Active localization of an intermittent sound source from a moving binaural sensor

Alban Portello, Gabriel Bustamante, Patrick Danès, Jonathan Piat, Jérôme Manhes

7 - 12 SEPTEMBER 2014, KRAKÓW
FORUM ACUSTICUM

Active localization of an intermittent sound source from a moving binaural sensor

Alban Portello, Gabriel Bustamante, Patrick Danès, Jonathan Piat
CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France.
Univ. de Toulouse, UPS, LAAS, F-31400 Toulouse, France.

Jérôme Manhès
CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France.

Summary
An unknown intermittent sound source, at rest or in motion, is considered. We describe an original method to detect its activity and determine its azimuth and range from a moving binaural sensor subject to measurement noise. The proposed source localization scheme is said "active", as it combines the sensed binaural signals with the sensor motion. It is composed of two stages: first, a pseudo-likelihood of the source azimuth is defined from the short-term time-frequency analysis of the binaural signals, and the source activity is detected by means of statistical identification; then, this information brought by the binaural signals is assimilated over time and fused with the sensor motor commands into a stochastic filtering strategy entailing a bank of noninteractive unscented Kalman filters. The method enjoys several important features. On the one hand, it is endowed with self-initialization and ensures the consistency of the covariances of the estimation errors. On the other hand, it can rigorously handle the effects induced by scatterers by explicitly exploiting the HRTFs to the microphones. A validation on simulated scenarios as well as on data coming from real experiments is included. This work has been partially supported by the EU FP7 FET-Open TWO!EARS project (2014–2016), whose goal is to develop a computational model of auditory perception and experience, in which binaural bottom up processing is interwoven with action and with top-down feedbacks originating from cortical levels. It takes place at the sensorimotor (reflex) level of this architecture. Current work concerns its extension to the multiple source case and to the definition of active motions which can improve perception.

PACS no. 43.60.Jn, 43.66.Pn

1. Introduction

Computational Auditory Scene Analysis (CASA) aims at extracting perceptual representations of sound sources from recordings of an acoustic scene [1]. Binaural sound localization is a key function of CASA systems. In this paper, an original binaural sound localization scheme is presented. It follows on from [2]–[3], but contrarily to these references, it explicitly allows and accounts for the presence of a head between the two microphones. Extraction of short-term spatial information is first performed along [4], which builds on [5]–[6]. Ulike several contributions to binaural localization in CASA (see [1] chapter 5 for an overview), our azimuth detection scheme is not based on physiology of auditory perception in humans (no cochlear processing is used), but is rather stated as an estimation problem. This binaural information is then fused with the motor commands of the sensor within a stochastic filtering framework. The joint exploitation of auditive and proprioceptive modalities (*i.e.* active audition) through filtering strategies has already been put forward in [7]. Nevertheless, our strategy differs in many aspects: modeling, spatial information extraction, Source Activity Detection (SAD), motor commands, filtering strategy, etc.

The paper is organized as follows. Section 2 is dedicated to the problem statement and its mathematical formulation. In §3, short-term spatial information extraction and source activity detection are presented. The way how this information is fused with the sensor motor commands is detailed in §4. The strategy is qualitatively evaluated in simulation in §5, while §6 shows some preliminary experimental results. A conclusion ends the paper.

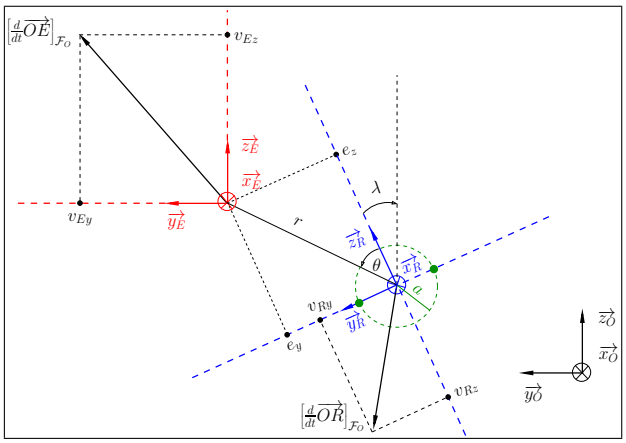

Figure 1. Geometrical representation of the considered planar problem.

2. Problem statement and mathematical formulation

2.1. 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_1 and R_2. A frame $\mathcal{F}_R : (R, \overrightarrow{x_R}, \overrightarrow{y_R}, \overrightarrow{z_R})$ is rigidly linked to the sensor, with R the midpoint of the line segment $[R_1; R_2]$, $\overrightarrow{y_R}$ the vector $\frac{\overrightarrow{R_2R_1}}{\|\overrightarrow{R_2R_1}\|}$ and $\overrightarrow{x_R}$ the downward vertical vector. The frame $\mathcal{F}_E : (E, \overrightarrow{x_O}, \overrightarrow{y_O}, \overrightarrow{z_O})$ attached to the source is parallel to the world reference frame $\mathcal{F}_O : (O, \overrightarrow{x_O}, \overrightarrow{y_O}, \overrightarrow{z_O})$, with $\overrightarrow{x_O} = \overrightarrow{x_R}$ (see Figure 1).

The source undergoes a translational motion (velocities v_{Ey}, v_{Ez} of $\mathcal{F}_E$ w.r.t. $\mathcal{F}_O$ expressed along axes $\overrightarrow{y_O}, \overrightarrow{z_O}$), 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 $\overrightarrow{y_R}, \overrightarrow{z_R}$; rotation velocity ω of $\mathcal{F}_R$ w.r.t. $\mathcal{F}_O$ around $\overrightarrow{x_O} = \overrightarrow{x_R}$). Assuming that v_{Ry}, v_{Rz}, ω are known, the aim is to localize the emitter w.r.t. the binaural sensor on the basis of the sensed data at R_1, R_2 and some prior knowledge on the source motion. Importantly, the audio sensor is not localized w.r.t. $\mathcal{F}_O$.

The source localization is performed in two steps: (1) spatial information extraction and Source Activity Detection (SAD) from short-term analysis of the binaural stream; (2) assimilation of this information and fusion with the sensor motor commands inside a stochastic filtering scheme. These two steps rely on a mathematical formulation of the problem, which is the topic of §2.2 and §2.3. Foundations of stochastic filtering are then presented in §2.4.

2.2. Kinematic model and state vector

The state space equation, which must account for the way how the source and sensor velocities affect the location variables, is obtained from rigid body kinematics. Denoting $e_y \triangleq \overrightarrow{RE}.\overrightarrow{y_R}$ and $e_z \triangleq \overrightarrow{RE}.\overrightarrow{z_R}$ as the cartesian coordinates of the emitter E in the sensor frame $\mathcal{F}_R$ and $\lambda \triangleq (\overrightarrow{z_R}, \overrightarrow{z_O})_{\overrightarrow{x_O}}$ as the angle between vectors $\overrightarrow{z_R}$ and $\overrightarrow{z_O}$ around $\overrightarrow{x_O}$, a continuous time state-space equation comes as [8]

$$\begin{bmatrix} \dot{e}_y(t) \\ \dot{e}_z(t) \\ \dot{\lambda}(t) \end{bmatrix} = \begin{bmatrix} -1 & 0 & e_z(t) & \cos\lambda(t) & -\sin\lambda(t) \\ 0 & -1 & -e_y(t) & \sin\lambda(t) & \cos\lambda(t) \\ 0 & 0 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_{Ry}(t) \\ v_{Rz}(t) \\ \omega(t) \\ v_{Ey}(t) \\ v_{Ez}(t) \end{bmatrix}. \quad (1)$$

When the sensor and source velocities are zero-order held at the sampling period T_s, equation (1) can be turned into an exact discrete-time state space equation of the form

$$x_k = f(x_{k-1}, u_{1k-1}, u_{2k-1}), \quad (2)$$

with f given in appendix 8, $x_k \triangleq [e_{yk}, e_{zk}, \lambda_k]'$, $u_{1k} \triangleq [v_{Ryk}, v_{Rzk}, \omega_k]'$, $u_{2k} \triangleq [v_{Eyk}, v_{Ezk}]'$, and, for any signal x, $x_k \triangleq x(kT_s)$. To account for uncertainty in the sensor motion, a random noise $\mathtt{w}_{k-1}$ is added to the right-hand side of (2), x_k and u_{2k} are turned into the random vectors $\mathtt{x}_k$ and $\mathtt{u}_{2k}$, while u_{1k} is still assumed deterministic and known.

