A THREE-STAGE FRAMEWORK TO ACTIVE SOURCE LOCALIZATION FROM A BINAURAL HEAD

Gabriel Bustamante[⋆†] *Alban Portello*[⋆†] *Patrick Danès*[⋆†]

[⋆] CNRS, LAAS, 7 avenue du Colonel Roche, F-31400 Toulouse, France
[†] Univ de Toulouse, UPS, LAAS, F-31400 Toulouse, France

ABSTRACT

This paper takes place within the field of binaural localization in robotics. The aim is to design "active" schemes, which combine the signals sensed by a binaural head with its motor commands so as to overcome limitations occuring in a static context: front-back confusion, non-observability of hidden variables, etc. A three-stage strategy is proposed, which entails: the short-term detection and localization of sources from the short-term analysis of the binaural stream; the assimilation of these data over time and the fusion with the motor commands of the binaural sensor; the improvement of this fusion through the feedback control of the binaural sensor. For each stage, the theoretical bases, some achievements and open problems are outlined.

Index Terms— Robot audition, ML estimation, Kalman filtering, information-based control, active localization.

1. INTRODUCTION

Since its emergence in the 2000s, robot audition has raised an increasing interest within and from outside robotics. Indeed, the combination of audition with other modalities —embedded or deployed in the environment—as well as the mobility offered by an auditory robot open unexpected problems and rich perspectives [1][2]. The challenges in terms of devices and algorithms (embeddability, real-time performance, ego-noise…), source and environment features (human voice, noise, reverberation…), tasks (acoustic effects due to motion, barge-in situations…) enriches the field [3].

Within the renewed interest for "active" binaural functions, which combine the binaural perception with the motor commands of the sensor, active binaural localization offers the perspective of overcoming limitations in a static context such as front-back confusion, range non-observability, etc. A sensorimotor view of this last problem, which entails no decisional/cognitive process, is proposed in this paper. The way how binaural sensing and sensor motion can be interwoven is described along a three-stage framework, conceptualized on Figure 1. Stage A and Stage B carry out the analysis

This work was partially supported by EU FET grant TWO!EARS, ICT-618075, www.twoears.eu.

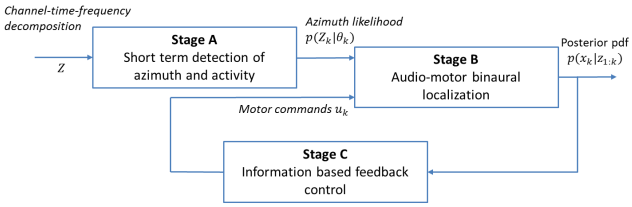

Fig. 1. Three-stage active binaural localization.

of the sensorimotor flow. Stage A ("Short-term detection") implements the estimation of the spatial arrangement of active sources—possibly with the detection of their number—from the analysis of the binaural stream over small time snippets. Stage B ("Audio-motor binaural localization") performs the assimilation of these data over time and their combination with the motor commands of the sensor, so as to get a first level of active localization. On this basis, Stage C ("Information-based feedback control") implements a feedback control of the sensor motion so as to improve the fusion performed in Stage B.

The following two main sections present Stage A and Stage B along a common organization: first, modeling issues are briefly sketched; then, the current state of the solutions for a binaural head are outlined; last, evaluations are shown. A conclusion discusses open problems for all three stages.

