Sensors **2014**, *14*, 2259-2279; doi:10.3390/s140202259

OPEN ACCESS

sensors

ISSN 1424-8220

www.mdpi.com/journal/sensors

Article

3D Joint Speaker Position and Orientation Tracking with Particle Filters

Carlos Segura [1,*] **and Javier Hernando** [2]

[1] Herta Security, Barcelona 08037, Spain

[2] Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona 08034, Spain; E-Mail: javier.hernando@upc.edu

* Author to whom correspondence should be addressed; E-Mail: cseguramail@gmail.com; Tel.: +34-936-020-888; Fax: +34-934-016-447.

Received: 24 December 2013; in revised form: 17 January 2014 / Accepted: 25 January 2014 / Published: 29 February 2014

Abstract: This paper addresses the problem of three-dimensional speaker orientation estimation in a smart-room environment equipped with microphone arrays. A Bayesian approach is proposed to jointly track the location and orientation of an active speaker. The main motivation is that the knowledge of the speaker orientation may yield an increased localization performance and *vice versa*. Assuming that the sound produced by the speaker is originated from his mouth, the center of the head is deduced based on the estimated head orientation. Moreover, the elevation angle of the head of the speaker can be partly inferred from the fast vertical movements of the computed mouth location. In order to test the performance of the proposed algorithm, a new multimodal dataset has been recorded for this purpose, where the corresponding 3D orientation angles are acquired by an inertial measurement unit (IMU) provided by accelerometers, magnetometers and gyroscopes in the three-axes. The proposed joint algorithm outperforms a two-step approach in terms of localization and orientation angle precision assessing the superiority of the joint approach.

Keywords: head pose; speaker orientation; acoustic source localization

Classification: PACS 43.60.Jn, 43.60.Fg, 43.70.Bk, 43.60.Bf

---

1. Introduction

In recent years, significant research efforts have been focused on developing human-computer interfaces in intelligent environments that aim to support human tasks and activities. The knowledge of the position and the orientation of the speakers present in a room constitutes valuable information allowing for better understanding of user activities and human interactions in those environments, such as the analysis of group dynamics or behaviors, deciding which is the active speaker among all present or determining who is talking to whom. In general, it can be expected that the knowledge about the orientation of human speakers would permit the improvement of speech technologies that are commonly deployed in smart-rooms. For instance, an enhanced microphone network management strategy for microphone selection can be developed based on both speaker position and orientation cues.

Very few methods have been proposed to solve the problem of speaker localization and speaker orientation estimation from acoustic signals. They differ mainly in how they approach the problem and can be coarsely classified in to two groups. The first group assumes the task of localization and orientation estimation as two separate and independent problems, working as a two-step algorithm: first locate the speaker, and then, the head orientation is estimated [1–6]. The main advantage of this approach is the simplicity and processing speed. However, the main drawback of this method is that the head orientation estimation process is highly dependent on the speaker tracking accuracy. This kind of approach does not take advantage of the fact that speaker orientation information could be used to improve the speaker localization precision.

The second group of approaches [7,8] considers the localization and the estimation of the orientation of the speaker as a joint process, which aims at improving the performance of the localization by proper weighting of the cross-correlation between microphone pairs, depending on their relative angle with the speaker, thus minimizing the degrading effects of the head orientation in the localization algorithm [9].

No previous work has been found that tackles the task of three-dimensional (3D) speaker orientation estimation with microphone arrays. This can be attributed to the fact that most smart environments have the microphones placed in nearly the same plane in order to maximize the localization performance in the xy coordinates, making it very difficult to estimate the head elevation angle, due to the low microphone placement diversity in the z-axis. Another possible cause may be the lack of acoustic databases with annotated speaker orientation and not even 3D orientation labels.

In this paper, a Bayesian approach is proposed to jointly track the location and orientation of a speaker. The main motivation is that the knowledge of the speaker orientation may yield to an increased localization performance and *vice versa*. The position and orientation of the speaker are estimated in the 3D space by means of a joint particle filter (PF) with coupled dynamic and observation models. Furthermore, the part from the vertical angle of the speaker's head can be inferred by the algorithm solely from the acoustic cues. In order to test the performance of the proposed algorithm, a new multimodal dataset has been purposely recorded, where the corresponding 3D orientation angles are acquired by an inertial measurement unit (IMU) provided by accelerometers, magnetometers and gyroscopes in the three axes. The position of the center of the head of the speaker is automatically provided by a video particle filter tracker from multiple cameras. The effectiveness of the proposed technique is assessed by means of a new proposed set of metrics derived from the multiple person tracking task [10] in

Section 6.2 over the cited database, showing an increased performance for the joint PF approach in relation to the two two-step algorithms that first estimate the position and then the orientation of the speaker.

The remainder of this paper is organized as follows. In Section 2, the head rotation representation is described. Section 3 introduces the speaker localization and orientation estimation algorithms as a two-step approach. Section 4 presents an alternative two-step algorithm employing a PF at each step. Section 5 describes the joint PF. Sections 6 and 7 show the experiments and results. Finally, Section 8 gives the conclusions.

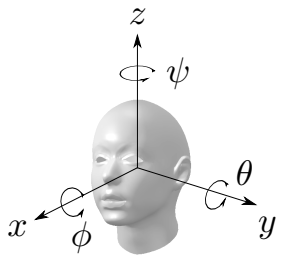
2. Head Rotation Representation

The parametrization of the head rotation in this work is based on the decomposition into Euler angles (ϕ, θ, ψ) with the $x-y-z$ convention of the rotation matrix of the head into the room's frame of reference, where (ϕ, θ, ψ) denote the three basic rotations, one for every axis. By the $x - y - z$ convention, the following rotations are chosen:

- Rotate by angle ψ about the head z-axis
- Rotate by angle θ about the head y-axis
- Rotate by angle ϕ about the head x-axis

These rotations are shown in Figure 1.

Figure 1. Euler angles, basic head rotations.



The rotation matrix, $\mathbf{R}(\phi, \theta, \psi)$, is given by:

$$\mathbf{R}(\phi, \theta, \psi) = \mathbf{R}_z(\psi)\mathbf{R}_y(\theta)\mathbf{R}_x(\phi), \tag{1}$$

where:

$$\mathbf{R}_x(\phi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\phi) & -\sin(\phi) \\ 0 & \sin(\phi) & \cos(\phi) \end{bmatrix} \tag{2}$$

$$\mathbf{R}_y(\theta) = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix} \tag{3}$$

$$\mathbf{R}_z(\psi) = \begin{bmatrix} \cos(\psi) & -\sin(\psi) & 0 \\ \sin(\psi) & \cos(\psi) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

The Euler angles (ϕ, θ, ψ) are also known as the *roll*, *tilt* and *pan*; or *roll*, *pitch* and *yaw* angles of the head. In this work, it seems not feasible to estimate the roll of the head with acoustic signals. Therefore, only the pan and tilt will be considered. Thus, the rotation of the head will be parametrized as $\mathbf{R}(\theta, \psi) = \mathbf{R}(0, \theta, \psi) = \mathbf{R}_z(\psi)\mathbf{R}_y(\theta)$. Nevertheless, the knowledge of the horizontal and vertical head angles, in addition to the head location, gives a good representation of the speaker in the 3D space. In order to estimate what the speaker may be referring to, the direction vector of his head in the 3D space can be computed from the rotation matrix as follows:

$$\mathbf{d}(\phi, \theta) = \mathbf{R}(\theta, \psi) \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} cos(\theta)cos(\psi) \\ cos(\theta)sin(\psi) \\ -sin(\theta) \end{bmatrix} \tag{5}$$

3. Two-Step Speaker Localization-Orientation Algorithm

The two-step algorithm to estimate the location and the orientation of speakers is based on the work presented in [11]. First, the position of the speaker is estimated by the steered response power-phase transform (SRP-PHAT) algorithm and the time difference of arrival (or time delay of arrival) (TDOA) for each microphone pair with respect to the detected position is computed. In the second step, the energy of the cross-correlation nearby the estimated time delay is used as the fundamental characteristic from where to derive the speaker orientation.