Since $\mathtt{u}_{2k}$ is unknown, a secondary dynamic system must be introduced to characterize its time evolution. In the scope of this paper, the source is supposed to follow a Gaussian random walk, *i.e.*, $\mathtt{u}_{20:k} \triangleq \mathtt{u}_{20}, \ldots, \mathtt{u}_{2k}$ is drawn from a zero-mean Gaussian distribution, with, for all i, j, $\mathbb{E}\{\mathtt{u}_{2i}\mathtt{u}_{2j}'\} = Q_{\mathtt{u}_2}\delta_{ij}$, δ_{ij} the Kronecker symbol, and $Q_{\mathtt{u}_2} = \sigma^2_{\mathtt{u}_2}\mathbb{I}_2$ a presumably known covariance matrix. Though not considered here, more involved source motion models can be dealt with.

2.3. Acoustic model and observation vector

The signals sensed over a T-width time interval associated with iteration k (say, the segment $I_k \triangleq [kT_s - T, kT_s]$) are supposed to be finite time samples of the following random processes)

$$\begin{cases} \mathtt{z}_1(t) = \mathtt{s}(t) + \mathtt{n}_1(t) \\ \mathtt{z}_2(t) = (\mathtt{s} * h_\theta)(t) + \mathtt{n}_2(t), \end{cases} \quad (3)$$

where the signal $\mathtt{s}$ (*i.e.*, the contribution of the emitter at R_1) and the noises $\mathtt{n}_1, \mathtt{n}_2$ are real, zero-mean band-limited, individually and jointly stationary random processes, h_θ denotes the impulse response from R_1 to R_2, and $*$ terms convolution. The source is assumed to lie in the far field, so that h_θ is only parametrized by the source azimuth θ. The noises $\mathtt{n}_1$ and $\mathtt{n}_2$ are supposed independent of $\mathtt{s}$. Importantly, the effect of motion is assumed negligible along any T-width interval, *i.e.*, $\theta(t) \approx \theta_k$ over I_k. The stationarity assumption is also implicitly related to the window length T. For instance, speech signals are generally assumed stationary over intervals of width within [20 ms–40 ms].

For those reasons, T generally does not exceed some tens of milliseconds. The observed signals are divided into N_f—possibly overlapped and windowed—frames of length L, and a Fourier series decomposition is applied to each such frame, yielding, for each channel and frame, a set of B coefficients, indexed by the integers $\ell_1, \ldots, \ell_B$, referring to integer multiples of the fundamental frequency $1/L$. The observation vector $\mathtt{z}$ then results in the channel-frame-frequency representation consisting in the stacking of these Fourier coefficients. From now on, $\mathtt{z}_k$ will refer to the observation vector at iteration k, and should not be confused with the signals $\mathtt{z}_1(t), \mathtt{z}_2(t)$ appearing in (3).

2.4. Fundamentals of stochastic filtering

The aim of the filtering strategy is to recursively compute—or approximate—the *posterior* probability density function (pdf) $p(x_k|z_{1:k})$ of the hidden stochastic state vector $\mathtt{x}_k$ given a sequence of audio measurements $z_{1:k} \triangleq z_1, \ldots, z_k$—considered samples of $\mathtt{z}_{1:k} \triangleq \mathtt{z}_1, \ldots, \mathtt{z}_k$—on the basis of a statistical description of its initial condition $p(x_0)$ (also termed *initial prior*), its *prior dynamics* $p(x_k|x_{k-1})$—fully defined by the kinematic model and the known or guessed statistics of $\mathtt{w}_k, \mathtt{u}_{2k}$—and its *likelihood function* $p(z_k|x_k)$. Theoretically, this is performed recursively *via* the Chapman-Kolmogorov equation (*time update* or *prediction*)

$$p(x_k|z_{1:k-1}) = \int p(x_k|x_{k-1}) p(x_{k-1}|z_{1:k-1}) dx_{k-1} \quad (4)$$

and Bayes' rule (*measurement update*)

$$p(x_k|z_{1:k}) = \frac{p(z_k|x_k) p(x_k|z_{1:k-1})}{p(z_k|z_{1:k-1})}. \quad (5)$$

So, our two step strategy can be reformulated as follows : (1) given a sample z_k of $\mathtt{z}_k$, computation of $p(z_k|x_k)$ from the acoustic model and a look-up table of the Interaural Transfer Function (ITF) $H_\theta(f)$—the Fourier transform of $h_\theta(t)$—; (2) practical implementation of equations (4)–(5) from the result of step (1) and the kinematic model. Step (1) is the topic of §3, while Step (2) will be tackled in §4.

3. Extraction of short-term spatial information from the binaural stream

The aim of this section is to define, at each iteration k of the filtering process, a likelihood function $p(z_k|x_k)$ to be used in the stochastic filter in charge of binaural data assimilation and fusion with motor commands. This likelihood is defined on the basis of the sensed acoustic signals at R_1, R_2 and the ITF $H_\theta(f)$, *i.e.*, the ratio of the Head Related Transfer Functions (HRTFs) to R_1 and R_2.

3.1. Maximum likelihood estimation of the source azimuth

The way how the Maximum Likelihood Estimate (MLE) of θ is obtained from a sample z of $\mathtt{z}$ is first described, assuming that signals and noises are jointly Gaussian.

When the source and noise spectra are flat over any $1/L$-width frequency range and when the autocorrelation time of h_θ is short compared to L—these hypotheses are generally abusive in real conditions, but their validity is seldom questioned in the literature—the mutual correlation of Fourier coefficients over distinct frequencies or nonoverlapped frames becomes negligible. Mutual correlation over overlapped frames can still be neglected if the autocorrelation of the used window function evaluated at the time delay between successive frames is negligible compared to its value at zero (*e.g.*, if a Hann window function is used and frame overlaps do not exceed 50%). Under such simplifications and assuming joint Gaussianity of the source and noises, the logarithm of the pdf of the observation vector—*i.e.*, of the channel-frame-frequency representation of the sensed signals—can be written as

$$\ln p(z; C_{\mathtt{z}}[\ell_1], \ldots, C_{\mathtt{z}}[\ell_B]) = c - N_f \sum_{\ell=\ell_1}^{\ell_B} \left(\ln\det C_{\mathtt{z}}[\ell] + \mathrm{tr}\left(C_{\mathtt{z}}^{-1}[\ell] \hat{C}_{\mathtt{z}}[\ell] \right) \right), \quad (6)$$

with $c \triangleq -2N_f B \ln\pi$ a constant, $C_{\mathtt{z}}[\ell]$ the theoretical spectral covariance matrix of the signals at frequency index ℓ, and $\hat{C}_{\mathtt{z}}[\ell]$ its empirical value built from the entries of z at frequency index ℓ (see [4]). Equation (6) is also referred to as the log-likelihood of matrices $C_{\mathtt{z}}[\ell_1], \ldots, C_{\mathtt{z}}[\ell_B]$ w.r.t. z. From (3) and associated hypotheses, $C_{\mathtt{z}}[\ell]$ can be expressed as

$$C_{\mathtt{z}}[\ell] = V_\theta[\ell] S_{\mathtt{ss}}[\ell] V_\theta[\ell]^\dagger + C_{\mathtt{n}}[\ell], \quad (7)$$

with $S_{\mathtt{ss}}$ the spectrum of $\mathtt{s}$, $V_\theta[\ell] \triangleq [1, H_\theta[\ell]]'$ the so-called sensor steering vector, and $C_{\mathtt{n}}[\ell]$ the acoustic noise spectral covariance matrix. To simplify the problem, first assume that noises $\mathtt{n}_1, \mathtt{n}_2$ are independent and identically distributed (iid), so that $C_{\mathtt{n}}[\ell] = \sigma^2[\ell] \mathbb{I}_2$. The pdf of $\mathtt{z}$ is hence parametrized by the unknown vector $\Theta = [\theta, \Theta'_{\mathrm{spec}}]'$, with Θ_{spec} the vector of spectral parameters of the problem, gathering $S_{\mathtt{ss}}[\ell_1], \ldots, S_{\mathtt{ss}}[\ell_B]$ and $\sigma^2[\ell_1], \ldots, \sigma^2[\ell_B]$. The MLE of Θ is the value $\hat{\Theta}$ maximizing (6), renamed $\ln p(z|\Theta)$. Though no analytical form of $\hat{\Theta}$ can be obtained, a *separable form* is available, *i.e.*, for a given guessed value of θ, the corresponding $\hat{\Theta}_{\mathrm{spec}}(\theta)$ can be derived analytically. Injecting back its expression —as a function of θ—in (6) yields the following criterion—named pseudo log-likelihood—to be maximized

w.r.t. θ [4]

