2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

HRTF-Based Source Azimuth Estimation and Activity Detection from a Binaural Sensor

Alban Portello, Patrick Danès, Sylvain Argentieri and Sylvain Pledel

***Abstract*— A theoretically grounded scheme to Direction of Arrival (DOA) estimation and Source Activity Detection (SAD) is proposed, on the basis of a pair of microphones. The method can capture the effects of the robot's scatterers, if any. The DOA estimator takes place within a probabilistic framework and outputs the Maximum Likelihood Estimate (MLE) of the DOA with respect to the collected audio data. Besides, the SAD relies on statistical identification. The behavior of the estimator is studied under various operating modes, considering free-field propagation and scattering by a rigid spherical head. Experimental results validate the approach.**

I. INTRODUCTION

The estimation of the Direction Of Arrival (DOA) of a broadband sound source from a pair of microphones has been widely dealt with. In most contributions, an estimation of the Time Delay Of Arrival (TDOA) between the microphones is first performed [1][2]—assuming no reflector nor scatterer—then the DOA is deduced, up to a front-back ambiguity, under the free-field assumption. However, in the context of humanoid robotics, the microphones are mounted on a head—possibly inside artificial pinnae—so that the free-field assumption no longer holds. Due to the scattering effect of the robot's head and torso, the phase difference between the two perceived signals is no longer linear w.r.t. frequency, and the amplitude difference becomes frequency dependent. Rigorously speaking, TDOA has thus no physical meaning.

Solutions to DOA estimation in the humanoid robotics context were recently proposed in [3][4], on the basis of the Interaural Level and Phase/Time Differences (ILD, IPD/ITD) computed from the recorded signals, together with a Head Related Transfer Function (HRTF) lookup database. In [5], the mapping between source positions and theoretical binaural cues is learnt beforehand experimentally from a human-like dummy head endowed with two microphones.

In this paper, a binaural head samples the wave induced by at most one broadband source into noise. Under some probabilistic assumptions, the Maximum Likelihood Estimate (MLE) of the source DOA w.r.t. the collected data is computed. It entails the transfer function between the two microphones in order to capture the effects of scatterers. Hence, this DOA estimator is asymptotically efficient. Contrarily to [3][4], it weights the frequency contributions according to the SNR, so it can be less sensitive to outliers appearing at low Signal-to-Noise Ratio (SNR) frequencies. In addition, it can cope with prior information about the environment noise. The Source Activity Detection (SAD) scheme relies on the Akaike Information Criterion (AIC) [6][7][8].

The forthcoming Section II introduces the model and working hypotheses. The DOA estimator and SAD scheme are then set up in Section III. The behavior of the DOA estimator is studied in Section IV when the binaural sensor is in free-field or mounted on a rigid sphere. Tests on real data recorded in this last scenario constitute Section V.

II. MODEL DEFINITION, WORKING HYPOTHESES

Consider a pointwise sound emitter E and two receivers R_1, R_2, possibly laid on a head. Define a frame $\mathcal{F}_R$ attached to them, whose center R is the midpoint of $[R_1R_2]$, and denote $\boldsymbol{\theta}$ a vector of parameters characterizing the position of E w.r.t. $\mathcal{F}_R$. Let $h_{\boldsymbol{\theta}}$ stand for the impulse response, parameterized by $\boldsymbol{\theta}$, between R_1 and R_2. $h_{\boldsymbol{\theta}}$ can capture head scattering. The signals x_1, x_2 at R_1, R_2 write as

$$\begin{cases} x_1(t) = s(t) + n_1(t) \\ x_2(t) = (s * h_{\boldsymbol{\theta}})(t) + n_2(t), \end{cases} \quad (1)$$

where the contribution s of the emitter at R_1 and the additive noises n_1, n_2 are real, band-limited, individually and jointly stationary random processes, and $*$ denotes convolution.

The signals x_1, x_2 are observed over a finite time window[1] $I_T \triangleq \left[-\frac{T}{2}, \frac{T}{2}\right]$. Define the set $\{I_n\}_{n=1,...,N}$ of N equal-length non-overlapping segments (or snapshots) covering I_T as $I_n = \left[-\frac{T}{2} + \frac{(n-1)T}{N}, -\frac{T}{2} + \frac{nT}{N}\right]$, $n = 1, ..., N$. For $j = 1, 2$ and $n = 1, ..., N$, define $x_{j,n}$, the observation of x_j on I_n, as the product of x_j with a window w symmetric over its T/N-width support I_n, *i.e.*,

$$x_{j,n}(t) = x_j(t)w(t - \tau_n), \text{ with } \tau_n \triangleq -\frac{T}{2} + \frac{(n-\frac{1}{2})T}{N}. \quad (2)$$

If s, n_1, n_2 are zero-mean jointly Gaussian, then the random processes defined for $j = 1, 2$ and $n = 1, ..., N$ as

$$X_{j,n}(f) = \sqrt{\frac{N}{T}} \int_{\mathbb{R}} x_{j,n}(t)e^{-2i\pi ft}dt, \quad (3)$$

are zero-mean jointly (circular complex) Gaussian[2].

*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.

A. Portello and P. Danès are with CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France, Univ de Toulouse, UPS, LAAS ; F-31400 Toulouse, France. {aportell,danes}@laas.fr

S. Argentieri is with UPMC Univ. Paris 06, 4 place Jussieu, F-75005, Paris, France. S. Argentieri and S. Pledel are with CNRS, ISIR UMR 7222, 4 place Jussieu, F-75005, Paris, France. {argentieri,pledel}@isir.upmc.fr

[1]Symmetry of I_T around 0 simplifies maths with no loss of generality.

[2]If the zero-mean assumption does not hold but s, n_1, n_2 can be considered mean-ergodic (*i.e.*, their time averages over I_T converge in the mean square sense to their statistical expectations as $T \to +\infty$), the problem can still be handled by removing from x_1, x_2 their time averages over I_T.