2. SHORT-TERM DETECTION

2.1. Modeling

The two microphones placed on the binaural head are named R_1 and R_2. They lie in the same horizontal plane as Q pointwise farfield sound sources $E_1, \dots, E_Q$. The aim is to estimate the source azimuths $\theta_1, \dots, \theta_Q$—*e.g.*, with respect to boresight—from the analysis of the binaural stream. To conduct this analysis, the left and right signals z_{R_1}, z_{R_2} are assumed to be finite-time samples of the random processes

$$\mathtt{z}_{R_1}(t)=\sum_{q=1}^{Q}\mathtt{s}_q(t)+\mathtt{n}_1(t),\ \mathtt{z}_{R_2}(t)=\sum_{q=1}^{Q}(\mathtt{s}_q(t)*h_{\theta_q}(t))+\mathtt{n}_2(t). \quad (1)$$

In (1), the noises $\mathtt{n}_1, \mathtt{n}_2$ and the contributions $\mathtt{s}_1, \dots, \mathtt{s}_Q$ of the emitters $E_1, \dots, E_Q$ at microphone R_1 are assumed real, zero-mean band-limited, jointly Gaussian. The noise vector

$\mathcal{T}$ time inverval $= N_g$ groups of N_f (possibly overlapped) L-sample frames	
$R_1, R_2 \bullet E_1, \ldots, E_Q$	left and right microphones $\bullet$ Q sources
$\mathtt{s}_1(t), \ldots, \mathtt{s}_Q(t)$	contributions of $E_1, \ldots, E_Q$ at R_1
$\mathtt{z}_R = [\mathtt{z}_{R_1}, \mathtt{z}_{R_2}]' \bullet \mathtt{n} = [\mathtt{n}_1, \mathtt{n}_2]'$	signals and noises vectors at R_1, R_2
$H_\theta(f) \bullet V_\theta(f) = [1, H_\theta(f)]'$	interaural transf. funct. $\bullet$ steering vector
$\mathtt{Z}_{R\,(n_g,n_f)}[k]$	Fourier transform of pre-windowed $\mathtt{z}_R$ on frame n_f of group n_g for $f = \frac{k}{L}$
$C_{n_g}[k]$	PSD of $\mathtt{z}_R$ on group n_g for $f = \frac{k}{L}$

Fig. 2. Notations

$[\mathtt{n}_1, \mathtt{n}_2]'$—with $'$ the transpose operator— is assumed independent of each $\mathtt{s}_q$. The interaural impulse response $h_\theta(t)$ accounts for scattering, and depends on the source azimuth θ. Its Fourier transform is the interaural transfer function $H_\theta(f)$, and $V_\theta(f) = [1, H_\theta(f)]'$ is the so-called steering vector.

The input data to Stage A is the channel-time-frequency decomposition of $z_R \triangleq [z_{R_1}, z_{R_2}]'$ on a given finite time interval $\mathcal{T}$. Each time signal z_{R_1}, z_{R_2} is divided into—possibly overlapping—frames of L samples each. N_f consecutive frames constitute a group of frames, and $\mathcal{T}$ corresponds to N_g groups. Each frame is modulated by a window function, then Discrete Fourier transformed. Let $X_{(n_g,n_f)}[k]$ stand for the value at $f = \frac{k}{L}$ of the transform $X_{(n_g,n_f)}(f)$ of the n_f^{th} frame in the n_g^{th} group of a pre-windowed signal $x(t)$. The channel-time(group of frames)-frequency decomposition Z of the binaural stream over $\mathcal{T}$ is then the stacking of the complex vectors $Z_{R\,(n_g,n_f)}[k_b]$ for $n_f = 1, \ldots, N_f$, $n_g = 1, \ldots, N_g$, and $b = 1, \ldots, B$, where the last subscripts define a "useful" frequency range. The variations of the means, autocorrelations and cross-correlations of the signals involved in (1) are assumed negligible over each group of frames. So, relative motion between the sensor and the emitters should be negligible as well.

