The Cocktail Party Robot: Sound Source Separation and Localisation with an Active Binaural Head

Antoine Deleforge, Radu Horaud

The Cocktail Party Robot: Sound Source Separation and Localisation with an Active Binaural Head *

Antoine Deleforge
INRIA Grenoble Rhône-Alpes
655, Av. Europe, 38330 Montbonnot, France
Antoine.Deleforge@inria.fr

Radu Horaud
INRIA Grenoble Rhône-Alpes
655, Av. Europe, 38330 Montbonnot, France
Radu Horaud@inria.fr

ABSTRACT

Human-robot communication is often faced with the difficult problem of interpreting ambiguous auditory data. For example, the acoustic signals perceived by a humanoid with its on-board microphones contain a mix of sounds such as speech, music, electronic devices, all in the presence of attenuation and reverberations. In this paper we propose a novel method, based on a generative probabilistic model and on active binaural hearing, allowing a robot to robustly perform sound-source separation and localization. We show how interaural spectral cues can be used within a constrained mixture model specifically designed to capture the richness of the data gathered with two microphones mounted onto a human-like artificial head. We describe in detail a novel EM algorithm, we analyse its initialization, speed of convergence and complexity, and we assess its performance with both simulated and real data.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics—*Multivariate statistics, Probabilistic algorithms*; I.2.7 [**Natural Language Processing**]: Speech Recognition and Synthesis

General Terms

Algorithms, Theory, Experimentation

Keywords

Blind source separation, computational auditory scene analysis, EM algorithm, learning

1. INTRODUCTION

There is an increasing interest in robots able to communicate with people in the most natural way, e.g., Fig. 1.

*This work was supported by the European project HUMAVIPS, under EU grant FP7-ICT-2009-247525.

HRI'12, March 5–8, 2012, Boston, Massachusetts, USA.

Figure 1: A *cocktail party robot* should be able to communicate with people in the most natural way. One fundamental task that such a robot should be able to accomplish is to localize the speakers in a room and to separate their emitted speech signals, all in the presence of music, background noise, and reverberations.

Such robots must be endowed with the ability to reliably process and understand sensory inputs, e.g, visual or auditory data, gathered in unconstrained physical situations. Within the field of computational auditory scene analysis (CASA) tremendous progress was made in speech recognition. Nevertheless, current approaches work well with a single sound source often recorded with a close-range microphone. A more natural setup, e.g., humanoids with their own on-board microphones, implies to deal with much more complex and challenging acoustic inputs involving auditory data of various kinds (speech, prosody, music, electronic devices, etc.) and originating from sparsely located multiple sound sources, all in the presence of noise, attenuation and reverberations.

A classical example illustrating the difficulty of modeling such situations is the well known *cocktail party problem* (CPP) [9]. While human listeners solve this problem routinely and effortlessly, we note that it has not been properly addressed from the perspective of *robot audition*. Two key aspects of CPP are localization and separation of several sound sources. We believe that principled solutions to these problems are some of the prerequisites for addressing higher-level tasks in HRI such as speech and music recognition, verbal communication, dialog handling, etc.

In this paper we propose a new method for solving for sound-source separation and localization based on a gen-

erative probabilistic model and on *active binaural hearing*. More precisely, we promote a novel robot audition paradigm based on interaural spectral features, namely the *interaural level difference* (ILD) and the *interaural phase difference* (IPD) and on a constrained mixture model specifically designed to capture the richness of the binaural data recorded with a robot head endowed with a human-like *head related transfer function* (HRTF). We formally derive an EM procedure that alternates between separation (E-step) and localization (M-step). We show how our system can be fully automatically and efficiently trained using an audiomotor map. We analyse our algorithm in detail and we assess its performance with both simulated and real data.

There is behavioral and physiological evidence that humans use binaural cues in order to infer the direction of a sound. Two such cues seem to play an essential role, namely the ILD (already mentioned) and the *interaural time difference* (ITD). A number of computational models have been recently developed for robust sound localization and sound tracking based on ITD and/or ILD [19, 24]. However, it is well known that the spatial information provided by interaural-difference cues within a restricted band of frequency is spatially ambiguous, particularly along a roughly vertical and front/back dimension [15]. To avoid these ambiguities, more accurate sound localization models incorporate the HRTF, e.g. [12]. These approaches are based on the fact that the particular shape of the head, pinna and torso act as a filter depending on the emitting 3D location of the sound source (distance, azimuth, and elevation). However, HRTF databases are subject-specific and room-dependent (noise, reverberations, room geometry, etc.), and only a handful of HRTF databases are available in practice [1, 21]. This makes them hardly applicable to a real robotic application. To overcome these issues, the HRTF of a specific robot can be automatically learnt using *audio-motor maps* [8, 10]. Such maps are built by recording a static full-spectrum sound source from different motor states of the robot.

The problem of sound source separation has been thoroughly studied in the last decades and several interesting approaches were proposed. For example, [4, 20] and many others achieve separation with a single microphone, based on known acoustic properties of speech signals, and are therefore limited to a specific type of input. Other techniques such as independent component analysis (ICA) [6] or multi-microphone techniques require as many microphones as the number of sources. Several other methods use binaural localization cues for source separation [11, 14, 23]. In [11] acoustic inputs at different frequency channels are clustered over time by means of some assumptions on the emitted signals, and an HRTF data look-up table is used to find their corresponding positions in space. Once exact locations are known, up to two sources can be separated using the HRTF at each frequency channel.

Our method is based both on clustering (localization) and on spectral masking (separation). Spectral masking, also called binary masking, allows the separation of an arbitrary number of sources from a mixed signal, with the only assumption that a single source is active at every frequency-time point (f, t). This is referred to as the *W-disjoint orthogonality* assumption [25] and it has been shown to hold, in general, for simultaneous speech signals; It is particularly well suited for binaural recordings in realistic environments. Recently, [14] proposed a probabilistic model for multiple sound source separation based on interaural spatial cues. For each sound source, a binary mask and a discrete distribution over interaural time delays is provided. This can be used to approximate the azimuth angle of the source with a front-back ambiguity, if the distance between the microphones is known.

