Localization of Multiple Sources from a Binaural Head in a Known Noisy Environment

Alban Portello, Gabriel Bustamante, Patrick Danès, Alexis Mifsud

Localization of Multiple Sources from a Binaural Head in a Known Noisy Environment

Alban Portello[1,2], Gabriel Bustamante[1,2], Patrick Danès[1,2] and Alexis Mifsud[1]

***Abstract*— This paper presents a strategy to the localization of multiple sound sources from a static binaural head. The sources are supposed W-Disjoint Orthogonal and their number is assumed known. Their most likely azimuths are computed by means of the Expectation-Maximization algorithm. Application of the method on simulated data is reported, as well as some evaluations of its $\mathcal{HARK}$ implementation on experimental data. Two important properties are observed: scattering effects can be coped with, thanks to the required prior knowledge of the (room-independent) head interaural transfer function; the environment noise statistics are handled separately.**

I. Introduction

Sound source localization is a fundamental function in Robot Audition [1]. Fast and accurate solutions relying on microphone arrays have been proposed, *e.g.*, broadband beamspace MUSIC and its mixed hardware/software implementation [2]. Binaural solutions are more difficult, and their performances drop when facing multiple sources in realistic environments (reverberation, interfering noise, etc.) [3][4].

Source localization is often turned into a maximization problem. If no prior knowledge on the source is available, it entails a multivariate function, which has no analytic expression and whose brute numerical optimization is prohibitive. So, simplifying assumptions are made and iterative maximization from an initial guess is targeted. For instance, [5] estimates Time Differences Of Arrival (TDOA) of multiple competing sources in a reverberant environment from a pair of microphones, on the basis of a diffuse noise covariance matrix capturing reverberation. The method is based on the Expectation-Maximization (EM) algorithm, and assumes that within a given "bin" of a time-frequency decomposition of the signals, only one source is dominant, what is often called "W-Disjointness Orthogonality" (WDO). In [6], K-means clustering is proposed rather than EM, though based on WDO. Methods based on the Degenerate Unmixing Estimation Technique (DUET) can be found in [7]. Ref [8] compares [6] and [5], and proposes an EM algorithm relaxing WDO. All these methods address TDOA estimation, and do not account for scatterers.

In [9], up to three sources are localized from an anthropomorphic binaural head, on the basis of EM and assuming WDO. Input data are Interaural Phase Differences (IPDs) and Interaural Level Differences (ILDs) over time-frequency bins. The Interaural Transfer Function is learnt beforehand along source directions and frequencies, but includes room reverberation and environment noise.

This paper estimates the spatial parameters of multiple sound sources by means of an EM algorithm under a parsimony assumption. Contrarily to [5][6], head induced scattering is learned beforehand and accounted for. A random source model is considered. Contrarily to [9], the interaural transfer function (without noise) and the statistics of the noise at the microphones are independently learned. In our opinion, this can contribute to limit the localization drop-off induced by noise correlation [4]. To our knowledge, to date no EM algorithm can cope independently with these two aspects.

In the following, Section II introduces modeling aspects as well as the Maximum Lihelihood estimation of the azimuth of a single source under known noise statistics. Section III extends this result to the multiple source case through the EM algorithm under WDO. Some theoretical aspects often overlooked in the literature are explicited. Sections IV and V report the application of the approach on synthetic data under MATLAB®, and on real data through the open-source library $\mathcal{HARK}$ [10]. A conclusion ends the paper.

II. Basics

Consider two microphones R_1, R_2 laid on a head, and denote z_1, z_2 the real-valued continuous-time signals perceived at R_1, R_2. Define $z \triangleq [z_1, z_2]'$, with $'$ the transpose operator. This section recalls the way how the Maximum Likelihood (ML) estimate of the azimuth of a single source emitting into noise can be computed from finite time records of z, on the basis of the head interaural transfer function (computed analytically or measured in an anechoic environment) and the noise statistics (given or learned beforehand). It builds on theoretical results from [11] and references [12][13].

The way how the data vector is generated from records of z is explained in §II-A. The hypothesized mathematical model and hypotheses underlying the prior statistics of z, as well as the ML estimation of the problem unknowns (including the source azimuth), are then presented in §II-B for this single source case. Definitions and results introduced in this section serve as the basis for theoretical developments of §III in the context of multiple sources.

A. Construction of the Data Vector from the Raw Signals

The signals z_1, z_2 are observed over a finite-length time interval, divided into P (possibly overlapping) *segments* of length L. The restrictions of z to these segments are called *frames*, and sets of N_f consecutive frames are termed *groups of frames*. The total number N_g of groups of frames is

*This work was supported by ANR/JST BINAAHR project, and by EU FET grant TWO!EARS, ICT-618075, www.twoears.eu.

[1]CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France
[2]Univ de Toulouse, UPS, LAAS, F-31400 Toulouse, France

such that $N_g N_f = P$. A subscript pair (n_g, n_f) henceforth refers to the n_f^{th} frame of the n_g^{th} group. Each frame indexed by (n_g, n_f) is first windowed and Fourier transformed according to

$$Z_{(n_g,n_f)}(f) = \frac{1}{\sqrt{L}} \int_{-\infty}^{\infty} z(t) w(t - \tau_{(n_g,n_f)}) e^{-2i\pi f t} dt. \quad (1)$$

Therein, w terms any window function symmetric over its L-width support, and $\tau_{(n_g,n_f)}$ is defined after (n_g, n_f). Define $Z_{(n_g,n_f)}[k] \triangleq Z_{(n_g,n_f)}(\frac{k}{L})$ and $Z_{n_g}[k] = [Z_{(n_g,1)}[k]', \ldots, Z_{(n_g,N_f)}[k]']'$. The $2N_f N_g B$-dimensional complex data vector Z, defined as the stacking of Short-Time Fourier Transforms (STFTs), is called *channel-time-frequency decomposition* of z, and comes as

$$Z = [Z_1[k_1]', ..., Z_{N_g}[k_1]', \ \ldots \ , Z_1[k_B]', ..., Z_{N_g}[k_B]']', \quad (2)$$

with $k_1, \ldots, k_B$ the B frequency indexes within the bandwidth of interest. Hereafter, $Z_{n_g}[k]$ is termed *data in the time-frequency bin*[1] *indexed by* (n_g, k).

