APPLICATION OF PARTICLE FILTERING TO AN INTERAURAL TIME DIFFERENCE BASED HEAD TRACKER FOR CROSSTALK CANCELLATION

Yesenia Lacouture-Parodi and Emanuël A. P. Habets

International Audio Laboratories Erlangen[†]
91058 Erlangen, Germany
{yesenia.lacouture,emanuel.habets}@audiolabs-erlangen.de

ABSTRACT

Crosstalk cancellation systems (CCSs) suffer from a rather narrow sweet spot and small head rotations introduce significant interaural time difference (ITD) errors at the ears of the listener that destroy the 3D experience. In a previous study, we proposed a CCS with a microphone-based head tracker that uses two microphones placed closed to the ears of the listener. The head orientation was estimated by minimizing the difference between the ITD of desired binaural signals and the ITD of the microphones signals. We present here an extension of the previously proposed system in which the tracking of the ITD error is performed using a particle filtering (PF) approach that allows to take into account the dynamics of the human head. The head dynamics are modelled as a Ornstein-Uhlenbeck process, which shows to be a closer approximation of the natural movements of the head. Experimental results shows that with the proposed PF approach, head rotations can be accurately tracked in the presence of noise and when multiple virtual sound sources are reproduced simultaneously.

Index Terms— Interaural time differences, particle filter, Ornstein-Uhlenbeck process, crosstalk cancellation, head tracker

1. INTRODUCTION

Crosstalk cancellation systems (CCSs) seek to compensate for the acoustic paths between the loudspeakers and the ears so that binaural signals can be reproduced accurately at the ears when rendering them through loudspeakers. When designed properly, CCSs are capable of an accurate 3D sound reproduction if the listener is within the sweet-spot. However, CCSs suffer in general from a very narrow sweet-spot and head rotations are likely to destroy the 3D sound experience [1, 2]. To ensure the desired experience, a head tracker is usually required. There are several different technologies to track the listener's head, such as magnetic trackers and video cameras (see [3] for a comprehensive summary). Many of the known technologies are however not accurate enough to properly update the crosstalk cancellation filters (CCFs).

Head-tracking in binaural reproduction systems can be seen as an acoustic source localization (ASL) problem, where the aim is to estimate the position or orientation angle of the listener relative to the loudspeakers. In this case the source positions are fixed and the microphone positions are varying. Surprisingly, there are very few studies that have attempted to use microphones to estimate the location of the listener. One of the few examples is the head tracker proposed in [4] where two microphones are placed at the ears of the listener and a set of anchor sources are used to estimate the location of the listener. In their approach, at least three anchor sources are needed to find an unique position. In [5, 6], the location of the listeners problem is solved by providing the listeners with binaural head-sets. The head orientation is then estimated using the time delay of arrival (TDOA) estimates between the listener and the speaker. The performance of this approach improves as the number of listeners/speakers increases and the proposed method shows root mean square errors (RMSE) in the range of 10°. For CCSs this error is however unacceptable, given that the introduced interaural time difference (ITD) error will be above the audibility threshold [7].

We proposed in [8] a head tracker that uses two microphones placed close to the ears of the listener. Using ITD error between the desired binaural signals and the microphones signals, the orientation angle with respect to the loudspeakers is estimated. Results showed that, for single virtual sound sources, the head orientation can be accurately tracked using a simple sign algorithm. However, in the presence of multiple virtual sound sources and noise, a more robust estimation algorithm is required.

This paper is an extension of the aforementioned CCS in which a particle filtering (PF) approach is used to estimate the orientation of the head in a robust manner. PF is a Bayesian filtering approach, which estimates the current state of a dynamic process based on current and previous measurements. Given that it is based on sequential Monte Carlo (MC) approximations, it is not subject to any assumption of linearity or Gaussianity of the model [9]. Thus, PF has been successfully used for ASL where the location of the source is time varying and the dynamics are usually non-linear [10–13].

In ASL applications microphone arrays are usually employed and a set of different measurements are available at each time frame. In the proposed CCS, we only have access to two microphones and thus only one measurement is available per time frame. In order to produce different measurements at each time frame, we propose to estimate the ITD error in subbands. A byproduct of this approach is that it improves the ability of the tracking system to accurately estimate the head orientation even when multiple virtual sound sources are reproduced. The measurements are then mapped into the ITD error model proposed in [8] in order to account for the current state of the CCFs and to derivate a localization function suitable for PF tracking. In [6], the dynamics of the head rotations are modelled using a simple Brownian motion model. We propose here to model the head rotations as an Ornstein-Uhlenbeck (OU) process, which better resembles the typical dynamics of head rotations [14, 15].

