NONCONCURRENT MULTIPLE SPEAKERS TRACKING BASED ON EXTENDED KALMAN PARTICLE FILTER

Xionghu Zhong and James R. Hopgood

Institute for Digital Communications,
School of Engineering and Electronics, University of Edinburgh, UK
x.zhong@ed.ac.uk and James.Hopgood@ed.ac.uk

ABSTRACT

Acoustic reverberation introduces multipath components into an audio signal, and therefore changes the source signal statistical properties. This causes problems for source localisation and tracking since reverberation generates spurious peaks in the time delay functions, and makes the subsequent location estimator hard to track the motion trajectory. Previous time delay based tracking methods, such as the extended Kalman filter and the particle filter, are sensitive to reverberation and are unable to follow sharp changes in the source positions. In this paper, the extended Kalman filter and the particle filter are combined to solve this problem. One of the advantages of this approach is that the optimal importance function can be obtained after extended Kalman filtering. Thus, the position samples are distributed in a more accurate area than using a prior importance function. Experiment results show that the proposed algorithm outperforms the sequential importance resampling particle filter by reducing the estimation error and following the switch of speakers quickly under a moderate reverberant environment (reverberation time $T_{60} < 0.3s$).

Index Terms— source tracking, reverberation, particle filter, extended Kalman filter

1. INTRODUCTION

Locating and tracking an acoustic source in a reverberant environment is an increasingly important research area in many applications such as teleconferencing, multimedia, hearing aids and hands-free teleconferencing systems. One popular way for this problem is the so-called indirect method, wherein the time difference of arrival (TDOA) of microphone pairs is estimated by, for example, employing generalised cross-correlation (GCC) function [1] or adaptive eigenvalue decomposition (AED) algorithm [2]. The TDOA is then used to triangulate the position using a maximum likelihood criterion [3]. This triangulation can also be achieved by a position estimator like the Kalman filter [4] or linear intersection algorithm [5]. If there is merely a time delay between the received signals and the TDOA estimates have a Gaussian distribution, and traditional indirect methods are able to track the source correctly.

However, the presence of reverberation and different kinds of noise in real-life often violate these assumptions. Thus, the performance of these TDOA estimation methods is seriously deteriorated. Recently, particle filtering is introduced into the acoustic source tracking problem to reduce the errors brought by false time delay estimates caused by the multipath reverberant components [6] and [7]. It is assumed that the received signal can be modeled by a free-field model, in which the reverberant signal is separated into direct path and multipath components. The later part is regarded as a noise term. The motion model of the speaker is then defined, and the likelihood function is constructed based on the assumption of a Gaussian distribution. Finally, the posterior distribution of the location is estimated using a particle filter. A full description of this method can be found in [6] and [7]. This sequential importance resampling (SIR) particle filter (PF) suffers from a tracking lag or even a track loss in following a sharp change of the position, which is a common case for nonconcurrent multiple speakers tracking.

The extended Kalman filter (EKF) is used in acoustic source localisation and tracking in [4]. The speaker's position is updated employing an EKF, wherein the observation and the states are associated with the TDOAs and the speaker's position separately. The results in [4] reveal that the EKF provides source accuracy superior to the linear intersection techniques. However, one drawback is that the EKF can't cope with the reverberant environment well. In this paper, we combine the particle filter and extended Kalman filter to room acoustic source tracking problem. The combined algorithm is an extended Kalman particle filter (EKPF) [8], through which the optimal importance function can be derived by estimating the posterior distribution of the states using an EKF. Thus during each iteration, samples are relocated with both the knowledge of the former state estimates and the current observations. These samples are more accurately distributed than using a prior importance function in [7] which only takes the past states into account. Furthermore, we can derive the variance of the Gaussian distribution in the likelihood function by one step prediction of the observation rather than empirical studies. These factors make the EKPF method more appropriate to the reverberant environments and complicated motion trajectories.

The rest of this paper is organised as follows. In section 2, a model for the reverberant signal and localisation problem is formulated. Section 3 summarizes the EKPF algorithm and exploits it for a tracking problem. The simulation experiments and the performance comparison with the SIR particle filter are described in section 4. Our conclusions are presented in section 5.

