Engineering Applications of Artificial Intelligence 69 (2018) 137–146

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Binaural lateral localization of multiple sources in real environments using a kurtosis-driven split-EM algorithm

P. Reche-Lopez [a,*], J.M. Perez-Lorenzo [b], F. Rivas [b], R. Viciana-Abad [a]

[a] Advanced Studies Center in Information and Communication Technologies (CEATIC), University of Jaen, Spain
[b] Multimedia & Multimodal Processing Group (M2P), University of Jaen, Campus Científico-Tecnológico de Linares, Avda. de la Universidad, s/n - 23700 Linares, Jaén, Spain

ARTICLE INFO

Keywords:
Robot audition
Multiple sound localization
Laplacian model mixture
KDS-EM
Mutational split

ABSTRACT

In this work a method for an unsupervised lateral localization of simultaneous sound sources is presented. Following a binaural approach, the kurtosis-driven split-EM algorithm (KDS-EM) implemented is able to estimate the direction of arrival of relevant sound sources without knowing a priori their number. Information about the localization is integrated within a period of observation time to serve as an auditory memory in the context of social robotics. Experiments have been conducted using two types of observation times, one shorter with the purpose of analyzing its performance in a reactive level, and other longer that allows the analysis of its contribution as an input of the building process of the sorroundings auditory models that serves to drive a more deliberative behavior. The system has been tested in real and reverberant environments, achieving a good performance based on an over-modeling process that is able to isolate the location of the relevant sources from adverse acoustic effects, such as reverberations.

1. Introduction

Hearing is a prominent sense for communication and socialization (Argentieri et al., 2015). Thus, auditory capabilities should be strengthened in those robots that aim to present a social behavior. Moreover, for both human and robots auditory capabilities, the stage of sound localization is one of the most important low-level auditory function (Argentieri et al., 2013). Human hearing allows to localize and distinguish individual sources from a complex mixture of sounds, which allows for example to listen to, and follows, one speaker in the presence of others (situation that has been called "the cocktail party problem" (Cherry, 1953, 1957). Indeed, following the approach of mimic human hearing system (bioinspired approach), research in robot audition has became an important topic for robotics (Argentieri et al., 2015, 2013). The implementation of such ability in a social robot would allow to establish conversations not only in noisy situations with several speakers, but also in residential environments with sounds originated from audio devices such as TV and radio equipments, and in presence of typical background noises that may be disturbing in Human–Robot Interaction (HRI). Moreover, robot audition can be exploited with the goal of building an acoustic map, which in turn is a typical approach widely used with other sensor inputs in robotics, such as visual information, odometry and laser data. Thus, an acoustic map can be used as an input to higher layers of an attentional mechanism, which is useful not only improving the interaction capabilities with a human speaker, but also in order to move towards a source of interest, track a mobile source, or merge the information with visual features (Ferreira et al., 2013; Viciana-Abad et al., 2014).

The localization of multiple sound sources is still an open problem in the Computational Auditory Scene Analysis field (CASA) (Wang and Brown, 2006), especially when two microphones are used with the aim of emulating the human auditory system. Within the scope of localization, binaural cues such as interaural time and level differences, together with monaural cues, can be used to determine azimuth and elevation angles, and even the distance to the source under certain conditions. The lateral localization cues are related with the interaural time difference (ITD) (Argentieri et al., 2015; Wang and Brown, 2006; Stern et al., 2006), since the wave sound arrives slightly earlier in time at the ear (or microphone) that is closer to the source, unless the source is located directly in front of the head. This difference is also known as time difference of arrival (TDOA) (May et al., 2013; Cobos et al., 2011) because it depends on the direction of arrival of the wave sound, and thus on the lateral localization of the source relative to the microphones.

* Corresponding author.
E-mail address: pjreche@ujaen.es (P. Reche-Lopez).

https://doi.org/10.1016/j.engappai.2017.12.013
Received 23 June 2017; Received in revised form 29 September 2017; Accepted 28 December 2017
Availableonline 12 January 2018
0952-1976/

The estimation of the ITD is probably the most critical aspect of binaural processing. Many models of binaural processing are based on the cross correlation of the signals arriving to the two ears after being processed by the auditory periphery, or based on other functions that are closely related to cross correlation (Stern et al., 2006). The notion of an interaural cross-correlation mechanism is broadly supported by physiological studies, which have revealed systematic arrangement of ITD-sensitive neurons in the auditory midbrain (Wang and Brown, 2006). While the first reported cells to be maximally sensitive to signals presented with a specific ITD were placed in the inferior colliculus in the brain stem, cells with a similar response have been later reported in other parts of it (Stern et al., 2006). However, although in some cases the goal is to better understand the human auditory system, and thus it is important to accurately simulate the functionality or even the physiological structure of the auditory pathway, this is not mandatory in the case of robotics. Indeed, in robotics, usually the knowledge about the human hearing system is only applied as far as it can improve the system performance (Kohlrausch et al., 2013).

Binaural signals have been already successfully used for the direction estimation of concurrent speakers (Dietz et al., 2010; Nikunen and Diment, 2016). Indeed, the source separation achieved by using direction of arrival (DOA) algorithms together with beamforming techniques has also allowed increasing the intelligibility in scenarios with more than one speaker (Nikunen and Diment, 2016). However, those approaches still require knowing *a priori* the number of sources, limiting therefore its applicability to tasks developed for a robot in a real scenario.

The purpose of this work is the unsupervised detection of the direction of arrival of multiple sound sources in reverberant environments, without *a priori* knowledge of the number of sources of interest. To achieve this goal, the method is divided into two steps as it is done in Wang and Brown (2006), May et al. (2013) and Bregman (1990). First, the distribution of direction of arrival is achieved by dividing the signal in different frames and second, a grouping stage is proposed to integrate the information belonging to each single sound source. This approach is inspired on the assumption of the existence of innate bottom-up auditory primitive grouping processes that contribute to form coherent audio streams in an acoustic scene (Wang and Brown, 2006; May et al., 2013; Bregman, 1990). In Bregman (1990) it is stated that these grouping processes are governed by mechanisms analogous to those proposed by the Gestalt psychologists for visual perception. While in the vision field, the Gestalt principles of grouping (also known as law of *prägnanz*) are based on visual properties such as proximity, similarity, continuity, etc. for objects perception, the auditory grouping processes rely on physical properties such as proximity in frequency and time, periodicity or common spatial location (Wang and Brown, 2006; Bregman, 1990). In particular, based on the common spatial location property, concurrent sounds originated from the same location in space tend to be grouped by these innate bottom-up processes in the human auditory system.

In our approach the grouping stage is based on an Expectation–Maximization (EM) algorithm. The EM algorithm has been widely used in many fields such as robot mapping (Thrun et al., 1998; Burgard et al., 1999) visual learning tasks (Wu and Huang, 2002), and location in radiocommunications (Roos et al., 2002). In this work the EM algorithm is employed under the assumption that the sources generate a Laplacian distribution of direction of arrival samples. Other works have also used this assumption to employ the EM technique for binaural multiple sound localization (Cobos et al., 2011; Zhang and Rao, 2010). However in Cobos et al. (2011), the number of sources must be known *a priori* and in Zhang and Rao (2010) the estimated number is based on a criterion that may penalize high order models, while in the method proposed in this paper, high order models are of interest since they can be used to isolate acoustic artifacts that may be detrimental for the localization task. In Escolano et al. (2014) the localization of multiple sources is achieved also based on Laplacian distributions, although via Bayesian inference, where the number of sources is estimated while the model is computed. However, the method presents a limit in terms of the maximum number of sources to be detected and it presents a high computational load.