$$L(\theta) = \tilde{c} - N_f \sum_{\ell=\ell_1}^{\ell_B} \mathrm{lndet}\Big(P_\theta[\ell]\hat{C}_{\mathbf{z}}P_\theta[\ell] + P_\theta^\perp[\ell]\mathrm{tr}\left(P_\theta^\perp[\ell]\hat{C}_{\mathbf{z}}[\ell]\right)\Big), \quad (8)$$

with $\tilde{c} \triangleq c - 2N_fB$ a constant independent of θ, $P_\theta[\ell] \triangleq V_\theta[\ell](V_\theta[\ell]^\dagger V_\theta[\ell])^{-1}V_\theta[\ell]^\dagger$ the orthogonal projector onto the space spanned by $V_\theta[\ell]$, and $P_\theta^\perp[\ell] \triangleq \mathbb{I}_2 - P_\theta[\ell]$ its orthogonal complement. Once the argmax $\hat{\theta}$ of (8) has been found, the spectral parameter vector estimate $\hat{\Theta}_{\mathrm{spec}}(\hat{\theta})$ can be deduced if necessary. In real acoustic conditions, the noise is generally spatially and temporally correlated, so that $C_{\mathbf{n}}[\ell] = \sigma^2[\ell]\mathbb{I}_2$ no longer holds. However, if $C_{\mathbf{n}}[\ell] = \sigma^2[\ell]\tilde{C}_{\mathbf{n}}[\ell]$ with $\tilde{C}_{\mathbf{n}}[\ell]$ some known Hermitian symmetric definite positive matrix, whose Cholesky decomposition is denoted $Q_{\mathbf{n}}[\ell]$, then the problem can still be handled by replacing $\hat{C}_{\mathbf{z}}[\ell]$ and $P_\theta[\ell]$ with

$$\tilde{\hat{C}}_{\mathbf{z}}[\ell] \triangleq Q_{\mathbf{n}}^{-1}[\ell]\hat{C}_{\mathbf{z}}[\ell]Q_{\mathbf{n}}^{-1}[\ell]^\dagger, \quad (9)$$
$$\tilde{P}_\theta[\ell] \triangleq Q_{\mathbf{n}}^{-1}[\ell]V_\theta[\ell](V_\theta[\ell]^\dagger\tilde{C}_{\mathbf{n}}^{-1}[\ell]V_\theta[\ell])^{-1}V_\theta[\ell]^\dagger Q_{\mathbf{n}}^{-1}[\ell]^\dagger$$

respectively in (8). We now present a procedure to detect source activity.

3.2. Information-theoretic source activity detection

So far, the observation vector statistics have been defined in (6), with the structure of $C_{\mathbf{z}}[\ell]$ given by (7). This "emitting source" model is referred to as $\mathcal{M}_e$. A second model $\mathcal{M}_s$ is now considered, in which the source is silent, *i.e.*, $S_{\mathbf{ss}} \equiv 0$, $C_{\mathbf{z}}[\ell]$ reduces to $C_{\mathbf{n}}[\ell]$, and Θ boils down to the noise spectral parameters. Their maximum likelihood estimates straightforwardy come as

$$\hat{\sigma}^2[\ell] = \frac{1}{2}\mathrm{tr}\left(\tilde{C}_{\mathbf{n}}^{-1}[\ell]\hat{C}_{\mathbf{z}}[\ell]\right). \quad (10)$$

The aim is to define a decision rule to detect one of the two competing models $\mathcal{M}_e, \mathcal{M}_s$, on the basis of the sample z. Both models appear in the form (6), but with different restrictions on the parameter vector Θ. Maximizing $\ln p(z|\Theta)$ for each model and selecting the most likely one is not a sound solution, in that the model with the highest number of free parameters is favored on average. Contrarily, the minimization of the Akaike Information Criterion (AIC) [9]

$$\mathrm{AIC}(\Theta) = -\ln p(z|\Theta) + P, \quad (11)$$

with P the number of free parameters in Θ, yields a sound model detection scheme. Here, no contraints bind the components of Θ, so that P is just its dimension.

3.3. Design of the likelihood function to be used in the filter

The first idea to fuse short-term spatial information with motor commands would be to assimilate the argmax of (8) in a stochastic filter, at times when the source is detected as "active" through (11). Unfortunately, this does not yields a sound solution. For microphones configurations that present a front-back symmetry (similar measured pressure when a source at azimut θ is shifted to azimuth $\pi - \theta$), *e.g.* for microphones mounted at the antipodes of a sphere or in free-field, the symmetry is also found in the pseudo log-likelihood, which admits two maxima with equal height[1]. Though the use of a nonsymmetric sensor (anthropomorphic head, presence of pinnæ) can theoretically break the front-back ambiguity, disambiguation remains difficult in practice due to acoustic noise, modeling errors, etc. (see figure 2), and the selection of the "wrong" peak can have drastic consequences on the filter consistency.

Rather than assimilating only the MLE of θ, we propose to build the likelihood function $p(z|\theta)$ to be used in the filter from the pseudo log-likelihood (8) itself. Though, strictly speaking, the exponential of (8) is not equal to the genuine likelihood function, we will nevertheless use it for $p(z|\theta)$. Yet, an important difficulty remains. The pseudo log-likelihood (8) has no analytical expression, as its numerical values are given only for a discrete set of azimuths (related to the limited resolution of the ITF). This induces problems to instantiate the Bayes inference. They could be solved in a particle filter, but as the generic particle filter can be proven to fail in this context we prefer a Gaussian sum filter. So, we propose to model the likelihood function as the unnormalized Gaussian mixture

$$p(z|\theta) = \sum_{j=1}^{J} \gamma^j \mathrm{e}^{-\frac{1}{2}\frac{\left(\theta - m^j\right)^2}{\phi^j}}, \quad (12)$$

where $\{\gamma^j, m^j, \phi^j\}_{j=1,\ldots,J}$ are to be deduced from (8). As it will be seen in §4, such a structure allows the derivation of a Gaussian mixture approximation to the posterior state pdf. The question is now to deduce the parameters in (12) from the evaluation of (8). Clearly, the prohibitive cost of nonlinear least-square methods precludes their use in online applications (though they have been used in [10]). We propose a rather simple alternative: (1) select the local maxima of the pseudo log-likelihood above a predefined threshold, and set the means $\{m^j\}_{j=1,\ldots,J}$ as the arguments of these maxima; (2) set the weights $\{\gamma^j\}_{j=1,\ldots,J}$ according to the peaks heights; (3) set the "variances" $\{\phi^j\}_{j=1,\ldots,J}$ to $\mathrm{res}^2/12$, with res the spatial resolution

[1] Note that in free-field, this problem is avoided by parametrizing the pdf of $\mathbf{z}$ in terms of Interaural Time Difference rather than azimuth.

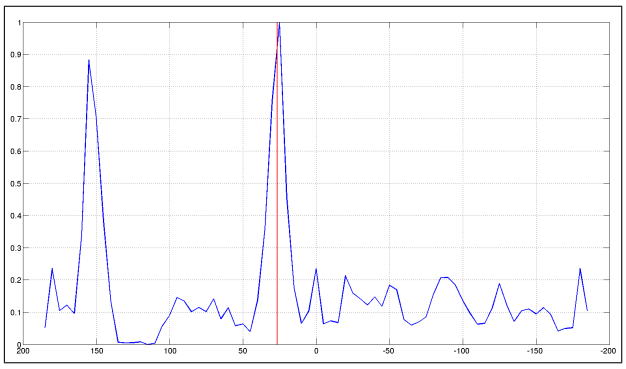

Figure 2. Normalized pseudo log-likelihood function computed from KEMAR® recordings (simulated). Abscissa is in degrees. True source azimuth θ_{true} is shown in red. The function exhibits two main peaks, one coinciding with θ_{true}, the other located near $180° - \theta_{\text{true}}$.

of the ITF (corresponding to the variance of a uniform distribution over a res-width interval).

4. Fusion of audio information with motor commands

4.1. Towards a well-posed filtering problem

To tackle binaural active localization through a well-posed filtering problem, a state space model must be defined where the state vector is minimal and gathers only the variables to be estimated. Define $\mathtt{e}_k \triangleq [\mathtt{e}_{y_k}, \mathtt{e}_{z_k}]'$. From the fact that observations do not depend on λ, and under some mild conditions on $\mathtt{w}, \mathtt{u}_2$ and the state initial conditions[2], one can deduce that $\mathtt{e}_k$ and λ_k are independent *a posteriori*, λ_k is independent from $\mathtt{z}_{1:k}$, and the time evolution of $\mathtt{e}_k$ can be described by the equation

$$\mathtt{e}_k = \tilde{f}(\mathtt{e}_{k-1}, u_{1\,k-1}) + \tilde{\mathtt{w}}_{k-1} \tag{13}$$

independently of λ_k, with $\tilde{f}$ given in appendix 8, and $\tilde{\mathtt{w}}$ a noise term whose statistics are deduced from those of $\mathtt{w}$ and $\mathtt{u}_2$.