B. Interaural Transfer Function Based Localization of a Single Source into Known Noise

1) Model definition: For the sake of simplicity, we restrict the problem to localization in the azimuthal plane, but this induces no loss of generality. It is assumed that the perceived signals have the form

$$\begin{aligned} z_1(t) &= s(t) + n_1(t) \\ z_2(t) &= (s * h_\theta)(t) + n_2(t), \end{aligned} \quad (3)$$

where the contribution s of the emitter at R_1 and the additive noises n_1, n_2 are real, zero-mean, band-limited, jointly Gaussian random processes, and * stands for convolution. The impulse response h_θ between R_1 and R_2 naturally depends on the source direction of arrival θ and captures head scattering. The source is assumed far enough from the head so that h_θ does not depend on its distance. Though the source signal is not assumed Wide Sense Stationary (WSS), its autocorrelation $R_{ss}(t, t-\tau) \triangleq \mathbb{E}\{s(t)s(t-\tau)\}$ over a group of time frames indexed by n_g is assumed not to vary significantly along t so that it and its Fourier transform (the power spectral density of s) can be denoted by $R_{n_g}(\tau)$ and $S_{n_g}(f)$. Under some conditions, the probability density function (pdf) of the data vector Z is shown to be faithfully approximated by

$$\begin{aligned} p_Z(z; C_1[k_1], ..., C_1[k_B], \ldots, C_{N_g}[k_1], ..., C_{N_g}[k_B]) = \\ \prod_{n_g=1..N_g \,;\, b=1..B} {}^{\mathbb{C}}\mathcal{N}\big(z_{n_g}[k_b]; 0, \text{blkdiag}(\underbrace{C_{n_g}[k_b], ..., C_{n_g}[k_b]}_{N_f \text{times}})\big), \end{aligned} \quad (4)$$

with $C_{n_g}[k]$ the covariance matrix of $Z_{n_g}[k]$—or, equivalently, the power spectral density evaluated at $\frac{k}{L}$ of the signal z windowed onto the n_g^{th} group of frames—and ${}^{\mathbb{C}}\mathcal{N}(.; m, P)$ the multivariate circular complex Gaussian distribution of mean m and covariance matrix P, and z (resp. $z_{n_g}[k]$) a sample of Z (resp. $Z_{n_g}[k]$).

[1] In "time-frequency", "time" thus stands for "group of frames".

Define the *steering vector* $V_\theta[k] \triangleq [1, H_\theta[k]]'$, where the *interaural transfer function* $H_\theta(f)$ stands for the Fourier transform of $h_\theta(t)$. When n_1, n_2 are independently and jointly WSS, independent identically distributed (iid), and independent of s, $C_{n_g}[k]$ becomes

$$C_{n_g}[k] \triangleq V_\theta[k] S_{n_g}[k] V_\theta[k]^\dagger + \sigma^2[k] \mathbb{I}_2 \quad (5)$$

with † the Hermitian transpose operator and $\mathbb{I}_2$ the 2×2 identity matrix. $H_\theta[k]$ is assumed known for at least a discrete set of frequencies $k = k_1, \ldots, k_B$ and azimuths θ, *e.g.*, through identification, analytical computation or continuous expansion. The variances $\sigma^2[k_1], \ldots, \sigma^2[k_B]$ of the noise are supposed learned beforehand as well.

2) The vector of unknowns parameters and its Maximum Likelihood Estimate (MLE): In the considered problem, θ is not the single unknown. Since no prior information about the source is given, the parameter vector to estimate boils down to

$$\Theta \triangleq [\theta, S_1[k_1], ..., S_1[k_B], \ldots, S_{N_g}[k_1], ..., S_{N_g}[k_B]]'. \quad (6)$$

Its MLE $\hat{\Theta}_{ML}$ is the argmax of the logarithm of (4), *i.e.*,

$$J(Z; \Theta) = c - N_f \sum_{n_g=1..N_g \,;\, b=1..B} \Big(\ln |C_{n_g}[k]| + \text{tr}(C_{n_g}[k]^{-1} \bar{C}_{n_g}[k]) \Big), \quad (7)$$

with c a constant, $|.|$ and $\text{tr}(.)$ the determinant and trace operators, and

$$\bar{C}_{n_g}[k] \triangleq \tfrac{1}{N_f} \textstyle\sum_{n_f=1}^{N_f} z_{(n_g,n_f)}[k] z_{(n_g,n_f)}[k]^\dagger \quad (8)$$

the sample covariance matrix of the data in the time-frequency bin (n_g, k). A closed-form expression of the MLE $\hat{\Theta}_{ML}$ cannot be obtained. However, in the considered single-source case, a *separable form* can be exhibited, which enables the determination of the most likely source azimuth $\hat{\theta}_{\text{ML}}$ through the maximization of a function of θ only. This is summarized in the following theorem:

Theorem 1: If, for each n_g^{th} group of frames constituting the considered time interval, the sample covariance matrices $\{\bar{C}_{n_g}[k_b]\}_{b=1,\ldots,B}$ are all full-rank, then the MLE $\hat{\theta}_{ML}$ of the source azimuth θ is given by

$$\hat{\theta}_{ML} = \arg\max_\theta L(\theta) \quad (9)$$

where the pseudo log-likelihood

$$\begin{aligned} L(\theta) = -N_f \sum_{n_g=1..N_g \,;\, b=1..B} \Big(\ln \big| P_\theta[k_b] \bar{C}_{n_g}[k_b] P_\theta[k_b] + \sigma^2[k_b] P_\theta^\perp[k_b] \big| \\ + \frac{1}{\sigma^2[k_b]} \text{tr}(P_\theta^\perp[k_b] \bar{C}_{n_g}[k_b]) \Big) \end{aligned} \quad (10)$$

is defined from the orthogonal projector $P_\theta[k] \triangleq V_\theta[k](V_\theta[k]^\dagger V_\theta[k])^{-1} V_\theta[k]^\dagger$ on the subspace spanned by $V_\theta[k]$ and from its orthogonal complement $P_\theta^\perp[k] \triangleq \mathbb{I}_2 - P_\theta[k]$.

