Acoustic Models and Kalman Filtering Strategies for Active Binaural Sound Localization

Alban Portello, Patrick Danès and Sylvain Argentieri

***Abstract*— This paper deals with binaural sound localization. An active strategy is proposed, relying on a precise model of the dynamic changes induced by motion on the auditive perception. The proposed framework allows motions of both the sound source and the sensor. The resulting stochastic discrete-time model is then exploited together with Unscented Kalman filtering to provide range and azimuth estimation. Simulations and experiments show the effectiveness of the method.**

I. Introduction

Within the large field of robot perception, Robot Audition is recent in comparison to vision or even haptics. The related topics have been widely envisioned for the last ten years inside a new paradigm, including original and specific constraints raised by the robotics context. This has led to various works dealing with speech recognition [1], speaker recognition [2], and/or sound source localization [3][4]. Two main approaches can be cited. On the one hand, array processing methods can be adapted so as to obtain effective and robust auditive systems [5][6]. On the other hand, recent works have been more concerned with binaural audition, based on only two microphones. But using only two microphones to perform the aforementioned auditive tasks is very difficult, and lots of work are now focusing on performances improvement.

Most of the above approaches have only focused on the auditory scene analysis from a static view of the world. Such an idealized situation greatly eases the problem, while it is clear that speech and hearing takes place in a world where none of the static assumptions hold [7]. Yet, active perception, exploiting the possible movement of the auditive system, may help in the environment analysis process and in the auditive tasks robustness. This paper is more concerned with this last topic. The movement of the binaural sensor is fused with the auditive perception to perform sound source localization, in the case when the source and/or the sensor are in motion. Though the static case has been widely studied in the literature, to our opinion the moving case is still open. Nevertheless, recent contributions have already proposed solutions to active auditive perception. In [8], a sound source tracking system integrates audition, vision and motor movements together with an adaptive ego-noise cancelling algorithm. This work has then been extended to active speech recognition in [9]. An active sound source localization scheme is also proposed in [10], where successive perceptions for various positions are exploited to train a Self-Organizing Map. The same approach is used in [11], relying on successive intersections of the cone of confusion. These two last contributions, though active, are fundamentally different from the one proposed in this paper, where the motion and the perception are fused to infer the sound location. This idea has been recently advanced in [12], where a particle filter is used for binaural tracking on the basis of interaural delay and motion parallax. In [13], an Extended Kalman Filter performs source tracking using interaural delay and instantaneous frequencies of perceived signals. Nevertheless, in both papers, the sensor position and velocity expressed in the world frame are assumed known, which implies to endow the sensor with an absolute localization system.

This paper aims at a generic framework to active binaural localization, *i.e.* fusing motion and perception. As modeling issues are often overlooked in the literature, a state space model is first proposed in Section II. The Kalman strategy underlying the estimation process is outlined in Section III. Auditive cues are then discussed in Section IV. The focus is mainly put on Interaural Time Difference (ITD), yet the results can be straightly extended to cope with Interaural Level Difference (ILD). Section V assesses the effectiveness of the proposed approach on realistic simulations and experiments. A conclusion and prospects end the paper.

II. Problem Statement and Objectives

One aim of this paper is to provide a correct mathematical framework to the binaural active localization of a moving sound source in a free-field environment. This section delves further into modeling issues. First, the problem is described in more accurate terms. Then, state space models are discussed, upon which Kalman filtering strategies can apply.

A. Problem statement

A pointwise sound emitter E and a binaural sensor move independently on a common plane parallel to the ground. The two transducers equipping the sensor are denoted by R_l and R_r. A frame $\mathcal{F}_R : (R, \boldsymbol{x_R}, \boldsymbol{y_R}, \boldsymbol{z_R})$ is rigidly linked to the sensor, with R the midpoint of the line segment $[R_l; R_r]$, $\boldsymbol{y_R}$ the vector $\frac{\boldsymbol{R_l R}}{\|\boldsymbol{R_l R}\|}$ and $\boldsymbol{x_R}$ the downward vertical vector. The frame $\mathcal{F}_E : (E, \boldsymbol{x_O}, \boldsymbol{y_O}, \boldsymbol{z_O})$ attached to the source is parallel to the world reference frame $\mathcal{F}_O : (O, \boldsymbol{x_O}, \boldsymbol{y_O}, \boldsymbol{z_O})$, with $\boldsymbol{x_O} = \boldsymbol{x_R}$ (Figure 1). $\|\boldsymbol{R_l R_r}\| = 2a$ terms the transducers interspace. The source undergoes a translational motion

A. Portello and P. Danès are with CNRS; LAAS; 7 avenue du colonel Roche, F-31077 Toulouse Cedex 4, France and Université de Toulouse; UPS, INSA, INP, ISAE ; UT1, UTM, LAAS; F-31077 Toulouse Cedex 4, France alban.portello@laas.fr,patrick.danes@laas.fr

S. Argentieri is with UPMC Univ. Paris 06, 4 place Jussieu, F-75005, Paris, France and ISIR - CNRS UMR 7222, F-75005, Paris, France sylvain.argentieri@upmc.fr

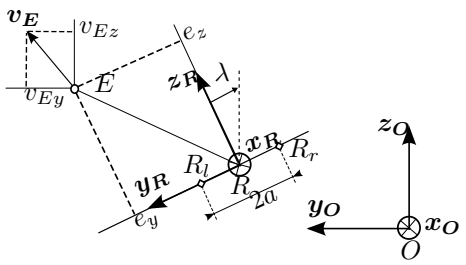


Fig. 1. The considered localization problem.

