MODELING OF MOTION DYNAMICS AND ITS INFLUENCE ON THE PERFORMANCE OF A PARTICLE FILTER FOR ACOUSTIC SPEAKER TRACKING

Eric A. Lehmann, Anders M. Johansson and Sven Nordholm

Western Australian Telecommunications Research Institute, Perth, Australia
{Eric.Lehmann,ajh}@watri.org.au

ABSTRACT

Methods for acoustic speaker tracking attempt to localize and track the position of a sound source in a reverberant environment using the data received at an array of microphones. This problem has received significant attention over the last few years, with methods based on a particle filtering principle perhaps representing one of the most promising approaches. As a Bayesian filtering technique, a particle filter relies on the definition of two main concepts, namely the measurement process and the transition equation (target dynamics). Whereas a significant research effort has been devoted to the development of improved measurement processes, the influence of the dynamics formulation on the resulting tracking accuracy has received little attention so far. This paper provides an insight into the dynamics modeling aspect of particle filter design. Several types of motion models are considered, and the performance of the resulting particle filters is then assessed with extensive experimental simulations using real audio data recorded in a reverberant environment. This paper demonstrates that the ability to achieve a reduced tracking error relies on both the chosen model as well as the specific optimization of its parameters.

1. INTRODUCTION

As a Bayesian filtering approach, the development of a particle filter (PF) for the acoustic speaker tracking (AST) problem requires the definition of two important concepts [1, 2]:

1) the measurement or observation PDF (probability density function), also known as the likelihood function; and
2) the transition PDF, based on a model describing the specific dynamics of the considered target (speaker).

In the currently available AST literature [3–5], a significant research effort focusses on the development of improved measurement densities, and very little attention is given to the type of motion model implemented in the algorithm. Originally defined in [3], the so-called "Langevin" dynamics constitutes the generic model of choice routinely implemented in these AST publications. However, no extensive justification was given as to why this specific model was originally selected. The present research provides some insight into the influence of the assumed dynamics (and the optimization of its parameters) on the overall tracking performance for AST. It must be emphasized that this work does not aim to find an optimal dynamics formulation for the problem at hand; this goal typically relies on a specialized parameter optimization process, which is currently the object of ongoing research.

This work was supported by National ICT Australia (NICTA). NICTA is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

There exist many different motion models suitable for an implementation in relation to the AST problem [6]; an exhaustive list can be found, for instance, in [7]. The present work considers a small subset of models which appear promising for such an implementation. Of particular interest is the specific behavior of the resulting PF algorithm during the silence gaps existing between separate utterances in a typical speech signal. A PF method was recently proposed in [4] which takes into account the measurements obtained with a voice activity detector (VAD). This algorithm, denoted PF-VAD, was developed on the basis of the usual Langevin dynamics, and as demonstrated in [8], this leads to the tracker effectively "freezing" its estimate and spreading the particles uniformly in all directions as soon as the speaker becomes silent. In essence, this corresponds to the assumption that a person is equally likely to move in any direction at any point in time; with a uniform spreading of the particles, the algorithm hence "tracks" *any* potential speaker motion while no observations are available.

This approach is however not fully relevant for practical scenarios: typically, speakers moving in a given environment rarely exhibit abrupt changes in direction and velocity. In other words, it is more realistic to assume that during (short) silence gaps, the speaker's motion remains similar to that displayed shortly before the speech interruption. Integrating this specific property of motion continuity within the tracking algorithm would hence lead to a superior tracking performance and an increased robustness against disturbances (noise, competing speakers, etc.). As shown in this paper, this can be achieved with a careful choice of dynamics model and an appropriate tuning of the model parameters.

In this work, the considered dynamics models are implemented in conjunction with the PF-VAD algorithm of [4]. The next section hence presents a brief review of this algorithm. Section 3 then describes the various models under consideration, and the performance results obtained from experimental simulations of the resulting PF algorithms are finally presented in Section 4.

