Acoustic Source Tracking in Reverberant Environment Using Regional Steered Response Power Measurement

Kai Wu and Andy W. H. Khong
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.
E-mail: wu0001ai@e.ntu.edu.sg; andykhong@ntu.edu.sg

***Abstract*—Acoustic source localization and tracking using a microphone array is challenging due to the presence of background noise and room reverberation. Conventional algorithms employ the steered response power (SRP) as the measurement function in a particle filter based tracking framework. The particle weight is updated according to a pseudo-likelihood derived from the SRP value of each particle position. The performance of this approach reduces in a noisy and reverberant environment. In this paper, instead of evaluating the SRP value for each discrete particle position, we propose to apply a regional SRP beamformer which takes into account a circular region centered on each particle position, in order to provide a more robust particle likelihood evaluation. In addition, a proper mapping function is proposed to transform the regional SRP value to the likelihood. Simulation results show that the proposed method achieves robustness in tracking a speech source in a noisy and reverberant environment.**

***Index Terms*—Acoustic localization and tracking, particle filter, steered response power, microphone array**

I. Introduction

Acoustic source localization and tracking (ASLT) involves estimating the position of an acoustic source using an array of distributed microphones. Recently, ASLT has become an active research area for applications including teleconferencing, automatic camera steering and surveillance. Localizing and tracking a speech source in an enclosed environment, however, is challenging due to the presence of background noise, room reverberation, sound interference and non-stationarity of the speech signal [1]. Therefore, developing a robust localization and tracking algorithm is necessary for real applications under an adverse environment.

ASLT algorithms aim to exploit the relative temporal/spatial information of the microphone received signals given the array geometry. In general, localization algorithms can be classified into two categories: single-step and dual-step approaches. The single-step approach estimates the source position directly by scanning a synthetic beamformer across all possible source locations and finding the maximum power corresponding to the source position estimate [2]. The dual-step approach, on the other hand, estimates the time-difference-of-arrival (TDOA) information across all microphone pairs in the first step [3]. These TDOAs are then used to perform localization in the second step by using a mapping from the TDOAs to the source location estimate [4].

One of the disadvantages of the above approaches is that the localization is performed independently across each time frame. Recently, the Bayesian approach which takes into account the temporal consistency of localization measurements by incorporating the source-dynamic model has been proposed [5]. The particle filter (PF), which does not require the need to satisfy linearity and Gaussianity assumptions, is one such approach that has been widely used for acoustic source tracking [6]. In PF, the source position at each time frame is defined by a state vector and propagated according to a source-dynamic model. The posterior probability density function (pdf) of the state vector is then updated by the measurement at current time frame. It was observed that the steered response power (SRP) beamformer can be used as a measurement function and it achieves better performance than TDOA-based measurement [7]. Instead of evaluating the SRP over the whole region, the PF constrains the estimation to within a relatively small number of positions (the particle set.) Such technique is often referred to as the *pseudo-likelihood* approach [7].

Although the pseudo-likelihood approach has been widely adopted in recent literature [8], [9], it still suffers from the effect of background noise and reverberation. In this paper, we propose a new PF framework which incorporates a regional SRP as its measurement function. Instead of evaluating the SRP for each discrete particle position, the proposed method takes into account a circular region centered around each particle position [10] so as to provide a more comprehensive evaluation of the likelihood function. The regional SRP value is used to compute the likelihood via a nonlinear mapping. As opposed to [10], the proposed method takes into account the temporal consistency of the source position and incorporates a source-dynamic model in the tracking scenario. Simulation results show that the proposed method achieves a performance that is more robust than that proposed in [8], [10] in a noisy and reverberant environment.

II. Review of PF Based Tracking Approach

A. Particle Filter Framework

In ASLT, the state-space model is used to describe the source position estimation problem in an iterative manner. Given a pre-defined Cartesian coordinate system, the source state vector is defined as $\boldsymbol{\alpha}_k = [x_k, y_k, \dot{x}_k, \dot{y}_k]^T$ at time frame

index k, where the first two elements x_k and y_k define the source position $\mathbf{r}_k = [x_k,\ y_k]$, $\dot{x}_k$ and $\dot{y}_k$ denote the source velocity in x and y direction, respectively. We also define the measurement variable $\mathbf{z}_k = [\widehat{x}_k, \widehat{y}_k]^T$ which contains the prior source position estimate. This variable $\mathbf{z}_k$ may be also defined by TDOA-based approach alternatively [7]. The state-space model can therefore be represented as