Because spatial short-term information extracted from the signals at the microphones is purely directional—*i.e.*, $p(z_k|e_k) = p(z_k|\theta_k)$—, it may be more convenient to parametrize the source relative position in terms of polar coordinates $r \triangleq [\rho, \theta]'$, related to cartesian coordinates through (see also figure 1)

$$\begin{bmatrix} e_y \\ e_z \end{bmatrix} = \begin{bmatrix} \rho \sin\theta \\ \rho \cos\theta \end{bmatrix}, \quad \begin{bmatrix} \rho \\ \theta \end{bmatrix} = \begin{bmatrix} \sqrt{e_y^2 + e_z^2} \\ \text{atan2}(e_y, e_z) \end{bmatrix}. \tag{14}$$

Hence, the stochastic state space equation governing polar coordinates becomes

$$\mathtt{r}_k = f_{\text{c2p}}(\tilde{f}(f_{\text{p2c}}(\mathtt{r}_{k-1}), u_{1\,k-1}) + \tilde{\mathtt{w}}_{k-1}), \tag{15}$$

[2] λ_0 and $\mathtt{e}_0$ are mutually independent, the component of $\mathtt{w}_k$ associated with λ_k is independent from the components related to $\mathtt{e}_k$.

with f_{c2p} and f_{p2c} the cartesian to polar and polar to cartesian transforms shown in equation (14). For the forthcoming sections, we denote $f_p(.,.,.) \triangleq f_{\text{c2p}}(\tilde{f}(f_{\text{p2c}}(.), .) + .)$.

Three specificities of our problem are to be noted: (1) the likelihood $p(z_k|r_k) = p(z_k|\theta_k)$ has no analytical expression, only its numerical values for a discrete set of azimuths are computed at each time step k (this has been discussed in §3); (2) f_p is a nonlinear function of the state and noise (in particular, noise is non-additive); and (3) little prior information is available at initial time about the state, *i.e.*, $p(r_0)$ is a «flat» function, whose associated 99% confidence region of minimum volume covers a wide region of the state space. Hereafter, an original filtering strategy, based on the *unscented transform*—a statistical linearization technique—together with a Gaussian mixture approximation of the likelihood function and the posterior state pdf, is proposed to circumvent the aforementioned difficulties.

4.2. A Gaussian mixture filter

Due to the specificites of the problem, the determination of the posterior state pdf $p(r_k|z_{1:k})$ through Bayesian inference is a difficult task, since no analytical form is available. In the linear Gaussian case, $p(r_k|z_{1:k})$ is shown to be a Gaussian, an hence fully caracterized by its two first moments $\hat{r}_{k|k} \triangleq \mathbb{E}\{\mathtt{r}_k|z_{1:k}\}$ and $P_{k|k} \triangleq \mathbb{E}\{(\mathtt{r}_k - \hat{r}_{k|k})(\mathtt{r}_k - \hat{r}_{k|k})'|z_{1:k}\}$, computed recursively and analytically through the celebrated Kalman Filter (KF) equations. Nonlinear extensions of the KF, such as the Extended Kalman Filter (EKF) and the Unscented Kalman Filter (UKF) approximate the two first moments of $p(r_k|z_{1:k})$ from linearization techniques, assuming that the initial prior is Gaussian. These extensions however lead to overconfident conclusions with a flat initial prior. To solve this problem, we propose here, in the vein of [11], to approximate the posterior state pdf as the Gaussian mixture

$$p(r_k|z_{1:k}) = \sum_{i=1}^{I_k} w_k^i \mathcal{N}\left(r_k; \hat{r}_{k|k}^i, P_{k|k}^i\right), \tag{16}$$

fully described by the set of weights and moments $\{w_k^i, \hat{r}_{k|k}^i, P_{k|k}^i\}_{i=1,\ldots,I_k}$, recursively computed through the following steps.

4.2.1. Time update step

It is assumed that at iteration $k-1$, the posterior state pdf is accurately approximated by a Gaussian mixture of the form

$$p(r_{k-1}|z_{1:k-1}) = \sum_{i=1}^{I_{k-1}} w_{k-1}^i \mathcal{N}\left(r_{k-1}; \hat{r}_{k-1|k-1}^i, P_{k-1|k-1}^i\right). \tag{17}$$

Then, according to the Chapman-Kolmogorov equation (4), the prediction pdf of $\mathtt{r}_k$ writes as

$$p(r_k|z_{1:k-1}) = \sum_{i=1}^{I_{k-1}} w_{k-1}^i \int p(r_k|r_{k-1}) \mathcal{N}\left(r_{k-1}; \hat{r}_{k-1|k-1}^i, P_{k-1|k-1}^i\right) dr_{k-1}. \quad (18)$$

To make the problem tractable, we approximate, for $i = 1, \ldots, I_{k-1}$, the i^{th} integral term in (18) as a Gaussian, whose moments $\hat{r}_{k|k-1}^i, P_{k|k-1}^i$ come from $\hat{r}_{k-1|k-1}^i, P_{k-1|k-1}^i$, f_p and from $Q_{\tilde{\mathtt{w}}}$. This computation relies on the so-called unscented transform [12]. As the dynamic noise is nonadditive, we consider the augmented vector $\mathtt{r}_k^a \triangleq [\mathtt{r}_k', \tilde{\mathtt{w}}_k']'$. Then we build, for $i = 1, \ldots, I_{k-1}$, the moments

$$\hat{r}_{k-1|k-1}^{i,a} \triangleq \left[\hat{r}_{k-1|k-1}^{i\,\prime}, 0, 0\right]', \quad P_{k-1|k-1}^{i,a} = \text{blockdiag}\left(P_{k-1|k-1}^i, Q_{\tilde{\mathtt{w}}}\right). \quad (19)$$

A deterministic sampling of the associated pdf, capturing the information on $\hat{r}_{k-1|k-1}^{i,a}, P_{k-1|k-1}^{i,a}$, is then performed through the computation of σ-points, along [12]. The σ-points are then passed through the nonlinear function f_p, and the empirical weighted mean and covariance matrix of the resulting points are then computed, yielding $\hat{r}_{k|k-1}^i, P_{k|k-1}^i$. So, the prediction pdf of $\mathtt{r}_k$ approximately writes as

$$p(r_k|z_{1:k-1}) = \sum_{i=1}^{I_{k-1}} w_{k-1}^i \mathcal{N}\left(r_k; \hat{r}_{k|k-1}^i, P_{k|k-1}^i\right). \quad (20)$$

4.2.2. Measurement update step

The likelihood function is supposed to have the following structure

$$p(z_k|r_k) = \sum_{j=1}^{J_k} \gamma_k^j \mathrm{e}^{-\frac{1}{2}\left(r_k - \mu_k^j\right)' \left(\Phi_k^j\right)^{-1} \left(r_k - \mu_k^j\right)}. \quad (21)$$

To model the fact that $p(z_k|r_k)$ does not depend on the range ρ_k, we set $\mu_k^j = [0, m_k^j]'$, $\Phi_k^j = \text{diag}(\infty, \phi_k^j)$, with $\{\gamma_k^j, m_k^j, \phi_k^j\}_{j=1,\ldots,J_k}$ obtained from the procedure described in 3.3. From classical results on products of Gaussians, it can be shown that Bayes' rule (5) yields

$$p(r_k|z_{1:k}) \propto \sum_{i=1}^{I_{k-1}} \sum_{j=1}^{J_k} \alpha_k^{i,j} \mathcal{N}\left(r_k; \hat{r}_{k|k}^{i,j}, P_{k|k}^{i,j}\right), \quad (22)$$

with

$$P_{k|k}^{i,j} = \left((P_{k|k-1}^i)^{-1} + (\Phi_k^j)^{-1}\right)^{-1}, \quad (23)$$

$$\hat{r}_{k|k}^{i,j} = P_{k|k}^{i,j}\left[(P_{k|k-1}^i)^{-1} \hat{r}_{k|k-1}^i + (\Phi_k^j)^{-1} \mu_k^j\right], \quad (24)$$

and

$$\alpha_k^{i,j} = w_{k-1}^i \gamma_k^j \left(\det P_{k|k}^{i,j}\right)^{\frac{1}{2}} \left(\det P_{k|k-1}^i\right)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}\left(\hat{r}_{k|k-1}^i - \mu_k^j\right)' \left(P_{k|k-1}^i + \Phi_k^j\right)^{-1} \left(\hat{r}_{k|k-1}^i - \mu_k^j\right)}. \quad (25)$$