For any continuous function of frequency $\phi(.)$, $\phi[k]$ henceforth stands for $\phi[k] \triangleq \phi(kf_N)$, with $f_N \triangleq \frac{N}{T}$, $k \in \mathbb{Z}$. Define the data vector $\boldsymbol{Z} \triangleq [\boldsymbol{X}[k_1]', ..., \boldsymbol{X}[k_B]']'$, with $k_1 f_N, ..., k_B f_N$ the set of B discrete frequencies within the bandwidth of s, $\boldsymbol{X}[k] \triangleq [\boldsymbol{X}_1[k]', ..., \boldsymbol{X}_N[k]']'$, and $\boldsymbol{X}_n[k] \triangleq [X_{1,n}[k], X_{2,n}[k]]'$.

Theorem 1. *If w is rectangular and $C(f)$, the Power Spectral Density (PSD) matrix of x_1, x_2, is smooth enough to be nearly constant over any N/T-width frequency interval, then the probability density function (pdf) of $\boldsymbol{Z}$ writes as*

$$p_{\boldsymbol{Z}}(\boldsymbol{z}; C[k_1], ..., C[k_B]) = \prod_{n,k} {}^{\mathbb{C}}\mathcal{N}(\boldsymbol{x}_n[k]; \boldsymbol{0}, C[k]), \quad (4)$$

where ${}^{\mathbb{C}}\mathcal{N}(.; \boldsymbol{m}, P)$ terms the multivariate circular complex Gaussian distribution of mean $\boldsymbol{m}$ and covariance matrix P, and $\boldsymbol{z}$ (resp. $\boldsymbol{x}_n[k]$) terms a sample of $\boldsymbol{Z}$ (resp. $\boldsymbol{X}_n[k]$).

Proof. Define, for $j, l \in \{1, 2\}$ and $m, n \in \{1, ..., N\}$,

$$R_{j,l}^{m,n}(f_1, f_2) \triangleq \mathbb{E}\{X_{j,n}(f_1) X_{l,m}(f_2)^*\}, \quad (5)$$

with $\mathbb{E}$ the mathematical expectation and * the complex conjugation. Injecting (3) in (5) and using the individual/joint stationarity of x_1, x_2 leads to, after some manipulations,

$$R_{j,l}^{m,n}(f_1, f_2) = \frac{N}{T} e^{-2i\pi(f_1 - f_2)\tau_n} \cdot \int_{\mathbb{R}} C_{j,l}(\nu + f_2) W(\nu) W(f_1 - f_2 - \nu) e^{-2i\pi\nu\tau_{m,n}} d\nu, \quad (6)$$

with W the Fourier transform of w, $C_{j,l}(f)$ the j^{th}-row l^{th}-column element of $C(f)$, and $\tau_{m,n} \triangleq \tau_m - \tau_n$. If C is smooth enough to be nearly constant over any frequency interval whose width is of the order of N/T (*i.e.*, the order of magnitude of the main lobe width of W), then

$$R_{j,l}^{m,n}(f_1, f_2) \approx \tfrac{N}{T} C_{j,l}(f_2) \overline{W}_{m,n}(f_1 - f_2) e^{-2i\pi(f_1 - f_2)\tau_n}, \quad (7)$$

with $\overline{W}_{m,n}(f)$ the Fourier transform of $w(t)w(t - \tau_{m,n})$. Since the segments $\{I_n\}_{n=1,...,N}$ do not overlap, (7) implies that $R_{j,l}^{m,n}(f_1, f_2)$ is null whatever f_1, f_2, j, l if $m \neq n$. When $m = n$, if w is taken to be a rectangular window then under the above assumptions,

$$R_{j,l}^{n,n}(f_1, f_2) \approx \operatorname{sinc}\big((f_1 - f_2)\tfrac{T}{N}\big) C_{j,l}(f_2) e^{-2i\pi(f_1 - f_2)\tau_n}, \quad (8)$$

with $\operatorname{sinc}(x) \triangleq \frac{\sin(\pi x)}{\pi x}$. Hence, when only integer multiples of f_N are considered, $R_{j,l}^{n,n}(f_1, f_2) \approx 0$ for $f_1 \neq f_2$, whatever j, l. One finally has

$$R_{j,l}^{m,n}[k_1, k_2] \approx \begin{cases} C_{j,l}[k_2] & \text{if } k_1 = k_2, m = n \\ 0 & \text{otherwise,} \end{cases} \quad (9)$$

so that the covariance matrices of $\boldsymbol{X}[k]$ and $\boldsymbol{Z}$ have the block-diagonal forms $C_{\boldsymbol{X}[k]} = \operatorname{blkdiag}(C[k], ..., C[k])$ and $C_{\boldsymbol{Z}} = \operatorname{blkdiag}(C_{\boldsymbol{X}[k_1]}, ..., C_{\boldsymbol{X}[k_N]})$. As uncorrelated Gaussian vectors are independent, Eq. (4) is deduced. □

When the noise is spatially white, uncorrelated with s and has the same PSD $\sigma^2[k]$, $k = k_1, ..., k_B$, at each receiver, the PSD matrix C can be expressed as follows

$$C[k] = \boldsymbol{V}_{\boldsymbol{\theta}}[k] \mathcal{S}[k] \boldsymbol{V}_{\boldsymbol{\theta}}[k]^\dagger + \sigma^2[k] \mathbb{I}_2, \quad (10)$$

with † the Hermitian transpose operator, $\mathcal{S}$ the PSD of s, $\boldsymbol{V}_{\boldsymbol{\theta}}[k] \triangleq [1, H_{\boldsymbol{\theta}}[k]]'$, $H_{\boldsymbol{\theta}}$ the Fourier transform of $h_{\boldsymbol{\theta}}$ and $\mathbb{I}_2$ the 2×2 identity matrix. Note that $H_{\boldsymbol{\theta}}$ comes as the ratio of the HRTFs to the two microphones (whatever the selected reference point). The matrix $C_S[k] \triangleq \boldsymbol{V}_{\boldsymbol{\theta}}[k] \mathcal{S}[k] \boldsymbol{V}_{\boldsymbol{\theta}}[k]^\dagger$ is positive semi-definite and rank one, so that $C[k]$ is positive definite, hence invertible. From (10), one can obtain sufficient conditions for C to be nearly constant on any N/T-width frequency interval to validate the approximation (7): the bandwidths of s and n must be large compared to N/T, $|H_{\boldsymbol{\theta}}|^2$ must vary slowly on any interval of length N/T, and one must have $\max_f |\tau_{\boldsymbol{\theta}}(f)| \ll T/N$, with $\tau_{\boldsymbol{\theta}} \triangleq -\frac{1}{2\pi} \frac{\partial}{\partial f} \arg H_{\boldsymbol{\theta}}$ the group delay of $H_{\boldsymbol{\theta}}$. Those conditions extend these described in [9], and are hereafter assumed to hold.