(velocities v_{Ey}, v_{Ez} of $\mathcal{F}_E$ w.r.t. $\mathcal{F}_O$ expressed along axes $\boldsymbol{y_O}, \boldsymbol{z_O}$), while the sensor is endowed with two translational and one rotational degrees-of-freedom (velocities v_{Ry}, v_{Rz} of $\mathcal{F}_R$ w.r.t. $\mathcal{F}_O$ expressed along axes $\boldsymbol{y_R}, \boldsymbol{z_R}$; rotation velocity ω of $\mathcal{F}_R$ w.r.t. $\mathcal{F}_O$ around $\boldsymbol{x_O} = \boldsymbol{x_R}$). v_{Ry}, v_{Rz}, ω, and, to simplify, v_{Ey}, v_{Ez} are assumed known. The aim is to localize the emitter ($\mathcal{F}_E$) w.r.t. the binaural sensor ($\mathcal{F}_R$) on the basis of the sensed data at R_l, R_r. Free-field condition is assumed. Importantly, the audio sensor is not localized w.r.t. $\mathcal{F}_O$.

B. Mathematical modeling

To tackle binaural active localization through a well-posed filtering problem, a state space model must be defined where the state vector is minimal. The state space equation, describing the way the source and sensor velocities affect the location variables, comes from rigid body kinematics. The output equation relates these spatial variables to acoustic cues extracted from the sensed data. This last topic will be shortly discussed here, then examined in more depth in Section IV.

The relative position and attitude of $\mathcal{F}_E$ w.r.t. $\mathcal{F}_R$ will henceforth be fully parameterized by means of a minimal set of three parameters $e_y \triangleq < \boldsymbol{RE}, \boldsymbol{y_R} >$, $e_z \triangleq < \boldsymbol{RE}, \boldsymbol{z_R} >$ (where $< ., . >$ stands for the scalar product), and $\lambda \triangleq \widehat{(\boldsymbol{z_R}, \boldsymbol{z_O})}_{\boldsymbol{x_O}}$. When the sensor and source velocities are zero-order held at the sampling period T_s, the exact discrete-time state space equation is given by

$$\boldsymbol{x}_{[k+1]} = \boldsymbol{F}\boldsymbol{x}_{[k]} + \boldsymbol{G_1}\boldsymbol{u_1}_{[k]} + \boldsymbol{G_2}(\boldsymbol{x}_{[k]})\boldsymbol{u_2}_{[k]} \quad (1)$$

$$\text{with } \boldsymbol{x}_{[k]} \triangleq \begin{pmatrix} e_{y\,[k]} \\ e_{z\,[k]} \\ \lambda_{[k]} \end{pmatrix},\ \boldsymbol{u_1}_{[k]} \triangleq \begin{pmatrix} v_{Ry\,[k]} \\ v_{Rz\,[k]} \\ \omega_{[k]} \end{pmatrix},\ \boldsymbol{u_2}_{[k]} \triangleq \begin{pmatrix} v_{Ey\,[k]} \\ v_{Ez\,[k]} \end{pmatrix},$$

$$\boldsymbol{F} = \begin{pmatrix} c_{[k]} & s_{[k]} & 0 \\ -s_{[k]} & c_{[k]} & 0 \\ 0 & 0 & 1 \end{pmatrix},\ \boldsymbol{G_1} = \begin{pmatrix} -\frac{s_{[k]}}{T_s} & \frac{c_{[k]}-1}{\omega_{[k]}} & 0 \\ -\frac{c_{[k]}-1}{\omega_{[k]}} & -\frac{s_{[k]}}{T_s} & 0 \\ 0 & 0 & -T_s \end{pmatrix},\ \begin{matrix} c_{[k]}=\cos(\omega_{[k]}T_s), \\ s_{[k]}=\sin(\omega_{[k]}T_s), \end{matrix}$$

$$\boldsymbol{G_2}(\boldsymbol{x}_{[k]}) = T_s \begin{pmatrix} \cos(\lambda_{[k]}-\omega_{[k]}T_s) & -\sin(\lambda_{[k]}-\omega_{[k]}T_s) \\ \sin(\lambda_{[k]}-\omega_{[k]}T_s) & \cos(\lambda_{[k]}-\omega_{[k]}T_s) \\ 0 & 0 \end{pmatrix}.$$

To account for differences w.r.t. this deterministic model—drifts, slips, etc.—the stochastic state equation

$$\boldsymbol{X}_{[k+1]} = \boldsymbol{F}\boldsymbol{X}_{[k]} + \boldsymbol{G_1}\boldsymbol{u_1}_{[k]} + \boldsymbol{G_2}(\boldsymbol{X}_{[k]})\boldsymbol{u_2}_{[k]} + \boldsymbol{W}_{[k]}, \quad (2)$$

is defined by inserting a Gaussian random dynamic noise $\boldsymbol{W}_{[k]}$ with known statistics. Recall that $\boldsymbol{u_1}, \boldsymbol{u_2}$ are given. The measurement vector $\boldsymbol{z}$ is made up with binaural cues, such as Interaural Time Difference (ITD) and Interaural Level Difference (ILD) [14][8]. It is viewed as a sample of the measurement process $\boldsymbol{Z}$, linked to $\boldsymbol{X}$ and the measurement noise process $\boldsymbol{V}$ by an output equation of the form

$$\boldsymbol{Z}_{[k]} = \boldsymbol{h}(\boldsymbol{X}_{[k]}, \boldsymbol{u_1}_{[k]}, \boldsymbol{u_2}_{[k]}) + \boldsymbol{V}_{[k]}. \quad (3)$$

A fundamental problem is often overlooked in the literature: due to the finite speed of sound, *the acoustic signal sensed by a microphone at time t depends on the signal emitted by the source at time $t-\tau$, with τ the time delay due to propagation; τ itself depends on the distance between the source at time $t-\tau$ and the microphone at t*. This fact precludes the writing of any output equation such as (2) when the source and/or the sensor are moving. Consequently, to base the localization strategy on a conventional state space model, as is required by any filtering scheme, simplifications are necessary. This will be the topic of Section IV. Before going into details, extensions to Kalman filtering are first briefly reviewed.