In this work a kurtosis-driven split-EM is proposed to be used in a robotic platform that will integrate the audio information in two different stages, being the period of observation used to run the algorithm the main difference among these two application levels. In one stage, the audio information related to the lateral localization of sources will be integrated within an observation time of 10 s. This level of processing can be used for a deliberative layer as, for example, within an inner model of the robot and its surroundings. The data about a source of interest will include not only its lateral localization relative to the robot, but also information of the percentage of time it has shown activity respect to the rest of sources within the observation time. These localization results are compared with those extracted from a second stage, which is featured with an observation time of 2 s. This new processing stage is run to endorse a robotic platform with the feasibility of exhibiting a more reactive behavior needed for HRI.

The rest of this paper is organized as follows: Section 2 presents the technique used to compute the ITDs and the directions of arrival for single frames based on the computation of the Generalized Cross Correlation-Phase Transform (GCC-PHAT) algorithm. Then, Section 3 describes with detail the proposed kurtosis driven split-EM method to achieve the localization of multiples sources. Section 4 presents experiments in two different scenarios and a comparison with a state of the art approach. Finally, Section 5 discusses the main conclusions and outlines future work.

2. Detection of the direction of arrival from Short-Time Fourier Transform

Different techniques have been proposed in the literature for the localization of sound sources based on the Short-Time Fourier Transform (STFT) of the signals registered at a pair of microphones (Wang and Brown, 2006; Cobos et al., 2011; Yilmaz and Rickard, 2004; Wang et al., 2016).

In an anechoic and noise-free model, where only the direct path between a source and a microphone is considered, the signals that arrive at two microphones from K acoustic sources can be expressed as:

$$y_1(t) = \sum_{k=1}^{K} a_{1,k}\ s_k(t - T_{1,k}) \tag{1}$$

$$y_2(t) = \sum_{k=1}^{K} a_{2,k}\ s_k(t - T_{2,k}), \tag{2}$$

being $s_k(t)$ the kth source signal, $a_{1,k}$ and $a_{2,k}$ the attenuation coefficients and $T_{1,k}$ and $T_{2,k}$ the time delays associated with the path from the kth source to microphones 1 and 2, respectively.

The microphone signals are synchronously sampled at frequency f_s, frame-decomposed and expressed into the time–frequency (T–F) domain via short-time Fourier transform (STFT). The microphone signal in the T–F domain can be expressed by means of:

$$Y_1(i,n) = \sum_{k=1}^{K} a_{1,k}\ S_k(i,n)\ e^{-j\,2\pi\,f_i\,T_{1,k}} \tag{3}$$

$$Y_2(i,n) = \sum_{k=1}^{K} a_{2,k}\ S_k(i,n)\ e^{-j\,2\pi\,f_i\,T_{2,k}}, \tag{4}$$

where $j = \sqrt{-1}$ and $S_k(i,n)$ is the STFT of the kth source signal at frequency index i in the framework index n. The term f_i is the analog frequency that corresponds to frequency index i.

Given this model, the inter-channel phase difference (IPD) is defined as:

$$\psi(i,n) = \angle\frac{Y_1(i,n)}{Y_2(i,n)}, \tag{5}$$

where $\psi(i,n)$ is the inter-channel phase difference.

Based on an frame-by-frame scheme, it is assumed that only one source is active in each frame. In this work, the TDOA is computed from

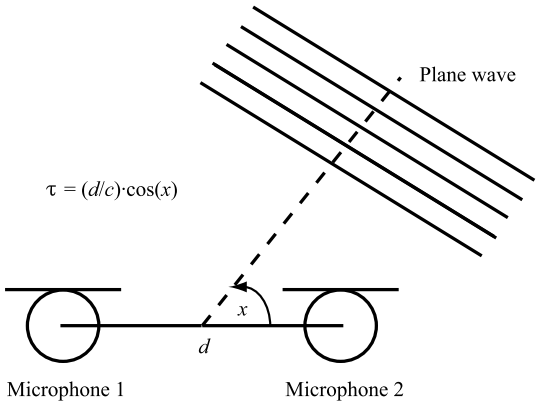

Fig. 1. Geometrical interpretation of the relation between direction of arrival, x, and time difference of arrival, τ.

IPD by means of the well known GCC-PHAT cross correlation algorithm (Knapp and Carter, 1976). The GCC-PHAT is computed as follows Wang et al. (2016):

$$R^{(n)}(\tau) = \left| \sum_i e^{j\psi(i,n)}\, e^{j2\pi f_i \tau} \right| = \left| \sum_i e^{-j2\pi f_i \cdot (\tau_k - \tau)} \right|, \tag{6}$$

From Eq. (6), the TDOA of the active source in the lth frame is estimated as $\tau^{(n)} = \arg\max_\tau R^{(n)}(\tau)$ such that $\tau^{(n)} \in \{\tau_k\}_{k=1}^{K}$. Notice that, by using this approach, the localization task can be performed in a specific bandwidth $[f_{i_{\min}}, f_{i_{\max}}]$. This feature will endorse this algorithm with the feasibility of being part of an attentional mechanism that follows a top-down approach (Basiri et al., 2016). Thus, the bandwidth can be selected as a decision of the high layers.

In general, if the sources are enough far to consider plane wave incidence, it can be shown (see Fig. 1) that the lateral localization of the active source in the nth frame is given by Eq. (7).

$$x^{(n)} = \arccos\left(\frac{c\ \tau^{(n)}}{d}\right), \tag{7}$$

being $x^{(n)}$ the direction of arrival (DOA) of the active source in the nth frame, d the distance between the microphones, and c the speed of sound.

Hence, the localization from the K sources can be estimated via the TDOA computations of the signals. This process has as the outcome a set of observated direction-of-arrival angles. This set can be used as the input of a grouping stage that may integrate the information belonging to each single sound source. So, similarly to the human auditory system, only those properties that are needed for the particular task of lateral localization are extracted, without the necessity of performing the separation and reconstruction of the individual signals (May et al., 2013).

To make more efficient the computation of Eqs. (3), (4) and (6), Fast Fourier Transform algorithm (FFT) is used. The blocks of power-of-two length L_w are padded trailing zeros to compute FFTs of size $2 \cdot L_w$. In this manner, the spectral resolution is improved. This scheme is summarized in Fig. 2.

Blocks with low energy usually correspond to silence periods. Thus, the DOAs that come from frames with energy below a threshold are not considered in the subsequent grouping stage.

3. The Kurtosis driven split-EM algorithm

For each frame with enough energy, a dominant source DOA estimation, $x^{(n)}$, is obtained as explained in Section 2. In Cobos et al. (2011) it is shown that the Laplacian mixture model fits the distribution of the arrival angles, where the mixture model order, K, coincides with the theoretical number of sources. Thus, our goal is to determine the mixture

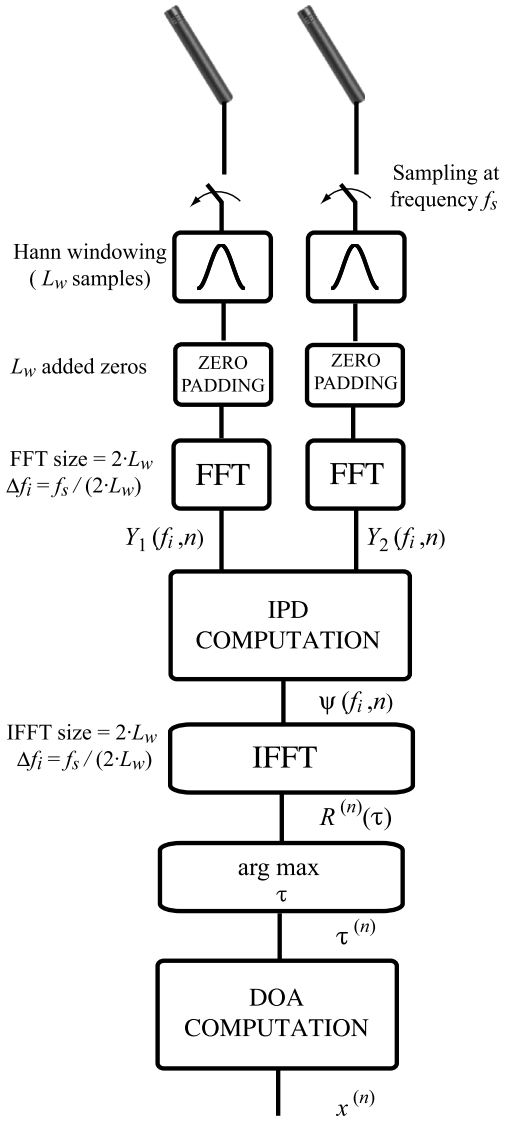

Fig. 2. Computation of the dominant direction of arrival in the nth frame.

parameters, that is to say, the values for $\Theta = \{\alpha_k, \mu_k, \sigma_k\}_{k=1}^{K}$, being α_k the mixing probabilities (*a priori* probability), μ_k the mean and σ_k the standard deviation of the kth Laplacian component.