Proof: [Sketch] The proof is an adaptation of the results of [11] to the broadband case, considering a single source and two sensors related by H_θ, instead of an array in the free field. Another important difference is that while s, n_1, n_2 are considered as snippets of jointly WSS signals over each

group of frames, the source autocorrelation is allowed to vary from one group of frames to another.

The idea is to express the first-order stationarity conditions which must be fulfilled by the MLEs $\{\hat{S}_{n_g}[k_b]\}_{n_g,b}$ of the spectral parameters of the signal s so that they can maximize (7). These necessary conditions are also sufficient here. Reporting them into (5) as functions $\{\hat{S}_{n_g}[k_b](\theta)\}_{n_g,b}$ of the azimuth θ, then reporting the result into (7), leads to $J(Z;\theta,\{\hat{S}_{n_g}[k_b](\theta)\}_{n_g,b}) = c + L(\theta)$, hence the result and the "pseudo likelihood" terminology. ∎

The more general case when the noises n_1, n_2 are stationary but not iid can still be handled if the covariance matrices $\{C_n[k_b]\}_{b=1,\dots,B}$ of $n = [n_1, n_2]'$ on the considered frequency bins can be learned. For instance, if $\forall b \in 1,\dots,B$, $C_n[k_b] \approx \tilde{C}_n[k_b] = \tilde{Q}_n[k_b]\tilde{Q}_n[k_b]^\dagger$, where $\{\tilde{C}_n[k_b]\}_{b=1,\dots,B}$ and their Cholesky roots are given (*e.g.*, computed from sample covariance matrices), then the covariance matrix $\tilde{C}_{n_g}[k]$ of $\tilde{X}_{n_g}[k] = (\tilde{Q}_n[k])^{-1}X_{n_g}[k]$ gets a structure similar to (5). Indeed,

$$\tilde{C}_{n_g}[k] \triangleq \tilde{V}_\theta[k]S_{n_g}[k]\tilde{V}_\theta[k]^\dagger + \mathbb{I}_2, \tag{11}$$

holds, with $\tilde{V}_\theta[k] = (\tilde{Q}_n[k])^{-1}V_\theta[k]$, and Theorem 1 applies.

III. An Expectation-Maximization approach to multiple source localization

When several sources emit simultaneously, the problem of maximizing (7) over the complete parameter vector Θ is no longer separable, so that a closed form of the MLEs of the source azimuths is out of reach. To tackle such cases without resorting to brute force, a general approach consists in adding simplifying assumptions which enable an iterative solution towards the optimum. The most recurrent hypothesis made in the literature certainly consists in postulating the *W-Disjointness Orthogonality (WDO)* of the sources, viz. at most one source is dominant in each "bin" of a time-frequency representation of the signals. Schematically, the algorithms founded on this assumption consist in iterating the two following steps:

- (*separation* or *masking*): gather the time-frequency bins onto which each source is supposed to be dominant, on the basis of the estimates of the spatial parameters of all the sources obtained at the preceding iteration;
- (*localization*): for each source, re-estimate its associated spatial parameters from the time-frequency bins onto which this source is assumed dominant.

This section constitutes the main theoretical contribution of the paper and is organized as follows. The Expectation-Maximization (EM) algorithm, which is the cornerstone of the approach, is briefly recalled in §III-A. Then, the likelihood defined in §II-B for the unknown parameters related to a single source is extended to the multiple source case in §III-B. Finally, §III-C shows how the EM algorithm can lead to the MLEs of the azimuths of several sources, and summarizes the whole localization strategy.

A. EM at a Glance

The Expectation-Maximization (EM) algorithm [14] is an iterative method to the Maximum Likelihood estimation of a parameter vector Θ from samples z of a random variable Z when facing incomplete data. A random vector X (continuous, discrete or hybrid) of *complete data* is introduced, such that

$$p(z|x,\Theta) = p(z|x), \tag{12}$$

i.e., such that Z is independent of Θ given X, and such that given a sample x of X the computation of $\arg\max_\Theta p(x|\Theta)$ can easily be done. The selection of X is not unique, and is a key point of the method efficiency. Quite often, one sets $X = [Z', Y']'$, with Y the random vector of *latent variables*.

Let Θ and Θ^* be two candidate values for the sought MLE. Defining $L(\Theta)$ as the log-likelihood $\ln p(z|\Theta)$ to be maximized, it can be shown that

$$L(\Theta) - L(\Theta^*) \geq Q(\Theta,\Theta^*) - Q(\Theta^*,\Theta^*), \tag{13}$$

where the bivariate *auxiliary function* $Q(.,.)$ is defined by

$$Q(\Theta,\Theta^*) = \mathbb{E}\big\{\ln p(X|\Theta)\big|Z=z,\Theta^*\big\} \tag{14}$$

$$= \int \ln p(x|\Theta)p(x|z,\Theta^*)dx. \tag{15}$$

So, given an initial guess Θ^*, if a value Θ can be found such that $Q(\Theta,\Theta^*) > Q(\Theta^*,\Theta^*)$, then, in view of (13), Θ is more likely (or at least no less likely) than Θ^* with respect to z. If, by a judicious choice of X, $Q(.,\Theta^*)$ can be made easy to maximize, then the following algorithm ensures to converge efficiently to a local maximum of the log-likelihood L:

1) Initialize Θ^*;
2) E-step: Compute $Q(.,\Theta^*)$ defined in (15);
3) M-step: Compute Θ as the argmax of $Q(.,\Theta^*)$;
4) If $L(\Theta) - L(\Theta^*)$ is greater than a predefined threshold η, then set $\Theta^* = \Theta$ and go back to step 2; otherwise output Θ as the sought MLE.

