Audio source localization by optimal control of a mobile robot

Emmanuel Vincent, Aghilas Sini, François Charpillet

AUDIO SOURCE LOCALIZATION BY OPTIMAL CONTROL OF A MOBILE ROBOT

Emmanuel Vincent[1,2,3]*, Aghilas Sini*[1,2,3] *and François Charpillet*[1,2,3]

[1]Inria, Villers-lès-Nancy, F-54600, France
[2]CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[3]Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
emmanuel.vincent@inria.fr

ABSTRACT

We consider the task of audio source localization using a microphone array on a mobile robot. Active localization algorithms have been proposed in the literature that can estimate the 3D position of a source by fusing the measurements taken for different poses of the robot. The robot movements are typically fixed, however, or they obey heuristic strategies, such as turning the head and moving towards the source, which may be suboptimal. In this paper, we propose to control the robot movements so as to locate the source as quickly as possible. We represent the belief about the source position by a discrete grid and we introduce a dynamic programming algorithm to find the optimal robot motion minimizing the entropy of the grid. We report initial results in a real environment.

***Index Terms*—** Source localization, occupancy grid, active sensing, mobile robot control.

1. INTRODUCTION

Robot audition is an emerging research field at the interface of audio signal processing, artificial intelligence, and control theory [1]. Today, assistive robots typically carry several microphones enabling them to locate and to recognize speech and other sound events. This enables the detection of visually hidden sound sources and efficient interaction with humans.

Source localization techniques fall into three classes [2]. One approach is to compute the time delay of arrival (TDOA) between every two microphones using generalized cross-correlation with phase transform (GCC-PHAT) [3] and to derive the source position by triangulation. It is typically outperformed by steered response power (SRP) [2] or multiple signal classification (MUSIC) [4] techniques that compute the pseudo-likelihood of each candidate position on a grid and pick the maxima on that grid. See [2, 5, 6] for experimental comparisons. Binaural extensions of the above techniques have been designed for situations when the array is mounted on a head [7]. Particle filtering-based tracking algorithms have also been studied for moving sources [8–10].

All three categories of techniques have been implemented on robots [7, 11–13]. They are most often used to estimate the source angle of arrival (AoA), as with a static far-field microphone array. Mobile robots are not restricted to this, however, and they can estimate the full 3D position of the source by taking measurements at different poses. So-called *active source localization* algorithms [7] have been proposed to integrate these measurements into a single location estimate by means of triangulation [14, 15], occupancy grids [11, 16], nonlinear extensions of Kalman filtering [17], or particle filtering [18]. Occupancy grids are one of the most successful frameworks for environment modeling in robotics [19]. The algorithms in [11, 16] estimate the posterior probability of each candidate position on a grid by rescaling the GCC-PHAT pseudo-likelihood. Such an *inverse sensor model* treats each cell of the grid separately and it results in probability values which do not integrate to 1, which leads to inconsistent maps [20].

The above algorithms operate by making the robot follow a fixed patrolling path [21] or heuristic motion strategies inspired from humans [22], such as turning the head towards the source [7,23] and getting closer to it [21,24]. Arm movements resulting in dynamic change of the array aperture have also been studied [25]. These strategies have been shown to improve localization performance experimentally but they may be suboptimal, especially for arrays of three or more microphones which do not fit the geometry of binaural hearing.

In this paper, we provide two contributions. Firstly, we represent the belief about the source position by a discrete grid and we update it over time using a rigorous *forward sensor model* [20]. Secondly, we quantify the information carried by the grid by its entropy and we introduce a dynamic programming algorithm to find the optimal robot motion minimizing the expected future entropy. The structure of the rest of the paper is as follows. The position mapping algorithm and the motion control algorithm are described in Sections 2 and 3, respectively. We report initial results in a real environment in Section 4 and we conclude in Section 5.

2. SOURCE POSITION MAPPING

We represent the environment as a discrete 3D grid whose cells are binary random variables encoding the presence or

absence of an audio source at each position. The goal of mapping is to estimate the posterior probability of these variables given a set of measurements. The number of parameters of the posterior grows exponentially with the size of the grid, hence approximations are required.

