PARTICLE FILTERING ALGORITHMS FOR TRACKING MULTIPLE SOUND SOURCES USING MICROPHONE ARRAYS

Mitsuru Kawamoto[1,2]*, Futoshi Asano*[1,2]*, Hideki Asoh*[1,2]*, and Kiyoshi Yamamoto*[1,2]

1. National Institute of Advanced Industrial Science and Technology (AIST),
Central 2, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
2. CREST, JST.

ABSTRACT

A particle filtering algorithm using the parameters in the EM (Expectation-Maximization) algorithm is proposed for tracking multiple sound sources. Differently from the conventional EM based algorithms, the proposed algorithm can track multiple sound sources without knowing their starting points. Moreover, an idea of the group tracking is applied to the particle filtering algorithm so that better tracking performances can be obtained. Experimental results show the validity of the proposed algorithm.

Index Terms— Particle filtering algorithms, EM algorithms, Tracking, Multiple sound sources, Microphone arrays

1. INTRODUCTION

Sound source tracking using microphone arrays has been one of the central problems in radar, sonar, navigation, speech interaction, and so on.

In this paper, we propose a method of tracking for multiple sound sources, using particle filtering algorithms. The particle filter is used to estimate sound positions and on/off audio status. Differently from the conventional particle filtering algorithms, e.g., [2, 3], the information used to handle the particle filter is only audio signals. In [9], a particle filtering algorithm utilizing only the information of audio signals has been proposed, but the number of tracking sound sources is only one. Hence, in this paper, for the tracking of multiple sound sources, we want to show a method where good tracking performances can be obtained by particle filters using only the information of audio signals.

To this objective, in our particle filter, as a function of estimating importance weights [4], a pseudo-likelihood function, which is calculated by the parameters used in Expectation-Maximization (EM) algorithms (EMAs), is proposed. Since an effect of signal separation is embedded in the EMA [1], the EMA based pseudo-likelihood function may be suitable for tracking multiple sound sources.

Some examples, in which EMAs are applied to sound localization and tracking problems, have been introduced until now [1, 5, 6]. In the EMA, given the initial value for estimating the sound location or the tracked point, and then by iterating the E-step and the M-step alternately, the localization or the tracking is achieved. This is one of the advantages of the EMA compared with other conventional localization methods such as MUSIC [7]. However, if the initial value is far from desired solutions, it cannot be guaranteed whether or not the EMA provides the desired solution (see Section 4). In the proposed algorithm, such a problem can be avoided using the particle filter (see Section 4).

Moreover, we consider applying an idea of the group tracking [8] to the particle filtering. Then we expect that better tracking performances can be obtained by the proposed algorithm. Experimental results show the validity of the proposed algorithm.

2. SOUND LOCALIZATION USING THE EM ALGORITHM (EMA)

In this section, the EMA based sound localization method is briefly introduced, because we adopt the idea of the EMA to the proposed algorithm and hence this explanation may be helpful for understanding the proposed algorithm.

2.1. Audio Signal Model

Throughout this paper, audio signals $\boldsymbol{z}(t)$ are treated in the frequency domain. The short-time Fourier transform (STFT) of the microphone input is defined as $\boldsymbol{y}(t,\omega) = [Y_1(t,\omega),\cdots, Y_M(t,\omega)]^T$ (input vector), where $Y_m(t,\omega)$ is the STFT of mth microphone input at time t and frequency ω, M is the number of microphones. Hereafter, the index of frequency ω is omitted for the simplicity of writing. The input vector can be modeled as

$$\boldsymbol{y}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \boldsymbol{n}(t), \tag{1}$$

where $\boldsymbol{A}$ is a location vector matrix defined as

$$\boldsymbol{A} = [\boldsymbol{a}(\theta_1), \cdots, \boldsymbol{a}(\theta_L)], \tag{2}$$

