Localizing an Intermittent and Moving Sound Source Using a Mobile Robot

Quan V. Nguyen[1,2,3], Francis Colas[1,2,3], Emmanuel Vincent[1,2,3], and François Charpillet[1,2,3]

***Abstract*— This paper addresses the problem of localizing and tracking one intermittent, moving sound source using a microphone array on a mobile robot. Robot motion provides a solution for estimating the distance to the source and avoiding front-back ambiguity. We propose a mixture Kalman filter (MKF) framework in order to fuse the robot motion information and the measurements taken at different poses of the robot. Experiments and statistical results demonstrate the ability of the proposed method to track one intermittent sound source in a reverberant environment where false measurements of the source angle of arrival (AoA) and the source activity often occur compared to a method that does not consider tracking source activity into account.**

I. Introduction

As assistive robots are becoming more popular, artificial hearing capabilities, *i.e.* robot audition [1], are widely recognized as essential for robot perception. Embedded microphones grant robots the ability to estimate and track the spatial location of sound sources over time [2], [3], [4]. Having a better knowledge about location of the sources can help the robot to localize itself with respect to known sound sources [5], [6], [7] as well as to improve the performance of source separation and speech recognition and, consequently, to efficiently interact with humans and the environment. For these reasons, source localization from multi-microphone recordings plays a central role in robot audition.

Broadly speaking, source localization techniques can be classified into three families [8]. One approach consists in computing the time delay of arrival (TDOA) between every pair of microphones using generalized cross-correlation with phase transform (GCC-PHAT) [9] and to derive the source position by triangulation. This approach is typically outperformed by steered response power (SRP) [8] or multiple signal classification (MUSIC) [10] techniques that compute the pseudo-likelihood of each candidate position on a grid, and pick the maxima on that grid (see [8], [11], [12] for experimental comparisons). Binaural variants of these techniques have been developed for situations when the array is mounted on the robot head [13]. In the typical situation when the distance to the source is larger than the array size, all of these techniques can only estimate the source angle of arrival (AoA) but not its distance.

Most of these approaches have been implemented on robots [13], [14], [15] and used to estimate the source AoA in a similar way as with a static microphone array. Robots can actually provide more information by exploiting motion: this is known as active audition. The absolute position of a source can be tracked by fusing the motion of the robot and the auditory perception. This is done using probabilistic filters such as nonlinear extensions of Kalman filtering [16], [3], particle filters [2], [17], [18], [19], or occupancy grids [20], [21], [4]. Compared to a static microphone array, the head or whole body movements performed in an active audition setting allow the robot to avoid front-back confusion and to estimate the distance to the source. Most of these techniques consider the situation when the sound source to be tracked is continuously active [16], [20], [4].

However, the sound sources in real life, *e.g.* speech, are not always active. Silence exists between successive words, successive sentences, and speaker turns. The problem of tracking one intermittent and moving source by using a nonlinear mixture Kalman filter (MKF) is presented in [3]. In this work, the source activity (i.e., whether the speaker is active or inactive in a given time frame) is supposed to be known a priori. The experimental evaluation in anechoic conditions showed promising results [3]. However, in the real world, reverberation (i.e., sound echoes on the walls of the room) and noise are present that severely degrade the performance of source localization [22], [12] and Source Activity Detection (SAD) [23], [24]. Joint estimation of the activity and false measurement from AoA measurements alone is introduced in [25]. The accuracy of this activity detection which depends on the AoA measurement is presumably poor compared to an independent SAD. We claim that joint estimation of the source location and activity from AoA measurement and SAD is required to achieve robust estimation in such real-world conditions.

The contributions of this paper are twofold. First, we formulate a framework that applies to any microphone array geometry for explicitly tracking both the location and activity of one intermittent moving sound source in a reverberant environment, with a nonlinear MKF. This choice of filter is motivated by the fact that the state vector consists both of continuous and discrete variables, as well as the fact that the observation model is nongaussian. Also, it facilitates comparison with the algorithm in [3]. Second, we provide an experimental validation of our algorithm in several realistic scenarios.

The remainder of the paper is organized as follows. Section II presents the mathematical model of MKF. The evaluation protocol and the results are respectively described in Sections III and IV. Section V concludes the paper.