A popular approximation is to assume independence and to estimate the occupancy of each cell separately [19]. This approach used in [11, 16] suffers from two drawbacks. First, the resulting maps are inconsistent: when the robot does not move, the probabilities will converge to 1 for all cells along the line from the microphone array to the source, while the source is present in one of those cells only. Secondly, the probabilities do not integrate to 1, which makes it difficult to quantify the amount of information carried by the grid.

2.1. Discrete grid

To address these issues, we adopt a rigorous approach based on a probabilistic forward sensor model [20]. We assume that there is a single active sound source located at absolute position s. We partition the time axis into time frames indexed by t and we denote by $p_t = [x_t, y_t, \theta_t]$ the pose of the microphone array at time t, i.e., its absolute position $[x_t, y_t, z]$ and its orientation θ_t w.r.t. the y-axis. We further denote by m_t the acoustic measurement at time t. The posterior probability of the source position to be estimated is then equal to

$$P(s = i|m_{1:T}, p_{1:T}) \tag{1}$$

where $m_{1:T} = \{m_1, \dots, m_T\}$, $p_{1:T} = \{p_1, \dots, p_T\}$, and $i = [x_i, y_i, z_i]$ is any cell of the grid.

2.2. Forward sensor model

The sensor model $p(m_t|s = i, p_t)$ defines the likelihood that a given localization technique applied to one signal frame provides a measurement m_t given the source position i and the robot pose p_t. In the case of a linear far-field microphone array, the measurements m_t are AoA estimates relative to the array and the sensor model can be expressed as

$$p(m_t|s = i, p_t) = p(m_t|d_{it}, \alpha_{it}) \tag{2}$$

where

$$d_{it} = [(x_i - x_t)^2 + (y_i - y_t)^2 + (z_i - z)^2]^{1/2} \tag{3}$$

$$\alpha_{it} = \arccos \frac{(y_i - y_t)\cos\theta_t - (x_i - x_t)\sin\theta_t}{d_{it}} \tag{4}$$

are the distance and the AoA of cell i relative to the array at time t, respectively. More general parameterizations can be found for nonlinear or nonplanar arrays.

As an example, let us consider the scenario of a Turtlebot[1] equipped with a Kinect. We estimate the source AoA from the Kinect's 4-microphone linear array output using MUSIC with generalized singular value decomposition (GSVD) [13] as implemented in HARK [1]. MUSIC-GSVD is a variant of MUSIC that is robust to spatially correlated noise and that operates in real time. The input covariance matrix $\mathbf{R}_{xx}(t, f)$ and the noise covariance matrix $\mathbf{R}_{nn}(f)$ are classically estimated in each time frame t and each frequency bin f by averaging the short time Fourier transform (STFT) of the input signal and a noise-only signal. The SVD of $\mathbf{R}_{nn}(f)^{-1}\mathbf{R}_{xx}(t, f)$ is then computed and the left singular vectors are used to compute the MUSIC angular spectrum, whose maximum yields the estimated source AoA m_t.

In order to learn the sensor model (2), we simulated speech recording in the presence of spatially isotropic Gaussian noise via the image method [26] using Roomsimove[2]. The room size, the reverberation time (250 ms), the intensity of speech and noise, and the noise spectrum roughly match those of the real environment in Section 4. The microphones were supposed to be omnidirectional since we did not have access to the head-related transfer functions (HRTFs) of the Kinect coupled with the Turtlebot. For each of 360 true AoAs (from 0° to 359°) and 5 distances (from 0.5 to 3 m), we built the histogram of AoAs estimated by MUSIC-GSVD over 50 time frames and 100 random robot poses.

The resulting probability density is illustrated in Fig. 1 for two AoAs. At small distance, it concentrates around the true AoA and its symmetric w.r.t. the microphone axis, a phenomenon known as *front-back confusion* for humans. As distance increases, it becomes smeared and spurious peaks appear at 0° and 180° due to lower signal-to-noise ratio.