2. PF-VAD ALGORITHM REVIEW

2.1. Basic Principle

Assume that an array of M acoustic sensors is set up in a given environment. Let $\boldsymbol{X}_k$ represent the state variable for time index k, corresponding to the position $\boldsymbol{\ell}_k = [x_k \;\; y_k]^{\mathrm{T}}$ and velocity $\dot{\boldsymbol{\ell}}_k = [\dot{x}_k \;\; \dot{y}_k]^{\mathrm{T}}$ of the speaker in the state space: $\boldsymbol{X}_k = [x_k \;\; y_k \;\; \dot{x}_k \;\; \dot{y}_k]^{\mathrm{T}}$. Also, let $\boldsymbol{Y}_k$ denote the measurement variable, which corresponds to the localization information obtained from the output $\mathcal{P}(\boldsymbol{\ell})$ of a delay-and-sum beamformer, steered to the location $\boldsymbol{\ell} = [x \;\; y]^{\mathrm{T}}$, and computed for each frame k of signal data from the sensors.

A Bayesian filtering approach to the tracking problem attempts to determine, for each time step k, the so-called posterior density $p(\boldsymbol{X}_k|\boldsymbol{Y}_{1:k})$, where $\boldsymbol{Y}_{1:k} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k\}$ represents the concate-

nation of all measurements. From a statistical viewpoint, the posterior PDF contains all the information available regarding the current condition of the state variable $\boldsymbol{X}_k$, and an estimate $\widehat{\boldsymbol{X}}_k$ of the state then follows, e.g., as the first-order moment of this density.

Particle filtering is an approximation technique that solves the Bayesian filtering problem by representing the posterior PDF as a set of N samples $\boldsymbol{X}_k^{(n)}$ of the state space (particles) with associated weights $w_k^{(n)}$, $n \in \{1, \ldots, N\}$ [1]. The PF-VAD algorithm is based on the so-called bootstrap PF [2], whose generic principle can be described as follows. Assume that the set of particles and weights $\{(\boldsymbol{X}_{k-1}^{(n)}, w_{k-1}^{(n)})\}_{n=1}^N$ is a discrete representation of the posterior $p(\boldsymbol{X}_{k-1}|\boldsymbol{Y}_{1:k-1})$. Given the observation $\boldsymbol{Y}_k$ obtained at time k, update each particle $n \in \{1, \ldots, N\}$ as follows:

1. *Prediction:* propagate the particles through the transition equation $\widetilde{\boldsymbol{X}}_k^{(n)} = g(\boldsymbol{X}_{k-1}^{(n)}, \boldsymbol{u}_k)$, where $\boldsymbol{u}_k$ is a noise variable.
2. *Update:* assign a likelihood weight to each new particle according to $\tilde{w}_k^{(n)} = w_{k-1}^{(n)} \cdot p(\boldsymbol{Y}_k|\widetilde{\boldsymbol{X}}_k^{(n)})$, then normalize the weights:
$$w_k^{(n)} = \tilde{w}_k^{(n)} \Big/ \Big(\sum\nolimits_{i=1}^{N} \tilde{w}_k^{(i)} \Big).$$
3. *Resampling:* draw N new samples $\boldsymbol{X}_k^{(n)}$ from the existing set of particles $\{\widetilde{\boldsymbol{X}}_k^{(i)}\}_{i=1}^N$ according to their weights $w_k^{(i)}$, then reset the weights to uniform values: $w_k^{(n)} = 1/N$, $\forall n$.