[1]Inria, Villers-lès-Nancy, F-54600, France
[2]CNRS, Loria, UMR n° 7503, Vandœuvre-lès-Nancy, F-54500, France
[3]Université de Lorraine, Loria, UMR n° 7503, Vandœuvre-lès-Nancy, F-54500, France
firstname.lastname@inria.fr

II. Mixture Kalman Filter

We consider the problem of estimating the absolute spatial position of a single intermittent, moving audio source by a moving robot. We represent the belief about the source position, motion and activity by a mixture of Gaussians. At the beginning, this mixture expresses all possible locations and activities of the source in the room. The goal of the MKF is to update the belief over time given a sequence of measurements. In this section, we first define the state vector and the observation vector, and then detail the proposed MKF algorithm.

A. State vector

By contrast with [3], we take the source activity into account in the state vector. We define the state vector as:

$$\begin{bmatrix} X \\ a \end{bmatrix} = \begin{bmatrix} X_r \\ X_s \\ a \end{bmatrix} = \begin{bmatrix} x_r \\ y_r \\ \theta_r \\ x_s \\ y_s \\ \theta_s \\ v_s \\ w_s \\ a \end{bmatrix} \tag{1}$$

where X_r is the pose of the robot, i.e., its absolute position $[x_r, y_r]$ and its orientation θ_r w.r.t. the x-axis; X_s is the state of the sound source, i.e., its absolute position $[x_s, y_s]$, its orientation θ_s w.r.t. the x-axis, and its linear and angular velocities $[v_s, w_s]$; a is source activity, where $a = 1$ indicates that the source is active, otherwise $a = 0$. We assume that the pose of the robot X_r is known and we need to estimate state of the source X_s and its activity a. With the state of the robot in the state vector, this model have potential to deal with non-zero process noise of the robot motion when we have additional observation for the robot.

B. Observation vector

As mentioned above, audio source localization techniques can estimate the source AoA but not its distance. Therefore, we assume that the observation vector Z_k in a given time frame k consists of one AoA measurement Z_k^{l} (obtained via a localization technique) and one source activity measurement Z_k^{a} (obtained via an SAD technique). The likelihood of the state vector w.r.t. this observation can be expressed as

$$P(Z_k|X_k, a_k) = \begin{cases} P_{\mathrm{sn}}(Z_k^{\mathrm{l}}|X_k)P(Z_k^{\mathrm{a}}|a_k) & \text{for } a_k = 1 \\ P_{\mathrm{n}}(Z_k^{\mathrm{l}})P(Z_k^{\mathrm{a}}|a_k) & \text{for } a_k = 0 \end{cases} \tag{2}$$

with P_{sn} and P_{n} denoting the distribution of the measured AoA when the source is active or inactive, respectively. In the latter case, it is supposed that the recorded signal consists of spatially diffuse noise, so P_{n} does not depend on X_k.

An example of P_{sn} is shown in Fig. 1 for our linear microphone array and the localization technique considered in Section III-A. The probability density concentrates around the true AoA and its symmetric w.r.t the microphone axis: this phenomenon is known as front-back confusion. The probability density for other AoAs is nonzero, but much smaller. Therefore we can approximate the observation model by a mixture of 2 Gaussians:

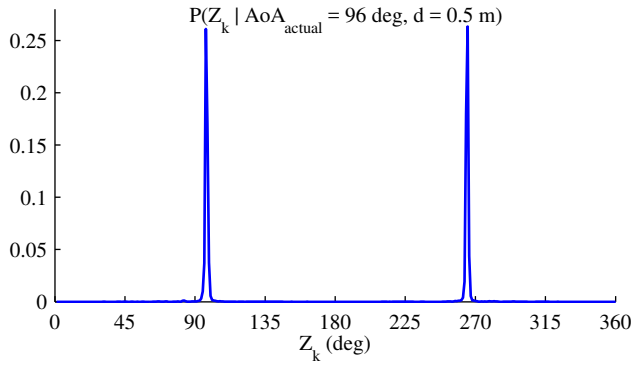

Fig. 1. Distribution of the measured AoA when the actual source is at 96° and 0.5 m from the microphone array.

$$P_{\mathrm{sn}}(Z_k^{\mathrm{l}}|X_k) = \sum_{j=1}^{M_k} \omega^j \mathcal{N}_{\mathcal{Z}}[h^j(X_k), R_k^{i,j}] \tag{3}$$

with $\sum_{j=1}^{M_k} \omega^j = 1$ and $M_k = 2$.

For a nonlinear microphone array, the observation model could be represented by only a single Gaussian.