3.1. Acoustic Source Localization

3.1.1. GCC-PHAT Algorithm

In a multi-microphone environment, one of the observable clues with positional information more commonly used in acoustic localization algorithms is the time difference of arrival of the signal between microphone pairs. Consider a smart-room provided with a set of M microphones from which we choose N microphone pairs. Let $\mathbf{x}$ denote a $\mathbb{R}^3$ position in space. Then, the time difference of arrival, $\tau_{\mathbf{p},i,j}$, of an hypothetical acoustic source located at $\mathbf{p}$ between two microphones, i, j, with positions $\mathbf{m}_i$ and $\mathbf{m}_j$ is:

$$\tau_{\mathbf{p},i,j} = \frac{\parallel \mathbf{p} - \mathbf{m}_i \parallel - \parallel \mathbf{p} - \mathbf{m}_j \parallel}{c}, \tag{6}$$

where c is the speed of sound in air.

The cross-correlation function is well-known as a measure of the similarity between signals for any given time displacement, and ideally, it should exhibit a prominent peak in correspondence to the delay between the pair of signals [12]. A commonly used weighting function in acoustic event localization is the phase transform (PHAT), also known in the literature as cross-power-spectrum phase technique [13], which is usually considered useful in reverberant conditions. It can be expressed in terms of the inverse Fourier transform of the estimated cross-power spectrum ($G_{ij}(f)$) with the following equation:

$$R_{ij}(\tau) = \int_{-\infty}^{\infty} U(f_1, f_2) \frac{G_{ij}(f)}{|G_{ij}(f)|} e^{j2\pi f\tau} df, \tag{7}$$

In practice, the frequency range used to compute $R_{ij}(\tau)$ can be reduced to the speech-band to increase the accuracy [14], employing the rectangular band-pass filter, $U(f_1, f_2)$, with a unitary value for frequencies $f_1 \leq |f| \leq f_2$, and zero otherwise.

The estimation of the TDOA for each microphone pair is computed as follows:

$$\hat{\tau}_{i,j} = \underset{\tau}{\operatorname{argmax}} R_{ij}(\tau) \tag{8}$$

3.1.2. SRP-PHAT Algorithm

The contributions of each microphone pair can be combined to derive a single estimation of the source position. However, in the general case, the availability of multiple TDOA estimations leads to a minimization of an over-determined and non-linear error function. A very efficient approach is the SRP-PHAT or global coherence field introduced in [14]. The SRP-PHAT algorithm performs very robustly in reverberant environments, due to the PHAT weighting, and actually, it has turned out to be one of the most successful state-of-the-art approaches to microphone array sound localization.

The basic operation of the SRP-PHAT algorithms consists of exploring the three-dimensional (3D) space, searching for the maximum of the global contribution of the PHAT-weighted generalized cross-correlations (GCC-PHAT) from all the microphone pairs. The 3D room space is quantized into a set of positions with a typical separation of 5–10 cm. The theoretical TDOA, $\tau_{\mathbf{p},i,j}$, from each exploration position to each microphone pair are precomputed and stored.

The set of GCC-PHAT functions are combined to create a spatial likelihood function (SLF) $F(\mathbf{p})$, which gives a score for each position, $\mathbf{p}$, in space by means of the following equation:

$$F(\mathbf{p}) = \sum_{i,j \in \mathbb{S}} R_{ij}(\tau_{\mathbf{p},ij}) \tag{9}$$

The estimated acoustic source location is the position of the quantized space that maximizes the contribution of the GCC-PHAT of all microphone pairs:

$$\hat{\mathbf{p}} = \underset{\mathbf{p}}{\operatorname{argmax}} F(\mathbf{p}), \tag{10}$$

where $\mathbb{S}$ is the set of microphone pairs. Then, the TDOA for each microphone pair, $\tau_{\hat{\mathbf{p}},i,j}$, is estimated using the obtained location.

3.2. Orientational Features

3.2.1. GCC-PHAT-A

The orientational cues used in this work are based on GCC-PHAT averaged peak (GCC-PHAT-A), described in [11]. It consists on computing the energy of the cross-correlation nearby the estimated time delay by the following equation:

$$\rho_{ij} \equiv \sum_{k=-\Delta}^{\Delta} |w(k) R_{ij}(k + \tau_{\hat{\mathbf{p}},ij})|^2, \tag{11}$$

where $\tau_{\hat{\mathbf{p}},ij}$ is the delay in samples and $w(k)$ is a window with length $L = 2\Delta + 1$. Different window types and lengths can be used in $w(k)$ with satisfactory performance, as addressed in [11].

Basically, the GCC-PHAT-A measure reduces to the sum of the energy of the band-filtered PHAT-weighted cross-correlation around the estimated TDOA, and essentially, it measures the proportion of the signal between frequencies f_1 and f_2 that contributes to the main peak in the localization. It is also important to note that this measure is commensurable across all microphone pairs independent of microphone gains, due to the PHAT weighting and, therefore, constitutes a valuable orientational feature.

3.2.2. Orientation Angle Estimation

In order to estimate the orientation of a speaker based on the GCC-PHAT-based orientational measures, a simple vectorial method is employed, similar to that described in [8]. The technique first needs the position of the active person to be known beforehand or estimated by means of the SRP-PHAT or any other source localization method. Then, the vectors, $\mathbf{v}_{ij}$, from the speaker to the center of each microphone pair are computed, adjusting their magnitude $|\mathbf{v}_{ij}|$ to the orientational measure of the microphone pair, ρ_{ij}. The orientational measures consists in the min-max-normalization scaled GCC-PHAT-A values, which fit in the range $[-\gamma, (1-\gamma)]$.

$$\bar{\rho}_{ij} = \frac{(\rho_{ij} - \rho_{min})}{(\rho_{ij} - \rho_{max})} - \gamma \tag{12}$$

$$\mathbf{v}_{ij} = \bar{\rho}_{ij} \frac{\hat{\mathbf{p}} - (\mathbf{m}_i + \mathbf{m}_j)/2}{\parallel \hat{\mathbf{p}} - (\mathbf{m}_i + \mathbf{m}_j)/2 \parallel}, \tag{13}$$

where ρ_{min} and ρ_{max} are the minimum and maximum value of the set of ρ_{ij}. Min-max normalization retains the original distribution of values, except for a scaling factor and transforms all values into the desired range [15]. The min-max normalization models the fact that the microphone pairs with the lowest orientational cue value are probably behind the speaker, and by giving those pairs a negative value, its resulting vector would help point to the correct direction. In our experiments, we obtained good results with $\gamma = 0.3$.

The sum of the vectors formed by all the orientational measures of each microphone pair is considered the estimated head direction, $\mathbf{v}_{sum}$, as follows:

$$\mathbf{v}_{sum} = \sum_{i,j \in \mathbb{S}} \mathbf{v}_{ij} \tag{14}$$

The estimated head orientation angle, $\hat{\psi}$, is computed as the angle of the projection of $\mathbf{v}_{sum}$ in the xy-plane with the x-axis.