$\boldsymbol{s}(t) = [S_1(t),\cdots, S_L(t)]^T$ is a source spectrum vector , and $\boldsymbol{n}(t) = [N_1(t),\cdots, N_M(t)]^T$ is a background noise spectrum vector. Here, L is the number of active sound sources and θ_l (l = 1,2, $\cdots$, L) represent the 2D directions of the sound sources. The noise is assumed to be zero mean Gaussian

noise. We assume that the covariance matrices of $\boldsymbol{s}(t)$ and $\boldsymbol{n}(t)$ are defined as, respectively, $E[\boldsymbol{s}(t)\boldsymbol{s}^H(t)] = \boldsymbol{K}_s = \mathrm{diag}\{\gamma_1, \cdots, \gamma_L\}$ and $E[\boldsymbol{n}(t)\boldsymbol{n}^H(t)] = \sigma\boldsymbol{I}$, where diag$\{\cdots\}$ denotes a diagonal matrix with the diagonal element $\{\cdots\}$, γ_l (l = 1,2 $\cdots$, L) denote the power spectrums of $S_l(t)$ (l = 1,2 $\cdots$, L), σ denotes the power of the noise $\boldsymbol{n}(t)$, and $\boldsymbol{I}$ denotes an identity matrix.

2.2. EMAs

In the EMA based approach, the input vector is decomposed into the following equation corresponding to each source signal, that is,

$$\boldsymbol{y}(t) = \sum_{l=1}^{L} \boldsymbol{x}_l(t) = \boldsymbol{H}\boldsymbol{x}(t), \tag{3}$$

where $\boldsymbol{x}_l(t) = \boldsymbol{a}(\theta_l)S_l(t) + \boldsymbol{n}_l(t)$, $\boldsymbol{x}(t) = [\boldsymbol{x}_1^T(t), \cdots, \boldsymbol{x}_L^T(t)]^T$ is a ML-column vector, and $\boldsymbol{H} = [\boldsymbol{I}, \cdots, \boldsymbol{I}]$ is an $M \times ML$ matrix. The symbol $\boldsymbol{n}_l(t)$ is an arbitrary decomposition of the noise vector satisfying $\sum_{l=1}^{L} \boldsymbol{n}_l(t) = \boldsymbol{n}(t)$ and $E[\boldsymbol{n}_l(t)\boldsymbol{n}_l^H(t)] = \frac{\sigma}{L}\boldsymbol{I}$. In the EMA based approach, a set of decomposed input vector, that is, $\boldsymbol{X} = [\boldsymbol{x}(1), \cdots, \boldsymbol{x}(N)]$, is called *complete data*, which is not directly observed. Under this decomposition, using a complete data $\boldsymbol{X}_l$ of $\boldsymbol{x}_l(t)$, the likelihood function for the complete data is defined as

$$L_{xl}(\theta_l, \gamma_l | \boldsymbol{X}_l) = \Psi_{xl} \exp\left(-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{C}_{xl}\boldsymbol{K}_{xl}^{-1}\right]\right), \tag{4}$$

where $\Psi_{xl} = (2\pi)^{-MN}[\det\boldsymbol{K}_{xl}]^{-N/2}$,

$$\boldsymbol{C}_{xl} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_l(n)\boldsymbol{x}_l^H(n), \tag{5}$$

$$\boldsymbol{K}_{xl} = \gamma_l \boldsymbol{a}(\theta_l)\boldsymbol{a}^H(\theta_l) + \frac{\sigma}{L}\boldsymbol{I}, \tag{6}$$

the symbol tr[$\boldsymbol{A}$] denotes the trace of the matrix $\boldsymbol{A}$ and the symbol det$\boldsymbol{A}$ denotes the determinant of the matrix $\boldsymbol{A}$. $\boldsymbol{C}_{xl}$ is the *sample covariance* matrix of the complete data $\boldsymbol{X}_l$. The covariance matrix of $\boldsymbol{y}(t)$ is written using (6) as $\boldsymbol{K}_y := E[\boldsymbol{y}(t)\boldsymbol{y}^H(t)] = \sum_{l=1}^{L} \boldsymbol{K}_{xl}$. In the E-step of the EMA, the conditional expectation of $\boldsymbol{C}_{xl}$ is estimated using the following equations;

$$\begin{aligned}\boldsymbol{C}_{xl}^p := E[\boldsymbol{C}_{xl}|\boldsymbol{C}_y; \hat{\boldsymbol{K}}_y^p] &= \hat{\boldsymbol{K}}_{xl}^p - \hat{\boldsymbol{K}}_{xl}^p(\hat{\boldsymbol{K}}_y^p)^{-1}\hat{\boldsymbol{K}}_{xl}^p \\ &\quad + \hat{\boldsymbol{K}}_{xl}^p(\hat{\boldsymbol{K}}_y^p)^{-1}\boldsymbol{C}_y(\hat{\boldsymbol{K}}_y^p)^{-1}\hat{\boldsymbol{K}}_{xl}^p \end{aligned} \tag{7}$$