The log-likelihood of the parameter vector $\boldsymbol{\Theta} \triangleq [\boldsymbol{\theta}', \sigma^2[k_1], ..., \sigma^2[k_B], \mathcal{S}[k_1], ..., \mathcal{S}[k_B]]'$ w.r.t. the measurement vector $\boldsymbol{Z}$ can now be derived, together with an information-theoretic SAD scheme.

III. Source localization and detection

A. Maximum likelihood source localization

Suppose $H_{\boldsymbol{\theta}}$ is known for all $\boldsymbol{\theta}$. The aim is to select the most likely value of $\boldsymbol{\theta}$ given a sample of $\boldsymbol{Z}$. Taking the logarithm of (4), one gets for the log-likelihood of $\boldsymbol{\Theta} \triangleq [\boldsymbol{\theta}', \sigma^2[k_1], ..., \sigma^2[k_B], \mathcal{S}[k_1], ..., \mathcal{S}[k_B]]'$:

$$L(\boldsymbol{Z}; \boldsymbol{\Theta}) = c_0 - N \sum_k \Big(\ln|C[k]| + \operatorname{tr}(C[k]^{-1} \hat{C}[k]) \Big), \quad (11)$$

with $c_0 \triangleq -2NB\ln(\pi)$ a constant, $\operatorname{tr}(.)$ and $|.|$ the trace and determinant, and $\hat{C}[k] \triangleq \frac{1}{N} \sum_n \boldsymbol{X}_n[k] \boldsymbol{X}_n[k]^\dagger$ the Sample Covariance Matrix (SCM) of x_1, x_2 at frequency index k.

When $\mathcal{S}$ and σ^2 are unknown, searching for the maximum value of L given a sample of $\boldsymbol{Z}$ involves a maximization procedure over $N_\Theta \triangleq N_{\boldsymbol{\theta}} + 2B$ parameters, with $N_{\boldsymbol{\Theta}}, N_{\boldsymbol{\theta}}$ the lengths of $\boldsymbol{\Theta}, \boldsymbol{\theta}$ respectively. However, it can be shown that $\hat{\boldsymbol{\theta}}_{ML}$, the vector gathering the first $N_{\boldsymbol{\theta}}$ entries of the maximum likelihood estimate $\hat{\boldsymbol{\Theta}}_{ML}$ of $\boldsymbol{\Theta}$, can be obtained by maximizing a function of $N_{\boldsymbol{\theta}}$ variables only, by carrying out analytically some first-order stationarity conditions[3] on L. This is summarized in the following theorem

Theorem 2. *If $\hat{C}[k]$ is full-rank whatever $k \in \{k_1, ..., k_B\}$, then the MLE $\hat{\boldsymbol{\theta}}_{ML}$ is obtained by maximizing*

$$J_2(\boldsymbol{\theta}) = c_2 - N \sum_k \Big(\ln|P_{\boldsymbol{\theta}}[k] \hat{C}[k] P_{\boldsymbol{\theta}}[k] + \hat{\sigma}^2_{\boldsymbol{\theta}}[k] P^\perp_{\boldsymbol{\theta}}[k]| \Big), \quad (12)$$

with $P_{\boldsymbol{\theta}}[k] \triangleq \boldsymbol{V}_{\boldsymbol{\theta}}[k] (\boldsymbol{V}_{\boldsymbol{\theta}}[k]^\dagger \boldsymbol{V}_{\boldsymbol{\theta}}[k])^{-1} \boldsymbol{V}_{\boldsymbol{\theta}}[k]^\dagger = P_{\boldsymbol{\theta}}[k]^\dagger = P_{\boldsymbol{\theta}}[k]^2$ the orthogonal projector onto the space spanned by $\boldsymbol{V}_{\boldsymbol{\theta}}[k]$, $P^\perp_{\boldsymbol{\theta}}[k] \triangleq \mathbb{I}_2 - P_{\boldsymbol{\theta}}[k]$ the orthogonal complement of $P_{\boldsymbol{\theta}}[k]$, $\hat{\sigma}^2_{\boldsymbol{\theta}}[k] \triangleq \operatorname{tr}(P^\perp_{\boldsymbol{\theta}}[k] \hat{C}[k])$, and $c_2 \triangleq c_0 - 2NB$.

Proof. The proof is omitted for space reasons. It adapts [10] to the considered broadband single-source case. The "pseudo log-likelihood" (12) also appears in [11], though restricted to free-field propagation. □

When the noise covariance matrix satisfies $C_N[k] = \sigma^2[k] \tilde{C}_N[k] \neq \sigma^2[k] \mathbb{I}_2$, with $\tilde{C}_N[k] = S_N[k] S_N[k]^\dagger$ some

[3] These conditions are indeed sufficient to get a maximum of the criterion.

known positive definite matrix, $S_N[k]$ its square-root—in the sense of its Choleski decomposition—, and $\sigma^2[k]$ an unknown scaling factor, the covariance matrix of $\tilde{\boldsymbol{X}}_n[k] \triangleq S_N[k]^{-1}\boldsymbol{X}_n[k]$ comes as $\tilde{C}[k]=\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}[k]\mathcal{S}[k]\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}[k]^\dagger+\sigma^2[k]\mathbb{I}_2$, with $\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}[k] \triangleq S_N[k]^{-1}\boldsymbol{V}_{\boldsymbol{\theta}}[k]$. Whenever the unknown σ^2 is frequency independent, the algorithm must be slightly modified [11], in that the noise PSD estimate has to be turned into $\hat{\sigma}^2_{\boldsymbol{\theta}} = \frac{1}{B}\sum_k \mathrm{tr}(P^{\perp}_{\boldsymbol{\theta}}[k]\hat{C}[k])$. The ML localization steps are shown in Algorithm 1.