2. MICROPHONE-BASED HEAD TRACKER

Fig. 1 shows a basic diagram of the proposed CCS and the geometry that will be used throughout the paper [8] . The angle ϕ_s corresponds to the span angle between loudspeakers, α is the orientation angle of the head with respect to the middle point between loudspeakers and θ_e is the angle of the ears with respect to the median plane. The

[†] A joint institution of the University of Erlangen-Nuremberg and Fraunhofer IIS

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

functions H_{ji} are the transfer functions between the *i*th loudspeakers and the *j*th ear. The CCS makes use of the ITD error between the input binaural signals d_i and the microphone signals v_i to estimate the orientation of the head. The estimated orientation angle $\hat{\alpha}$ is then used to calculate the corresponding acoustic transfer functions (ATFs) $\hat{H}_{ji}$. The inputs to the CCS are the desired binaural signals d_i and the estimated ATFs $\hat{H}_{ji}$, from which new CCFs are calculated. In this work, we make use of the spherical head model (SHM) to calculate the ATFs [16].

3. PARTICLE FILTER

The basic idea of PF is to recursively compute a posterior probability density function (PDF) of the current state [11] . Let us first define the state vector at time m for our head-tracking problem:

$$\mathbf{x}_m = [\mathbf{q}, \boldsymbol{\omega}] \,, \tag{1}$$

where $\mathbf{q}$ and $\boldsymbol{\omega}$ represent the rotation and angular velocity vectors respectively. In the state vector, the orientation is encoded by the vector part of the unit quaternion, i.e. $\mathbf{q} = \vec{\mathbf{a}} \sin({}^{\alpha}\!/_{2})$, where α is the orientation angle and $\vec{\mathbf{a}}$ is the unit vector along the axis of rotation [15]. The evolution state can be defined as a first-order Markov process

$$\mathbf{x}_m = T\left(\mathbf{x}_{m-1}, \boldsymbol{\xi}_{x_m}\right) \,, \tag{2}$$

where $T(\cdot)$ is a (possibly non-linear) function of the state $\mathbf{x}_{m-1}$ and $\boldsymbol{\xi}_{x_m}$ is an i.i.d. noise process. We seek to recursively estimate the current state $\mathbf{x}_m$ from the measurements

$$y_m = S\left(\mathbf{x}_m, \boldsymbol{\xi}_{y_m}\right) \,, \tag{3}$$

where $S(\cdot)$ is an unknown (possibly non-linear) function and $\boldsymbol{\xi}_{y_m}$ is an i.i.d. measurement noise process. The Bayesian approach to the tracking problem is to recursively estimate the posterior PDF $p(\mathbf{x}_m|\mathbf{y}_{1:m})$, where $\mathbf{y}_{1:m} = [y_1 \dots y_m]$ is the concatenation of all measurements up to time m. The estimate $\hat{\mathbf{x}}_m$ can then be computed as the mean or mode of this PDF. This density is usually unavailable, but can be estimated using a "prediction and update" scheme, using the posterior probability $p(\mathbf{x}_{m-1}|\mathbf{y}_{1:m-1})$ at time $m-1$ and the transition PDF $p(\mathbf{x}_m|\mathbf{x}_{m-1})$ [10],

$$p(\mathbf{x}_m|\mathbf{y}_{1:m-1}) = \int p(\mathbf{x}_m|\mathbf{x}_{m-1}) p(\mathbf{x}_{m-1}|\mathbf{y}_{1:m-1}) d\mathbf{x}_{m-1}. \tag{4}$$

The prior PDF can be updated using Bayes' rule to obtain the posterior PDF of the current state:

$$p(\mathbf{x}_m|\mathbf{y}_{1:m}) = \frac{p(\mathbf{y}_m|\mathbf{x}_m) p(\mathbf{x}_m|\mathbf{y}_{1:m-1})}{p(\mathbf{y}_m|\mathbf{y}_{1:m-1})} , \tag{5}$$

where $p(\mathbf{y}_m|\mathbf{y}_{1:m-1}) = \int p(\mathbf{y}_m|\mathbf{x}_m) p(\mathbf{x}_m|\mathbf{y}_{1:m-1}) d\mathbf{x}_m$. The likelihood function $p(\mathbf{y}_m|\mathbf{x}_m)$ is defined by (3) and measures the probability of receiving the data $\mathbf{y}_m$ given the state $\mathbf{x}_m$.