The estimation of the model parameters $\hat{\Theta}$ is performed according to the *maximum likelihood* (ML) criterion by maximizing the log-likelihood function, $\mathcal{L}(X^{obs}, \Theta)$. Given a set of N DOA observations $X^{obs} = \{x^{(1)}, \ldots, x^{(N)}\} = \{x^{(n)}\}_{n=1}^{N}$, the log-likelihood function corresponding to a K-order Laplacian mixture model is:

$$\mathcal{L}(X^{obs}, \Theta) = \sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} \alpha_k \frac{\exp\left(-\sqrt{2} \cdot \frac{|x^{(n)} - \mu_k|}{\sigma_k}\right)}{\sqrt{2} \cdot \sigma_k} \right) \tag{8}$$

The optimization problem $\hat{\Theta} = \arg\max_\Theta \left(\mathcal{L}(X^{obs}, \Theta)\right)$ has no analytical solution. In Dempster et al. (1977) the Expectation–Maximization (EM) algorithm was used to provide a numerical solution. This algorithm produces a sequence of estimations of the model parameters $\{\hat{\Theta}^{(t)}\}_{t=0,1,2,\ldots}$ by iteratively applying two steps (first Expectation and then Maximization) until convergence of the log-likelihood is achieved or the maximum number of iteration, $T_{\max}$, is reached. Alternating these two steps, the EM algorithm increases monotonically the likelihood of the observations X^{obs}, yielding thus an optimum in the ML sense (Redner

and Mixture densities, 1984). Although the classic EM algorithm reaches the ML solution, this technique has two main drawbacks to consider:

- The number of components, K should be known beforehand.
- The algorithm success relays on the correct initialization. Thus, the number of iterations until convergence may depend on the initial solution, i.e. $\hat{\Theta} = \left\{\alpha_k^{(0)}, \mu_k^{(0)}, \sigma_k^{(0)}\right\}_{k=1}^{K}$. In addition, an inappropriate initialization may lead to a premature convergence to a local optimum.

In the literature, several methods have been proposed to select among a set of candidate models (Rissanen, 1989; Schwarz, 1978) the one that contains the optimal number of components. An approach based on information theory concept is Rissanen's minimum description length (MDL) (Rissanen, 1989). This selection criteria considers the maximized log-likelihood of each model, the number of adjustable parameters and the number of samples. The Schwarz's Bayesian inference criterion (BIC) (Schwarz, 1978) is a similar approach that formally coincides with MDL. The main limitation of the MDL criterion is the assumption that the distribution of interest can be approximated by a mixture of a unique parametric family (for instance, all Gaussians or all Laplacians). Thus, if the data distribution corresponding to a particular component of the mixture drastically deviates from the general family shape, the MDL criterion may fails choosing the optimal model. Another limitation of the MDL criterion is the need of estimating the maximized log-likelihood as it may be difficult when the number of samples is not large enough or EM algorithm prematurely converges. To overcome this limitation, in Lu and Traore (2006) the number of components of a Gaussian mixture has been obtained from a genetic EM algorithm by defining a proper entropy-based fitness function. In Vlassis and Likas (1999) the number of Gaussian components has been dynamically obtained from the overall kurtosis of the mixture. In our approach, the kurtosis is also used to obtain the model order as it is described in Section 3.1.

3.1. The Kurtosis-driven split-EM (KDS-EM) algorithm

In order to overcome the drawbacks of the classic EM above mentioned, we propose a kurtosis-driven split-EM (KDS-EM) algorithm, specifically designed for the localization of acoustic sources in reverberant environments. As in Vlassis and Likas (1999) the splitting process is driven for the value of the kurtosis of all the components that build the mixture. The main novelty of our approach is the use of a stochastic mutation procedure to initialize different order classic-EM executions. Unlike (Vlassis and Likas, 1999) we use real time processed data approximately fitted to Laplacian mixtures instead of a synthetic Gaussian mixture data. Our approach also incorporates a stopping criterion and a mechanism for discarding the non relevant kernels both particularized to our real acoustic scenario.

The initialization procedure is based on splitting the least Laplacian kernel of a previous solution obtained by EM. It is well known that the excess kurtosis of a Gaussian distribution is equal to zero, whereas for a Laplacian distribution is three, therefore, we can state that the further the excess kurtosis of a given Kernel is from three, the less Laplacian the kernel is.

The technique presented here is based on the following main processes: (1) exploring a set of candidate models for a range of values of the mixture order, K, and selecting the solution that better models not only the acoustic sources but also the reflections and noisy effects of the environment; (2) initializing the m order model computation with the modification of the previous solution (reached for the $(m-1)$ order model), by applying a mutational split to the component exhibiting a excess kurtosis value furthest than 3; and (3) discarding those kernels of the finally selected model which are less relevant and can be attributed to undesired effects, such as reflections.

The KDS-EM algorithm is described by the flow diagram shown in Fig. 3. The computation of the different models is performed by running the classic EM. This algorithm finishes when t, the number of iterations, has reached the maximum allowed value ($T_{\max} = 50$) or the following convergence criteria is satisfied:

$$\left|\mathcal{L}\left(X^{obs}, \hat{\Theta}^{(t)}\right) - \mathcal{L}\left(X^{obs}, \hat{\Theta}^{(t-1)}\right)\right| < 0.5 \times 10^{-4}$$

The main procedures of the KDS-EM algorithm are described as follows.

Initialization based on the analysis of the histogram. As it can be seen in the first stage of Fig. 3, the first execution of the classical EM requires an initialization that cannot be based on lower order models. Our initial hypotheses are: (1) each acoustic source has generated at least the 5% of the whole observation samples; (2) all the kernels corresponding to each acoustic source have similar standard deviation; and (3) the means of each kernel are approximately equal to the positions of the histogram local maximums.