It appears that there are very few methods that formally combine 2D localization and separation. The first originality of our approach is to formally cast the localization and separation tasks into a generative probabilistic model that is solved very efficiently with a novel EM algorithm. Full details of this algorithm and its initialization as well as favorable comparisons with recent binaural-based separation methods are presented in a companion paper [7]. The second originality of our approach is to use an active robotic head in order to automatically learn *audio-motor maps* for a very large set of sound-source directions located in the far field. Such a training phase incorporates an implicit model of the HRTF and leads to a high precision both in terms of source separation and of localization. A thorough evaluation of the method with simulated and real data is provided, and puts forwards this approach as a promising future tool for auditory human-robot interactions.

The remainder of this paper is organized as follows: section 2 describes a binaural sound representation, section 3 presents in detail the formal model as well as its associated EM algorithm, section 4 describes the data acquisition and recording technique, section 5 presents our validation method, experiments and results. Concluding remarks and directions for future work are discussed in section 6.

2. BINAURAL SOUND REPRESENTATION

Both sound source localization and separation require a proper representation of the perceived data. For localization, one needs a content-independent representation that contains as much spatial information as possible. For separation, one needs a representation preserving all the richness of the original signals. In this paper we put forward two *binaural* representations, namely ILD and IPD spectrograms.

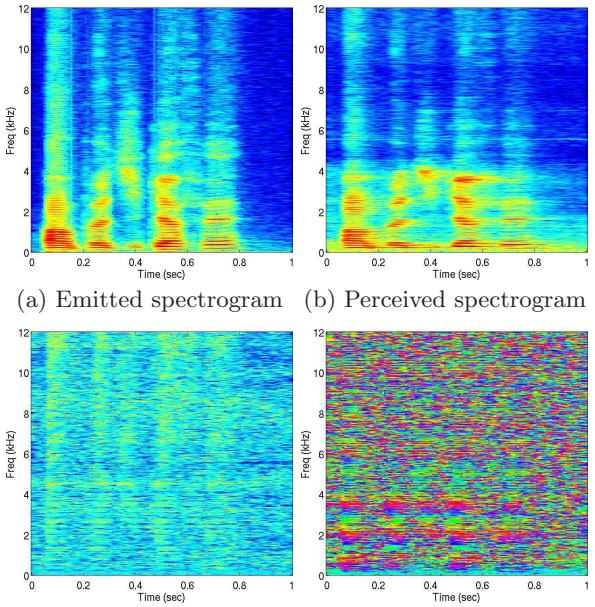
As already mentioned in section 1, our sound source separation method is based on binary masking. This technique consists in "filtering" the original signal by weighting all frequency-time points corresponding to the target source with 1 and all the other points with 0. Consequently, the perceived signal needs to be described within a time-varying spectral representation, a *spectrogram*. Spectrograms associated with each one of the two microphones are computed using a short-term fast Fourier transform (FFT) algorithm [2]. We use a time window of 64ms with 8ms overlap between two consecutive windows, thus yielding $T = 126$ time-windows for a signal lasting 1s. Since sounds are recorded at a sample rate of 24,000Hz, each time-window contains 1,536 samples, multiplied by a Hann window padded with zeros, for a total length of 2,048 samples. Each window is then transformed via FFT to obtain complex coefficients of $F = 1,024$ positive frequency channels between 0 and 12,000Hz. We denote with $s_{f,t}^{(k)} \in \mathbb{C}$ the spectrogram value at frequency-time point (f, t) of a signal emitted by a sound source k and with $s_{f,t}^{(\mathrm{L})}$ and $s_{f,t}^{(\mathrm{R})}$ the spectrogram values perceived by the left- and right-microphone respectively. The spectrogram of the *emitted acoustic level* is defined by:

$$a_{f,t}^{(k)} = 10 \log(|s_{f,t}^{(k)}|^2) \tag{1}$$

while the spectrogram of the *perceived acoustic level* is defined by

$$a_{f,t}^{(\mathrm{LR})} = 10\log(|s_{f,t}^{(\mathrm{L})}|^2 + |s_{f,t}^{(\mathrm{R})}|^2) \quad (2)$$

Fig. 2-(a) and (b) show an example of spectrograms of emitted and perceived acoustic levels.



(a) Emitted spectrogram (b) Perceived spectrogram

(c) ILD spectrogram (d) IPD spectrogram

Figure 2: Spectrograms corresponding to a male utterance emitted from the left-hand side of the binaural head.

The W-disjoint orthogonality assumption implies that a single sound source emits at a given frequency-time point (f,t). Let k be that source, the HRTF model provides a relationship between emitted and perceived spectrogram points at (f,t):

$$s_{f,t}^{(\mathrm{L})} = h^{(\mathrm{L})}(\boldsymbol{x}_k, f)s_{f,t}^{(k)} \text{ and } s_{f,t}^{(\mathrm{R})} = h^{(\mathrm{R})}(\boldsymbol{x}_k, f)s_{f,t}^{(k)} \quad (3)$$

where $\boldsymbol{x}_k \in \mathbb{R}^3$ denotes the 3D position of sound source k in a robot-centered coordinate frame (the origin of this frame is the midpoint between the two microphones and the x, y, and z axes pointing towards the left microphone, the head-top and in front of the head) and $h^{(\mathrm{L})}$ and $h^{(\mathrm{R})}$ denote the left and right HRTF. The latter functions depend on both the emitter's position and the frequency and they act as linear filters on the emitted signal. The HRTFs are mainly determined by the shape of the head, pinna and torso of the listener, e.g, the robot-mounted dummy-head in our case. The *interaural transfer function* (ITF) $\hat{I}$ is defined by the ratio between the left- and right-HRTF:

$$I(\boldsymbol{x}_k, f) = \frac{h^{(\mathrm{R})}(\boldsymbol{x}_k, f)}{h^{(\mathrm{L})}(\boldsymbol{x}_k, f)} \in \mathbb{C} \quad (4)$$