Despite the fact that no closed form solution exists to these equations they can be approximated through MC simulations of a set of particles. The PDF $p(\mathbf{x}_m|\mathbf{y}_{1:m-1})$ can be approximated with a discrete distribution using a set of N random samples of the state space $\{\mathbf{x}_m^{(p)}\}_{p=1}^N$ with associated likelihood weights $\{w_m^{(p)}\}_{p=1}^N$. $\mathbf{x}_m^{(p)}$ represents a particle and is defined as the sampled representation of the source state [10, 17].

3.1. Head dynamics

In ASL problems, the source dynamics are typically modeled using stochastic approaches to describe the dynamic equations. Brownian motion models are commonly used for this purpose, where the dynamics are modelled as a Wiener process (WP) [6, 13]. Another commonly used approach is the Langenvin model [17]. However, such models do not necessarily resemble the dynamics of head rotations. Here we make use of an stochastic model that describes the dynamics of the head as an OU process [14, 15]. In an OU process there is an inherent tendency of the particles to move back towards a central location, which describes better the nature of head movements.

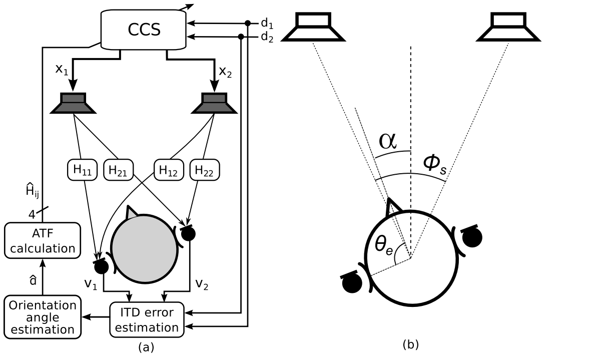

Fig. 1. Simplified diagram of the proposed CCS with a microphone-based head-tracker and geometry used throughout the paper (see [8] for a detailed description).

Modelling the head movement as an OU process, the set of stochastic differential equations is given by [15]

$$\dot{\mathbf{q}} = \frac{q\omega}{2}, \quad \dot{\boldsymbol{\omega}} = -\mathrm{B}_\omega \boldsymbol{\omega} + \sqrt{2\mathrm{B}_\omega} \mathrm{G}_\omega \mathrm{w}_\omega , \tag{6}$$

where w_ω is a vector whose elements are independent unit-variance zero-mean Gaussian white noise processes and G_ω is a covariance matrix of the accelerations along each axis. The matrix $\mathrm{B}_\omega = \operatorname{diag}\left(\left[\beta_{\omega_x}, \beta_{\omega_y}, \beta_{\omega_z}\right]\right)$ contains the correlation coefficients on each axis. Given that in our CCS we assume rotations in only one axis, the OU discrete update equations for this case are [15]

$$\begin{aligned} q_m &= q_{m-1} \exp({}^{\Theta_{m-1}}\!/_{2}) \exp({}^{\xi_{q_m-1}}\!/_{2}), & (7)\\ \omega_m &= \omega_{m-1} \exp(-\beta_\omega \Delta T) + \xi_{\omega_{m-1}}, & (8) \end{aligned}$$

where $\Theta_{m-1} = \omega_m \beta_\omega^{-1} \left(1 - \exp\left(-\beta_\omega \Delta T\right)\right)$. $\xi_{q_{m-1}}$ and $\xi_{\omega_{m-1}}$ are the accumulated effects of random accelerations of the state vector:

$$\begin{aligned} \xi_{q_m} &= \int_0^{\Delta T} \beta_\omega^{-1} \left(1 - \exp\left(-\beta_\omega u\right)\right) \sqrt{2\beta_\omega} \sigma \, \mathrm{w}_\omega(u) \, \mathrm{d}u, & (9)\\ \xi_{\omega_m} &= \int_0^{\Delta T} \exp\left(-\beta_\omega u\right) \sqrt{2\beta_\omega} \sigma \, \mathrm{w}_\omega(u) \, \mathrm{d}u, & (10) \end{aligned}$$

where σ is the variance of the acceleration. From the noise characteristics, the transition PDF for this model can be defined as