Fig. 1 further shows that, as observed in [7], the probability of correct localization is higher for AoAs close to 90°, a region known as the *auditory fovea*. It is yet even higher for AoAs close to 0° or 180°, which do not suffer from front-back confusion. Human-inspired motion strategies based on steering the auditory fovea [7, 23] may hence be suboptimal.

2.3. Updating the grid

The grid (1) is initialized with uniform probability. It is then recursively updated after each new measurement using Bayes law as

$$P(s = i|m_{1:T}, p_{1:T}) = \frac{P(m_T|s = i, p_T)P(s = i|m_{1:T-1}, p_{1:T-1})}{\sum_{i'} P(m_T|s = i', p_T)P(s = i'|m_{1:T-1}, p_{1:T-1})} \tag{5}$$

3. ROBOT MOTION CONTROL

The mapping algorithm in Section 2 makes it possible to estimate the 3D coordinates of the source to a certain extent. We now turn to the question of controlling the robot motion so as to estimate it to the best extent possible in a given time.

[1] http://www.turtlebot.com/

[2] http://www.loria.fr/~evincent/Roomsimove.zip

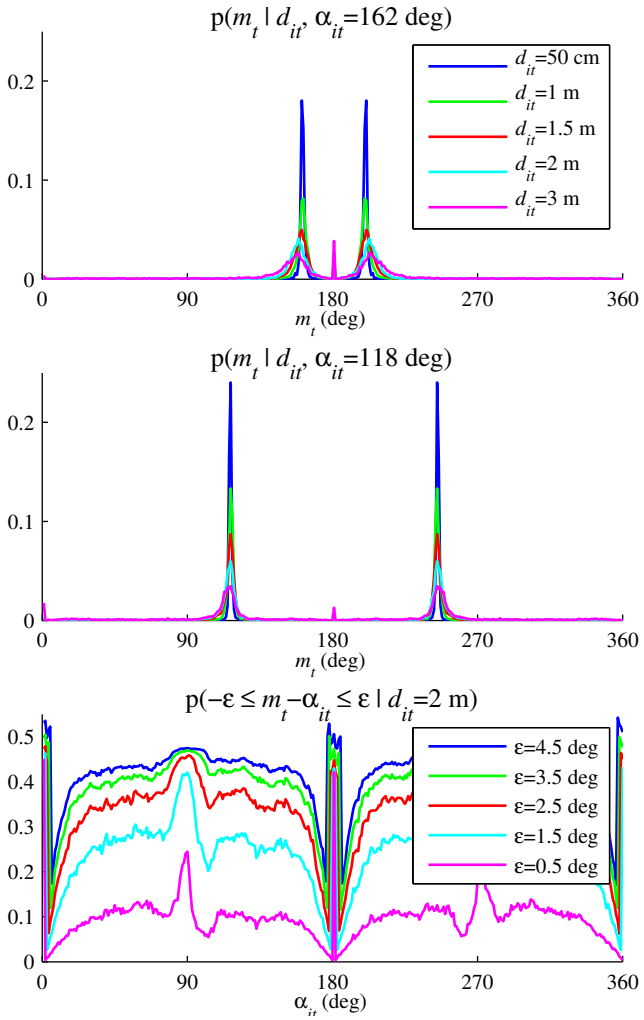

Fig. 1. Top and middle: example sensor model distributions. Bottom: probability of correct localization up to tolerance ϵ.

3.1. Cost function

For a given set of measurements, we quantify the amount of information carried by the grid by its entropy:

$$H(s|m_{1:T}, p_{1:T}) = \\ -\sum_i P(s=i|m_{1:T}, p_{1:T}) \log P(s=i|m_{1:T}, p_{1:T}). \quad (6)$$

The lower the entropy, the greater the amount of information.