As a result, the new set $\{(\boldsymbol{X}_k^{(n)}, w_k^{(n)})\}_{n=1}^N$ is approximately distributed as the posterior $p(\boldsymbol{X}_k|\boldsymbol{Y}_{1:k})$. An estimate of the speaker's position is then obtained as $\hat{\boldsymbol{\ell}}_k = \sum_{n=1}^N w_k^{(n)} \boldsymbol{\ell}_k^{(n)}$, where $\boldsymbol{\ell}_k^{(n)}$ corresponds to the location of the n-th particle: $\boldsymbol{X}_k^{(n)} \triangleq [\boldsymbol{\ell}_k^{(n)} \;\; \dot{\boldsymbol{\ell}}_k^{(n)}]^{\mathrm{T}}$.

In the PF-VAD method, the likelihood $p(\boldsymbol{Y}_k|\boldsymbol{X}_k)$ is defined as a mixture density involving the beamformer output $\mathcal{P}(\cdot)$, as well as the current VAD state. The algorithm also requires a model representing the state dynamics in terms of the transition equation $\boldsymbol{X}_k = g(\boldsymbol{X}_{k-1}, \boldsymbol{u}_k)$. The specific definition of this function, and the analysis of its influence on the tracking results, constitutes the object of focus in the present work. Readers are referred to [4] for more information regarding the PF-VAD implementation.

2.2. Performance Assessment Parameters

The PF estimation error for the k-th frame is $\varepsilon_k = \big\|\boldsymbol{\ell}_{\mathrm{S},k} - \hat{\boldsymbol{\ell}}_k\big\|$, where $\boldsymbol{\ell}_{\mathrm{S},k}$ is the ground-truth source position. In order to assess the global tracking performance of the algorithm over K frames of audio data, the average error is computed as the RMSE (root-mean-square error) parameter

$$\bar{\varepsilon} = \sqrt{\frac{1}{K} \sum\nolimits_{k=1}^{K} \varepsilon_k^2}.$$

Due to the partially random nature of PF implementations, statistical averaging over a number D of algorithm runs is used in the results presentation. A parameter of particular interest to AST is the percentage of these runs for which the tracking algorithm completely loses track of the target during the simulation, typically due to significant silence gaps in the speech or an incorrect setting of the model parameters. For each simulation run $d \in \{1, \ldots, D\}$, a track loss parameter is thus defined as

$$\zeta_d = \begin{cases} 1 & \text{if } \big(\sum_{k=K-k^*}^{K} \varepsilon_{k,d}\big)/(k^*-1) > \delta\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

where $k^* = \lceil 0.5/T \rceil$ and T represents the time update period (from time k to $k+1$). The parameter ζ_d effectively determines whether the average estimation error over the last 0.5s of audio data is smaller than some threshold $\delta = 0.1$m, i.e., whether the algorithm is still correctly tracking the target at the end of the simulation run. The global track loss percentage (TLP) $\bar{\zeta}$ (in %) for a given audio sample is then defined as $\bar{\zeta} = (100/D) \cdot \sum_{d=1}^{D} \zeta_d$.

3. DYNAMICS MODELING

3.1. Model Types

As mentioned previously, several dynamics models represent potential candidates for an implementation in the frame of AST [6, 7]. In the following, two main model types are investigated:

1) Coordinate-uncoupled (CU) dynamics represent the target's velocity in a Cartesian coordinate setting, with the state vector typically defined as $\boldsymbol{X}_k \triangleq [x_k \;\; y_k \;\; \dot{x}_k \;\; \dot{y}_k]^{\mathrm{T}}$. By definition, this type of manoeuvre model assumes a negligible coupling between coordinates, and only one generic variable (chosen here to be x) needs to be considered in the derivations.
2) Curvilinear (CL) models represent the target's velocity vector $\boldsymbol{v}_k$ using a polar coordinate system, i.e., in terms of its magnitude $v_k = \|\boldsymbol{v}_k\|$ and its orientation angle φ_k with respect to the x-axis. The dynamics for φ_k can be considered through the target's normal acceleration a_k using the kinematics equation of a uniform circular motion: $a_k = v_k \cdot \dot{\varphi}_k$. With this approach, the state vector follows as $\boldsymbol{X}_k \triangleq [x_k \;\; y_k \;\; v_k \;\; a_k]^{\mathrm{T}}$, with the target's heading angle then resulting indirectly as $\varphi_k = \varphi_{k-1} + Ta_k/v_k$, and the target's position as $x_k = x_{k-1} + Tv_k \cos(\varphi_k)$ and $y_k = y_{k-1} + Tv_k \sin(\varphi_k)$.