C. Recursive Bayesian estimation

The transition probability between time steps is given by

$$\begin{aligned} &P(X_k, a_k|X_{k-1}, a_{k-1}) = \\ &\quad P(a_k|X_{k-1}, a_{k-1})P(X_k|X_{k-1}, a_{k-1}, a_k). \end{aligned} \tag{4}$$

The source activity a_k and the state X_k are conditionally independent so we can rewrite the above equation as

$$P(X_k, a_k|X_{k-1}, a_{k-1}) = P(a_k|a_{k-1})P(X_k|X_{k-1}). \tag{5}$$

The state transition probability $P(X_k|X_{k-1})$ is defined by the dynamic model

$$X_k = f(X_{k-1}, u) + d \tag{6}$$

where u are the robot commands and d is the process noise of the robot and the sound source with covariance matrix Q. Note that the control input u for the robot is given during the estimation, so for simplicity it is later omitted from the equations.

The source activity transition probability $P(a_k|a_{k-1})$ is defined by $P_{\mathrm{appear}} = P(a_k = 1|a_{k-1} = 0)$ which is the source appearance probability or $P_{\mathrm{disappear}} = P(a_k = 0|a_{k-1} = 1)$ which is the source disappearance probability.

The posterior probability of the state vector can be recursively computed by alternating these two steps:

- prediction: compute $P(X_k, a_k|Z_{1:k-1})$ given the previous belief $P(X_{k-1}, a_{k-1}|Z_{1:k-1})$ and the state transition model, as shown in (7),

$$\begin{aligned} P(X_k, a_k|Z_{1:k-1}) &= \sum_{a_{k-1}} \int P(X_k, a_k|X_{k-1}, a_{k-1}) P(X_{k-1}, a_{k-1}|Z_{1:k-1}) d(X_{k-1}) \\ &= \sum_{a_{k-1}} \int P(a_k|a_{k-1}) P(X_k|X_{k-1}) P(X_{k-1}, a_{k-1}|Z_{1:k-1}) d(X_{k-1}) \end{aligned} \tag{7}$$

$$P(X_k, a_k|Z_{1:k-1}) = \sum_{i=1}^{N_{k|k-1}} \omega_{k|k-1}^i \mathcal{N}(\hat{X}_{k|k-1}^i, \hat{P}_{k|k-1}^i)\delta(a_k = 0) + \sum_{i=N_{k|k-1}+1}^{N_{k|k-1}(M_{k-1}+1)} \omega_{k|k-1}^i \mathcal{N}(\hat{X}_{k|k-1}^i, \hat{P}_{k|k-1}^i)\delta(a_k = 1) \tag{8}$$

$$\omega_{k|k-1}^i = \begin{cases} \omega_{k-1|k-1}^i P(a_k = 0|a_{k-1} = 0), & \text{if } i \in \left[1, N_{k-1|k-1}\right] \\ \omega_{k-1|k-1}^i P(a_k = 0|a_{k-1} = 1), & \text{if } i \in \left[N_{k-1|k-1}+1, N_{k-1|k-1}(M_{k-1}+1)\right] \\ \omega_{k-1|k-1}^{i-N_{k-1|k-1}(M_{k-1}+1)} P(a_k = 1|a_{k-1} = 0), & \text{if } i \in \left[N_{k-1|k-1}(M_{k-1}+1)+1, N_{k-1|k-1}(M_{k-1}+2)\right] \\ \omega_{k-1|k-1}^{i-N_{k-1|k-1}(M_{k-1}+1)} P(a_k = 1|a_{k-1} = 1), & \text{if } i \in \left[N_{k-1|k-1}(M_{k-1}+2)+1, N_{k-1|k-1}(2M_{k-1}+2)\right] \end{cases} \tag{9}$$

$$\begin{aligned} P(X_k, a_k|Z_{1:k}) =& \eta \sum_{i=1}^{N_{k|k-1}} \omega_{k|k-1}^i P(Z_k^{\mathrm{a}}|a_k) P_{\mathrm{n}}(Z_k^{\mathrm{l}}) \mathcal{N}(\hat{X}_{k|k-1}^i, \hat{P}_{k|k-1}^i)\delta(a_k = 0) + \\ & \eta \sum_{i=N_{k|k-1}+1}^{2N_{k|k-1}} \omega_{k|k-1}^i P(Z_k^{\mathrm{a}}|a_k) P_{\mathrm{sn}}(Z_k^{\mathrm{l}}|X_k) \mathcal{N}(\hat{X}_{k|k-1}^i, \hat{P}_{k|k-1}^i)\delta(a_k = 1) \end{aligned} \tag{10}$$