Let us assume that the robot has taken measurements up to a certain time T and that we wish to move it to a new pose at time $T+1$. In order to find the optimal pose at time $T+1$, we need to compute the entropy conditionally to the motion sequence $p_{T+1:T+K}$ for all possible motion sequences up to a fixed horizon $T+K$. This entropy cannot be deterministically computed, since future measurements $m_{T+1:T+K}$ are unavailable, but its expectation can be expressed as

$$\mathbb{E}_{m_{T+1:T+K}}[H(s|m_{1:T+K}, p_{1:T+K})] = \sum_{m_{T+1:T+K}} \\ P(m_{T+1:T+K}|m_{1:T}, p_{1:T+K}) H(s|m_{1:T+K}, p_{1:T+K}) \quad (7)$$

with

$$P(m_{T+1:T+K}|m_{1:T}, p_{1:T+K}) = \\ \sum_i P(s=i|m_{1:T}, p_{1:T}) \prod_{k=1}^{K} P(m_{T+k}|s=i, p_{T+k}). \quad (8)$$

The expression in (7) is intractable due to combinatorial explosion of the set of future measurements with increasing K.

In order to address this issue, we compute the expectation of the entropy separately for each future pose p_{T+k} instead:

$$c_{p_{T+k}} = \mathbb{E}_{m_{T+k}}[H(s|m_{1:T}, m_{T+k}, p_{1:T}, p_{T+k}) \quad (9)$$

$$= \sum_{m_{T+k}} P(m_{T+k}|m_{1:T}, p_{1:T}, p_{T+k}) \\ H(s|m_{1:T}, m_{T+k}, p_{1:T}, p_{T+k}) \quad (10)$$

where

$$P(m_{T+k}|m_{1:T}, p_{1:T}, p_{T+k}) = \\ \sum_i P(s=i|m_{1:T}, p_{1:T}) P(m_{T+k}|s=i, p_{T+k}) \quad (11)$$

and $H(s|m_{1:T}, m_{T+k}, p_{1:T}, p_{T+k})$ is obtained via (5) and (6) for each m_{T+k} and $p_{1:T}$. We interpret the quantity (9) as the cost of a future pose and we denote it as $c_{p_{T+k}}$.

3.2. Dynamic programming algorithm

We assume that the cost of moving from one pose p_{T+k} to the next p_{T+k+1} is 0 when this motion is feasible and $+\infty$ otherwise. The optimal motion sequence is given by:

$$\widehat{p}_{T+1:T+K} = \min_{p_{T+1:T+K}} \sum_{k=1}^{K} c_{p_{T+k}}. \quad (12)$$

This is a conventional dynamic programming problem. In the terminology of hidden Markov models, the cost of each pose is the log-observation probability and the cost of each move is the log-transition probability. The optimal sequence is obtained via the Viterbi algorithm (a.k.a. the Bellman algorithm in robotics). Once it has been found, the robot moves to the optimal pose $\widehat{p}_{T+1}$, takes a new measurement m_{T+1}, reestimates the optimal sequence up to $T+K+1$, and so on.

4. EXPERIMENTAL EVALUATION

4.1. Protocol

We evaluated our position mapping and motion control algorithms for the localization of a speech source in the *smart*

home at Inria Nancy. A Turtlebot equipped with a Kinect is placed at a fixed initial pose. A speech signal is emitted by a small loudspeaker at the same height and at a given distance and angle from the robot. Two different distances (1.2 and 2.4 m) and three angles (45°, 90° and 135°) are tested.

The room is discretized into a grid with 5 cm resolution. At each iteration, the robot moves to a new position at 30 cm distance and obtains one AoA estimate using HARK [1]. The robot orientation is constrained by the direction of movement from one position to the next. The actual robot pose after movement is measured with a Sokuiki laser. The source position is eventually estimated as the point with maximum probability in the grid. We implemented the proposed algorithms in C++ and interfaced them with HARK and the robot actuators using ROS.

4.2. Results

Fig. 2 illustrates one test case. After a first measurement, the optimal strategy is not to move towards the estimated source position, which addresses front-back confusion only, but to move in a position slightly aside of it, which makes it possible to estimate the source distance too. The robot eventually passes by the source and moves around it. A similar trajectory was followed in the other test cases.