Additionally, for any given state variable ξ (e.g., target's velocity or acceleration), the transition equation can be defined either as:

1) a random-walk (RW) process with variance σ_ξ^2, that is, $\xi_k = \xi_{k-1} + \sigma_\xi \cdot u_k$, with $u_k \sim \mathcal{N}(0,1)$; or
2) a time-correlated (TC) process with variance σ_ξ^2 and correlation time constant $1/\beta_\xi$, whose discrete-time formulation is
$$\xi_k = \mathrm{e}^{-\beta_\xi T} \cdot \xi_{k-1} + \sigma_\xi \sqrt{1 - \mathrm{e}^{-2\beta_\xi T}} \cdot u_k\,, \quad u_k \sim \mathcal{N}(0,1)\,.$$

Finally, different representations also result depending on the considered model order, i.e., whether the model makes use of the target's velocity or acceleration in the state vector.

The different combinations of the above choices (model type, model order, and time-correlation vs. random-walk) lead to a prohibitively large number of dynamics formulations to assess. This work only considers a handful of models, whose implementation in relation to the AST problem definition is deemed promising or of some interest. The following subsection enumerates these different models, with the generic noise variables $u_k, u'_k \sim \mathcal{N}(0,1)$.

3.2. Considered Models

1) **CU-RWV**: CU model with RW velocity,
$$\begin{bmatrix} x_k \\ \dot{x}_k \end{bmatrix} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_{k-1} \\ \dot{x}_{k-1} \end{bmatrix} + \begin{bmatrix} T/2 \\ 1 \end{bmatrix} \cdot \sigma_v \, u_k\,.$$

2) **CU-TCA**: CU model with time-correlated acceleration [7],
$$\begin{bmatrix} x_k \\ \dot{x}_k \\ \ddot{x}_k \end{bmatrix} = \begin{bmatrix} 1 & T & \alpha_1 \\ 0 & 1 & \alpha_2 \\ 0 & 0 & \mathrm{e}^{-\beta_a T} \end{bmatrix} \cdot \begin{bmatrix} x_{k-1} \\ \dot{x}_{k-1} \\ \ddot{x}_{k-1} \end{bmatrix} + \begin{bmatrix} T^2/2 \\ T \\ 1 \end{bmatrix} \cdot \sigma_a \sqrt{1 - \mathrm{e}^{-2\beta_a T}} \cdot u_k,$$
with $\alpha_1 = (\beta_a T - 1 - \mathrm{e}^{-\beta_a T})/\beta_a^2$, and $\alpha_2 = (1 - \mathrm{e}^{-\beta_a T})/\beta_a$.

3) **CU-LAN**: CU model with Langevin dynamics [3],

$$\begin{bmatrix} x_k \\ \dot{x}_k \end{bmatrix} = \begin{bmatrix} 1 & T\,\mathrm{e}^{-\beta_v T} \\ 0 & \mathrm{e}^{-\beta_v T} \end{bmatrix} \cdot \begin{bmatrix} x_{k-1} \\ \dot{x}_{k-1} \end{bmatrix} + \begin{bmatrix} T \\ 1 \end{bmatrix} \cdot \sigma_v \sqrt{1-\mathrm{e}^{-2\beta_v T}} \cdot u_k \,.$$

4) **CL-RWV-RA**: CL model, RW velocity, random acceleration,

$$\begin{bmatrix} v_k \\ a_k \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} v_{k-1} \\ a_{k-1} \end{bmatrix} + \begin{bmatrix} \sigma_v\, u_k \\ \sigma_a\, u'_k \end{bmatrix}.$$