$$\hat{\boldsymbol{K}}_y^p = \sum_{l=1}^{L} \hat{\boldsymbol{K}}_{xl}^p \tag{8}$$

$$\hat{\boldsymbol{K}}_{xl}^p = \hat{\gamma_l}^p \boldsymbol{a}(\hat{\theta}_l^p)\boldsymbol{a}(\hat{\theta}_l^p)^H + \frac{\sigma}{L}\boldsymbol{I} \tag{9}$$

In the M-step, the parameters θ_l and γ_l are estimated using

$$\hat{\theta}_l^{p+1} = \arg\max_{\theta_l} \frac{\boldsymbol{a}^H(\theta_l^p)\boldsymbol{C}_{xl}^p\boldsymbol{a}(\theta_l^p)}{|\boldsymbol{a}(\theta_l^p)|^4} \tag{10}$$

$$\hat{\gamma}_l^{p+1} = \frac{\boldsymbol{a}^H(\theta_l^{p+1})\boldsymbol{C}_y\boldsymbol{a}(\theta_l^{p+1})}{|\boldsymbol{a}(\theta_l^{p+1})|^4}, \tag{11}$$

so that the likelihood function (4) is maximized using the conditional expectation of $\boldsymbol{C}_{xl}$. The symbol $\hat{}$ (hat) indicates the estimate in the EMA. $\boldsymbol{C}_y$ is the sample covariance matrix of the input vector $\boldsymbol{y}(t)$. The superscript p denotes the iteration number of the EMA. Therefore, by iterating the E-step and the M-step alternately, the EMA works such that sound source locations can be estimated.

In our proposed algorithm, (7) through (9), and (11) are used to localize sound sources. However, since $\boldsymbol{C}_{xl}^p$ in (7) is calculated at a frequency bin, if there are no information in the frequency bin, then the accuracy of the estimation of $\boldsymbol{C}_{xl}^p$ becomes wrong and the good estimate of $\hat{\theta}_l^{p+1}$ cannot be obtained. Hence, when (7) is utilized in the proposed algorithm, the following $\hat{\boldsymbol{C}}_{xl}^p$ is used instead of $\boldsymbol{C}_{xl}^p$ in (7).

$$\hat{\boldsymbol{C}}_{xl}^p = \sum_{\omega} \boldsymbol{C}_{xl}^p(\omega). \tag{12}$$

This is our original novel point not possessed by the conventional EMAs. Note that the way of estimating θ_l in the proposed algorithm is shown in the next section.

Remark 1 *In (11), the EMA in [1] does not use $\boldsymbol{C}_y$ but uses $\boldsymbol{C}_{xl}^p$. However, from our simulation results, we can see that (11) with $\boldsymbol{C}_y$ provides more stable calculation than $\boldsymbol{C}_{xl}^p$. Therefore, we use $\boldsymbol{C}_y$ to calculate $\hat{\gamma}_l$.*

3. PARTICLE FILTERING ALGORITHMS (PFAS)

In this section, we propose a particle filtering algorithm (PFA) for estimating θ_l, where a pseudo-likelihood function calculated by the parameters $\boldsymbol{C}_{xl}^p$ in (7) and $\hat{\boldsymbol{K}}_{xl}^p$ in (9) is utilized.

Before explaining our proposed algorithm, let us formulate the tracking problem considered in this paper. This problem can be formulated in a framework of Bayesian estimation of hidden state sequences. As in [4], let the hidden variable vector be

$$\boldsymbol{\chi}(t) = [m(t), \chi_1(t), \cdots, \chi_{m(t)}(t)], \tag{13}$$

where $m(t)$ is the number of tracking targets and $\chi_i(t) = (\theta_i, s_i)$ is the configuration of ith target. The discrete variable θ_i is the same as in (2) and the Boolean variable s_i denotes audio activity (on/off). Observation variables $\boldsymbol{Z}(t)$ are composed of audio signals $\boldsymbol{z}(t)$.

