SOURCE TRACKING USING MOVING MICROPHONE ARRAYS FOR ROBOT AUDITION

Christine Evers, Yuval Dorfan, Sharon Gannot*, and Patrick A. Naylor*

Department of Electrical and Electronic Engineering, Imperial College, London, SW7 2AZ, UK
* Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel
c.evers@imperial.ac.uk

ABSTRACT

Intuitive spoken dialogues are a prerequisite for human-robot interaction. In many practical situations, robots must be able to identify and focus on sources of interest in the presence of interfering speakers. Techniques such as spatial filtering and blind source separation are therefore often used, but rely on accurate knowledge of the source location. In practice, sound emitted in enclosed environments is subject to reverberation and noise. Hence, sound source localization must be robust to both diffuse noise due to late reverberation, as well as spurious detections due to early reflections. For improved robustness against reverberation, this paper proposes a novel approach for sound source tracking that constructively exploits the spatial diversity of a microphone array installed in a moving robot. In previous work, we developed speaker localization approaches using expectation-maximization (EM) approaches and using Bayesian approaches. In this paper we propose to combine the EM and Bayesian approach in one framework for improved robustness against reverberation and noise.

Index Terms— Bayesian estimation; Expectation-Maximization; Particle filter; Acoustic Signal Processing; Sound Source Tracking.

1. INTRODUCTION

The ability of robots to engage in verbal dialogues is a fundamental prerequisite for intuitive interaction between humans and machines. To focus on desired sound sources subject to interference and noise, autonomous systems, such as robots, rely on beamforming [1] in the direction of salient acoustic events. The source directions are estimated using sound source localization techniques [2]. However, in realistic environments, reverberation causes localization errors and spurious detections due to dominant early reflections [3].

For improved robustness of source localization, spatial diversity of microphones installed on moving platforms can be exploited constructively in order to infer the source-sensor distance and to disambiguate the direction of impinging sound waves due to the direct path of a source [4, 5, 6]. In previous work [7], we compared the two paradigms of maximum likelihood (ML) and Maximum *a posteriori* (MAP) estimation for sound source localization from moving microphones in reverberant environments.

The ML estimator was implemented using an iterative expectation-maximization (EM) approach that maximized the likelihood of measured pair-wise relative phase ratios (PRPs) in order to estimate the position of the source within a pre-defined, discrete grid over the room region. The EM algorithm provides a natural approach to fit data from multi-modal distributions. The results in [7] therefore demonstrated that the iterative EM approach robustly estimates the source positions from a batch of data by clustering direct-path PRPs from noisy PRPs due to reverberation. For real-time processing, the iterative EM in [7] can be extended to recursive EM (REM) algorithms [8] for online processing. Nevertheless, the performance of the EM approach is limited by the resolution of the discrete grid of source hypotheses. Consequently, for applications where high localization resolution is required, the use of a discrete grid can be computationally wasteful. Furthermore, due to the lack of temporal models within the ML framework, parameters are estimated based on the current data, independent of the source trajectories.

The MAP estimator in [7] was implemented using a particle filter, propagating in time a cloud of random variates, or particles, of source positions using prior information of temporal models of the source dynamics. Information is inferred from the PRPs by evaluating the likelihood of each particle. Resampling ensures that only stochastically relevant particles are retained. The particle filter therefore estimates smoothed trajectories of source positions across time. Furthermore, the particle filter propagates only a meaningful cloud of randomly sampled source positions, therefore avoiding the need for discretized grids. Nevertheless, particle filters rely on the abstraction of the raw data to low-level features such as PRPs. Although the speech signals contain potentially crucial information encapsulated in the received speech signals, a transformation that maps the source positions to the acoustic transfer function between the source and sensors is unknown in practice. As a consequence, given the particles and raw data, the likelihood cannot be evaluated directly.

Nevertheless, it was shown in [9] that the EM can be used to evaluate and maximize the raw data likelihood. Thus, the particle filter would benefit from the likelihood function evaluated within the EM framework. Simultaneously, the EM algorithm would benefit from the adapative grid of source positions provided by the particles.