5) **CL-RWV-RWA**: CL model with RW velocity and acceleration,

$$\begin{bmatrix} v_k \\ a_k \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} v_{k-1} \\ a_{k-1} \end{bmatrix} + \begin{bmatrix} \sigma_v\, u_k \\ \sigma_a\, u'_k \end{bmatrix}.$$

3.3. Parameter Optimization

Each of the above models contains at least one free parameter that requires to be optimized (variance σ and/or rate constant β). In the frame of AST, these parameters should ideally be optimized by considering a large number of speakers, rooms, array setups and speaker trajectories, in various environments with different SNR and reverberation levels. The acquisition of such a large amount of real-world data constitutes a significant practical challenge. A simplified approach is used in the present paper, whose main goal is to provide some insight into how the use of a specific dynamics model can influence the tracking accuracy of a PF algorithm. For this purpose, the model parameters are coarsely optimized using a series of real audio samples recorded in an environment with reverberation time $T_{60} \approx 0.27$s and approximate background noise SNR $\approx$ 20dB (white noise). The results obtained from this parameter tuning process are summarized in Table 1. The model denoted CU-LAN-ORIG corresponds to the Langevin formulation given in Section 3.2.3) with its parameter values set as defined in most of the current AST literature (i.e., non-optimized Langevin dynamics); this specific definition is included here as a comparative benchmark for other models.

4. EXPERIMENTAL SIMULATIONS

4.1. Practical Recording Setup

An array of $M = 8$ microphones is placed at a constant height in a room with dimensions 3.5m $\times$ 3.1m $\times$ 2.2m, with one sensor pair centered on each wall. The distance between the sensors in each pair is 0.8m. The walls are partially covered with acoustic foam, leading to a practical reverberation time $T_{60} \approx 0.27$s (frequency-averaged up to 24kHz). Audio samples of background noise are recorded separately from the speech signals, and used in the simulation phase to generate specific values of SNR. The availability of clean speech signals further allows the precise measurement of the speaker trajectory directly from the audio data in each scenario, using the method presented in [8]: the microphone signals are processed with a high-precision beamformer delivering accurate localization estimates, with outliers easily discarded based on the approximate knowledge of the source trajectory.

4.2. Tracking Example

Each of the considered dynamics models is implemented within the PF-VAD framework with $N = 75$ particles. These different algorithms are then simulated using an example of speaker trajectory, with an approximate SNR level of 20dB (white noise, long-term average). Fig. 1 presents the tracking results obtained with each of the considered models, with Fig. 2 showing an example of audio signal recorded with one of the array sensors for this simulation. The plots in Fig. 1 show the PF's x-location estimate averaged over 50 simulation runs for each model (results for the y dimension are similar).

Model	σ_v	β_v	σ_a	β_a
CU-RWV	0.05	-	-	-
CU-TCA	-	-	2	30
CU-LAN	0.7	0.2	-	-
CU-LAN-ORIG	0.8	10	-	-
CL-RWV-RA	0.07	-	1.5	-
CL-RWV-RWA	0.3	-	2	-

Table 1: Selected values for each model parameter. β values are given in Hz, σ_v values in m/s, and σ_a values in m/s^2.

These results clearly illustrate that some of the considered models are able to implement the desired property of "continued motion", i.e., maintain a proper heading and velocity during silence gaps. Based on the assumption that the speaker is unlikely to exhibit abrupt direction and velocity changes, this consequently results in a superior tracking performance. On the other hand, the non-optimized model CU-LAN-ORIG essentially stops tracking whenever no measurements are available, as previously observed in [4, 8]. The model CL-RWV-RWA also presents the same weakness, perhaps due to a poor parameter optimization or an inappropriate representation of the target dynamics. Interestingly, the optimized version of the Langevin model CU-LAN achieves very good results, which leads to the important observation that an improved tracking behavior depends not only on the type of dynamics model, but also strongly relies on an optimal tuning of its parameters.