2.2. The single-source case ($Q = 1$)

In view of the above, though the source signal $\mathtt{s} \triangleq \mathtt{s}_1$ is not wide sense stationary (WSS) over the whole interval $\mathcal{T}$, a power spectral density (PSD) $S_{n_g}[k]$ corresponding to each n_g^{th} group of frames can be defined. Also suppose that: a Hann (resp. rectangular) window function with less than 50% frame overlap (resp. with no overlap) is used; the source and noise spectra are roughly constant over any $\frac{1}{L}$-width frequency range; the autocorrelation time of $h_\theta(t)$ is much lower than L. Then the Fourier coefficients of the pre-windowed binaural signal vector $\mathtt{z}_R = [\mathtt{z}_{R_1}, \mathtt{z}_{R_2}]'$ can be shown to be mutually independent at distinct frequencies or frames, and to satisfy $\mathbb{E}\{\mathtt{Z}_{R\,(n_g,n_f)}[k]\mathtt{Z}_{R\,(n_g,n_f)}[k]^\dagger\} = C_{n_g}[k]$, with $\dagger$ the Hermitian transpose and $C_{n_g}[k]$ the PSD matrix of the random process $\mathtt{z}_R$ corresponding to group n_g evaluated at $f = \frac{k}{L}$. Let $\bar{C}_{n_g}[k] \triangleq \frac{1}{N_f}\sum_{n_f=1}^{N_f} Z_{R\,(n_g,n_f)}[k]Z_{R\,(n_g,n_f)}[k]^\dagger$ be the sample covariance matrix over the time(group of frames)-frequency bin (n_g, k). Then, the maximum likelihood estimate (MLE) of the source azimuth θ is given by the following theorem, assuming $\mathtt{n} = [\mathtt{n}_1, \mathtt{n}_2]'$ is i.i.d. with PSD $C_{\mathtt{n}}[k] = \sigma^2[k]\mathbb{I}_2$ (extension to unknown $\{\sigma^2[k_b]\}_b$ is easy).

Theorem 1 *If $\{\bar{C}_{n_g}[k_b]\}_{n_g,b}$ are full-rank, then the MLE $\hat{\theta}_{ML}$ comes as the* $\arg\max$ *w.r.t. θ of the* pseudo log-likelihood

$$L(\theta) = \sum_{n_g=1..N_g\,;\,b=1..B} J_{n_g}[k_b](\theta), \text{ with} \tag{2}$$

$$J_{n_g}[k_b](\theta) = -N_f\Big(\ln\big|P_\theta[k_b]\bar{C}_{n_g}[k_b]P_\theta[k_b] + \sigma^2[k_b]P_\theta^\perp[k_b]\big| + \frac{1}{\sigma^2[k_b]}\mathrm{tr}(P_\theta^\perp[k_b]\bar{C}_{n_g}[k_b])\Big),$$

$$P_\theta[k] \triangleq V_\theta[k](V_\theta[k]^\dagger V_\theta[k])^{-1}V_\theta[k]^\dagger \text{ and } P_\theta^\perp[k] \triangleq \mathbb{I}_2 - P_\theta[k].$$

Proof This is an adaptation of [4] to the broadband case, considering a single source and two sensors related by H_θ instead of a freefield array, and allowing source autocorrelation changes along groups of frames. First, $\{C_{n_g}[k_b]\}_{n_g,b}$ is written as a function of $C_{\mathtt{n}}[k]$ and the unknowns vector $\Theta = [\theta, \{S_{n_g}[k_b]\}_{n_g,b}]'$, which entails $V_\theta[k]$. Then, $p(Z|\Theta)$ is expressed as a circular complex Gaussian pdf. Though its $\arg\max \hat{\Theta}_{\text{ML}}$ w.r.t. Θ has no closed form, the problem is separable, *i.e.*, the MLEs $\{\hat{S}_{n_g}[k_b]\}_{n_g,b}$ of $\{S_{n_g}[k_b]\}_{n_g,b}$ can be expressed as functions of θ. It follows that

$$\max_\Theta \ln p(Z|\Theta) = \max_\theta \underbrace{\ln p\big(Z|\theta, \{\hat{S}_{n_g}[k_b](\theta)\}_{n_g,b}\big)}_{=\text{constant}+L(\theta)}$$

hence the result and the "pseudo likelihood" terminology. □

Moreover, the source activity can be checked through AIC or BIC information-theoretic criteria [5].