Therefore, in this paper, we propose a novel approach that combines the EM algorithm within the Bayesian framework, for mutually improved performance. We show that the particle filter can be used to estimate and propagate an adaptive grid of source positions. The particle positions are used within the EM algorithm to estimate and maximize the likelihood of reverberant data, which is subsequently used in the particle filter to assign weights to the particles. Room simulations for realistic conditions demonstrate high accuracy in source localization using a single, moving pair of microphones.

This paper is structured as follows: Section 2 introduces the signal models. The proposed method is derived in Section 3. Performance results using room simulations are presented in Section 4, and conclusions are drawn in Section 5.

The research leading to these results has received funding from the European Unions Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 609465.

2. SYSTEM MODEL

2.1. Source motion model

The state, $\mathbf{s}(t) \triangleq \left[x(t), y(t), \dot{x}(t), \dot{y}(t)\right]^T$, of a source at time step t, and located at position $(x(t), y(t))$ with velocity $(\dot{x}(t), \dot{y}(t))$, can be modelled over time using a Langevin model [10], i.e.,

$$\mathbf{s}(t) = \mathbf{F}(t)\mathbf{s}(t-1) + \mathbf{u}(t), \quad \mathbf{u}(t) \sim \mathcal{N}\left(\mathbf{0}_{4\times 1},\, \mathbf{Q}(t)\right) \tag{1}$$

where the dynamics, $\mathbf{F}(t)$, and process noise covariance, $\mathbf{Q}(t)$, are

$$\mathbf{F}(t) \triangleq \begin{bmatrix} \mathbf{I}_2 & a\Delta_t \mathbf{I}_2 \\ \mathbf{0}_{2\times 2} & a\mathbf{I}_2 \end{bmatrix} \quad \text{and} \quad \mathbf{Q}(t) \triangleq \begin{bmatrix} b^2\,\Delta_t^2 \mathbf{I}_2 & \mathbf{0}_{2\times 2} \\ \mathbf{0}_{2\times 2} & b^2 \mathbf{I}_2 \end{bmatrix} \tag{2}$$

where Δ_t is the time step, $a \triangleq e^{-\beta\,\Delta_t}$ and $b \triangleq \bar{v}\sqrt{1-a^2}$, and where β is the rate constant and $\bar{v}$ is the steady-state velocity. Therefore, the transition density of the source states, $p\left(\mathbf{s}(t) \mid \mathbf{s}(t-1)\right)$, is given by probability transformation of (1) as:

$$p\left(\mathbf{s}(t) \mid \mathbf{s}(t-1)\right) = \mathcal{N}\left(\mathbf{s}(t) \,\middle|\, \mathbf{F}(t)\mathbf{s}(t-1),\, \mathbf{Q}(t)\right). \tag{3}$$

2.2. Robot motion model

The microphone array used in this paper consists of one microphone pair with Cartesian positions, $\mathbf{p}_m(t) \triangleq \left[x_m(t), y_m(t)\right]^T$ for $m = 1, 2$ and with an inter-sensor distance of 0.5 m. The microphone pair moves at constant speed along a straight line within the room. In this paper, the positions of the microphones are assumed known. For unknown robot positions, the source localization approach proposed in this paper can be integrated in the acoustic Simultaneous Localization and Mapping (SLAM) approach in [4, 5, 6].

2.3. Signal model

The short-time Fourier transform (STFT) of the clean speech signal emitted by a single source at time t and frequency k is given by $y(t,k)$. The source signal is convolved with the Acoustic Impulse Response (AIR) of the reverberant room, such the STFT, $z_m(t,k)$, at each of the two microphones is expressed as:

$$z_m(t,k) = h_m(t,k)\, y(t,k), \tag{4}$$

where $h_m(t,k)$ is the Acoustic Transfer Function (ATF) between the source and microphone $m \in 1, 2$ at time t and frequency k. The ATF can be separated into the Relative Direct Transfer Function (RDTF), $h_m^{(d)}(t,k)$, and the transfer function due to early reflections and late reverberation, $h_m^{(r)}$, such that (4) is equivalent to,