Algorithm 1 The DOA MLE and SAD algorithm

$[J_2, \hat{\boldsymbol{\theta}}_{ML}, \hat{d}_{AIC}] = \mathsf{LOC\_DETECT}[x_1(I_T), x_2(I_T), \boldsymbol{V}_{\boldsymbol{\theta}}, S_N^{-1}]$

Assuming mean-ergodicity of the signals, make $x_1(I_T), x_2(I_T)$ zero-mean by removing them their time averages over I_T.
FOR $j = 1, 2$ **DO**
  Using a short-time Fourier transform algorithm (STFT) based on non-overlapping rectangular windows, compute the time-frequency representation of x_j: $X_{j,1,\ldots,N}[k_1, ..., k_B] = \mathsf{STFT}[x_j(I_T)]$.
END FOR
FOR $k = k_1, \ldots, k_B$ **DO**
  Define $\boldsymbol{X}_{1,\ldots,N}[k] \triangleq [X_{1,1,\ldots,N}[k]', X_{2,1,\ldots,N}[k]']'$.
  Perform the transform $\tilde{\boldsymbol{X}}_{1,\ldots,N}[k]=S_N[k]^{-1}\boldsymbol{X}_{1,\ldots,N}[k]$.
  Compute the Sample Covariance Matrix (SCM): $\hat{C}[k]=\frac{1}{N}\tilde{\boldsymbol{X}}_{1,\ldots,N}[k]\tilde{\boldsymbol{X}}_{1,\ldots,N}[k]^\dagger$.
  Perform an eigendecomposition of the SCM $\hat{C}[k]$: $[\tilde{\boldsymbol{U}}_{1,k}, \tilde{\boldsymbol{U}}_{2,k}, l_{1,k}, l_{2,k}] = \mathsf{EIG}[\hat{C}[k]]$, with $l_{1,k} \geq l_{2,k}$.
  FOR $\boldsymbol{\theta} = \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta}_{N_{\boldsymbol{\theta}}}$ **DO**
    Perform the transform $\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}[k] = S_N[k]^{-1}\boldsymbol{V}_{\boldsymbol{\theta}}[k]$.
    Compute the orthogonal projector onto the space spanned by $\tilde{\boldsymbol{V}}_{\theta}[k]$: $P_{\boldsymbol{\theta}}[k] \triangleq \tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}[k](\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}[k]^\dagger\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}[k])^{-1}\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}}[k]^\dagger$.
  END FOR
END FOR
FOR $d = 0, 1$ **DO**
  Compute the Akaike Information Criterion $\mathrm{AIC}(d)$ at d according to (16)–(17)–(18).
END FOR
FOR $\boldsymbol{\theta} = \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_{\boldsymbol{\theta}}}$ **DO**
  Compute the pseudo log-likelihood at $\boldsymbol{\theta}$ according to (12).
END FOR
Output the MAICE of d: $\hat{d}_{AIC} \triangleq \mathrm{argmin}_d \mathrm{AIC}(d)$.
(if $d \neq 0$) Output the MLE of $\boldsymbol{\theta}$: $\hat{\boldsymbol{\theta}}_{ML} \triangleq \mathrm{argmax}_{\boldsymbol{\theta}} J_2(\boldsymbol{\theta})$.

B. Information-theoretic source activity detection

Define the binary index d by $d=1$ (resp. $d=0$) iff the emitter is active (resp. inactive). The SAD consists in computing the Minimum Akaike Information Criterion Estimate (MAICE) $\hat{d}_{\mathrm{AIC}}$ of d from the measurement vector $\boldsymbol{Z}$. The proposed method is related—but not equivalent—to [7][8].

Depending on the value of d, the eigendecomposition of the PSD matrix $C[k]$ at frequency index k—henceforth denoted as $C_d[k]$—writes as

$$C_d[k] = \lambda_k \boldsymbol{U}_k\boldsymbol{U}_k{}^\dagger + \sigma^2[k](\mathbb{I}_2 - \boldsymbol{U}_k\boldsymbol{U}_k{}^\dagger) \quad \text{if } d=1 \quad (13)$$

$$C_d[k] = \sigma^2[k]\mathbb{I}_2 \quad \text{if } d=0. \quad (14)$$

Therein, $\boldsymbol{U}_k$ is the unit-norm eigenvector of $C_1[k]$ associated with the eigenvalue λ_k, which spans the so-called "signal subspace"; $\boldsymbol{U}_k\boldsymbol{U}_k{}^\dagger$ is the orthogonal projector onto the signal subspace; $\mathbb{I}_2 - \boldsymbol{U}_k\boldsymbol{U}_k{}^\dagger$ is its orthogonal complement, or projector onto the "noise subspace". Note that $\lambda_k - \sigma^2$ is the single positive eigenvalue of $C_S[k]$, so $\lambda_k > \sigma^2$. From (13)–(14), a parameter vector fully characterizing the pdf of $\boldsymbol{Z}$ comes as $\boldsymbol{\rho}_1 = [\lambda_{k_1}, ..., \lambda_{k_B}, \sigma^2[k_1], ..., \sigma^2[k_B], \boldsymbol{U}'_{k_1}, ..., \boldsymbol{U}'_{k_B}]'$ for $d=1$ and $\boldsymbol{\rho}_0 = [\sigma^2[k_1], ..., \sigma^2[k_B]]'$ for $d=0$. Whatever $d \in \{0,1\}$ the log-likelihood of $\boldsymbol{\rho}_d$ w.r.t. $\boldsymbol{Z}$ is

$$L(\boldsymbol{Z};\boldsymbol{\rho}_d) = c_0 - N\sum_k \Big(\ln|C_d[k]| + \mathrm{tr}(C_d[k]^{-1}\hat{C}[k])\Big). \quad (15)$$

For $k \in \{k_1, ..., k_B\}$, define $l_{1,k}, l_{2,k}$ as the eigenvalues of $\hat{C}[k]$, ordered such that $l_{1,k} \geq l_{2,k}$, and $\tilde{\boldsymbol{U}}_{1,k}, \tilde{\boldsymbol{U}}_{2,k}$ their corresponding eigenvectors.