III. Estimation strategies

An original estimation scheme is presented, which can ensure a convenient localization with no prior knowledge.

A. Extended and Unscented Kalman filtering

Consider a stochastic state space model such as (2)–(3). In the linear Gaussian case, the Kalman filter enables the recursive closed-form computation of the first two moments $\hat{x}_{[k|k]}, P_{[k|k]}$ and $\hat{x}_{[k|k-1]}, P_{[k|k-1]}$ of the probability density functions (pdfs) $p(\boldsymbol{X}_{[k]}|\boldsymbol{Z}_{[1:k]} = \boldsymbol{z}_{[1:k]})$ and $p(\boldsymbol{X}_{[k]}|\boldsymbol{Z}_{[1:k-1]} = \boldsymbol{z}_{[1:k-1]})$ for a given sequence of measurements $z_{[1:k]} = z_{[1]}, \ldots, z_{[k]}$. In the nonlinear case, the extended Kalman filter (EKF) propagates over time approximations relying on first-order Taylor expansions. Though widely used, it is often overconfident, *i.e.* it outputs too small covariances. Contrarily, the unscented Kalman filter (UKF) provides approximations of $P_{[k|k]}, P_{[k|k-1]}$ up to an order of accuracy which can be analyzed theoretically, with no additional complexity [15]. For numerical efficiency and stability, the square-root UKF (SRUKF) has been implemented [16].

B. A multiple hypothesis filter

Two difficult problems remain. The first one concerns the characterization of the initial state prior mean and covariance, to be used in the SRUKF initialization. Indeed, an initialization unconsistent with the values of the genuine hidden state can cause filter divergence. A tightly connected second problem may occur even if the SRUKF is well-initialized: the unscented transform of widely spread distributions may lead to overconfident conclusions. This partly comes from the difficulty of tuning the so-called "scaling parameter" of the UT.

To avoid these pitfalls, J independent—noninteracting—SRUKFs $\{F_j\}_{j=1,\ldots,J}$ are started in parallel, each one propagating the posterior moments $\{\hat{x}^j_{[k|k]}, P^j_{[k|k]}\}_{j=1,\ldots,J}$. Their initial conditions $\{\hat{x}^j_{[0|0]}, P^j_{[0|0]}\}_{j=1,\ldots,J}$ are defined from a partition of the admissible state space—*i.e.* of the admissible relative sensor-source locations—into J overlapping cells $\{\mathcal{C}_j\}_{j=1,\ldots,J}$, *e.g.* so that each 99 %

Algorithm 1 The MH-SRUKF.

$\left[\hat{x}_{[k|k]}, P_{[k|k]}, \{P(F_j|z_{1:k})\}_{j=1..J}, \{\hat{x}^j_{[k|k]}, P^j_{[k|k]}\}_{j=1..J}\right] = \text{MH-SRUKF}(z_k, \{P(F_j|z_{1:k-1})\}_{j=1..J}, \{\hat{x}^j_{[k-1|k-1]}, P^j_{[k-1|k-1]}\}_{j=1..J}, \boldsymbol{u_1}_{[k]}, \boldsymbol{u_2}_{[k]})$

IF $k = 0$ **THEN**
  Define the initial conditions $\{\hat{x}^j_{[0|0]}, P^j_{[0|0]}\}_{j=1,\dots,J}$ and the weights $\{W^j_0 \triangleq P(F_j)\}_{j=1,\dots,J}$ of the filters $\{F_j\}_{j=1,\dots,J}$, with $\sum_{j=1}^J P(F_j) = 1$.
END IF
IF $k \geq 1$ **THEN**
  FOR $j = 1, \dots, J$ **DO**
    Predict the moments $\hat{x}^j_{[k|k-1]}, P^j_{[k|k-1]}$ from $\hat{x}^j_{[k-1|k-1]}, P^j_{[k-1|k-1]}$ through F_j according to (2) (SRUKF Time Update).
    Inside F_j, on the basis of (3) fuse $\boldsymbol{z}_{[k]}$ with the moments $\hat{x}^j_{[k|k-1]}, P^j_{[k|k-1]}$ so as to get $\hat{x}^j_{[k|k]}, P^j_{[k|k]}$ (SRUKF Measurement Update).
    (Filter Likelihood) Apply the UT on $\mathcal{N}(\boldsymbol{X}_{[k]}; \hat{x}^j_{[k|k-1]}, P^j_{[k|k-1]})$ through (3), and get the predicted mean and covariance $\hat{z}^j_{[k|k-1]}, S^j_{[k|k-1]}$ of the output. Then, set $p(\boldsymbol{Z}_{[k]} = \boldsymbol{z}_{[k]}|F_j, \boldsymbol{Z}_{[1:k-1]} = \boldsymbol{z}_{[1:k-1]}) = \frac{1}{\sqrt{(2\pi)^{n_{\boldsymbol{x}}} \det(S^j_{[k|k-1]})}} \exp\left(-\frac{1}{2}(\boldsymbol{z}_{[k]} - \hat{z}^j_{[k|k-1]})^T (S^j_{[k|k-1]})^{-1} (\boldsymbol{z}_{[k]} - \hat{z}^j_{[k|k-1]})\right)$, with $n_{\boldsymbol{x}} = \dim(\boldsymbol{x} \text{ or } \boldsymbol{X}) = 3$ after (1)–(2).
  END FOR
  FOR $j = 1, \dots, J$ **DO**
    (Filters posterior probabilities) Update $W^j_k \triangleq P(F_j|\boldsymbol{Z}_{[1:k]} = \boldsymbol{z}_{[1:k]})$ of F_j according to (5).
  END FOR
  (Filters collapsing) If some W^j_k are lesser than a given threshold γ, then suppress the corresponding filters F_j. Decrease J and renormalize all the filters posterior probabilities accordingly.
  (Output processing) From the set of active filters, compute the overall posterior mean and covariance according to (7).
END IF