$$z_m(t,k) = h_m^{(d)}(t,k)\, y(t,k) + n_m(t,k), \tag{5}$$

where the non-direct component, $n_m(t,k)$, captures the effects of early reflections and late reverberation, i.e.,

$$n_m(t,k) \triangleq h_m^{(r)}\, y(t,k). \tag{6}$$

Furthermore, $h_m^{(d)}(t,k)$ in (5) denotes the RDTF between sensor m and the source, modeled as a function of the source angle, $\vartheta_m(t)$, relative to the array of microphones via the plane-wave approximation [11],

$$h_m^{(d)}(t,k) = \exp\left\{\frac{2\pi\, \jmath\, k\, d_m\, \cos\vartheta_m(t)}{K\, T_s\, c}\right\}, \tag{7}$$

where c is the speed of sound, K is the number of frequency bins, T_s is the sampling period, and d_m is the distance between microphone m and the reference microphone.

The microphone signals, $\mathbf{z}(t,k) \triangleq \left[z_1(t,k), z_2(t,k)\right]^T$, can be synonymously expressed in vector form as

$$\mathbf{z}(t,k) = \mathbf{h}^{(d)}(t,k)\, y(t,k) + \mathbf{n}(t,k), \tag{8}$$

where $\mathbf{h}^{(d)}(t,k) \triangleq \left[h_1^{(d)}(t,k), h_2^{(d)}(t,k)\right]^T$ with noise term $\mathbf{n}(t,k) \triangleq \left[n_1(t,k), n_2(t,k)\right]^T$. Assuming $\mathbf{n}(t,k)$ is white Gaussian, the likelihood of the reverberant signals in (8) is given by

$$p\left(\mathbf{z}(t,k) \mid \mathbf{s}(t)\right) = \mathcal{N}^c\left(\mathbf{z}(t,k) \,\middle|\, \mathbf{0}_{M\times 1},\, \mathbf{\Phi}(t,k)\right), \tag{9}$$

where $\mathcal{N}^c$ denotes the complex Gaussian density, and $\mathbf{0}_{M\times 1}$ is the $M \times 1$ zero vector. The covariance in (9) is given by the Power Spectral Density (PSD), $\mathbf{\Phi}(t,k)$:

$$\mathbf{\Phi}(t,k) = \mathbf{h}^{(d)}(t,k)\left[\mathbf{h}^{(d)}(t,k)\right]^H \phi_y(t,k) + \mathbf{\Phi}_r(t,k), \tag{10}$$

where the direct-path PSD is denoted as $\phi_y(t,k) \triangleq \mathbb{E}\left[|y(t,k)|^2\right]$, and $\mathbf{\Phi}_r(t,k) \triangleq \mathbb{E}\left[\mathbf{n}(t,k)\,\mathbf{n}(t,k)^H\right]$ is the reverberation PSD matrix, which can be modelled in terms of its spatial incoherence matrix, $\mathbf{\Gamma}(t,k)$, and reverberation level, $\phi_R(t,k)$, as:

$$\mathbf{\Phi}_r(t,k) = \mathbf{\Gamma}(t,k)\, \phi_R(t,k). \tag{11}$$

Assuming reverberation can be modelled as a spatially homogenous, spherically isotropic sound field [12], element (i,j) for each $\{i,j\} \in \{1,2\}$ of the matrix $\mathbf{\Gamma}(t,k)$ can be modelled as

$$\Gamma_{i,j}(t,k) = \operatorname{sinc}\left(\frac{2\pi k\, d_{i,j}}{K\, T_s\, c}\right) + \epsilon\, \delta(i-j), \quad \{i,j\} = 1,2 \tag{12}$$

where ϵ is the diagonal loading factor, and $d_{i,j}$ is the Euclidean distance between microphone i and j.

