NONLINEAR FILTERING FOR SPEAKER TRACKING IN NOISY AND REVERBERANT ENVIRONMENTS

J. Vermaak and A. Blake

Microsoft Research Cambridge, Cambridge CB2 3NH, UK
{jacov, ablake}@microsoft.com

ABSTRACT

This paper addresses the problem of speaker tracking in a noisy and reverberant environment using time delay of arrival (TDOA) measurements at spatially distributed microphone pairs. The tracking problem is posed within a state-space estimation framework, and models are developed for the speaker motion and the likelihood of the speaker location in the light of the TDOA measurements. The resulting state-space model is non-linear and non-Gaussian, and consequently no closed-form solutions exist for the filtering distributions required to perform tracking. Here Sequential Monte Carlo (SMC) methods are applied to approximate the true filtering distribution with a set of samples. The resulting tracking algorithm requires no triangulation, is computationally efficient, and can straightforwardly be extended to track multiple speakers.

1. INTRODUCTION

Speaker tracking involves determining and following the position of a speaker within some acoustic environment. The need for robust speaker tracking is becoming increasingly important with the increase in multimedia applications. A popular strategy for speaker localisation is performing triangulation based on time delay of arrival (TDOA) measurements at spatially distributed microphone pairs [1, 2]. This works well in acoustic environments characterised by low noise and reverberation, but breaks down even in moderately reverberant conditions. Some heuristic modifications to reduce the effects of reverberation have been proposed in *e.g.* [3, 4, 5], but these are reliant on either specific array configurations, or rather strong assumptions about the source signals and acoustic environment, and are far from robust in general scenarios.

In the approach taken here the speaker tracking problem is formulated within a state-space estimation framework. This framework requires a model for the speaker motion and a likelihood model for the speaker location in the light of the TDOA measurements. To cope with the effects of reverberation a multi-hypothesis likelihood model similar to those proposed before for radar [6] and vision based [7, 8] tracking is developed here. Tracking then amounts to estimating the distribution of the speaker location recursively in time based on all the past and present TDOA measurements. The resulting state-space model is non-linear and non-Gaussian, and consequently no closed-form solutions exist for the filtering distributions. Under these circumstances Sequential Monte Carlo (SMC) methods [9], also known as particle filtering methods, provide accurate, yet simple and computationally efficient, estimation strategies.

The multi-hypothesis model together with the SMC estimation strategy provide a number of important advantages over more conventional speaker localisation strategies. Most important is the ability of the system to operate robustly in adverse acoustic environments. Furthermore, the methodology can straightforwardly be extended to track multiple speakers. More subtly, the need to perform triangulation, which is especially susceptible to reverberation, is completely eliminated. This is due to the fact that in the likelihood model the distribution of the TDOA measurements is conditioned on the hypothesised source location, or equivalently, the corresponding hypothesised TDOA values. The non-linear transformation between the source location and the corresponding TDOA values is elegantly and accurately accommodated within the SMC estimation framework. Finally, the strategy is applicable to arbitrary array configurations.

The remainder of this paper is organised as follows. Section 2 formulates a model for the source motion. Section 3 describes the measurement system and develops the likelihood model for the source location based on the TDOA measurements. The SMC tracking algorithm is briefly described in Section 4, with a tracking example following in Section 5. Finally, the findings of the paper are summarised in Section 6.

2. SOURCE MODEL

The problem considered is that of tracking a source in the XY-plane. It should be stated, however, that the methodology can easily be extended to perform 3D tracking. The source state at discrete time k is defined as $\boldsymbol{\alpha}_k \triangleq (x_k, y_k, \dot{x}_k, \dot{y}_k)$, where (x_k, y_k) and $(\dot{x}_k, \dot{y}_k)$ are the source position and velocity, respectively. The source motion in the X and Y coordinates are assumed to be independent and identical. Although this is a rather strong assumption, it was found to work well in practice, even for trajectories that clearly violate this assumption (see Section 5). The source motion is modelled as a Langevin process, which in the X coordinate is specified by $\frac{d^2x}{dt^2} + \beta_x \frac{dx}{dt} = F_x$, with β_x the rate constant and F_x a thermal excitation process. It corresponds to the discrete process