2. SIGNAL MODEL AND SOURCLOCALISATIONON

Let $\boldsymbol{p}_{m,i}, \mathbf{x}_t \in \mathbb{R}^3$ denote the position of ith microphone of mth microphone pair and the position of the source, respectively. The discrete time signal from a single source received can be modeled as

$$x_{m,i}(t) = s(t) * h(\boldsymbol{p}_{m,i}, \mathbf{x}_t) + n_{m,i}(t) \tag{1}$$

where $s(t)$ is the source signal, $h(\boldsymbol{p}_{m,i}, \mathbf{x}_t)$ is the overall impulse response cascading the room and the microphone channel response, $n_{m,i}(t)$ is additive noise which often assumed to be uncorrelated with the source signal and from different sensors, and $*$ denotes convolution. To formulate TDOA estimates, we rewrite the impulse

response in terms of the direct path and multipath components as

$$x_{m,i}(t) = \frac{1}{r_{m,i}} s(t - \tau_{m,i}) + s(t) * g(\boldsymbol{p}_{m,i}, \mathbf{x}_t) + v_{m,i}(t) \quad (2)$$

where $r_{m,i}$ is the distance between the source and microphone, $\tau_{m,i}$ is the direct path time delay, and $g(\boldsymbol{p}_{m,i}, \mathbf{x}_t)$ is the new impulse response which is defined as the original response minus the direct path component. In fact, this model is a free field model in that it regards the reverberation part as a noise term.

The signal model contains the parameter of interest, namely the time delay $\tau_{m,i}$. The time difference of arrival of the microphone pair can be expressed as

$$\tau_m(\mathbf{x}_t) = \tau_{m,1} - \tau_{m,2} = \frac{\|\mathbf{x}_t - \boldsymbol{p}_{m,1}\| - \|\mathbf{x}_t - \boldsymbol{p}_{m,2}\|}{c} \quad (3)$$

where c is the sound velocity, and $\|\cdot\|$ is the Euclidean norm denoting the distance between two positions. Given a series of time delay estimates $\hat{\tau}_m(t)$, the maxmum likelihood (ML) criterion [3] for location can be estimated as

$$\hat{\mathbf{x}}_t = \underset{\mathbf{x}_t}{\operatorname{argmin}} \sum_{m=1}^{M} \left(\hat{\tau}_m(t) - \tau_m(\mathbf{x}_t)\right)^2 \quad (4)$$

The evaluation of $\mathbf{x}_t$ at each time step involves the optimization of a non-linear function and necessitates the use of search methods, since no close form solution exists to equation (4).

3. EKPF FOR TRACKING

3.1. Extended Kalman filter

For the operation of the Kalman filter, we present the process and observation equation involved in the state space model. The process equation is governed by a random walk motion model, wherein the speaker is moving only under the control of the process noise

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \Delta T \mathbf{v}_t \quad (5)$$

where ΔT is the time period between two neighbouring steps, and $\mathbf{v}_t$ is white noise with variance $\mathbf{Q}_t$ representing the velocity component of the motion. According to the motion model, the transition probability density function (PDF) then can be expressed as

$$p(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)}) = \mathcal{N}(\mathbf{x}_t^{(i)}; \mathbf{x}_{t-1}^{(i)}, \mathbf{Q}_t) \quad (6)$$

To be related with the observations, we approximate the time delay $\tau_m(\mathbf{x}_t)$ with a first-order Taylor expansion [4]. That is:

$$\tau_m(\mathbf{x}_t) = \tau_m(\mathbf{x}_{t-1}) + \mathbf{c}_m^T(t)[\mathbf{x}_t - \mathbf{x}_{t-1}] + \bar{n}_t \quad (7)$$

where superscript T denotes transposepose, $\bar{n}_t = O(\mathbf{x}_t)$ is the higher order of the time delay expansion, and $\mathbf{c}_m^T(t)$ is the coefficient vector of Taylor expansion

$$\mathbf{c}_m^T(t) = \frac{1}{c}\left[\frac{\mathbf{x}_t - \boldsymbol{p}_{m,1}}{\|\mathbf{x}_t - \boldsymbol{p}_{m,1}\|} - \frac{\mathbf{x}_t - \boldsymbol{p}_{m,2}}{\|\mathbf{x}_t - \boldsymbol{p}_{m,2}\|}\right]^T_{\mathbf{x}_t = \bar{\mathbf{x}}_{t-1}} \quad (8)$$