$$\begin{aligned} P(X_k, a_k|Z_{1:k}) =& \eta \sum_{i=1}^{N_{k|k-1}} \omega_{k|k-1}^i P(Z_k^{\mathrm{a}}|a_k) P_{\mathrm{n}}(Z_k^{\mathrm{l}}) \mathcal{N}(\hat{X}_{k|k-1}^i, \hat{P}_{k|k-1}^i)\delta(a_k = 0) + \\ & \eta \sum_{i=N_{k|k-1}+1}^{2N_{k|k-1}} \sum_{j=1}^{M_k} \omega_{k|k-1}^i \omega^j P(Z_k^{\mathrm{a}}|a_k) \mathcal{N}_{\mathcal{Z}}[h^j(X_k), R_k^{i,j}] \mathcal{N}(\hat{X}_{k|k-1}^i, \hat{P}_{k|k-1}^i)\delta(a_k = 1) \end{aligned} \tag{11}$$

$$\lambda^{i,j} = \frac{1}{\sqrt{\left|2\pi(H P_{k|k-1}^i H^T + R_k^{i,j})\right|}} e^{-\frac{1}{2}\left[Z_k^j - h(\hat{X}_{k|k-1}^i))\right]^T \left[H P_{k|k-1}^i H^T + R_k^{i,j}\right]^{-1} \left[Z_k^j - h(\hat{X}_{k|k-1}^i))\right]} \tag{12}$$

$$\begin{aligned} P(X_k, a_k|Z_{1:k}) =& \eta \sum_{i=1}^{N_{k|k-1}} \omega_{k|k-1}^i P(Z_k^{\mathrm{a}}|a_k) P_{\mathrm{n}}(Z_k^{\mathrm{l}}) \mathcal{N}(\hat{X}_{k|k}^i, \hat{P}_{k|k}^i)\delta(a_k = 0) \\ & + \eta \sum_{i=N_{k|k-1}+1}^{2N_{k|k-1}} \sum_{j=1}^{M_k} \omega_{k|k-1}^i \omega^j P(Z_k^{\mathrm{a}}|a_k) \mathcal{N}_{\mathcal{Z}}[h^j(\hat{X}_k), R_k^{i,j}] \mathcal{N}(\hat{X}_{k|k-1}^i, \hat{P}_{k|k-1}^i)\delta(a_k = 1) \\ =& \sum_{i=1}^{N_{k|k}} \omega_{k|k}^i \mathcal{N}(\hat{X}_{k|k}^i, \hat{P}_{k|k}^i)\delta(a_k = 0) + \sum_{i=N_{k|k}+1}^{N_{k|k}(M_k+1)} \omega_{k|k}^i \mathcal{N}(\hat{X}_{k|k}^i, \hat{P}_{k|k}^i)\delta(a_k = 1) \end{aligned} \tag{13}$$

$$\begin{aligned} P(X_{k-1}, a_{k-1}|Z_{1:k-1}) = & \sum_{i=1}^{N_{k-1|k-1}} \omega_{k-1|k-1}^i \mathcal{N}(\hat{X}_{k-1|k-1}^i, \hat{P}_{k-1|k-1}^i)\delta(a_{k-1} = 0) + \\ & \sum_{i=N_{k-1|k-1}+1}^{N_{k-1|k-1}(M_{k-1}+1)} \omega_{k-1|k-1}^i \mathcal{N}(\hat{X}_{k-1|k-1}^i, \hat{P}_{k-1|k-1}^i)\delta(a_{k-1} = 1) \end{aligned} \tag{14}$$

- update: recompute the belief $P(X_k, a_k|Z_{1:k})$ given the prediction and the new measurement Z_k

$$P(X_k, a_k|Z_{1:k}) = \eta P(Z_k|X_k, a_k) P(X_k, a_k|Z_{1:k-1}) \tag{15}$$

where η is a normalizing constant.

D. Derivation of the MKF

Since the state vector includes both continuous and discrete variables and the observation model is a mixture of Gaussians, we propose a MKF to address these two issues.

1) Prediction step: Assume that at the previous time step $k-1$ the belief about the state (X_{k-1}, a_{k-1}) is given by the

mixture of Gaussians in (14) with weights $\omega^i_{k-1|k-1}$ such that $\sum_i \omega^i_{k-1|k-1} = 1$. Applying the prediction rule (7) to this density yields the predicted density (8). This is a mixture of Gaussians, whose weights $\omega^i_{k|k-1}$ are expressed in (9) and means and variances are given by

$$\hat{X}^i_{k|k-1} = f(\hat{X}^i_{k-1|k-1}, u) \tag{16}$$

$$F^i_{k-1} = \frac{\partial f_{k-1}(X,u)}{\partial X}|_{X=\hat{X}_{k-1|k-1}} \tag{17}$$