4. Two-Step Particle Filter Tracking

In this section, a two-step approach to estimate the location and orientation of the speaker is proposed, employing a particle filter in each stage, which is introduced here to enable a fair comparison with the joint particle filter approach.

4.1. Particle Filter Tracking

The concept of tracking can be defined as the recursive estimation of the hidden state of a target based on the partial observations at every time instant. Assuming that the evolution of the state sequence is defined by a Markov process of first order, the dynamics of the state can be described by the transition equation:

$$\mathbf{x}_k = \mathbf{f}_k(\mathbf{x}_{k-1}, \mathbf{v}_{k-1}) \tag{15}$$

where $\mathbf{f}_k$ is a possibly non-linear function of the previous state, $\mathbf{x}_{k-1}$, and an independent and identically distributed (i.i.d.) process noise, $\mathbf{v}_{k-1}$. At every time instant, k, the observation of the state, $\mathbf{x}_k$, is defined by the observation equation:

$$\mathbf{z}_k = \mathbf{h}_k(\mathbf{x}_k, \mathbf{n}_k) \tag{16}$$

where, again, $\mathbf{h}_k$ is, in general, a non-linear function of the state and an i.i.d. measurement noise sequence, $\mathbf{n}_k$.

Tracking aims to estimate $\mathbf{x}_k$ based on the set of all available measurements $\mathbf{z}_{1:k} = \{\mathbf{z}_i, i = 1, \ldots, k\}$ up to time k. One solution is to use the Bayesian approach to reconstruct the probability density function (pdf) of $\mathbf{x}_k$ given all the data, $\mathbf{z}_{1:k}$, up to time k, or in a compact notation, $p(\mathbf{x}_k|\mathbf{z}_{1:k})$. The pdf, $p(\mathbf{x}_k|\mathbf{z}_{1:k})$, is known as the *posterior density* and contains all statistical information gathered by the measurements up to time k. The posterior density may be obtained recursively by means of the Bayesian approach based on two fundamental iteration steps, namely, prediction and update.

In the prediction step, the prior pdf, $p(\mathbf{x}_k|\mathbf{z}_{1:k-1})$, is obtained making use of the *transition* pdf, $p(\mathbf{x}_k|\mathbf{x}_{k-1})$, which is derived from transition Equation (15):

$$p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1}) d\mathbf{x}_{k-1} \tag{17}$$

In the update stage, the new measurement, $\mathbf{z}_k$, is used to update the prior pdf via the Bayes' rule and obtain the required posterior density of the current state:

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) = \frac{p(\mathbf{z}_k|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} \tag{18}$$

where the denominator:

$$p(\mathbf{z}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{z}_k|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) d\mathbf{x}_k \tag{19}$$

is a normalizing constant, which depends on the pdf, $p(\mathbf{z}_k|\mathbf{x}_k)$, defined by observation Equation (16).

Particle filters (PF) [16] approximate the Bayesian filter approach by representing the probability distribution recursively with a finite set of samples, known as particles, that are updated according to their measured likelihood for a given dynamical and observational model. Applications of PF to acoustic localization can be found in [17–19] with a comprehensive research in [20].

Let $\{\mathbf{x}_k^i\}_{i=1}^{N_s}$ denote a set of N_s random samples of the state with associate weights $\{w_k^i\}_{i=1}^{N_s}$, normalized such that $\sum_i w_k^i = 1$. Then, the posterior density, $p(\mathbf{x}_k|\mathbf{z}_{1:k})$, can be approximated as:

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \approx \sum_{i=1}^{N_s} w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i) \tag{20}$$

Considering that the samples, $\mathbf{x}_k^i$, are drawn from a sampling distribution, $q(\mathbf{x}_k|\mathbf{x}_{k-1}^i, \mathbf{z}_k)$, called *importance density*, and taking some widely accepted assumptions [16], the weights can be computed recursively by the following expression:

$$w_k^i \propto w_{k-1}^i \frac{p(\mathbf{z}_k|\mathbf{x}_k^i)p(\mathbf{x}_k|\mathbf{x}_{k-1}^i)}{q(\mathbf{x}_k|\mathbf{x}_{k-1}^i, \mathbf{z}_k)} \tag{21}$$

In the literature regarding other domains, some techniques aim at constructing efficient importance density functions through Markov Chain Monte Carlo methods [21] or exploiting independence among variables in the state space using Rao-Blackwellized particle filters [22]. Although there is a large number of methods to compute the associated particle weights, one approach that is the most largely accepted, in part for its convenience, is to choose the importance density to be the prior:

$$q(\mathbf{x}_k|\mathbf{x}_{k-1}^i, \mathbf{z}_k) = p(\mathbf{x}_k|\mathbf{x}_{k-1}^i) \tag{22}$$

reducing the weight recursion to:

$$w_k^i \propto w_{k-1}^i p(\mathbf{z}_k|\mathbf{x}_k^i) \tag{23}$$

A common problem with the PF is the degeneracy phenomenon, where, after a few iterations, all the weight concentrates in just one particle, and the rest of the particles have almost zero contribution to the approximation of the posterior. A measure of the degeneracy of the PF is the *effective sample size* introduced in [23] and [24], defined as:

$$\widehat{N_{eff}} = \frac{1}{\sum_i^{N_s} (w_k^i)^2} \tag{24}$$

where $\widehat{N_{eff}} \leq N_s$ and a small $\widehat{N_{eff}}$ is a symptom of severe degeneracy. Although this problem could be tackled by using a very large N_s, a common approach, whenever a significant degeneracy is observed, is to make use of particle resampling techniques, which consist of discarding the particles with lower weight and proportionally replicating those with a higher one, while still representing the posterior density.

The best estimation of the state at time k, $\hat{\mathbf{x}}_k$, is derived based on the discrete approximation of Equation (20). The most common solution is the Monte Carlo approximation of the expectation:

$$\hat{\mathbf{x}}_k = \mathbb{E}[\mathbf{x}_k|\mathbf{z}_{1:k}] \approx \sum_{i=1}^{N_s} w_k^i \mathbf{x}_k^i \tag{25}$$

The design parameters of the PF are the state model, the dynamical model and the observational model, which are defined in the following sections.

4.2. Location Tracking

4.2.1. State and Dynamical Models

A common approach is to characterize the human movement dynamics as a *Langevin process* [25], since it is reasonably simple and has been proven to work well in practical applications [19,25]. In this case, the state variable, $\mathbf{x}_k$, is defined as:

$$\mathbf{x}_k = \begin{bmatrix} \mathbf{p}_k \\ \dot{\mathbf{p}}_k \end{bmatrix} \tag{26}$$

where $\mathbf{p}_k = [x_k\ y_k\ z_k\ ]^T$ denotes the position and $\dot{\mathbf{p}}_k = [\dot{x}_k\ \dot{y}_k\ \dot{z}_k\ ]^T$ denotes the velocity of the target. The addition of the velocity component in the state variable aims to improve the representation of the movement dynamics.