$$\dot{x}_k = a_x \dot{x}_{k-1} + b_x F_{x_k}, \quad x_k = x_{k-1} + \Delta T \dot{x}_k, \tag{1}$$

where $F_{x_k} \overset{iid}{\sim} \mathcal{N}(0,1)$, ΔT is the discretisation time step, and

$$a_x = \exp(-\beta_x \Delta T), \quad b_x = \overline{v}_x \sqrt{1 - a_x^2},$$

with $\overline{v}_x$ the steady-state root-mean-square velocity. Thus, the source dynamics follow a first-order Markov process of the form

$$p(\boldsymbol{\alpha}_k | \boldsymbol{\alpha}_{k-1}) = p(x_k | x_{k-1}, \dot{x}_k)\, p(\dot{x}_k | \dot{x}_{k-1}) \\ p(y_k | y_{k-1}, \dot{y}_k)\, p(\dot{y}_k | \dot{y}_{k-1}). \tag{2}$$

The chosen dynamical model is general enough to capture many different kinds of motion. Alternatively, more accurate models for human motion can be designed and trained on a set of representative motion trajectories. However, the performance of such models often degrades rapidly when encountering trajectories that deviate from those in the training set. In the experiments performed here the simple model was retained, with its parameters fixed to $\beta_x = 10\ s^{-1}$ and $\overline{v}_x = 1\ ms^{-1}$, and was found to give good results.

3. MEASUREMENT MODEL

Measurement System. The measurement system consists of M spatially distributed microphone pairs, with all the microphones omni-directional. For the m-th pair the microphone locations are specified by $\mathbf{m}_1^{(m)}$ and $\mathbf{m}_2^{(m)}$, respectively. At each microphone pair candidates for the TDOA are taken to be the positions of the peaks in the generalised cross-correlation function (GCCF) [10] between the signals received at the microphones comprising the pair. Apart from the true source, "ghost sources" due to reverberation also lead to the peaks in the GCCF. Peaks not due to the true source will be referred to as clutter. Suppressing the time index k, the measurement vector is defined as $\mathbf{D} \triangleq (\mathbf{D}^{(1)}, \ldots, \mathbf{D}^{(M)})$, with $\mathbf{D}^{(m)} \triangleq (D_1^{(m)}, \ldots, D_{N^{(m)}}^{(m)})$ the $0 \leq N^{(m)} \leq N_{\max}$ candidate TDOA measurements at the m-th microphone pair. The maximum TDOA that can be measured at the m-th microphone pair is $D_{\max}^{(m)} = c^{-1}\|\mathbf{m}_1^{(m)} - \mathbf{m}_2^{(m)}\|$, with c the speed of sound (normally taken to be 342 ms^{-1}), and $\|\cdot\|$ the Euclidean norm. The true TDOA associated with the source state $\boldsymbol{\alpha}$ at the m-th microphone pair is given by

$$D_{\boldsymbol{\alpha}}^{(m)} = c^{-1}\left(\left\|\mathbf{p}_{\boldsymbol{\alpha}} - \mathbf{m}_1^{(m)}\right\| - \left\|\mathbf{p}_{\boldsymbol{\alpha}} - \mathbf{m}_2^{(m)}\right\|\right), \qquad (3)$$

where $\mathbf{p}_{\boldsymbol{\alpha}} \triangleq (x, y)$ is the source location.

Likelihood Model. The aim is to develop a likelihood model for the source state based on the TDOA measurements, *i.e.* $p(\mathbf{D}|\boldsymbol{\alpha})$. Given the state of the source $\boldsymbol{\alpha}$, the vector of the true TDOAs at each of the microphone pairs $\mathbf{D}_{\boldsymbol{\alpha}} \triangleq (D_{\boldsymbol{\alpha}}^{(1)}, \ldots, D_{\boldsymbol{\alpha}}^{(M)})$ can be computed using (3). Since this is a deterministic mapping the likelihood satisfies $p(\mathbf{D}|\boldsymbol{\alpha}) = p(\mathbf{D}|\mathbf{D}_{\boldsymbol{\alpha}})$. The latter form will be used in the development that follows.