probability ellipsoid defined from the Gaussian prior $p(\boldsymbol{X}_{[0]}|F_j) = \mathcal{N}(\boldsymbol{X}_0; \hat{x}^j_{[0|0]}, P^j_{[0|0]})$ covers $\mathcal{C}_j$. Each filter F_j is assigned a given initial probability $W^j_0 \triangleq P(F_j)$. So, the initial state prior pdf is described by the Gaussian mixture

$$p(\boldsymbol{X}_{[0]}) = \sum_{j=1}^J P(F_j)\mathcal{N}(\boldsymbol{X}_{[0]}; \hat{x}^j_{[0|0]}, P^j_{[0|0]}). \quad (4)$$

At each time k, after the measurement update, the posterior probability —or "weight"— $W^j_{[k]} \triangleq P(F_j|\boldsymbol{Z}_{[1:k]} = \boldsymbol{z}_{[1:k]})$ of each filter F_j is computed from its likelihood $p(\boldsymbol{Z}_{[k]} = \boldsymbol{z}_{[k]}|F_j, \boldsymbol{Z}_{[1:k-1]} = \boldsymbol{z}_{[1:k-1]})$ w.r.t. the measurement $\boldsymbol{z}_{[k]}$ and from the weights at time $k-1$ $\{P(F_j|\boldsymbol{Z}_{[1:k-1]} = \boldsymbol{z}_{[1:k-1]})\}_{j=1,\dots,J}$ through (see [17])

$$\begin{aligned} W^j_k &\triangleq P(F_j|\boldsymbol{Z}_{[1:k]} = \boldsymbol{z}_{[1:k]}) \quad (5)\\ &= \frac{p(\boldsymbol{Z}_{[k]}=\boldsymbol{z}_{[k]}|F_j, \boldsymbol{Z}_{[1:k-1]}=\boldsymbol{z}_{[1:k-1]})P(F_j|\boldsymbol{Z}_{[1:k-1]}=\boldsymbol{z}_{[1:k-1]})}{\sum_{l=1}^J p(\boldsymbol{Z}_{[k]}=\boldsymbol{z}_{[k]}|F_l, \boldsymbol{Z}_{[1:k-1]}=\boldsymbol{z}_{[1:k-1]})P(F_l|\boldsymbol{Z}_{[1:k-1]}=\boldsymbol{z}_{[1:k-1]})}\\ &= \frac{p(\boldsymbol{Z}_k=\boldsymbol{z}_k|F_j, \boldsymbol{Z}_{[1:k]}=\boldsymbol{z}_{[1:k]})W^j_{[k-1]}}{\sum_{l=1}^J p(\boldsymbol{Z}_k=\boldsymbol{z}_k|F_l, \boldsymbol{Z}_{[1:k]}=\boldsymbol{z}_{[1:k]})W^l_{[k-1]}}. \end{aligned}$$

Then, those filters whose weights fall below a given threshold γ (*e.g.* $\gamma = 0.01$) collapse, and the weights of the remaining ones are renormalized. The posterior pdf then writes as

$$\begin{aligned} p(\boldsymbol{X}_{[k]}|\boldsymbol{Z}_{[1:k]}=\boldsymbol{z}_{[1:k]}) &= \sum_{j=1}^J P(F_j|k)p(\boldsymbol{X}_{[k]}|F_j, k) \quad (6)\\ &= \sum_{j=1}^J P(F_j|k)\mathcal{N}(\boldsymbol{X}_{[k]}; \hat{x}^j_{[k|k]}, P^j_{[k|k]}), \end{aligned}$$

where ${}^P\!/_p(.|k)$ are shortcuts for ${}^P\!/_p(.|\boldsymbol{Z}_{[1:k]} = \boldsymbol{z}_{[1:k]})$. The overall posterior mean and covariance follow:

$$\begin{aligned} \hat{x}_{[k|k]} &= \sum_{j=1}^J W^j_{[k]}\hat{x}^j_{k|k} \quad (7)\\ P_{[k|k]} &= \sum_{j=1}^J W^j_{[k]}\left(P^{(j)}_{k|k} + (\hat{x}^j_{[k|k]} - \hat{x}_{[k|k]})(\hat{x}^j_{[k|k]} - \hat{x}_{[k|k]})^T\right). \end{aligned}$$

The consequent Multiple Hypothesis SRUKF (MH-SRUKF) is summarized in Algorithm 1.

IV. Acoustic Cues

A. Assumptions and General Equations

The air medium is assumed linear, and free of reflectors or scatterers. The signal $s_T(t)$ received at time t by a transducer $T \in \{R_l, R_r\}$ is a propagated and attenuated transform of the signal $s_E(t)$ emitted by E according to

$$s_T(t) = \frac{1}{c\tau(t)} s_E(t - \tau(t)), \quad (8)$$

with $c = 340\text{m.s}^{-1}$ the speed of sound and

$$\tau(t) = \frac{1}{c}\|\boldsymbol{OT}(t) - \boldsymbol{OE}(t - \tau(t))\| \quad (9)$$

the propagation delay from E to T. Define $\boldsymbol{v_E}$ and $\boldsymbol{v_T}$ as the velocity vectors of E and T w.r.t. frame $\mathcal{F}_O$. Deriving $(c\tau(t))^2$ w.r.t. time leads to, after some manipulations,

$$\dot{\tau}(t) = \frac{\left\langle \left(\boldsymbol{v_E}(t-\tau(t)) - \boldsymbol{v_T}(t)\right), \boldsymbol{n_{T(t)\to E(t-\tau)}}\right\rangle}{c + \left\langle \boldsymbol{v_E}(t-\tau(t)), \boldsymbol{n_{T(t)\to E(t-\tau)}}\right\rangle} \quad (10)$$

with

$$\boldsymbol{n_{T(t)\to E(t-\tau)}} \triangleq \frac{\boldsymbol{OE}(t-\tau(t)) - \boldsymbol{OT}(t)}{\|\boldsymbol{OE}(t-\tau(t)) - \boldsymbol{OT}(t)\|}. \quad (11)$$