$$p(\mathbf{x}_m|\mathbf{x}_{m-1}) = F\left(\mathbf{x}_m; \mathbf{Q}_m, \boldsymbol{\Sigma}\right) , \tag{11}$$

where $F(\cdot)$ is the normal cumulative distribution function of a Gaussian variable evaluated at $\mathbf{x}_m$, with mean and covariance matrices:

$$\mathbf{Q}_m = \operatorname{diag}\left(\left[q_{m-1} \exp\left({}^{\Theta_{m-1}}\!/_{2}\right) , \, \omega_{m-1} \exp\left(-\beta_\omega \Delta T\right)\right]\right) , \tag{12}$$

$$\boldsymbol{\Sigma} = \operatorname{diag}\left(\left[\beta_\omega^{-1} \left(1 - e^{-\beta_\omega \Delta T}\right) \sqrt{2\beta_\omega} \sigma \, , \, e^{-\beta_\omega \Delta T} \sqrt{2\beta_\omega} \sigma\right]\right) . \tag{13}$$

3.2. Localization function

The localization function transforms the measurements into a function that usually exhibits a peak at the estimated location of the sound source. Traditionally, these measurements are obtained directly from a beamformer output [18] or indirectly from TDOAs estimated by different receivers [19]. In both cases, ASL methods commonly rely on more than one measurement per time frame. In our head-tracker application, the head orientation is estimated based on ITD errors instead of TDOAs. That is, the measurement is not only a function of the current head-orientation, but also of the head orientation used to calculate the CCFs. Furthermore, we only have access to two microphone signals. Thus, we need to redefine our measurements in a way that different measurements become available at each time frame and that a proper localization function can be derived.

To obtain different ITD error estimates at each time frame, we propose to estimate the ITD error in subbands. Let the discrete functions $r_{\text{in}}^{(m)}(n)$ and $r_{\text{out}}^{(m)}(n)$ be, respectively, the ITD estimation functions at time frame m of the input binaural signals and of the microphone signals, with their respective short-time frequency responses $\mathcal{R}_{\text{in}}(m,k)$ and $\mathcal{R}_{\text{out}}(m,k)$. The ITD is estimated as the time at which $r_{\text{in}}^{(m)}(n)$ and $r_{\text{out}}^{(m)}(n)$ have their maximum value. These functions could be obtained for example from the generalize cross-correlation (GCC) method or ATFs ratios [20,21]. Now, let us divide the spectrum into L partitions, where $b = 1 \ldots L$ is the partition index. We define the subband ITD error as

$$\text{ITD}_{\text{error}}(m,b) = \arg\max_{n} \left\{ r_{\text{in}}^{(m,b)}(n) \right\} - \arg\max_{n} \left\{ r_{\text{out}}^{(m,b)}(n) \right\}, \tag{14}$$

where $r_{\text{in}}^{(m,b)}(n)$ and $r_{\text{out}}^{(m,b)}(n)$ are, respectively, the time domain representations of $\mathcal{R}_{\text{in}}(m,k)$ and $\mathcal{R}_{\text{out}}(m,k)$ for $k \in [E_{b-1}, E_b)$, where E_b is defined by the partition bandwidth. Let us use the coherence at the microphones and the magnitude of the ITD estimation function as a measure of the closeness of the estimated ITD error to the true value as follows:

$$w(b) = \overline{C}_{\text{out}}(m,b) \frac{\overline{\mathcal{R}}_{\text{out}}(m,b)}{\|\overline{\boldsymbol{\mathcal{R}}}_{\text{out}}(m,:)\|_\infty}, \tag{15}$$

where $\overline{\mathcal{R}}_{\text{out}}(m,b) = \frac{1}{E_b - E_{b-1}} \sum_{k=E_{b-1}}^{E_b} |\mathcal{R}_{\text{out}}(m,k)|$ is the mean value of the magnitude of the ITD estimation function for partition b, $\overline{\boldsymbol{\mathcal{R}}}_{\text{out}}(m,:) = [\ \overline{\mathcal{R}}_{\text{out}}(m,1) \ \ldots \ \overline{\mathcal{R}}_{\text{out}}(m,L)\ ]$ and $\overline{C}_{\text{out}}(k,b) = \frac{1}{E_b - E_{b-1}} \sum_{k=E_{b-1}}^{E_b} C_{\text{out}}(m,k)$ is the mean value of the coherence between microphones evaluated in partition b. We need now to map our measurements into particles. In [8], we proposed a model for the ITD error as a function of the head orientation angle. According to that model, the ITD error for a given state $\mathbf{x}_m^{(p)}$ is defined as