Numerical evaluation of (25) requires the computation of $(P_{k|k-1}^i + \Phi_k^j)^{-1}$, which is shown to simplify to

$$(P_{k|k-1}^i + \Phi_k^j)^{-1} = \text{diag}(0, ((P_{k|k-1}^i)_{2,2} + \phi_k^j)^{-1}), \quad (26)$$

with $(.)_{2,2}$ terming the 2^{nd} row 2^{nd} column element of a matrix. The normalizing constant $p(z_k|z_{1:k-1})$ in (22) is given by marginalizing the joint pdf $p(r_k, z_k|z_{1:k-1})$ w.r.t. r_k

$$C_k \triangleq p(z_k|z_{1:k-1}) = \int p(z_k|r_k) p(r_k|z_{1:k-1}) dr_k = \sum_{i=1}^{I_{k-1}} \sum_{j=1}^{J_k} \alpha_k^{i,j}, \quad (27)$$

so that the posterior state pdf at k finally becomes

$$p(r_k|z_{1:k}) = \sum_{i=1}^{I_{k-1}} \sum_{j=1}^{J_k} w_k^{i,j} \mathcal{N}\left(r_k; \hat{r}_{k|k}^{i,j}, P_{k|k}^{i,j}\right) \quad (28)$$

with $w_k^{i,j} = \alpha_k^{i,j}/C_k$. Reindexing the $I_k = I_{k-1} + J_k$ terms as $i \leftarrow i + I_{k-1}(j-1)$, (28) can be reformulated as (16).

4.2.3. Pruning and merging hypotheses

As equation (28) shows, the number of hypotheses in the mixture increases with time, which prevents any practical implementation without an efficient hypotheses management. The used procedure consists in pruning Gaussians whose weights are below a given threshold, and then merging Gaussians that are "close" to each other in the sense of a normalized quadratic distance between their means (this is performed heuristically). Finally, if the number of hypotheses is still larger than a maximum allowed number $I_{\max}$, then only the $I_{\max}$ Gaussians with the highest weights can be kept.

5. Simulation

To qualitatively study the behaviour of the filtering scheme, simulations have been conducted with MATLAB®. The perceived binaural signals have been synthetized by using the KEMAR® dummy-head head-related impulse response (HRIR) measurements (large pinnae, 0° elevation) made available by MIT Media Lab[3]. The emitted signal was a 15 seconds-long male speaker utterance record from french radio, sampled at 44.1 kHz. As in the presence of relative motion,

[3] http://sound.media.mit.edu/resources/KEMAR.html

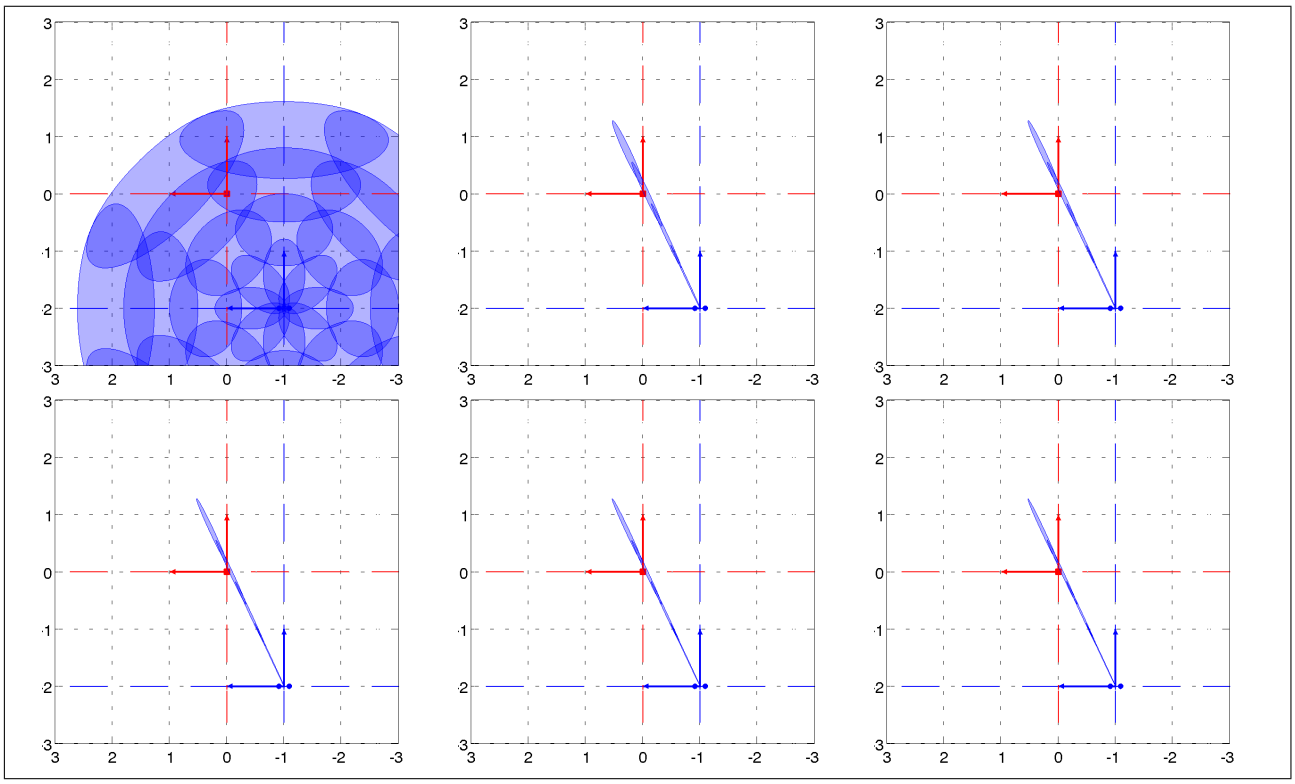

Figure 3. Localization results in the world reference frame $\mathcal{F}_O$ at times $\{0\text{ s}, 3\text{ s}, 6\text{ s}, 9\text{ s}, 12\text{ s}, 15\text{ s}\}$; static source and sensor.

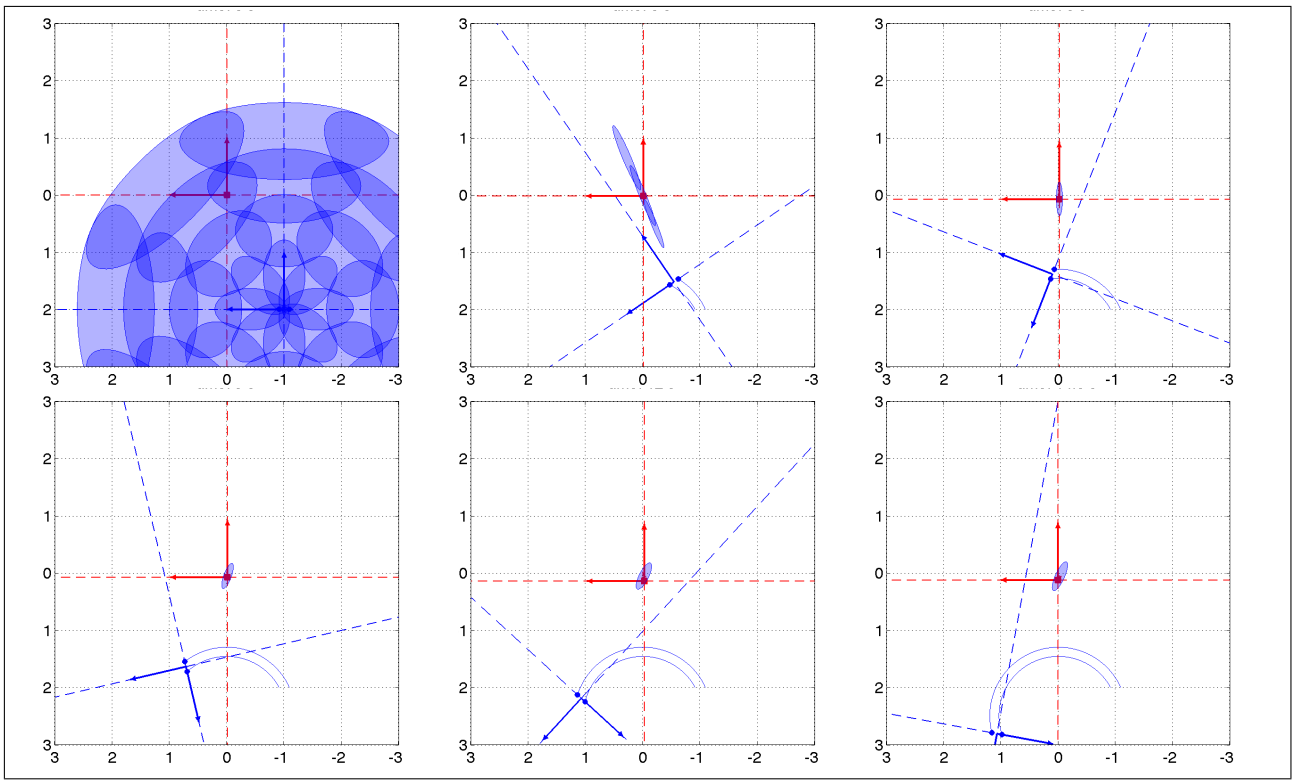

Figure 4. Localization results in the world reference frame $\mathcal{F}_O$ at times $\{0\text{ s}, 3\text{ s}, 6\text{ s}, 9\text{ s}, 12\text{ s}, 15\text{ s}\}$; moving source and sensor.