with $\bar{\mathbf{x}}_{t-1}$ denoting the state at the last time step. Defining

$$\mathbf{C}_t = \begin{bmatrix} \mathbf{c}_1^T(t) \\ \mathbf{c}_2^T(t) \\ \vdots \\ \mathbf{c}_M^T(t) \end{bmatrix}, \hat{\boldsymbol{\tau}}_t = \begin{bmatrix} \hat{\tau}_1(t) \\ \hat{\tau}_2(t) \\ \vdots \\ \hat{\tau}_M(t) \end{bmatrix}, \boldsymbol{\tau}(\mathbf{x}_t) = \begin{bmatrix} \tau_1(\mathbf{x}_t) \\ \tau_2(\mathbf{x}_t) \\ \vdots \\ \tau_M(\mathbf{x}_t) \end{bmatrix} \quad (9)$$

then following equation (7), define:

$$\mathbf{y}_t = \hat{\boldsymbol{\tau}}_t - \boldsymbol{\tau}(\bar{\mathbf{x}}_{t-1}) + \mathbf{C}_t \bar{\mathbf{x}}_{t-1} \quad (10)$$

such that the observation equation can be written as

$$\mathbf{y}_t \approx \mathbf{C}_t \mathbf{x}_t + \mathbf{n}_t \quad (11)$$

Here $\hat{\boldsymbol{\tau}}_t$ is estimated by GCC method, and $\mathbf{n}_t$ is the measurement noise which is assumed to be zero-mean Gaussian process with a variance of $\mathbf{R}_t$ representing the higher order expansion of the time delay vector.

The key idea of the EKF is equation (7), the implementation of a minimum square error (MMSE) estimator through Taylor series expansion of the nonlinear functions around the estimates. Regarding equations (5) and (11) as process and observation equation respectively, the Gaussian approximation of the posterior distribution of the states by EKF is easily derived as following:

$$\bar{\mathbf{x}}_{t|t-1} = \bar{\mathbf{x}}_{t-1|t-1} \quad (12a)$$

$$\mathbf{P}_{t|t-1} = \mathbf{P}_{t-1|t-1} + \Delta T^2 \mathbf{Q}_t \quad (12b)$$

$$\bar{\mathbf{y}}_{t|t-1} = \mathbf{C}_t \bar{\mathbf{x}}_{t|t-1} \quad (12c)$$

$$\mathbf{S}_t = \mathbf{R}_t + \mathbf{C}_t \mathbf{P}_{t|t-1} \mathbf{C}_t^T \quad (12d)$$

$$\mathbf{K}_t = \mathbf{P}_{t|t-1} \mathbf{C}_t^T \mathbf{S}_t^{-1} \quad (12e)$$

$$\bar{\mathbf{x}}_{t|t} = \bar{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \tau(\bar{\mathbf{x}}_{t|t-1})) \quad (12f)$$

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{K}_t \mathbf{C}_t \mathbf{P}_{t|t-1} \quad (12g)$$

Evidently, the filtered distribution of the states is $p(\mathbf{x}_t|\mathbf{x}_{0:t-1}\, \mathbf{y}_{1:t}) = \mathcal{N}(\mathbf{x}_t; \bar{\mathbf{x}}_{t|t}, \mathbf{P}_{t|t})$. This distribution is used as the importance function in the particle filter in next subsection.

3.2. Tracking algorithm

As the introduction of the particle filter can easily be found in many open literature [8],[9], here we only summarize, without deduction, the particle filter algorithm combining with EKF to track the source trajectory.

Extended Kalman particle filter

1. Initialization: $t = 0$

• For $i = 1, \ldots, N$, draw the position samples (particles) $\mathbf{x}_0^{(i)}$ from the prior $p(\mathbf{x}_0)$.