For the sake of simplicity, consider the Langevin process in the x-coordinate as follows:

$$x_k = x_{k-1} + T\dot{x}_k \tag{27}$$

$$\dot{x}_k = a\dot{x}_{k-1} + \sigma_x n_x \tag{28}$$

where $n_x \sim \mathcal{N}(0,1)$ is a normally distributed random variable, T is the time step unit between consecutive updates of the state vector and the two constants are defined as:

$$a = e^{-\beta T} \tag{29}$$

$$\sigma_x = v_x\sqrt{1-a^2} \tag{30}$$

where v denotes the steady-state root mean square velocity and β is the rate constant. The motion model in the x and y coordinates is assumed to be independent and identically distributed, which yields to identical model parameters in both coordinates. The random variable, n_z, for the z-axis is set to have a normal distribution with a very low variance, σ_z^2. Equations (27) and (28) can be rewritten following the form of transition Equation (15):

$$\mathbf{x}_k = \underbrace{\begin{bmatrix} 1 & 0 & 0 & aT & 0 & 0 \\ 0 & 1 & 0 & 0 & aT & 0 \\ 0 & 0 & 1 & 0 & 0 & aT \\ 0 & 0 & 0 & a & 0 & 0 \\ 0 & 0 & 0 & 0 & a & 0 \\ 0 & 0 & 0 & 0 & 0 & a \end{bmatrix}}_{\mathbf{F}} \mathbf{x}_{k-1} + \mathbf{v}_{k-1} \tag{31}$$

with $\mathbf{v}_{k-1}$ characterized as a zero-mean Gaussian noise variable with covariance matrix $\mathbf{Q}_{k-1}$:

$$\mathbf{v}_{k-1} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \underbrace{\begin{bmatrix} \sigma_x^2T^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_y^2T^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_z^2T^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_x^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_y^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_z^2 \end{bmatrix}}_{\mathbf{Q}_{k-1}} \right) \tag{32}$$

4.2.2. Observational Model

The particle filter approach requires the definition of the likelihood function, $p(\mathbf{z}_k|\mathbf{x}_k^i)$, in order to update the weight of every particle. In this case, the observation, $\mathbf{z}_k$, is not limited to the estimated source location [19], and the full SRP-PHAT SLF generated by Equation (9) or a modification thereof can be employed [26]. Other works [17] construct the likelihood function employing solely the TDOA estimations.

In this work, the localization likelihood is derived from a spatial likelihood function $F(\mathbf{p})$ obtained by the SRP-PHAT algorithm with the PHAT-weighted cross-correlation smoothed by the convolution with a triangular window, $\Omega(\tau)$, of five samples:

$$\tilde{R}_{ij}(\tau) = \left(R_{ij}^2(\tau) * \Omega(\tau)\right)^{1/2} \quad (33)$$

$$F(\mathbf{p}) = \left(\sum_{i,j \in \mathbb{S}} \tilde{R}_{ij}(\tau_{\mathbf{p},i,j})\right)^2 \quad (34)$$

Given the iterative nature of the PF, this smoothed SLF enables a faster convergence of the particles to its global maximum, while avoiding being trapped around local maxima. Since the position that maximizes $F(\mathbf{p})$ determines the most probable location of the sound source, the localization observation likelihood function is constructed from the estimated position of the speaker's mouth, $\mathbf{t}_k$, and the SLF:

$$p(\mathbf{z}_{k,loc}|\mathbf{x}_k) = F(\mathbf{t}_k), \quad (35)$$

where $\mathbf{z}_{k,loc}$ denotes the observation of the localization.

The likelihood function, $F(\mathbf{p})$, is usually precomputed for a discrete set of space positions for every audio frame in order to gain speed in the evaluation of $p(\mathbf{z}_{k,loc}|\mathbf{x}_k)$ in the case of a PF with a large number of particles, at the expense of localization precision. In this work, the quantization step is set to 5 cm.

4.3. Orientation Tracking

4.3.1. State and Dynamical Models

The state vector of the particle filter used to estimate the orientation consists only of the pan angle and the dynamical model as follows:

$$\mathbf{x}_k = \psi_k \quad (36)$$

$$\psi_k = \psi_{k-1} + n_\psi \quad (37)$$

where $n_\psi \sim \mathcal{N}(0, \sigma_\psi^2)$ is a normally distributed random variable.

The state head direction vector in 3D space $\mathbf{d}_k(\psi_k) = \begin{bmatrix} cos(\psi_k) & sin(\psi_k) & 0 \end{bmatrix}^T$.

4.3.2. Observational Model

The orientation likelihood is obtained from the GCC-PHAT averaged peak features described in Section 3.2. A vector, $\mathbf{v}_n$, is created from the estimated speaker's position, $\mathbf{p}_k$, to the center of each microphone pair, adjusting their magnitude $|\mathbf{v}_n|$ to the normalized orientational measure of the microphone pair as defined in Section 3.2.2. The orientation observation is formed by the resulting vector, $\mathbf{v}_{sum}$, of the vectorial sum of $\mathbf{v}_n$. The orientation likelihood function is then defined as the scalar product of the state head direction vector and the normalized resulting vector as follows:

$$p(\mathbf{z}_{k,ori}|\mathbf{x}_k) = \left(\frac{\langle \mathbf{d}_k(\psi_k), \frac{\mathbf{v}_{sum}^T}{|\mathbf{v}_{sum}|} \rangle + 1}{2}\right)^{C|\mathbf{v}_{sum}|} \quad (38)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, $\mathbf{z}_{k,ori}$ is the observation of the orientation and C is a constant to control the width of the observation probability function.

The scalar product of the two unitary vectors is scaled into the range $[0, 1]$ to better resemble a likelihood function. The exponent, $n|\mathbf{v}_{sum}|$, is used as a *confidence factor* for the orientational observation, with the constant, n, set empirically to four. The magnitude of the observation vector, $|\mathbf{v}_{sum}|$, models the likelihood function, where a very small value of the vector length yields to the constant likelihood function independent of the state. On the other hand, higher values of the observation vector magnitude will narrow the likelihood function to observation angles close to the state angle.

5. Joint Localization-Orientation Particle Filter Tracker

In this work, a particle filter approach is proposed to jointly track the location and orientation of a speaker. The main motivation is that the knowledge of the speaker orientation may yield to an increased localization performance and *vice versa*. The position and orientation of the speaker are estimated in the 3D space by means of a joint particle filter with coupled dynamic and observation models. The proposed system makes the assumption that the voice of a speaker is produced around the mouth, and the knowledge about the orientation yields to a better estimate of the head position. On the other hand, in this work, it is proposed to assume that the person movement is dependent on his orientation and *vice versa*. Next sections describe the proposed state and coupled dynamic and observation models.

5.1. State Model

The state of the particles is composed by the position of the center of the speaker's head $\mathbf{p}_k = [x_k \; y_k \; z_k]^T$, the velocity of the speaker $\dot{\mathbf{p}}_k = [\dot{x}_k \; \dot{y}_k \; \dot{z}_k]^T$ and the tilt and pan of his head.

$$\mathbf{x}_k = \begin{bmatrix} \mathbf{p}_k \\ \dot{\mathbf{p}}_k \\ \psi_k \\ \theta_k \end{bmatrix} \tag{39}$$

The estimated head rotation at any time is defined by:

$$\mathbf{R}_k(\theta_k, \psi_k) = \begin{bmatrix} \cos(\theta_k)\cos(\psi_k) & -\sin(\psi_k) & \sin(\theta_k)\cos(\psi_k) \\ \cos(\theta_k)\sin(\psi_k) & \cos(\psi_k) & \sin(\theta_k)\sin(\psi_k) \\ -\sin(\theta_k) & 0 & \cos(\theta_k) \end{bmatrix} \tag{40}$$

The estimation of the position of the speaker's mouth $\mathbf{t}_k$ is determined at every instant by the state vector, and it is synthesized from the head center position and the rotation angles as follows:

$$\mathbf{t}_k = \mathbf{p}_k + \mathbf{R}_k(\theta_k + \alpha, \psi_k) \begin{bmatrix} r & 0 & 0 \end{bmatrix}^T \tag{41}$$

where it has been assumed that the mouth lies at r distance from the head center with an inclination angle of α. A preliminary radius of r = 15 cm and an inclination of α = 45 degrees have been set experimentally.