Finally, an aspect of particular importance to consider in Fig. 1 is the resulting standard deviation of the particle set during silence periods. The process of spreading the particles when no observations are available is what allows the algorithm to successfully resume tracking once the speaker becomes active again, in the eventuality that the target has slightly changed its course during the silence gap. It can be seen from Fig. 1 that most models achieve a satisfactory performance from this point of view, which suggests that the various model parameters in Table 1 are set to meaningful values; setting these values too tightly might result in the PF algorithm not being able to spread the particles fast enough.

4.3. Average Performance vs. SNR

While most models have shown to achieve successful tracking results at 20dB SNR, a risk exists that the chosen parameter values restrict their performance when the SNR decreases. Fig. 3 presents the performance results for each model as a function of the SNR level. For each SNR value, these results were averaged over a total of 240 simulation runs, corresponding to 60 runs carried out for each of four different audio recordings representing different speaker trajectories and speech signals.

Fig. 3 shows a similar trend for most of the considered models, with a breakdown of the tracking performance as the SNR decreases below approximately 5 or 10dB. This suggests that none of these methods suffers from a drastically erroneous setting of its parameters (except perhaps for CL-RWV-RA). The non-optimized Langevin model CU-LAN-ORIG has a decreased overall tracking performance (larger RMSE results) due to an increased tracking error during periods of speech inactivity. Here too, a distinct improvement in the tracking accuracy is observed for CU-LAN when compared to the non-optimized version CU-LAN-ORIG.

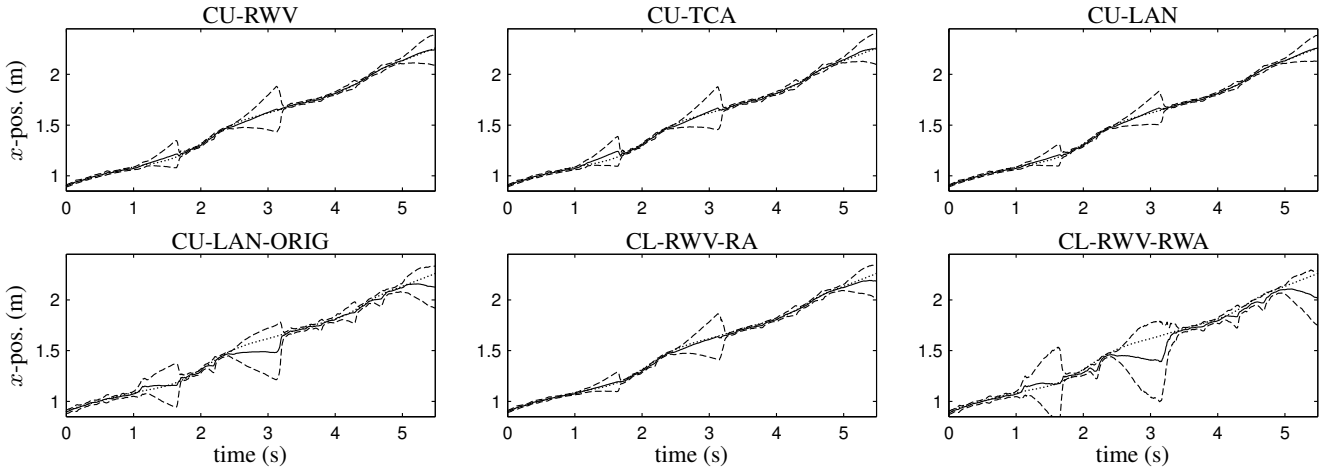

Figure 1: Example of tracking results with each considered model. Dotted lines correspond to the true source location, solid lines show the PF estimates, and dashed lines represent plus/minus one standard deviation of the particle set (average particle spread in the x dimension).