Therefore, we have analyzed the histogram to estimate the number of acoustic sources (initial mixture order) and compute the initial values of the parameters of the mixture kernels (mixing probabilities, means and standard deviations). The initialization is made in accordance with the following rules:

- The initial mixing probabilities are proportional to the height of the peaks. The peaks corresponding with kernels with *a priori* probabilities lower than $U_{ini_\alpha} = 0.05$ are discarded and initially considered irrelevant.
- The number of relevant local maximum of the histogram is the initial mixture order, K_0.
- The initial means of each kernel are initialized with the positions of the relevant local maximum.
- The initial standard deviations of each kernel are initialized with the same value, which is computed as the standard deviation of the set of observations divided by the squared root of the initial mixture order, i.e.:

$$\sigma_k = \frac{\sqrt{\frac{\sum_{n=1}^{N}\left(x^{(n)}\right)^2}{N} - \left(\frac{\sum_{n=1}^{N} x^{(n)}}{N}\right)^2}}{\sqrt{K_0}} \tag{9}$$

Mutational splitting driven by Kurtosis analysis. The splitting procedure provides an initialization of the next order model computation via running the EM algorithm (see Fig. 3). The goal of this procedure is minimizing the number of iterations until convergence. The procedure, which has a random component, consists in obtaining a m order model by slightly modifying the $(m-1)$ order model. This m order model is used to re-start the execution of the EM algorithm. As previously mentioned, the splitting operation is driven by the excess kurtosis of the $(m-1)$ order model kernels. Once the computation of the $(m-1)$ order model is completed, the excess kurtosis of the kth component is obtained by Eq. (10).

$$\kappa_k = \left[\frac{1}{N_k}\sum_{n=1}^{N} P\left(k\middle|x^{(n)}\right)\left(\frac{x^{(n)}-\mu_k}{\sigma_k}\right)^4\right] - 3 \tag{10}$$

The kernel with the excess kurtosis furthest from 3 is selected to obtain two new components from it, as it is considered the least Laplacian kernel. The splitting process of the kernel with a set of parameters $\{\alpha_{ini}, \mu_{ini}, \sigma_{ini}\}$ has as output two kernels with parameters $\{\alpha_{s_1}, \mu_{s_1}, \sigma_{s_1}\}$ and $\{\alpha_{s_2}, \mu_{s_2}, \sigma_{s_2}\}$ according to the following rules:

- The mean of one new kernel is initialized to the mean of the selected kernel, i.e. $\mu_{s_1} = \mu_{ini}$.
- The mean of the second kernel μ_{s_2}, is obtained by means of a stochastic process similar to the mutation operators used in

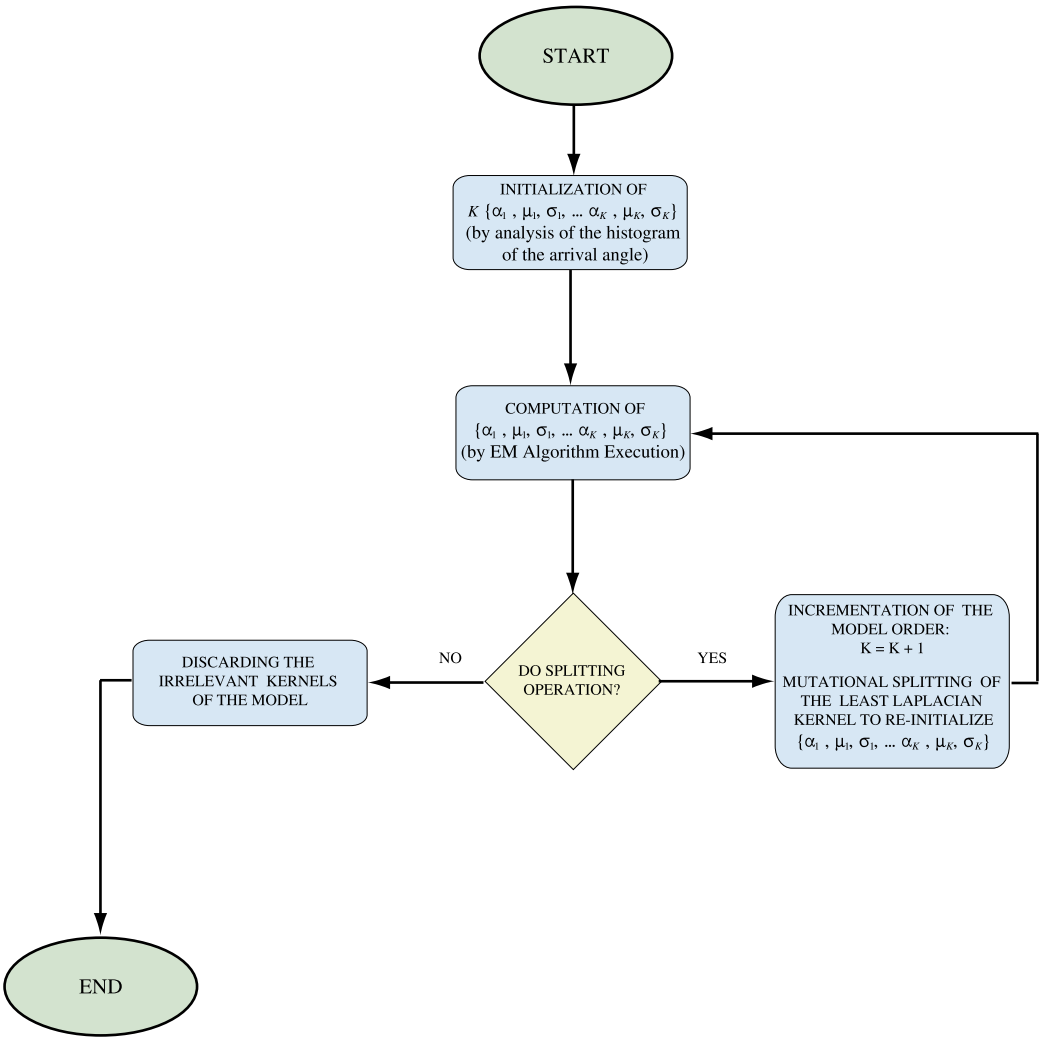

Fig. 3. Flow diagram for the KDS-EM algorithm.

evolutionary algorithms (Beyer and Schwefel, 2002). Thus, the initial value for the mean of the second kernel is obtained as:

$$\mu_{s_2} = \mu_{ini} \pm s \cdot u \cdot \sigma_{ini} \tag{11}$$

where u is a random number from a uniform distribution on the unit interval and s is the mutation strength that in this work has been fixed to 1. The positive sign in Eq. (11) is selected whenever the angle of arrival distribution shows more observations with values higher than μ_{ini}. Otherwise the negative sign is used.

- The mixing probabilities and the standard deviations of the new kernels are randomly chosen but keeping the following conditions: (1) $\alpha_{s_1} + \alpha_{s_2} = \alpha_{ini}$ and (2) $\sigma_{s_1}^2 + \sigma_{s_2}^2 = \sigma_{ini}^2$.

Selection of the best model. In the proposed algorithm the selection of the best model is a consequence of the stopping criterion used in the splitting process. The model selected is the one that maximizes the number of components, with just the limitation of having a minimum distance between two means greater than 10°, i.e.: $\left|\mu_i - \mu_j\right| > 10° \quad \forall i \neq j$. This criterion is proposed since in real applications it is reasonable to assume that two speakers are not usually placed so close to each other (Wang et al., 2016; May et al., 2013). Moreover, this limitation is also biologically plausible since it is known that the smallest detectable change in angular position is about 7° in the human localization task of single sinusoidal sources that get away from the median plane (Stern et al., 2006).

Discarding of the irrelevant kernels. As it can be seen in Fig. 3, the last stage is the responsible of choosing the relevant sources. The irrelevant kernels discarded are considered to be the consequence of scattering, reflections and other undesired noisy effects. Thus, after completing the model selection and the splitting operations, the model is refined by removing those kernels having a mixing probability α_k lower than a threshold (computed as the 25% of the other kernel mixing probability average value). We have also considered irrelevant those kernels with a height lower than the 40% of the other kernels average height, being the height of the kth kernel defined by $h_k = \alpha_k / \left(\sigma_k \sqrt{2}\right)$.

As an example, the kernels estimation performed within a fragment of 2 s (see Section 4) is represented in Fig. 4a. This figure shows that the two sources are estimated together with reflections and other noisy artifacts, by means of six kernels. Then the irrelevant ones are discarded, obtaining an accurate detection as their observations have not been included in the kernels chosen as sources of interest (see Fig. 4b). Table 1 shows in more detail the values of mean, mixing probability and height of the kernels involved in the discarding process. The resulting model shown in Fig. 4b is obtained after applying the discarding process that removes kernels 1, 2, 4 and 6, highlighting therefore as the relevant sources those ones having directions of arrival $\mu_3 = 77.8°$ and $\mu_5 = 133.0°$.