$$\hat{P}^i_{k|k-1} = F^i_{k-1} P^i_{k-1|k-1} F^{iT}_{k-1} + Q_{k-1}. \tag{18}$$

2) Update step: By applying the update rule (15) to the predicted density and replacing the observation model by the mixture of 2 Gaussians in (3), we obtain the new belief in (10) and (11). The product of every 2 Gaussians $\mathcal{N}_{\mathcal{Z}}[h^j(X_k), R^{i,j}_k]\mathcal{N}(\hat{X}^i_{k|k-1}, \hat{P}^i_{k|k-1})$ can be computed in closed form as $\lambda^{i,j}\mathcal{N}(\hat{X}^i_{k|k}, \hat{P}^i_{k|k})$ where $\lambda^{i,j}$ is a constant defined in (12). The new belief can therefore be expressed as in (13), where

$$\omega^i_{k|k} = \begin{cases} \omega^i_{k|k-1} P_{\mathrm{n}}(Z^{\mathrm{l}}_k) P(Z^{\mathrm{a}}_k | a_k) \eta & \text{if} \quad a_k = 0 \\ \omega^i_{k|k-1} \omega^j P(Z^{\mathrm{a}}_k | a_k) \lambda^{i,j} \eta & \text{if} \quad a_k = 1 \end{cases} \tag{19}$$

If $a_k = 0$, then $\hat{X}^i_{k|k} = \hat{X}^i_{k|k-1}$. If $a_k = 1$, then

$$\hat{X}^i_{k|k} = \hat{X}^i_{k|k-1} + G^i_k \left[Z^j_k - h(\hat{X}^i_{k|k-1}) \right] \tag{20}$$

$$H^i_k = \frac{\partial h(X)}{\partial X}|_{X=\hat{X}_{k|k-1}} \tag{21}$$

$$\hat{P}^i_{k|k} = \hat{P}^i_{k|k-1} - G^i_k H^i_k \hat{P}^i_{k|k-1} \tag{22}$$

$$S^i_k = H^i_k \hat{P}^i_{k|k-1} H^{iT}_k + R^{i,j}_k \tag{23}$$

$$G^i_k = \hat{P}^i_{k|k-1} H^{iT}_k (S^i_k)^{-1}. \tag{24}$$

3) Pruning: From (11), we can realize that the number of hypotheses in the MKF increases over time, which will consume a lot of memory. To deal with this problem, when the number of hypotheses is larger than $N_{\max}$ we keep only the $N_{\min}$ hypotheses with the highest weights and prune the other hypotheses which have much lower weights.

III. Experimental protocol

We conducted numerical experiments to evaluate our MKF algorithm for tracking one intermittent and moving speech source with room reverberation and noise. Our experimental settings mimic the *smart room* at Inria Nancy, where the robot is a Turtlebot equipped with a Kinect sensor. In this work, we are only using the linear array of 4 microphones included in the Kinect.

A. Data

Due to the statistical nature of false measurements, a large number of experiments is needed to obtain statistically meaningful results. Such a large number of experiments can hardly be conducted with a real robot. Therefore, we resort to simulation of the robot movements, the source movements, and the resulting location and activity measurements. We employ state-of-the-art techniques for the simulation of reverberation and acoustic noise, whose parameters are

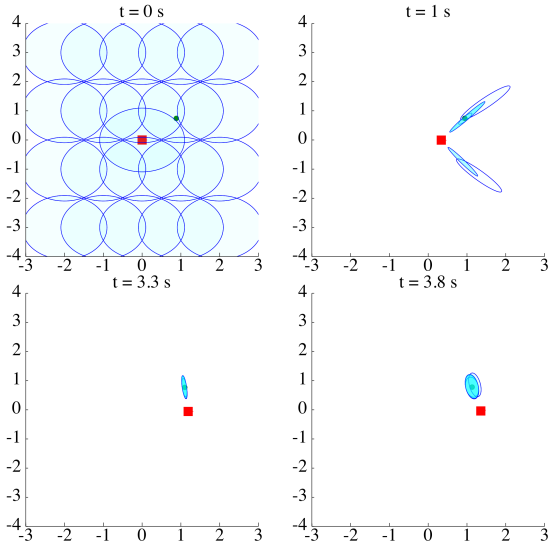

Fig. 2. Visualization of our MKF in the example scenario. Robot positions are shown as red squares, and the actual source position as a green circle. Blue ellipses represent 95% confidence regions of source location estimation of various hypotheses in the mixture with a transparency proportional to the weight of the components.

