Swarm Intelligence Based Particle Filter for Alternating Talker Localization and Tracking Using Microphone Arrays

Kai Wu, V. G. Reju, *Member, IEEE,* Andy W. H. Khong, *Member, IEEE,* Shu Ting Goh

***Abstract*—We address the problem of localizing and tracking alternating (moving or stationary) talkers using microphone arrays in a room environment. One of the main challenges is the frequent (and possibly abrupt) change of talker positions which requires the algorithm to capture the active talker rapidly. In addition, the presence of interference, background noise and room reverberation degrades the tracking performance. We propose a new algorithm that jointly exploits the advantages of the particle filter (PF) and particle swarm intelligence. The PF is used as a general tracking framework which incorporates a proposed alternating source-dynamic model for recursive estimation of talker position. Unlike the conventional PF where particles operate independently in the particle sampling stage, the use of swarm intelligence allows particles to interact with each other, thereby improving convergence toward the active talker location. In addition, the memory mechanism in swarm intelligence allows particles to remain at their previous best-fit state estimate when signals are corrupted by interference, noise and/or reverberation. Simulations and experiments were conducted to demonstrate the effectiveness of the proposed algorithm.**

***Index Terms*—Talker localization and tracking, microphone arrays, particle filter, particle swarm intelligence**

I. Introduction

TALKER localization and tracking (TLT) involves estimating the location of an active talker using an array of microphones. TLT is an active research area in recent years due to its widespread applications including teleconferencing [1], [2], automatic camera steering, [3], speech enhancement [4], and human-robot interaction [5]. In this work, we consider the problem of localizing and tracking the position of a dominant talker which alternates frequently between multiple people. This may occur, for example, in an interactive classroom environment where the instructor is moving around and the occupants (instructor and students) speak alternatively during an interaction session. In such a scenario, the dominant talker needs to be tracked in the presence of interferers and noise.

To achieve TLT, location measurements are, in general, independently computed from the received signals at each time frame. Well-known approaches include beamforming [6]–[8] and time-difference-of-arrival (TDOA) [9]–[11]. Given location measurements across successive frames, the temporal consistency of location measurements is then exploited via Bayesian filtering [12], [13] with incorporation of an assumed source motion model. This results in a recursive computation of location estimates using successive data frames.

Within the Bayesian filter family, although the Kalman filter has been proposed for TLT [14], the particle filter (PF) [15], [16] has attracted considerable attention. The sequential importance resampling PF (SIRPF) was first introduced in TLT to track a single source [13], [17]. A voice activity detector was subsequently integrated into this SIRPF framework to mitigate degradation in performance due to the non-continuous nature of speech signals [18]. In [19], the multiple-hypothesis model was incorporated to deal with erroneous measurements caused by reverberation. Besides single-source tracking, research has also been focused on addressing issues pertaining to multiple simultaneously active sources [20]–[27]. To deal with the addition/removal of active/inactive sources, hidden Markov model [25] and random finite set theory [27] have been exploited. Although significant progress has been achieved, tracking of alternating dominant source is still challenging due to (i) the rapid change of position of interest, where the algorithm is required to switch from the previous talker to the active (desired) talker rapidly; (ii) the presence of data frames with low signal-to-interference ratio (SIR), signal-to-noise ratio (SNR) and signal-to-reverberation ratio (SRR), where the tracking performance is to be maintained.

Given the above challenges, direct application of the SIRPF [13], [18] for tracking of alternating sources will not achieve good performance. This is because SIRPF only employs a prior propagation density for particle sampling, and its performance is highly dependent on the predefined source-dynamic model. The SIRPF algorithms based on the single-source dynamic model [13], [18] may result in a lag in detection of the newly active source when an alternation occurs. The multi-source tracking algorithms [20]–[27] are also not suitable for the considered scenario since these algorithms may wrongly take the interferer as the desired source. For example, the algorithm proposed in [25] employs a bank of Kalman filters where each filter can be initialized to track one of the talkers if he/she keeps active. This may result in the wrong initialization of filters due to active interferers. In addition, the multi-source tracking algorithms, in general, do not have any mechanism to achieve fast convergence to the newly active source [28]. To address the alternating-source problem, the extended Kalman PF (EKPF) was proposed where the importance sampling (IS) density is estimated by an extended Kalman filter using the latest measurements [28], [29]. Although the EKPF has shown to achieve fast conver-

The authors are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (Email: eewukai@gmail.com, reju@ntu.edu.sg, andykhong@ntu.edu.sg, shuting@ntu.edu.sg).

gence, it suffers from performance degradation in a noisy and reverberant environment. This is due to the sub-optimal particle sampling when erroneous TDOA measurements are used during the low SNR, SIR and/or SRR frames.

To address the above challenges, we propose a swarm intelligence based PF (SWIPF). Inspired by the foraging behavior of a bird flock, particle swarm optimization (PSO) was originally proposed to optimize a static fitness function [30], [31], which has been applied for localizing and tracking of sources in video [32], [33] and acoustics [34], [35]. While PF incorporates a source-dynamic model for predicting the particles, making it attractive for recursive state estimation in TLT, its performance is limited by the approximation of IS density and particle sampling. In addition, particles propagate independently without any interaction between themselves. On the contrary, particles of the PSO are endowed with the ability to remember their best-fit positions and interact within the population. We therefore exploit this swarm intelligence capability found in PSO for tracking alternating sources. In essence, the memory mechanism allows particles to remain at their previous best-fit positions when TDOA measurement of the current time frame is erroneous. Convergence of the particles is expected to be improved since, with inter-particle interaction, particles can now be directed towards the active source region by sharing the fitness information among themselves.

As opposed to [35]–[38] where either PSO or PF is partially utilized, our proposed framework jointly exploits advantages of both PF and PSO. In [35], prediction of particles is replaced by the swarm move, which excludes the capability of incorporating any source-dynamic model for the alternating talker scenario. While [37], [38] exploit PSO for particle sampling, swarm intelligence is confined within each PF iteration such that memory of the best-fit information cannot be inherited across time frames to compensate the effect of any erroneous measurement. On the contrary, in the unified framework of the proposed SWIPF, PF is used for sequential state estimation, which, in turn, allows us to incorporate the proposed alternating source-dynamic model for predicting the particles effectively. The interaction mechanism from PSO is then exploited for particle convergence. Due to a newly introduced fitness decay mechanism, the memory mechanism from PSO is exploited across time frames such that memory of the previous best-fit information can be utilized if a source-dynamic model mismatch occurs or if the current signal frame is corrupted.

This paper is organized as follows: Section II formulates the TLT problem. The proposed alternating source-dynamic model and measurement likelihood are formulated in Section III. Section IV discusses the proposed SWIPF algorithm. In Section V, the performance of the proposed algorithm is compared with existing techniques while Section VI concludes the paper.

II. Problem Formulation

A. TDOA Measurement

Consider a desired source signal $s(t)$ emanating from position $\mathbf{x} = [x,\ y]^T$ and a number of microphones with known positions being deployed within a room. These microphones are distributed in the form of several linear arrays and, for each array, we define consecutive microphones as a pair indexed by the variable m. Therefore, the known microphone positions can be denoted as $\mathbf{r}_{m,i}$, where $m = 1, \ldots, M$ and $i = 1, 2$ denotes the left and right microphone of that pair, respectively. The microphone received signal can be expressed as

$$y_{m,i}(t) = \beta s(t - \Delta t_{m,i}) + n_{m,i}(t), \tag{1}$$

where $\beta \propto\ 1/||\mathbf{x} - \mathbf{r}_{m,i}||$ is the attenuation parameter, $\Delta t_{m,i}$ is the source-to-microphone propagation time and $n_{m,i}(t)$ denotes signal disturbance due to background noise and/or interference. The TDOA between the mth pair microphones can then be defined as

$$\begin{aligned}\tau_m &= \Delta t_{m,2} - \Delta t_{m,1} \\ &= \big(||\mathbf{x} - \mathbf{r}_{m,2}|| - ||\mathbf{x} - \mathbf{r}_{m,1}||\big)/c \\ &= \tau_m(\mathbf{x}),\end{aligned} \tag{2}$$

where $||\cdot||$ is the Euclidean distance, c is the speed of sound and $\tau_m(\mathbf{x})$ is to signify that the TDOA is a nonlinear function of the source position. We assume that the microphone arrays are deployed along the room perimeter and hence there is no front-back TDOA ambiguity which would otherwise exist for a single linear array [4].

In this work, the TDOA is measured using phase-transformed generalized cross-correlation (GCC) due to its computational efficiency. Defining k as the frame index, the GCC function is given by [10]

$$R_{k,m}(\tau) = \frac{1}{2\pi}\int_{\Omega} \Phi(k,\omega) Y_{m,1}(k,\omega) Y^*_{m,2}(k,\omega) e^{\jmath\omega\tau} d\omega, \tag{3}$$

where $Y_{m,i}(k,\omega)$ is the short-time Fourier transform of $y_{m,i}(t)$, ω is the angular frequency, Ω is the frequency range of interest, $\Phi(k,\omega) = 1/\|Y_{m,1}(k,\omega)Y^*_{m,2}(k,\omega)\|$ [10] and $(\cdot)^*$ denotes complex conjugate. The TDOA of the mth microphone pair can generally be estimated via [9]

$$\widehat{\tau}_{k,m} = \arg \max_{\tau \in [-\tau_{\max}, \tau_{\max}]} R_{k,m}(\tau), \tag{4}$$

where $\tau_{\max} = ||\mathbf{r}_{m,2} - \mathbf{r}_{m,1}||/c$ is the maximum admissible TDOA value.

B. Review of Sequential State Estimation and Particle Filtering

We consider estimating the source position in two-dimension with the state vector being defined as $\mathbf{x}_k = [x_k,\ y_k]^T$ at frame index k, where the elements correspond to the x and y source location coordinates. The measurement vector $\mathbf{z}_k = [\widehat{\tau}_{k,1},\ \widehat{\tau}_{k,2}, \ldots \widehat{\tau}_{k,M}]^T$ is defined by concatenating TDOA estimates across all the M microphone pairs. The state-space model can therefore be given by

$$\mathbf{x}_k = \mathcal{G}(\mathbf{x}_{k-1}, \mathbf{u}_k), \tag{5}$$

$$\mathbf{z}_k = \mathcal{H}(\mathbf{x}_k, \mathbf{w}_k), \tag{6}$$

where $\mathbf{u}_k$ and $\mathbf{w}_k$ denote the process and measurement noise, respectively. In (5), the state-transition function $\mathcal{G}(\cdot)$ defines the state evolution across time frames and will be specified

by a new alternating source-dynamic model in Section III-A. The measurement function $\mathcal{H}(\cdot)$ in (6) introduces nonlinearity such that

$$\begin{aligned}\mathbf{z}_k &= [\widehat{\tau}_{k,1},\ \widehat{\tau}_{k,2}, \dots \widehat{\tau}_{k,M}]^T \\ &= [\tau_1(\mathbf{x}_k),\ \tau_2(\mathbf{x}_k),\ \dots \tau_M(\mathbf{x}_k)]^T + \mathbf{w}_k, \end{aligned} \tag{7}$$

where $\tau_m(\cdot)$ is the nonlinear function defined in (2). Equations (5) and (6) formulate the Bayesian filter for the underlying tracking problem. Given $p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})$, we assume knowledge of $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ derived from $\mathcal{G}(\cdot)$ and $p(\mathbf{z}_k|\mathbf{x}_k)$ derived from $\mathcal{H}(\cdot)$ at the kth frame. The objective is to estimate the posterior probability density function (pdf) $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ recursively.