We can now define the interaural spectrogram as the ratio between the left- and right-microphone spectrograms:

$$\hat{I}_{f,t} = \frac{s_{f,t}^{(\mathrm{R})}}{s_{f,t}^{(\mathrm{L})}} \quad (5)$$

From (3), (4) and (5) one may notice that if the source k emitting at time-frequency point (f,t) is located at $\boldsymbol{x}_k$ then $\hat{I}_{f,t} \approx I(\boldsymbol{x}_k, f)$. The equality is only approximate due to the sensor noise and FFT errors. Therefore, at a given frequency-time point, the interaural spectrogram value $I_{f,t}$ does not depend on the emitted spectrogram value $s_{f,t}^{(k)}$ but only on the source position $\boldsymbol{x}_k$. We finally define the *ILD spectrogram* α and the *IPD spectrogram* ϕ as the log-amplitude and phase of the complex interaural spectrogram $\hat{I}_{f,t}$, e.g., Fig. 2-(c) and (d):

$$\alpha_{f,t} = 20\log|\hat{I}_{f,t}| \in \mathbb{R} \text{ and } \phi_{f,t} = \arg(\hat{I}_{f,t}) \in]-\pi, \pi] \quad (6)$$

3. SEPARATION AND LOCALIZATION

This section starts by describing how a training set of interaural parameters can be built out of a sound dataset annotated with sources positions. After formally stating the problem, we then introduce our novel probabilistic model for interaural parameters in a mixture of point sound sources. We finally detail a new version of the EM algorithm for this model, allowing to achieve simultaneous sound sources separation and localization based on learned positions.

3.1 Building a Training Set

We first explain how a training set of interaural parameters can be built out of a sound dataset annotated with sources positions. Let $\mathcal{X} = \{\boldsymbol{x}_n\}_{n=1}^N$ be a set of 3D coordinates in the robot centered frame that we already described above. If a single sound source n emits white noise from $\boldsymbol{x}_n \in \mathcal{X}$, let $\{\alpha_{f,t}^n\}_{f=1,t=1}^{F,T}$ and $\{\phi_{f,t}^n\}_{f=1,t=1}^{F,T}$ be the perceived ILD and IPD spectrograms. The *mean ILD vector* $\boldsymbol{\mu}(\boldsymbol{x}_n) = (\mu_1^n \dots \mu_f^n \dots \mu_F^n)^\top \in \mathbb{R}^F$ associated with $\boldsymbol{x}_n$ is defined by taking the temporal mean of α at each frequency channel:

$$\mu_f(\boldsymbol{x}_n) = \frac{1}{T}\sum_{t=1}^{T} \alpha_{f,t}^n \quad (7)$$

Similarly, the *mean IPD vector* $\boldsymbol{\xi}(\boldsymbol{x}_n) = (\xi_1^n \dots \xi_f^n \dots \xi_F^n)$ is the temporal mean of the IPD spectrogram ϕ at each frequency channel:

$$\xi_f(\boldsymbol{x}_n) = \arg\frac{1}{T}\sum_{t=1}^{T} \exp(j\phi_{f,t}^n) \in]-\pi, \pi] \quad (8)$$

The components of the mean IPD vector are estimated in the complex domain in order to avoid problems due to phase circularity, as suggested in [16]. White noise is used because it contains equal power within a fixed bandwidth at any center frequency: the source n is therefore the only source emitting at each frequency-time point (f,t), and μ_f and ξ_f are thus approximating the log-amplitude and phase of $\hat{I}(.,f)$. The set $\mathcal{X}$ of 3D sound-source positions as well as the mappings $\boldsymbol{\mu}$ and $\boldsymbol{\xi}$ will be referred to as the training data to be used in conjunction with the sound-source separation and localization algorithm described below.

3.2 Problem Formulation

We suppose that there are K sound sources randomly located in the robot's environment and that these K sources emit simultaneously sounds with unknown spectrograms from the unknown locations $\{\boldsymbol{x}_k\}_{k=1}^K \subset \mathcal{X}$. From the acoustic inputs perceived by the robot's microphone pair, one can build the ILD and IPD spectrograms $\{\alpha_{f,t}\}_{f=1,t=1}^{F,T}$ and $\{\phi_{f,t}\}_{f=1,t=1}^{F,T}$ as already described. The goal of the sound-source separation and localization algorithm is to associate

each perceived frequency-time point (f,t) with one of the sound sources and to estimate their 3D locations.

The observed $\alpha_{f,t}$ and $\phi_{f,t}$ are useful only if there is a significant signal at (f,t). To identify such *significant observations* we estimate the perceived acoustic level $a^{\mathrm{LR}}_{f,t}$ with (2) and we retain only those frequency-time points for which the acoustic level is above a threshold, $a^{\mathrm{LR}}_{f,t} > \epsilon_f$. One empirical way to choose the thresholds (one at each frequency) is to measure the highest perceived acoustic level at each f in the absence of emitting sound sources. These thresholds are typically very low compared to perceived acoustic levels due to natural sounds, and allow to filter out frequency-time points corresponding to the "room silence". We call *level mask* the associated spectral mask. We denote by M_f the number of significant observation at f and by $\alpha_{f,m}$ and $\phi_{f,m}$ the m-th significant ILD and IPD observations at f. The *observed data* will be denoted by $\mathbf{A} = \{\alpha_{f,m}\}_{f=1,m=1}^{F,M_f}$ and $\mathbf{\Phi} = \{\phi_{f,m}\}_{f=1,m=1}^{F,M_f}$.