Note that in the case when $X = [Z', Y']'$ the auxiliary function becomes $Q(\Theta,\Theta^*) = \mathbb{E}\big\{\ln p(z,y|\Theta)\big|Z=z,\Theta^*\big\}$ $= \int \ln p(z,y|\Theta)p(y|z,\Theta^*)dy$.

B. Log-likelihood for the Multiple WDO Sources Model

All the data and hypotheses similar to §II are in effect in this multiple sources case, except that the signal s is traded for the contributions $s_1,\dots,s_Q$ due to the emitters $q = 1,\dots,Q$ at the microphone R_1. As before, each q^{th} such signal, though non necessarily WSS, is assumed "locally" WSS over each n_g^{th} group of frames, with $R_{n_g}^{(q)}(\tau)$ and $S_{n_g}^{(q)}(f)$ its "local" autocorrelation and power spectral density. The W-Disjointness Orthogonality (WDO) assumption is mathematically turned into the fact that within the time-frequency bin indexed by (n_g,k), at most one source, say the q^{th} one, is dominant. The covariance matrix (5) of the data vector in the considered bin (n_g,k) then becomes

$$C_{n_g}^{(q)}[k] \triangleq V_{\theta_q}[k]S_{n_g}^{(q)}[k]V_{\theta_q}[k]^\dagger + \sigma^2[k]\mathbb{I}_2. \tag{16}$$

Let Q be the (*a priori* known) number of active sources. The vector Θ of the parameters to be estimated is defined as (in the vein of the single-source case (6))

$$\begin{aligned}\Theta &\triangleq [\theta', S_1[k_1]',...,S_1[k_B]',\ldots,S_{N_g}[k_1]',...,S_{N_g}[k_B]']' \\ &\text{with } \theta \triangleq [\theta_1,\ldots,\theta_Q]', \\ &\text{and}, \forall(n_g,b),\ S_{n_g}[k_b] \triangleq [S_{n_g}^{(1)}[k_b],\ldots,S_{n_g}^{(Q)}[k_b]]'. \quad (17)\end{aligned}$$

With the above assumptions and notations in mind, the log-likelihood of Θ with respect to the data can be written

$$\ln p(z|\Theta) = \sum_{n_g=1..N_g\,;\,b=1..B} \ln p\big(z_{n_g}[k_b]\big|\theta, S_{n_g}[k_b]\big). \quad (18)$$

The prior probability of dominance in any time-frequency bin indexed by (n_g,k) is assumed to be evenly distributed (*i.e.*, equal to $\frac{1}{Q}$) among the Q sources. So, the conditional pdf of $Z_{n_g}[k]$ comes as the mixture of Q equiprobable hypotheses

$$p\big(z_{n_g}[k]\big|\theta, S_{n_g}[k]\big) = \sum_{q=1..Q} \frac{1}{Q}\, p\big(z_{n_g}[k]\big|\theta_q, S_{n_g}^{(q)}[k]\big), \quad (19)$$

each one being a circular complex Gaussian

$$\begin{aligned}&p\big(z_{n_g}[k]\big|\theta_q, S_{n_g}^{(q)}[k]\big) = \\ &\qquad {}^{\mathbb{C}}\mathcal{N}\big(z_{n_g}[k]; 0, \mathrm{blkdiag}(C_{n_g}^{(q)}[k],...,C_{n_g}^{(q)}[k])\big) \quad (20)\end{aligned}$$

along (4), where the covariance matrix $C_{n_g}^{(q)}[k]$ defined in (16) is parameterized by the power spectral density $S_{n_g}^{(q)}[k]$ in the time-frequency bin (n_g,k) and by the azimuth θ_q, both associated to the q^{th} source.

C. Latent Variables and Main Theorem

To extract the argmax of (18), with (19)–(20) therein, by means of the EM algorithm, the latent discrete random vector

$$Y \triangleq [Y_1[k_1],...,Y_1[k_B],\ldots,Y_{N_g}[k_1],...,Y_{N_g}[k_B]]' \quad (21)$$

is introduced, each entry $Y_{n_g}[k_b]$ of which accounts for the origin of $Z_{n_g}[k_b]$, *i.e.*, accounts for which q^{th} source is dominant in the corresponding time-frequency bin:

$$\begin{aligned}&\{Y_{n_g}[k]=q\} \\ &\quad \Rightarrow z_{n_g}[k] \sim {}^{\mathbb{C}}\mathcal{N}\big(.; 0, \mathrm{blkdiag}(C_{n_g}^{(q)}[k],...,C_{n_g}^{(q)}[k])\big). \quad (22)\end{aligned}$$

This selection of latent variables random vector is quite natural since if its value was known, then it would be sufficient to partition the set of time-frequency bins along their associated dominant sources, and, for each source, to compute the related azimuth through a single-source Maximum Likelihood estimation. The random vectors $\{[Y_{n_g}[k_b], X_{n_g}[k_b]]'\}_{n_g=1..N_g\,;\,b=1..B}$ are assumed mutually independent, and the complete data random vector can be written as $X=[Z',Y']'$. The following theorem holds.