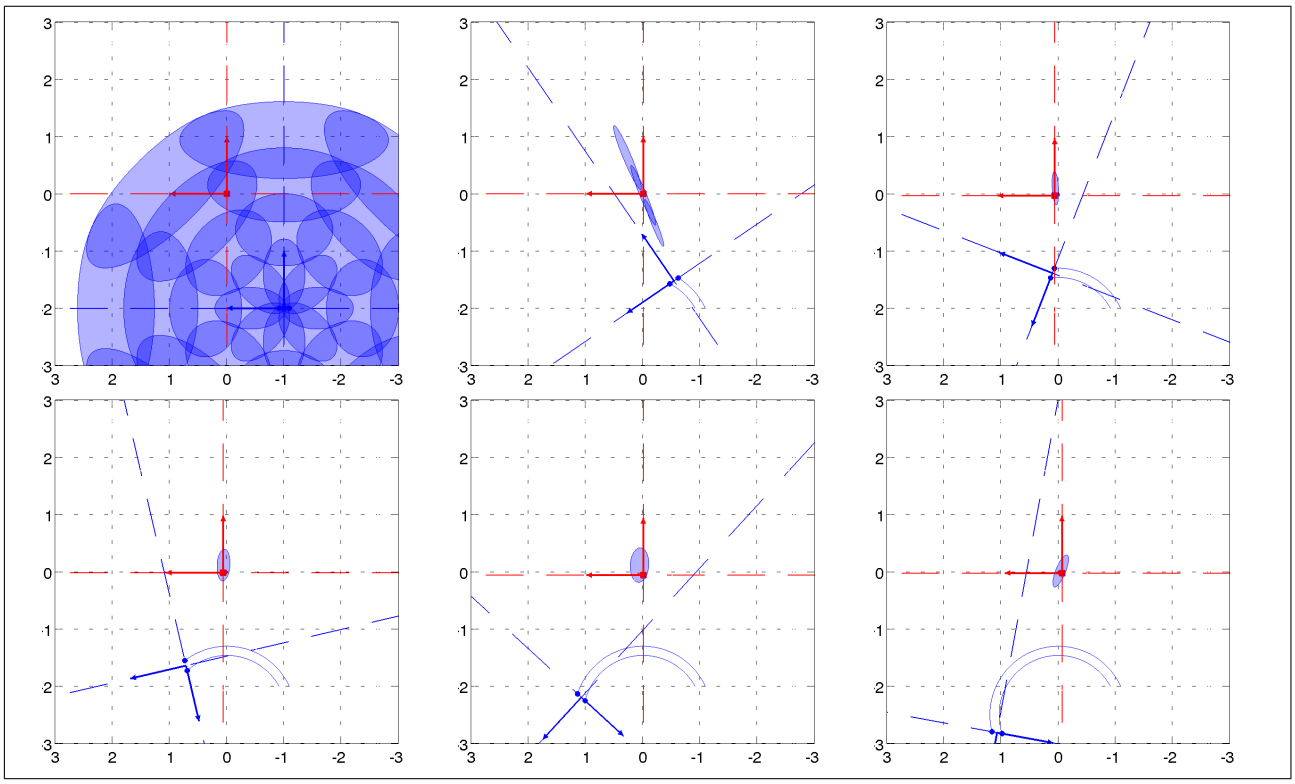

Figure 5. Localization results in the world reference frame $\mathcal{F}_O$ at times $\{0\text{ s}, 3\text{ s}, 6\text{ s}, 9\text{ s}, 12\text{ s}, 15\text{ s}\}$; moving source and sensor, with source intermittence.

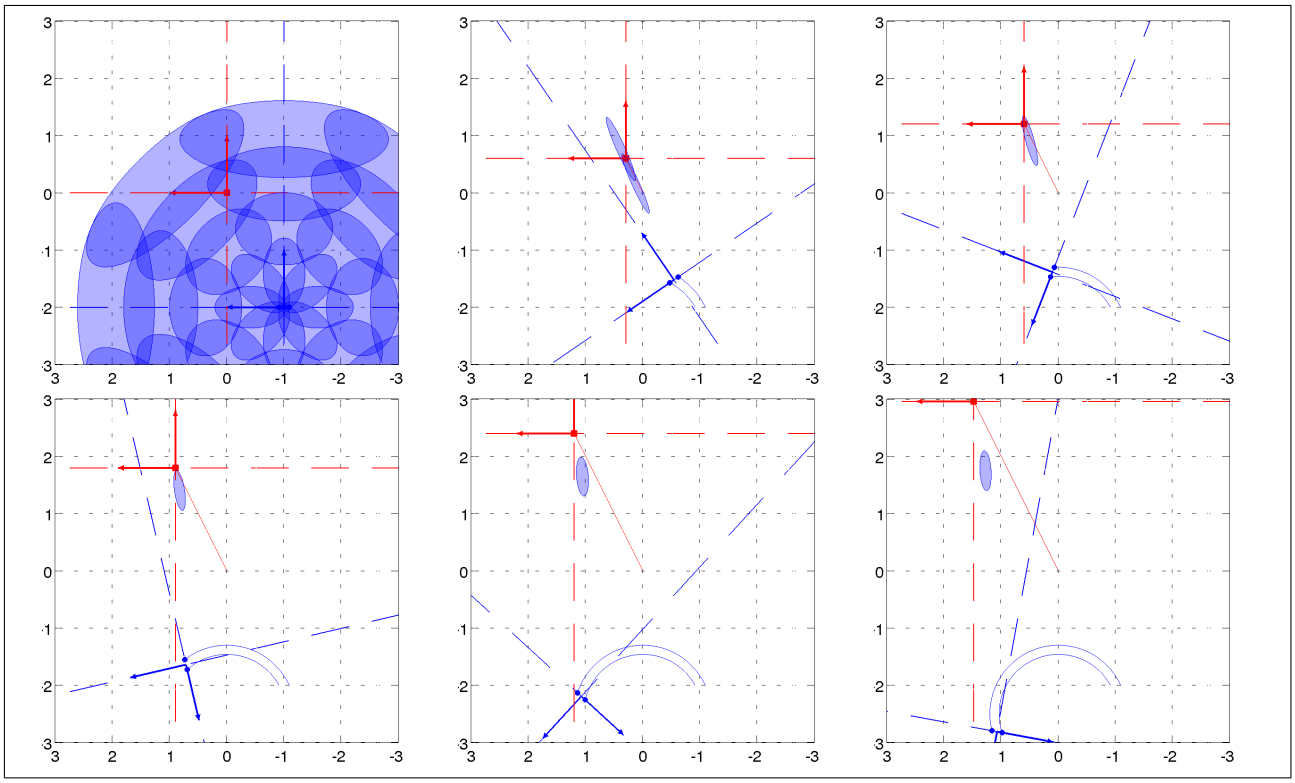

Figure 6. Localization results in the world reference frame $\mathcal{F}_O$ at times $\{0\text{ s}, 3\text{ s}, 6\text{ s}, 9\text{ s}, 12\text{ s}, 15\text{ s}\}$; moving source and sensor. Mismatch induced by an incorrect source motion model.

the source azimuth varies continously with time, and since the HRIR spatial resolution is only 5°, a linear interpolation was performed to get the HRIR at any desired azimuth. To synthetize signals while the source and sensor move, azimuth was zero-order held over 512 samples long frames (≈ 12 ms). Because the HRIR was measured for a single distance value (≈ 1.4 m), the effect of source distance on the perceived signals has not been taken into account in the simulation. Finally, additive iid white Gaussian noises have been added to the synthetized signals, so that the Signal-to-Noise Ratio at microphone R_1 was 13 dB. No room reverberation has been simulated.