fixed as in [4] and closely match the real conditions in that room. More specifically, the reverberation time (250 ms), the intensity of speech and noise, and the noise spectrum match those of the real environment. The source AoA is estimated by MUSIC with generalized singular value decomposition (GSVD) [15] as implemented in HARK [26]. The probability distribution of AoAs estimated by MUSIC-GSVD for each of 360 true AoAs (from 0° to 359°) and 5 distances (from 0.5 to 3 m) was constructed and used to simulate the observed source location. We consider an SAD error rate of 5%. We considered four different scenarios, depending whether the sound source is static or mobile ($v_s = 0.07$ m/s, $w_s = 8°$/s) and inactive for many short time intervals (0.5 s) or a long time interval (2 s). For each scenario, we randomly generated 100 source trajectories for a duration of 10 s. The robot trajectory was fixed in all experiments with a maximum speed of 0.38 m/s.

B. Algorithm settings

We set the parameter values of the MKF as follows. The time step is 0.1 s. The covariance matrix Q was set as $Q = \mathrm{diag}(0, 0, 0, 0.00095\,\mathrm{m}^2, 0.00062\,\mathrm{m}^2, 6.2°\,^2, 0, 0)$, the initial covariance of the pose of the robot is $P^R_{0|0} = \mathrm{diag}(0, 0, 0)$, the variance $R^{i,j}$ varied as a function of the source distance between $0.8°\,^2$ at 0.3 m and $4.5°\,^2$ at 3 m, the source appearance/disappearance probabilities were set to $P_{\mathrm{appear}} = 0.5$ and $P_{\mathrm{disappear}} = 0.5$, and the number of hypotheses in the MKF to $N_{\max} = 300$ and $N_{\min} = 24$. By testing our model with different values of the hyper-parameters (*e.g.*, P_{appear}, $P_{\mathrm{disappear}}$ and SAD error rate) in separated experiments, we see that our model is robust to those hyper-parameters.

We evaluate the performance of our MKF method with vs. without tracking of the source activity. Our method without

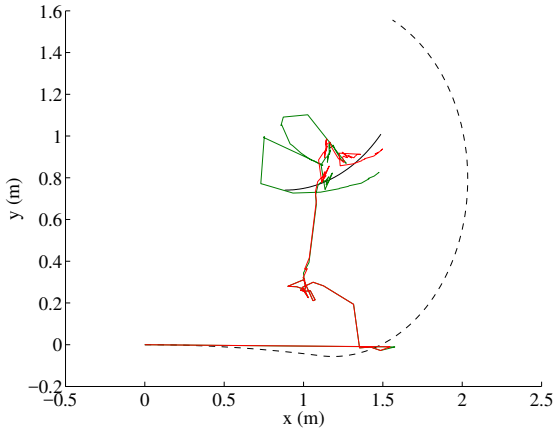

Fig. 3. Estimated trajectories in the example scenario. Dashed black line: trajectory of the robot, Solid black line: ground truth trajectory of the sound source, Green line: estimate of Portello MKF, Red line: estimate of our MKF.

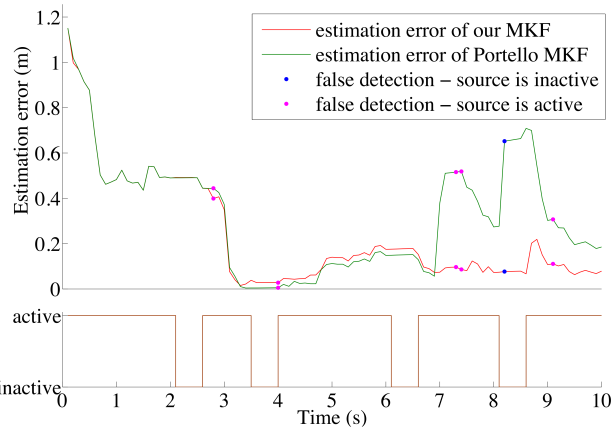

Fig. 4. Top: Estimation error of our MKF vs Portello MKF. Bottom: Ground truth of source activity.

tracking of the source activity is similar to Portello et al.'s [3][1] therefore we call it Portello MKF in the following.