Theorem 2: Under all the assumptions made so far, including the prior knowledge of the number Q of the active sources, the MLE $\hat{\theta}_{ML} = [\hat{\theta}_1,\ldots,\hat{\theta}_Q]'$ of the vector $\theta = [\theta_1,\ldots,\theta_Q]'$ of their azimuths can be obtained from an initial guess $\theta^{(\text{init})}$ by means of the EM Algorithm 2. Beforehand, the data and steering vectors have been normalized taking into account the characteristics of the noise through Algorithm 1. Importantly, the most likely azimuth of each source comes from an independent maximization process, and no initial guess is needed for the spectral parameters of the sources.

Algorithm 1: Data Conditioning

Inputs: $\{\{Z_{n_g}[k_b]\}_{n_g=1,..,N_g}, Q_n[k_b], \{V_\theta[k_b]\}_{\theta=\theta_1,..,\theta_{N_\theta}}\}_{b=1,..,B}$
Outputs: $\{J_{n_g}[k_b](\theta)\}_{n_g=1,..,N_g,b=1,..,B,\theta=\theta_1,..,N_\theta}$

for $b=1,\ldots,B$ **do**
  for $n_g=1,\ldots,N_g$ **do**
    Transformation making noises normed and iid applied on data vector
    $\tilde{Z}_{\text{temp}} = (Q_n[k_b])^{-1}\, Z_{n_g}[k_b]$
    Sample covariance matrix at bin (n_g,k_b)
    $\bar{C}_{n_g}[k_b] = \frac{1}{N_f}\left[\tilde{Z}_{\text{temp}}\right]\left[\tilde{Z}_{\text{temp}}\right]^\dagger$
  end
  for $\theta=\theta_1,\ldots,\theta_{N_\theta}$ **do**
    Transformation making noises normed and iid applied on steering vector
    $\tilde{V}_\theta[k_b] = (Q_n[k_b])^{-1}\, V_\theta[k_b]$
    Projector onto the space spanned by $\tilde{V}_\theta[k_b]$
    $P_\theta[k_b] = \tilde{V}_\theta[k_b](\tilde{V}_\theta[k_b]^\dagger \tilde{V}_\theta[k_b])^{-1}\tilde{V}_\theta[k_b]^\dagger$
    Orthogonal complement of $P_\theta[k_b]$
    $P_\theta^\perp[k] = \mathbb{I}_2 - P_\theta[k]$
    for $n_g=1,\ldots,N_g$ **do**
      elements necessary to the computation of auxiliary functions (step 8 of Algorithm 2)
      $J_{n_g}[k_b](\theta) =$
        $-N_f\Big(\ln|P_\theta[k_b]\bar{C}_{N_g}[k_b]P_\theta[k_b] + P_\theta^\perp[k_b]|$
        $+\mathrm{tr}(P_\theta^\perp[k_b]\bar{C}_{N_g}[k_b])\Big)$
    end
  end
end

Proof: In the following, point-mass functions (pmf) are equally used to term probabilities of discrete random variables, and mixed probability density and mass functions (mixed pdf-pmf's) are handled. By definition, the auxiliary function can be written as

$$Q(\Theta,\Theta^*) = \sum_{y\in\{1,\ldots,Q\}^{N_g B}} \ln p(z,y|\Theta)p(y|z,\Theta^*). \quad (23)$$

As $p(z_{n_g}[k]|y_{n_g}[k],\theta,S_{n_g}[k]) = p(z_{n_g}[k]|\theta_{y_{n_g}[k]}, S_{n_g}^{(y_{n_g}[k])}[k])$, one gets:

$$\begin{aligned}&\ln p(z,y|\Theta) = \sum_{n_g=1..N_g\,;\,b=1..B} \ln p\big(z_{n_g}[k_b], y_{n_g}[k_b]\ \big|\ \theta, S_{n_g}[k_b]\big) \\ &= \sum_{n_g,b} \ln \tfrac{1}{Q} p\big(z_{n_g}[k_b]\ \big|\ \theta_{y_{n_g}[k_b]}, S_{n_g}^{(y_{n_g}[k_b])}[k_b]\big) \quad (24) \\ &= \sum_{n_g,b} \ln {}^{\mathbb{C}}\mathcal{N}\big(z_{n_g}[k_b]; 0, \mathrm{blkdiag}(C_{n_g}^{(y_{n_g}[k_b])}[k_b],...,C_{n_g}^{(y_{n_g}[k_b])}[k_b])\big)... \\ &\qquad\qquad -N_g B \ln Q \\ &= c(Q) - N_f \sum_{n_g=1..N_g\,;\,b=1..B} \big(\ln|C_{n_g}^{(y_{n_g})}[k_b]| + \mathrm{tr}(C_{n_g}^{(y_{n_g})}[k_b]^{-1}\bar{C}_{n_g}[k_b])\big)\end{aligned}$$

with $\bar{C}_{n_g}[k]$ the sample covariance matrix defined in (8) and $c(Q)$ a constant depending on Q. Besides, the likelihood of Θ^* with respect to the latent variables given the data can be written as (where a pmf $p(y|.)$ terms the probability $\mathbb{P}(Y=y|.)$)

$$p\big(y|z,\Theta^*\big) = \prod_{n_g,b} p\big(y_{n_g}[k] \,\big|\, z_{n_g}[k],\theta^*,S_{n_g}[k]^*\big)$$

$$= \prod_{n_g,b}\left(\frac{\frac{1}{Q}p\left(z_{n_g}[k]|\theta^*_{y_{n_g}[k]},S_{n_g}^{(y_{n_g}[k])}[k]^*\right)}{\sum_q \frac{1}{Q}p\left(z_{n_g}[k]|\theta^*_q,S_{n_g}^{(q)}[k]^*\right)}\right) \quad (25)$$

$$= \prod_{n_g,b}\left(\frac{{}^{\mathbb{C}}\mathcal{N}\Big(z_{n_g}[k_b];0,\mathrm{blkdiag}\Big(C_{n_g}^{(y_{n_g}[k])*}[k_b],\ldots,C_{n_g}^{(y_{n_g}[k_b])*}[k_b]\Big)\Big)}{\sum_q {}^{\mathbb{C}}\mathcal{N}\Big(z_{n_g}[k_b];0,\mathrm{blkdiag}\Big(C_{n_g}^{(q)*}[k_b],\ldots,C_{n_g}^{(q)*}[k_b]\Big)\Big)}\right).$$