Theorem 3. *The MAICE $\hat{d}_{AIC}$ is the argument minimizing*

$$\mathrm{AIC}(d) = -2\tilde{L}(\boldsymbol{Z};\hat{\boldsymbol{\rho}}_d) + 2P(d), \quad (16)$$

with $\hat{\boldsymbol{\rho}}_d$ the MLE of $\boldsymbol{\rho}_d$,

$$\tilde{L}(\boldsymbol{Z};\hat{\boldsymbol{\rho}}_d) = \sum_k \ln\left(\frac{\prod_{i=d+1}^{2} l_{i,k}^{\frac{1}{2-d}}}{\frac{1}{2-d}\sum_{i=d+1}^{2} l_{i,k}}\right)^{(2-d)N}, \quad (17)$$

and $P(d) = (d+1)B + 4dB - 2dB.$ (18)

Proof. Injecting (13) into (15) leads to

$$L(\boldsymbol{Z};\boldsymbol{\rho}_1) = c_0 - N\sum_k \Big(\ln(\lambda_k\sigma^2[k]) + \tfrac{1}{\sigma^2[k]}\mathrm{tr}(\hat{C}[k]) - \big(\tfrac{1}{\sigma^2[k]} - \tfrac{1}{\lambda_k}\big)\mathrm{tr}(\hat{C}[k]\boldsymbol{U}_k\boldsymbol{U}_k{}^\dagger)\Big), \quad (19)$$

which is maximized (under $\boldsymbol{U}_k{}^\dagger\boldsymbol{U}_k = 1$) at $\hat{\boldsymbol{\rho}}_1$ defined by (see [7]) $\hat{\boldsymbol{U}}_k = \tilde{\boldsymbol{U}}_{1,k}$, $\hat{\lambda}_k = \mathrm{tr}(\hat{C}[k]\tilde{\boldsymbol{U}}_{1,k}\tilde{\boldsymbol{U}}_{1,k}{}^\dagger) = l_{1,k}$, $\hat{\sigma}^2[k] = \mathrm{tr}(\hat{C}[k](\mathbb{I}_2 - \tilde{\boldsymbol{U}}_{1,k}\tilde{\boldsymbol{U}}_{1,k}{}^\dagger)) = l_{2,k}$, so that

$$L(\boldsymbol{Z};\hat{\boldsymbol{\rho}}_1) = c_2 - 2N\sum_k \ln(\sqrt{l_{1,k}l_{2,k}}). \quad (20)$$

Similarly, injecting (14) into (15) leads to

$$L(\boldsymbol{Z};\boldsymbol{\rho}_0) = c_0 - 2N\sum_k \ln(\sigma^2[k]) - N\sum_k \tfrac{\mathrm{tr}(\hat{C}[k])}{\sigma^2[k]}, \quad (21)$$

which is maximum for $\hat{\sigma}^2[k] = \frac{1}{2}\mathrm{tr}(\hat{C}[k]) = \frac{l_{1,k}+l_{2,k}}{2}$ so that

$$L(\boldsymbol{Z};\hat{\boldsymbol{\rho}}_0) = c_2 - 2N\sum_k \ln\Big(\tfrac{l_{1,k}+l_{2,k}}{2}\Big). \quad (22)$$

Setting $\tilde{L}(\boldsymbol{Z};\hat{\boldsymbol{\rho}}_d) \triangleq L(\boldsymbol{Z};\hat{\boldsymbol{\rho}}_d) - L(\boldsymbol{Z};\hat{\boldsymbol{\rho}}_1) - c_2$ leads to (17).

Minimizing the negative log-likelihood $-\tilde{L}(\boldsymbol{Z},\hat{\boldsymbol{\rho}}_d)$ over d would favor the model which entails the maximum number of free parameters. This is why the AIC includes a correction term depending on the number $P(d)$ of free entries in $\boldsymbol{\rho}_d$ [6]. In (18), $(d+1)B$ is the number of distinct eigenvalues in the set $\{C_d[k_1], ..., C_d[k_B]\}$, $4dB$ is the number of coefficients of the complex entries of the signal subspace eigenvectors, and $-2dB$ is the reduction of degrees of freedom in $\boldsymbol{\rho}_d$ due to the normalization of these eigenvectors. □

This SAD can also extend to non spatially white noise.

IV. Behavior of the DOA estimator

Suppose that the source and microphones lie on a common plane and that the source is in the farfield, *i.e.*, the source range $r \triangleq |RE|$ is sufficiently high compared to the microphones interspace $2a$ so that the source wavefronts can be considered as planar in the vicinity of the microphone pair. The transfer between the left (R_1) and right (R_2) sensors depends, in this case, on a single spatial parameter θ, namely the angle between the line perpendicular to (R_1R_2) passing through R and (RE). In this section, some properties of the MLE of θ are put forward for two acoustic models: the free-field propagation and the scattering by a rigid sphere with two antipodal sensors. Because with these two models the problem shows a front-back symmetry ($\theta_1 = \pi - \theta_2 \Rightarrow H_{\theta_1} = H_{\theta_2}$), a unique localization cannot be expected if θ lies within the whole interval $]-\pi, \pi]$ [3]. The range of θ is henceforth restricted to $[-\frac{\pi}{2}, \frac{\pi}{2}]$.

The statistics of the MLE—in addition to varying with the observation time—strongly depend on the signal and noise PSDs. To simplify the discussion, the case where s, n have flat spectra on a common frequency range is considered, *i.e.*, $\mathcal{S}, \sigma^2$ are constant on this range. The parameters playing upon the MLE statistics then boil down to the signal bandwidth, its central frequency, and the Signal-to-Noise Ratio (SNR), defined in dB as $\mathrm{SNR} \triangleq 10\log(S/\sigma^2)$.