2. For $t = 1, 2, \ldots$

• For $i = 1, \ldots, N$:

Importance sampling step

- Update the particles with the EKF according to (12).
- Sample $\hat{\mathbf{x}}_t^{(i)} \sim q(\hat{\mathbf{x}}_t^{(i)}|\mathbf{x}_{0:t-1}^{(i)}, \mathbf{y}_{1:t}) \triangleq \mathcal{N}(\hat{\mathbf{x}}_t; \bar{\mathbf{x}}_{t|t}^{(i)}, \mathbf{P}_{t|t}^{(i)})$
- Set $\hat{\mathbf{x}}_{0:t}^{(i)} \triangleq (\mathbf{x}_{0:t-1}^{(i)}, \hat{\mathbf{x}}_t^{(i)})$ and $\hat{\mathbf{P}}_{0:t}^{(i)} \triangleq (\mathbf{P}_{0:t-1}^{(i)}, \mathbf{P}_{t|t}^{(i)})$
- Evaluate the importance weights

$$\omega_t^{(i)} \propto \frac{p(\mathbf{y}_t|\hat{\mathbf{x}}_t^{(i)}) p(\hat{\mathbf{x}}_t^{(i)}|\mathbf{x}_{t-1}^{(i)})}{q(\hat{\mathbf{x}}_t^{(i)}|\mathbf{x}_{0:t-1}^{(i)}, \mathbf{y}_{1:t})} \quad (13)$$

where $q(\cdot)$ is the importance function and will be given by equation (15).

• For $i = 1, \ldots, N$:

(a) Normalizing the importance weight

$$\widetilde{\omega}_t^{(i)} = \frac{\omega_t^{(i)}}{\sum_{i=1}^{N} \omega_t^{(i)}} \quad (14)$$

(b) Selection step

Multiply/Discard particles $(\hat{\mathbf{x}}_{0:t}^{(i)}, \hat{\mathbf{P}}_{0;t}^{(i)})$ with high/low importance weights $\widetilde{\omega}_t^{(i)}$.

3. Output: MSE estimate of the state $E(\mathbf{x}_t) = \sum_{i=1}^{N} \widetilde{\omega}_t^{(i)} \hat{\mathbf{x}}_t^{(i)}$.

Importance function: There are several choices for selecting the importance function (proposal distribution) [9]. In this paper, we use optimal importance function, which is conditional upon the trajectory $\mathbf{x}_{0:t-1}^{(i)}$ and the observations $\mathbf{y}_{0:t}$. This is formed as

$$q(\mathbf{x}_t^{(i)}|\mathbf{x}_{0:t-1}^{(i)}, \mathbf{y}_{1:t}) = p(\mathbf{x}_t^{(i)}|\mathbf{x}_{0:t-1}^{(i)} \mathbf{y}_{1:t}) \tag{15}$$

where $p(\mathbf{x}_t^{(i)}|\mathbf{x}_{0:t-1}^{(i)} \mathbf{y}_{1:t}) = \mathcal{N}(\mathbf{x}_t^{(i)}; \bar{\mathbf{x}}_{t|t}^{(i)}, \mathbf{P}_{t|t}^{(i)})$. $\bar{x}_{t|t}^{(i)}$ and $\mathbf{P}_{t|t}^{(i)}$ are the filtered state and variance using EKF respectively.

Likelihood function: Because of reverberation and noise, the likelihood model $p(\mathbf{y}_t|\mathbf{x}_t^{(i)})$ can no longer be expressed in a simple way. Let K be the number of potential delays obtained from the time-delay estimation function. Using one-step prediction of the observation and following the approaches used in [6], the likelihood function of the pth microphone pair can be

$$f_p(\mathbf{y}_t^{(i)}|\mathbf{x}_t^{(i)}) = \sum_{\kappa=1}^{K} q_\kappa \mathcal{N}(\mathbf{y}_t^{(i)}; \mathbf{y}_{t|t-1}^{(i)}, \mathbf{S}_t^{(i)}) + q_0 \tag{16}$$

where $\mathbf{y}_{t|t-1}^{(i)}$ and $\mathbf{S}_t^{(i)}$ is the prediction mean and variance of observation respectively. $q_\kappa < 1$, $\kappa = 1, \ldots, K$ is the prior probability denoting that the κth potential time delay is associated the true position, and $q_0 < 1$ denotes the probability that none of the delays will contribute to the true source. We assume that the measurements across all microphone pairs are independent. If P sensor pairs are used, the complete likelihood function becomes

$$p(\mathbf{y}_t^{(i)}|\mathbf{x}_t^{(i)}) = \prod_{p=1}^{P} f_p(\mathbf{y}_t^{(i)}|\mathbf{x}_t^{(i)}) \tag{17}$$