The particle filter [15], [16] determines a solution by approximating $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ via a set of particles and their associated weights $\{(\mathbf{x}_k^{(p)}, w_k^{(p)})\}_{p=1}^{N_p}$, i.e.,

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \approx \sum_{p=1}^{N_p} w_k^{(p)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(p)}), \tag{8}$$

where $p = 1, \dots, N_p$ denotes the particle index, N_p is the number of particles, $\mathbf{x}_k^{(p)}$ is the pth particle in the state space, $w_k^{(p)}$ is its associated weight, and $\delta(\cdot)$ is the Dirac delta function. At each iteration, the PF consists of a prediction and an update stage. In the prediction stage, particles of the previous time frame $\mathbf{x}_{k-1}^{(p)}$ are propagated to the current time frame by sampling the IS density described by

$$\mathbf{x}_k^{(p)} \sim p^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k), \tag{9}$$

where the superscript $\cdot^{(\mathrm{IS})}$ denotes for the importance sampling density. In the update stage, each particle weight is updated according to

$$w_k^{(p)} \propto\ w_{k-1}^{(p)}\ \frac{p(\mathbf{z}_k|\mathbf{x}_k^{(p)})\ p(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)})}{p^{(\mathrm{IS})}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k)}. \tag{10}$$

In general, the PF requires resampling to mitigate the problem of degeneration [15].

C. Review of Particle Swarm Optimization

PSO derives an optimal solution of a fitness function $\mathcal{F}(\mathbf{x})$ within a given search space [30], [31]. In contrast to PF where each particle moves independently in the particle propagation step, PSO allows each particle to retain memory of its best-fit position and communicate with other particles. Given $\mathcal{F}(\mathbf{x})$ to be maximized, PSO is initialized using a group of randomly distributed particles $\{\mathbf{x}_0^{(p)}\}_{p=1}^{N_p}$. Two important factors are introduced: the previous best position $\mathbf{x}_{\mathrm{pb}}^{(p)}$ that each particle has found so far, and the global best position $\mathbf{x}_{\mathrm{gb}}$ found within the entire population. Particle movement can then be described as [30]

$$\begin{aligned}\mathbf{v}_k^{(p)} = \chi\Big[&\mathbf{v}_{k-1}^{(p)} + \varphi_1\boldsymbol{\gamma}_1 \odot (\mathbf{x}_{\mathrm{pb}}^{(p)} - \mathbf{x}_{k-1}^{(p)}) \\ &+ \varphi_2\boldsymbol{\gamma}_2 \odot (\mathbf{x}_{\mathrm{gb}} - \mathbf{x}_{k-1}^{(p)})\Big], \end{aligned} \tag{11}$$

$$\mathbf{x}_k^{(p)} = \mathbf{x}_{k-1}^{(p)} + \mathbf{v}_k^{(p)}, \tag{12}$$

where $\mathbf{v}_k^{(p)}$ denotes the particle velocity, φ_1 and φ_2 are the acceleration parameters, $\boldsymbol{\gamma}_1$, $\boldsymbol{\gamma}_2$ are vectors with elements uniformly sampled within the range $[0, 1]$ and $\odot$ denotes element-wise product. Defining $\varphi = \varphi_1 + \varphi_2$ and $\varphi > 4$, a constriction parameter $\chi = \frac{2}{|2-\varphi-\sqrt{\varphi^2-4\varphi}|}$ is used to generate a damping effect on each particle's oscillation. Such a constriction can prevent misconvergence — a phenomenon where the particles increasingly oscillate until out-of-bound error occurs [30].

In (11) and (12), the movement of each particle mainly depends on the individual cognitive knowledge $\varphi_1\boldsymbol{\gamma}_1 \odot (\mathbf{x}_{\mathrm{pb}}^{(p)} - \mathbf{x}_{k-1}^{(p)})$ and the social influence $\varphi_2\boldsymbol{\gamma}_2 \odot (\mathbf{x}_{\mathrm{gb}} - \mathbf{x}_{k-1}^{(p)})$ which drive the particle to $\mathbf{x}_{\mathrm{pb}}^{(p)}$ and $\mathbf{x}_{\mathrm{gb}}$, respectively, using oscillated forces introduced by $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$. The variables φ_1 and φ_2 define the relative weights on these two forces. In general, $\varphi_1 = \varphi_2$ is assumed if no prior knowledge of the relative weights is available [30]. After (11) and (12), $\mathbf{x}_{\mathrm{pb}}^{(p)}$ and $\mathbf{x}_{\mathrm{gb}}$ are updated by

$$\mathbf{x}_{\mathrm{pb}}^{(p)} \Leftarrow \begin{cases} \mathbf{x}_k^{(p)}, & \text{if } \mathcal{F}(\mathbf{x}_k^{(p)}) \geq f_{\mathrm{pb}}^{(p)}; \\ \mathbf{x}_{\mathrm{pb}}^{(p)}, & \text{otherwise,} \end{cases} \tag{13}$$

$$\mathbf{x}_{\mathrm{gb}} = \arg\max_{\mathbf{x}_{\mathrm{pb}}^{(p)}} \mathcal{F}(\mathbf{x}_{\mathrm{pb}}^{(p)}), \tag{14}$$

where $\mathcal{F}(\mathbf{x}_k^{(p)})$ denotes fitness value evaluated at the propagated particle and $f_{\mathrm{pb}}^{(p)} = \mathcal{F}(\mathbf{x}_{\mathrm{pb}}^{(p)})$ is the fitness value of the previous best-fit position.

The PSO presented above is well suited for a static optimization problem where $\mathcal{F}(\mathbf{x})$ and its optimal solution are time invariant. Application of the PSO for TLT, however, requires further modification since the TLT involves a dynamic search of the source position which varies across time.

III. Proposed Models for Alternating-source Tracking

We first propose a new source-dynamic model $\mathcal{G}(\cdot)$ that is suitable for our alternating-source scenario and the corresponding $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ is derived. The measurement likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ will then be formulated by taking into account the reliability of $\widehat{\tau}_{k,m}$ due to noise, interference and/or reverberation. These two sources of information will then be used in SWIPF described in Section IV.

A. Alternating Source-dynamic Model

The state-transition function $\mathcal{G}(\cdot)$ in (5) determines the state evolution across time frames. The Langevin and random walk processes have commonly been used to model this state transition for single-source tracking application [13], [18]. While these processes have also been extended for tracking alternating sources by simply increasing the process noise to account for alternation uncertainty [19], [28], [29], the underlying continuous moving-source assumption may not be valid for the alternating-source scenario.

The function $\mathcal{G}(\cdot)$ that we are proposing better suits the

alternating-source scenario by introducing two hypotheses:

$\mathcal{S}_k^0$: An alternation has not occurred at the kth frame;

$\mathcal{S}_k^1$: An alternation has occurred at the kth frame.

Under hypothesis $\mathcal{S}_k^0$, indicating that the source location is consistent with the previous frames, either a Langevin process model or a random walk model can be used to describe the continuous talker motion. In this work, due to computational efficiency, the random walk model, which models perturbation of the source location within the neighborhood region of the last iteration, is used[1]. Therefore, the state is predicted using

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{u}_k, \quad \text{if} \quad \mathcal{S}_k^0, \tag{15}$$

where $\mathbf{u}_k \sim \mathcal{N}(\cdot; \mathbf{0}_{2\times 1}, \mathbf{\Sigma}_{2\times 2})$, $\Sigma_{2\times 2} = \sigma_{\mathbf{u}}^2 \mathbf{I}_{2\times 2}$ is the covariance of the Gaussian distribution and $\sigma_{\mathbf{u}}^2$ defines the variance of the human motion both in x and y directions. Given (15), the conditioned state-transition probability can be written as

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathcal{S}_k^0) = \mathcal{N}(\mathbf{x}_k; \mathbf{x}_{k-1}, \mathbf{\Sigma}_{2\times 2}). \tag{16}$$

On the other hand, under hypothesis $\mathcal{S}_k^1$, an alternation is expected to have occurred. This leads the state being re-initialized in a uniformly distributed manner within the enclosed domain as

$$\mathbf{x}_k = \mathcal{U}_{\mathcal{D}}(\mathbf{x}_k), \quad \text{if} \quad \mathcal{S}_k^1, \tag{17}$$

where the function $\mathcal{U}_{\mathcal{D}}(\cdot)$ denotes a multivariate uniform distribution over the enclosure domain $\mathcal{D}$. In this work, we consider a rectangular shoe-box room such that $\mathcal{D} = \{x_k, y_k | x_{\min} \leq x_k \leq x_{\max}, y_{\min} \leq y_k \leq y_{\max}\}$, where the variables $x_{\min}$, $x_{\max}$ and $y_{\min}$, $y_{\max}$ denote boundaries in the x and y directions, respectively. The corresponding conditioned state-transition probability can be written as

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathcal{S}_k^1) = \frac{1}{(x_{\max} - x_{\min})(y_{\max} - y_{\min})}. \tag{18}$$

Apart from $\mathcal{U}_{\mathcal{D}}(\cdot)$ in (17), other state distributions may also be used to reflect the occurrence of any new active source, e.g., a lower probability near the room boundaries. However, determination of this pdf is environment dependent and beyond the scope of this paper.

Given the two hypotheses, the state-transition probability in (10) can then be computed by

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}) = \sum_{i=0}^{1} p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathcal{S}_k^i) p(\mathcal{S}_k^i|\mathbf{x}_{k-1}), \tag{19}$$

where $p(\mathcal{S}_k^i|\mathbf{x}_{k-1})$ denotes the prior probability of $\mathcal{S}_k^i$ and we define $p(\mathcal{S}_k^i|\mathbf{x}_{k-1}) = P_{\text{alt}}$ and $p(\mathcal{S}_k^0|\mathbf{x}_{k-1}) = 1 - P_{\text{alt}}$ where P_{alt} is a predefined probability value for alternation.

In the above derivation, the source is assumed to alternate at every frame with an empirical probability of P_{alt}. This assumption may be violated when the existing source is continuously active. Hence, direct incorporation of the proposed $\mathcal{G}(\cdot)$ into the SIRPF will still lead to model mismatch. This model mismatch problem will be addressed using the proposed memory mechanism within SWIPF described in Section IV.

[1]Additional simulation shows that the use of Langevin process will not bring significant performance improvement compared to the random walk model since the effect of Langevin process can be reduced by the swarm update in (30).

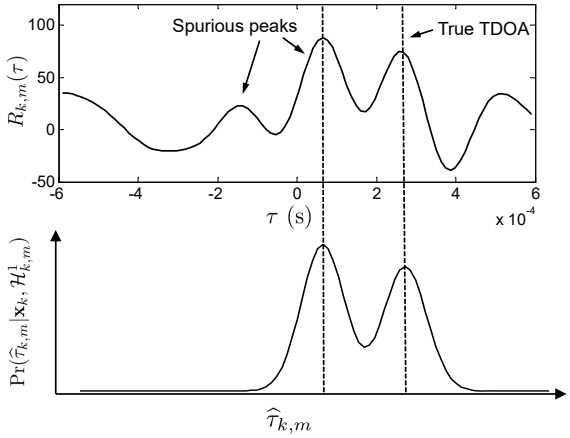

Fig. 1: Upper plot: GCC function computed using (3) for actual recorded signals with the same room-source setup described in Sec. V at an estimated SNR = 15 dB and $T_{60} = 0.35$ s; Lower plot: conditioned measurement likelihood $p(\widehat{\tau}_{k,m}|\mathbf{x}_k, \mathcal{H}_{k,m}^1)$ computed as a mixture of Gaussian pdfs as in (20) using the peaks in upper plot.