IV. Results

A. Example scenario – Visualization

Fig. 2 shows the first few seconds of tracking one intermittent, moving source. At time $t = 0\,\text{s}$, the mixture is initialized with several components evenly distributed over the room in order to approximate a uniform prior. After 1 s, half of the hypotheses for the source position are distributed along the direction from the source to the robot and the rest are symmetric w.r.t. the microphone axis. This symmetrical uncertainty is due to the front-back confusion phenomenon illustrated in Fig. 1. These symmetrical hypotheses become smaller and disappear after 3 s, thanks to the robot motion. More precisely, the motion of the source for the symmetric components is bigger, less coherent and therefore less probable than of the correct components. For a nonlinear microphone array, the transitory phase with two directions would not exist.

B. Example scenario – Estimated trajectories

Fig. 3 compares the source trajectory with the estimations of our MKF and Portello MKF. As both models are mixture models, the posterior distribution is usually not unimodal. In order to generate a single point estimate, we simply compute the mean of the distribution. After the first few seconds discussed above, both trajectories follow the sound source. We can also observe that, there are some moments when the estimated source location of Portello MKF is far away from actual source location, however, our MKF still can track the source location with lower estimation error.

Fig. 4 shows the estimation error, that is the distance between the estimated source position and the true position. During the first 3 s, both MKFs have high estimation error due to the front-back ambiguity. Between 3.5 and 3.9 s, the source is inactive and the SAD is correct. During this period, the uncertainty about the source position increases because of the lack of measurements. The uncertainty gets smaller when source becomes active again.

At time $t = 2.8\,\text{s}$, a false measurement of the source activity occurs: the source is active but SAD detects it as inactive. The estimation error of Portello MKF becomes larger than ours but only for one time step. Conversely, from $t = 3.2\,\text{s}$ to 7 s, the estimation error of Portello MKF is lower than ours but by 2 cm only.

A false measurement of AoA occurs at time $t = 4.8\,\text{s}$. The AoA difference between observation and ground truth is $9°$, this is not a big value. As a result, both the estimated error of our MKF and Portello MKF have a small rise.

At time $t = 7\,\text{s}$, a false AoA measurement occurs: the ground truth AoA is $81°$, but the measured AoA is $62°$. Although such a false measurement can occur with very low probability, it can have a major impact. Indeed, the estimation error of Portello MKF increases drastically and remains large. By contrast, the estimation error of our MKF does not change much. This is an unexpected benefit of the proposed approach: when a false AoA measurement occurs, the weight of the hypotheses corresponding to an inactive source increases, so that the belief is little affected.

At time $t = 8.2\,\text{s}$, a false measurement of the source activity occurs: the source is inactive but SAD detects it as active. Again, it appears that our MKF can handle this situation but Portello MKF severely suffers from it. This was expected, since the tracking of the source activity implemented by our method was designed precisely to address this issue.

C. Statistical analysis

In order to conduct a statistical comparison of our method and Portello's, we compare the distributions of errors between the estimate and the ground truth. These distributions, shown in Fig. 5, accumulate the error for all time steps and all experimental runs. We can see that the error for our MKF is mostly concentrated below 1 m while Portello MKF shows a bigger error distribution with a heavy tail. This is confirmed

[1]The difference with their method lies in the number of Gaussians used to approximate the observation likelihood.

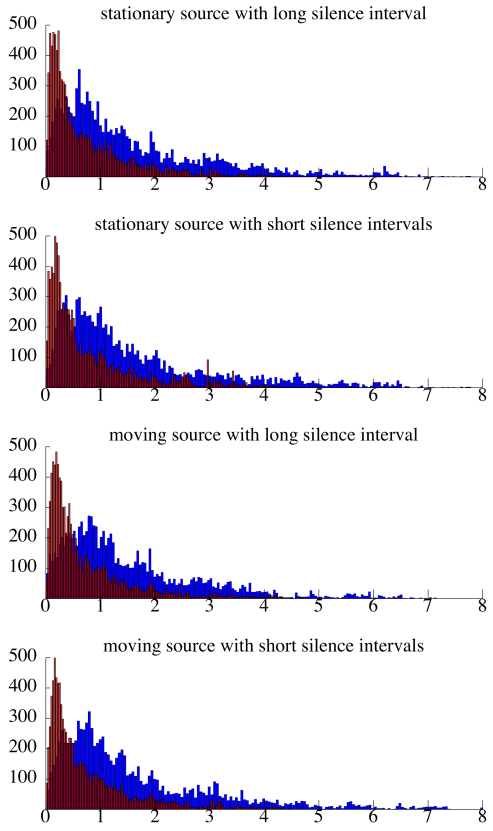

Fig. 5. Estimation error distribution of our MKF method in red and Portello MKF method in blue.

with a Wilcoxon signed-rank test that assesses that our MKF has a significantly smaller error than Portello with $p < 0.01$. Interestingly, our MKF outperforms Portello MKF in all four scenarios, whether the source is static or moving and includes short or long silences.