2.3. The multiple-source case

Turning back to (1), where each signal $\mathtt{s}_q$ is WSS over each n_g^{th} group of frames and has "local" autocorrelation $R_{n_g}^{(q)}(\tau)$ and PSD $S_{n_g}^{(q)}(f)$, the vector of unknowns $\Theta = [\Theta_1', \ldots, \Theta_Q']'$ is constituted of Q $(1 + N_gB)$-element subvectors $\Theta_q = [\theta_q, \{S_{n_g}^{(q)}[k_b]\}_{n_g,b}]'$ similar to the single-source case. The maximum likelihood estimation is no longer separable. To make it tractable, W-Disjoint Orthogonality is assumed, *i.e.*, within any time(groups of frames)-frequency bin (n_g, k), at most one source is dominant. Assuming that the prior probability of source dominance is evenly distributed on each bin, $p(Z|\Theta)$ now takes the form of a mixture—over the source indexes—of circular complex Gaussian pdfs.

The extraction of the MLE $\hat{\theta}_{\text{ML}} = [\hat{\theta}_1, \ldots, \hat{\theta}_Q]'$ of the azimuths vector $\theta = [\theta_1, \ldots, \theta_Q]'$ can be run iteratively through the Expectation-Maximization algorithm [6], by introducing the vector of latent random variables $Y \triangleq [\{Y_{n_g}[k_b]\}_{n_g,b}]'$ such that $Y_{n_g}[k_b] = q$ iff the q^{th} source is dominant on bin (n_g, k_b)—*i.e.*, is at the origin of $Z_{n_g}[k_b]$—and by assuming mutual independence of $\{Y_{n_g}[k_b]\}_{n_g,b}$ and $\{X_{n_g}[k_b]\}_{n_g,b}$.

Theorem 2 *Given a number Q of active sources, the MLE $\hat{\theta}_{ML}$ of the azimuths vector θ can be obtained from an initial guess $\theta^{(init)}$ by iterating* [E-step,M-step] *sequences shown in Algorithm 1 until* Stop condition *holds. Importantly, the most*

Algorithm 1: Multiple-source azimuth estimation (one iteration of the EM algorithm)

Inputs: Initial guess θ^* issued from the previous iteration
Outputs: Most likely $\hat{\theta}$ generated by the current iteration

E-step ("Source separation")

for $n_g = 1, \ldots, N_g,\ b = 1, \ldots, B$ **do**
  For $q = 1, \ldots, Q$ **do** $\bar{\gamma}_{n_g}^{(q)}[k_b] = \exp\left(J_{n_g}[k_b](\theta_q^*)\right)$ **end**
  $\gamma_{n_g}^{(q)}[k_b] = \frac{\bar{\gamma}_{n_g}^{(q)}[k_b]}{\sum_{\ell=1}^{Q} \bar{\gamma}_{n_g}^{(\ell)}[k_b]}$
end

M-step ("Source localization")

for $q = 1, \ldots, Q$ **do**
  $\hat{\theta}_q = \arg\max_{\vartheta} \sum_{n_g=1..N_g\,;\,b=1..B} \gamma_{n_g}^{(q)}[k_b]\, J_{n_g}[k_b](\vartheta)$
end

Log-Likelihood computation and Stop condition

$$L(\hat{\theta}) = \sum_{n_g=1..N_g\,;\,b=1..B} \ln\Big(\sum_q \frac{1}{Q} J_{n_g}[k_b](\hat{\theta}_q)\Big)$$

The algorithm stops when $\Delta L = \frac{L(\hat{\theta}) - L(\theta^*)}{L(\theta^*)} < \eta$, for given η.

likely azimuth of each source comes from a separate maximization, and no initial guess is needed for the sources PSDs.