B. Measurement Likelihood

The GCC function $R_{k,m}(\tau)$ defined in (3) is used to obtain $\widehat{\tau}_{k,m}$, which will then be used to formulate $p(\mathbf{z}_k|\mathbf{x}_k)$. As shown in the upper plot of Fig. 1, the peak corresponding to the true TDOA may be lower than the spurious peaks caused by interference, noise and reverberation. Therefore, as opposed to using only the maximum of $R_{k,m}(\tau)$ described in (4), we consider the use of multiple peaks in $R_{k,m}(\tau)$ in order to increase the probability of including the true peak. In this work, for each $R_{k,m}(\tau)$, peaks that are higher than 0.7 of the maximum peak will be used as TDOA candidates. These candidates are denoted by $\{\widehat{\tau}_{k,m}^{\ell}\}_{\ell=1}^{N_{k,m}}$, where ℓ is the peak index and $N_{k,m}$ is the number of peaks for that pair.

Furthermore, due to variation in speech energy, the received signal power may vary across different microphone pairs and time frames. Two hypotheses are introduced to describe the reliability of estimated TDOA:

$\mathcal{H}_{k,m}^0$: TDOA estimate $\widehat{\tau}_{k,m}$ being unreliable;

$\mathcal{H}_{k,m}^1$: TDOA estimate $\widehat{\tau}_{k,m}$ being reliable.

Under hypothesis $\mathcal{H}_{k,m}^1$, one of the elements in $\{\widehat{\tau}_{k,m}^{\ell}\}_{\ell=1}^{N_{k,m}}$ will correspond to the true TDOA. Given that it is uncertain as to which $\widehat{\tau}_{k,m}^{\ell}$ corresponds to the true TDOA, we propose to include all the elements in $\{\widehat{\tau}_{k,m}^{\ell}\}_{\ell=1}^{N_{k,m}}$ such that the conditioned measurement likelihood is formulated as a mixture of Gaussian pdfs given by [13]

$$p(\widehat{\tau}_{k,m}|\mathbf{x}_k, \mathcal{H}_{k,m}^1) = \sum_{\ell=1}^{N_{k,m}} \bar{a}_{k,m}^{\ell} \mathcal{N}(\widehat{\tau}_{k,m}^{\ell}; \tau_m(\mathbf{x}_k), \sigma_{\tau}^2), \tag{20}$$

where $\bar{a}_{k,m}^{\ell} = a_{k,m}^{\ell} / \sum_{\ell=1}^{N_{k,m}} a_{k,m}^{\ell}$ is the normalized amplitude of each TDOA candidate given that $a_{k,m}^{\ell}$ is the amplitude of the ℓth peak of $R_{k,m}(\tau)$, $\tau_m(\cdot)$ is the nonlinear function defined in (2), and σ_{τ}^2 is the variance of the Gaussian

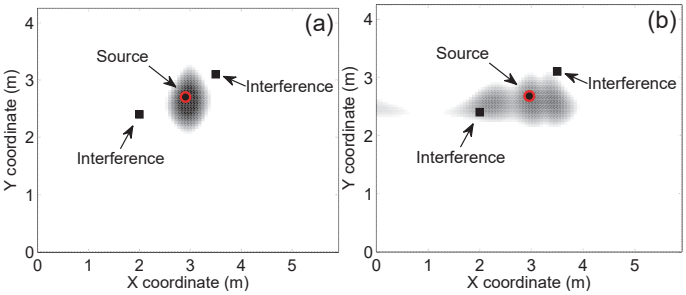

Fig. 2: A simulation example of $p(\mathbf{z}_k|\mathbf{x}_k)$ in the log-domain as a function of $\mathbf{x}_k$ for a room environment with $T_{60} = 300$ ms and SNR = 15 dB. The desired talker and two interferers are present with the SIR = 10 dB for each interferer. (a) at the 45th frame (high source energy frame); (b) at the 47th frame (low source energy frame).

component. The normalization after (20) ensures that the integral of the derived pdf equals to unity. It also implies that, as opposed to using only the maximum within the set $\{\widehat{\tau}^{\ell}_{k,m}\}_{\ell=1}^{N_{k,m}}$, amplitude-based weightings are being applied on each candidate $\widehat{\tau}^{\ell}_{k,m}$. These weights do not compromise the fact that $\widehat{\tau}^{\ell}_{k,m}$ with a higher amplitude, in general, has a higher probability that it corresponds to the true TDOA of the desired dominant source. The lower plot of Fig. 1 shows the derived $p(\widehat{\tau}_{k,m}|\mathbf{x}_k, \mathcal{H}^1_{k,m})$ computed using the two peaks of $R_{k,m}(\tau)$ in the upper plot. Although the maximum of the pdf may not correspond to the true TDOA, a high probability is derived for the true TDOA. It is worth noting that a high density caused by reverberation, noise and interference will not be consistent spatially (across microphone pairs) and temporally (across time frames) due to inconsistency of the spurious peaks in $R_{k,m}(\tau)$ [19]. These biased effects can therefore be minimized since the proposed algorithm employs multiple microphone pairs and incorporates tracking and memory mechanism during successive frames.

For the case of hypothesis $\mathcal{H}^0_{k,m}$, the conditioned measurement likelihood can be modeled as a uniform distribution within the maximum admissible TDOA range, i.e.,

$$p(\widehat{\tau}_{k,m}|\mathbf{x}_k, \mathcal{H}^0_{k,m}) = \frac{1}{2\tau_{\max}}, \tag{21}$$

where $\tau_{\max}$ has been defined after (4). Therefore, the measurement likelihood of the mth microphone pair is

$$p(\widehat{\tau}_{k,m}|\mathbf{x}_k) = \sum_{i=0}^{1} p(\widehat{\tau}_{k,m}|\mathbf{x}_k, \mathcal{H}^i_{k,m}) p(\mathcal{H}^i_{k,m}|\mathbf{x}_k), \tag{22}$$

where $p(\mathcal{H}^i_{k,m}|\mathbf{x}_k)$ is the prior hypothesis probability. This probability can be determined using the instantaneous power ratio $\lambda_{k,m} = 10\log_{10}\left\{\frac{P(k,m)}{P_n(m)}\right\}$ where $P(k,m)$ is the instantaneous power of the mth microphone pair at frame k computed by averaging over the two channels in that pair. The variable $P_n(m)$ denotes the noise power at the mth microphone pair evaluated during non-speech periods obtained by the voice-activity detection algorithm [18]. Assuming that noise is stationary across a few frames during the tracking process, we have $P(k,m) \geq P_n(m)$ giving $\lambda_{k,m} \geq 0$. A higher value of $\lambda_{k,m}$ implies close source-sensor distance, absence of occlusion between source and sensors or high signal power. This translates to a higher probability that $\widehat{\tau}_{k,m}$ is more reliable. We therefore adopt, similar to [18], a monotonic mapping function $p(\mathcal{H}^1_{k,m}|\mathbf{x}_k) = \frac{2}{\pi}\arctan(\lambda_{k,m})$ for the pdf formulation, and $p(\mathcal{H}^0_{k,m}|\mathbf{x}_k) = 1 - p(\mathcal{H}^1_{k,m}|\mathbf{x}_k)$.

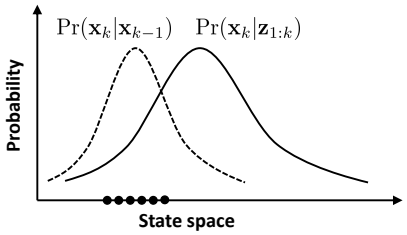

Fig. 3: An illustration of the sampling impoverishment problem in the PF.

With reference to the definition of $\mathbf{z}_k$ in (7) and assuming that $\widehat{\tau}_{k,m}$ from each microphone pair is independent [27], [29], the overall measurement likelihood can be written as

$$p(\mathbf{z}_k|\mathbf{x}_k) = \prod_{m=1}^{M} p(\widehat{\tau}_{k,m}|\mathbf{x}_k). \tag{23}$$

Figure 2 shows examples of the computed $p(\mathbf{z}_k|\mathbf{x}_k)$ as function of $\mathbf{x}_k$. It can be observed that a high density (denoted by dark color) has been achieved at the desired source position in Fig. 2 (a) due to high SIR, SNR and/or SRR at that frame, while during the low source energy frames in Fig. 2 (b), the density is comparatively significant for the interferers. The pdf $p(\mathbf{z}_k|\mathbf{x}_k)$ therefore serves as a weighting function in (10) only during signal frames with high SIR, SNR and/or SRR. For other frames, the proposed algorithm exploits the memory mechanism as described in the following section.

IV. Proposed Swarm Intelligence Based PF for Alternating Source Tracking

Performance of the PF is mainly determined by the sampling of particles. Conventional SIRPF approximates the IS density as $p^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_k) \approx p(\mathbf{x}_k|\mathbf{x}_{k-1})$ [13], [18]. However, this sub-optimal IS density may result in the sampling impoverishment problem. As illustrated in Fig. 3, when a mismatch between the assumed state-transition model and actual source motion occurs, $p(\mathbf{x}_k|\mathbf{x}_{k-1}) \neq p(\mathbf{x}_k|\mathbf{z}_{1:k})$ and most of the particles that are sampled from $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ contributes insignificantly for the approximation of $p(\mathbf{x}_k|\mathbf{z}_{1:k})$. In the alternating-source tracking scenario, this implies that if a single-source model (e.g. the Langevin process model [18]) is used, the particles may wrongly be sampled near the state space of the previous talker position when another talker has become active. While this lag effect can be reduced by applying the proposed alternating source-dynamic model $\mathcal{G}(\cdot)$, model mismatch may still occur where alternation is wrongly assumed for a continuously active source. As a result, particles may wrongly be sampled to detect any non-existent "new" source. The EKPF addresses the model mismatch problem by approximating $p^{(\mathrm{IS})}(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_k)$ using the latest measurement $\mathbf{z}_k$ [28]. However, particle sampling may still be compromised if the latest measurement $\mathbf{z}_k$ is erroneous. For the talker tracking scenario, this problem occurs in frames with low SNR, SIR, and/or SRR.

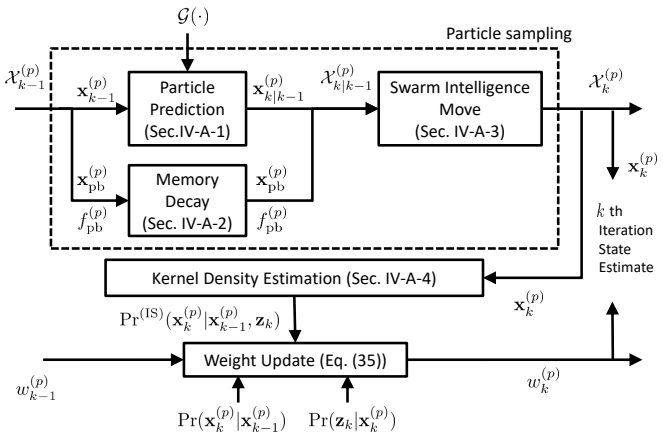

Fig. 4: Schematic diagram of the proposed SWIPF algorithm for a single iteration.

To address the above problems, our proposed SWIPF exploits the advantages of PF and PSO jointly for optimal particle sampling. In the following, without loss of generality, the proposed SWIPF is first introduced in the context of a generic state estimation problem. The solution of SWIPF for alternating source tracking will then be discussed.