By using the relationships

$$\textstyle\sum_{n_g,b}\ln p\big(z_{n_g}[k_b]\,\big|\,\theta_{y_{n_g}[k_b]},S_{n_g}^{(y_{n_g}[k_b])}[k_b]\big) = \sum_{n_g,b}\sum_{q=1}^{Q}\delta_{q,y_{n_g}[k_b]}\ln p\big(z_{n_g}[k_b]\,\big|\,\theta_q,S_{n_g}^{(q)}[k_b]\big), \quad (26)$$

and

$$\textstyle\sum_{y_{n_g}[k]=1}^{Q}\delta_{q,y_{n_g}[k]}p(y_{n_g}[k]|z_{n_g}[k],\theta^*,S_{n_g}[k]^*) = p(q|z_{n_g}[k],\theta^*,S_{n_g}[k]^*), \quad (27)$$

where δ (resp. the pmfs $p(y|.)$ and $p(q|.)$) stands for the Kronecker symbol (resp. $\mathbb{P}\{Y_{n_g}[k]=y|.\}$ and $\mathbb{P}\{Y_{n_g}[k]=q|.\}$), the auxiliary function $Q(\Theta,\Theta^*)$ can be written as

$$Q(\Theta,\Theta^*) = \sum_{q=1..Q} Q^{(q)}(\Theta_q,\Theta^*) \quad (28)$$

where the spatial and spectral parameters associated to the q^{th} source are gathered into $\Theta_q \triangleq [\theta_q, S_1^{(q)}[k_1],\ldots,S_1^{(q)}[k_B],\ldots,S_{N_g}^{(q)}[k_B],\ldots,S_{N_g}^{(q)}[k_B]]'$. After some algebraic manipulations, one gets

$$Q^{(q)}(\Theta_q,\Theta^*) = \sum_{n_g=1..N_g\,;\,b=1..B}\gamma_{n_g}^{(q)}[k_b]\ln p\big(z_{n_g}[k_b]\,\big|\,\theta_q,S_{n_g}^{(q)}[k_b]\big), \quad (29)$$

$$\gamma_{n_g}^{(q)}[k] = \mathbb{P}\{Y_{n_g}[k]=q|z_{n_g}[k],\theta^*,S_{n_g}[k]^*\}. \quad (30)$$

It is then sufficient to maximize each function $Q^{(q)}(\Theta_q,\Theta^*)$ with respect to its associated decision vector Θ_q in order to get the argmax of $Q(\Theta,\Theta^*)$ along Θ. Similarly to (7) in the single-source case, $\ln p(z_{n_g}[k]|\theta_q,S_{n_g}^{(q)}[k])$, already expressed in (20), has the form $c(Q) - N_f\sum_{n_g,b}\big(\ln|C_{n_g}^{(q)}[k]| + \mathrm{tr}(C_{n_g}^{(q)}[k]^{-1}\bar{C}_{n_g}[k])\big)$. In view of earlier developments in §II-B leading to Theorem 1, the MLE of each azimuth θ_q is obtained through step 8 of Algorithm 2 with J defined in steps 11–12 of Algorithm 1. The difference with Theorem 1 is that step 8 involves the weighting factors $\gamma_{n_g}^{(q)}[k]$ in the sum over the considered time-frequency bins. Each such weight accounts for the probability that the q^{th} source is dominant in the bin indexed by (n_g,k), on the basis of the data and making the "naive" hypothesis that the parameter vector is Θ^*.

Algorithm 2: Localization

Inputs: $\theta^{(\text{init})}, \{J_{n_g}[k_b](\theta)\}_{n_g=1,..,N_g,b=1,..,B,\theta=\theta_1,..,N_\theta}$
Outputs: $\hat{\theta}$
Parameters: η, maxit

1 count $= 0, \Delta L = \infty$
2 **while** $\Delta L \geq \eta$ | count $<$ maxit **do**
3 count $=$ count $+ 1$
  `Maximization step`
4 **for** $q = 1,\ldots,Q$ **do**
5 **if** count $= 1$ **then**
6 $\hat{\theta}_q = \theta_q^{(\text{init})}$
7 **else**
8 $\hat{\theta}_q = \arg\max_\theta \sum_{n_g=1..N_g\,;\,b=1..B}\gamma_{n_g}^{(q)}[k_b]\,J_{n_g}[k_b](\theta)$
9 **end**
10 **end**
  `Log-Likelihood computation`
11 $L(\hat{\theta}) = \sum_{n_g=1..N_g\,;\,b=1..B}\ln\Big(\sum_q \frac{1}{Q}J_{n_g}[k_b](\hat{\theta}_q)\Big)$
12 **if** count > 1 **then**
13 $\Delta L = \frac{L(\hat{\theta})-L(\theta^*)}{L(\theta^*)}$
14 **end**
15 $\theta^* = \hat{\theta}$
  `Expectation step`
16 **for** $n_g = 1,\ldots,N_g,\ b = 1,\ldots,B$ **do**
17 **for** $q = 1,\ldots,Q$ **do**
18 $\bar{\gamma}_{n_g}^{(q)}[k] = \exp\big(J_{n_g}[k](\theta_q^*)\big)$
19 **end**
20 $\gamma_{n_g}^{(q)}[k] = \frac{\bar{\gamma}_{n_g}^{(q)}[k]}{\sum_{\ell=1}^{Q}\bar{\gamma}_{n_g}^{(\ell)}[k]}$
21 **end**
22 **end**