Proof The proof is omitted for space reasons, see [7]. Compared with Theorem 1, the auxiliary function maximized in the M-step involves the weights $\{\gamma_{n_g}^{(q)}[k_b]\}_{n_g,b}$ computed in the E-step. Each one equals $\mathbb{P}\{Y_{n_g}[k_b]{=}q|z_{n_g}[k_b], \theta^*, S_{n_g}[k_b]^*\}$, the probability that the q^{th} source is dominant in the bin (n_g, k_b), conditioned on the data and the "naive" hypothesis that the azimuth vector is θ^*—from which the most likely corresponding source PSDs $\{S_{n_g}^{(q)}[k_b]\}_{q,n_g,b}^*$ are determined thanks to a "local" separability. So, no initial guess is required for the source PSDs. Besides, the global computational cost is linear with the number of sources. □

2.4. Evaluations

Theorem 2 was extensively evaluated in simulation. Binaural signals were synthesized from the TUB anechoic KEMAR® HRIR database [8] with 1° resolution. The emitted signals were 15 seconds-long male and female speakers utterance records from french radio, sampled at 44.1 kHz. Some stationary fan noise at predefined azimuth was added to the ear signals by convolving it with the left and rights HRIRs.

FFTs were performed on Hanning windowed frames of $L = 1024$ samples, with a $L/2$-overlap. Sample covariance matrices $\{\bar{C}_{n_g}[k_b]\}_{n_g,b}$ were computed from groups of $N_f = 4$ successive frames (≈ 60 ms). N_g was set to 50, so $\mathcal{T} \approx 3$ s. B was defined so that the useful frequency range is 7 kHz. To reduce the risk of convergence to local minima of the log-likelihood, 20 instances of Algorithm 1, with different initializations, were running in parallel at each localization step, and the most likely estimate was then kept.

About 90% of the azimuths were estimated with errors less than 3° for $Q \leq 3$ active sources. Performances were degraded in reverberant environments, yet good results could be recovered by replacing anechoic HRIRs/HRTFs by BRIRs.

3. AUDIO-MOTOR BINAURAL LOCALIZATION

This section adresses active/audio-motor binaural localization in the single-source case. It is shown how an azimuth likelihood defined from above can be combined with the motor commands of the head so as to infer its relative situation to a static source. A Gaussian mixture square-root unscented Kalman filter (GM-srUKF) is advocated. Contrarily to several particle filters, it ensures self-initialisation as well as posterior covariance consistency. Handling of false measurements and source intermittency are reported in [9].

3.1. Modeling

From now on, the scalar k indexes the localization time, *e.g.*, the timestamp of a group of frames, and T_s terms the localization period. The measurement is still the channel-time-frequency decomposition but is now denoted by Z_k. As the spatial information carried by Z_k is purely directional, the 2-dimensional state vector $\mathtt{r}_k$ to be estimated is the relative head-to-source translation, expressed in polar coordinates. The 3-dimensional control input u_k of the sensor is the stacking of its translation and rotation velocities. An exact discrete-time nonlinear state space equation with Gaussian dynamic noise can be obtained, considering that u is zero-order-held at T_s. However, no closed-form measurement equation is available. Instead a likelihood $p(Z|\theta)$ can be built from Section 2, where θ stands for the head-to-source azimuth in $\mathtt{r}_k$. An approximation of $p(Z|\theta)$ is assumed to be determined in an *ad hoc* way, in the form of the following unnormalized mixture with parameters $\{\gamma^j, m^j, \phi^j\}_{j=1,\ldots,J}$:

$$p(Z|\theta) \approx \sum_{j=1}^{J} \gamma^j \mathrm{e}^{-\frac{1}{2}\frac{\left(\theta - m^j\right)^2}{\phi^j}}. \tag{3}$$