Suppressing the superscript (m), consider first the measurements at any one of the microphone pairs. These are assumed to be independent, so that

$$p\left(\mathbf{D}|\, D_{\boldsymbol{\alpha}}\right) = \prod_{i=1}^{N} p\left(D_i|\, D_{\boldsymbol{\alpha}}\right). \qquad (4)$$

In practice, however, clutter measurements due to reverberation are expected to be strongly coherent with the true source, thus violating the independence assumption. Accurate modelling of reverberation requires detailed knowledge about the composition and acoustic properties of the environment, which is difficult to obtain in practice, and thus not attempted here. Notwithstanding, the model still performed well. Of the measurements at most one is associated with the true source, while the rest is associated with clutter. To distinguish between the two cases a classification label c_i is introduced, such that $c_i = T$ if D_i is associated with the true source, and $c_i = C$ if D_i is associated with clutter. The likelihood for a measurement from the true source is taken to be

$$p\left(D_i|\, D_{\boldsymbol{\alpha}}, c_i = T\right) = c_{\boldsymbol{\alpha}} \mathcal{N}\left(D_i; D_{\boldsymbol{\alpha}}, \sigma_D^2\right) \mathbb{I}_{\mathcal{D}}\left(D_i\right),$$

where $\mathcal{D} \triangleq [-D_{\max}, D_{\max}]$ is the set of admissible TDOA values for the microphone pair, $c_{\boldsymbol{\alpha}}$ is a normalising constant that can be obtained using the Gaussian error function, and $\mathbb{I}_{\mathcal{D}}(\cdot)$ is the indicator function for the set $\mathcal{D}$. Thus, within the range of admissible TDOA values, the measurement is assumed to be the true TDOA corrupted with additive Gaussian observation noise of variance σ_D^2. Empirical studies showed this to be a reasonable assumption. Similar to what was done in *e.g.* [8], the likelihood for measurements associated with clutter is taken to be

$$p\left(D_i|\, c_i = C\right) = \mathcal{U}_{\mathcal{D}}\left(D_i\right).$$

Thus, the clutter is assumed to be uniformly distributed within the admissible interval, independent of the true source TDOA. For N measurements there are $N+1$ possible hypotheses. Either all the measurements are due to clutter, or one of the measurements correspond to the true source, and the rest to clutter. More formally,

$$\begin{aligned} &\mathcal{H}_0 \triangleq \{c_i = C : i = 1, \ldots, N\} \\ &\mathcal{H}_i \triangleq \{c_i = T, c_j = C : j = 1, \ldots, N, j \neq i\}, \end{aligned}$$

with $i = 1, \ldots, N$. The likelihoods for these hypotheses follow straightforwardly from (4), and are given by

$$\begin{aligned} &p\left(\mathbf{D}|\, \mathcal{H}_0\right) = \mathcal{U}_{\mathcal{D}^N}\left(\mathbf{D}\right) \\ &p\left(\mathbf{D}|\, D_{\boldsymbol{\alpha}}, \mathcal{H}_i\right) = c_{\boldsymbol{\alpha}} \mathcal{N}\left(D_i; D_{\boldsymbol{\alpha}}, \sigma_D^2\right) \mathbb{I}_{\mathcal{D}}\left(D_i\right) \mathcal{U}_{\mathcal{D}^{N-1}}\left(\mathbf{D}_{-i}\right), \end{aligned}$$

where $\mathbf{D}_{-i}$ is $\mathbf{D}$ with D_i removed. However, for any set of measurements, the correct hypothesis is not known beforehand, and the final likelihood for the microphone pair should be obtained by summing over all the possible hypotheses, *i.e.*

$$p\left(\mathbf{D}|\, D_{\boldsymbol{\alpha}}\right) = \sum_{i=0}^{N} q_i p\left(\mathbf{D}|\, D_{\boldsymbol{\alpha}}, \mathcal{H}_i\right), \qquad (5)$$

where $q_i \triangleq p(\mathcal{H}_i|D_{\boldsymbol{\alpha}})$, $i = 0, \ldots, N$, are the prior probabilities of the hypotheses. These are commonly assumed to be equal and independent of the true source TDOA $D_{\boldsymbol{\alpha}}$. They can, however, be adjusted to reflect the confidence in the measurements. In the case where no measurements are available the likelihood is simply set to $p(\mathbf{D}|D_{\boldsymbol{\alpha}}) \propto 1$. Thus, no new information about the source state can be obtained from the measurements.