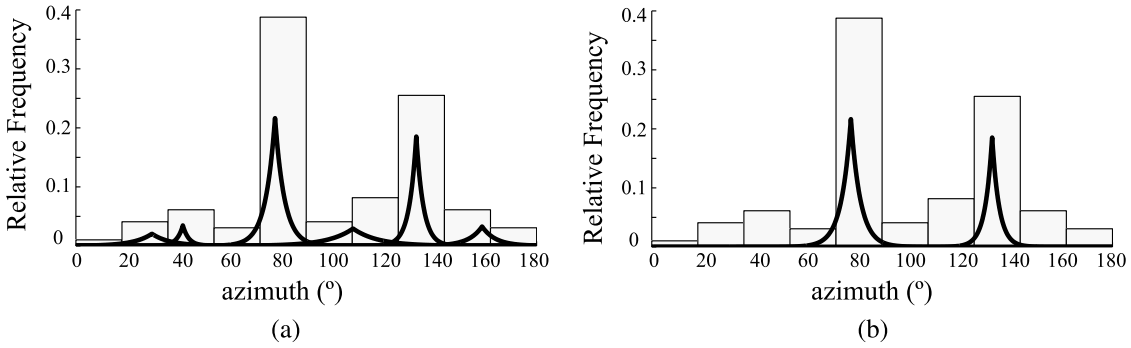

Fig. 4. Relative frequency histogram and probability density function for Laplacian mixture distribution. (a) Model with six kernels for an observation time of 2 s with two real sources in a reverberant environment. (b) Selected kernels after the discarding process.

Table 1
Mean, mixing probability and height of the kernels (μ_k, α_k, h_k) included in the model described as an example in Fig. 4a. The parameters of the relevant kernels, which can be seen in Fig. 4b, are marked in bold.

kth kernel	μ_k	α_k	h_k
1	$\mu_1 = 29.7°$	$\alpha_1 = 0.07$	$h_1 = 0.005$
2	$\mu_2 = 41.8°$	$\alpha_2 = 0.04$	$h_2 = 0.009$
3	$\boldsymbol{\mu_3 = 77.8°}$	$\boldsymbol{\alpha_3 = 0.40}$	$\boldsymbol{h_3 = 0.054}$
4	$\mu_4 = 108.3°$	$\alpha_4 = 0.14$	$h_4 = 0.007$
5	$\boldsymbol{\mu_5 = 133.0°}$	$\boldsymbol{\alpha_5 = 0.26}$	$\boldsymbol{h_5 = 0.046}$
6	$\mu_6 = 158.7°$	$\alpha_6 = 0.09$	$h_6 = 0.008$

4. Experiments and results

The main goal of the experiments described is to evaluate the performance of the proposed algorithm (described in Section 3) estimating the localization of the active sound sources in the surroundings of the robot. In particular, the algorithm is tested considering its usage for two purposes, and therefore as part of two different stages of a robotic platform architecture. First, the feasibility of employing the algorithm as a type of auditory sensor will be evaluated. This is essential to integrate acoustic information perceived for a long period of time (Lu et al., 1992), in this case with the goal of performing the lateral localization of the sources. In a robotics context this information may be useful as an input for an inner model of the robot and its surroundings, being the purpose of this inner model the internalization of the perceived information coming from multiple sensory sources as a support for a cognitive architecture (Calderita et al., 2014). Thus, the mean μ_k, mixing probability α_k and height h_k values of the relevant sources are interesting parameters to characterize the sound sources included in this inner model. For such purpose, an analysis time of 10 s has been established.

Second, the algorithm may also exhibit enough accuracy detecting sources in shorter temporal segments. Through this second evaluation, we can also measure the extent to which the KDS-EM algorithm may serve in a reactive component of a robotic platform. Thus, the estimation of the audio source in short periods of time can be exploited as an input for a reactive mechanism (Viciana-Abad et al., 2014).

Considering this experimental rationale, for each experiment presented, the localization results for a single analysis of 10 s will be shown (named as long-term analysis or LT-A) together with the results for shorter segments of 2 s for the same time frame (named as short-term analysis or ST-A). Shorter observation times could not provide enough number of samples to obtain accurate results from the EM algorithm. In the second analysis the results obtained in some of the 2 s segments are reported.

In addition, a comparison with an alternative Bayesian inference approach is made with audio recordings that contains two, three and four simultaneous sources. In this case, 30 realizations of the experiments are made, with the purpose of analyzing both average and variance values in the localization task. This comparison also considers the computational cost and other limitations.

4.1. Experimental setup

A similar robotic head configuration has been used that in Viciana-Abad et al. (2014). The audio hardware is formed by two AKG C 417 PP omnidirectional microphones that are separated by 13.5 cm, and connected via an M-Audio USB Interface A/D device working at a sample frequency of 44.1 kHz. The frames are obtained by windowing with a Hann window with a 50% overlap. The size of the block L_w has been set to 1024 samples, being equivalent to 23.2 ms, that guarantees the stationarity property of a voice signal. In the frequency domain, the setup achieves a spectral resolution of $\Delta f_i \approx 21.5$ Hz. Also a minimum energy threshold has been established in such a way that just the DOA estimations of frames with energy over the threshold are considered. The threshold has been established as the 75% of the average value considering all the frames within the observation period.

Through the experiments, several speakers have been placed in front of the robotic head at a distance of 1.5 m and with different azimuth angles. The speakers have been requested to be simultaneously talking along the time of the experiments. In addition, a single computer speaker has been used in some cases in order to introduce in the environment the sound of an internet streaming radio at a fixed position.

4.2. Experiment 1

This experiment is carried out with different numbers of active sound sources placed in a large room used as a computing laboratory (from now on referred as Scenario 1). The Scenario 1 setup is sketched in Fig. 5.a. Although there are computers, tables, chairs and tool-boxes, this experiment has been made in a non-highly reverberant environment and with free space between the sound sources and the microphone array. The different configurations (type and number of sources) considered and the results obtained are detailed below.

One speaker. The speaker is placed in S_1 as it is shown in Fig. 5.a. As can be seen in Table 2, in this situation, results of the LT-A (10 s) report the same sources positions than results obtained with the ST-A (2 s). The execution of a KDS-EM algorithm leads to a small number of splitting processes (1 or 2). Thus, the model obtained was built upon just one or two kernels due to the low reverberant conditions of scenario 1, the output of irrelevant kernel discarding process consists of just one kernel placed at the direction-of-arrival of the speaker, approximately 63°.

Two speakers. The speakers are placed in S_1 and S_3 as it is indicated in Fig. 5.a. In this experiment (see Table 2), results of the LT-A are not coincident with the output of the ST-A in some cases, as in some of these short periods one of the speakers was silent in nearly all the frames. The algorithm execution yields to a small number of splitting processes (1 or 2) with an initial model of 2 kernels built upon the histogram analysis. As in the previous test, the final model does not have a number of kernels much higher than the number of active sound sources. Thus, in this case the LT-A exhibits a more accurate estimation of possible speakers positions (or active sources) while the ST-A highlights the more relevant source for a reactive behavior.