We also introduce the *missing data*, i.e., a set of *unobserved variables* $\boldsymbol{z}_{f,m} \in \{0,1\}^K$, such that $z_{f,m,k} = 1$ if observations $\alpha_{f,m}$ and $\phi_{f,m}$ were generated by source k, and $z_{f,m,k} = 0$ otherwise. The W-disjoint orthogonality constraint can therefore be written as:

$$\sum_{k=1}^{K} z_{f,m,k} = 1 \quad \forall (f,m) \tag{9}$$

Variables $\boldsymbol{z}_{f,m}$ are also called *data-to-source assignments* at (f,m), and $\mathcal{M}_k = \{z_{f,m,k}\}_{f=1,m=1}^{F,M_f}$ corresponds to the binary spectral mask associated with sound-source k. Finally, we denote with $\mathbf{Z} = \{\boldsymbol{z}_{f,m}\}_{f=1,m=1}^{F,M_f}$ the set of all missing data. The problem of simultaneous localization and separation amounts to estimate the sound-source locations $\{\boldsymbol{y}_k\}_{k=1}^K$ and the masking variables $\mathbf{Z}$, given the number of sound sources K and the observed data $\mathbf{A}$ and $\mathbf{\Phi}$.

3.3 A Constrained Mixture Model

We assume that both the ILD and IPD observed data are drawn from normal distributions. The conditional likelihood of the observation $\alpha_{f,m}$ given its assignment to sound source k, i.e., $z_{f,m,k} = 1$ and located at $\boldsymbol{x}_k$ is therefore drawn from a 1D Gaussian distribution centered at $\mu_f(\boldsymbol{x}_k)$ and with variance $\sigma^2_{f,k}$:

$$\begin{aligned} P(\alpha_{f,m}|z_{f,m,k},\boldsymbol{x}_k,\sigma_{f,k}) &= \mathcal{N}(\alpha_{f,m}|\mu_f(\boldsymbol{x}_k),\sigma^2_{f,k}) \\ &= \frac{1}{(2\pi)^{1/2}\sigma_{f,k}} \exp\left(-\frac{(\alpha_{f,m}-\mu_f(\boldsymbol{x}_k))^2}{2\sigma^2_{f,k}}\right) \end{aligned} \tag{10}$$

For simplicity, the expression $z_{f,m,k} = 1$ is replaced by $z_{f,m,k}$ in probabilities. Since the IPD data lie on the circle $]-\pi,\pi]$, they should be modeled by a circular normal distribution. As proposed in [14], we approximate the wrapped normal distribution with a 1D Gaussian. The conditional likelihood of the observation $\phi_{f,m}$ given its assignment to sound source k located at $\boldsymbol{x}_k$ is given by:

$$\begin{aligned} P(\phi_{f,m}|z_{f,m,k},\boldsymbol{x}_k,\rho_{f,k}) &= \angle\mathcal{N}(\phi_{f,m}|\xi_f(\boldsymbol{x}_k),\rho^2_{f,k}) \\ &= \frac{1}{(2\pi)^{1/2}\rho_{f,k}} \exp\left(-\frac{\Delta(\phi_{f,m},\xi_f(\boldsymbol{x}_k))^2}{2\rho^2_{f,k}}\right) \end{aligned} \tag{11}$$

where $\rho^2_{f,k}$ is the IPD variance associated with source k at frequency f and the Δ function is defined by

$$\Delta(x,y) = \arg(e^{j(x-y)}) \in]-\pi,\pi] \tag{12}$$

The distribution $\angle\mathcal{N}(\xi,\rho^2)$ approximates the normal distribution on the circle $]-\pi,\pi]$ when ρ is small relative to 2π. Preliminary experiments on IPD spectrograms of white noise showed that this assumption holds in the general case.

Note that in our model, the source positions act as latent constraints on Gaussian means and thus *tight* the observations at different frequency channels. Such an approach can also be used to tight auditory and visual observations, as recently done in [13].

As emphasized in [14], the well known correlation between ILD and IPD does not contradict the assumption that Gaussian noises corrupting the observations are independent. Under this assumption, the conditional likelihood of the observed data $(\alpha_{f,m},\phi_{f,m})$ given its assignment to source k located at $\boldsymbol{x}_k$ and with variances $\sigma^2_{f,k}$ and $\rho^2_{f,k}$ is

$$\begin{aligned} &P(\alpha_{f,m},\phi_{f,m}|\boldsymbol{z}_{f,m},\boldsymbol{x}_k,\sigma_{f,k},\rho_{f,k}) = \\ &\mathcal{N}(\alpha_{f,m}|\mu_f(\boldsymbol{x}_k),\sigma^2_{f,k}) \times \angle\mathcal{N}(\phi_{f,m}|\xi_f(\boldsymbol{x}_k),\rho^2_{f,k}) \end{aligned} \tag{13}$$

The prior probability of assigning an observation to a sound source writes:

$$P(z_{f,m,k}) = \pi_{f,k} \tag{14}$$

Finally, we denote by Θ the set of all the parameters of our binaural mixture model:

$$\Theta = \left\{\{\boldsymbol{x}_k\};\ \{\pi_{f,k}\};\ \{\sigma^2_{f,k}\};\ \{\rho^2_{f,k}\}\right\}_{f=1,k=1}^{F,K} \tag{15}$$

3.4 The Separation-Localization Algorithm

The problem of both source separation and source localization can now be expressed as an optimal parameter estimation problem, namely the maximization of the observed-data log-likelihood over the parameters Θ:

$$\widetilde{\Theta} = \arg\max_{\Theta} \mathcal{L}(\mathbf{A},\mathbf{\Phi};\Theta) \tag{16}$$

In order to keep the model as general as possible, there is no assumption on the emitted sounds as well as the way their spectra are spread across the frequency-time points. Therefore, we assume that all the observations are independent:

$$\begin{aligned} \mathcal{L}(\mathbf{A},\mathbf{\Phi};\Theta) &= \log P(\mathbf{A},\mathbf{\Phi}|\Theta) \\ &= \sum_{f=1}^{F}\sum_{m=1}^{M_f} \log P(\alpha_{f,m},\phi_{f,m}|\Theta) \end{aligned} \tag{17}$$

It is well established that the direct optimization of (17) is difficult because of the presence of many local maxima. Therefore we recast the problem within the framework of *maximum likelihood with missing data*, i.e., $\mathbf{Z}$ in (13), which is traditionally solved via expectation-maximization (EM). The EM algorithm alternates between estimating the *expected complete-data log-likelihood* (E-step) using the current model parameters and maximizing this likelihood over the model parameters given the current posterior probabilities (M-step). The posterior probability $r_{f,m,k}$ of $z_{f,m,k}$ is defined by:

$$r_{f,m,k} = P(z_{f,m,k}|\alpha_{f,m},\phi_{f,m};\Theta) \tag{18}$$

As it will be detailed below, the E-step of the proposed algorithm computes the posterior probabilities of assigning each

spectrogram point to a sound source k (*separation-step*), while the M-step maximizes the expected complete-data log-likelihood with respect to the model parameters Θ, namely the priors, the variances and, most notably, the source locations $\{\boldsymbol{x}_k\}_{k=1}^K$ (*localization-step*).