Concerning the filtering strategy, short term spatial information extraction and source activity detection have been performed, at each iteration of the localization process, from 4 frames of 1024 samples with 50 % overlap between frames (total length 58 ms). The localization process sampling period T_s has been set to 200 ms. Though a linear interpolation of the HRIR was used to synthetize signals, azimuths tested in the procedure were limited to the azimuths available in the original HRIR dataset (−180° to 180° by steps of 5°).

Several scenarios have been considered to show some interesting features of the strategy. Results are reported in figures 3 to 6. In these figures, ground truth positions as well as localization results are depicted at 6 time steps (chronological order from left to right and top to bottom). Axes are in meters. The sensor frame is plotted in blue, while the source frame is depicted in red. Blue ellipses represent 99% cofidence regions associated with the various hypotheses of the mixture characterizing the posterior state pdf. Note that the filter presented in §4 computes the joint posterior state pdf of polar coordinates relative to the sensor frame $\mathcal{F}_R$, while results are here shown in terms of cartesian coordinates relative to the world frame $\mathcal{F}_O$. This was done for pure reading convenience. To do so, the polar to cartesian transform is performed through the use of the unscented transform (see §4), then absolute cartesian coordinates are obtained from ground truth positions of the sensor and source. Again, one must be aware that the source localisation is relative, and that absolute positions are shown because we think it is a more "natural" way to depict results.

In figure 3, the source and sensor are perfectly static. The hypotheses of the posterior state pdf spread along the sensor-to-source direction, depicting that there is a large uncertainty on the distance to the source. This comes from the fact that short-term spatial information is purely directional, and that in the absence of relative motion, one cannot expect to recover the observability of the source distance. Note that the hypotheses do not spread exactly along the sensor-to-source direction, and that the source lies slightly outside the confidence region of the filter. This is due to the limited spatial resolution of the KEMAR® HRIR (5°). More consistency could be obtained by enlarging slightly the ϕ_j in (12). Also, note that if the sensor had a symmetric configuration, a symmetry w.r.t. the (R_1R_2) axis would be reported in the posterior state pdf. In this last case, motion (*e.g.* head rotation) would be required to disambiguate front and back. In figure 4, both source and sensor are in motion ($Q_{\tilde{\mathbf{w}}} = \mathbb{O}$, $\sigma_{\mathbf{u}_2} = 0.05$ m/s, $u_1 \equiv [0.1 \text{ m/s}, 0.2 \text{ m/s}, -0.2 \text{ rad/s}]'$). As it can be seen, quite rapidly, a single sharp hypothesis remains, fitting the true source position, meaning that distance observability has been recovered. Figure 5 shows the same scenario as in figure 4, but the source stops emitting within the interval [7 s, 13 s]. Source silence is correctly detected by the SAD scheme, so that only prediction is performed during this period. As a consequence, during prediction, one can see the confidence region spreading more and more with time, due to the integration of dynamic noise. When the source emits again, this region gets sharp again, thanks to the assimilation of new audio information. Finally, figure 6 shows the importance of kinematic modeling. It points out the consequences of a mismatch between the genuine source motion model and the one used in the filter. The source is thought to follow a Gassian randow walk (as in figure 4), while it actually follows a rectilinear uniform motion, with $u_2 \equiv [0.1 \text{ m/s}, 0.2 \text{ m/s}]'$. A consistency loss is reported, and the source range is clearly underestimated. This conclusion highlights an important problem in robotics: the modeling of human motion. Random walks and Langevin processes are often used in the literature [13], but human dynamics is clearly more involved, time varying, etc. Modeling issues and multiple models approaches [14] are subject to future work.

6. Experiments

Live experiments shown in this section are limited to the case when microphones are assumed in free-field, observations are maximum likelihood estimates of the Interaural Time Difference (ITD), and no SAD is performed (*i.e.* short-term spatial information is systematically assimilated into the stochastic filter). Experimental conditions are quite favorable—the source signal is white—and limited—only rotation of the sensor is performed. Implementation of the complete strategy, as well as experimentation in more adverse and realistic conditions are in progress.

In the experiments, the microphones are placed on a Pan-Tilt unit (see figure 8) whose motion is limited to pan rotation (0° tilt angle). The signals perceived by the microphones are relayed to a computer performing sound localization and controlling the pan-tilt unit according to a predefined motion. Localization results are displayed in real time. §6.1 provides details on the

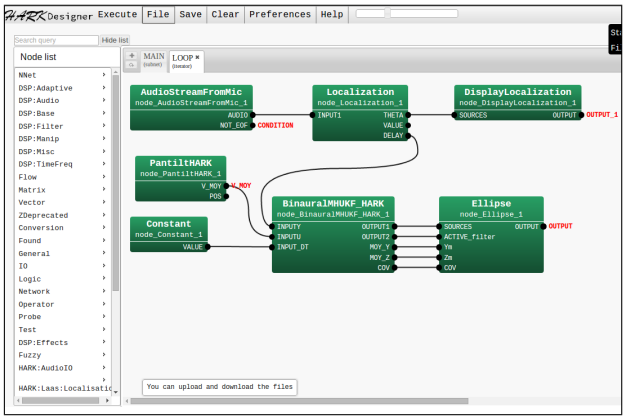

Figure 7. $\mathcal{HARK}$ network for active localization.

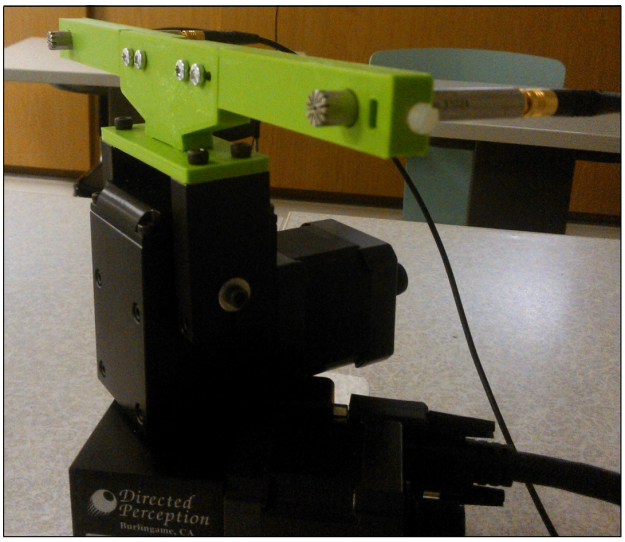

Figure 8. Pan-Tilt Unit with two GRAS® microphones.

software part of the experiment, while §6.2 shows the hardware part and results.

6.1. $\mathcal{HARK}$ implementation

The software part of the experiment is held by the comprehensive open source robot audition suite $\mathcal{HARK}$ [15]. This software provides several online implementations of canonical auditory functions such as localization, separation, or recognition. It relies on the open-source middleware Flowdesigner which enables the design of complex functions. These are described on the basis of elementary blocks, whose interconnections depict the dataflow circulating between them (`http://sourceforge.net/apps/mediawiki/flowdesigner`).

After being prototyped under MATLAB®, the algorithms were coded into C++ so as to be encapsulated into $\mathcal{HARK}$, which enables online data processing. New nodes called Localization, BinauralMHUKF_HARK, PantilHARK and Ellipse were synthetized.

The main loop of the program depends on the node AudioStreamFromMic which reads the ALSA buffer. At each step of the loop, the node performs the channel-frame-frequency decomposition of the binaural input signals (streamed from the microphones, or possibly from an audio file) on the basis of the fftw library (which provides a robust and fast way to compute the Dsicrete Fourier Transform), and outputs an estimated ITD. The node PantiltHARK controls the Pan-Tilt motion and outputs its speed. The node BinauralMHUKF_HARK fuses the information delivered by Localization and PantiltHARK with the Multi-Hypothesis Unscented Kalman Filter strategy [2]. The results are observed in real time from the node Ellipse which displays the 99% confidence region associated with each hypothesis in the frame $\mathcal{F}_R$ linked to the binaural sensor.

6.2. Scenario and results

The $\mathcal{HARK}$ live experiment of the has been performed with two G.R.A.S® 40PQ microphones spaced out from 170 mm and a loudspeaker emitting a white noise. The digital sound source acquisition is obtained by connecting these microphones to the RME® Babyface which provides a low latency audio interface with the computer. Then the audio stream is held by ALSA. The node AudioStreamFromMic is configured to work with ALSA and it is tuned at the 44100 Hz sampling frequency. The microphones are fixed on a Directed Perception® Pan-Tilt Unit which undergoes a two-way cycle from 40° to −40° at constant velocity (see Figure 8).