The state head direction vector in the 3D space, $\mathbf{d}_k(\theta_k, \psi_k)$, is computed, rotating the head direction vector in the head coordinate reference to the 3D space reference:

$$\mathbf{d}_k(\theta_k, \psi_k) = \mathbf{R}_k^i(\theta_k, \psi_k) \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T \tag{42}$$

5.2. Dynamical Model

Similarly to Section 4.2.1, a *Langevin process* is chosen to characterize the speaker movement dynamics. Usually, the motion model in the x and y coordinates is assumed to be independent and identically distributed, which yields to identical model parameters in both coordinates. However, in this work, it is assumed that the movement in the x and y coordinates is dependent on the *pan* orientation angle of the person. It is expected as a more probable event that the speaker moves to his forward direction than to his sideways or backward directions. This is modeled as a Rayleigh distribution probability in the speaker's forward direction and a normal distribution in his sideways direction. The Rayleigh distribution, $\mathcal{R}(0, 1)$, is scaled and centered in order to have a zero mean expectation and unity variance. The variance of the distributions determined by the σ factor is also different for the forward and sideways directions.

$$n_{forward} \sim \mathcal{R}(0, \sigma^2_{forward}) \tag{43}$$

$$n_{sideway} \sim \mathcal{N}(0, \sigma^2_{sideway}) \tag{44}$$

The random variable, n_x, from Equation (28) for the x and y coordinates are obtained by the rotation of $n_{forward}$ and $n_{sideway}$ by the *pan* angle ψ_k:

$$\begin{bmatrix} n_x \\ n_y \end{bmatrix} = \begin{bmatrix} \cos(\psi_k) & -\sin(\psi_k) \\ \sin(\psi_k) & \cos(\psi_k) \end{bmatrix} \begin{bmatrix} n_{forward} \\ n_{sideway} \end{bmatrix} \tag{45}$$

The random variable, n_z, for the z-axis is set to have a normal distribution with a very low variance, σ_z^2.

In this work, the horizontal orientation angle of the speaker is assumed to be dependent on his velocity. It is expected that the faster the person moves, the more probable it is that the person is looking to his moving direction. This is modeled by predicting the next state head direction as the weighted sum of the current state head direction vector in the xy plane and the normalized moving direction vector plus a normally distributed random variable, $\mathbf{n}_d$, where the weight factor, α_ψ, depends on the person's velocity and the maximum expected velocity, v_{max}, as follows:

$$\alpha_\psi = \frac{|\dot{\mathbf{p}}_{k-1}|}{v_{max}} \tag{46}$$

$$\begin{bmatrix} d_x \\ d_y \\ d_z \end{bmatrix} = (1 - \alpha_\psi)\mathbf{d}_{k-1}(0, \psi_{k-1}) + \alpha_\psi \frac{\dot{\mathbf{p}}_{k-1}}{|\dot{\mathbf{p}}_{k-1}|} + \mathbf{n}_d \tag{47}$$

$$\mathbf{n}_d \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_d^2 & 0 & 0 \\ 0 & \sigma_d^2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \right) \tag{48}$$

d_x, d_y and d_z being the x, y and z components of $\mathbf{d}_{k-1}$.

Finally, the next state *yaw* orientation is the angle formed by the y and x components of the head direction:

$$\psi_k = \arctan(d_y, d_x) \tag{49}$$

The *pitch* orientation angle recursion equation, assuming independence with other state variables, is defined as:

$$\theta_k = \alpha_\theta \theta_{k-1} + n_\theta \tag{50}$$

where α_θ is a forgetting factor accomplishing $|\alpha_\theta| \leq 1$ and $n_\theta \sim \mathcal{N}(0, \sigma_\theta^2)$ is a normally distributed random variable. The *pitch*, θ_k, determines the height of the mouth of the speaker in relation to the head center position. The variables, α_θ, n_θ and n_z, are adjusted, so that short-term vertical head movements are inferred by θ_k, whereas long-term smooth head height changes are incorporated into the state head height, due to the forgetting factor.

5.3. Observational Model

The observation likelihood, $p(\mathbf{z}_k|\mathbf{x}_k)$, is composed from the localization, $\mathbf{z}_{k,loc}$, from Equation (35) and orientation $\mathbf{z}_{k,ori}$ from Equation (38) feature observations as follows:

$$p(\mathbf{z}_k|\mathbf{x}_k) = p(\mathbf{z}_{k,loc}|\mathbf{x}_k)p(\mathbf{z}_{k,ori}|\mathbf{x}_k) \tag{51}$$

where it is assumed that these observations are conditionally independent, given the current state, $\mathbf{x}_k$.

6. Experiments

6.1. Experimental Setup and Database Description

The joint PF tracker performance will be compared with the two two-step algorithms introduced in Sections 3.2.2. and 4 in the task of estimating the position and orientation of the speaker's head. Since the two-step approaches are only able to estimate the horizontal orientation angle, the pitch and roll hypothesis are set to 0 for all time frames. The comparison with the two-step PF approach assesses that the performance increase obtained by the joint method is due to the joint dynamic and observation models and not the filtering itself.

Figure 2. Smart-room sensor setup used in this database, with 5 cameras (Cam1-Cam5) and 6 T-shaped microphone clusters (T-Cluster 1 -6).