One of the most interesting properties of EM algorithms is that the log-likelihood $\mathcal{L}$ in (16) is increased at each EM iteration and hence it converges to a local maximum. An outline of the proposed EM method is provided in Algorithm 1. The finally estimated posterior probabilities allow to compute the binary spectral masks $\mathcal{M}_k$ associated with each source k while the final parameters $\boldsymbol{x}_k$ provide estimates for the source locations. $v^{(p)}$ denotes the value of variable v at iteration p, and $v^{(\mathrm{P})}$ denotes its final value. The E-step, M-step, initialization and convergence check are detailed below.

Algorithm 1 Separation-Localization EM

Input: $\mathbf{A}$, $\mathbf{\Phi}$, $\{\boldsymbol{\mu}(\boldsymbol{x}_n)\}_{n=1}^N$, $\{\boldsymbol{\xi}(\boldsymbol{x}_n)\}_{n=1}^N$, K
Output: $\{\boldsymbol{x}_k^{(\mathrm{P})}\}_{k=1}^K$, $\{\mathcal{M}_k\}_{k=1}^K$
$\Theta^{(0)}$:= **initialize**
$p := 0$
while **!converged do**
  $p := p + 1$
  $\{r_{f,m,k}^{(p)}\}$:= **E-step**$(\Theta^{(p-1)})$
  $\Theta^{(p)}$:= **M-step**$(\{r_{f,m,k}^{(p)}\}_{f,m,k})$
end while
$\mathcal{M}_k := (k == \arg\max_{k'} r_{f,m,k'}^{(\mathrm{P})})_{f=1,m=1}^{F,M_f}$

Based on Bayes' formula and on marginalization rules, the E-step computes the current posterior probabilities conditioned by the previously estimated parameters, i.e., (13):

$$r_{f,m,k}^{(p)} := \frac{\pi_{f,k} P(\alpha_{f,m}, \phi_{f,m} | \boldsymbol{z}_{f,m}; \Theta^{(p-1)})}{\sum_{i=1}^K \pi_{f,i} P(\alpha_{f,m}, \phi_{f,m} | \boldsymbol{z}_{f,m}; \Theta^{(p-1)})} \tag{19}$$

The expected complete-data log-likelihood can now be written as:

$$\begin{aligned} Q(\Theta|\Theta^{(p-1)}) &= E_{(\mathbf{Z}|\mathbf{A},\mathbf{\Phi};\Theta)}[\log P(\mathbf{A}, \mathbf{\Phi}, \mathbf{Z}|\Theta)] \\ &= \sum_{f=1}^F \sum_{m=1}^{M_f} \sum_{k=1}^K r_{f,m,k}^{(p)} \log \pi_{f,k} P(\alpha_{f,m}, \phi_{f,m} | \boldsymbol{z}_{f,m}; \Theta) \end{aligned} \tag{20}$$

The M-step maximizes (20) with respect to Θ:

$$\Theta^{(p)} := \widetilde{\Theta} = \arg\max_{\Theta} Q(\Theta|\Theta^{(p-1)}) \tag{21}$$

By combining (13) with (20) the problem becomes equivalent to minimizing:

$$\begin{aligned} \sum_{f=1}^F \sum_{m=1}^{M_f} r_{f,m,k}^{(p)} \Bigg(\log\left(\frac{\sigma_{f,k}^2}{\pi_{f,k}}\right) + \log\left(\frac{\rho_{f,k}^2}{\pi_{f,k}}\right) + \\ \frac{(x_{f,m} - \mu_f(\boldsymbol{x}_k))^2}{\sigma_{f,k}^2} + \frac{\Delta(\phi_{f,m}, \xi_f(\boldsymbol{x}_k))^2}{\rho_{f,k}^2} \Bigg) \end{aligned} \tag{22}$$

which can be easily differentiated with respect to $\{\pi_{f,k}\}_f$, $\{\sigma_{f,k}\}_f$ and $\{\rho_{f,k}\}_f$ to obtain closed-form expression, conditioned by $\boldsymbol{x}_k$, for the optimal values of these parameters. These expressions are then subtituted in (22), which is evaluated for all $\boldsymbol{x}_k \in \mathcal{X}$ to find the optimal position $\hat{\boldsymbol{x}}_k$, and deduce all the other optimal parameters. Interested readers can find a detailed solution in [7].

As already mentioned, EM converges to a local maximum of the observed data log-likelihood function $\mathcal{L}$. However, the non-injectivity of the interaural functions μ_f and ξ_f leads to a very large number of these maxima, especially when the set of learned positions $\mathcal{X}$, i.e., section 3.1, is large. This makes the algorithm to be very sensitive to initialization. A common way to avoid being trapped in local maxima may be to initialize the parameters at random, but such a strategy cannot be directly applied here: First, because the cardinality of the parameter set Θ is very large and second, because there is no straightforward way to initialize the variances $\sigma_{f,k}^2$ and $\rho_{f,k}^2$. Another possibility may be to randomly initialize the source assignment variables $\mathbf{Z}$ and then proceed with the M-step, but extensive experiments with simulated data revealed that the algorithm very rarely converged to a global maximum (in less than 0.1% of the cases). We therefore decided to adopt a method that combines these two initialization strategies by randomly perturbating both the source locations and the source assignments.