A. Swarm Intelligence Based PF

As opposed to the conventional particle filter where only $\mathbf{x}_k^{(p)}$ is used for frame k, the proposed SWIPF incorporates swarm intelligence attributes that comprise of historical best-fit particle $\mathbf{x}_{\mathrm{pb}}^{(p)}$ and its previous best-fit fitness $f_{\mathrm{pb}}^{(p)}$ for memorizing historical information. This results in the full information set of a particle given by

$$\mathcal{X}_k^{(p)} = \{\mathbf{x}_k^{(p)},\ \mathbf{x}_{\mathrm{pb}}^{(p)},\ f_{\mathrm{pb}}^{(p)}\}. \tag{24}$$

Furthermore, a fitness function is to be formulated for directing the particles towards the optimal position in state space. By considering the measurement likelihood $p(\mathbf{z}_k|\mathbf{x})$ as a function of $\mathbf{x}$, the fitness function is proposed as

$$\mathcal{F}_k(\mathbf{x}) \triangleq \mathcal{M}\{p(\mathbf{z}_k|\mathbf{x})\}, \tag{25}$$

where $\mathcal{M}(\cdot)$ denotes a monotonic function. Unlike the static function $\mathcal{F}(\cdot)$ in (13) and (14), the subscript k in $\mathcal{F}_k(\cdot)$ implies that the proposed fitness function varies across time frames for a dynamic state tracking problem for which, examples have been shown in Fig. 2. The subscript k in $\mathbf{x}$, however, has been temporarily omitted to imply that this time-varying function is defined for a general space of $\mathbf{x}$. The definition of (25) is motivated by the fact that the maximum of $\mathcal{F}_k(\mathbf{x})$ corresponds to the maximum likelihood estimate of $\mathbf{x}$ from $p(\mathbf{z}_k|\mathbf{x})$ and the particles will converge to this state estimate. A schematic diagram of the proposed SWIPF is shown in Fig. 4 and the detailed steps are described in the following.

1) Particle Prediction: Given the full information set of a particle at previous frame $\mathcal{X}_{k-1}^{(p)} = \{\mathbf{x}_{k-1}^{(p)},\ \mathbf{x}_{\mathrm{pb}}^{(p)},\ f_{\mathrm{pb}}^{(p)}\}$, we now derive the prediction of $\mathcal{X}_{k|k-1}^{(p)} = \{\mathbf{x}_{k|k-1}^{(p)},\ \mathbf{x}_{\mathrm{pb}}^{(p)},\ f_{\mathrm{pb}}^{(p)}\}$ where the subscript $k|k-1$ signifies prediction. Firstly, the state component $\mathbf{x}_{k|k-1}^{(p)}$ can be predicted using the assumed state-transition model $\mathcal{G}(\cdot)$ given by

$$\mathbf{x}_{k|k-1}^{(p)} = \mathcal{G}(\mathbf{x}_{k-1}^{(p)}). \tag{26}$$

Similar to conventional PF, this step propagates the particles in order to explore any potential state candidate according to knowledge of state-transition statistics.

2) Memory Decay Mechanism: To obtain $\mathbf{x}_{\mathrm{pb}}^{(p)}$ and $f_{\mathrm{pb}}^{(p)}$ in $\mathcal{X}_{k|k-1}^{(p)}$, we know that both of them represent the historical best-fit information of the particle, which is crucial to direct the particle towards the "memorised" best state estimate. Nevertheless, to deal with the time-varying nature of $\mathcal{F}_k(\cdot)$ where the past information has to be gradually reduced while the recent information should be emphasized, a linear decay of $f_{\mathrm{pb}}^{(p)}$ is first performed at each iteration as

$$f_{\mathrm{pb}}^{(p)} \Leftarrow f_{\mathrm{pb}}^{(p)} - \Delta f, \tag{27}$$

where Δf denotes the fitness decay amount — a lower value of Δf indicates a slower memory decay such that historical information will be weighted more.

After fitness decay, comparison can be made between the latest fitness value of the predicted $\mathbf{x}_{k|k-1}^{(p)}$ evaluated using $\mathcal{F}_k(\cdot)$ and the reduced version of $f_{\mathrm{pb}}^{(p)}$ from (27). If $\mathbf{x}_{k|k-1}^{(p)}$ is closer to the true state than $\mathbf{x}_{\mathrm{pb}}^{(p)}$, and provided that $\mathcal{F}_k(\cdot)$ is reliable, we have $\mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}) > f_{\mathrm{pb}}^{(p)}$. Hence, $\mathbf{x}_{\mathrm{pb}}^{(p)}$ and $f_{\mathrm{pb}}^{(p)}$ should be timely updated using the latest best-fit information. Otherwise, historical information should be preserved. The above can be described using

$$\mathbf{x}_{\mathrm{pb}}^{(p)} \Leftarrow \begin{cases} \mathbf{x}_{k|k-1}^{(p)}, & \text{if } \mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}) \geq f_{\mathrm{pb}}^{(p)}; \\ \mathbf{x}_{\mathrm{pb}}^{(p)}, & \text{otherwise,} \end{cases} \tag{28}$$

$$f_{\mathrm{pb}}^{(p)} \Leftarrow \begin{cases} \mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}), & \text{if } \mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}) \geq f_{\mathrm{pb}}^{(p)}; \\ f_{\mathrm{pb}}^{(p)}, & \text{otherwise.} \end{cases} \tag{29}$$

Here, the memory mechanism of particle swarm intelligence has been exploited by taking the time-varying nature of $\mathcal{F}_k(\cdot)$ into account. The historical information $f_{\mathrm{pb}}^{(p)}$ and $\mathbf{x}_{\mathrm{pb}}^{(p)}$ will be preserved if $\mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}) < f_{\mathrm{pb}}^{(p)}$. This can occur if 1) $\mathbf{x}_{k|k-1}^{(p)}$ is wrongly predicted by $\mathcal{G}(\cdot)$, or if 2) the fitness function $\mathcal{F}_k(\cdot)$ is erroneous at the kth frame.

3) Swarm Intelligence Move for Optimal Particle Sampling: Given $\mathcal{X}_{k|k-1}^{(p)}$, the particle information for the kth frame $\mathcal{X}_k^{(p)} = \{\mathbf{x}_k^{(p)},\ \mathbf{x}_{\mathrm{pb}}^{(p)},\ f_{\mathrm{pb}}^{(p)}\}$ can be achieved by the swarm intelligence move. Similar to (11) and (12), $\mathbf{x}_k^{(p)}$ in $\mathcal{X}_k^{(p)}$ can be obtained using

$$\begin{aligned} \mathbf{x}_k^{(p)} = \ & \mathbf{x}_{k|k-1}^{(p)} + \chi\big[\varphi_1\boldsymbol{\gamma}_1 \odot (\mathbf{x}_{\mathrm{pb}}^{(p)} - \mathbf{x}_{k|k-1}^{(p)}) \\ & + \varphi_2\boldsymbol{\gamma}_2 \odot (\mathbf{x}_{\mathrm{gb}} - \mathbf{x}_{k|k-1}^{(p)})\big], \end{aligned} \tag{30}$$

where $\mathbf{x}_{\mathrm{gb}}$ is evaluated among all $\mathbf{x}_{\mathrm{pb}}^{(p)}$, i.e.,

$$\mathbf{x}_{\mathrm{gb}} = \arg\max_{\mathbf{x}_{\mathrm{pb}}^{(p)}} f_{\mathrm{pb}}^{(p)}. \tag{31}$$

It can be observed in (30) that the particle has been driven to $\mathbf{x}_k^{(p)}$ to explore for a better state estimate, according to a joint force contributed by its individual historical information $\mathbf{x}_{\mathrm{pb}}^{(p)}$

and social interaction information $\mathbf{x}_{\rm gb}$. Since a movement has been made for $\mathbf{x}_k^{(p)}$ from $\mathbf{x}_{k|k-1}^{(p)}$, the latest fitness value of $\mathbf{x}_k^{(p)}$ needs to be evaluated using $\mathcal{F}_k(\cdot)$ and the variables $\mathbf{x}_{\rm pb}^{(p)}$, $f_{\rm pb}^{(p)}$ in $\mathcal{X}_k^{(p)}$ need to be updated as

$$\mathbf{x}_{\rm pb}^{(p)} \Leftarrow \begin{cases} \mathbf{x}_k^{(p)}, & \text{if } \mathcal{F}_k(\mathbf{x}_k^{(p)}) \geq f_{\rm pb}^{(p)}; \\ \mathbf{x}_{\rm pb}^{(p)}, & \text{otherwise,} \end{cases} \tag{32}$$

$$f_{\rm pb}^{(p)} \Leftarrow \begin{cases} \mathcal{F}_k(\mathbf{x}_k^{(p)}), & \text{if } \mathcal{F}_k(\mathbf{x}_k^{(p)}) \geq f_{\rm pb}^{(p)}; \\ f_{\rm pb}^{(p)}, & \text{otherwise.} \end{cases} \tag{33}$$

Finally, while $\mathbf{x}_k^{(p)}$ obtained in (30) represents the latest particle, the variable $\mathbf{x}_{\rm pb}^{(p)}$ records the best-fit state estimate among $\mathbf{x}_{k-1}^{(p)}$, $\mathbf{x}_{k|k-1}^{(p)}$ and $\mathbf{x}_k^{(p)}$ in terms of closeness to the true state. We therefore propose to use $\mathbf{x}_{\rm pb}^{(p)}$ as the sampled particle for the kth iteration by performing the following assignment

$$\mathbf{x}_k^{(p)} \Leftarrow \mathbf{x}_{\rm pb}^{(p)}. \tag{34}$$

The swarm intelligence based particle sampling is summarized in Table I. Note that the above can be viewed as a hierarchical sampling for achieving $\mathbf{x}_k^{(p)} \sim p^{(\rm IS)}(\mathbf{x}_k|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k)$ where $p^{(\rm IS)}(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_k) \approx p(\mathbf{x}_k|\mathbf{z}_{1:k})$. That is, given $\mathbf{x}_{k-1}^{(p)}$ for approximating $p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})$, sampling at the current frame has been achieved via prediction using $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ in (26), followed by an update step using the latest measurement $\mathbf{z}_k$ when evaluating $\mathcal{F}_k(\cdot)$ in swarm intelligence described in (28)-(34).

4) Weight Update: Similar to conventional PF, computation of $w_k^{(p)}$ is given by

$$w_k^{(p)} \propto w_{k-1}^{(p)} \frac{p(\mathbf{z}_k|\mathbf{x}_k^{(p)})\ p(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)})}{p^{(\rm IS)}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k)}, \tag{35}$$

where $p(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)})$ can be computed by substituting $\mathbf{x}_k^{(p)}$ in $\mathcal{X}_k^{(p)}$ and $\mathbf{x}_{k-1}^{(p)}$ in $\mathcal{X}_{k-1}^{(p)}$ into (19). Furthermore, given the fitness definition in (25) and the availability of $f_{\rm pb}^{(p)}$, $p(\mathbf{z}_k|\mathbf{x}_k^{(p)}) = \mathcal{M}^{-1}(f_{\rm pb}^{(p)})$ where $\mathcal{M}^{-1}(\cdot)$ denotes the inverse function of $\mathcal{M}(\cdot)$ defined in (25).

For the computation of $p^{(\rm IS)}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k)$, since particles have been well-sampled as $\{\mathbf{x}_k^{(p)}\}_{p=1}^{N_p}$ by swarm intelligence, $p^{(\rm IS)}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k)$ can be estimated from these sampled particles using kernel density estimation (KDE) [39], [40] as

$$p^{(\rm IS)}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k) = \frac{1}{N_p} \sum_{i=1}^{N_p} \mathcal{K}_{\mathbf{H}}(\mathbf{x}_k^{(p)} - \mathbf{x}_k^{(i)}), \tag{36}$$

where $\mathcal{K}_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2}\mathcal{K}(\mathbf{H}^{-1/2}\mathbf{x})$ is a scaled kernel density function specified by the bandwidth matrix $\mathbf{H}$ while $\mathcal{K}(\cdot)$ is the basic kernel density function. The choice of $\mathcal{K}(\cdot)$ is not crucial to the accuracy of density estimation and $\mathcal{K}(\cdot) = \mathcal{N}(\cdot; \mathbf{0}, \mathbf{I})$ is commonly used [39]. In addition, $\mathbf{H}$ controls the the amount of smoothing in the estimated pdf. Here, we consider the Silverman's rule of thumb [41], $\mathbf{H} = (N_p)^{-1/3}\text{diag}([\sigma_x^2,\ \sigma_y^2])$, where σ_x and σ_y denote the variance of particles in x and y directions, respectively.