V. Conclusions

We presented an MKF that applies to any microphone array geometry for tracking one intermittent, moving sound source in a reverberant environment using a mobile robot. The main theoretical contribution of our method is the explicit estimation of the source activity, which allows us to cope with imperfect SAD algorithms.

Experimental results have demonstrated the significantly better performance of our algorithm compared to recent work in the literature [3]. An additional advantage of our algorithm is its ability to track the location of the source in the presence of false AoA measurements.

This work is for now done in a realistic simulation but evaluation on a real robot should highlight the advantage of using a realistic sensor model in the estimation. Another perspective of this work is its extension to multiple sound sources, integrating signal characteristics to help disambiguate the data association issue.

References

[1] H. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *Proc. ICASSP*, vol. 2015-August, 8 2015, pp. 5610–5614.

[2] I. Marković, A. Portello, P. Danès, I. Petrović, and S. Argentieri, "Active speaker localization with circular likelihoods and bootstrap filtering," in *Proc. IROS*, 2013, pp. 2914–2920.

[3] A. Portello, G. Bustamante, P. Danès, J. Piat, and J. Manhès, "Active localization of an intermittent sound source from a moving binaural sensor," in *Forum Acusticum*, 2014.

[4] E. Vincent, A. Sini, and F. Charpillet, "Audio source localization by optimal control of a mobile robot," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 5630–5634.

[5] K. Nakadai, H. G. Okuno, and H. Kitano, "Real-time sound source localization and separation for robot audition." in *Proc. INTERSPEECH*, 2002, pp. 193–196.

[6] J. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, and H. Okuno, "Robust recognition of simultaneous speech by a mobile robot," *Robotics, IEEE Transactions on*, vol. 23, no. 4, pp. 742–752, Aug 2007.

[7] E. Martinson and A. Schultz, "Discovery of sound sources by an autonomous mobile robot," *Auton. Robots*, vol. 27, pp. 221–237, 2009.

[8] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localisation in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[9] C. Knapp and G. Carter, "The generalized cross-correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[10] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[11] A. Badali, J.-M. Valin, F. Michaud, and P. Aarabi, "Evaluating real-time audio localization algorithms for artificial audition in robotics," in *Proc. IROS*, 2009, pp. 2033–2038.

[12] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Process.*, vol. 92, no. 8, pp. 1950–1960, 2012.

[13] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proc. AAAI*, 2000, pp. 832–839.

[14] U.-H. Kim, J. Kim, D. Kim, H. Kim, and B.-J. You, "Speaker localization using the TDOA-based feature matrix for a humanoid robot," in *Proc. RO-MAN*, 2008, pp. 610–615.

[15] K. Nakamura, K. Nakadai, and G. Ince, "Real-time super-resolution sound source localization for robots," in *Proc. IROS*, 2012, pp. 694–699.

[16] A. Portello, P. Danès, and S. Argentieri, "Acoustic models and Kalman filtering strategies for active binaural sound localization," in *Proc. IROS*, 2011, pp. 137–142.

[17] E. A. Lehmann and R. C. Williamson, "Particle filter design using importance sampling for acoustic source localization and tracking in reverberant environment," *EURASIP J. Adv. Signal Process.*, vol. 2006, p. 017021, 2006.

[18] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robot. Auton. Syst.*, vol. 55, no. 3, pp. 216–228, 2007.

[19] M. F. Fallon and S. J. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1409–1415, 2012.

[20] E. Martinson and A. Schultz, "Auditory evidence grids," in *Proc. IROS*, 2006, pp. 1139–1144.

[21] B. P. DeJong, "Auditory occupancy grids with a mobile robot," *J. Autom., Mobile Robot., Intell. Syst.*, vol. 6, no. 3, 2012.

[22] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. ICASSP*, vol. 5, 2001, pp. 3021–3024 vol.5.

[23] J. Ramírez, J. M. Górriz, and J. C. Segura, "Voice activity detection. fundamentals and speech recognition system robustness," in *Robust Speech Recognition and Understanding*, 2007.

[24] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 697–710, 2013.

[25] A. Portello, P. Danes, and S. Argentieri, "Active binaural localization of intermittent moving sources in the presence of false measurements." in *IROS*, 2012, pp. 3294–3299.

[26] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and implementation of robot audition system 'HARK' — open source software for listening to three simultaneous speakers," *Adv. Robot.*, vol. 24, no. 5–6, pp. 739–761, 2010.