$$\text{ITD}_{\text{error}}^{(p,m)} \approx \left(\tau_{22}^{(m-1)} - \tau_{11}^{(m-1)} \right) - \left(\tau_{22}^{(p,m)} - \tau_{11}^{(p,m)} \right), \tag{16}$$

where $\tau_{ii}^{(p,m)}$, $i \in \{1,2\}$, are the delays of the direct paths between the ith loudspeaker to the ith ear corresponding to the state vector $\mathbf{x}_m^{(p)}$ and $\tau_{ii}^{(m-1)}$, $i \in \{1,2\}$, are the delays corresponding to the ATFs used to calculated the CCFs. We model these delays as [8,22]

$$\tau_{22} - \tau_{11} = \tag{17}$$
$$\frac{r_{\text{s}}}{c} \begin{cases} \theta_e - \frac{\phi_{\text{s}}}{2} - \alpha + \Gamma\left(\frac{\phi_{\text{s}}}{2} - \alpha\right) & -\frac{\phi_{\text{s}}}{2} \le \alpha \le -\theta_{\text{lim}} \\ -2\,\alpha & -\theta_{\text{lim}} \le \alpha \le \theta_{\text{lim}} \\ -\theta_e + \frac{\phi_{\text{s}}}{2} - \alpha - \Gamma\left(\frac{\phi_{\text{s}}}{2} + \alpha\right) & \theta_{\text{lim}} \le \alpha \le \frac{\phi_{\text{s}}}{2} \end{cases},$$

where $\Gamma(\Psi) = -\theta_0 + \sqrt{\rho^2 - 1} - \sqrt{\rho^2 - 2\rho\cos(\theta_e - \Psi) + 1}$, $\rho = {}^{r_{\text{s}}}\!/_{r}$ is the normalized distance of the loudspeaker (r_{s}) with respect to the radius of the sphere (r), $\theta_0 = \cos^{-1}({}^{1}\!/_{\rho})$ and $\theta_{\text{lim}} = \theta_e - \theta_0 - \frac{\phi_{\text{s}}}{2}$. In this study we assume a constant distance between the loudspeakers and the centre of the head.

Making use of the subband ITD error defined in (14) and the model (16), we define our localization function for the state vector $\mathbf{x}_m^{(p)}$ as

$$y_m(p) = \sum_{b=1}^{L} w(b) S\left(\mathbf{x}_m^{(p)}, b\right), \tag{18}$$

where

$$S\left(\mathbf{x}_m^{(p)}, b\right) = 1 - \tanh\left[\kappa\left(|\text{ITD}_{\text{error}}^{(p,m)} - \text{ITD}_{\text{error}}(m,b)| - \sigma_\tau\right)\right]$$

is a function that maps the estimated ITD error into the model (16). The constant κ controls the width of the peaks in the localization function y_m and σ_τ is the standard deviation of (16). We have that if $\mathbf{w}(m)$ contains only one significant value at a certain partition b,

$$y_m(p) \approx \mathcal{N}\left(\text{ITD}_{\text{error}}^{(p,m)}; \text{ITD}_{\text{error}}(m,b), \sigma_\tau\right), \tag{19}$$

where $\mathcal{N}(\cdot)$ is a Gaussian distribution with mean value equal to $\text{ITD}_{\text{error}}(m,b)$ and variance σ_τ, evaluated at $\text{ITD}_{\text{error}}^{(p,m)}$.

3.3. Likelihood function

The likelihood function should reflect the fact that peaks in the localization function correspond to likely head orientations [17]. Given that our localization function is actually a mapping of one measurement into different possible states, we use in this study a pseudo-likelihood function, which uses directly the localization function as proposed in [23], namely

$$p(\mathbf{y}_m|\mathbf{x}_m^{(p)}) = \left(\max\left\{y_m(p), \epsilon_0\right\}\right)^r, \tag{20}$$

where $r > 0$ shapes the localization function to make it tractable for recursive implementation and ϵ_0 ensures that the likelihood function is always positive [23]. Another purpose of ϵ_0 is to reflect the probability that none of the peaks in the localization function correspond to the true orientation of the head.