The extension of the likelihood for a single microphone pair in (5) to M microphone pairs is straightforward. The transformation of source state to the corresponding true TDOAs at the microphone pairs effectively decouples the measurements, so that the likelihood for M microphone pairs becomes

$$p\left(\mathbf{D}|\, \mathbf{D}_{\boldsymbol{\alpha}}\right) = \prod_{m=1}^{M} p\left(\mathbf{D}^{(m)}\middle|\, D_{\boldsymbol{\alpha}}^{(m)}\right), \qquad (6)$$

where each $p(\mathbf{D}^{(m)}|D_{\boldsymbol{\alpha}}^{(m)})$ is computed according to (5). Note that no assumptions have been made regarding relations between the individual microphone pairs, so that the likelihood model is applicable to arbitrary array configurations.

4. TRACKING ALGORITHM

The general tracking problem involves the recursive estimation of the filtering distribution $p(\boldsymbol{\alpha}_k|\mathbf{D}_{1:k})$, with $\mathbf{D}_{1:k} \triangleq (\mathbf{D}_1, \ldots, \mathbf{D}_k)$, from which estimates of the source state can be obtained. The general recursions to compute the filtering distribution are given by

$$p\left(\boldsymbol{\alpha}_k|\,\mathbf{D}_{1:k-1}\right) = \int p\left(\boldsymbol{\alpha}_k|\,\boldsymbol{\alpha}_{k-1}\right) p\left(\boldsymbol{\alpha}_{k-1}|\,\mathbf{D}_{1:k-1}\right) d\boldsymbol{\alpha}_{k-1}$$
$$p\left(\boldsymbol{\alpha}_k|\,\mathbf{D}_{1:k}\right) \propto p\left(\mathbf{D}_k|\,\boldsymbol{\alpha}_k\right) p\left(\boldsymbol{\alpha}_k|\,\mathbf{D}_{1:k-1}\right). \tag{7}$$

The first, or prediction, step uses the dynamic model in (2) and the filtering distribution at the previous time step to compute the one-step ahead prediction distribution of the state. This then acts as the prior for the state in the second, or update, step where it is combined with the likelihood in (6) to obtain the desired filtering distribution.

For the model considered here no closed-form solutions exist for the general recursions in (7). Furthermore, analytic approximations, like the Extended Kalman Filter, fail due to the inherent multi-modality of the problem. Under these circumstances one particularly attractive solution strategy is Sequential Monte Carlo (SMC), or particle filtering, methods [9]. These methods are conceptually simple, computationally efficient, and do not degrade in performance as the dimensionality of the state-space increases. They are also well-suited to cope with the multi-modality due to clutter.

Standard particle filtering is essentially a Monte Carlo implementation of the recursions in (7). The filtering distribution is approximated by a large number of samples, or particles, with associated importance weights. At every time step each of the particles is propagated according to the dynamical model (prediction) and reweighted with its likelihood (filtering). The particles are then resampled according to their new importance weights to ensure a uniform weight distribution.

One limitation of standard particle filtering is the fact that particles are propagated without taking account of the new measurement. Thus, many may be needed to accurately represent the filtering distribution. This is especially the case for narrow likelihood functions, or cases where the likelihood has significant mass in the tails of the prior. The Auxiliary Particle Filter (APF) [11] solves this problem by resampling the particles using an importance function that incorporates some knowledge of the new measurement prior to propagation. This effectively directs particles towards the modes, and dramatically reduces the number needed to accurately represent the filtering distribution. This is the strategy adopted here.

5. SIMULATION EXAMPLE

Here the algorithm performance is illustrated on a particularly difficult artificial tracking problem. The simulated acoustic environment is depicted in Figure 1. The dimensions of the enclosure are 3 m × 3 m × 2.5 m, with a reverberation time of 0.3 s and a background noise level of 30 dB. Two omni-directional microphone pairs were used, each with a separation of 60 cm. For the source, also omni-directional, the utterance *"Draw every outer line first, then fill in the interior."* was taken from the TIMIT database. The clean signal at each of the microphones was obtained using the imaging method [12], with the source following a simulated semi-circular trajectory. These signals were subsequently corrupted by adding white Gaussian noise of the desired level.