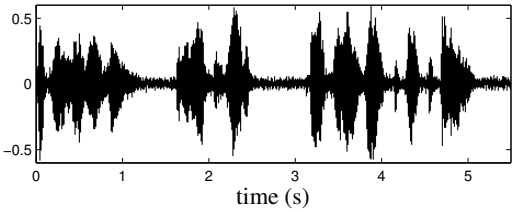

Figure 2: Example of sensor signal used for the results of Fig. 1.

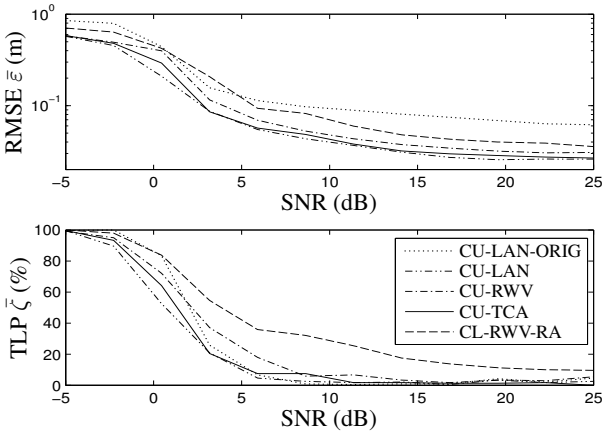

Figure 3: Tracking performance versus SNR level.

5. CONCLUSION AND FUTURE RESEARCH

The work presented in this paper provides a preliminary study on the use of various dynamics models in the frame of a Bayesian approach to acoustic source tracking. It was shown that the dynamics model implemented in the tracking algorithm can play a potentially significant part in achieving superior tracking results, with both the CU as well as the CL model types having the potential to provide a more accurate tracking of the speaker during periods of speech inactivity. Rather than simply freezing its estimate, the tracking algorithm can be made to "blindly" track the speaker when no measurements are available. This ultimately leads to a decreased chance of track loss when the speech resumes, and consequently, an improved robustness against noise, reverberation, and competing speakers. Finally, it was also demonstrated that the successful implementation of a tracking algorithm does not solely rely on choosing a specific model type; the process of optimizing the model parameters also plays a crucial part in the accuracy of the resulting algorithm. The critical issue is hence to obtain a set of parameter values achieving a robust tracking performance, while being able to deal successfully with a wide range of target motions; a tradeoff might have to be found between these two factors in practice. An ongoing research effort is currently focussing on a rigorous and efficient optimization method for dynamics models in the design of an algorithm for acoustic source tracking.

6. REFERENCES

[1] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Sig. Proc.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[2] N. Gordon, D. Salmond, and A. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEE Proc. Radar and Sig. Proc.*, vol. 140, no. 2, pp. 107–113, Apr. 1993.

[3] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. IEEE ICASSP*, vol. 5, Salt Lake City, UT, USA, may 2001, pp. 3021–3024.

[4] E. Lehmann and A. Johansson, "Particle filter with integrated voice activity detection for acoustic source tracking," *EURASIP J. Adv. Sig. Proc.*, vol. 2007, Article ID 50870.

[5] D. Ward and R. Williamson, "Particle filter beamforming for acoustic source localization in a reverberant environment," in *Proc. IEEE ICASSP*, vol. 2, Orlando, FL, USA, May 2002, pp. 1777–1780.

[6] S. Blackman and R. Popoli, *Design and analysis of modern tracking systems*. Boston: Artech House, 1999.

[7] X. Li and V. Jilkov, "Survey of maneuvering target tracking. Part I: dynamic models," *IEEE Trans. Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1333–1364, Oct. 2003.

[8] E. Lehmann and A. Johansson, "Experimental performance assessment of a particle filter with voice activity data fusion for acoustic speaker tracking," in *Proc. IEEE Nordic Sig. Proc. Symposium*, Reykjavik, Iceland, Jun. 2006, pp. 126-129.