The following conclusions are of fundamental importance.

A. Free-field propagation

When there is no head between R_1, R_2, the noise-free binaural signals differ only by a time shift, which depends on θ. More precisely, $H_\theta(f) = e^{-2i\pi f . \frac{2a}{c}\sin\theta}$.

Some simulation results are given in Fig.1. The pseudo log-likelihood (12), normalized so as to lie between 0 and 1 for each true bearing value, is depicted as a function of true and tested bearings for various signal bandwidths and SNRs. The data vector $\boldsymbol{Z}$ was randomly drawn according to (4), and the central frequency was set to 3kHz. Fig.1-left, for which the SNR was set to 20dB and the bandwidth to 260Hz, depicts the well-known phase ambiguity problem: when the signal is narrowband (*i.e.*, its bandwidth is small compared to its central frequency), the pseudo log-likelihood admits multiple peaks having nearly equal height. Due to the presence of noise, the highest one does not necessarily correspond to the true source bearing: the estimation error pdf is multimodal, and the Mean Square Error (MSE) is quite large in this mode of operation. When reasoning in terms of TDOA rather than DOA, the tightest lower bound of the MSE in this mode is the so-called Barankin bound [12]. Note that the peak corresponding to the true source location is broader when the source is near $\pm 90°$: bearing estimation is more accurate when the source is facing the microphone pair, as acknowledged in [4]. The aforementioned phase ambiguity vanishes as the SNR rises, as depicted in Fig.1-center, in which the SNR was raised to 60dB: the main peak gets sharper as the SNR increases, and the secondary peaks get lower. Importantly, the narrower the signal bandwidth, the higher the required SNR to fully disambiguate the problem. A second way to avoid phase ambiguity is to have a higher signal bandwidth. This is shown in Fig.1-right, in which the bandwidth was set to 1kHz. As depicted, the secondary peaks are strongly attenuated. A third option would be to increase the observation time. When the SNR, bandwidth and observation time increase, the estimator's pdf becomes unimodal on $[-\frac{\pi}{2}; \frac{\pi}{2}]$ and approaches a Gaussian distribution, in virtue of the asymptotic properties of ML estimators. In this mode of operation, the tightest bound of the MSE is the Cramér-Rao Lower Bound (CRLB). Finally, though not depicted in the figures, when the SNR and bandwidth become too small, the estimator's pdf approaches a uniform distribution on $[-\frac{\pi}{2}, \frac{\pi}{2}]$, along [12].

Fig.2 shows the influence of the central frequency of a narrowband signal (90Hz bandwidth) on the distance between adjacent peaks and their width. As the central frequency decreases, the peaks broaden and stray from each other. At 300Hz (Fig.2-right), secondary peaks have totally disappeared and the system is no longer prone to phase ambiguity. However, the main peak is quite large at this frequency, and so is the MSE.

B. Scattering by a rigid sphere

When the microphones are mounted at the antipodes of a rigid sphere of radius a, the HRTFs to the left and right channels under farfield conditions are obtained from acoustic field theory, see for instance [13] for a decomposition on the basis of Legendre polynomials. Unlike the free-field model, in which only the phase brings information on the source position, the amplitude of H_θ contains useful information.

Some simulation results are shown in Fig.3. In Fig.3-left (260Hz bandwidth, 3kHz central frequency, 20dB SNR), the effect of ILD information on the pseudo log-likelihood can be observed. When comparing with Fig.1-left (for which the conditions of operation are the same), one can see that secondary peaks are somehow attenuated. It depicts that ILD contributes to disambiguate—though not totally—the problem. When the source is close to boresight (0°), the pseudo log-likelihood shows little difference compared to the free-field case. In fact, as the source approaches 0°, the IPD tends to be linear w.r.t. frequency, and the ILD approaches 0dB. The effect of the head hence becomes negligible. Note that, in the same conditions of operation, the pseudo log-likelihood is narrower near $\pm 90°$ with the spherical head model. Hence, the estimator is expected to be more accurate near those DOAs when a spherical head is used. Fig.3-center (90Hz bandwidth, 340Hz central frequency, 20dB SNR) is to be compared with Fig.2-right. It shows the low frequency behavior of the estimator when the head is present. At low frequencies, the sphere ILD does not vary significantly with frequency and bearing. Hence the effect of ILD is negligible. Furthermore, the low frequency IPD can be approximated by $-2i\pi f \frac{3a}{c}\sin\theta$. Hence, the behavior of the estimator in this case is expected to be the same as in the free-field case, excepted that the frequencies are are scaled by a factor $\frac{3}{2}$. This explains the differences between Fig.3-center and Fig.2-right. Fig.3–right, to be compared with Fig.1–right, shows the

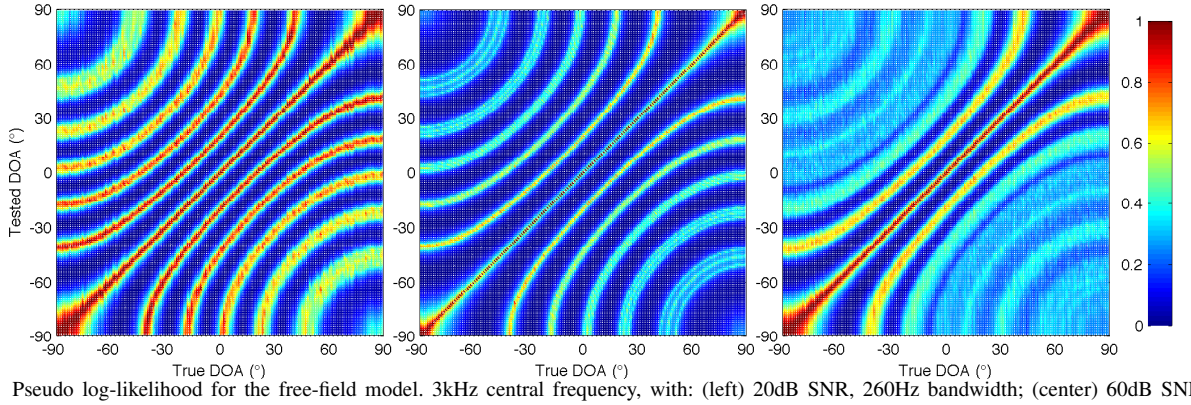

Fig. 1. Pseudo log-likelihood for the free-field model. 3kHz central frequency, with: (left) 20dB SNR, 260Hz bandwidth; (center) 60dB SNR, 260Hz bandwidth; (right) 20dB SNR, 1kHz bandwidth.