$$\boldsymbol{\alpha}_k = g(\boldsymbol{\alpha}_{k-1}, \mathbf{u}_k), \tag{1a}$$

$$\mathbf{z}_k = h(\boldsymbol{\alpha}_k, \mathbf{w}_k), \tag{1b}$$

where $g(\cdot)$ denotes the state-transition process, $\mathbf{u}_k$ is the process noise, $h(\cdot)$ denotes the measurement function, and $\mathbf{w}_k$ is the measurement noise. Similar to [7]–[9], we employ the Langevin process which had been proposed as a source-dynamic model for simulating a realistic human motion. Equation (1a) can then be rewritten as

$$\boldsymbol{\alpha}_k = \begin{bmatrix} 1 & 0 & aT & 0 \\ 0 & 1 & 0 & aT \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & a \end{bmatrix} \boldsymbol{\alpha}_{k-1} + \begin{bmatrix} bT & 0 \\ 0 & bT \\ b & 0 \\ 0 & b \end{bmatrix} \mathbf{u}_k, \tag{2}$$

where $\mathbf{u}_k \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the noise variable, T is the time interval between consecutive frames while $\boldsymbol{\mu} = [0, 0]^T$ and $\boldsymbol{\Sigma} = \mathbf{I}_{2\times 2}$ denote the mean vector and covariance matrix, respectively. The parameters a and b are defined as

$$a = \exp(-\beta T), \tag{3a}$$

$$b = \bar{v}\sqrt{1 - a^2}, \tag{3b}$$

where $\bar{v}$ is the steady-state velocity and β is the rate constant. In this paper, we have used, similar to [8], $\bar{v} = 0.8$ m/s, $\beta = 10$ Hz.

The bootstrap PF is commonly used in ASLT due to its simplicity [6]. Defining p as the particle index and N_p as the total number of particles, the posterior pdf $\Pr(\boldsymbol{\alpha}_k|\mathbf{z}_k)$ is approximated using a set of particles of the state space with associated weights $\{\boldsymbol{\alpha}_k^{(p)}, w_k^{(p)}\}_{p=1}^{N_p}$. Each particle goes through a propagation followed by an update step. The bootstrap PF is summarized in Table I and will be adopted in this paper. The source position estimate $\widehat{\mathbf{r}}_k$ corresponds to the first two elements of the estimated state $\widehat{\boldsymbol{\alpha}}_k$.

TABLE I: Summary of the bootstrap PF.

At time $k-1$, a set of particles $\{\boldsymbol{\alpha}_{k-1}^{(p)}, w_{k-1}^{(p)}\}_{p=1}^{N_p}$ is a discrete representation of $\Pr(\boldsymbol{\alpha}_{k-1}|\mathbf{z}_{k-1})$.

For the kth frame:

1) *Particle propagation*: Propagate each particle through the source-dynamic model described by (2),
$$\boldsymbol{\alpha}_k^{(p)} = g(\boldsymbol{\alpha}_{k-1}^{(p)}, \mathbf{u}_k).$$
2) *Update*: Each particle is then assigned a weight according to its likelihood
$$\widetilde{w}_k^{(p)} = w_{k-1}^{(p)} \Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k^{(p)}),$$
followed by a normalization step $w_k^{(p)} = \widetilde{w}_k^{(p)} (\sum_{i=1}^{N_p} \widetilde{w}_k^{(i)})^{-1}$.
3) *Resampling*: Resample the particles if the effective sample size is below a threshold, $N_{\text{eff}} < N_{\text{t}}$, where $N_{\text{eff}} = (\sum_{p=1}^{N_p} (w_k^{(p)})^2)^{-1}$.
4) *Result*: The particle set $\{\boldsymbol{\alpha}_k^{(p)}, w_k^{(p)}\}_{p=1}^{N_p}$ is obtained for approximation of $\Pr(\boldsymbol{\alpha}_k|\mathbf{z}_k)$. The state estimate at the kth frame is $\widehat{\boldsymbol{\alpha}}_k = \sum_{p=1}^{N_p} w_k^{(p)} \boldsymbol{\alpha}_k^{(p)}$.