3. METHODOLOGY

3.1. Sequential Importance Sampling

The MAP estimate of the source position is obtained from the posterior Probability Density Function (pdf), $p\left(\mathbf{s}(t) \mid \mathbf{Z}_t\right)$, via

$$\hat{\mathbf{s}}^{\text{MAP}}(t) = \underset{\mathbf{s}(t)}{\arg\max}\, p\left(\mathbf{s}(t) \mid \mathbf{Z}_{1:t}, \phi_y(t,k), \phi_R(t,k)\right), \tag{13}$$

where $\mathbf{Z}_{1:t} = \left[\mathbf{Z}_1^T, \ldots, \mathbf{Z}_t^T\right]^T$ with $\mathbf{Z}_t \triangleq \left[\mathbf{z}(t,1)^T, \ldots, \mathbf{z}(t,K)^T\right]^T$. The posterior density is given via Bayes's theorem as

$$\begin{aligned} &p\left(\mathbf{s}(t) \mid \mathbf{Z}_t, \phi_y(t,k), \phi_R(t,k)\right) \\ &\quad = \frac{p\left(\mathbf{Z}_t \mid \boldsymbol{\theta}_t\right) p\left(\mathbf{s}(t) \mid \mathbf{s}(t-1)\right)}{\int p\left(\mathbf{Z}_t \mid \boldsymbol{\theta}_t\right) p\left(\mathbf{s}(t) \mid \mathbf{s}(t-1)\right) d\mathbf{s}(t)}, \end{aligned} \tag{14}$$

where $\boldsymbol{\theta}_t \triangleq \left[\mathbf{s}^T(t), \phi_y(t,k), \phi_R(t,k)\right]^T$, the likelihood is denoted as $p\left(\mathbf{Z}_t \mid \boldsymbol{\theta}_t\right)$ and $p\left(\mathbf{s}(t) \mid \mathbf{s}(t-1)\right)$ is the prior. However, due to the denominator in (14), the posterior pdf is analytically intractable.

Nevertheless, the posterior pdf can be approximated by sampling from a proposal density, $\pi\left(\mathbf{s}(t) \mid \hat{\mathbf{s}}^{(j)}(t-1), \mathbf{Z}_t\right)$, that includes the support of $p\left(\mathbf{s}(t) \mid \mathbf{Z}_{1:t}\right)$. Hence,

$$p\left(\mathbf{s}(t) \mid \mathbf{Z}_{1:t}, \phi_y(t,k), \phi_R(t,k)\right) \approx \sum_{j=1}^{J} \tilde{w}^{(j)}(t)\, \delta_{\hat{\mathbf{s}}^{(j)}(t)}(\mathbf{s}(t)), \tag{15}$$

where $\tilde{w}^{(j)}(t) \triangleq w^{(j)}(t)/\sum_{j=1}^{J} w^{(j)}(t)$ are the normalized importance weights, and the particles, $\hat{\mathbf{s}}^{(j)}(t)$, are drawn from

$$\hat{\mathbf{s}}^{(j)}(t) \sim \pi\left(\mathbf{s}(t) \mid \hat{\mathbf{s}}^{(j)}(t-1), \mathbf{Z}_t\right). \tag{16}$$

Using prior importance sampling from (3), i.e.,

$$\pi\left(\mathbf{s}(t) \mid \hat{\mathbf{s}}^{(j)}(t-1), \mathbf{Z}_t\right) = p\left(\mathbf{s}(t) \mid \hat{\mathbf{s}}^{(j)}(t-1)\right), \tag{17}$$

the unnormalized importance weights, $w^{(j)}(t)$ are given by [13]:

$$w^{(j)}(t) = w^{(j)}(t-1)\, p\left(\mathbf{Z}_t \mid \boldsymbol{\theta}_t^{(j)}\right). \tag{18}$$

Assuming Independent and Identically Distributed (IID) frequency bins, the likelihood of the reverberant measurements corresponding to particle, $\hat{\mathbf{s}}^{(j)}(t)$, can be obtained from (9) as:

$$p\left(\mathbf{Z}_t \mid \boldsymbol{\theta}_t^{(j)}\right) = \prod_{k=1}^{K} p\left(\mathbf{z}(t,k) \mid \boldsymbol{\theta}_t^{(j)}\right). \tag{19}$$

However, the direct-path and reverberant PSDs, $\phi_y(t,k)$ and $\phi_R(t,k)$ are required in order to evaluate the PSD in (10). As $\phi_y(t,k)$ and $\phi_R(t,k)$ are unknown in practice, the likelihood in (19) and hence the particle weights in (18) cannot be evaluated directly from the particles, $\{\hat{\mathbf{s}}^{(j)}(t)\}_{j=1}^{J}$.