In the Bayesian framework, the relationship between $\boldsymbol{\chi}_{1|T} = \boldsymbol{\chi}(1), \cdots, \boldsymbol{\chi}(T)$ and $\boldsymbol{Z}_{1|T} = \boldsymbol{Z}(1), \cdots, \boldsymbol{Z}(T)$ is modeled by a joint probability distribution $P(\boldsymbol{\chi}_{1|T}, \boldsymbol{Z}_{1|T})$. It is assumed that the joint probability distribution can be decomposed as

$$P(\boldsymbol{\chi}_{1|T}, \boldsymbol{Z}_{1|T}) = \prod_{t=1}^{T} P(\boldsymbol{Z}(t)|\boldsymbol{\chi}(t))P(\boldsymbol{\chi}(t)|\boldsymbol{\chi}(t-1)), \tag{14}$$

This means that the state transition probabilities $P(\boldsymbol{\chi}(t)|\boldsymbol{\chi}(t-1))$ and the observation probabilities (the likelihood of the observation) $P(\boldsymbol{Z}(t)|\boldsymbol{\chi}(t))$ are needed to specify the joint probability distribution. Then, when observations $\boldsymbol{z}_{1|t}$ are given,

according to Bayes' theorem, the posterior probability distribution $P(\boldsymbol{\chi}(t)|\boldsymbol{Z}(t))$ is computed. Therefore, the tracking problem can be formulated as follows; Estimate $P(\boldsymbol{\chi}(t)|\boldsymbol{Z}(t))$, that is, the configurations of the targets, using $P(\boldsymbol{\chi}(t)|\boldsymbol{\chi}(t-1))$ and $P(\boldsymbol{Z}(t)|\boldsymbol{\chi}(t))$.

From the formulation, we utilize such a particle filtering algorithm that the state transition probability is used as the proposal distribution and the (pseudo-) likelihood of the observation is used as the importance weight [4].

3.1. State Transition Probability

As the state transition model, each target is assumed to move independently from its current location according to a Gaussian distribution with zero mean and common variance σ_l. Then, based on an idea of the group tracking, the following transformation is applied to each particle;

$$\hat{\boldsymbol{\chi}}_i(t) = \mathbf{h}(\dot{\boldsymbol{\chi}}_i(t), \boldsymbol{B}(t)), \tag{15}$$

$\hat{\boldsymbol{\chi}}_i(t) = [\hat{\chi}_{ix}(t), \hat{\chi}_{iy}(t)]^T$ might be an affine transformed state vector, $\boldsymbol{B}(t)$ is a bulk component state vector, (on the details of the transformation, see [8], p. 268), and $\dot{\boldsymbol{\chi}}_i(t) = [\cos(\theta_i(t)), \sin(\theta_i(t))]^T$. Then $\theta_i = tan^{-1}(\hat{\chi}_{iy}(t)/\hat{\chi}_{ix}(t))$ becomes a new hidden variable in order to estimate a target. This transformation might be expected to provide better tracking performances compared with the case when this transformation is not used (see Section 4). Speech activity s changes randomly according to transition probability.

3.2. Likelihood for observations

In order to compute the importance weight, the following pseudo-likelihood function calculated with $\boldsymbol{C}^p_{xl}$ in (7) and $\hat{\boldsymbol{K}}^{-1}_{xl}$ in (9) is used;

$$L(\boldsymbol{y}_{t|t+N}|\boldsymbol{\chi}(t)) = \exp\left(-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{C}^p_{xl}\hat{\boldsymbol{K}}^{-1}_{xl}\right]\right). \tag{16}$$

Then the observation probability for the broadband signal is computed as

$$P(\boldsymbol{z}_{t|t+N}|\boldsymbol{\chi}(t)) = \prod_\omega L(\boldsymbol{y}_{t|t+N}(\omega)|\boldsymbol{\chi}(t)). \tag{17}$$

This is one of the novel points in the proposed algorithm.

Therefore, the proposed PFA is as follows:

1. Particles corresponding to each $\chi_i(t)$ are given according to the rule described in **3.1**.
2. Importance weights are computed using (17).
3. Normalize the weights.
4. Resample.