B. Steered Response Power Measurement

The key step in bootstrap PF-based acoustic source tracking is to determine the measurement likelihood $\Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k)$ so that a proper weight can be assigned to each particle. A *pseudo-likelihood* approach which incorporates a SRP beamformer as the measurement function has been proposed in [7]. More specifically, the SRP beamformer defines the energy of an assumed (look) position $\mathbf{r}'$ as [2], [11]

$$\mathcal{P}(\mathbf{r}') = \sum_{\omega_l \in \Omega} \left| \sum_{i=1}^{M} W_i(\omega_l) Y_i(\omega_l) e^{\jmath \omega_l \tau_i(\mathbf{r}')} \right|^2, \tag{4}$$

where i is the microphone index, M is the number of microphones, $Y_i(\omega_l)$ is the frequency-domain received signal of the ith microphone, $\omega_l = 2\pi l/L$ is the angular frequency of the lth frequency bin, L is the number of frequency bins, Ω is the frequency range of interest such that $\Omega = [0,\ 6]$ kHz is often chosen for a speech source [9], $\tau_i(\mathbf{r}') = \|\mathbf{r}' - \mathbf{r}_i^{\text{m}}\|_2/c$ is the time-of-arrival from $\mathbf{r}'$ to the ith microphone, c is the speed of sound, and $W_i(\omega_l)$ is a weighting function. The phase transform (PHAT) weighting $W_i(\omega_l) = 1/|Y_i(\omega_l)|$ is commonly used in ASLT due to its robustness to reverberation and noise [8], [12]. In general, the SRP beamformer is employed to scan the assumed source position $\mathbf{r}'$ across the whole surveillance region such that the source position estimate corresponds to that having the maximum power. However, this search process requires high computational complexity for realistic applications.

The pseudo-likelihood PF approach mitigates this drawback based on the concept of "pseudo-likelihood." In PF, the likelihood $\Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k)$ defines the probability of obtaining the measurement $\mathbf{z}_k$ given the state $\boldsymbol{\alpha}_k$. The SRP value, representing the power for each discrete point, can be used as an approximate version of this likelihood during the voiced frame, i.e.,

$$\Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k) = \begin{cases} \mathcal{P}^{\gamma}(\mathbf{r}'_k), & \text{for voiced frame} \\ \mathcal{U}_{\mathcal{D}}(\mathbf{r}'_k), & \text{for unvoiced frame} \end{cases}, \tag{5}$$

where $\mathbf{r}'_k = [x'_k\ y'_k]^T$ represents the first two elements of the state vector $\boldsymbol{\alpha}_k$, $\gamma = 2$ is a control parameter to regulate the fusion of the SRP function to the likelihood [8], and $\mathcal{U}_{\mathcal{D}}(\cdot)$ is the uniform pdf over the considered enclosure domain $\mathcal{D} = \{x_k, y_k | x_{\min} \le x_k \le x_{\max}, y_{\min} \le y_k \le y_{\max}\}$.

By using the pseudo-likelihood PF approach, the SRP evaluation $\mathcal{P}^{\gamma}(\mathbf{r}'_k)$ is thus constrained within a relatively small number of positions (the particle set.) However, this approach still suffers in terms of performance in the presence of background noise and reverberation due to the lack of robustness for the SRP [7], [8]; noise and reverberation may flatten the SRP spatial spectrum and cause the location corresponding to the maximum power to deviate from the true source position.

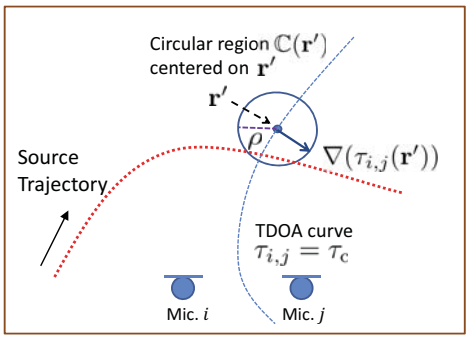

Fig. 1: Regional steered response power for a circle region.

The performance of ASLT algorithm can be improved if a robust measurement function is adopted in the PF tracking framework.

III. The Proposed Method

A. Regional SRP Measurement

We propose to employ a regional SRP beamformer [10] as a measurement function in order to mitigate the effect of reverberation and noise. Due to the energy integration over a square grid centered on an assumed position, the regional SRP beamformer has shown to be more robust than the conventional SRP [11] in a noisy and reverberant environment.