3.4. Importance function

The aim of the importance function is to relocate particles based on the current measurements instead of propagate them from the previous state [9,12]. It can be interpreted as the PDF $q(\mathbf{x}_m|\mathbf{y}_{1:m})$ which gives a rough estimation of the state-space regions from where the particles are to be generated. Here we make use of (19) to derive an importance function that depends on the current measurements and the model as follows:

$$q(\mathbf{x}_m^{(p)}|\mathbf{y}_{1:m}) = \tag{21}$$
$$\frac{1}{L}\sum_{b=1}^{L}(1-\epsilon_q)\mathcal{N}\left(\text{ITD}_{\text{error}}^{(p,m)}; \text{ITD}_{\text{error}}(m,b), \sigma_{\text{imp}}\right) + \epsilon_q,$$

where ϵ_q serves the same purpose as ϵ_0 and σ_{imp} controls the width of the state-space region from where particles will be sampled. The unormalized importance weights are calculated as [9]

$$\tilde{w}_m^{(n)} = w_{m-1}^{(n)} \frac{p(\mathbf{y}_m|\mathbf{x}_m^{(p)})p(\mathbf{x}_m^{(p)}|\mathbf{x}_{m-1}^{(p)})}{q(\mathbf{x}_m^{(p)}|\mathbf{x}_{m-1}, \mathbf{y}_m)}. \tag{22}$$

3.5. Algorithm

Ideally, the importance function should be a function of current measurements and previous states [9]. However, the importance function defined in (21) takes into account only current measurements. Thus, to generate some particles based on previous states, we apply an approach similar to the one proposed in [12]. The PF algorithm used in this study can be summarized as follows:

At time frame m and particle index $p = 1 \ldots N$, draw a random number l from a normal distribution and generate new particles with one of the following approaches:

1. For $Pr \leq l < P_r + P_s$: sample the particle $\mathbf{x}_m^{(p)}$ from the importance function $q(\mathbf{x}_m^{(p)}|\mathbf{x}_{m-1}, \mathbf{y}_m)$ and compute the unormalized weights according to (21);
2. For $l \geq Pr + Ps$: generate a new particle $\mathbf{x}_m^{(p)}$ according to the dynamic model (7) and (8) and set the unormalize weights to $\tilde{w}_m^{(p)} = p(\mathbf{y}_m|\mathbf{x}_m^{(p)})$;
3. For $l < P_r$ (reinitialisation): sample the particle $\mathbf{x}_m^{(p)}$ from the importance function $q(\mathbf{x}_m^{(p)}|\mathbf{y}_{1:m})$ and set the unormalize weights to $\tilde{w}_m^{(p)} = p(\mathbf{y}_m|\mathbf{x}_m^{(p)})$;

where P_s is the probability of the importance function being suitable for sampling and P_r is the probability of reinitialisation. Reinitialisation is introduced to deal more efficiently with silences [12]. The weights are then normalized so that they add up to unity, i.e. $w_m^{(p)} = \tilde{w}_m^{(p)} / \sum_{i=1}^{N} w_m^{(i)}$. The head orientation is estimated as

$$\hat{\alpha}_m = 2\sin^{-1}\left(\sum_{p=1}^{N} w_m^{(p)} q_m^{(p)}\right). \tag{23}$$

Finally, a systematic resampling scheme was incorporated to account for weight degeneracy [24].

4. EXPERIMENTAL RESULTS

We simulated a CCS with a microphone-based head tracker as depicted in Fig. 1. The simulated loudspeakers were spanning 30° and were located symmetrically with respect to the center of the listener's head at a distance of $r_s = 1.2$ m. The simulated listener rotated his head from left to right at random intervals with an average speed of 4 degrees per second. The binaural input signal consisted of three virtual sources: a female singer located at 45°, a saxophone located at $-30°$ and a crowd sound located at 0°. All angles are relative to the median plane and negative angles denote sources at the right of the listener. All virtual sources were placed at 0.75 m from the listener and had a duration of 20 s, with a sampling frequency of 48 kHz.

The adaptation was done on a frame-by-frame basis with a frame size of $M = 2048$ resulting in an update interval $\Delta T = 42.7$ ms. To estimate the ITD error, we used ATFs ratios as proposed in [21]. The first form of stationarity was used and the recursive least square (RS1) was implemented. The spectrum was divided into $L = 20$ partitions uniformly spaced between 500 Hz and 8000 Hz on an ERB scale. Sensor noise at the microphones was simulated using white Gaussian noise and the signal-to-noise ratio was set to 20 dB.