The second part of Algorithm 2 (steps 16–20) is a separation stage, for it consists in computing the allocation of the sources in the various time-frequency bins. Note that the logarithm of $\bar{\gamma}_{n_g}^{(q)}[k]$ is the log-likelihood of $\theta_q^*, S_{n_g}^{(q)}[k]^*$ with respect to $z_{n_g}[k]$, where $\theta_q^*, S_{n_g}^{(q)}[k]^*$ contributed to maximize the auxiliary function at the precedent iteration. Because of the "local" separability property enabling to express the most likely spectral parameters of the sources as functions of their azimuths, no initial guess is required for these parameters. This fact and the linearity of the global computational cost with the number of sources are commendable properties. ■

IV. MATLAB Experiments on Synthetic Data

A. Scenario

To assess the validity of the approach, results obtained from signals synthesized on MATLAB® are shown. From a database of French male and female 15 s speech records, the perceived binaural signals have been artificially generated by using the KEMAR dummy-head head-related impulse response (HRIR) measurements made available by MIT Media Lab[2]. HRIRs symmetrical w.r.t. the median plane, derived from the large pinna responses, have been used. Sources

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

have been placed in the azimutal plane ($0°$ elevation), with random azimuths uniformly distributed within the set $\mathcal{S} = \{-90°, -85°, \ldots, 90°\}$. Some non-iid noises, obtained by convolving a recorded fan noise with the $85°$ HRIRs, have been added to the left and right channels. The noise statistics have been learnt from a 2 s initial noise-only sequence. The signal-to-noise ratio (SNR) at the reference left microphone is about 33 ± 3 dB, and the power ratio between the most and least powerful sources ranges within $[0\,\text{dB}; 5\,\text{dB}]$.

As for the algorithm, FFTs are performed on Hanning windowed frames of 1024 samples at $F_s = 44.1$ kHz, with a 512-sample length overlap between frames. Sample covariance matrices are computed from $N_f = 4$ successive frames, and $N_g = 50$ groups of frames are used. Hence, the signal processed by the algorithm to produce each localization is approximately 2.3 s long. The frequencies bandwidth exploited for localization ranges from 0 to about 8 kHz. Though the algorithm works on 2.3 s signal sequences, a localization result is output at every available new group of N_f frames, *i.e.*, 46 ms. The EM iterations are stopped and a result is produced when the log-likelihood increase between two EM steps falls below $2\,\%$.

B. Evaluation criteria

The evaluation must take into account that the entries of the produced MLE $\hat{\theta}_{\text{ML}} = [\hat{\theta}_1, \ldots, \hat{\theta}_Q]'$ of the vector $\theta = [\theta_1, \ldots, \theta_Q]'$ of the genuine source azimuths appear in random order. So, all the possible entry permutations σ are first applied to $\hat{\theta}_{\text{ML}}$, and for each of them the Euclidean norm $\|\varepsilon_\sigma\|$ of the corresponding estimation error vector $\varepsilon_\sigma \triangleq (\sigma(\hat{\theta}_{\text{ML}}) - \theta)$ is computed. The permutation σ^* leading to $\varepsilon^* = \arg\min_\sigma \|\varepsilon_\sigma\|$ is kept.

Three criteria have been defined. c_1 and c_2 are the averages of the minimum and maximum values, respectively, of the entries of ε^* over all the experiments and localization outputs (*i.e.*, group of N_f frames). c_3 is defined as the percentage of the total number of entries of ϵ^* falling within $[-5°; 0°]$ or $[0°; +5°]$ over all the experiments and results. This criterion, similar to the "Recall" measure of [8], turns to be the ratio of the number of conveniently localized sources.

C. Results

Simulations have been performed with $Q = 2$ and $Q = 3$ sources. At each localization step, the algorithm is initialized with $\theta^{(\text{init})}$ uniformly drawn on $\mathcal{S}^Q$, with Q the true (known) number of sources. The results are as follows.

Q	c_1	c_2	c_3
2 sources	$1.7°$	$23.8°$	76%
3 sources	$3.0°$	$38.2°$	58%

As expected, the higher the number of sources, the worse the quality of the results.

The next table sketches the incidence of the initial guess on the results for an experiment involving $Q = 3$ sources. The above random initialization scheme is compared to the initialization of the EM algorithm at each time (*i.e.*, group of frames) with the localization estimated at the precedent localization time. The impact of initialization on performances is depicted below.

$\theta^{(\text{init})}$ at time $\#n_g$	c_1	c_2	c_3
Random	$1.8°$	$19.8°$	64%
$\hat{\theta}_{\text{ML}}$ output at time $\#(n_g - 1)$	$0.0°$	$1.0°$	97%

Additional sound files and videos of the simulated behavior throughout time are available on `http://homepages.laas.fr/danes/IROS2014`.

V. $\mathcal{HARK}$ Experiments on Real Data

$\mathcal{HARK}$, an english word for "listen", also stands for "HRI-JP Audition for Robots with Kyoto University". This is a comprehensive open source robot audition suite which provides several online implementations of canonical auditory functions such as localization, separation, or recognition [10]. An important feature of $\mathcal{HARK}$ is that it is oriented towards the transparent use of many devices. Sound files can be processed as well as if they were streamed signals. The overall design eases robotic audition prototyping and deployment.

A. $\mathcal{HARK}$ implementation

$\mathcal{HARK}$ relies on a the open-source middleware Flowdesigner which design complex functions, described on the basis of elementary blocks, whose interconnections depict the dataflow circulating between them (`http://sourceforge.net/apps/mediawiki/flowdesigner`). Since recent releases, $\mathcal{HARK}$supports the graphical user interface $\mathcal{HARK}$designer.