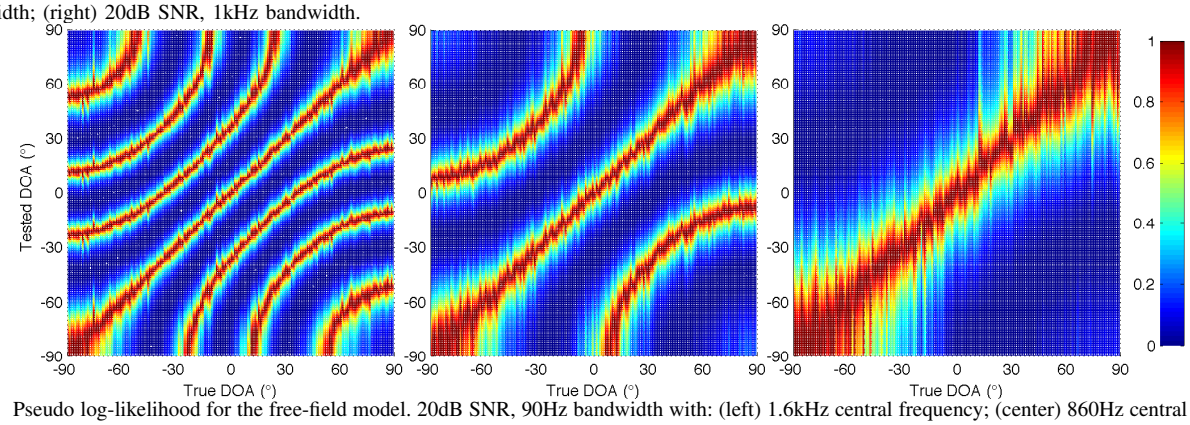

Fig. 2. Pseudo log-likelihood for the free-field model. 20dB SNR, 90Hz bandwidth with: (left) 1.6kHz central frequency; (center) 860Hz central frequency; (right) 340Hz central frequency.

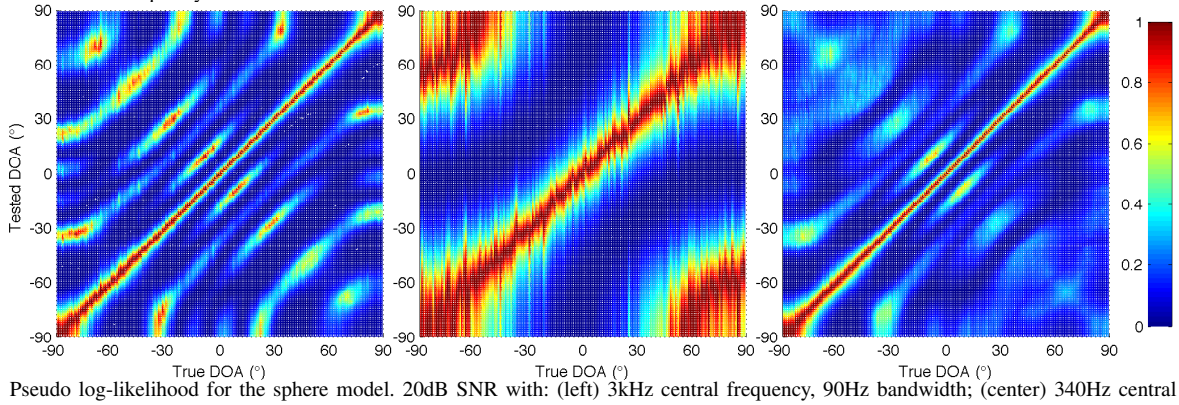

Fig. 3. Pseudo log-likelihood for the sphere model. 20dB SNR with: (left) 3kHz central frequency, 90Hz bandwidth; (center) 340Hz central frequency, 90Hz bandwidth; (right) 3kHz central frequency, 1kHz bandwidth.

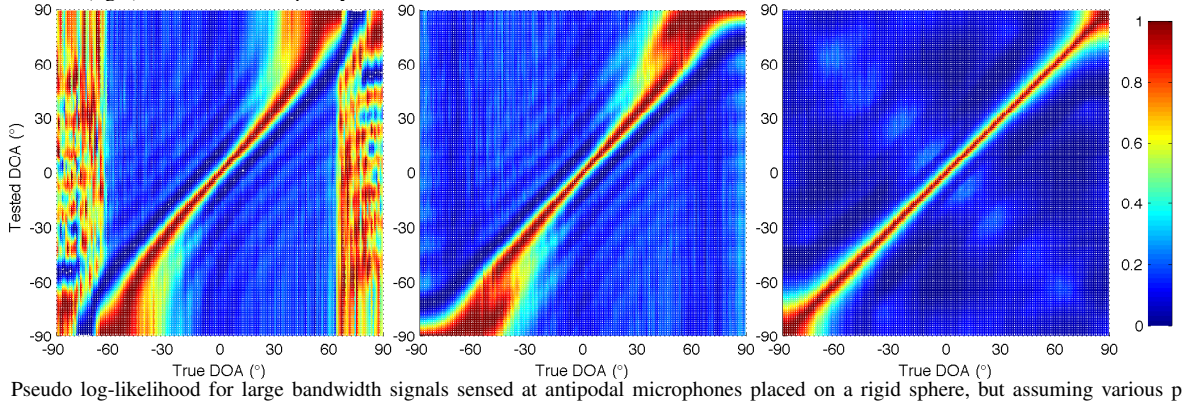

Fig. 4. Pseudo log-likelihood for large bandwidth signals sensed at antipodal microphones placed on a rigid sphere, but assuming various propagation models: (left) mismatch induced by a free-field propagation assumption; (center) mismatch induced by the use of the Woodworth approximation for ITD and the neglection of ILD variations; (right) matching when the scattering is assumed to comply with the genuine model.