The celebrated formula of the Doppler shift straightly follows from (10). Indeed, defining the instantaneous frequency f of a signal as 2π times the derivative of its phase, one gets

$$f_T(t) = f_E(t - \tau(t))\left(\frac{c + \left\langle \boldsymbol{v_T}(t), \boldsymbol{n_{T(t)\to E(t-\tau)}}\right\rangle}{c + \left\langle \boldsymbol{v_E}(t-\tau(t)), \boldsymbol{n_{T(t)\to E(t-\tau)}}\right\rangle}\right). \quad (12)$$

Eqs. (8)–(9)–(10)–(12) are graphically represented on Figure 2. As they link the characteristics of the emitted and received signals to their spatial positions and velocities, they are important for sound source localization. Nevertheless, as outlined before, they cannot be used as they are in the output equation of the state space model for they do not express a memoryless mapping between the output and the state/control vectors. The forthcoming section discusses some approximations.

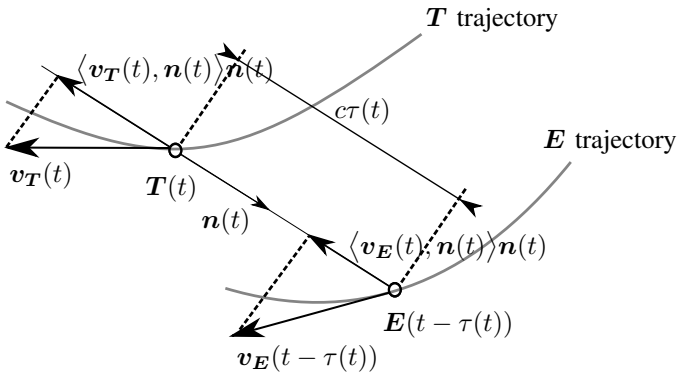


Fig. 2. Geometry of sound propagation

B. Approximations

Developing the square of (9) and using the relationship

$$\boldsymbol{OE}(t-\tau(t)) = \boldsymbol{OE}(t) - \int_{t-\tau(t)}^{t} \boldsymbol{v_E}(u)\,\mathrm{d}u \tag{13}$$

leads to the equation

$$\begin{aligned} c^2\tau^2(t) =& \|\boldsymbol{OT}(t)\|^2 + \|\boldsymbol{OE}(t)\|^2 - 2\big\langle \boldsymbol{OT}(t), \boldsymbol{OE}(t)\big\rangle \\ &+ \| \textstyle\int_{t-\tau(t)}^{t} \boldsymbol{v_E}(u)\,\mathrm{d}u\|^2 \\ &+ 2\big\langle (\boldsymbol{OT}(t) - \boldsymbol{OE}(t)), \textstyle\int_{t-\tau(t)}^{t} \boldsymbol{v_E}(u)\,\mathrm{d}u\big\rangle, \end{aligned} \tag{14}$$

which can be approximated in various ways.

1) Zero-th order approximation: For a slowly moving source, considering that the vector $\boldsymbol{OE}$ is constant during the time delay τ in (14) leads to

$$c\tau(t) \approx \sqrt{\|\boldsymbol{OT}(t)\|^2 + \|\boldsymbol{OE}(t)\|^2 - 2\big\langle \boldsymbol{OT}(t), \boldsymbol{OE}(t)\big\rangle}. \tag{15}$$

2) First order approximation: If the emitter velocity is not small but constant during the time delay τ, then taking $\boldsymbol{v_E}(t) \equiv \boldsymbol{v_E}$ out of integral terms in (14) results in

$$\begin{aligned} &\alpha\tau^2 + \beta\tau(t) + \gamma \approx 0 \\ &\text{with} \quad \alpha = \|\boldsymbol{v_E}(t)\|^2 - c^2 \\ &\qquad \beta = 2\big\langle (\boldsymbol{OT}(t) - \boldsymbol{OE}(t)), \boldsymbol{v_E}(t)\big\rangle \\ &\qquad \gamma = \|\boldsymbol{OT}(t)\|^2 + \|\boldsymbol{OE}(t)\|^2 - 2\big\langle \boldsymbol{OT}(t), \boldsymbol{OE}(t)\big\rangle. \end{aligned} \tag{16}$$

As only subsonic motions are considered, (16) always admits a single positive solution for τ, which depends both on positions and velocities.