Nevertheless, considering the cloud of particles as an adaptive grid of source position hypotheses, the ML framework can be used to estimate $p(\mathbf{Z}_t \mid \boldsymbol{\theta}_t)$ for the importance weights in (18).

3.2. EM algorithm

As the likelihood, $p(\mathbf{Z}_t \mid \boldsymbol{\theta}_t)$, is generally high-dimensional and multi-modal, it is often difficult to maximize in practice. Rather than maximizing the likelihood directly, the EM algorithm in [9] maximizes the joint density of the observed data and a set of latent, unobserved and discrete variables, $\mathbf{X}_t$,

$$p(\mathbf{Z}_t \mid \boldsymbol{\theta}_t) = \frac{p(\mathbf{Z}_t, \mathbf{X}_t \mid \boldsymbol{\theta}_t)}{p(\mathbf{X}_t \mid \boldsymbol{\theta}_t)}, \tag{20}$$

with $\mathbf{X}_t \triangleq \left[x(t,1,\vartheta), \ldots, x(t,K,\vartheta)\right]^T$ and where $x(t,k,\vartheta)$ is an IID indicator that (t,k) is solely associated with a source in the direction of $\vartheta = \gamma(t) - \tan^{-1}\left(x_r(t)/y_r(t)\right)$, where $(x_r(t), y_r(t))$ is the source position relative to the sensor with orientation $\gamma(t)$. Moreover, the joint posterior, $p(\mathbf{Z}_t, \mathbf{X}_t \mid \boldsymbol{\theta}_t)$, can be expressed using (9) as [9]:

$$p(\mathbf{Z}_t, \mathbf{X}_t \mid \boldsymbol{\theta}_t) = \prod_{k} \sum_{j=1}^{J} x(t,k,\vartheta_j)\, \mathcal{N}^c\left(\mathbf{z}(t,k) \,\middle|\, \mathbf{0}_{2\times 1},\, \boldsymbol{\Phi}^{(j)}(t,k)\right). \tag{21}$$

In [9] the indicator is evaluated over a predetermined, discrete grid of source directions in $[0, 2\pi]$. This paper proposes to use instead the directions of the particles, denoted by $\{\vartheta_j\}_{j=1}^{J}$, as an adaptive grid of $P = J$ source directions.

The log-likelihood corresponding to (20) is given by:

$$\ln p(\mathbf{Z}_t \mid \boldsymbol{\theta}_t) = \ln p(\mathbf{Z}_t, \mathbf{X}_t \mid \boldsymbol{\theta}_t) - \ln p(\mathbf{X}_t \mid \boldsymbol{\theta}_t) \tag{22}$$

where $p(\mathbf{Z}_t, \mathbf{X}_t \mid \boldsymbol{\theta}_t)$ is the joint pdf of the complete data and $p(\mathbf{X}_t \mid \boldsymbol{\theta}_t)$ is marginal density of the indicator.

The pdf of the complete data can be written as [14]:

$$\ln p(\mathbf{Z}_t, \mathbf{X}_t \mid \boldsymbol{\theta}_t) = Q(\boldsymbol{\theta}_t, \boldsymbol{\theta}_t^{(\ell-1)}) - \sum_{k,j} p\left(x(t,k,\vartheta_j) \mid \mathbf{Z}_t, \boldsymbol{\theta}_t^{(\ell-1)}\right) \ln p\left(x(t,k,\vartheta_j) \mid \mathbf{Z}_t, \boldsymbol{\theta}_t^{(\ell-1)}\right), \tag{23}$$

where

$$Q(\boldsymbol{\theta}_t, \boldsymbol{\theta}_t^{(\ell-1)}) \triangleq \sum_{k,j} p\left(x(t,k,\vartheta_j) \mid \mathbf{Z}_t, \boldsymbol{\theta}_t^{(\ell-1)}\right) \ln p(\mathbf{Z}_t, x(t,k,\vartheta_j) \mid \boldsymbol{\theta}_t), \tag{24}$$