This lead us to develop a *stochastic initialization* procedure similar to the stochastic EM (SEM) algorithm [5]. The idea of exploiting stochasticity to escape from local maxima is a commonly used principle in global optimization [26]. The SEM algorithm includes a stochastic step (S) in between the E and the M steps, during which random samples $R_{f,m,k} \in \{0, 1\}$ are drawn from the posteriors $r_{f,m,k}$. These samples are then used instead of the posterior probabilities during the M-step. To initialize our algorithm, we first set all the posterior probabilities to $1/K$ and then proceed through the following step sequence: S M* E S M, where the M*-step is a variation of the M-step in which the sources' positions are drawn randomly from $\mathcal{X}$ instead of computing $\{\hat{\boldsymbol{x}}_k\}_k$. Experiments with simulated data showed that this technique converged to a global optimal solution in over 10% of the cases. More advanced techniques may also be used to improve the convergence rate, and are detailed in [7].

In practice, twenty stochastic initializations are used in order to increase the chances of correct convergence, and only the one providing the best log-likelihood after two iterations is eventually iterated until the convergence criteria is satisfied. The algorithm stops either when the log-likelihood gain is less than 1%, or after $p_{\max} = 20$ iterations. Indeed, from two hundred simulated experiments (see section 5.2), we concluded that, on an average, the algorithm converges in eleven iterations.

3.5 Algorithm Complexity

Both the E- and S-steps are linear in the total number of significant observations, $\sum_f M_f$ and in the number of sound sources, K. The M-step is linear in the number of frequency channels F and in the number N of source locations available in the training set $\mathcal{X}$. In the case of simulated data using one-second long signals composed of two sources, the values of these parameters are: $K = 2$, $F = 1024$, $\sum_f M_f = 70,000$, and $N = 10,800$. With a Matlab implementation executed on a 2.53GHz Intel-Xeon processor, we obtained the following average running times: 49ms (E), 1030ms (S), 2520ms (M) and 24ms (M*). The most time-consuming part of the algorithm is its initialization (20 stochastic initializations iterated 2 times) which takes 195.9 seconds, while each subsequent EM iteration

takes 2.569 seconds, amongst which 98% corresponds to selecting all optimal $\hat{\boldsymbol{x}}_k$ in $\mathcal{X}$. Careful algorithm and software optimization will allow, however, to obtain realistic execution times needed for human-robot interactions scenarios, e.g., one to two seconds.

4. DATA ACQUISITION

In order to build the training set introduced in section 3.1, we developed a technique to learn a large number of sound source locations in an entirely unsupervised and automated way using the motor system of a binaural robot head. This technique is initially inspired from the sensorimotor theories of early development in psychology, suggesting that experiencing the sensory consequences of voluntary motor actions was necessary for an organism to learn the perception of space [17]. In particular [3] argued that naive organisms such as humans and echo-locating bats could learn sound localization based solely on acoustic inputs and their relation to motor states. This idea was experimentally validated using a robot system in [8].

Figure 3: A binaural head is placed onto an agile device that can perform precise and reproducible pan and tilt motions (left). The emitter (a loud-speaker) is placed in front of the robot head at approximately 2.7 meters (right).

Sound acquisition is performed with a Sennheiser MKE 2002 acoustic dummy-head linked to a computer via a Behringer ADA8000 Ultragain Pro-8 digital external sound card. The head is mounted onto a robotic system with two rotational degrees of freedom: a pan motion and a tilt motion (see Fig. 3). This device was specifically designed to achieve precise and reproducible movements. The emitter – a loud-speaker – is placed at approximately 2.7 meters ahead of the robot, as showed on Fig. 3. The loud-speaker's input and the microphones' outputs were handled by two synchronized sound cards in order to simultaneously record and play. All the experiments were carried out in real-world conditions, i.e., a room with natural reverberations and background noise due to computer fans. All the recordings are publicly available[1]. We believe that such a large audio-motor data set has no equivalent today, and is therefore a contribution in its own right.

Rather than placing the emitter at known 3D locations around the robot, it was kept in a fixed reference position while the robot recorded emitted sounds from different motor states. Consequently, a sound source direction is directly associated to a pan-tilt motor state (ψ, θ) rather than a 3D point in space. A robot trained in such a way would therefore be able to perform the head movement pointing toward an emitting sound source in an entirely unsupervised way, without needing the inverse kinematics, the distance between microphones, or any other parameters.

Recordings were made from $10,800$ uniformly spread motor states: 180 pan rotations ψ in the range $[-180^\circ, 180^\circ]$ (left-right) and 60 tilt rotations θ in the range $[-60^\circ, 60^\circ]$ (top-down). The associated set of 3D source positions $\mathcal{X} \subset \mathbb{R}^3$ could be deduced using the direct kinematic model of the robot given in [8]. However, the speaker was located in the far field of the head during experiments (> 1.8 meters), and [18] showed that HRTFs mainly depend on the sound source direction while the distance has fewer impact in that case. That is why sound source locations will be expressed with angles in the rest of the paper.

At each motor state, five binaural recordings of one second each were made while the speaker emitted different sounds. Sound 1 corresponds to white noise, and was used to build the training set (section 3.1). Sounds 2, 3 and 4 form the test set (see section 5.3), and correspond respectively to a woman pronouncing *"Bonjour!"*, a man pronouncing *"Un petit café?"*, and a flute melody. Sound 0 corresponds to "room silence" and was used to determine the perceived acoustic level threshold ϵ_f during tests (see section 3.2).

5. EXPERIMENTS AND RESULTS

5.1 Performance Evaluation

Algorithm 1 was tested and validated using the database described in the previous section. In all presented results, a source is considered as *correctly localized* if the algorithm could locate it within 2° of absolute angular error in both azimuth and elevation. For real mixtures (section 5.2), the separation was evaluated using standard metrics, namely Source-to-Distortion Ratio (SDR) and Source-to-Interferences Ratio (SIR), proposed in [22]. These metrics are based on the decomposition of the estimated signal into the target signal, the error term coming from other interfering signals, and the error term coming from unexplained artifacts. Another term can be added to evaluate background noise reduction, but this is outside the scope of this work. Once the decomposition is achieved, the SDR is defined by the ratio of energy between the target and the error terms, while the SIR is the ratio of energy between the target and the interference term only. Both these metrics are expressed in decibels. As they are computed from 1D signals, the masked left and right microphone spectrograms were converted back to temporal signals using the inverse FFT, and concatenated to evaluate the quality of binaural-based separation.