TABLE I: Swarm intelligence based particle sampling.

Input: $\{\mathcal{X}_{k-1}^{(p)}\}_{p=1}^{N_p} = \{\mathbf{x}_{k-1}^{(p)},\ \mathbf{x}_{\rm pb}^{(p)},\ f_{\rm pb}^{(p)}\}_{p=1}^{N_p}$

For each particle:

- Predict $\mathbf{x}_{k|k-1}^{(p)}$ using the state-transition model, i.e., $\mathbf{x}_{k|k-1}^{(p)} = \mathcal{G}(\mathbf{x}_{k-1}^{(p)})$.
- Apply fitness decay on $f_{\rm pb}^{(p)}$ to de-emphasize the past information, i.e., $f_{\rm pb}^{(p)} \Leftarrow f_{\rm pb}^{(p)} - \Delta f$.
- Evaluate the fitness for the predicted $\mathbf{x}_{k|k-1}^{(p)}$ using $\mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)})$, compare with $f_{\rm pb}^{(p)}$ and update $\mathbf{x}_{\rm pb}^{(p)}$ and $f_{\rm pb}^{(p)}$ using (28) and (29).

End

- Obtain global best particle using $\mathbf{x}_{\rm gb} = \arg\max_{\mathbf{x}_{\rm pb}^{(p)}} f_{\rm pb}^{(p)}$.

For each particle:

- Apply swarm intelligence move to obtain $\mathbf{x}_k^{(p)}$ using

$$\begin{aligned}\mathbf{x}_k^{(p)} =& \mathbf{x}_{k|k-1}^{(p)} + \chi\big[\varphi_1\boldsymbol{\gamma}_1 \odot (\mathbf{x}_{\rm pb}^{(p)} - \mathbf{x}_{k|k-1}^{(p)}) \\ &+ \varphi_2\boldsymbol{\gamma}_2 \odot (\mathbf{x}_{\rm gb} - \mathbf{x}_{k|k-1}^{(p)})\big]\end{aligned}$$

- Evaluate the fitness for $\mathbf{x}_k^{(p)}$ using $\mathcal{F}_k(\mathbf{x}_k^{(p)})$, compare with $f_{\rm pb}^{(p)}$ and update $\mathbf{x}_{\rm pb}^{(p)}$ and $f_{\rm pb}^{(p)}$ using (32) and (33).
- Assign $\mathbf{x}_k^{(p)}$ by the best-fit particle using $\mathbf{x}_k^{(p)} \Leftarrow \mathbf{x}_{\rm pb}^{(p)}$.

End

Output: $\{\mathcal{X}_k^{(p)}\}_{p=1}^{N_p} = \{\mathbf{x}_k^{(p)},\ \mathbf{x}_{\rm pb}^{(p)},\ f_{\rm pb}^{(p)}\}_{p=1}^{N_p}$.

Table II summarizes the proposed SWIPF algorithm. Given the estimated $\{\mathcal{X}_k^{(p)},\ w_k^{(p)}\}_{p=1}^{N_p}$, the state estimate at the kth frame is then given by

$$\widehat{\mathbf{x}}_k = \sum_{p=1}^{N_p} w_k^{(p)}\mathbf{x}_k^{(p)}. \tag{37}$$

The proposed SWIPF incorporates the resampling step similar to conventional SIRPF. When the effective sample size is below a threshold, i.e., $N_{\rm eff} < N_{\rm thr}$, $\{\mathcal{X}_k^{(p)}\}_{p=1}^{N_p}$ is resampled. This step again improves the particle sampling since any ineffective particles will be replaced by other effective particles. While mathematical analysis to determine the biasness of the state estimate is beyond the scope of this paper, Monte Carlo simulations suggest that the state estimate generally exhibits low error as will be shown in Sec. V.

B. Application on Alternating Source Tracking

To apply the proposed SWIPF for the alternating source tracking scenario, the monotonic function $\mathcal{M}(\cdot)$ in (25) is to be defined. Since each $\widehat{\tau}_{k,m}$ in $\mathbf{z}_k$ spanning over a small TDOA range (in ms), the density value of $p(\mathbf{z}_k|\mathbf{x})$ in (23) potentially has a high permissible range. We therefore employ a logarithmic function for $\mathcal{M}(\cdot)$ to achieve a lower permissible range. The fitness function in (25) can therefore be rewritten

TABLE II: Summary of the SWIPF algorithm.

For all particles, initialize $\mathcal{X}_0^{(p)} = \{\mathbf{x}_0^{(p)},\ \mathbf{x}_{\rm pb}^{(p)},\ f_{\rm pb}^{(p)}\}$ where $\mathbf{x}_0^{(p)} \sim \mathcal{N}(\widehat{\mathbf{x}}_0, \mathbf{I}_{2\times 2})$, $\widehat{\mathbf{x}}_0$ is initial state, $\mathbf{I}_{2\times 2}$ is identity matrix, $\mathbf{x}_{\rm pb}^{(p)} = \mathbf{x}_0^{(p)}$ and $f_{\rm pb}^{(p)} = -\infty$. The weight is initialized as $w_k^{(p)} = 1/N_p$.

For the kth frame:

Input: $\{\mathcal{X}_{k-1}^{(p)},\ w_{k-1}^{(p)}\}_{p=1}^{N_p}$

1) *Particle sampling*: obtain optimally sampled $\{\mathcal{X}_k^{(p)}\}_{p=1}^{N_p}$ using Table. I.
2) *IS density estimation*: compute $p^{(\rm IS)}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k)$ using kernel density estimation as described in (36), i.e.,
$$p^{(\rm IS)}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k) = \frac{1}{N_p}\sum_{i=1}^{N_p} \mathcal{K}_{\mathbf{H}}(\mathbf{x}_k^{(p)} - \mathbf{x}_k^{(i)}).$$
3) *Weight update*: compute $p(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)})$ using (19), obtain $p(\mathbf{z}_k|\mathbf{x}_k^{(p)}) = \mathcal{M}^{-1}(f_{\rm pb}^{(p)})$ and then update $w_k^{(p)}$ using
$$w_k^{(p)} \propto w_{k-1}^{(p)} \frac{p(\mathbf{z}_k|\mathbf{x}_k^{(p)})\ p(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)})}{p^{(\rm IS)}(\mathbf{x}_k^{(p)}|\mathbf{x}_{k-1}^{(p)}, \mathbf{z}_k)},$$
followed by a normalization step $w_k^{(p)} \Leftarrow w_k^{(p)}(\sum_{i=1}^{N_p} w_k^{(i)})^{-1}$.
4) *Resampling*: resample $\{\mathcal{X}_k^{(p)}\}_{p=1}^{N_p}$ if the effective sample size is below a threshold, i.e., $N_{\rm eff} < N_{\rm thr}$, where $N_{\rm eff} = (\sum_{p=1}^{N_p}(w_k^{(p)})^2)^{-1}$.

Output: $\{\mathcal{X}_k^{(p)},\ w_k^{(p)}\}_{p=1}^{N_p}$

End

as
$$\mathcal{F}_k(\mathbf{x}) \triangleq \ln\{p(\mathbf{z}_k|\mathbf{x})\}, \quad \mathbf{x} \in \mathcal{D}. \tag{38}$$
In addition, for the considered talker tracking application, a fitness decay amount $\Delta f = \alpha(f_{\max} - f_{\rm v})$ is applied for (27) such that
$$f_{\rm pb}^{(p)} \Leftarrow f_{\rm pb}^{(p)} - \alpha(f_{\max} - f_{\rm v}), \tag{39}$$
where $f_{\max}$ is the maximum of the fitness values evaluated during an initial calibration period in which the desired source is active, $f_{\rm v}$ is the average fitness value during an initial period in which the desired source is silent, and α is a control parameter that regulates the decay speed. Therefore, for an illustrative case where $\alpha = 0.2$, the fitness value can reduce from $f_{\max}$ to $f_{\rm v}$ within five iterations. The value of α needs to be empirically determined according to the speed of the moving source — a lower value of α indicates a slower memory decay.

The behavior of the proposed SWIPF for the alternating source tracking can be described as follows. The particles are first propagated using $\mathbf{x}_{k|k-1}^{(p)} = \mathcal{G}(\mathbf{x}_{k-1}^{(p)})$ where the alternating-source model $\mathcal{G}(\cdot)$ defined in (15) and (17) is used.

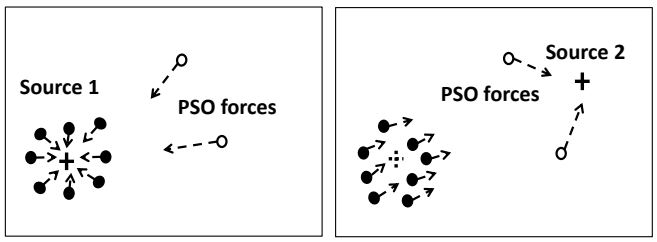

Fig. 5: The proposed swarm intelligence based particle sampling. (a) When the first talker is active, both drifted particles and newly initiated particles are driven by the PSO force. (b) When the second talker is active, the newly initiated particles become the early members to detect the new active source.

In practice, a random sample r is drawn from a uniform distribution and if $r > P_{\rm alt}$, indicating that alternation does not occur, $\mathbf{x}_{k|k-1}^{(p)}$ is obtained by using (15). Otherwise, if $r \le P_{\rm alt}$, $\mathbf{x}_{k|k-1}^{(p)}$ is obtained from (17). Figure 5 illustrates the particle propagation process based on the proposed source-dynamic model. From a particle sampling point of view, $P_{\rm alt}$ regulates the division of particles. Therefore, if $P_{\rm alt} = 0.2$, 80% of the particles (according to (15) and shown by the solid dots) will be perturbed within the neighborhood of its previous estimated source position. The remaining 20% of the particles (according to (17) and shown by the hollow circles) will be re-initialized to facilitate the detection of any new active source. As opposed to the use of single-source motion model, this propagation step increases the likelihood of detecting any new active source while preserving the capability of tracking an existing source.

Although the use of alternating-source model $\mathcal{G}(\cdot)$ improves particle prediction, model mismatch may still occur. Firstly, an alternation is assumed with probability $P_{\rm alt}$ for every frame and this assumption may not be valid when the source is continuously active. As shown in Fig. 5 (a), some of the particles $\mathbf{x}_{k|k-1}^{(p)}$ (denoted by the hollow circle) are uniformly re-initiated resulting in them being far away from an existing source. In this case, the memory mechanism will be enabled to compensate the model mismatch effect: since the wrongly predicted $\mathbf{x}_{k|k-1}^{(p)}$ is not in line with the location information obtained from $\mathbf{z}_k$, $\mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}) < f_{\rm pb}^{(p)}$ and the historical information $\mathbf{x}_{\rm pb}^{(p)}$ and $f_{\rm pb}^{(p)}$ (from (28) and (29)) will be preserved. The swarm intelligence will then move the wrongly predicted $\mathbf{x}_{k|k-1}^{(p)}$ back to its previous best-fit positions in (30). Otherwise, if alternation has occurred, the uniformly re-initiated particles (shown by the hollow circle in Fig. 5 (b)) become the early members to detect the new source. For these particles, $\mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}) \ge f_{\rm pb}^{(p)}$ and the historical information will be updated according to (28) and (29). The interaction mechanism will share the information among all the particles using $\mathbf{x}_{\rm gb}$ and converge the particles to the global best-fit position given by (30).