Evaluation of the regional SRP over a square grid proposed in [10] requires the computation of the distance from the center to each boundary along a certain direction. We however consider a circular region centered on each particle, in order to reduce the computational complexity given that the distance from the center to the circular circumference is a constant. Before defining the regional SRP function, we note that the relationship between the conventional SRP function in (4) and the GCC function is given by [13]

$$\begin{aligned}\mathcal{P}(\mathbf{r}') &= \sum_{\omega_l} \left| \sum_{i=1}^{M} W_i(\omega_l) Y_i(\omega_l) e^{j\omega_l \tau_i(\mathbf{r}')} \right|^2 \\ &= 2\pi \sum_{i=1}^{M} \sum_{j=1}^{M} R_{i,j}(\tau_{i,j}(\mathbf{r}')),\end{aligned} \tag{6}$$

where

$$R_{i,j}(\tau_{i,j}(\mathbf{r}')) = \frac{1}{2\pi} \sum_{\omega_l} \Psi_{i,j}(\omega_l) Y_i(\omega_l) Y_j^*(\omega_l) e^{j\omega_l \tau_{i,j}(\mathbf{r}')} \tag{7}$$

is the GCC function between the ith and jth microphones,

$$\begin{aligned}\tau_{i,j}(\mathbf{r}') &= \tau_j(\mathbf{r}') - \tau_i(\mathbf{r}') \\ &= \frac{\|\mathbf{r}' - \mathbf{r}_j^{\mathrm{m}}\| - \|\mathbf{r}' - \mathbf{r}_i^{\mathrm{m}}\|}{c}\end{aligned} \tag{8}$$

is the TDOA between the ith and jth microphones, and

$$\Psi_{i,j}(\omega_l) = \frac{1}{|Y_i(\omega_l) Y_j^*(\omega_l)|} \tag{9}$$

is the PHAT weighting. Expanding (6) and removing the fixed energy terms and symmetries [13], one can define a modified SRP function for a discrete assumed position $\mathbf{r}'$ in terms of the summation of GCC functions:

$$\mathcal{P}^{\mathrm{m}}(\mathbf{r}') = 2\pi \sum_{i=1}^{M} \sum_{j=i+1}^{M} R_{i,j}(\tau_{i,j}(\mathbf{r}')). \tag{10}$$

where the superscript "m" in (10) denotes for the modified SRP function. Equation (10) indicates that instead of using (4), the power at $\mathbf{r}'$ can also be computed from the summation of GCC functions in which the TDOAs are determined by the discrete assumed position.

Now, instead of considering $\mathbf{r}'$, we take into account a circular region $\mathbb{C}(\mathbf{r}')$ centered at $\mathbf{r}'$, as illustrated in Fig. 1. The regional SRP is defined by accumulating the power within $\mathbb{C}(\mathbf{r}')$, i.e.,

$$\mathcal{P}^{\mathrm{c}}(\mathbf{r}') = 2\pi \sum_{i=1}^{M} \sum_{j=i+1}^{M} \sum_{\mathbf{r}'' \in \mathbb{C}(\mathbf{r}')} R_{i,j}(\tau_{i,j}(\mathbf{r}'')), \tag{11}$$

where the superscript "c" denotes for the circular region. It has been shown in [10] that the GCC function for points within a region takes only values in the TDOA range $\tau_{i,j}(\mathbf{r}') \in [\tau_{i,j}^{\mathrm{l}}(\mathbf{r}'), \tau_{i,j}^{\mathrm{h}}(\mathbf{r}')]$ for each microphone pair, where the TDOA range limits $\tau_{i,j}^{\mathrm{l}}(\mathbf{r}'), \tau_{i,j}^{\mathrm{h}}(\mathbf{r}')$ are only determined by the region boundary. In this paper, since we are considering a circular region $\mathbf{r}'' \in \mathbb{C}(\mathbf{r}')$ in (11), $\tau_{i,j}^{\mathrm{l}}(\mathbf{r}'), \tau_{i,j}^{\mathrm{h}}(\mathbf{r}')$ can be determined by the boundary of the circular region. In order to compute these TDOA range limits, we first evaluate the TDOA gradient along which the TDOA exhibits the highest rate of increase. By taking the gradient of (8), the TDOA gradient $\nabla(\tau_{i,j}(\mathbf{r}'))$ at position $\mathbf{r}'$ can be derived as