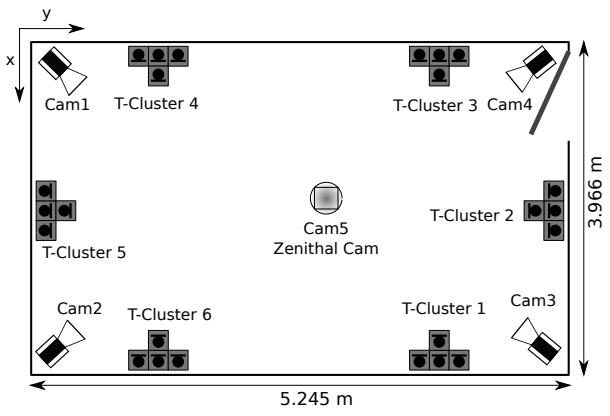
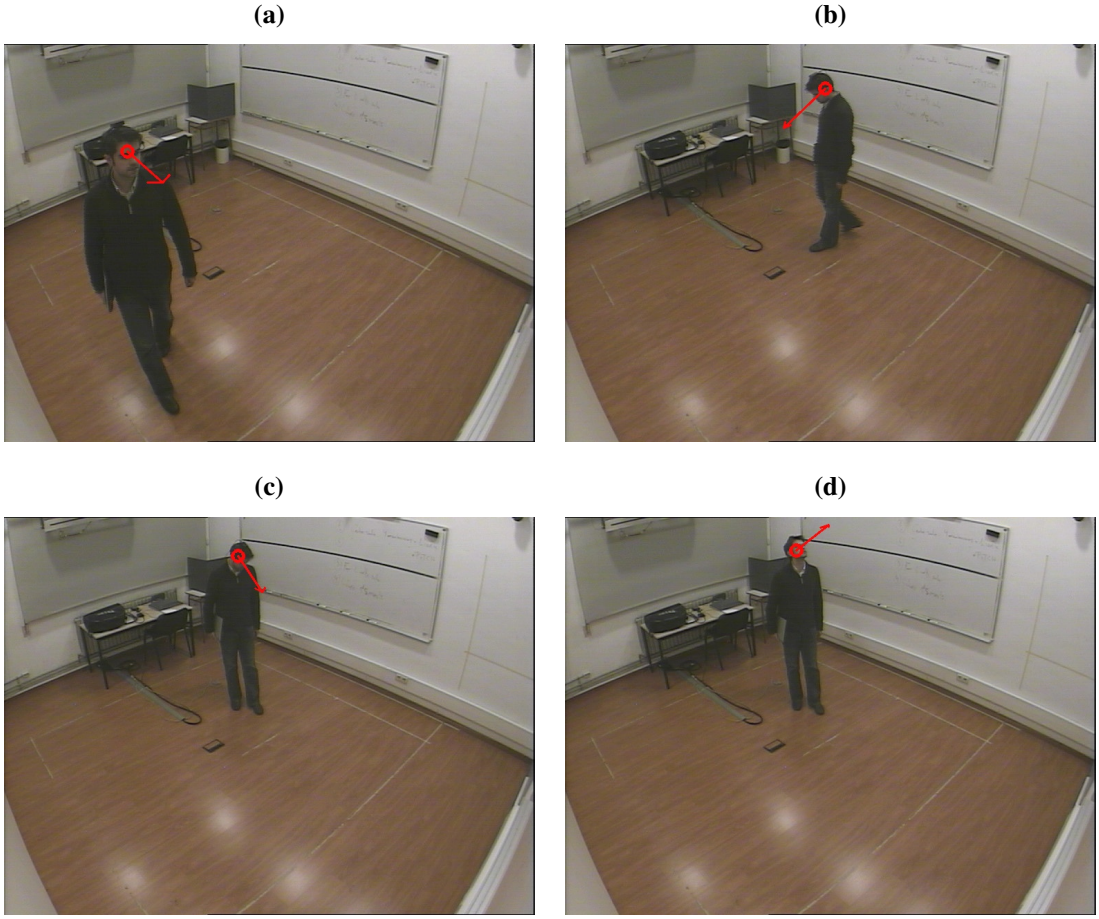


Figure 3. Single person dataset snapshots, with superposed head position and rotation annotations.



The performance of the proposed head orientation estimation algorithm was evaluated using a purposely recorded database collected in the smart-room at the Universitat Politècnica de Catalunya. It is a meeting room equipped with several multimodal sensors, such as microphone arrays, table-top microphones and fixed or pan-tilt-zoom video cameras. The room dimensions are $3,966 \times 5,245 \times 4,000$ mm, which correspond to the x, y and z coordinates, respectively, and its measured reverberation time is approximately 400 ms. A schematic figure of the room setup can be observed in Figure 2.

The database is composed of one single person dataset involving the recording of multi-microphone audio, multi-camera video and IMU data for seven people moving freely in a smart room speaking most of the time and another multi-person dataset consisting in the recording of a group discussion with four participants. Only the simple person dataset will be considered in this work, since it is oriented toward the person tracking task, while the multi-person dataset is oriented toward the group analysis task. A sample of the database is shown in Figure 3.

The ground truth provided by the database consists in the annotations of the center of the head and the Euler rotation angles of every participant. The center of the head was obtained automatically by means of a multi-camera video PF tracker and the Euler orientation angles are acquired by an inertial measurement unit (IMU) provided by accelerometers, magnetometers and gyroscopes in the three axes.

6.2. Metrics

The metrics proposed in [10] for acoustic and audiovisual person-tracking are considered for evaluation and comparison purposes. These metrics have been used in international evaluation contests [27] and have been adopted by several research projects, such as the European Computers in the Human Interaction Loop (CHIL) [28] or the U.S. VACE [29] thus, they allow an objective and fair comparison with other acoustic tracking methods and with methods from other modalities.

In [10], two metrics are defined for an acoustic and audiovisual person-tracking task. Multiple object tracking precision (MOTP), which shows the trackers ability to estimate precise object positions, whereas multiple object tracking accuracy (MOTA) expresses the performance for estimating the correct number of objects and keeping to consistent trajectories. Additionally, the acoustic multiple object tracking accuracy (A-MOTA) score is defined for the acoustic tracking task, evaluated only for the active speaker at each time instant, k. A new metric is proposed in this work, multiple head orientation tracking precision (MHOTP), which determines the performance for estimating the head orientation of multiple persons.

6.2.1. Multiple Object Tracking Accuracy (MOTA) (%)

This is the accuracy of the tracker when it comes to keeping correct correspondences over time, estimating the number of people, recovering tracks, *etc*., the tracker, false positives, misses and mismatches, over all frames, divided by the total number of ground truth points.

$$MOTA = 1 - \frac{\sum_k ms_k + \sum_k fp_k + \sum_k mm_k}{\sum_k g_k} \tag{52}$$

where ms_k, fp_k and mm_k denote, respectively, the number of misses, false positives and mismatches, and g_k is the number of ground truth objects at time instant k. A distance threshold of 1 m is used to associate a track with the ground truth. Distances above this threshold will be treated as either false positives or mismatches. A more detailed description of this metric can be found in [10].

6.2.2. Multiple Object Tracking Precision (MOTP) (mm)

This is the precision of the tracker when it comes to determining the exact position of a tracked person in the room.

$$MOTP = \frac{\sum_{i,k} d_{i,k}}{\sum_k c_k} \tag{53}$$

where c_k is the number of correspondence matches found for time frame k and $d_{i,k}$ is the distance between the ground truth position and its corresponding hypothesis.

6.2.3. Multiple Head Orientation Tracking Precision (MHOTP) (degrees)

This is the precision of the tracker when it comes to determining the exact orientation of a tracked person in the room. It is the Euclidean angle error for matched *ground truth-hypothesis* pairs over all frames, averaged by the total number of matches made. It shows the ability of the tracker to estimate the correct orientation and is independent of its tracking accuracy. The Euclidean angle is computed as the

angle between the estimated head direction vector, $\hat{\mathbf{d}}(\hat{\theta}, \hat{\psi})$, and the ground truth vector, $\mathbf{d}(\theta, \psi)$. The multiple head orientation tracking precision can be also detailed by three sub-metrics, which account for the angle error in every axis.

$$MHOTP_{\psi} = \frac{\sum_{i,k} |\hat{\psi}_{i,k} - \psi_{i,k}|}{\sum_k c_k} \quad (54)$$

$$MHOTP_{\theta} = \frac{\sum_{i,k} |\hat{\theta}_{i,k} - \theta_{i,k}|}{\sum_k c_k} \quad (55)$$

$$MHOTP_{\phi} = \frac{\sum_{i,k} |\hat{\phi}_{i,k} - \phi_{i,k}|}{\sum_k c_k} \quad (56)$$

$$MHOTP = \frac{\sum_{i,k} \arccos(\langle \mathbf{d}(\theta_{i,k}, \psi_{i,k}), \mathbf{d}(\hat{\theta}_{i,k}, \hat{\phi}_{i,k}) \rangle)}{\sum_k c_k} \quad (57)$$

$$\mathbf{d}(\theta, \psi) = \begin{bmatrix} \cos(\theta)\cos(\psi) \\ \cos(\theta)\sin(\psi) \\ -\sin(\theta) \end{bmatrix} \quad (58)$$

where $\phi_{i,k}$, $\theta_{i,k}$ and $\psi_{i,k}$ are the ground truth Euler angles for the target, i, at the time instant, k, and $\hat{\phi}_{i,k}$, $\hat{\theta}_{i,k}$ and $\hat{\psi}_{i,k}$ are the estimated Euler angles for the corresponding hypothesis.