4. EXPERIMENTS

4.1. Conditions

In order to evaluate the proposed algorithm, many experiments were conducted in a medium-sized meeting room with a reverberation time of approximately 0.5 (sec.). In this section, some results of the experiments are shown. As shown in Fig. 1, there were two sound sources; one is a human and the other is a loudspeaker. The standing human uttered intermittently a number of sentences in Japanese. The loudspeaker played music continuously. A microphone array was mounted on the head of a humanoid robot HRP-2 developed in the AIST. The robot head was placed on a computer controlled turntable. Then the moving sound sources were made by rotating the turntable at a constant speed. The configuration of the robot head, the human, and the loudspeaker is shown in Fig. 1.

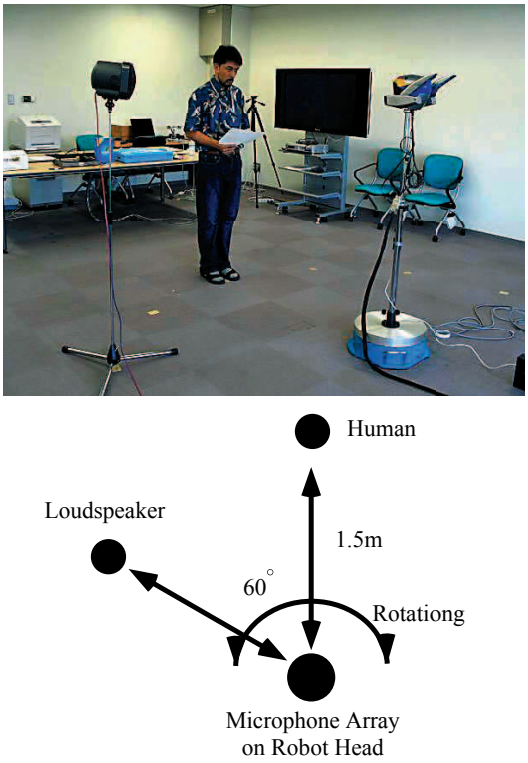
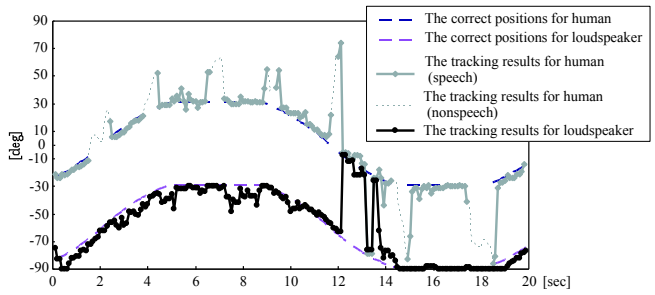


Fig. 1. Experimental setup.

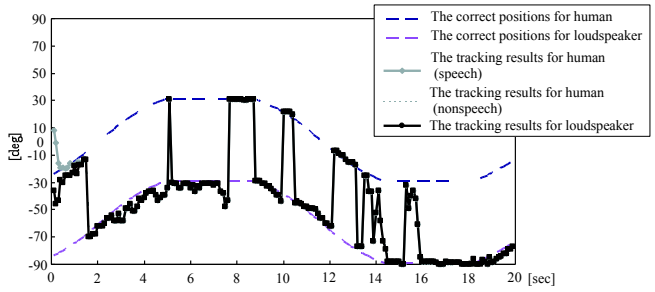
4.2. Results

Audio signals from 8 microphones were sampled at 16 kHz. The length of the Fourier transform window was 512 and the frame shift was 128. The range of frequency ω was [800, 1600] Hz. The direction of targets θ was quantized into 1-degree segment, i.e., 181 direction bins were created. We chose the averaging interval $N = 9$ (about 0.1 sec.) for balancing stability and trackability. For every 0.1 sec., the estimates were calculated and the algorithms were executed. About the particle filter, the number of particles was 100. The bulk component state vector $\boldsymbol{B}(t)$ used the one in [8] (see p.268) and the parameters of $\boldsymbol{B}(t)$ were appropriately chosen.