Finally, for frames with low SIR, SNR and/or SRR, $p(\mathbf{z}_k|\mathbf{x})$ and hence $\mathcal{F}_k(\mathbf{x})$ will not provide accurate source-location information for particles to converge. Knowing that the current erroneous $\mathbf{z}_k$ will cause the newly evaluated fitness $\mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)})$ to be lower than the previous best-fit fitness de-

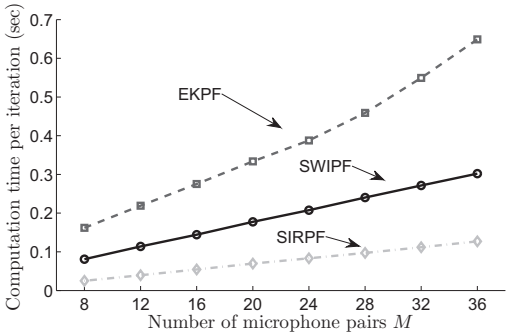

Fig. 6: Averaged computation time per iteration versus number of microphone pairs M. The number of particles is fixed at $N_p = 100$.

rived from the reliable measurement, i.e., $\mathcal{F}_k(\mathbf{x}_{k|k-1}^{(p)}) < f_{\text{pb}}^{(p)}$, the memory mechanism will also be exploited and the particles will remain at their previous best-fit state estimates.

C. Algorithmic Complexity

This section compares the computational complexity of the proposed SWIPF with that of SIRPF [13], [18] and EKPF [28]. Since the dimension of state vector is usually fixed for talker tracking application (e.g., $\mathbf{x}_k = [x_k, y_k]$) while the dimension of measurement vector $\mathbf{z}_k = [\tau_{k,1}, \tau_{k,2}, \ldots, \tau_{k,M}]$ varies depending on the number of microphone pairs M to be used, the complexity is therefore estimated with respect to M.

Given N_p particles, the algorithmic complexity per iteration of the proposed SWIPF algorithm is $\mathcal{O}(2N_pM)$. This results from fact that evaluation of the fitness function $\mathcal{F}_k(\cdot)$, i.e., $p(\mathbf{z}_k|\mathbf{x}_k)$, involves complexity $\mathcal{O}(M)$ as in (23), and such fitness function needs to be evaluated twice (in (28)/(29) and (32)/(33), respectively) for each particle per iteration. For the SIRPF algorithm, the algorithmic complexity per iteration is $\mathcal{O}(N_pM)$ because computation of $p(\mathbf{z}_k|\mathbf{x}_k)$ involves complexity of $\mathcal{O}(M)$ similar to (23) for each particle. The EKPF algorithm requires a complexity of $\mathcal{O}(N_pM^3)$ per iteration. This is because a matrix inverse is required for an $M \times M$ measurement covariance matrix in the Kalman filtering step, which generally requires $\mathcal{O}(M^3)$, and such step has to be repeated for every particle. In view of the above, the SIRPF requires the least computation while the EKPF algorithm requires the highest computational load as M increases. The proposed SWIPF algorithm offers a good tradeoff between computational load and performance.

Figure 6 shows the averaged computation time per iteration given a fixed number of particles $N_p = 100$. The computation time is evaluated on Matlab 2013b platform using a desktop PC with a quad-core processor of 3.07 GHz. As expected, the computation time of both SWIPF and SIRPF increases linearly with M. The computation time of the proposed SWIPF is less than that of EKPF. The SWIPF requires 100 ms processing time per iteration for $M = 12$, which is close to a typical frame length of 64 ms (1024 samples when $f_s = 16$ kHz) to 128 ms (2048 samples when $f_s = 16$ kHz), implying that the algorithm is suitable for real-time processing.

V. Simulation and Experiment Results

The performance of SWIPF is evaluated both in simulated environment with room impulse responses (RIRs) generated

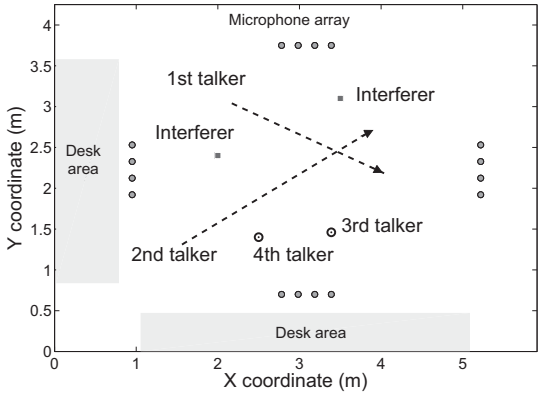

Fig. 7: Room setup for simulation and experiment. Four microphone arrays with total of 16 microphones are employed. Four alternating talkers including two moving talkers and two stationary talkers are active in turns. A silence period is introduced after the third talker becomes silent and before the fourth talker becomes active. A fan noise interferer and a speech interferer are always present during the entire duration.

using the method of images [42], [43] and an actual room environment. The proposed SWIPF algorithm is compared with SIRPF [18] and EKPF tracking algorithms [28]. For SIRPF, both the Langevin process (single-source) [18] and the proposed alternating-source model have been applied to evaluate the particle sampling impoverishment problem. These algorithms are denoted as SIRPF-single and SIRPF-alternating, respectively. In addition, TDOA measurements are considered for SIRPF as in [13] for consistency with other two algorithms. For EKPF, the IS density is estimated using the extended Kalman filter for tracking alternating talkers. Fig. 7 shows a $5.9\ \text{m} \times 4.25\ \text{m} \times\ 2.3\ \text{m}$ room environment where we conducted our experiments. Similar to [28], four microphone arrays with total of 16 microphones are deployed and each array consists of three pairs of adjacent microphones with 0.2 m inter-microphone spacing. Four talkers (two moving and two stationary) are used and they are activated alternatively to generate an alternating-source scenario. A fan noise and an interfering speech signal are present during the entire duration. In addition, a silence period is introduced between the third and fourth talkers' active periods, during which only interferers are present. This is to examine whether the tracking system is able to resume tracking the fourth talker after being interrupted by interferers during the long break. For moving talkers, similar to [13], [18], [28], the speed of the talkers was set to approximately 0.3 m/s. Since a person is expected to move considerably slower indoors compared to outdoors, the speed used was approximately a quarter that of a pedestrian's[2] [44]. Speech signals were obtained from the TIMIT database [45] and were of duration 26 s for both simulation and experiment. All audio data are sampled at 16 kHz. For simulations, synthetic RIRs were also generated using the same room dimension and microphone-source configuration. To simulate the moving talker, RIRs were generated in a manner similar to [13], [18], [22], [28] for discrete talker positions sampled frame-wise along a pre-defined path trajectory (code available online [42], [43]). The speech signals

[2]Additional simulations show that an increase in walking speed will reduce the performance of SIRPF and EKPF while the impact for the proposed SWIPF algorithm is modest due to SWIPF's high rate of convergence

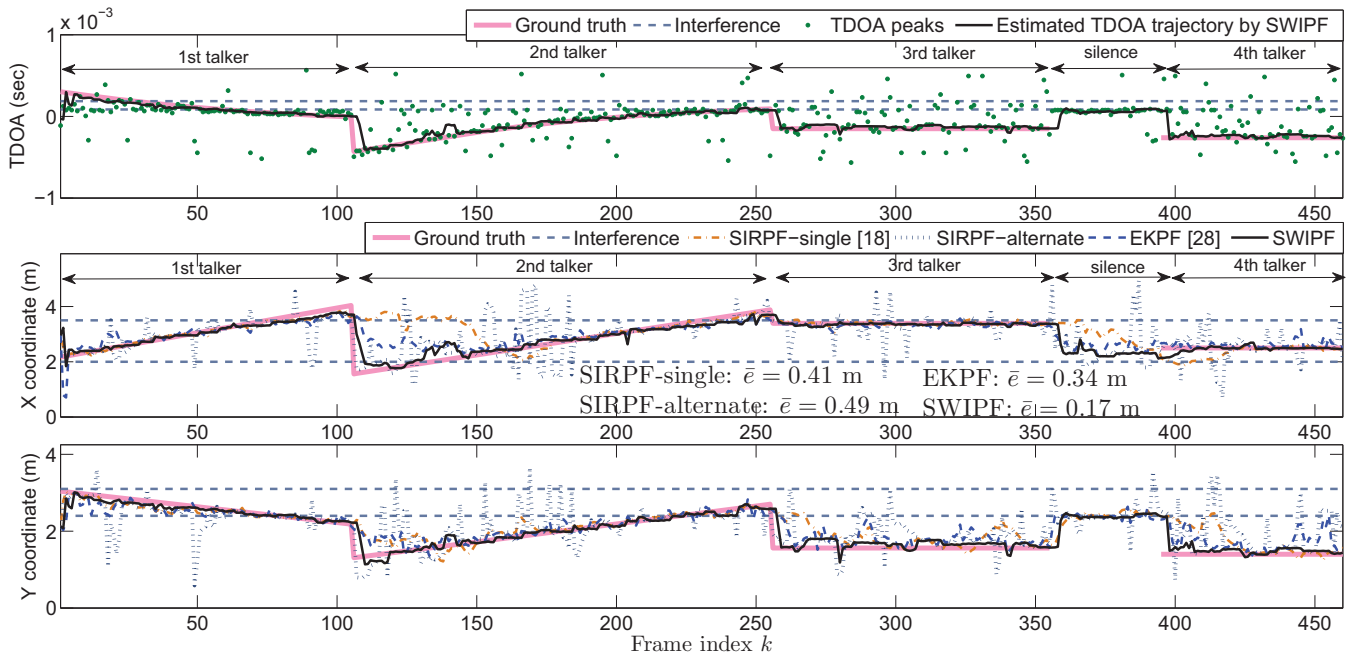

Fig. 8: A single-trial result using the simulated data when $T_{60} = 0.5$ s, SNR = 10 dB and SIR = 9 dB for each interferer. A silence period (during frames $355 < k < 395$) is introduced during which the desired talker is absent and only two interferers are present. (a) measured TDOA versus estimated TDOA trajectory by SWIPF for one microphone pair; (b) x coordinate of the estimated trajectory; (c) y coordinate of the estimated trajectory.

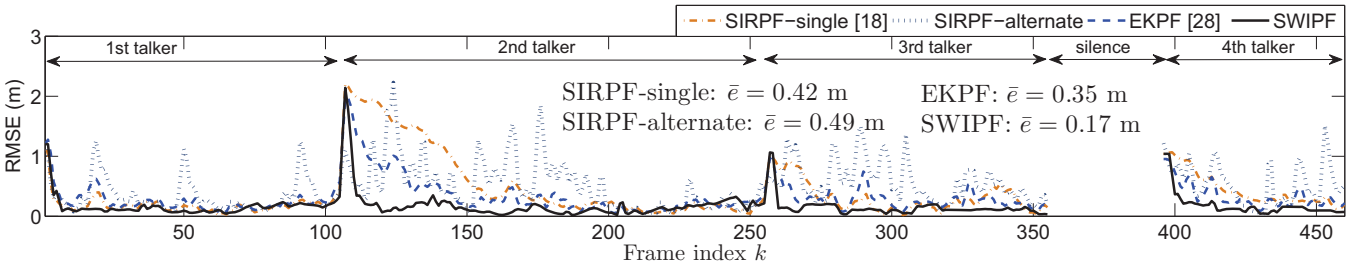

Fig. 9: RMSE over 50 Monte Carlo trials where particle propagation and swam moves are realized differently in probabilistic approach for each trial. The environment is simulated with $T_{60} = 0.5$ s, SNR = 10 dB and SIR = 9 dB for each interferer.

were then convolved with these RIRs frame-wise to generate signals corresponding to moving talkers.