SDR and SIR scores obtained with our algorithm are given together with upper and lower bounds. The upper bound corresponds to the SDR of the ground truth mask or *oracle mask* [25], which is set to 1 at every spectrogram point in which the target signal is at least as loud as the combined other signals and 0 everywhere else. The lower bound corresponds to the SDR and SIR in the original mixture. The level mask described in section 3.2 was applied to all signals so that only separation of significant observations could be evaluated.

For tests made on simulated spectrograms, since there is no 1D signal to compare with, the separation was evaluated by a pointwise $\neq$ operation between estimated and oracle binary masks: we define the *mask error* by the ratio of points in the estimated mask that differs from the oracle mask.

[1] `http://perception.inrialpes.fr/~Deleforge/CAMIL_Dataset/`

5.2 Simulated Data

To validate our model we first tested the algorithm on simulated data. We simulated ILD and IPD spectrograms with $F = 1024$ frequency channels from 0 to 12,000Hz, and 10 to 126 significant observations per channel (M_f randomly drawn for each f). First, K source positions $\boldsymbol{x}_1 \dots \boldsymbol{x}_K$ are randomly drawn from $\mathcal{X}$. Then, each spectrogram point (f, m) is randomly assigned to one of the K sources, and we generate ILD and IPD observations $\alpha_{f,m}$ and $\phi_{f,m}$ using the distributions $\mathcal{N}(\mu_f(\boldsymbol{x}_k), \sigma_f^2(\boldsymbol{x}_k))$ and $\angle\mathcal{N}(\xi_f(\boldsymbol{x}_k), \rho_f^2(\boldsymbol{x}_k))$, where $\sigma_f^2(\boldsymbol{x}_k)$ and $\rho_f^2(\boldsymbol{x}_k)$ correspond to ILD and IPD variances of white noise emitted from $\boldsymbol{x}_k$, and are estimated from the dataset. A total of 200 mixtures composed of 2 to 5 sources were generated, and K was set to the correct number of sources for each test. Table 1 shows the percentage of correctly localized sources, the mean mask error obtained over all sources, and the mean mask error obtained with randomly generated binary masks.

Table 1: Mean localization and separation results for 200 simulated mixtures of 2 to 5 sources.

K	Correctly Localized(%)	Estimated Mask Error(%)	Random Mask Error(%)
2	99.2	20.7	50.0
3	91.2	25.4	55.6
4	54.1	28.9	62.5
5	14.4	29.3	68.0

These results show that our model is very well suited for both separation and localization tasks. A very high localization accuracy is achieved for mixtures composed of 2 and 3 sources. When the number of sources is higher, the number of observations per source decreases while the number of local maxima of the log-likelihood function increases, leading to poorer results.

5.3 Real Mixtures from Learned Positions

Three sounds were used to test the algorithm with real data: male speech, female speech, and flute melody (see section 4). Mixtures were obtained by summing raw binaural signals from the dataset, corresponding to randomly drawn positions in $\mathcal{X}$. ILD and IPD spectrograms were computed from these mixtures, and only significant observations corresponding to points with a higher acoustic level than the corresponding background mixtures were kept, as described in section 2.

First, 200 mixtures for each possible pair of test sounds were generated, resulting in 1,200 localization-separation tasks. Mean results obtained with our method for all these tasks as well as mean results for correctly localized sources only are shown in Table 2. SDR and SIR scores are compared to those of the original mixture and oracle mask.

A very high localization rate is obtained, with 84.6% of the 1,200 test sound sources localized with less that 2° error in both azimuth and elevation. In addition, significant improvements of SDR and SIR ratios over the original mixtures are achieved. This is particularly true for SIR ratios that almost reached oracle ratios in some tests. This shows that although our algorithm generates some artifacts in the target sound signal during the spectral masking process, it is able to considerably reduce the volume of the interferer and thus significantly improve the perceived quality. Finally, we note that SDR and SIR ratios are always better when the algorithm could correctly localize the sound source, which demonstrates the importance of localization cues for sound sources separation.

Table 2: Mean angle error, SDR (Source-to-Distortion Ratio) and SIR (Source-to-Interferences Ratio) for real mixtures of 2 sound sources.

	Ang Err(°)	SDR(dB)	SIR(dB)
Original	-	0.05	0.05
Us (All sources)	**11.5**	**4.11**	**14.3**
Us (Loc: 84.6%)	**0.18**	**5.17**	**15.2**
Oracle	-	17.9	39.6

A second experiment was made with 200 mixtures of the three test sounds emitting altogether from random positions in $\mathcal{X}$, resulting in 600 localization-separation tasks. Results are displayed in a similar fashion in Table 3.

Table 3: Mean angle error, SDR and SIR for real mixtures of 3 sound sources.

	Ang Err(°)	SDR(dB)	SIR(dB)
Original	-	-3.61	-3.61
Us (All sources)	**35.5**	**-4.03**	**3.21**
Us (Loc: 45.2%)	**0.18**	**0.09**	**7.29**
Oracle	-	14.5	32.7

Although performances significantly decrease for this very challenging task (a 1s mixture of 3 equally loud sounds is hard to decipher even for humans), we note that the algorithm could still accurately localize almost half of the 600 individual sources, while significantly improving their SDR and SIR with respect to the original mixture.

5.4 Real Mixtures from Unlearned Positions

The last experiment is also the most challenging one. Indeed, we tested the ability of our algorithm to separate sound sources emitting from positions outside the training set. More precisely, we made several test recordings in which two loud speakers were simultaneously emitting different sounds while the robot stayed in its reference position $(0°, 0°)$. One of the loud speaker was placed in a *frontal* position corresponding to the training dataset, while the second one was manually placed in 21 *side* positions around the robot, with a 10° to 90° azimuth distance and 0° to 30° elevation distance. These experiments were repeated for the six possible pairs of test sounds, resulting in 252 separation-localization tasks. The frontal loud speaker was correctly localized at $(0°, 0°)$ in 95.2% of the mixtures tested. Localization performances for the side loud speaker were not evaluated, as no ground truth was available. Table 4 shows the mean SDR and SIR scores of each source obtained with our approach, the original mixture, and the oracle mask.