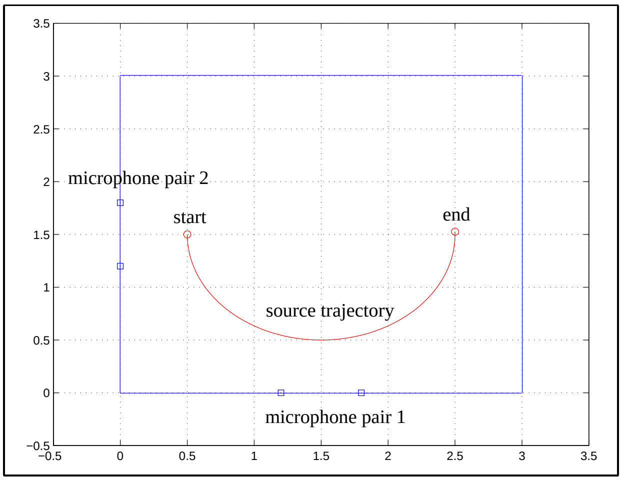

Fig. 1. Experimental setup for evaluating the tracking performance.

Figure 2 shows the original speech signal and the evolution of the GCCF at each of the microphone pairs, computed every 32 ms over frames of 64 ms. From these graphs it is clear that the GCCF exhibits multiple peaks, and that the strongest peak at each time step is not necessarily associated with the true source (see Figure 3). In fact, the peak associated with the true source often disappears for one or a number of time steps. Also, contrary to expectation, the clutter appears to be largely uncorrelated with the true source, thus validating (at least empirically) the likelihood model.

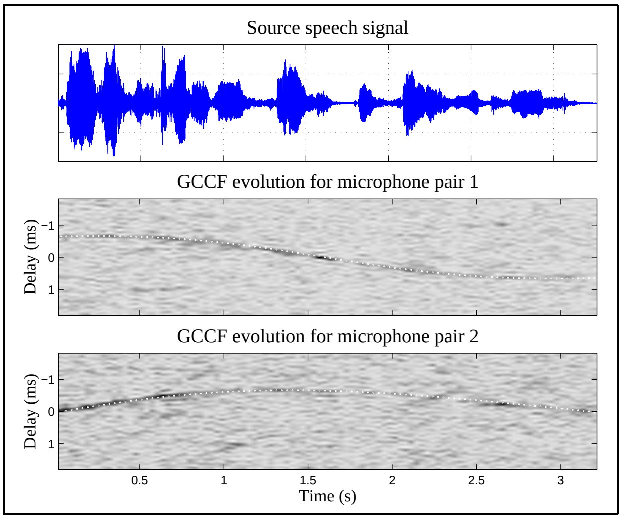

Fig. 2. Source speech signal and GCCF evolution at each of the microphone pairs. The GCCF is overlayed by the true TDOA (white dots). At each time step the GCCF exhibits multiple peaks, with the strongest peak not necessarily associated with the true source. Furthermore, the clutter is largely uncorrelated with the true source.

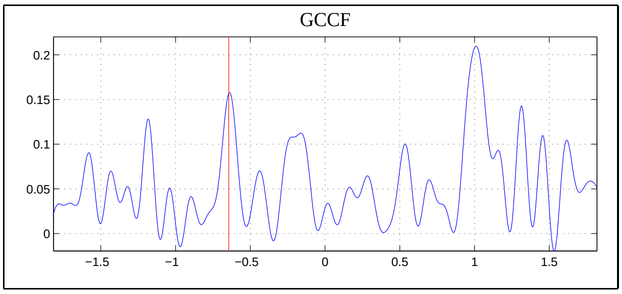

Fig. 3. GCCF for the first microphone pair at $t = 0.5\ s$. The existence of multiple peaks due to reverberation is clear, with the peak corresponding to the true source (indicated by the vertical line) being notably smaller than the largest peak.