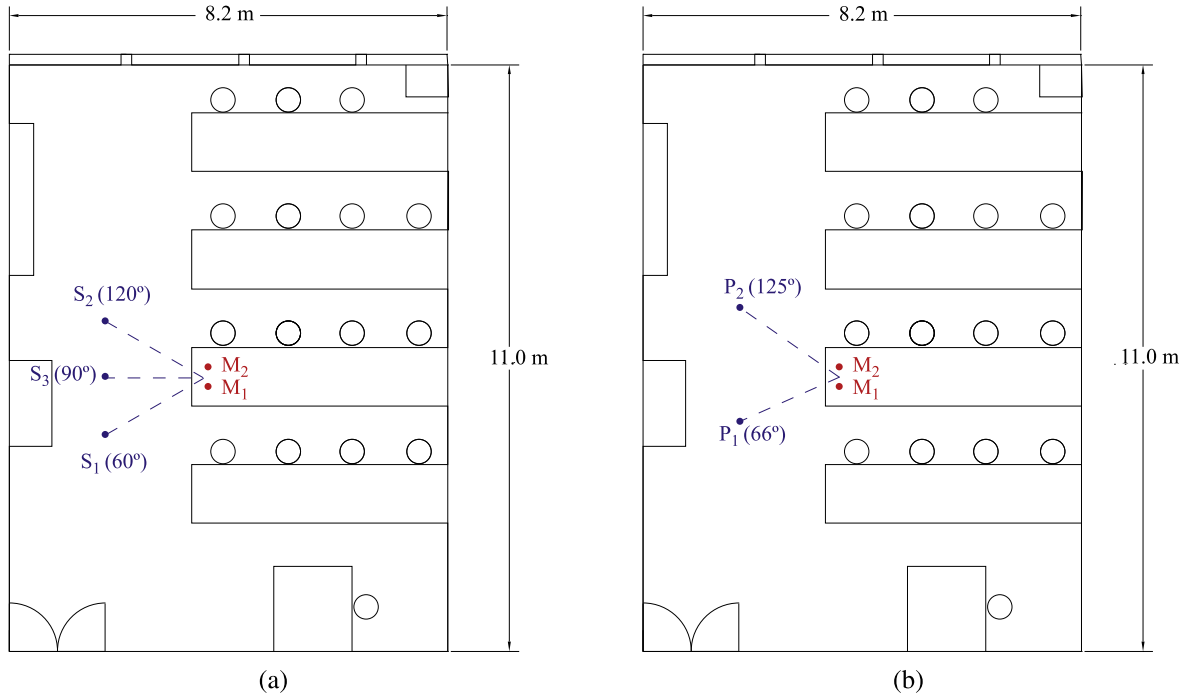

Fig. 5. Floor plan of Scenario 1. The microphones are placed in M_1 and M_2 (distance d between microphones is not represented at scale, $d = 13.5$ cm). (a) In the experiment 1, the sound sources can be placed in some of the following positions S_1, S_2 and S_3, depending of the speakers' number. (b) In the experiment 2, the speaker is placed in the position P_1 and the musical sound source (loudspeaker) is placed in the position P_2.

Table 2
Results obtained from some realizations of Experiment 1.

Number of sources	Analysis time	Number of kernels/splits	Detected sources S_i and kernel parameters (μ_i, α_i, h_i)
One speaker	LT-A:	2/2	$S_1(62.2°, 0.85, 0.49)$
	ST-A:	1/1	$S_1(61.7°, 1, 0.27)$
	ST-A:	1/1	$S_1(63.7°, 1, 0.05)$
Two speakers	LT-A:	3/2	$S_1(59.6°, 0.38, 0.20)$ $S_3(121.9°, 0.48, 0.37)$
	ST-A:	3/2	$S_1(59.2°, 0.38, 0.21)$ $S_3(121.4°, 0.48, 0.55)$
	ST-A:	2/1	$S_1(60.2°, 0.41, 0.09)$
Three speakers	LT-A:	3/1	$S_1(59.2°, 0.31, 0.05)$ $S_2(90.9°, 0.18, 0.07)$ $S_3(122.4°, 0.50, 0.05)$
	ST-A:	3/1	$S_1(58.9°, 0.40, 0.05)$ $S_2(91.0°, 0.14, 0.04)$ $S_3(121.6°, 0.46, 0.05)$
	ST-A:	3/1	$S_1(58.5°, 0.44, 0.56)$ $S_2(90.8°, 0.17, 0.07)$ $S_3(123.4°, 0.39, 0.03)$
	ST-A:	**4/3**	$S_3(122.0°, 0.49, 0.69)$
	ST-A:	3/1	$S_2(91.1°, 0.33, 0.16)$ $S_3(123.1°, 0.54, 0.04)$

Long-term analysis (10 s): LT-A; Short-term analysis (2 s): ST-A; Detected Source: S_i; Mean, mixing probability and height of the kernel corresponding to S_i: (μ_i, α_i, h_i).

Three speakers. The speakers are placed in S_1, S_2 and S_3 as it is indicated in Fig. 5.a. As in the two speakers' case, all the speaker have been correctly detected at their positions 60°, 90° and 120° in the LT-A (see Table 2). However, in the ST-A, there are cases where some of the sources are considered irrelevant, appearing 1, 2 or 3 of them.

In general, in this scenario the ST-A only requires one splitting step. An exception is marked in bold in Table 2, here the final model needs 3 splitting steps because the initial model (obtained from the histogram) has a low number of kernels.

Table 3
Results obtained from some realizations of Experiment 2.

Analysis time	Detected sources P_i kernel parameters (μ_i, α_i, h_i) without LPF	Detected sources P_i and kernel parameters (μ_i, α_i, h_i) with LPF
LT-A	$P_2(124.7°, 0.93, 2.05)$	$P_1(66.4°, 0.44, 0.06)$ $P_2(123.4°, 0.46, 0.09)$
ST-A	$P_2(124.7°, 0.94, 1.96)$	$P_1(59.7°, 0.58, 0.03)$ $P_2(123.1°, 0.42, 0.07)$
ST-A	$P_2(124.7°, 1, 3.04)$	$P_1(65.9°, 0.54, 0.07)$ $P_2(123.5°, 0.42, 0.06)$

Scenario 1 with one speaker (66°) and one musical source (125°) without low pass filter (LPF) and with LPF (cut frequency at 3.4 kHz). Detected Source: P_i; Mean, mixing probability and height of the kernel corresponding to P_i: (μ_i, α_i, h_i).

4.3. Experiment 2

The purpose of this second experiment is to evaluate the performance of the KDS-EM algorithm with musical sources. Thus, with this goal two tests have been made in Scenario 1 (computing laboratory already used for experiment 1) with the sound sources placed as it is shown in Fig. 5.b. Both tests share the same experimental set-up and the only difference is the use of a low-pass filter (LPF) with a cut frequency of 3.4 kHz. The cut frequency is selected to keep only the main spectral components in the human voice signal (Argentieri et al., 2015).

One speaker and one musical source without filtering. Due to differences in the bandwidth and the sparsity property, the musical source is the dominant one in the DOAs estimation in both, LT-A and ST-A. As can be seen in Table 3, for all the ST-A results the voice source is not detected and moreover, the mixing probability of the musical source α_k is nearly 1.0.

One speaker and one musical source with low-pass filtering. In this case, the results are similar to those obtained with two speakers in the same scenario (see Table 2); being therefore the low-pass-filtering necessary to properly discriminate the active speakers from other sound sources.

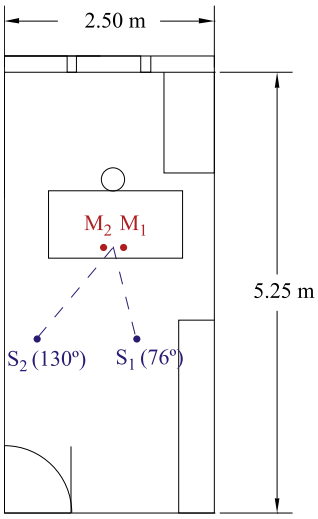

Fig. 6. Floor plan of Scenario 2. The microphones are placed in M_1 and M_2 (distance d between microphones is not represented at scale, $d = 13.5$ cm). The sound sources are placed in S_1 and S_2.