C. Measurement equation

Recall that two transducers R_l, R_r are used, and that the source signal is assumed unknown. The theoretical binaural Interaural Time Difference (ITD) then comes as

$$\mathrm{ITD}(t) = \tau_l(t) - \tau_r(t), \tag{17}$$

with $\tau_l(t), \tau_r(t)$ the propagation delays from the source to R_l, R_r, respectively. Define $T \in \{R_l, R_r\}$, $\varepsilon_{R_l} = -1$, $\varepsilon_{R_r} = +1$. From (15)–(16) and the additional relations

$$\begin{pmatrix} <\boldsymbol{OT}, \boldsymbol{y_O}> \\ <\boldsymbol{OT}, \boldsymbol{z_O}> \end{pmatrix} = \begin{pmatrix} <\boldsymbol{OR}, \boldsymbol{y_O}> \\ <\boldsymbol{OR}, \boldsymbol{z_O}> \end{pmatrix} - \varepsilon_T\, a \begin{pmatrix} cos\lambda \\ -sin\lambda \end{pmatrix}, \tag{18}$$

$$\begin{pmatrix} e_y \\ e_z \end{pmatrix} = \begin{pmatrix} cos\lambda & -sin\lambda \\ sin\lambda & cos\lambda \end{pmatrix} \begin{pmatrix} <\boldsymbol{OE}-\boldsymbol{OT}, \boldsymbol{y_O}> \\ <\boldsymbol{OE}-\boldsymbol{OT}, \boldsymbol{z_O}> \end{pmatrix}, \tag{19}$$

approximate static equations linking the time delays with the location variables can be obtained. On this basis, an ordinary—static—output equation can be built for ITD cues. This equation will then be part of the state space model underlying the filtering based localization. The approximate output equations are as follows:

1) Zero-th order approximation:

$$c\tau_{l,r} = \sqrt{a^2 + e_y^2(t) + e_z^2 + \varepsilon_{R_{l,r}} 2ae_y}. \tag{20}$$

2) First order approximation:

$$\begin{aligned} &\tau_{l,r} \text{ is the positive root of } \alpha\tau_{l,r}^2 + \beta\tau_{l,r} + \gamma \\ &\quad \text{with } \alpha = v_{Ey}^2 + v_{Ez}^2 - c^2 \\ &\quad \beta = 2\left\langle \left(-\begin{pmatrix} cos\lambda & sin\lambda \\ -sin\lambda & cos\lambda \end{pmatrix} \begin{pmatrix} e_y \\ e_z \end{pmatrix} - \varepsilon_{R_{l,r}} a \begin{pmatrix} cos\lambda \\ -sin\lambda \end{pmatrix}\right), \begin{pmatrix} v_{Ey} \\ v_{Ez} \end{pmatrix} \right\rangle \\ &\quad \gamma = a^2 + e_y^2 + e_z^2 + \varepsilon_{R_{l,r}} 2ae_y. \end{aligned} \tag{21}$$

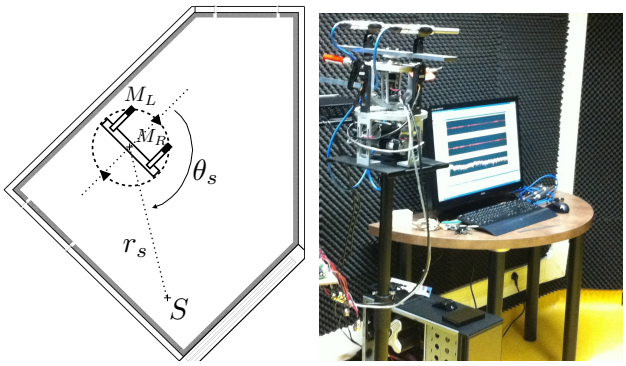
D. Extraction of the ITD from the raw signals

Despite an output equation can now be defined, the way the measurements—viz. the binaural cues—can be elaborated still need to be discussed. This process is based on the estimation of the temporal shift between the raw signals sensed by the binaural sensor while it is moving. One approach to measure a constant time delay between the signals $s_{R_l}(t)$ and $s_{R_r}(t)$, provided they are individually and jointly stationary, consists in extracting the argmax of their cross-correlation $C_{lr}(u) = \mathbb{E}[s_{R_l}(t) s_{R_r}(t-u)]$. From the data gathered on a finite time window of length T during a single experiment, an estimate of C_{lr} can be computed as follows:

$$\hat{C}_{lr}(u) = \tfrac{1}{T} \textstyle\int_{\tau}^{T} s_{R_l}(t) s_{R_r}(t-u)\mathrm{d}t. \tag{22}$$

However, a number of potential issues underlying the position of the peak in $\hat{C}_{lr}$ must be taken into account. First, the time delay to be estimated must vary slowly within the observation time T. This is a fundamental concern when the source and sensor move. Next, the finiteness of the time window may result in a wide and inaccurate peak of $\hat{C}_{lr}$, *e.g.* when the measurement noise is important or if the source is narrowband. As a solution, [18] suggests to use the so-called "Generalized Cross Correlation" (GCC), which consists in weighting the signals Cross Power Spectral Density. Among all the frequency-weighting functions ψ proposed in [18], the Phase Transform (PHAT) processor may be the most popular. An enhancement, called Reliability-Weighted Phase Transform (RW-PHAT) consists in reducing the weighting function at frequencies where the signal-to-noise ratio is low [19]. One can also cite the Maximum Likelihood (ML) processor [18], which exhibits an asymptotic Gaussianity, making possible to consider the measurement noise in (3) as Gaussian.

In all the following, the considered sound source will emit a Gaussian white noise so as to simplify the auditory cues estimation problem. Consequently, only the digital version of (22) will be considered. But it is known that the sampling frequency f_s affects the temporal shift estimator statistics, introducing quantization errors in the estimation,



(a) (b)

Fig. 3. (a) Experimental setup scheme. The two left (M_L) and right (M_R) omnidirectional microphones are circularly moved by a motoreductor. The loudspeaker (S) emits a white noise during the movement. (b) Experimental setup picture, showing the two microphones and the acquisition computer.

as noninteger multiples of sampling period delays cannot be reached. This problem can be partially solved by the simple approximation of fitting a parabola in the neighborhood of the cross correlation peak [20].