4. SIMULATION EXPERIMENTS

In this section, the performance of the algorithm is illustrated on two typical motion trajectories; motion as a line or "switch-speaker". For the line trajectory case, there is only one speaker moving along the diagonal line, which is marked as trajectory 1 in Figure 1. The switch-speaker case involves a source change at the time center of the whole voice period; the motion orientation and the break position are denoted in trajectory 2. The length of the audio file for the both cases is $7.6s$, and the corresponding reverberant signal at each of microphones are generated using the image method [10]. In our experiment 4 microphone pairs each with a separation of $40cm$ at symmetrical position were employed. The room dimesion is $5m \times 5m \times 2.7m$ with background noise yielding a SNR level of 30dB. Different reflection coefficients are set from 0 to 0.9 with an increment of 0.1 to simulate various reverberant environments. The reverberation time T_{60} corresponding to the different reflection coefficients can be found in table 1. The audio signal is split into 120 frames for each trajectory. The whole experimental setup is depicted in Fig. 1.

Table 1. Reflection coefficient β and its corresponding reverberation time T_{60}.

β	0	0.1	0.2	0.3	0.4
T_{60}	0	0.11	0.12	0.13	0.15
β	0.5	0.6	0.7	0.8	0.9
T_{60}	0.19	0.23	0.28	0.41	0.56

Here we give the tracking results for both cases under a reflection coefficient of 0.5, that is $T_{60} = 0.19s$. The tracking algorithm is run with $N = 100$ particles, and particles are initialised around the center position of the room. q_0 was set to 0.4 according the setting in [7] . Fig. 2 shows the speech signal from the single source and the tracking result for line trajectory, which is represented by trajectory 1. Both the SIR particle filter and the EKPF track the source trajectory well.

Fig. 3 shows the speech signal from two nonconcurrent speakers denoted by trajectory 2 and the estimated result. This result demonstrate that EKPF is able to track the trajectory with a satisfactory accuracy, and quickly locks on to the source when the source switches.

For testing the performance of the algorithm in different reverberant environment, a Monte Carlo experiment with 50 runs is implemented under the different reflection coefficients. Fig. 4 gives the root mean square error (RMSE) [7] obtained from each trajectory with the SIR particle filter and the EKPF. As depicted in the figure, both the algorithms do well for line trajectory, but the tracking result of EKPF is much better for switch-speaker trajectory. This shows that our method is more effective for the sharp change of the position or the switch of the speakers in the moderate reverberant environment (reflection coefficients ≤ 0.6). However the performance degrades quickly when the reflection coefficient is greater than 0.6. This is because GCC algorithm collapses under the strong reverberation environment, and thus the observed TDOAs are far away from the true time delays.

5. CONCLUSIONS

A new approach to source tracking in a reverberant environment is presented in this paper. By linearizing the time delay function and using an extended Kalman filter, the optimal importance function can be derived. The particles thus can be relocated in a more accu-

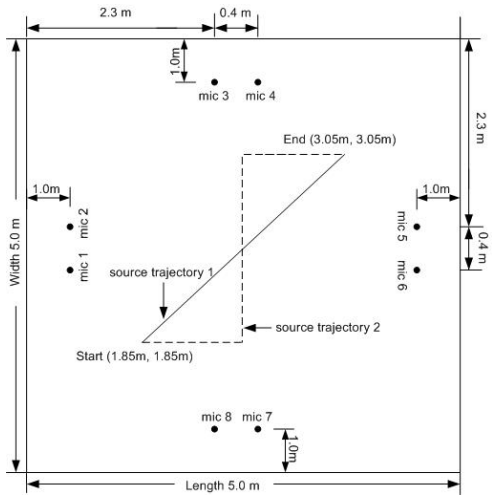

Fig. 1. Experiment setup. Black dots numbered from 1 to 8 are the position of microphones, solid line and dash-dotted line represent the line trajectory and switch-speaker trajectory respectively.