3.2. A Gaussian mixture unscented Kalman filter

In view of the nonlinearity of the prior state dynamics and the frequent multimodality of the likelihood—*e.g.*, due to front-back ambiguity—a bank of unscented Kalman filters (UKFs) is used, implemented in their numerically robust square-root form [10]. After assimilating Z_k, this bank contains I_k filters, each i^{th} one handling a Gaussian distribution $\mathcal{N}(r_k; \hat{r}_{k|k}^i, P_{k|k}^i)$. The filters run in a non-interactive manner, but their posterior probabilities $\{w_k^i\}$ are recursively updated in view of their likelihoods w.r.t. the measurements—*i.e.*, of their ability to predict the available data. By developing classical computations [11][12], one gets the following theorem.

Theorem 3 *The posterior pdf of* $\mathtt{r}_k$ *is approximated by*

$$p(r_k|\mathtt{Z}_{1:k} = Z_{1:k}) = \sum_{i=1}^{I_k} w_k^i \mathcal{N}(r_k; \hat{r}_{k|k}^i, P_{k|k}^i). \tag{4}$$

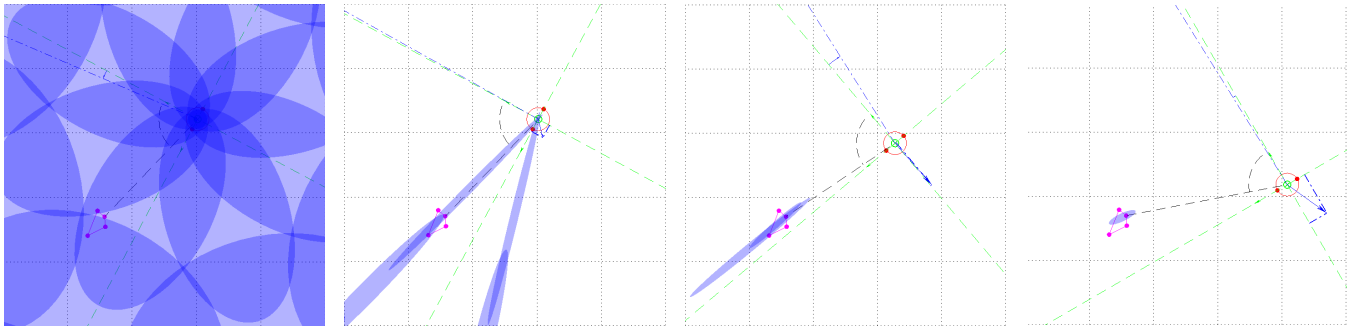

Fig. 3. Audio-motor localization of a loudspeaker (top right edge of the pentagon) by a moving spherical binaural head. (a): Self-initialization. (b): Front-back ambiguity when no head motion. (c)-(d): Disambiguation thanks to the head motion.

Its initalization—i.e., *the prior* $p(x_0)$*—satisfies* $w_k^0 = \frac{1}{I_k^0}$ *and is such that the union of the 99%-probability confidence ellipsoids corresponding to* $\{\mathcal{N}(\hat{r}_{0|0}^i, P_{0|0}^i)\}_{i=1,\dots,I_0}$ *covers the admissible initial head-to-source situations. The parameters* $\{w_k^i, \hat{r}_{k|k}^i, P_{k|k}^i\}$ *are updated according to Algorithm 2.*

Algorithm 2: Overview of audio-motor localization

Inputs: Parameters $\{\gamma_k^j, m_k^j, \phi_k^j\}_1^{J_k}$ of $p(Z_k|\theta_k)$
Parameters $\{w_{k-1}^i, \hat{r}_{k-1|k-1}^i, P_{k-1|k-1}^i\}_{i=1..I_{k-1}}$ of $p(r_k|Z_{1:k-1})$
Control vector u_{k-1}
Outputs: Parameters $\{w_k^i, \hat{r}_{k|k}^i, P_{k|k}^i\}_{i=1..I_k}$ of $p(r_k|Z_{1:k})$