7. Results

Experiments were conducted over the cited database to compare the performance of the joint PF tracker and the two two-step approaches. A tight relationship between the tracking accuracy (MOTA) and precision (MOTP and MHOTP) has been observed in the three algorithms, since it is possible to output a localization and orientation hypothesis only when the confidence of the algorithm is above a threshold and achieve a high precision at the expense of tracking accuracy, and *vice versa*. In order to ensure a fair comparison between the three algorithms, the peak value of the SLF is selected as the confidence for all methods, where a sweep threshold parameter is used to obtain the curve of all possible accuracy and precision results.

Figure 4a shows the position tracking error in relation to the tracking accuracy for the three methods. The two-step PF approach is slightly better than the two-step algorithm. However, the proposed joint PF approach obtains a notable performance increase in the localization precision with respect to the two-step PF approach, that ranges from 7% to 24% error reduction depending on the A-MOTA working point. This increased localization precision is due to the fact that the database position annotations correspond with the head center position (this is a general fact for almost all tracking databases), whereas the acoustic localization algorithm detects the position of the mouth of the speaker. The proposed joint algorithm takes advantage of the knowledge of the mouth position and head orientation to estimate the center of the head, thus obtaining better localization results. Two A-MOTA working points have been selected to show numerical values, which can be observed in Tables 1 and 2.

Figure 4. Curve of all possible tracking accuracy (acoustic multiple object tracking accuracy (A-MOTA)), localization tracking precision (multiple object tracking precision (MOTP)) (a) and 3D orientation angle precision (multiple head orientation tracking precision (MHOTP)) (b) results, employing a sweep threshold parameter on the algorithm confidence.

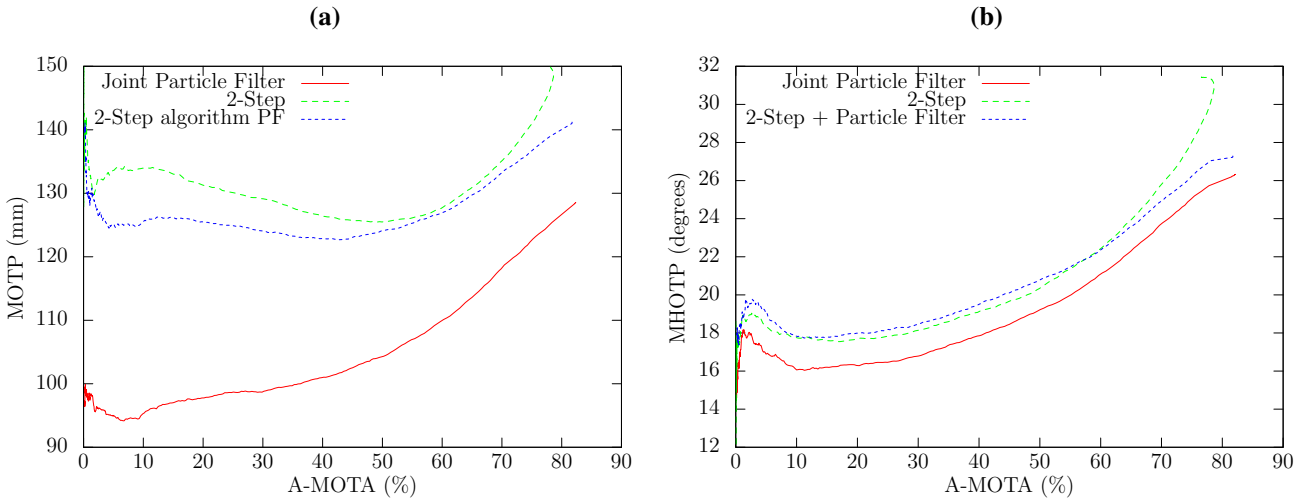


Table 1. Tracking performance joint and two-step approaches for an A-MOTA working point of 10%. PF, particle filter.

System	MOTP	MHOTP	$MHOTP_{\psi}$	$MHOTP_{\theta}$
2-Step	133.94 mm	17.76^{o}	11.27^{o}	10.63^{o}
2-Step PF	125.58 mm	17.84^{o}	11.53^{o}	10.53^{o}
Joint PF	95.30 mm	16.06^{o}	10.38^{o}	9.39^{o}

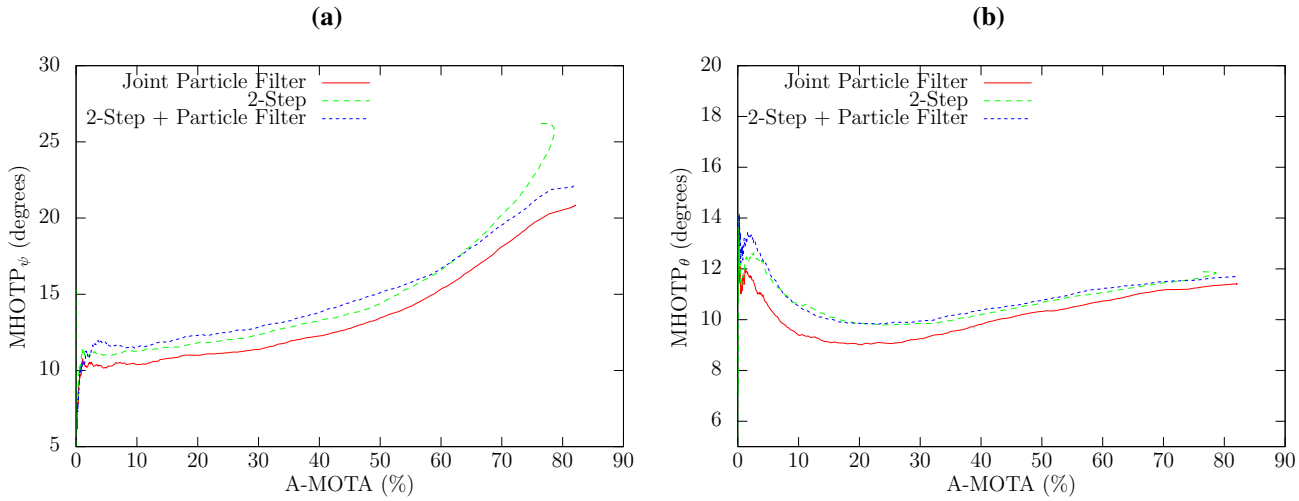
Table 2. Tracking performance joint and two-step approaches for an A-MOTA working point of 75%.

System	MOTP	MHOTP	$MHOTP_{\psi}$	$MHOTP_{\theta}$
2-Step	140.86 mm	28.04^{o}	22.64^{o}	11.54^{o}
2-Step PF	136.62 mm	26.25^{o}	21.01^{o}	11.58^{o}
Joint PF	122.67 mm	25.08^{o}	19.60^{o}	11.30^{o}

The overall precision of the estimation of 3D direction of the head in relation to the tracking accuracy is shown in Figure 4b for all methods. The joint PF approach exhibits an overall reduction of 1.4 degrees in the 3D angle estimation error with respect to the two-step approaches, which have a very similar performance. The 3D angle error can be split in the horizontal and vertical angle error, which are shown in Figure 5, respectively. The proposed joint method has a horizontal angle error reduction of 8.2% to 9.1%, depending on the selected confidence threshold in comparison to both two-step approaches, which, again, have a very similar angle error. Interestingly, the results obtained for the vertical angle, which

are similar to the localization results, have better precision when only high confidence SLF values are employed. This can be explained by the fact that the proposed method estimates the elevation angle from the small term height changes produced by the acoustic localization algorithm and that high confidence SLF values provide a more accurate acoustic source position.