Despite a decrease of performances with respect to sound mixtures from learned positions, the significant improvement of SIR obtained in realistic cocktail-party like scenarios puts forward our approach as a promising tool for auditory human-robot interactions.

Table 4: Mean SDR and SIR of frontal and side speakers

	Frontal SDR(dB)	Frontal SIR(dB)	Side SDR(dB)	Side SIR(dB)
Original	-0.28	-0.28	0.32	0.32
Us	**-1.09**	**3.01**	**-3.30**	**5.69**
Oracle	18.29	40.64	18.63	40.13

6. CONCLUSION AND FUTURE WORK

Traditionally, computational auditory scene analysis was addressed with either a close-range or an array of microphones and using simulated or anechoic audio data. We propose to bridge the gap between constrained and unconstrained audio analysis and to apply CASA to HRI. We propose a system integrating the audio-motor abilities of a robot within a unified framework, and performing sound-source separation and localization in a realistic cocktail-party like scenario. By providing a genuine audio-motor database and presenting encouraging results obtained from these data, we presented a benchmark for the unexplored field of sensorimotor learning for robot audition.

One of the most interesting and promising directions will be to extend our model to a continuous space of sound source positions. This could be done using the manifold structure of interaural parameters studied in detail in [8]. By approximating this manifold by local tangent spaces, the size of the training set could be considerably reduced, thus speeding up the M-step, while improving the localization of sound sources from unknown places. Dynamic models incorporating moving sound sources and head movements could also be included based on this idea. Finally, a "garbage" source class could be added to our model in order to better deal with background and non-point sources. We believe that these ideas combined with careful algorithm and software optimization could lead to a novel *robot hearing* paradigm within the emerging field of human-robot interaction.

References

[1] R. V. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF Database. *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 92–102, Oct. 2001.

[2] J. Allen. Short-term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Trans. Acous., Speech and Signal Process.*, 25(3):235–238, 1977.

[3] M. Aytekin, C. F. Moss, and J. Z. Simon. A sensorimotor approach to sound localization. *Neural Computation*, 20(3):603–635, 2008.

[4] S. Bensaid, A. Schutz, and D. T. M. Slock. Single microphone blind audio source separation using EM-Kalman filter and short+long term AR modeling. In *Latent Variable Analysis and Signal Separation*, pages 106–113, 2010.

[5] G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3):315–332, 1992.

[6] P. Comon and C. Jutten. *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press (Elsevier), Feb. 2010.

[7] A. Deleforge and R. Horaud. A latently constrained mixture model for audio source separation and localization. In *Latent Variable Analysis and Signal Separation*, Tel Aviv, Israel, March 2012.

[8] A. Deleforge and R. P. Horaud. Learning the direction of a sound source using head motions and spectral features. Technical Report RR-7529, INRIA, Feb. 2011.

[9] S. Haykin and Z. Chen. The cocktail party problem. *Neural Computation*, 17:1875–1902, 2005.

[10] J. Hörnstein, M. Lopes, J. Santos-Victor, and F. Lacerda. Sound localization for humanoid robots – building audio-motor maps based on the HRTF. In *Proc. of IEEE/RSJ IROS*, pages 1170–1176, 2006.

[11] F. Keyrouz, W. Maier, and K. Diepold. Robotic localization and separation of concurrent sound sources using self-splitting competitive learning. In *Proc. of IEEE CIISP*, pages 340–345, Hawaii, Apr. 2007.

[12] F. Keyrouz, Y. Naous, and K. Diepold. A new method for binaural 3D localization based on HRTFs. In *Proc. of IEEE ICASSP*, volume 5, May 2006.

[13] V. Khalidov, F. Forbes, and R. P. Horaud. Conjugate mixture models for clustering multimodal data. *Neural Computation*, 23(2):517–557, Feb. 2011.

[14] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis. Model-based expectation-maximization source separation and localization. *IEEE Trans. on Audio, Speech and Lang. Proc.*, 18:382–394, Feb. 2010.

[15] J. C. Middlebrooks and D. M. Green. Sound localization by human listeners. *Annual Review of Psychology*, 42:135–159, January 1991.

[16] J. Nix and V. Hohmann. Sound source localization in real sound fields based on empirical statistics of interaural parameters. *Journal of the Acoustical Society of America*, 119(1):463–479, 2006.

[17] J. K. O'Regan and A. Noe. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24:939–1031, 2001.

[18] M. Otani, T. Hirahara, and S. Ise. Numerical study on source-distance dependency of head-related transfer functions. *Journal of the Acoustical Society of America*, 125(5):3253–61, 2009.

[19] N. Roman and D. Wang. Binaural tracking of multiple moving sources. *IEEE Trans. on Acoust., Speech and Signal Process.*, 16(4):728–739, 2008.

[20] S. T. Roweis. One microphone source separation. In *Advances in Neural Information Processing Systems*, volume 13, pages 793–799. MIT Press, 2000.

[21] B. Shinn-Cunningham, N. Kopco, and T. J. Martin. Localizing nearby sound sources in a classroom: Binaural room impulse responses. *Journal of the Acoustical Society of America*, 117(5):3100–3115, 2005.

[22] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech & Language Processing*, 14(4):1462–1469, 2006.

[23] H. Viste and G. Evangelista. On the use of spatial cues to improve binaural source separation. In *Proc. Int. Conf. on Digital Audio Effects*, pages 209–213, 2003.

[24] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Koerner. A probabilistic model for binaural sound localization. *IEEE Transactions on Systems, Man, and Cybernetics–Part B*, 36(5):982–994, 2006.

[25] O. Yılmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52:1830–1847, 2004.

[26] A. Zhigljavsky and A. Žilinskas. *Stochastic Global Optimization*. Springer, 2008.