$$\nabla(\tau_{i,j}(\mathbf{r}')) = [\nabla_{x'}(\tau_{i,j}(\mathbf{r}')), \nabla_{y'}(\tau_{i,j}(\mathbf{r}'))], \tag{12}$$

where $\nabla_{x'}(\cdot) = \partial(\cdot)/\partial x'$ such that

$$\nabla_{x'}(\tau_{i,j}(\mathbf{r}')) = \frac{1}{c}\left(\frac{x' - x_j^{\mathrm{m}}}{\|\mathbf{r}' - \mathbf{r}_j^{\mathrm{m}}\|} - \frac{x' - x_i^{\mathrm{m}}}{\|\mathbf{r}' - \mathbf{r}_i^{\mathrm{m}}\|} \right), \tag{13a}$$

$$\nabla_{y'}(\tau_{i,j}(\mathbf{r}')) = \frac{1}{c}\left(\frac{y' - y_j^{\mathrm{m}}}{\|\mathbf{r}' - \mathbf{r}_j^{\mathrm{m}}\|} - \frac{y' - y_i^{\mathrm{m}}}{\|\mathbf{r}' - \mathbf{r}_i^{\mathrm{m}}\|} \right). \tag{13b}$$

In (13), x' and y' denote the two-dimensional components of $\mathbf{r}'$ while x_i^{m} and y_i^{m} denote the two-dimensional components of the ith microphone location. The lower and upper limits of the TDOA can be computed by considering the product of the gradient magnitude and the distance along the gradient, i.e.,

$$\tau_{i,j}^{\mathrm{l}}(\mathbf{r}') = \tau_{i,j}(\mathbf{r}') - \|\nabla(\tau_{i,j}(\mathbf{r}'))\|\rho, \tag{14a}$$

$$\tau_{i,j}^{\mathrm{h}}(\mathbf{r}') = \tau_{i,j}(\mathbf{r}') + \|\nabla(\tau_{i,j}(\mathbf{r}'))\|\rho, \tag{14b}$$

where ρ is the radius of the circular region. With the obtained TDOA range limits, the regional SRP in (11) can then be evaluated as

$$\mathcal{P}^{\mathrm{c}}(\mathbf{r}') = 2\pi \sum_{i=1}^{M} \sum_{j=i+1}^{M} \sum_{\tau_{i,j}(\mathbf{r}') = \tau_{i,j}^{\mathrm{l}}(\mathbf{r}')}^{\tau_{i,j}^{\mathrm{h}}(\mathbf{r}')} R_{i,j}(\tau_{i,j}(\mathbf{r}')). \tag{15}$$

B. Distribution of Regional SRP Values

The regional SRP value computed from (15) cannot be directly used as a measurement likelihood. We seek for some mapping function $\mathcal{M}(\cdot)$ to map the regional SRP value into the likelihood $\Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k)$ that is within the range of $[0, 1]$.

$$\Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k) = \mathcal{M}(\mathcal{P}^{\mathrm{c}}(\mathbf{r}_k')). \tag{16}$$

In order to develop a proper mapping function, we first analyze the distribution of regional SRP values. Substituting (7) and (9)

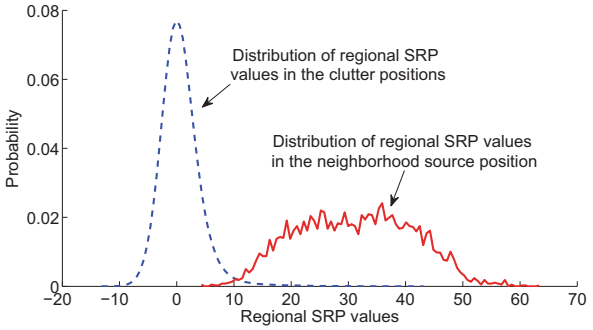

Fig. 2: Distribution of the regional SRP values.