where $\psi^{(j)}$ is the probability to have a source in the j^{th} direction. The EM algorithm therefore iteratively estimates the maximum likelihood in a two-stage process. Using (21), the E-step [9] evaluates:

$$\mu^{(\ell-1)}(t,k,j) \triangleq \mathbb{E}\left[x(t,k,\vartheta_j) \mid \mathbf{z}(t,k), \boldsymbol{\theta}^{(\ell-1)}\right] = \frac{\psi_j^{(\ell-1)} \mathcal{N}^c\left(\mathbf{z}(t,k) \,\middle|\, \mathbf{0}_{2\times 1},\, \boldsymbol{\Phi}_j(t,k)\right)}{\sum_{j=1}^{J} \psi_j^{(\ell-1)} \mathcal{N}^c\left(\mathbf{z}(t,k) \,\middle|\, \mathbf{0}_{2\times 1},\, \boldsymbol{\Phi}_j(t,k)\right)} \tag{25}$$

where $\boldsymbol{\Phi}_j(t,k) = \mathbf{h}^{(r)}(t,k)\left[\mathbf{h}^{(r)}(t,k)\right]^H \phi_{S,j}(t,k) + \boldsymbol{\Gamma}(t,k)\,\phi_{R,j}(t,k)$.

In the M-step, $p(\mathbf{Z}_t, x(t,k,\vartheta_j) \mid \boldsymbol{\theta}_t)$ is maximized. According to (24), this maximization is equivalent to separately maximizing $\ln \psi_m$ and the log-likelihood. By constrained maximization:

$$\psi_j^{(\ell)} = \frac{1}{K} \sum_{k=1}^{K} \mu^{(\ell-1)}(t,k,j). \tag{26}$$

Furthermore, maximization of the log-likelihood reduces to [15],

$$\phi_{R,j}(t,k) = \mathbf{z}^H(t,k)\left[\mathbf{I}_2 - \mathbf{b}_j\, \mathbf{h}_j^H(t,k)\right] \boldsymbol{\Gamma}^{-1}(t,k)\, \mathbf{z}(t,k) \tag{27a}$$

$$\phi_{S,j}(t,k) = \mathbf{b}_j^H\left[\mathbf{z}(t,k)\, \mathbf{z}^H(t,k) - \phi_{R,j}(t,k)\, \boldsymbol{\Gamma}(t,k)\right] \mathbf{b}_j \tag{27b}$$

where $\mathbf{b}_j$ is the Minimum Variance Distortionless Response beamformer in the direction ϑ_j, i.e.,