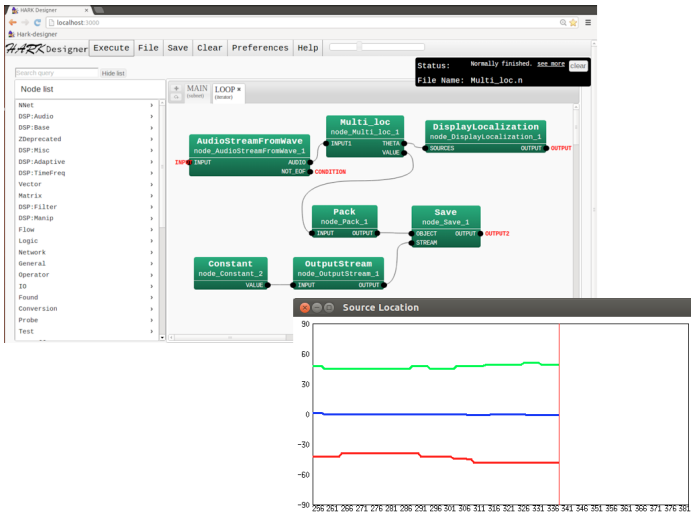
After being prototyped under MATLAB®, the aforementioned algorithms were coded into C++ so as to be encapsulated into $\mathcal{HARK}$, which enables online data processing. A new node called Multi_loc was created, see Figure 1-top. The input to this node is a stereo signal which can either be stored into an audio file, or streamed from microphones. The output port THETA delivers the localization result in the format defined by the Source class dedicated to the $\mathcal{HARK}$DisplayLocalization viewer. The port VALUE outputs the same data so that it can be stored in a text file, by the node Save. The STREAM input to this node specifies the path to the file.

The MAIN sheet of the $\mathcal{HARK}$ diagram reads the audio file and streams it into the LOOP sheet which handles a *frame* of the audio file. The online data processing is performed by the $\mathcal{HARK}$ node AudioStreamFromWave, which can be replaced by AudioStreamFromMic if microphones are used.

At each step of LOOP, the Multi_loc node performs the FFT of the new snapshot on the basis of the C fftw library so as to improve performance. The algorithm is configured so that the *frames* and *groups of frames* durations are 40 ms and 2 s, respectively.

B. Scenario and results

The $\mathcal{HARK}$ live experimentation of the algorithm, has been performed from audio files, recorded in an acoustically prepared (nearly anechoic) room at ISIR, Paris. Natural



1. Overview of the $\mathcal{HARK}$ implementation of the multiple source binaural localization into the node Multi_loc. $\mathcal{HARK}$ visualization, through the DisplayLocalization node, of the localization of three simultaneously uttering speakers at -45/0/45 degrees w.r.t. boresight.

sounds have been recorded by two microphones laid on a 176mm-diameter spheric head, by means of a NI 9234 acquisition module tuned at the 52100 Hz sampling frequency.

Three speakers equally distant to the head, are simultaneously uttering from $\{-45°; 0°; 45°\}$ with respect to boresight. As the problem is restricted to azimuthal plane, the speakers and microphones heights are similar to each other. Figure 1-bottom visualizes the result of the localization. As the utterances are not rigorously continuous, the localization may deviate a bit from the genuine location of the speakers. The means and standard deviations of the estimated source locations over 30 s of speech processing are $\{-40.4°; -0.8°; 47.1°\}$ and $\{7.6°; 5.3°; 7.3°\}$.

VI. Conclusion and Prospective Work

An EM algorithm for the localization of multiple sound sources from a binaural head has been presented. It is based on the sources WDO assumption, and relies on prior independent learnings of the interaural transfer function and the environment noise statistics. The algorithm has been prototyped and evaluated from synthetic data on MATLAB®, implemented on the open source $\mathcal{HARK}$ library and tested on some real data recorded in an anechoic room.

Ongoing works concern larger scale experiments from real data with ground-truth positions for a better global evaluation of the method against various conditions: reverberation, dynamic noise, etc. Fine tuning of the algorithm (selection of $L, N_f, N_g, B, ...$) has also been investigated. As an extension, source number detection will be included through statistical identification. The algorithm will be implemented on the EAR architecture. Longer term prospectives include integration of this azimuth detector into a stochastic filtering strategy, fusing a moving robot motor commands with audio perception to infer "active" multiple sound sources localization.

VII. Acknowledgements

The authors thank Dr. Sylvain Argentieri and colleagues (UPMC and ISIR-CNRS, Paris 06) who kindly provided audio sequences acquired in their acoustically prepared room.

References

[1] H. Okuno, T. Ogata, K. Komatani, and K. Nakadai, "Computational auditory scene analysis and its application to robot audition," in *IEEE Int. Conf. on Informatics Research for Development of Knowledge Society Infrastructure (ICKS'2004)*.

[2] V. Lunati, J. Manhès, and P. Danès, "A versatile system-on-a-programmable-chip for array processing and binaural robot audition," in *IEEE/RSJ Int. Conf. on Intell. Robots and Systems (IROS'2012)*.

[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. Springer Berlin Heidelberg, 2013, pp. 225–254.

[4] 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. Springer Berlin Heidelberg, 2013, pp. 397–425.

[5] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2CH BSS using the EM algorithm in reverberant environment," in *IEEE Wkshop on Appl. of Signal Processing to Audio and Acoustics 2007*.

[6] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. on Audio, Speech, and Lang. Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

[7] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[8] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," INRIA, http://hal.inria.fr/inria-00576297, Tech. Rep. RR-7566, 2011.

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

[10] K. Nakadai, T. Takahashi, H. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and implementation of robot audition system HARK," *Advanced Robotics*, vol. 24, no. 5–6, pp. 739–761, 2010.

[11] A. 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.

[12] A. Portello, P. Danès, S. Argentieri, and S. Pledel, "HRTF-based source azimuth estimation and activity detection from a binaural sensor," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'2013)*, Tokyo, Japan, 2013.

[13] A. Portello, "Active binaural localization of sound sources in humanoid robotics," Ph.D. dissertation, Univ. Toulouse III Paul Sabatier (in French), 2013.

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