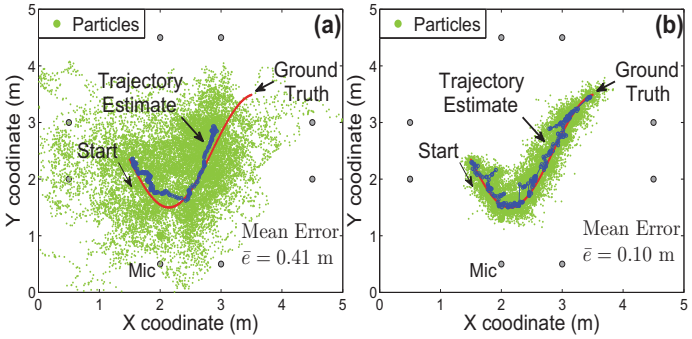

Fig. 3: Comparison of tracking results with $T_{60} = 450$ ms and SNR = 10 dB. (a) Conventional PF-SRP tracking method [8]. (b) Proposed PF-regional SRP tracking method.

into (15), we obtain

$$\mathcal{P}^{\mathrm{c}}(\mathbf{r}') = \sum_{i=1}^{M}\sum_{j=i+1}^{M}\sum_{\tau_{i,j}(\mathbf{r}')=\tau_{i,j}^{1}(\mathbf{r}')}^{\tau_{i,j}^{\mathrm{h}}(\mathbf{r}')}\sum_{\omega_l} e^{-\jmath\omega_l\tau_{i,j}(\mathbf{r})+\jmath\omega_l\tau_{i,j}(\mathbf{r}')}, \tag{17}$$

where $\mathbf{r}$ is the true source position. Equation (17) is useful for analysis of the distribution of the regional SRP values. We split the whole surveillance area $\mathcal{D}$ into two areas.

Distribution of regional SRP values in the neighborhood of source position: The neighborhood of source position is defined as positions with distance from the true source position being less than a threshold, i.e., $\|\mathbf{r}' - \mathbf{r}\| \leq d_{\mathrm{t}}$. In this simulation $d_{\mathrm{t}} = 0.2$ m was used. For positions in this area, $\mathcal{P}^{\mathrm{c}}(\mathbf{r}')$ in (17) reaches the maximum due to the compensation of phase delays of the received signals.

Distribution of regional SRP values in the clutter positions: The clutter positions are defined as the positions which are at some distant away from the source position such that $\|\mathbf{r}' - \mathbf{r}\| \geq d_{\mathrm{t}}$. For those clutter positions, due to the unmatch in the phase compensation, we assume that the phase follows a uniform distribution [9], given by

$$\begin{aligned} O &= e^{-\jmath\omega_l\tau_{i,j}(\mathbf{r})+\jmath\omega_l\tau_{i,j}(\mathbf{r}')} \\ &= e^{\jmath\theta},\ \theta \sim \mathcal{U}[-\pi,\pi). \end{aligned} \tag{18}$$

In addition, due to the identically independent distributions of the phases and the sufficient number of summations for the phases, we deduce, based on central limit theorem, that the regional SRP power values for the clutter positions follow a Gaussian distribution, i.e.,

$$\mathcal{P}^{\mathrm{c}}(\mathbf{r}') \sim \mathcal{N}(0,\sigma^2),\ \|\mathbf{r}' - \mathbf{r}\| \geq d_{\mathrm{t}}. \tag{19}$$

where σ^2 is the variance of distribution of regional SRP values in clutter positions.

Figure 2 shows the two distributions of the regional SRP values in these two areas. The distribution of regional SRP values in the neighborhood of source position is indicated by the solid line, while the distribution of SRP values in clutter positions is indicated by the dashed line. The figure shows that the distribution of SRP values in clutter positions corresponds approximately to a zero mean Gaussian distribution as expected. The variance σ^2 depends on the TDOA summation boundary and number of microphone pairs used in (17). In our simulation, $\sigma^2 = 25$ was observed when $M = 8$ and $\rho = 0.1$ m was used. On the other hand, the regional SRP values corresponding to the neighborhood of source position are generally higher than the values corresponding to the clutter positions due to the phase compensation in (17). We therefore choose a threshold to distinguish between these two distributions of regional SRP values. In this work, we set an ad-hoc threshold $\mathcal{P}_{\mathrm{t}} = 20$ in order to eliminate the effect of clutter positions as much as possible. This threshold should be modified accordingly if different M and ρ are used.