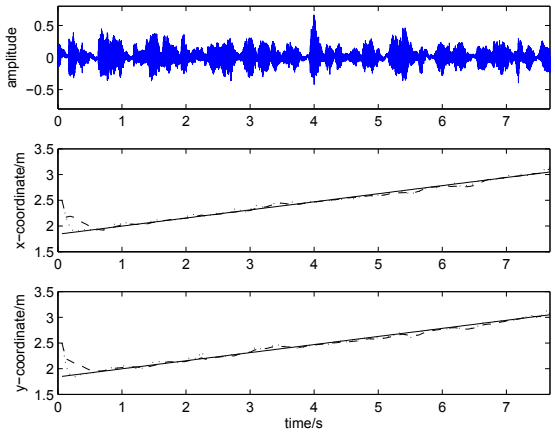

Fig. 2. Line trajectory (trajectory 1) estimation result under the reflection coefficient 0.5. Solid lines are true position, dashed lines and dotted lines denote the estimates by PF and EKPF respectively.

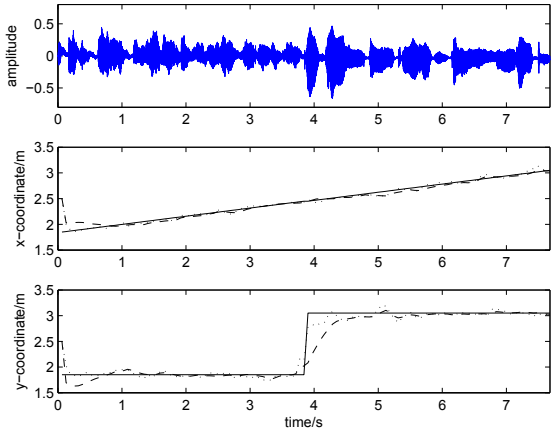

Fig. 3. Switch speaker (trajectory 2) estimation result under the reflection coefficient 0.5. Solid lines are true position, dashed lines and dotted lines denote the estimates by PF and EKPF respectively.

rate area, and helps the algorithm easily to recover from any tracking loss and detect the switch of speakers. The simulation results show that the tracking performance is robust against the reverberation and background noise, and even in a complicated motion case. As an initial stage of multiple simultaneously active sources tracking, nonconcurrent speakers tracking provide a lot of knowledge for our future work.

6. REFERENCES

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

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

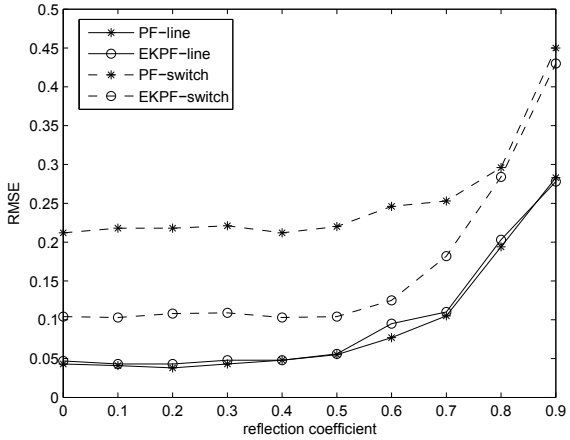

Fig. 4. Average RMSE for proposed method and general PF with different trajectories *vs.* different reflection coefficients. Solid lines and dashed lines are for line trajectory (trajectory 1) and switch trajectory (trajectory 2) respectively; circle and star denote the estimates based on EKPF and PF separately.

[3] M. Brandstein and D. Ward, *Microphone Arrays. Signal Processing Techniques and Applications*, Berlin, Germany: Springer-Verlag, 2001.

[4] Ulrich Klee, Tobias, and John McDonough, "Kalman filters for time delay of arrival-based source localization," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–15, 2006.

[5] M.S. Brandstein, J.E. Adcock, and H.F. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE trans. on Speech and Audio Processing*, vol. 5, no. 1, pp. 45–50, Jan. 1997.

[6] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 3021–3024, May 2001.

[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. on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, Nov. 2003.

[8] R. van der Merwe, A. Doucet, N. de Freitas, and E. A. Wan, "The unscented particle filter," Tech. Rep., Cambridge University Engineering Department, Aug. 2000.

[9] Godsill S. Doucet, A. and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, pp. 197–208, 2000.

[10] J. B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Jul. 1979.