For the algorithms, TDOA measurements are computed using a frame size of 1024 samples and no overlapping is applied between consecutive frames. The frequency range of interest is set as $\Omega = \{2\pi \times 100 \text{ Hz} \leq \omega \leq 2\pi \times 6 \text{ kHz}\}$ corresponding to frequencies where speech components mainly exist [46]. The PF resampling threshold is set as $N_{\text{thr}} = 37.5$ [13], [18]. In addition, no prior knowledge of initial talker location is assumed for all three algorithms and the initial state $\widehat{\mathbf{x}}_0$ in Table II is initialized at the center of the room. The parameters of the proposed alternating source-dynamic model was set as $\sigma^2_{\mathbf{u}} = 0.01$ and $P_{\text{alt}} = 0.2$. For SWIPF, $\sigma^2_{\tau} = 5 \times 10^{-5}$ is used for (20) and the fitness decay factor was $\alpha = 0.2$. No priority is assumed for the acceleration on individual influence and social influence, i.e., $\varphi_1 = \varphi_2 = 2.1$ are applied.

The tracking performance is evaluated using

$$e_k = ||\widehat{\mathbf{x}}_k - \mathbf{x}^{\text{s}}_k||, \tag{40}$$

where $\widehat{\mathbf{x}}_k$ is the estimated position, and $\mathbf{x}^{\text{s}}_k$ is the true talker position. The average tracking error $\bar{e} = \frac{1}{K}\sum_{k=1}^{K} e_k$ quantifies the performance across all audio frames, where K is the total number of frames.

A. Simulation Results

Tracking results for a single trial with reverberation time $T_{60} = 0.5$ s, SIR = 9 dB, and SNR = 10 dB are presented in Fig. 8, where the SNR is computed by involving only the background diffuse noise without the interferers. Figure 8 (a) shows the maximum peaks in GCC function (i.e., the measurements) for one pair of microphones and the tracked TDOA trajectory that obtained from the talker position trajectory estimated by SWIPF. The solid bold line denotes the TDOA trajectory corresponding to the actual (active) talker and the two dashed horizontal lines denote the ones with stationary fan and speech interferers. Three alternation instances between the talkers are introduced at the 106th , 256th and 395th frame (after the silence break). It can be observed that the TDOA measurements (shown by the dots) correspond to the desired dominant talker for most of the frames due to his/her highest energy. The occasional outliers (e.g., during frames $250 < k < 350$) are mainly caused by the low SIR. The SWIPF algorithm tracks the desired talker and exhibits robustness to outliers by taking into account both the signal

energy and historical activeness information. It can also be observed that during the silence period between the 355th and 395 frames, the tracking system inevitably tracks the interferer due to the absence of the desired talker. The SWIPF algorithm resumes tracking the fourth talker immediately after he/she becomes active since the corresponding measurements have been obtained. Note that to distinguish the active and silent phases of the speech source, algorithms such as described in [47], [48] can be considered such that the proposed tracking algorithm will be only enabled when the presence of speech is detected.

Figures. 8 (b) and (c) compare tracking results in terms of position trajectory estimates from a single trial. The SIRPF-single algorithm (dash-dotted line) requires the longest transition time to track the new active talker after alternation is introduced. This is caused by the mismatch between the assumed single-source model and the actual alternation between talkers. When another talker is activated at a new position, the particles still remain in the neighborhood of the previous talker position. Higher number of iterations is therefore required to resample the particles into an area where the active talker is located. For the same reason, long transition time is required to resume tracking the fourth talker after being interrupted by the interference during the silence period. The SIRPF-alternating algorithm (dotted line) shows the effectiveness of the proposed alternating-source model in terms of the transition time. This algorithm, however, exhibits high error during periods where only a single talker is active (i.e., no alternation). This is because the alternating-source model assumes possibility of alternation at each iteration and this assumption is violated during periods when a single talker is continuously moving/stationary. When this occurs, particles are poorly sampled and they converge toward a non-existent "new" talker caused by interferers, noise and/or reverberation. The EKPF algorithm (dashed line) is able to strike a balance between transition time and robustness in continuous tracking period. This is because EKPF approximates a better IS density by taking into account information derived from measurements. However, particles may not be appropriately sampled in EKPF due to the incorporation of erroneous measurements in frames with low SNR, SIR and/or SRR. The resultant trajectory therefore frequently deviates from the ground truth. The proposed SWIPF algorithm (solid line) achieves the shortest transition time as well as the least fluctuation. This is because SWIPF utilizes the proposed alternating source-dynamic model and interaction mechanism to achieve higher rate of convergence to the new active talker. In addition, the memory mechanism is exploited to compensate the mismatch effect brought by the alternating-source model and disturbance from interference, noise and reverberation. The particles remain at their best-fit positions when no alternation occurs, leading to reduced fluctuation.

Figure 9 shows the root-mean-square error (RMSE) over fifty Monte Carlo simulation trials where, for each trial, sampling of particles is realized differently in terms of particle propagation from the probabilistic source-dynamic model and random swarm moves. The room and source configuration setup remain the same as the previous simulation. Three high

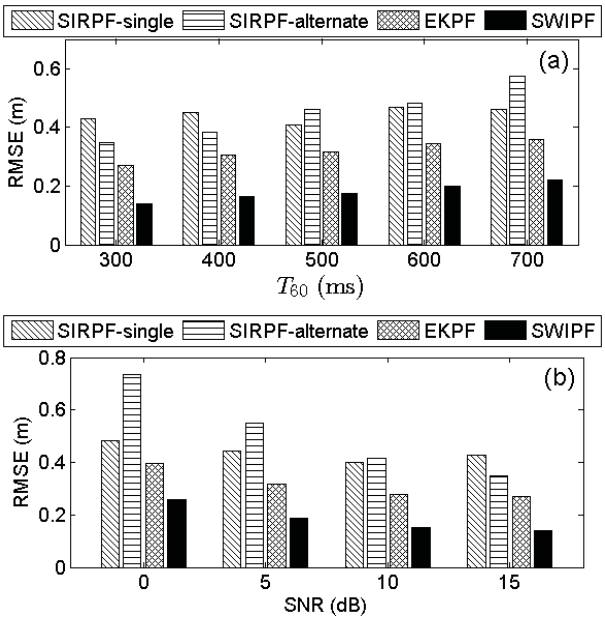

Fig. 10: Variation of RMSE for (a) different reverberation time at SNR=15 dB and SIR = 9 dB for each interferer, and (b) different SNRs at $T_{60} = 0.3$ s and SIR = 9 dB for each interferer.

RMSE peaks indicate the discontinuity in RMSE when the alternation is introduced. The RMSE between the 355th and 395th frames is not plotted due to absence of the speech source. These multiple trials validate the performance of SWIPF compared to the other three algorithms. The SWIPF algorithm requires less than 0.64 s (10 frames with frame length of 1024 samples in 16 kHz sampling rate) on average to successfully switch to the second source. This implies that that SWIPF is expected to perform well provided that the interval between any alternation occurrences is beyond 0.64 s, compared to 2.56 and 4.48 s for the EKPF and SIRPF-single algorithms, respectively. In addition, the tracking error remains low for the period after the 256th frame, indicating that SWIPF is also suitable for stationary sources.

Figure 10 (a) shows how the RMSE varies with reverberation time when SNR = 15 dB and SIR = 9 dB. This result shows that the tracking error increases with reverberation time, as expected. Both SIRPF-single and SIRPF-alternating algorithms exhibit high tracking error due to the model mismatch problem. Comparatively, the EKPF algorithm achieves lower tracking error. The proposed SWIPF algorithm achieves the lowest tracking error in all the cases being considered since it utilizes memory information in frames with low SRR. Figure 10 (b) shows how the RMSE varies with SNR for $T_{60} = 0.3$ s and SIR = 9 dB. The RMSE increases with reducing SNR for all algorithms as expected. As before, both SIRPF algorithms suffer from the highest tracking error due to particle sampling impoverishment. The SWIPF algorithm achieves the lowest tracking error in a noisy environment.

B. Experiment Results

An experiment was conducted in an actual room environment (see Fig. 7). Four speech signals were generated, in a manner similar to [19], [28], using loudspeakers to

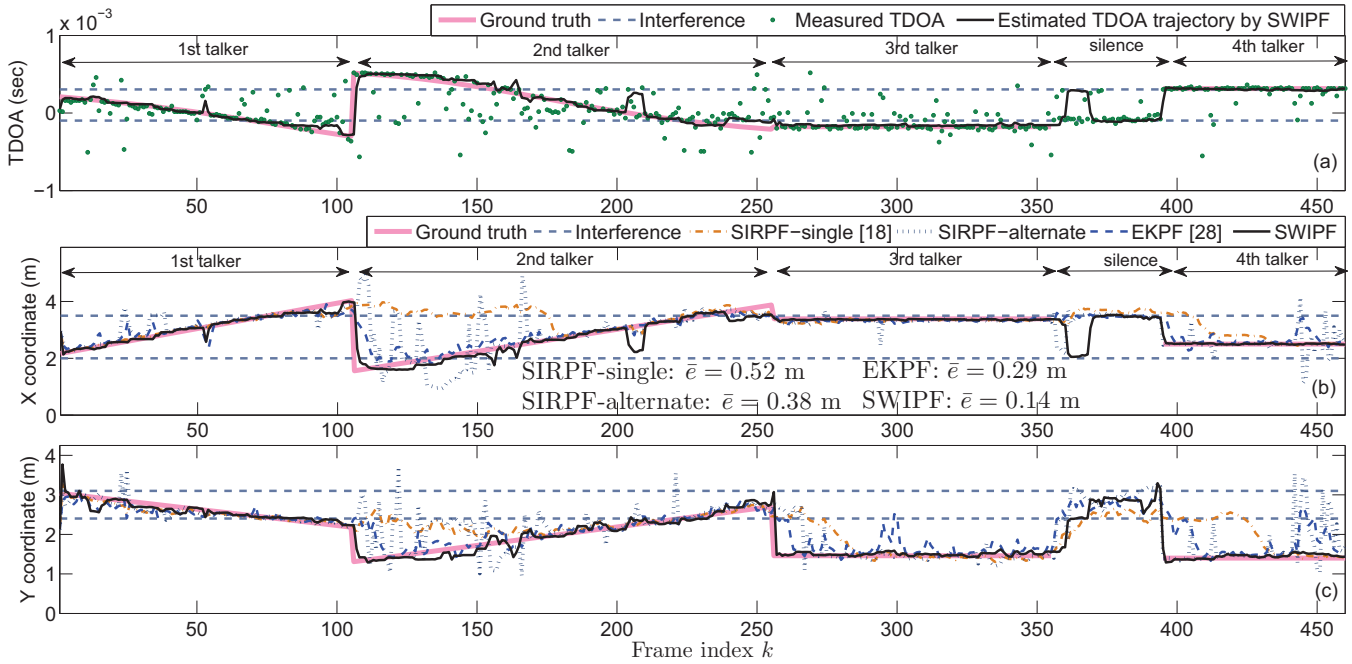

Fig. 11: A single-trial result using recorded data from an actual environment with an estimated $T_{60} = 0.35$ s and SNR = 7 dB, SIR = 12 dB and 15 dB for the interfering speech signal and fan, respectively. A silence period (during frames $355 < k < 395$) is introduced during which the desired talker is absent and only two interferers are present. (a) measured TDOA versus estimated TDOA trajectory by SWIPF for one microphone pair; (b) x coordinate of the estimated trajectory; (c) y coordinate of the estimated trajectory.