A normal cumulative distribution function (cdf) can be applied as the mapping function:

$$\mathcal{M}(\mathcal{P}^{\mathrm{c}}(\mathbf{r}'_k)) = \Phi(\mathcal{P}^{\mathrm{c}}(\mathbf{r}'_k), \mathcal{P}_{\mathrm{t}}, \sigma^2_{\mathcal{P}}), \tag{20}$$

where $\Phi(\cdot)$ is a normal cdf. As discussed, the threshold $\mathcal{P}_{\mathrm{t}} = 20$ is chosen so that the regional SRP values of clutter positions are mapped onto the lower end of $\Phi(\cdot)$, while those corresponding to the neighborhood of the source position are mapped onto the higher end of $\Phi(\cdot)$. The variable $\sigma^2_{\mathcal{P}}$ is the variance of the normal cdf which determines its steepness. In this work, $\sigma^2_{\mathcal{P}} = 12$ was chosen and performs well in our simulation. The likelihood $\Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k)$ thus can be defined as

$$\Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k) = \begin{cases} \mathcal{M}(\mathcal{P}^{\mathrm{c}}(\mathbf{r}'_k)), & \text{for voiced frame} \\ \mathcal{U}_{\mathcal{D}}(\mathbf{r}'_k), & \text{for unvoiced frame} \end{cases}. \tag{21}$$

The remaining procedures follow the standard PF framework in Table I. The position estimate at each iteration $\widehat{\mathbf{r}}_k$ correspond to the first two elements of the state estimate $\widehat{\boldsymbol{\alpha}}_k$.

IV. Simulation Results

Simulations were conducted in a room of dimension 5 m $\times$ 5 m $\times$ 2.5 m. Eight microphones were distributed 0.5 m away from the perimeter of the room (see Fig. 3.) A 13 s speech signal sampled at 16 kHz from the TIMIT database [14] was used as a source signal. The microphone signals were generated by the method of images [15]. White Gaussian noise (WGN) at different signal-to-noise ratio (SNR) was added to the microphone signals. The positions of speech source were computed using a frame size of 1024 samples with $N_p = 80$ particles. The radius of the circular region centered on each particle was $\rho = 0.1$ m. The effective sample size threshold in PF was $N_{\mathrm{t}} = 37.5$.

The proposed method is compared with the conventional PF-SRP tracking method [8] where the simple binary voice/unvoice detector was implemented and the regional SRP localization method without PF framework [10]. We quantify their performance using $e_k = ||\widehat{\mathbf{r}}_k - \mathbf{r}_k||_2$, where $\widehat{\mathbf{r}}_k$ is the estimated position at the kth frame, and $\mathbf{r}_k$ is the true source position. The average tracking error $\bar{e} = \frac{1}{K}\sum_{k=1}^{K} e_k$

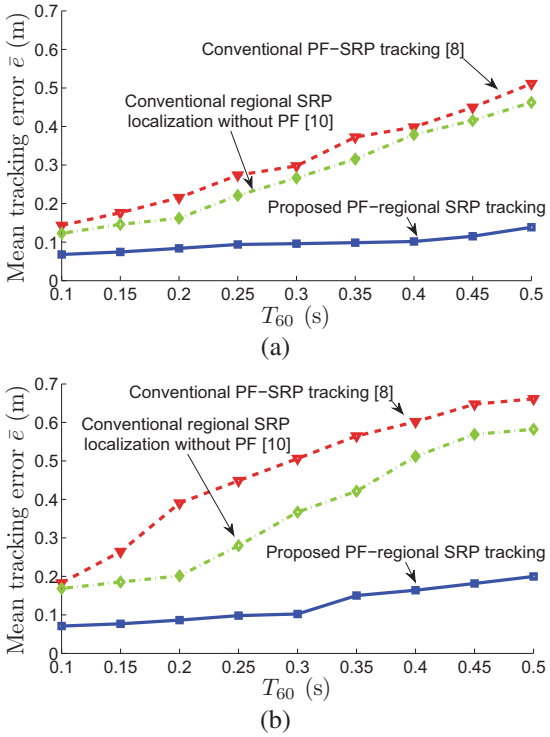

Fig. 4: Variation of average tracking error with reverberation time for (a) SNR = 10 dB and (b) SNR = 3 dB.

quantifies the performance across all audio frames, where K is the number of frames.