Two scenarios have been studied. One with the sound source in front of the microphones (*i.e.* 0°) and the other with an angle of 30°. Due to the front-back symmetry of the sensor configuration, the hypotheses of the mixture spread along two directions at the very first iterations. However, rotation of the sensor allows front-back disambiguation. Because no translational motion of the sensor is performed, the uncertainty on the source distance remains large.

7. Conclusions

In this paper, an "active" binaural localization scheme has been presented. It jointly uses auditive and proprioceptive modalities of a binaural sensor to localize a possibly moving sound source. The proposed strategy, combined with a suitable motion, is shown to allow automatic triangulation of the source position.

Several prospective issues have to be dealt with. First, a quantitative evaluation of performances in realistic experimental conditions has to be performed. Particularly, the robustness of the spatial information extraction and SAD schemes to modeling errors (reverberation, etc.) has to be tested. The influence of false detections and false measurements on the filter consistency is a crucial point that has to be investigated. Secondly, the effect of a realistic (*i.e.* correlated and nonstationary) acoustic noise has to be analysed.

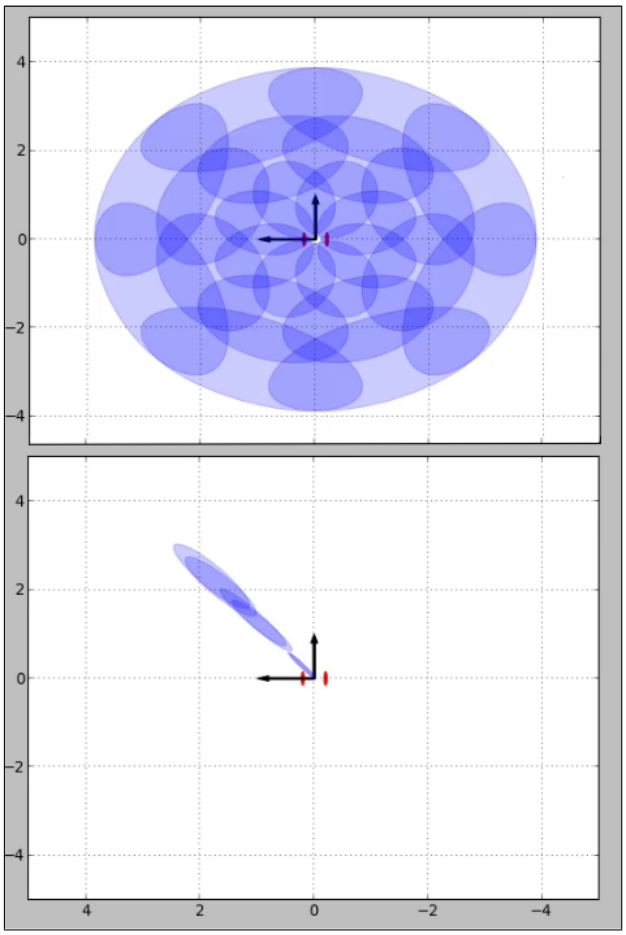

Figure 9. MHUKF result : 99% confidence region of source localization at initial step and after few seconds.

When the sensor is embeded on a robotic platform, the noise generated by its actuators is a additional problem that has to be tackled. The way how the statistics of the various noises can be learned and used has to be investigated. Other issues concern modeling of the source dynamics (as stated in §5), extension to multiple sound sources, and definition of active control stategies of the sensor motion to obtain localization results as informative as possible (in the idea of moving to improve perception).

8. Appendix : discrete time state space equation

The function f appearing in equation (2) has the following expression

$$f(x_k, u_{1k}, u_{2k}) = F(\omega_k)x_k + G_1(\omega_k)u_{1k} + G_2(\lambda_k, \omega_k)u_{2k}, \quad (29)$$

with

$$F(\omega_k) = \begin{bmatrix} \cos(\omega_k T_s) & \sin(\omega_k T_s) & 0 \\ -\sin(\omega_k T_s) & \cos(\omega_k T_s) & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$G_1(\omega_k) = \begin{bmatrix} -\frac{\sin(\omega_k T_s)}{\omega_k} & \frac{\cos(\omega_k T_s)-1}{\omega_k} & 0 \\ -\frac{\cos(\omega_k T_s)-1}{\omega_k} & -\frac{\sin(\omega_k T_s)}{\omega_k} & 0 \\ 0 & 0 & -T_s \end{bmatrix},$$

$$G_2(\lambda_k, \omega_k) = T_s \begin{bmatrix} \cos(\lambda_k - \omega_k T_s) & -\sin(\lambda_k - \omega_k T_s) \\ \sin(\lambda_k - \omega_k T_s) & \cos(\lambda_k - \omega_k T_s) \\ 0 & 0 \end{bmatrix}. \quad (30)$$

The proof can be found in [8]. The function $\tilde{f}$ in equation (13) comes as

$$\tilde{f}(e_k, u_{1k}) = \tilde{F}(\omega_k)e_k + \tilde{G}_1(\omega_k)u_{1k} \quad (31)$$

with $\tilde{F}(\omega_k)$ (resp. $\tilde{G}_1(\omega_k)$) obtained by removing the last column and row (resp. last row) from $F(\omega_k)$ (resp. $G_1(\omega_k)$).

Acknowledgement

This research has been supported by EU FET grant TWO!EARS, ICT-618075, `www.twoears.eu`, and BINAAHR (BINaural Active Audition for Humanoid Robots) project funded by ANR (France) and JST (Japan) under Contract n°ANR-09-BLAN-0370-02.

References

[1] D. Wang and G. J. Brown: Computational Auditory Scene Analysis: Principles, Algorithms and Applications. John Wiley and Sons, New York, 2006.

[2] A. Portello, P. Danès and S. Argentieri: Acoustic models and Kalman filtering strategies for active binaural sound localization. Int. Conf. on Intelligent Robots and Systems, pp. 137-142, 2011.

[3] A. Portello, P. Danès and S. Argentieri: Active binaural localization of intermittent moving sources in the presence of false measurements. Int. Conf. on Intelligent Robots and Systems, pp. 3294-3299, 2012.

[4] A. Portello, P. Danès, S. Argentieri, and S. Pledel: HRTF-based source azimuth estimation and activity detection from a binaural sensor. Int. Conf. on Intelligent Robots and Systems, pp. 2908-2913, 2013.

[5] A.G. Jaffer: Maximum likelihood direction finding of stochastic sources: a separable solution. Int. Conf. on Acoustics, Speech, and Signal Processing, vol. 5, pp. 2893-2896, 1988.

[6] M.A. Doron, A.J. Weiss and H. Messer: Maximum-Likelihood Direction Finding of Wide-Band Sources. IEEE Transactions on Signal Processing, vol. 41, n° 1, pp. 411-.

[7] Y. Lu and M. Cooke: Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners. Jour. of Speech Communication, 2010.

[8] A. Portello: Active Binaural Localisation of Sound Sources in Humanoid Robotics. Ph.D. dissertation, Univ. Toulouse III Paul Sabatier (in French), 2013.

[9] H. Akaike: A new look at the statistical model identification. IEEE Trans. on Automatic Control, vol. 19, n° 6, pp. 716-723, 1974.

[10] I. Marković, A. Portello, P. Danès, I. Petrović and S. Argentieri: Active Speaker Localization with Circular Likelihoods and Bootstrap Filtering. Int. Conf. on Intelligent Robots and Systems, pp. 2914-2920, 2013.

[11] D.L. Alspach and H.W. Sorenson: Nonlinear Bayesian estimation using Gaussian sum approximations. IEEE Trans. on Automatic Control, vol. 17, n° 4, pp. 439-448, 1972.

[12] S. J. Julier and J. K. Uhlmann: New extension of the Kalman filter to nonlinear systems. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 3068, pp. 182-193, 1997.

[13] I. Marković and I. Petrović: Speaker localization and tracking with a microphone array on a mobile robot using von Mises distribution and particle filtering. Journal of Robotics and Autonomous Systems, vol. 58, n° 11, pp. 1185-1196, 2010.

[14] Y. Bar-Shalom and X.R. Li: Estimation and Tracking : Principles, Techniques and Software. Artech House, 1993.

[15] K. Nakadai, T. Takahashi, H.G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino: Design and Implementation of Robot Audition System HARK. Jour. of Advanced Robotics, vol. 24, n° 5-6, pp. 739-761, 2010.