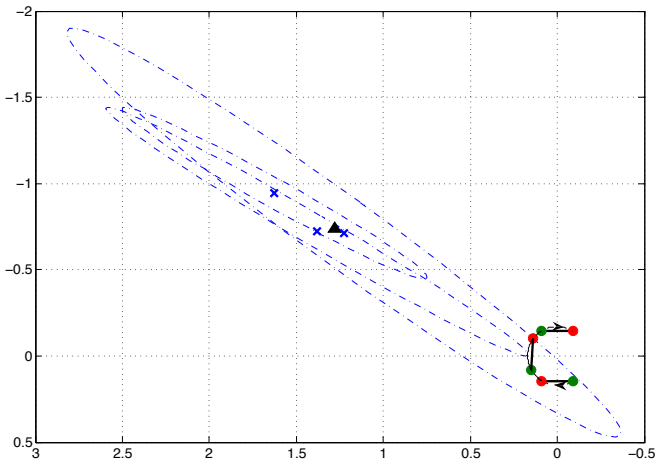
V. Case Study

In this section, experimental and simulation results are shown to confirm the proposed approach. In simulations as in the experiment, the position of a static pointwise source is estimated w.r.t a moving binaural sensor. The results reveal that the proposed filtering strategy, by using information coming from relative motion, overcomes front-back ambiguity problems inherent to ITD, and can provide range estimation despite ITD does not carry information about distance.

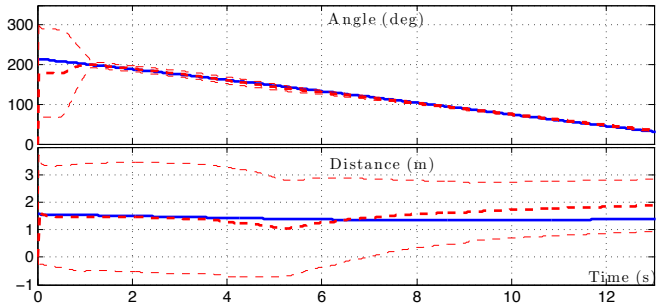
A. Experimental results

1) Experimental setup: In order to assess the proposed approach with real binaural signals, several experiments have been performed in an acoustically prepared room, equipped with 3D pyramidal pattern studio foams placed on the roof and on the walls. Two omnidirectional microphones, spaced by 18cm, are then placed on the top of a simple mechanical system made of a single rotation moved by a motoreductor. The angular axis position is recorded during the movement with a $0.5°$-precision, producing the proprioception of the platform. The two microphone outputs are simultaneously acquired by a National Instruments PCI acquisition card through 24 bits delta-sigma converters operating at the sampling frequency $f_s = 44.1$kHz. A loudspeaker, diffusing a white gaussian noise, is placed in front of the mechanical system at the position $(r_s, \theta_s) = (1.48\text{m}, 120°)$ (see Figure 3). The recording of the proprioception and of the auditive perception is simultaneoulsy started at the begining of the movement, which consists in a rotation performed at successive various constant angular velocities ω.

2) Sound source estimation results: Figure 4b shows the source location estimate (azimuth and distance) as well as the 99% confidence interval as a function of time in the sensor frame. The azimuth estimate exhibits good performance, reaching its steady state in about two seconds, with a $3°$ standard deviation. The same does not apply for the distance estimate, which shows a 1m standard deviation. This large uncertainty in distance is depicted in Figure 4a, where very flat confidence ellipses (dotted lines) are reported, extending towards the source (triangle)-sensor (line segment) direction. In fact, some filters of the MH-SRUKF which were initialized with an azimuth close to the real source azimuth but with a different range might produce a likely predicted output w.r.t the measurement, because, again, ITD does not provide much information about range. If the type of sensor motion does not bring sufficient additional information, the likelihood weights of these filters will remain above the minimum threshold, they will not collapse, and the global estimate, which is a combination of all active filters estimates, will remain inaccurate in distance. Thus, the motion considered in this experiment is not suited to segregate all inconsistent estimates w.r.t range. Because of limitations of the mechanical system, which is only able to perform a simple rotation of the interaural axis, a realistic simulation has been performed with a different type of motion, in order to bring to the fore the sensor motion influence.

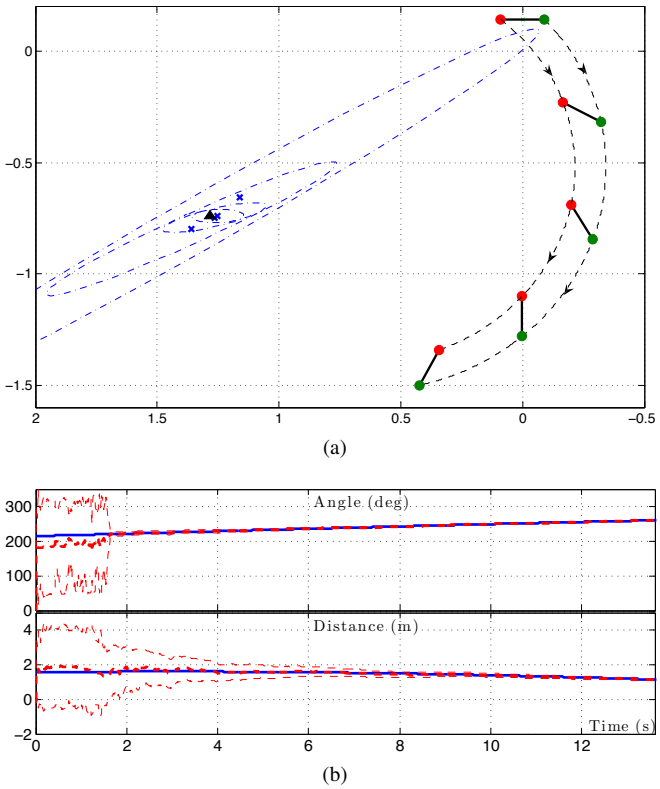


(a)



(b)

Fig. 4. Experimental results. (a) Sensor motion and emitter position estimate in the world frame (b) Range and azimuth estimates in the sensor frame as a function of time.