Figure 5. Curve of all possible tracking accuracy (A-MOTA), horizontal orientation angle precision (MHOTP$_\psi$) (a) and vertical orientation angle precision (MHOTP$_\theta$) (b) results, employing a sweep threshold parameter on the algorithm confidence.



8. Conclusions

A PF approach for joint head position and 3D orientation estimation has been presented in this article. Experiments conducted over the purposely recorded database with Euler angles and head center annotations for seven different people in a smart room showed an increased performance for the joint PF approach in relation to two two-step algorithms that first estimate the position and then the orientation of the speaker. Both two-step approaches have a very similar angle estimation error, with a small increase in the localization precision (MOTP) for the two-step PF. The proposed joint algorithm outperforms both two-step algorithms in terms of localization precision and orientation angle precision (MHOTP), assessing the superiority of the joint approach. Furthermore, by means of the definition of a joint dynamical model, part of the the elevation angle of the head is inferred by the algorithm. Future work will be devoted to extending the joint PF to track multiple speakers and to study the fusion with video approaches with a focus on 3D orientation estimation.

Acknowledgments

This work has been partially funded by the Spanish project SARAI (TEC2010-21040-C02-01).

Author Contributions

This work was carried out as part of the research for the Ph.D. dissertation of C. Segura, who actively participated in the study conception, the literature review, data collection and annotation, analysis

and interpretation. J. Hernando participated as thesis advisor in the study conception, analysis and interpretation, and critical revision of manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Brutti, A.; Omologo, M.; Svaizer, P. Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays. *Proc. Interspeech* **2005**, 2337–2340.
2. Segura, C.; Canton-Ferrer, C.; Abad, A.; Casas, J.; Hernando, J. Multimodal Head Orientation towards Attention Tracking in Smartrooms. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), Honolulu, HI, USA, 16–20 April 2007; pp. 681–684.
3. Segura, C.; Abad, A.; Hernando, J.; Nadeu, C. Speaker Orientation Estimation Based on GCC-PHAT and HLBR Hybridation. In Proceedings of International Conference on Spoken Language Processing (ICSLP'08), Brisbane, Australia, 22–26 September 2008; pp. 1325–1328.
4. Levi, A.; Silverman, H. A New Algorithm for the Estimation of Talker Azimuthal Orientation Using a Large Aperture Microphone Array. In Proceedings of 2008 IEEE International Conference on Multimedia and Expo, Hannover, Germany, 23–26 June 2008; pp. 565–568.
5. Sachar, J.; Silverman, H. A Baseline Algorithm for Estimating Talker Orientation Using Acoustical Data from a Large-Aperture Microphone Array. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Quebec, Canada, 17–21 May 2004; pp. 65–68.
6. Levi, A.; Silverman, H. A robust method to extract talker azimuth orientation using a large-aperture microphone array. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 277 –285.
7. Mungamuru, B.; Aarabi, P. Enhanced sound localization. *IEEE Trans Syst. Man Cybern. Part B* **2004**, *34*, 1526–1540.
8. Abad, A.; Segura, C.; Nadeu, C.; Hernando, J. Audio-Based Approaches to Head Orientation Estimation in a Smart-Room. In Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007), Antwerp, Belgium, 27–31 August 2007; pp. 590–593.
9. Abad, A.; Macho, D.; Segura, C.; Hernando, J.; Nadeu, C. Effect of Head Orientation on the Speaker Localization Performance in Smart-Room Environment. In Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH 2005—Eurospeech), Lisbon, Portugal, 4–8 September 2005; pp. 145–148.

10. Bernardin, K.; Elbs, A.; Stiefelhagen, R. MultipleObject Tracking Performance Metrics and Evaluation in a Smart Room Environment. In Proceedings of the Sixth IEEE International Workshop on Visual Surveillance, in conjunction with 9th European Conference on Computer Vision (ECCV 2006), Graz, Austria, 7–13 May 2006.
11. Segura, C. Speaker Localization and Orientation in Multimodal Smart Environments. Ph.D. Thesis, Technical University of Catalonia (UPC), Barcelona, Spain, May **2011**.
12. Svaizer, P.; Matassoni, M.; Omologo, M. Acoustic Source Location in a Three-Dimensional Space Using Crosspower Spectrum Phase. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1997), Munich, Germany, 21–24 April 1997; pp. 231–234.
13. Omologo, M.; Svaizer, P. Acoustic Event Localization Using a Crosspower-Spectrum Phase Based Technique. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1994), Adelaide, Australia, 19–22 April 1994; pp. 273–276.
14. DiBiase, J.; Silverman, H.; Brandstein, M. Microphone Arrays. In *Robust Localization in Reverberant Rooms*; Springer: Berlin, Germany, **2001**.
15. Jain, A.; Nandakumar, K.; Ross, A. Score normalization in multimodal biometric systems. *Pattern recognition* **2005**, *38*, 2270–2285.
16. Arulampalam, M.; Maskell, S.; Gordon, N.; Clapp, T. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.* **2002**, *50*, 174–188.
17. Vermaak, J.; Blake, A. Nonlinear Filtering for Speaker Tracking in Noisy and Reverberant Environments. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001), Salt Lake City, UT, USA, 7–11 May 2001; pp. 3021–3024.
18. Ward, D.; Williamson, R. Particle Filter Beamforming for Acoustic Source Localization in a Reverberant Environment. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002), Orlando, FL, USA, 13–17 May 2002; pp. 1777–1780.
19. Ward, D.; Lehmann, E.; Williamson, R. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Trans. Speech Audio Process.* **2003**, *11*, 826–836.
20. Lehmann, E. Particle Filtering Methods for Acoustic Source Localisation and Tracking. Ph.D. Thesis, Australian National University, Canberra, Australia, July **2004**.
21. Gilks, W.; Richardson, S.; Spiegelhalter, D. *Markov Chain Monte Carlo in Practice*; Chapman and Hall: London, UK, **1996**.
22. Doucet, A.; De Freitas, N.; Gordon, N. *Sequential Monte Carlo Methods in Practice*; Springer Verlag: New York, NY, USA, **2001**.
23. Bergman, N. *Recursive Bayesian Estimation: Navigation and Tracking Applications*; Department of Electrical Engineering, Linköping University: Linköping, Sweden, **1999**.
24. Liu, J.; Chen, R. Sequential Monte Carlo methods for dynamic systems. *J. Am. Stat. Assoc.* **1998**, *93*, 1032–1044.
25. Vermaak, J.; Gangnet, M.; Blake, A.; Perez, P. Sequential Monte Carlo Fusion of Sound and Vision for Speaker Tracking. In Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV 2001), Vancouver, Canada, 7–14 July 2001; pp. 741–746.

26. Pertilä, P.; Korhonen, T.; Visa, A. Measurement combination for acoustic source localization in a room environment. *EURASIP J. Audio Speech Music Process.* **2008**, *2008*, 1–14.
27. CLEAR - Classification of Events, Activities and Relationships Evaluation and Workshop. Available online: http://www.clear-evaluation.org (accessed on 24 January 2014).
28. Waibel, A.; Stiefelhagen, R. *Computers in the Human Interaction Loop*; Springer: Berlin, Germany, **2009**.
29. VACE—Video Analysis and Content Extraction. Available online: http://www.informedia.cs.cmu.edu/arda/vaceII.html (accessed on 24 January 2014).