The variance σ and correlation coefficient β_ω of the dynamic model were set to 0.5 rad s^{-1} and 16.4 s^{-1} respectively [15]. The probabilities P_s and P_r were varied from frame to frame and set to reflect the probability of a signal being present. For that purpose, we used the maximum value of $\mathbf{w}(m) = [w(1) \ldots w(L)]$, i.e. $P_s = 0.25\max\{\mathbf{w}(m)\}$ and $P_r = 0.05\max\{\mathbf{w}(m)\}$. To evaluate the dynamic model (7) and (8), we evaluated the tracking performance of the proposed PF when modelling the dynamics as a WP using the same values for σ and β_ω [15]. The settings of the PF algorithm are summarized in Table 1. The selected values were found to be optimal for the PF performance with both dynamic models. As a reference, the tracking performance is also compared with the sign-error algorithm presented in [8]. The input to the latter is set to the estimated ITD error of the partition where $\mathbf{w}(m)$ is maximum.

Variable	N	κ	σ_τ	σ_{imp}	ϵ_0	ϵ_q	r
Value	100	$1e6$	$5e{-6}$	$7e{-5}$	0.5	0.25	3

Table 1. Settings of the PF algorithm used in the experiments

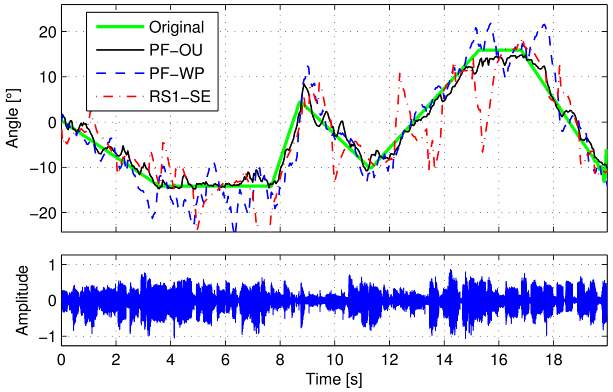

Fig. 2. Angle estimation (upper panel) and amplitude of the input signal (lower panel) as a function of time. PF-OU: proposed PF using OU dynamic model (RMSE = 1.78°); PF-WP: proposed PF modelling dynamics as a Wiener process (RMSE = 4.03°); RS1-SE: sign-error algorithm presented in [8] (RMSE = 5.35°).

Fig. 2 shows the estimated head orientation angle as a function of time for the proposed PF algorithm when modelling the dynamics according to (7) and (8) (PF-OU), as a Wiener process (PF-WP) and when estimating the head orientation using a simple sign-error algorithm (RS1-SE). The input binaural signal is plotted below the results to highlight its time variations. The RMSE values obtained for the algorithms PF-OU, PF-WP and RS1-SE were, respectively, 1.78°, 4.03° and 5.35°. These values were calculated over all frames. It is clear from the results, that the dynamic model plays an important role in the PF tracking algorithm. When modelling the dynamics as a Wiener process, some particles are propagated into unlikely head orientations, leading the PF algorithm to overestimate the movements and decreasing in that manner the performance substantially. We can also see that while the sign-error algorithm looses track due to variations in the signal and sudden changes of the head orientation, the propose PF tracker is not only able to cope with multiple virtual sources and noise, but also with silences and variations of the input signal. Comparing the RMSE values obtained with each method, we can see that in general, the PF-OU algorithm outperforms the PF-WP and the RS1-SE approaches.

5. DISCUSSION

In this paper, we presented a PF approach to track the head orientation of the listener for a CCS using two microphones placed near the ears of the listener. We derived a localization function that uses the estimated ITD error between the input binaural signals and the signals at the microphone in subbands and converts the measurements to an ITD error model based on the SHM. The head dynamics were modelled as a Ornstein-Uhlenbeck process, which showed to be a better approximation of the typical motions of the head. Based on the localization function and the statistics of the noise of the dynamic model, an importance function that depends on the current measurements and the model was derived.

The proposed algorithm was evaluated throughout simulations, where the head tracking was done while multiple virtual sound sources were reproduced simultaneously and noise was present. As opposed to a simple sign-error algorithm, the proposed PF algorithm achieves a rather accurate tracking performance even when multiple virtual sound sources are simultaneously reproduced. The selection of dynamic model showed to be a critical design criteria for the proposed head-tracker. Head dynamics are in the simulations better described as an OU process than as the commonly used WP.

6. REFERENCES

[1] Y. Lacouture-Parodi and P. Rubak, “Objective evaluation of the sweet spot size in spatial sound reproduction using elevated loudspeakers,” *J. Acoust. Soc. Am.*, vol. 128, no. 3, pp. 1045 – 1055, September 2010.