(a)

(b)

Fig. 5. Simulation : the sensor follows a circular trajectory. (a) Sensor motion and emitter position estimate in the world frame (b) Range and azimuth estimates in the sensor frame as a function of time.

B. Simulation results

In the simulation reported on Figure 5a, the sensor center follows a circular trajectory, with a constant interaural axis velocity. This motion enables a much more accurate range estimation, as shown in Figure 5b. Indeed, the confidence ellipsoid shrinks along all directions during the estimation. An auditive feedback control of the sensor motion could be designed in order to improve source localization accuracy.

VI. Conclusion

A new active approach to binaural sound source localization has been proposed. It relies on an accurate modeling of the acoustic cues, providing ground credible equations for Kalman filtering applications. Theoretical developments have been supported by simulation and experiments.

Although an accurate model for ITD has been put forward in this paper, some limitations appear in the ITD extraction when the source-sensor relative velocity becomes important. In fact, high time delay variations induced by fast relative motions decrease the accuracy of classical cross correlation methods, which have been mainly designed for slowly varying ITDs [18]. An interesting future work could consist in using an enhanced cross correlation method accounting for source or sensor motion, as described in [21].

Another improvement will study the coupling of the proposed filtering scheme with detections of wrong ITDs/ILDs values computed during source silences (*e.g.* based on residual monitoring or hypotheses testing). Various re-initialization strategies will then be evaluated.

Other prospective issues would consider a more sophisticated binaural system including a head, and account for room reverberation. Acoustic considerations then would have to be revisited accordingly. Finally, MUSIC-based super-resolution time delay estimation methods developed in [22] would be a possible alternative to Generalized Cross Correlation, exhibiting a higher accuracy in a reverberant or multi-source context.

Acknowledgment

This work was conducted within the BINAAHR (BINaural Active Audition for Humanoid Robots) project funded by ANR (France) and JST (Japan) under Contract n°ANR-09-BLAN-0370-02. Detailed analytical computations are not included for space reasons, but can be obtained on request.

References

[1] T. Takahashi, K. Nakadai, K. Komatani, T. Ogata, and H. G. Okuno, "An improvement in automatic speech recognition using soft missing feature masks for robot audition," *IEEE/RSJ IROS'2010*, pp. 964–969.

[2] K. Youssef, S. Argentieri, and J.-L. Zarader, "From monaural to binaural speaker recognition for humanoid robots," in *IEEE Humanoids'2010*, pp. 580–586.

[3] T. Rodemann, "A study on distance estimation in binaural sound localization," in *IEEE/RSJ IROS'2010*, pp. 425–430.

[4] F. Keyrouz and K. Diepold, "An enhanced binaural 3D sound localization algorithm," in *IEEE Symp. on Sig. Proc. and Info. Tech. 2006*.

[5] J.M. Valin, F. Michaud, J. Rouat and D. Létourneau, *Robust sound source localization using a microphone array on a mobile robot*, in *IEEE/RSJ IROS'2003*, pp. 1228–1233.

[6] H. L. Van Trees, *Optimum Array Processing*, ser. Detection, Estimation, and Modulation Theory. John Wiley & Sons, Inc., 2002, vol. IV.

[7] M. Cooke, Y.-C. Lu, Y. Lu, and R. P. Horaud, "Active hearing, active speaking," in *Int. Symp. on Auditory and Audiological Res.*, 2007.

[8] K. Nakadai, T. Lourens, H. Okuno, and H. Kitano, "Active audition for humanoid," in *17th Nat. Conf. on Artificial Intelligence*, 2000.

[9] K. Nakadai, H. G. Okuno, and H. Kitano, "Robot recognizes three simultaneous speech by active audition," in *IEEE ICRA'2003*.

[10] E. Berglund and J. Sitte, "Sound source localisation through active audition," in *IEEE/RSJ IROS'2005*, pp. 509–514.

[11] L. Kneip and C. Baumann, "Binaural model for artificial spatial sound localization based on interaural time delays and movements of the interaural axis," *Jour. Acoust. Soc. America*, 124(5), pp. 3108–19, 2008.

[12] Y.-C. Lu and M. Cooke, "Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners," *Speech Comm.*, 2010.

[13] M. Kumon and S. Uozumi, "Binaural localization for a mobile sound source," *Jour. of Biomechanical Science and Engineering*, 6(1), 2011.

[14] H. Viste and G. Evangelista, "Binaural source localization," in *Int. Conf. on Digital Audio Effects (DAFx'04)*.

[15] S. Julier and J. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of IEEE*, 92(3), 2004.

[16] R. Van der Merwe and E. Wan, "The square-root unscented kalman filter for state and parameter estimation," in *IEEE ICASSP'01*.

[17] Y. Bar-Shalom and X. Li, *Estimation and Tracking : Principles, Techniques and Software*. YBS editions, Norwood, 1993.

[18] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. ASSP*, 24(4), pp. 320–327.

[19] K. Nakadai, H. Okuno, and H. Kitano, "Auditory fovea based speech separation and its application to dialog system," *IEEE/RSJ IROS'2002*.

[20] R. E. Boucher and J. C. Hassab, "Analysis of discrete implementation of generalized cross correlator," *IEEE Trans. ASSP*, 29(3).

[21] C. Knapp and G. Carter, "Time delay estimation in the presence of relative motion," *IEEE Trans. ASSP*, pp. 280–283, 1977.

[22] F. Ge, D. Shen, Y. Peng, and V. Li, "Super-resolution time delay estimation in multipath environments," *IEEE Wireless Communications and Networking Conf.*, pp. 1121–1126, 2004.