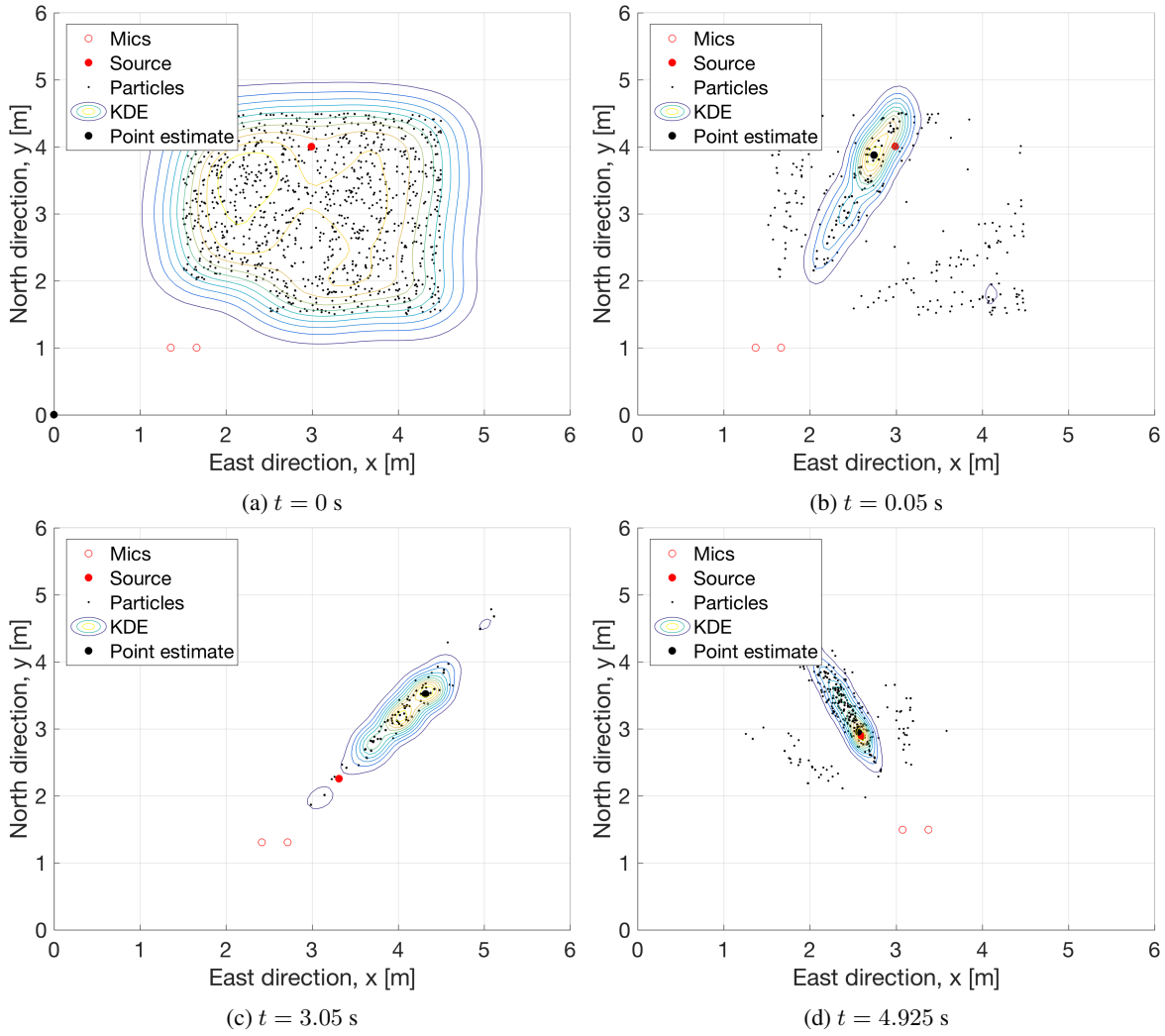
$$\mathbf{b}_j = \frac{\boldsymbol{\Gamma}^{-1}(t,k)\, \mathbf{h}_j(t,k)}{\mathbf{h}_j^H(t,k)\, \boldsymbol{\Gamma}^{-1}(t,k)\, \mathbf{h}_j(t,k)}. \tag{28}$$

Therefore, instead of using a deterministic grid of source positions, the positions sampled within the particle filter are used to maximize the source directions in (25)-(27). Simultaneously, the probabilities, $\psi_j^{(L)}$, are used to weight the particles in lieu of (18):

$$w^{(j)}(t) = w^{(j)}(t-1)\, \psi_j^{(L)}\, \mathcal{N}^c\left(\mathbf{z}(t,k) \,\middle|\, \mathbf{0}_{2\times 1},\, \boldsymbol{\Phi}_j(t,k)\right). \tag{29}$$

4. RESULTS

The trajectory of a moving pair of microphones is simulated over 10 s along a straight line in a $6 \times 6 \times 2.5$ m^3 room with the initial and final position of the origin of the pair at $(1.5, 1, 1.5)$ m and $(5, 2, 1.5)$ m respectively. The two microphones are offset by 0.15 m in x-direction to the left and right of the origin respectively. The



(a) $t = 0$ s (b) $t = 0.05$ s

(c) $t = 3.05$ s (d) $t = 4.925$ s

Fig. 1: Distribution of particles across time.

trajectory of a moving source is simulated using (1) with $\beta = 2$, $\bar{v} = 1$ m/s. Using the room impulse response (RIR) generator in [16] the RIR for each source-sensor geometry is simulated for $T_{60} = 500$ ms at sampling frequency $f_s = 8$ kHz. The resulting RIRs are convolved with a 10 s anechoic speech signal from a female speaker constructed from the TIMIT database. The STFT of the signal is evaluated using a rectangular window for each microphone for a frame length of 50 ms. The proposed approach is evaluated for $\Delta_t = 0.375$ s using 1000 particles. The initial particles are drawn from a uniform distribution with a minimum distance of 1.5 m to each of the walls and at least 1 m from the microphone origin.