[2] Y. Lacouture-Parodi, *A systematic study of binaural reproduction systems through loudspeaker: a multiple stereo-dipole approach*, Ph.D. thesis, Aalborg University, 2010, ISBN: 978-87-92328-47-2.

[3] W. Hess, “Head-tracking techniques for virtual acoustics applications,” in *133rd AES Conv.*, San Francisco, C.A., October 2012.

[4] M. Karjalainen, M. Tikander, and A. Härmä, “Head-tracking and subject positioning using binaural headset microphones and common modulation anchor sources,” in *IEEE Int. Conf. on Acoust., Speech and Signal Proc., (ICASSP).*, 2004.

[5] H. Gamper, S. Tervo, and T. Lokki, “Head orientation tracking using binaural headset microphones,” in *131st. AES Convention*, New York, NY, USA, 2011.

[6] H. Gamper, S. Tervo, and T. Lokki, “Speaker tracking for teleconferencing via binaural headset microphone,” in *Int. Workshop on Acoust. Sig. Enhancement (IWAENC)*, Aachen, July 2012, pp. 1–4.

[7] Y. Lacouture-Parodi and P. Rubak, “Sweet spot size in virtual sound reproduction: a temporal analysis,” in *Principles and App. of Spatial Hearing*. World Scientific, Singapore, February 2011, ISBN: 978-981-4313-87-2.

[8] Y. Lacouture-Parodi and E. A. P. Habets, “Crosstalk cancellation system using a head tracker based on interaural time differences,” in *Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept. 2012.

[9] A. Doucet, S. Godsill, and C. Andrieu, “On sequential Monte Carlo sampling methods for Bayesian filtering,” *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000.

[10] E. A. Lehmann, *Particle filtering methods for acoustic source localization and tracking*, Ph.D. thesis, The Australian National University, Canberra ACT 0200, Australia, 1995.

[11] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, “A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking,” *IEEE Trans. on Sig. Proc.*, vol. 50, no. 2, pp. 174–187, February 2002.

[12] E. A. Lehmann and R. C. Williamson, “Particle filter design using importance sampling for acoustic source localization and tracking in reverberant enviroments,” *IEURASIP Journal on Applied Sig. Proc.*, 2006.

[13] A. Levy, S. Gannot, and E. A. P Habets, “Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments,” *IEEE Trans. on Audio, Speech, and Language Proc.*, no. 99, pp. 1540–1555, 2011.

[14] G. Uhlenbeck and L. Ornstein, “On the theory of the Brownian motion,” *Physical Review*, vol. 36, no. 5, pp. 823–841, Sept. 1930.

[15] D. W. Heuring, and J. J. Murray, “Modeling and copying human head movements,” *IEEE Trans. on Robotics and Automation*, vol. 15, no. 6, pp. 1095–1108, 1999.

[16] R. O. Duda and W. L. Martens, “Range dependence of the response of a spherical head model,” *J. Acoust. Soc. Am.*, pp. 3048–3057, 1998.

[17] D. Ward, E. A. Lehmann, and R. C. Williamson, “Particle filtering algorithms for acoustic source localization,” *IEEE Trans. on Speech and Audio Proc.*, 2002.

[18] D. Ward and R. C. Williamson, “Particle filter beamforming for acoustic source localization in a reverberant environment,” in *IEEE Int. Conf. of Acoust., Speech, Sig. Proc. (ICASSP)*, Orlando, FL, May 2002.

[19] J. Chen, J. Benesty, and Y. Huang, “Time delay estimation in room acoustic environments: an overview,” *EURASIP J. on Applied Sig. Proc.*, 2006.

[20] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[21] T. G. Dvorkind and S. Gannot, “Time difference of arrival estimation of speech source in a noisy and reverberant environment,” *Signal Processing*, vol. 85, no. 1, pp. 177–204, Jan. 2005.

[22] J. Blauert, *Spatial hearing*, Hirzel-Verlag, 3rd edition, 2001.

[23] 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 Audio Process.*, vol. 11, no. 6, pp. 826–836, Nov. 2003.

[24] R. Douc, O. Cappe, and E. Moulines, “Comparison of resampling schemes for particle filtering,” in *Proceedings of the 4th International Symp. on Image and Sig. Proc. and Analysis (ISPA)*, 2005, p. 64.