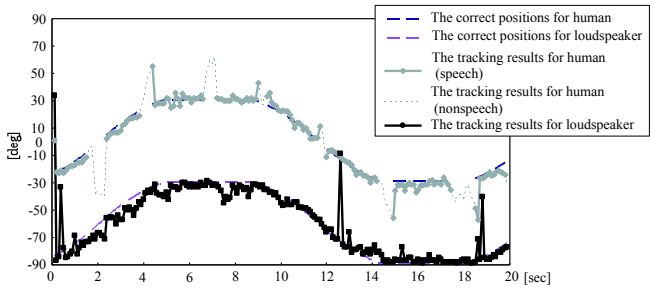
Fig. 2 shows the tracking performance results for each algorithm. For each figure, the upper and lower lines are the results with respect to the human and the loudspeaker, respectively. The horizontal and vertical axes represent time (sec.) and angles (deg.), respectively. The dashed lines represent



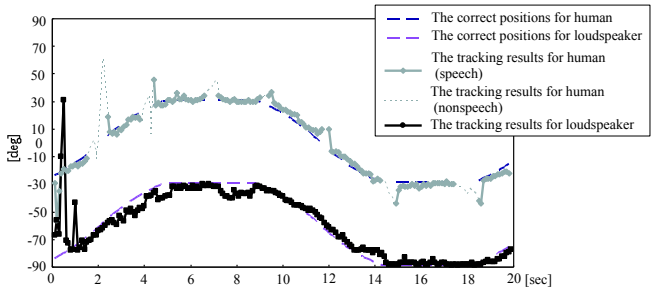
(a) Tracking results using the EMA with an initial value



(b) Tracking results using the EMA with an initial value



(c) Tracking results using the PFA



(d) Tracking results using the PFA with an idea of group tracking

Fig. 2. The tracking results.

the correct positions of the human and loudspeaker with respect to time t. The dots in the gray lines are the tracking results in the case where the human did not speak. Fig. 2(a) shows the tracking results using the EMA, where the number of iterations for every 0.1 (sec.) was $p = 5$ and the initial set positions for the human and the loudspeaker were, respectively, $\theta_h = -21$ and $\theta_s = -56$. In this case, since those initial angles are near to the initial positions of the human and the loudspeaker, the tracking performances are not so bad, where the tracking of the loudspeaker between about 12 and 14 (sec.) failed. However, if only the initial angle θ_h was changed to $\theta_h = 29$, which was far from the initial position of the human, the tracking performances were too bad (see Fig. 2(b)). When the proposed PFA was used for the tracking, the performances became better than the EMA (see Fig. 2(c)), where the bulk component state vector was not used in the particle filter. Moreover, by using a particle filter with a bulk component state vector, the proposed algorithm provided much better performances (see Fig. 2(d)), where at the beginning of the tracking, it takes a little time to get on the track.

5. CONCLUSIONS

We have proposed a particle filtering algorithm (PFA) for tracking multiple sound sources. Although the proposed PFA has an idea of the EMA based sound localization method, the proposed PFA does not possess such a drawback of the EMA that tracking results are affected by their starting points. Moreover, in order to obtain better tracking performances, a PFA with an idea of group tracking has been proposed. Experimental results have demonstrated the effectiveness of the proposed algorithm.

6. REFERENCES

[1] F. Asano and H. Asoh, "Sound source localization and separation based on the EM algorithm," Proc. SAPA 2004.

[2] H. Asoh, I. Hara, F.Asano, and K. Yamamoto, "Tracking human speech events using a particle filter," Proc. ICASSP2005, 2005.

[3] N. Checka, K. W. Wilson, and M. R. Siracusa, and T. Darerll, "Multiple person and speaker activity tracking with a particle filter," Proc. ICASSP2004, 2004.

[4] A. Doucet, N. Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, 2001.

[5] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithms," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. 36, No. 4, pp. 477-489, 1988.

[6] M. Miller and D. Fuhrmann, "Maximum-likelihood narrow-band direction finding and the EM algorithm," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. 38, No. 9, pp. 1560-1577, 1990.

[7] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Trans. Antennas Propag., Vol. AP-34, No. 3, pp. 276-280, March 1986.

[8] B. Ristic, S. Arulampalam, N. Gordon, *Beyond the Kalman Filter, Particle Filters for Tracking Applications*, Artech House, 2004.

[9] 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 Processing, Vol. 11, No. 6, pp. 826-836, November 2003.