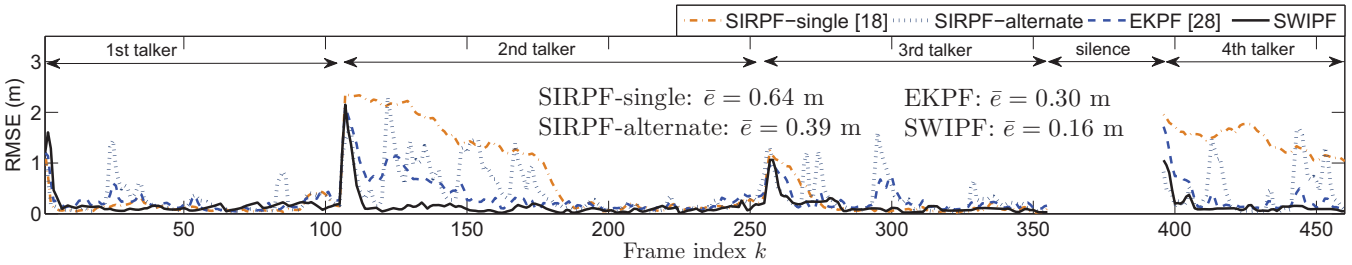

Fig. 12: RMSE over 50 Monte Carlo trials where particle propagation and swam moves are realized differently in probabilistic approach for each trial. Signals are recorded in an actual environment with an estimated $T_{60} = 0.35$ s and SNR = 7 dB, SIR = 12 dB and 15 dB for the interfering speech signal and fan, respectively.

simulate two moving and two stationary talkers. The moving loudspeakers were faced in the direction of motion, while the stationary loudspeaker was faced towards the center of the room. Additionally, both interferers were also played through the loudspeakers towards the center of the room. For this experiment, $T_{60} \approx 0.35$ s and the background noise without any interference was estimated to be at SNR$\approx$ 7 dB. The SIR was estimated to be 12 dB and 15 dB for the interfering speech and fan noise, respectively. Parameters for the algorithms were configured similar to that described in the simulation setup.

Figure 11 (a) shows the tracking result of SWIPF in terms of TDOA trajectory versus TDOA measurement. The SWIPF algorithm filters out the unreliable measurements and achieves short transition time for alternations and after the silence period. Figure 11 (b) and (c) show tracking results among three algorithms in terms of position trajectory estimation. Similar to that of simulations, the SIRPF-single algorithm fails to "lock on" to the new talker after alternation has occurred. While the SIRPF-alternating algorithm achieves shorter transition time, it suffers from high fluctuations during the period when the talker is continuously active. The EKPF algorithm achieves a shorter transition time than SIRPF-single and less fluctuation than SIRPF-alternating. The proposed SWIPF algorithm achieves the highest tracking accuracy among the considered algorithms in terms of the convergence rate and tracking stability. Similar results can also be found in Monte Carlo trials as in Fig. 12.

VI. Conclusion

An SWIPF algorithm is proposed to track alternating talkers. The proposed algorithm exploits PF and swarm intelligence jointly to achieve optimal particle sampling. As opposed to propagating the particles independently in PF, SWIPF incorporates the interaction mechanism in swarm intelligence to improve the particle convergence to the active talker region. In addition, the memory mechanism enables particles to be retained at the previous best-fit positions when signals are corrupted by interference, noise and reverberation. Simulation and experiment results show that SWIPF can locate and track the alternating talkers with short transition period, resulting in the lowest tracking error compared to EKPF and SIRPF in a noisy and reverberant environment.

REFERENCES

[1] J. Chen Y. Huang and J. Benesty, "Immersive audio schemes," *IEEE Signal Process. Magazine*, vol. 28, pp. 20–32, Jan. 2011.

[2] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[3] A. Marti, M. Cobos, and J. J. Lopez, "Real time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'11)*, May, pp. 2592–2595.

[4] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, vol. 1, Springer Science & Business Media, 2008.

[5] K. Wu and A. W. H. Khong, "Sound source localization and tracking," in *Context Aware Human-Robot and Human-Agent Interaction*, pp. 55–78. Springer, 2016.

[6] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *ASSP Magazine, IEEE*, vol. 5, no. 2, pp. 4–24, 1988.

[7] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, Nov. 2007.

[8] J. Dmochowski, J. Benesty, and S. Affes, "Linearly constrained minimum variance source localization and spectral estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1490–1502, Nov. 2008.

[9] J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP J. Adv. Signal Process.*, vol. 2006, 2006.

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

[11] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, 2000.

[12] X. Zhong, *A Bayesian framework for multiple acoustic source tracking*, Ph.D. thesis, The University of Edinburgh, 2010.

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

[14] S. Gannot and T. G. Dvorkind, "Microphone array speaker localizers using spatial-temporal information," *EURASIP J. on Applied Signal Process. (special issue on microphone arrays)*, vol. 2006, pp. 1–17, 2006.

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

[16] B. Ristic, S. Arulampalm, and N. J. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*, Artech House Publishers, 2004.

[17] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'01)*, 2001, pp. 3021–3024.

[18] E. A. Lehmann and A. M. Johansson, "Particle filter with integrated voice activity detection for acoustic source tracking," *EURASIP J. on Adv. Signal Process.*, vol. 2007, 2007.

[19] 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, Lang. Process.*, vol. 19, no. 6, pp. 1540–1555, Aug. 2011.

[20] A. Brutti and F. Nesta, "Tracking of multidimensional TDOA for multiple sources with distributed microphone pairs," *Computer Speech & Language*, vol. 27, no. 3, pp. 660–682, 2013.

[21] X. Zhong and J.R. Hopgood, "A time–frequency masking based random finite set particle filtering method for multiple acoustic source detection and tracking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2356–2370, 2015.

[22] A. Masnadi-Shirazi and B.D. Rao, "An ICA-SCT-PHD filter approach for tracking and separation of unknown time-varying number of sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 828–841, 2013.

[23] M. F. Fallon and S. J. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1409–1415, 2012.

[24] A. Quinlan and F. Asano, "Tracking a varying number of speakers using particle filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'08)*, 2008, pp. 297–300.

[25] Y. Oualil and D. Klakow, "Multiple concurrent speaker short-term tracking using a Kalman filter bank," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'14)*. IEEE, 2014, pp. 1444–1448.

[26] T. Gehrig and J. McDonough, "Tracking multiple speakers with probabilistic data association filters," in *Proc. Classification of Events, Activities and Relationships (CLEAR)*. Springer, 2006, pp. 137–150.

[27] W. K. Ma, B. N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: a random finite set approach," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3291–3304, 2006.

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

[29] X. Zhong and J.R. Hopgood, "Nonconcurrent multiple speakers tracking based on extended Kalman particle filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'08)*, 2008, pp. 293–296.

[30] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization," *Swarm intelligence*, vol. 1, no. 1, pp. 33–57, 2007.

[31] A. P. Engelbrecht, *Fundamentals of Computational Swarm Intelligence*, vol. 1, Wiley Chichester, 2005.

[32] X. Zhang, W. Hu, W. Qu, and S. Maybank, "Multiple object tracking via species-based particle swarm optimization," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1590–1602, 2010.

[33] M. Thida, H. Eng, D. N. Monekosso, and P. Remagnino, "A particle swarm optimisation algorithm with interactive swarms for tracking multiple targets," *Applied Soft Computing*, vol. 13, no. 6, pp. 3106–3117, 2013.

[34] R. Parisi, P. Croene, and A. Uncini, "Particle swarm localization of acoustic sources in the presence of reverberation," in *Proc. IEEE Int. Symposium on Circuits and Systems. (ISCAS 2006)*. IEEE, 2006, pp. 4739–4742.

[35] E. Antonacci, D. Riva, A. Sarti, M. Tagliasacchi, and S. Tubaro, "Tracking of two acoustic sources in reverberant environments using a particle swarm optimizer," in *Proc. IEEE Int. Conf. Adv. Video and Signal Based Surveillance (AVSS' 07)*. IEEE, 2007, pp. 567–572.

[36] M. Hirakawa and K. Suyama, "Multiple sound source tracking by two microphones using pso," in *Proc. Intelligent Sig. Process. and Comm. Systems (ISPACS 2013)*, 2013, pp. 467–470.

[37] X. Zhang, W. Hu, S. Maybank, X. Li, and M. Zhu, "Sequential particle swarm optimization for visual tracking," in *Proc. Computer Vision and Pattern Recognition (CVPR 2008)*. IEEE, 2008, pp. 1–8.

[38] X. Zhang, W. Hu, and S. Maybank, "A smarter particle filter," in *Asian Conference on Computer Vision*. Springer, 2009, pp. 236–246.

[39] M. P. Wand and M. C. Jones, *Kernel Smoothing*, CRC Press, 1994.

[40] C. Musso, N. Oudjane, and F. LeGland, "Improving regularised particle filters," in *Sequential Monte Carlo Methods in Practice*, pp. 247–271. Springer, 2001.

[41] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, vol. 26, CRC press, 1986.

[42] E. A. Lehmann, "Room impulse response generator," www.eric-lehmann.com, (Accessed: 03/04/2016).

[43] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 269–277, July 2008.

[44] K. Aspelin, "Establishing pedestrian walking speeds," *Portland State University*, pp. 5–25, 2005.

[45] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgrena, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Philadelphia, PA, 1993.

[46] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time Processing of Speech Signals*, Wiley-IEEE Press, 2000.

[47] K. Wu, S. T. Goh, and A. W. H. Khong, "Speaker localization and tracking in the presence of sound interference by exploiting speech harmonicity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'13)*, 2013.

[48] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Process. Letters*, vol. 20, no. 3, pp. 197–200, 2013.

Kai Wu received his B.Eng. degree in electronic and information engineering from University of Electronic Science and Technology of China in 2010. He is now a Ph.D. candidate in the field of information engineering in Nanyang Technological University (NTU), Singapore, with a research topic on the algorithm development for talker localization and tracking in room environment. He was also an R&D engineer for audio signal processing applications in Panasonic R&D Center Singapore from 2015 to 2017. He is currently a research scientist in Agency for Science, Technology and Research (A*STAR), Singapore. His research interests involve developing machine learning based algorithms on statistical signals and data.

V. G. Reju received the B.Tech. degree in electrical and electronics engineering from University of Kerala, India, in 1992, M.Tech. degree in electronics engineering from Cochin University of Science and Technology, India, in 1994 and Ph.D. from Nanyang Technological University, Singapore in 2010. He was with Cochin University of Science and Technology from 1995 to 2002 as a Lecturer. He is currently a Senior Research Fellow at Centre for Infocomm Technology, Nanyang Technological University. His research interests include array signal processing and blind source separation.

Andy W. H. Khong is currently an Associate Professor in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. Prior to that, he obtained his Ph.D. ('02-'05) from the Department of Electrical and Electronic Engineering, Imperial College London, after which he also served as a research associate ('05-'08) in the same department. He obtained his B.Eng. ('98-'02) from the Nanyang Technological University, Singapore. His research interest includes adaptive filtering, tangible human-computer interfaces, speech enhancement, acoustic source localization and tracking. More recently, he is working in the area of machine learning and data mining applied to education data. Andy was a visiting professor at UIUC in 2012 under the Tan Chin Tuan Fellowship and is the author/co-author of two "Best Student Paper Awards" paper.

Shu Ting Goh received his Ph.D. degree at the Mechanical Engineering Engineering Mechanics Department, Michigan Technological University in 2012. He is currently a Principle Engineer in Satellite Technology and Research Centre, National University of Singapore. He previously worked in Satellite Research Centre, Nanyang Technological University as Research Fellow. He was the mission planner for both VELOX-I and VELOX-II nanosatellites. He currently works for satellite mission orbit design and analysis for upcoming nanosatellite development. His research interests include the data assimilation, attitude determination, and Kalman filter based parameter estimation, target tracking, and satellite navigation.