The tracking algorithm was run with $N = 50$ particles. The particles were uniformly randomly initialised within the enclosure, with zero initial velocity. The parameters for the motion model were fixed to the values in Section 2. A maximum of ten TDOA measurements were allowed for each microphone pair. The probability for $\mathcal{H}_0$ was set to $q_0 = 0.3$, reflecting the fact that the true measurement was often not among the candidates. All the other hypotheses were assumed to be equally likely. The standard deviation of the TDOA measurement error was set to $\sigma_D = 2T_s$, with $T_s = 125\ \mu s$ the sampling period. This value was empirically determined from training data. The tracking, however, proved to be robust to changes in the parameters of the motion and likelihood models. At each time step an estimate of the source position was computed as the mean of the particles. The tracking results are depicted in Figure 4. The particles quickly lock on to the source, and follow its trajectory to a satisfactory degree of accuracy. This is remarkable given that the particles were randomly initialised, and that no effort has been made to compensate for the background noise or the reverberation.

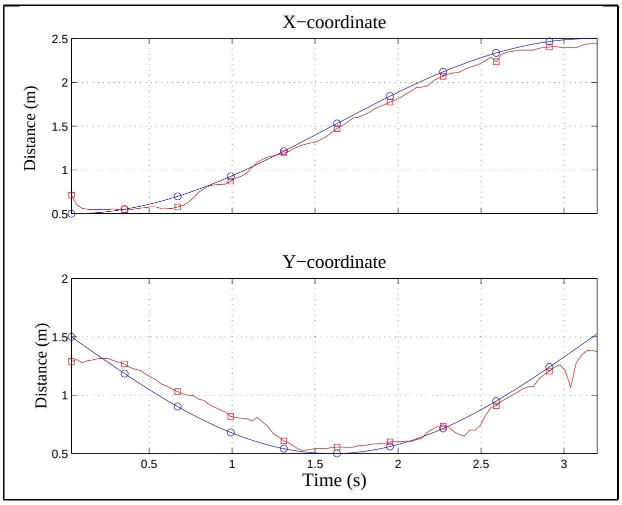

Fig. 4. True (circles) and estimated (squares) source trajectory. The randomly initialised tracking algorithm quickly locks on to the true source and follows its trajectory to a satisfactory degree of accuracy. Due to the array configuration tracking is better in the X coordinate.

6. CONCLUSIONS

This paper developed a SMC method based on the APF to perform speaker tracking in a noisy and reverberant environment using TDOA measurements at a number of spatially distributed microphone pairs. Models were developed for the speaker motion and the likelihood of the speaker location in the light of the TDOA measurements. The latter elegantly accounts for the multiple hypotheses due to clutter measurements resulting from reverberation. With as few as 50 particles the tracking performance proved to be robust under challenging acoustic conditions.

7. REFERENCES

[1] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer, Speech and Language*, vol. 11, no. 2, pp. 91–126, 1997.

[2] H. F. Silverman and E. Kirtman, "A two-stage algorithm for determining talker location from linear microphone array data," *Computer Speech and Language*, vol. 6, pp. 129–152, 1992.

[3] M. S. Brandstein, "Time-delay estimation of reverberant speech exploiting harmonic structure," *Journal of the Acoustic Society of America*, vol. 105, no. 5, pp. 2914–2919, 1999.

[4] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, 1997, pp. 375–378.

[5] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, 1997, pp. 187–190.

[6] N. Gordon, "A hybrid bootstrap filter for target tracking in clutter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 33, no. 1, pp. 353–358, 1997.

[7] M. Isard and A. Blake, "CONDENSATION – conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 28, no. 1, pp. 5–28, 1998.

[8] J. MacCormick and A. Blake, "Probabilistic exclusion and partitioned sampling for multiple object tracking," *International Journal of Computer Vision*, vol. 39, no. 1, pp. 57–71, 2000.

[9] A. Doucet, J. F. G. de Freitas, and N. J. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, New York, 2000, To Appear.

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

[11] M. K. Pitt and N. Shephard, "Filtering via simulation: Auxiliary particle filter," *Journal of the American Statistical Association*, vol. 94, pp. 590–599, 1999.

[12] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.