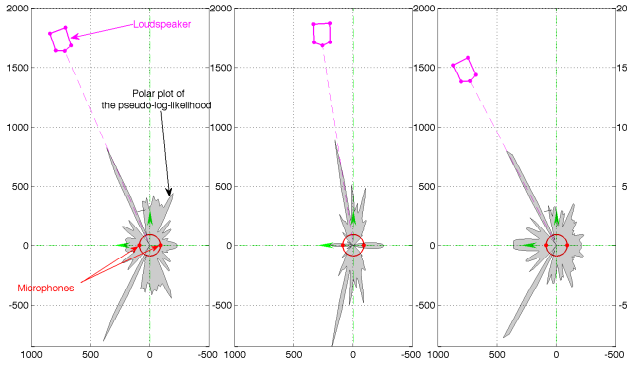

Fig. 5. Experimental results. The ground truth positions (in mm) are depicted in the sensor frame at three time instants. At each time, a polar plot of the pseudo log-likelihood is shown.

influence of the head scattering on the pseudo-log-likelihood for a 1kHz bandwidth 3kHz central frequency signal.

Fig.4 shows the influence of modeling error. Fig.4-left shows the pseudo log-likelihood for a large signal bandwidth when free-field propagation is assumed despite a spherical head is actually present between the microphones. Near boresight, modeling error has little influence on the estimation error, since head scattering is negligible in this case. However, as the true bearing strays from $0°$, a bias appears and increases. An erratic behavior is reported when $|\theta| > 60°$. In Fig.4-center, the sphere is still actually present, but the IPD is approximated from the Woodworth ITD formula [14]: $\text{IPD} = -2i\pi f \frac{a}{c}(\sin\theta + \theta)$, and the ILD variations are just neglected. Although the bias is reduced compared to Fig.4-left, important errors are still reported when the source encloses $\pm 90°$. Fig.4-right show the results when the hypothesized model matches the actual sphere model.

V. Experimental results

In order to assess the proposed estimator with real recordings, experiments were performed in an acoustically prepared room, equipped with 3D pyramidal pattern studio foams placed on the roof and on the walls. Two surface microphones were mounted at the antipodes of a 8.9cm radius plastic rigid sphere, itself placed on a tripod. The two microphones outputs were simultaneously acquired at 44.1kHz. The sphere tripod was moved manually with a wheeled cart while the source, a loudspeaker placed at the same height as the microphones, was emitting various types of signals. The true source and sensor positions were acquired at 200Hz with a motion capture system, providing a less than 1mm position error. Small infrared active markers were placed on the sphere and the loudspeaker, so that their signals could be beamed to three infrared camera units placed at the corners of the room.

The localization and detection algorithms were assessed on various experiments. Results can be found at the URL http://homepages.laas.fr/danes/IROS2013/portello-et-al_videoIROS2013.mp4. Fig.5 shows some of these results for a white noise source. As shown, the pseudo log-likelihood exhibits two main peaks, one corresponding to the true source position, the other one to the symmetric direction w.r.t. $(R_1 R_2)$.

VI. Conclusion

In this paper, a theoretically grounded scheme to DOA estimation and SAD was proposed, which explicitly takes into account the scattering effect of the head. The DOA estimator was studied in simulation for two kinds of models: the free-field model and the rigid sphere model. In this last case, it was shown to be more accurate than other model assumptions. In addition, the ILD entailed in the sphere model was shown to refine, to some extent, the information brought by the IPD only. The proposed approach was tested on real binaural signals from microphones mounted on a plastic sphere, and validated under clean acoustic conditions.

Ongoing work consist in assessing this instantaneous detection-localization algorithm into real environments, and in using it within a stochastic filter for active localization, along [15]. This is a first step towards a theoretically sound algorithm enabling the fusion of binaural perception and sensor motion for any kind of robotics head.

Some proofs were not included for space reasons, but they can be easily be obtained on request.

References

[1] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[2] Y. Chan, R. Hattin, and J. Plant, "The least squares estimation of time delay and its use in signal detection," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'78)*.

[3] A. Handzel and P. Krishnaprasad, "Biomimetic sound-source localization," *IEEE Sensors Journal*, vol. 2, no. 6, pp. 607–616, 2002.

[4] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, 2010.

[5] A. Deleforge and R. Horaud, "The cocktail party robot: Sound source separation and localization with an active binaural head," in *IEEE/ACM Int. Conf. on Human Robot Interaction*, 2012.

[6] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[7] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.

[8] P. Danès and J. Bonnal, "Information-theoretic detection of broadband sources in a coherent beamspace MUSIC scheme," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'2010)*.

[9] V. Macdonald and P. Schultheiss, "Optimum passive bearing estimation in a spatially incoherent noise environment," *Jour. of the Acoustic Society of America*, vol. 46, pp. 37–43, 1969.

[10] A. Jaffer, "Maximum likelihood direction finding of stochastic sources: A separable solution," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'98)*.

[11] M. Doron, A. Weiss, and H. Messer, "Maximum-likelihood direction finding of wide-band sources," *IEEE Trans. on Signal Processing*, vol. 41, no. 1, p. 411, 1993.

[12] A. Weiss and E. Weinstein, "Fundamental limitations in passive time delay estimation - Part I: Narrowband systems," *IEEE Trans. on Acoustics, Speech and Signal Processing*, 1983.

[13] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *Jour. of the Acoustic Society of America*, vol. 104, pp. 3048–3058, 1998.

[14] R. Woodworth and H. Schlosberg, *Experimental Psychology*, 3rd ed. Holt, Rinehart and Winston, 1971.

[15] 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 Intelligent Robots and Systems (IROS'2012)*.