Time update
for $i = 1, \dots, I_{k-1}$ **do**
  Compute the moments $\hat{r}_{k|k-1}^i, P_{k-1|k-1}^i$ of the i^{th} hypothesis of $p(r_k|Z_{1:k-1})$ from $\hat{r}_{k-1|k-1}^i, P_{k-1|k-1}^i$, the polar state space equation, and u_{k-1}, with a UKF time update (see [10]).
end

Measurement update
for $i = 1, \dots, I_{k-1}$ **do**
  for $j = 1, \dots, J_k$ **do**
    Fuse $\mathcal{N}(r_k; \hat{r}_{k|k-1}^i, P_{k|k-1}^i)$ with the j^{th} hypothesis of $p(Z_k|r_k) = p(Z_k|\theta_k)$ in (3) through the Bayes rule. From classical results on products of Gaussians, this yields $\mathcal{N}(r_k; \hat{r}_{k|k}^{i,j}, P_{k|k}^{i,j})$ (see [13]).
  end
end

Weights update (and possible pruning)
Compute the weights $\{w_k^{i,j}\}_{i,j}$ of the Gaussian mixture $p(r_k|Z_{1:k})$ from those of $p(r_{k-1}|Z_{1:k-1})$, $p(Z_k|r_k)$ (see [13]). The posterior finally writes as $p(r_k|Z_{1:k}) = \sum_{i,j} w_k^{i,j} \mathcal{N}(r_k; \hat{r}_{k|k}^{i,j}, P_{k|k}^{i,j})$. Prune hypotheses whose weights $w_k^{i,j}$ fall below a given threshold.

3.3. Evaluations and Open problems

Experiments were conducted in an anechoic room with a binaural sphere. Figure 3 shows in the world frame at four times (from left to right and top to bottom) how the head motion enables front-back disambiguation. The pentagon depicts the loudspeaker. The 99%-probability confidence ellipses are associated to each hypothesis of the posterior state pdf.

4. OPEN PROBLEMS

Concerning Stage A, a thorough evaluation is in progress. The algorithm will be compared, under reverberant conditions, to suboptimal but computationally faster aproaches such as the GCC-PHAT together with a time integration method (GCC averaging, histogram), or bio-inspired approaches [14]. The learning of the environment noise statistics will be studied. Multi-conditional training as per [14] will be evaluated to "desensitize" H_θ to reverberations. Influence of the algorithm's parameters $L, N_f, N_g, \dots$ will also be investigated. The detection of the number of multiple sources will be adressed.

Stage B will be extended to the multiple-source case. To address data association, filtering in the Random Finite Sets paradigm will be studied [15][16].

Ongoing developments concern Stage C. To simplify, consider the case when the posterior pdf (4) at time $k-1$ reduces to a single Gaussian, and a control input u_{k-1} is sought so as to maximize the information at time k while respecting constraints on u_{k-1}, x_k, etc. The problem then boils down to maximizing the "size" of the information matrix $I_{k|k} = P_{k|k}^{-1}$, through the maximization of its log-determinant or trace for instance [12][17]. When using the Unscented Kalman Filter with a closed-form measurement equation $\mathbf{z}_k = h(\mathbf{r}_k) + \mathbf{v}_k$, an information update equation can be set up in the form $I_{k|k} = I_{k|k-1} + \mathcal{H}_k^T R^{-1} \mathcal{H}_k$ where both $I_{k|k-1}$ and $\mathcal{H}_k$ depends on the decision variable u_{k-1} and on the function $h(.)$, but not on the measurement z_k [18]. Current work consists in selecting a function $h(.)$ to guide the exploration, and in rephrasing the control problem as a convex optimization problem consisting in the maximization of the log-determinant of $I_{k|k}$ subject to linear matrix inequalities constraints on u_{k-1} [19]. To reach this aim, the fact that the prior dynamics is a rigid motion is used together with changes of variables and embedding of nonlinearities into uncertainties. The challenge is to check if the induced conservativeness does not prevent the applicability of this MAXDET solution to our genuine real problem.