4.4. Experiment 3

The third group of tests has been carried out in a different room (Scenario 2, which is depicted in Fig. 6). In this case, the room is much smaller and with objects of different types placed near the microphones, which causes the apparition of acoustic artifacts in the computation of the direction of arrival values (Perez-Lorenzo et al., 2012) (see Fig. 4). In this experiment, it has been simulated the situation where two people are speaking at the same time, with the goal of analyzing the performance of the KDS-EM algorithm in an scenario featured with more adverse auditory conditions.

One speaker. Table 4 shows the results with just one voice source active in the room. Both the LT-A and ST-A detect the source at a position close to 130°. Compared with the situation of just one source in Scenario 1 (Table 2), in the Scenario 2 the method has used a greater number of kernels for the final model. It is said that the algorithm has over-modeled the active sources. This over-modeling is due to the presence of acoustic artifacts that correspond with kernels considered irrelevant in the discarding process. Thanks to this process, they can be isolated to avoid a lower accurate detection of the source of interest.

Two speakers. As can be seen in Table 4, the KDS-EM algorithm was able to properly detect and locate the two active sources in both cases, ST-A and LT-A. Indeed, the performance is similar to that exhibited detecting two speakers in Scenario 1. Again, the main difference between experiment 1 and 3 is mainly due to the different auditory conditions of the scenarios. Thus, due to the more adverse conditions of this scenario (reverberant, undesired effects, etc.) the algorithm has employed models with a higher number of kernels (3, 4 and even 6) for just two sources. One of the models obtained in the ST-A has been represented in Fig. 4, where four splitting processes have been required to obtain the final model.

4.5. Comparison with a Bayesian inference method

Methods that extend the binaural GCC-PHAT algorithm to simultaneously determine the number of speech sources and their lateral localization in an unsupervised way are not very common in the literature. Here, the proposed method is compared with the one presented in Escolano et al. (2014), where the localization of multiple speech sources is achieved via Bayesian inference to estimate the model that best fits with a localization histogram, also under the assumption of a mixture of Laplacian distributions. Using a nested sampling method to calculate the Bayesian evidence, both the number and position of the sources are inferred, achieving an accurate localization on binaural recordings in a real environment. In Escolano et al. (2014), a robotic head with a configuration similar to the previous subsections was used in a reverberant room with up to four speakers distributed in the localizations depicted in Table 5. The experiments are based on the analysis of binaural recordings with two, three and four simultaneous speakers with ambient noise, and with a time duration between 10 and 12 s. The algorithm was executed 30 times, in an offline mode, to compute the mean and variance in the localization task of each recording, so both robustness and accuracy in the task could be checked. For comparison purposes, the KDS-EM method has been tested in the same way. The results obtained to compare the two methods are shown in Table 6.

As it can be seen, both methods are able to properly detect the sources present in the experiment. In terms of accuracy, the Bayesian inference method is more accurate. Thus, the variance exhibited by the results of this method is at most 0.17 and the highest error committed

Table 4
Results obtained from some realizations of Experiment 3.

Number of sources	Analysis time	Detected sources S_i and kernel parameters (μ_i, α_i, h_i)	Number of kernels/splits
One speaker	LT-A:	$S_2(132.4°, 0.64, 0.17)$	4/1
	ST-A:	$S_2(128.8°, 0.93, 0.04)$	3/2
	ST-A:	$S_2(132.6°, 0.55, 0.13)$	3/2
	ST-A:	$S_2(130.2°, 0.79, 0.07)$	3/1
	ST-A:	$S_2(130.4°, 0.89, 0.06)$	2/1
Two speakers	LT-A:	$S_1(76.5°, 0.52, 0.02)$	4/1
		$S_2(130.5°, 0.36, 0.01)$	
	ST-A	$S_1(76.8°, 0.44, 0.06)$	4/3
		$S_2(138.0°, 0.30, 0.02)$	
	ST-A	$S_1(76.6°, 0.32, 0.15)$	4/3
		$S_2(131.0°, 0.47, 0.03)$	
	ST-A	$S_1(77.8°, 0.40, 0.05)$	6/4
		$S_2(133.0°, 0.26, 0.05)$	
	ST-A	$S_1(76.8°, 0.45, 0.38)$	3/1

Long-term analysis (10 s): LT-A; Short-term analysis (2 s): ST-A; Detected Source: S_i; Mean, mixing probability and height of the kernel corresponding to S_i: (μ_i, α_i, h_i).

Table 5
Sources lateral positions in the experiment performed to compare KDS-EM and a Bayesian inference method (Escolano et al., 2014).

S_1	S_2	S_3	S_4
$\mu_1 = 53°$	$\mu_2 = 76°$	$\mu_3 = 104°$	$\mu_4 = 127°$

Table 6
Comparison results for two, three and four simultaneous sources. The algorithms are executed 30 times for each experiment, and both mean and variance of the sources localization are computed.

Number of sources	Method	Detected sources (mean, variance)
Two speakers	Nested sampling:	$S_1(52.24°, 0.08)$
		$S_2(76.38°, 0.05)$
	KDS-EM:	$S_1(53.84°, 0.55)$
		$S_2(79.00°, 0.25)$
Three speakers	Nested sampling:	$S_1(52.51°, 0.07)$
		$S_2(76.55°, 0.04)$
		$S_3(102.51°, 0.01)$
	KDS-EM:	$S_1(49.43°, 0.35)$
		$S_2(77.07°, 0.10)$
		$S_3(102.77°, 0.04)$
Four speakers	Nested sampling:	$S_1(53.00°, 0.15)$
		$S_2(76.90°, 0.05)$
		$S_3(103.29°, 0.06)$
		$S_4(127.23°, 0.17)$
	KDS-EM:	$S_1(51.57°, 2.04)$
		$S_2(78.19°, 0.65)$
		$S_3(102.56°, 0.16)$
		$S_4(129.60°, 3.42)$

is 1.49°. However, the performance of the KDS-EM method in terms of robustness and accuracy is fairly close. Results show a highest value of variance of 3.42 and a maximum average error in the localization of 3.57°, which can be considered acceptable in robotics application where the purpose is to direct the attention to a particular location.

In terms of the computational cost however the KDS-EM method outperforms the Bayesian inference method. Indeed, it is concluded in Escolano et al. (2014) that although the nestled sampling approach presents a better computational cost compared with other Bayesian sampling methods such as Metropolis–Hastings and importance sampling, it is still not suitable for real-time applications, and some optimizations should be done to reduce the actual computational time for an online approach. Thus, the main advantage of the KDS-EM proposal is that it is suitable for an online situation, since the measured computation time for a single analysis of these recordings is around 2 s in a standard Intel i7 CPU (time depends on the number of kernels of the selected model).

Also, the work in Escolano et al. (2014) presents two additional limitations in the localization task. One is the minimum separation of 10° between sources, that actually is also used in the KDS-EM proposal as the stop criterion. The second limitation corresponds to the requirement of establishing *a priori* the number of maximum sources considered in order to delimit the search space (it has been set to five in the experiments) of the Bayesian inference method.

5. Conclusions and future works

This study proposes a kurtosis driven split-EM algorithm for the unsupervised lateral localization of simultaneous sound sources using a binaural approach. The main novelty of the algorithm is the inclusion of splitting steps of those sources with a distribution far from being ideal Laplacian, which is measured by the kurtosis. In this way, the initial estimation of the number of sources can be corrected and also the undesired effects for localization tasks can be isolated from the real sources and subsequently discarded at a final step. Also, when computing a new model, thanks to the splitting process from a previous model, less EM iterations are needed until convergence compared to a completely random initialization.