Figure 3 compares the tracking results of the two PF based tracking methods when $T_{60} = 450$ ms. Figure 3 (a) shows that the performance of the conventional PF-SRP method [8] is significantly affected by room reverberation. The particles, indicated by the dotted points, are scattered around the surveillance region due to the poor performance of the conventional SRP measurements. The conventional PF-SRP method has an average tracking error of 0.41 m. Figure 3 (b) shows the performance of the proposed PF-regional SRP method. The regional SRP measurements result in well-propagated particles which are concentrated along the true source trajectory. The proposed method achieves an averaged tracking error of 0.10 m, indicating that it outperforms the conventional PF-SRP method in this reverberant condition.

Figure 4 presents the average tracking error of the conventional PF-SRP method [8], the regional SRP without PF method [10] and the proposed PF-regional SRP method, for various reverberation time. Two cases of SNR = 10 and 3 dB were examined. The performance of these three methods reduces with reverberation time, as expected. The conventional PF-SRP method and the regional SRP without PF method consistently exhibit higher tracking error than the proposed PF-regional SRP method. The lower SNR condition further degrades the performance of the conventional methods. Due to the improved regional SRP evaluation, the regional SRP without PF method performs modestly better than the PF-SRP method, even though it does not exploit the temporal consistency of source positions. By incorporating the PF framework and taking into account the temporal consistency of source

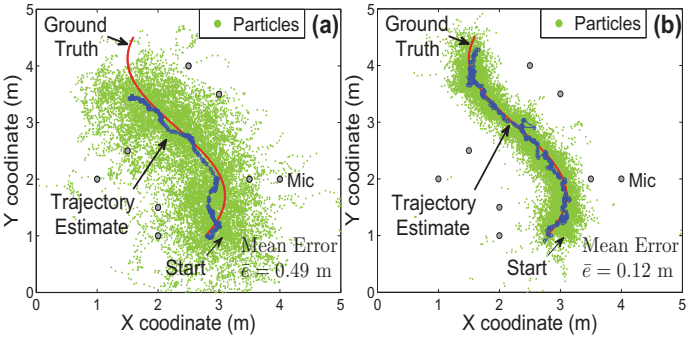

Fig. 5: Comparison of tracking results with $T_{60} = 450$ ms and SNR = 10 dB using randomly distributed microphones. (a) Conventional PF-SRP tracking method [8]. (b) Proposed PF-regional SRP tracking method.

positions, the proposed PF-regional SRP results in a mean error of less than 0.2 m, indicating that it outperforms both of the two conventional methods for the environments being examined. The improvement over the conventional methods becomes more significant at lower SNR and higher reverberant condition.

To further examine the validity of the algorithm in different microphone array configuration, we consider microphones that are randomly distributed as illustrated in Fig. 5. The remaining parameters were the same as the previous simulations. The conventional PF-SRP method [8], shown in Fig. 5 (a), results in the particles scattered around the room enclosure and poor performance is exhibited. The proposed PF-regional SRP method, shown in Fig. 5 (b), can achieve good tracking performance by reducing the tracking error from 0.49 m to 0.12 m. This simulation indicates that the algorithm is not limited to the case where the microphones have to be placed along the parameter of the room enclosure.

V. Conclusion

We propose a PF based acoustic source tracking framework by using a regional SRP measurement function. Instead of evaluating the power of discrete particle positions, the proposed method takes into account a circular region centered on each particle by accumulating the power within each region to provide a more comprehensive likelihood evaluation. Simulation results show that the proposed method achieves lower tracking error than the conventional methods in a noisy and reverberant environment.

References

[1] 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.

[2] 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.

[3] 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.

[4] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: a practical linear-correction least-squares approach," *IEEE Trans. Speech, Audio Process.*, vol. 9, no. 8, pp. 943–956, Nov. 2001.

[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] M. F. Fallon and S. Godsill, "Acoustic source localization and tracking using track before detect," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1228–1242, 2010.
[10] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Process. Letters*, vol. 18, no. 1, pp. 71–74, 2011.
[11] J. DiBiase, H. Silverman, and M Brandstein, "Robust localization in reverberant rooms," *Microphone Arrays: Signal Processing Techniques and Applications.*, pp. 157–180, 2001.
[12] D. Florencio C. Zhang and Z. Zhang, "Why does PHAT work well in low noise, reverberant environment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'08)*, 2008, pp. 2565–2568.
[13] J. H. DiBiase, *A High Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments using Microphone Arrays*, Ph.D. thesis, Brown Univ., 2000.
[14] 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.
[15] 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.