5. REFERENCES

[1] K. Nakadai, T. Lourens, H.G. Okuno, and H. Kitano, "Active audition for humanoid," in *Nat. Conf. on Artificial Intelligence (AAAI'2000)*, Austin, TX, 2000.

[2] M. Cooke, Y.C. Lu, Y. Lu, and R. Horaud, "Active hearing, active speaking," in *Int. Symp. on Auditory and Audiological Research (ISAAR'07)*, Marienlyst, Helsigør, Denmark, 2007.

[3] S. Argentieri, A. Portello, M. Bernard, P. Danès, and B. Gas, "Binaural systems in robotics," in *The Technology of Binaural Listening*, J. Blauert, Ed., pp. 225–254. Springer, 2013.

[4] A.G. Jaffer, "Maximum likelihood direction finding of stochastic sources: a separable solution," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'1988)*, New York, NY, 1988.

[5] D. Williams, "Detection: Detecting the number of sources," in *The Digital Signal Processing Handbook*, V.K. Madisetti and D.B. Williams, Eds., chapter 67. CRC Press, 1999.

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Jour. of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[7] A. Portello, G. Bustamante, P. Danès, and A. Mifsud, "Localization of multiple sources from a binaural head in a known noisy environment," in *IEEE/RSJ Int. Conf. on Intell. Robots and Systems (IROS'2014)*, Chicago, IL, 2014.

[8] H. Wierstorf, M. Geier, and S. Spors, "A free database of head related impulse response measurements in the horizontal plane with multiple distances," in *Audio Engineering Society Convention 130*, 2011.

[9] A. Portello, P. Danès, and S. Argentieri, "Active binaural localization of intermittent moving sources in the presence of false measurements," in *IEEE/RSJ Int. Conf. on Intell. Robots and Systems (IROS'2012)*, 2012.

[10] R. Van der Merwe and E.A. Wan, "The square-root unscented Kalman filter for state and parameter-estimation," in *IEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'2001)*, 2001.

[11] B.D.O. Anderson and J.B. Moore, *Optimal filtering*, Englewood Cliffs, N.J. Prentice-Hall, 1979.

[12] Y. Bar-Shalom and Xiao-Rong Li, *Estimation and Tracking : Principles, Techniques, and Software*, Artech House, 1998.

[13] A. Portello, G. Bustamante, P. Danès, J. Piat, and J. Manhès, "Active localization of an intermittent sound source from a moving binaural sensor," in *Forum Acustium (FA'2014)*, Krakow, Poland, 2014.

[14] T. May, S. van de Par, and A. Kohlrausch, "Binaural localization and detection of speakers in complex acoustic scenes," in *The Technology of Binaural Listening*, J. Blauert, Ed., pp. 397–425. Springer, 2013.

[15] A. Masnadi-Shirazi and B. Rao, "An ICA-SCT-PHD filter approach for tracking and separation of unknown time-varying number of sources," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 21, no. 4, pp. 828–841, 2013.

[16] W.-K. Ma, B.-N. Vo, S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach," *IEEE Trans. on Signal Processing*, vol. 54, no. 9, pp. 3291–3304, 2006.

[17] B. Grocholsky, *Information-Theoretic Control of Multiple Sensor Platforms*, Ph.D. thesis, Univ. of Sydney, 2006.

[18] D.J. Lee, "Nonlinear estimation and multiple sensor fusion using unscented information filtering," *IEEE Signal Processing Letters*, vol. 15, pp. 861–864, 2008.

[19] L. Vandenberghe, S. Boyd, and S. Wu, "Determinant maximization with linear matrix inequality constraints," *SIAM Jour. on Matrix Analysis and Applications*, vol. 19, pp. 499–533, 1998.