Figure 1 shows the distribution of particles at four time steps. For each time step, the point estimate of the source position is extracted from the particles as the peak of the weighted Kernel Density Estimate (KDE). The KDE is shown in Figure 1 as the contour plot, highlighting concentrations of particles with high weight. It can be seen that the peak of the KDE converges to the true source angle. Triangulation of the source position is highly dependent on the source-sensor geometry, resulting in estimation errors of under 0.4 m at $t = 3.05, 4.925$ s, and an average position error across all time steps of 0.805 m. The main reason for the estimation error is the unmeasured source-sensor distance. The scenario corresponds to source-sensor distances between $[0.688, 4.464]$ m. However, the source position is triangulated from a pair of microphones with only 0.3 m inter-microphone distance and relatively small displacement of approximately 0.15 m between time steps. Nevertheless, high accuracy is achieved in the estimated source angle, with an average of 0.319 deg accuracy.

5. CONCLUSION

We proposed a novel approach to sound source tracking in reverberant environments using a single moving pair of microphones. A particle filter is used to propagate hypotheses of source positions across time. At each time step, the EM algorithm uses the particles to estimate and maximize the likelihood of reverberant measurements. The resulting probabilities are used in the particle filters as importance weights. Results for 500 ms reverberation time using two microphones separated by 0.3 m demonstrated estimation accuracy of 0.805 m in position and 0.319 deg in the source direction of arrival.

6. REFERENCES

[1] B. D. van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoustics, Speech and Signal Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[2] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, ser. Digital Signal Processing, M. Brandstein and D. Ward, Eds. Springer Berlin Heidelberg, 2001, pp. 157–180.

[3] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Springer, 2010.

[4] C. Evers, A. H. Moore, and P. A. Naylor, "Acoustic simultaneous localization and mapping (a-SLAM) of a moving microphone array and its surrounding speakers," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016.

[5] ——, "Localization of moving microphone arrays from moving sound sources for robot audition," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Budapest, Hungary, 2016.

[6] ——, "Towards informative path planning for acoustic simultaneous localization of microphone arrays and mapping of surrounding sound sources (a-SLAM)," in *DAGA*, Aachen, Germany, mar 2016.

[7] Y. Dorfan, C. Evers, S. Gannot, and P. A. Naylor, "Speaker localization with moving microphone arrays," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Budapest, Hungary, Aug. 2016.

[8] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 392–402, Feb. 2014.

[9] O. Schwartz, Y. Dorfan, E. Habets, and S. Gannot, "Multiple doa estimation in reverberant conditions using EM," in *Proc. Intl. Workshop for Acoustic Echo Cancellation and Noise Control (IWAENC)*, Xi'an, China, Sep. 2016.

[10] E. A. Lehmann and R. C. Williamson, "Particle filter design using importance sampling for acoustic source localisation and tracking in reverberant environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, pp. 1–9, 2006.

[11] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," in *Proc. Intl. Symposium on Signal Processing and Its Applications*, vol. 2, Jul. 2003, pp. 411–414.

[12] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Am.*, vol. 122, no. 6, pp. 3464–3470, Dec. 2007.

[13] A. Doucet, S. Godsill, and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000.

[14] C. Bishop, *Pattern Recognition and Machine Learning*, ser. Information Science and Statistics. New York, USA: Springer, 2006.

[15] H. Ye and D. DeGroat, "Maximum likelihood DOA estimation and asymptotic Cramer-Rao bounds for additive unknown colored noise," *IEEE Trans. Signal Process.*, vol. 43, no. 4, pp. 938–949, Apr. 1995.

[16] E. Habets, "Room impulse response (RIR) generator," http://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator, 2010.