE of the tracking algorithms for speech source signal.

for speech is more challenging due to the nonstationary characteristic of the speech signal. Fig. 6 shows the TDOA estimation from the microphone pair $\#4$ under $T_{60} = 0.3$s (case $\#1$) and $T_{60} = 0$s (case $\#3$). The SNR is set as 20dB. Even in the anechoic environment (i.e., $T_{60} = 0$s), the TDOA measurements are degraded significantly. The PCRBs are presented in Fig. 7. Compared to the PCRB for simulated TDOAs and WGN generated TDOAs, the PCRB for real speech signal is higher. However, the variation of the PCRB under a specific environment matches that under simulated TDOA and WGN generated TDOA scenarios well.

VI. Conclusions

In this paper, a Bayesian tracking performance bound is derived for an acoustic source tracking in a room environment. Multiple TDOA measurement model that considers the false alarms and miss detection is considered to make the derived bound more practical. Simulations under different reverberant environments and different source signals are organized to validate the proposed PCRBs. However, the probability of detection on each TDOA measurement is equally distributed. In our future work, real probability of detection on different TDOA measurements will be incorporated and more achievable PCRB will be developed.

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. 2000 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. Hopgood, "Nonconcurrent multiple speakers tracking based on extended kalman particle filter," in *Proc. 2008 IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 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] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. 2001 IEEE Int. Conf. Acoust., Speech, and Signal Process.*, vol. 5, May 2001, pp. 3021–3024.

[13] X. Zhong and J. R. Hopgood, "Particle filtering for TDOA based acoustic source tracking: Nonconcurrent multiple talkers," *Signal Process.*, vol. 96, pp. 382–394, 2014.

[14] T. Gustafsson, B. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *Speech Audio Process, IEEE Transactions on*, vol. 11, no. 6, pp. 791–803, 2003.

[15] P. Tichavsky, C. H. Muravchik, and A. Nehorai, "Posterior Cramér-Rao bounds for discrete-time nonlinear filtering," *IEEE Trans. Signal Process.*, vol. 46, no. 5, pp. 1386–1396, 1998.

[16] X. Zhong, "A Bayesian framework for multiple acoustic source tracking," Ph.D. dissertation, The University of Edinburgh, Edinburgh, U.K., 2010.

[17] 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.

[18] 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.

[19] 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.

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

[21] S. M. Kay, *Foundamentals of Statistical Signal Process.* Prentice Hall, Englewood Cliffs, 1993.

[22] 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.

[23] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): specification and initial experiments," in *Proc. of the 2005 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU-05)*, 2005, pp. 357 – 362.