The algorithm has been used in a robotic head in two different stages of its architecture. The algorithm has been successfully integrated in a high-level stage implemented with the goal of supporting deliberative behaviors that may depend of the task being accomplished. In this case, the algorithm has been tested within what we have called a long-term analysis in two scenarios and with different number of speakers. Results obtained have outlined this approach as useful to incorporate auditory information within an inner model about the robot and its surroundings, which is usually built to support autonomous behavior. Indeed, results obtained of this long-term analysis suggest that the model extracted about the direction of the active sources could be used as an input to other architecture stages. For example, to complement an attentional mechanism based on more reactive responses or a binaural rendering system, as that proposed to increase the intelligibility of speech in Nikunen and Diment (2016). The short-term analysis made to serve in a more reactive stage of the architecture has highlighted the usefulness of the algorithm ignoring irrelevant or less active sound sources.

Experiments have been carried out in two different environments and results showed that the KDS-EM algorithm adapts by itself to reverberant environments on detriment of increasing the complexity of the model. Also, it has been shown the need of using a vocal band filter for the simultaneous localization of voice sounds and musical sources, which present a higher bandwidth and less temporal sparsity than voice. Finally, the results of a comparison made with a state of the art approach has probed that KDS-EM algorithm is more suitable for robotics purposes where the computational demands must be constrained by the capacity to assist the development of tasks performed in real time, and where there is not *a priori* information about the maximum numbers of sources that can be present.

Future work will include the effective integration of the auditory system into a robotics middleware such as Robocomp (Calderita et al., 2014). This inclusion will allow inferring design strategies to better exploit the capabilities of this algorithm not only at the deliberative level but also as additional support for a reactive behavior.

Acknowledgments

This work has been supported by Economy and Competitiveness Department of the Spanish Government and European Regional Development Fund under the project TIN2015-65686-C5-2-R (MINECO/FEDER, UE). In particular, this work is associated to the development of a person perceptor agent.

References

Argentieri, S., Dansè, P., Souères, P., 2015. A survey on sound source localization in robotics: From binaural to array processing methods. Comput. Speech Lang. 34 (1), 87–112.

Argentieri, S., Portello, A., Bernard, M., Danès, P., Gas, B., 2013. Binaural systems in robotics. In: The Technology of Binaural Listening, Ch. 9. Springer, pp. 225–254.

Basiri, M., Schill, F., Lima, P., Floreano, D., 2016. On-board relative bearing estimation for teams of drones using sound. IEEE Robotics Autom. Lett. 1 (2), 820–827.

Beyer, H.G., Schwefel, H.P., 2002. Evolution strategies – A comprehensive introduction. Nat. Comput. 1 (1), 3–52.

Bregman, A.S., 1990. Auditory Scene Analysis: The Perceptual Organization of Sound. The MIT Press, Cambridge, MA, USA.

Burgard, W., Fox, D., Jans, H., Matenar, C., Thrun, S., 1999. Sonar-based mapping with mobile robots using EM. In: Proceedings of the International Conference Machine Learning, Slovenia, pp. 67–76.

Calderita, L.V., Manso, L.J., Bustos, P., Suarez-Mejias, C., Fernandez, F., Bandera, A., 2014. Therapist: Towards an autonomous socially interactive robot for motor and neurorehabilitation therapies for children. JMIR Rehabil. Assist. Technol. 1 (1).

Cherry, E.C., 1953. Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. 25 (5), 975–979.

Cherry, E.C., 1957. On Human Communication: A Review, A Survey, and A Criticism. MIT Press, Cambridge, MA.

Cobos, M., Lopez, J.J., Martinez, D., 2011. Two-microphone multi-speaker localization based on a laplacian mixture model. Digit. Signal Process. 21 (1), 66–76.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. *B* Stat. Methodol. 39 (1), 1–38.

Dietz, M., Ewert, S.D., Hohmann, V., 2010. Auditory model based direction estimation of concurrent speakers from binaural signals jmir rehabilitation and assistive technologies. Speech Commun. 53 (5), 592–605.

Escolano, J., Xiang, N., Perez-Lorenzo, J.M., Cobos, M., Lopez, J.J., 2014. A Bayesian direction-of-arrival model for an undetermined number of sources using a two-microphone array. J. Acoust. Soc. Am. 135 (2), 742–753.

Ferreira, J.F., Lobo, J., Bessière, P., Castelo-Branco, M., Dias, J., 2013. A Bayesian framework for active artificial perception. IEEE Trans. Cybern. 43 (2), 699–711.

Knapp, C., Carter, G., 1976. The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. 24 (4), 320–327.

Kohlrausch, A., Braasch, J., Kolossa, D., Blauert, J., 2013. An introduction to binaural processing. In: The Technology of Binaural Listening, Ch. 1. Springer, pp. 1–32.

Lu, W., Traore, I., 2006. A genetic EM algorithm for learning the optimal number of components of mixture models. WSEAS Trans. Comput. 5 (9), 1795–1802.

Lu, Z.L., Williamson, S.J., Kaufman, L., 1992. Behavioral lifetime of human auditory sensory memory predicted by physiological measures. Science-New York then Washington 258, 1668–1668.

May, T., Van de Par, S., Kohlrausch, A., 2013. Binaural localization and detection of speakers in complex acoustic scenes. In: The Technology of Binaural Listening, Ch. 15. Springer, pp. 397–425.

Nikunen, J., Diment, A., 2016. Binaural rendering of microphone array captures based on source separation. Speech Commun. 76, 157–169.

Perez-Lorenzo, J.M., Viciana-Abad, R., Reche-Lopez, P., Rivas, F., Escolano, J., 2012. Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments. Appl. Acoust. 73 (8), 698–712.

Redner, R.A., Walker, H.F., 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. 26 (2), 195–239.

Rissanen, J., 1989. In: Stochastic Complexity in Statistical Inquiry, Vol 15, World Scientific.

Roos, T., Myllymaki, P., Tirri, H., 2002. A statistical modeling approach to location estimation. IEEE Trans. Mob. Comput. 99 (1), 59–69.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464.

Stern, R., Brown, G.J., Wang, D., 2006. Binaural sound localization. In: Computational Auditory Scene Analysis, Principles, Algorithms and Applications, Ch.5. John Wiley & Sons, pp. 147–178.

Thrun, S., Burgard, W., Fox, D., 1998. A probabilistic approach to concurrent mapping and localization for mobile robots. Auton. Robots 5 (3–4), 253–271.

Viciana-Abad, R., Marfil, R., Perez-Lorenzo, J.M., Bandera, J.P., Romero-Garces, A., Reche-Lopez, P., 2014. Audio-visual perception system for a humanoid robotic head. Sensors 14 (6), 9522–9545.

Vlassis, N., Likas, A., 1999. A kurtosis-based dynamic approach to Gaussian mixture modeling. IEEE Trans. Syst. Man Cybern.-Part A: Syst. Hum. 29 (4), 393–399.

Wang, D., Brown, G.J., 2006. Fundamentals of computational auditory scene analysis. In: Computational Auditory Scene Analysis, Principles, Algorithms and Applications, Ch.1. John Wiley & Sons, pp. 1–44.

Wang, L., Tsz-Kin, H., Reiss, J.D., Cavallaro, A., 2016. An iterative approach to source counting and localization using two distant microphones. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (6), 1079–1093.

Ying, Wu, Thomas S., Huang, 2002. Towards self-exploring discriminating features for visual learning. Eng. Appl. Artif. Inteligence 15 (2), 139–150.

Yilmaz, O., Rickard, S., 2004. Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process. 52 (7), 1830–1847.

Zhang, W., Rao, D.B., 2010. A two microphone-based approach for source localization of multiple speech sources. IEEE Trans. Acoust. Speech Lang. Process. 18 (8), 1913–1928.