Table 1 compares the proposed motion control strategy with a random motion strategy, where the next position is chosen randomly among all positions at 30 cm distance. The proposed strategy achieves consistently smaller localization error, down to 0.07 m after 8 to 12 measurements, compared to 0.19 to 0.29 m for the random strategy.

Distance (m)	1.2		2.4	
Motion	random	proposed	random	proposed
2 poses	0.45	0.33	0.91	0.71
4 poses	0.36	0.25	0.77	0.49
6 poses	0.27	0.13	0.62	0.32
8 poses	0.19	0.07	0.48	0.19
10 poses	N/A	N/A	0.36	0.13
12 poses	N/A	N/A	0.29	0.07

Table 1. Average localization error (m) as a function of the initial distance to the source and the number of robot poses.

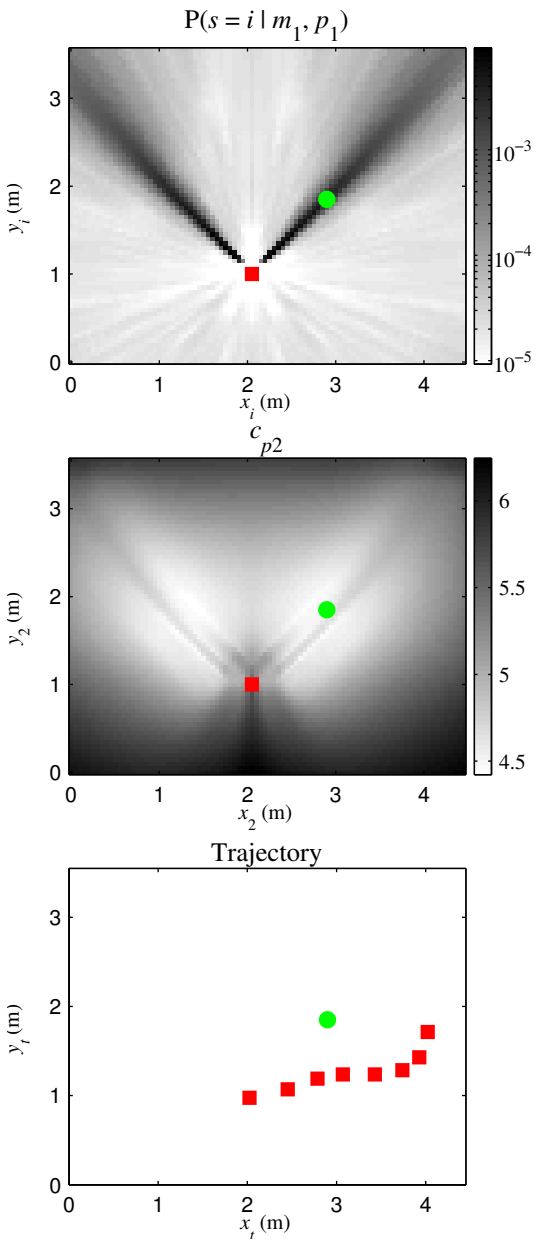

Fig. 2. Top: Source location probability after a first measurement at position p_1 ($\theta_1 = 0$). Middle: Expected entropy at the next position p_2 ($\theta_2 = 0$). Bottom: Complete robot trajectory. Initial and successive robot positions are shown as red squares, and the true source position as a green circle.

5. CONCLUSION

We proposed a motion control strategy for audio source localization by a mobile robot based on discrete grid mapping with a forward sensor model and on the estimation of the grid entropy resulting from each possible movement. We showed that the optimal motion reduces the average localization error up to a factor of 4 compared to random motion. Future work will focus on the design of a simultaneous localization and mapping (SLAM) algorithm applicable when the robot pose is unknown, on the handling of moving or intermittent sources, and on audiovisual integration.

6. ACKNOWLEDGMENTS

This work was supported by the AEN Inria PAL project.

7. REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

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

[8] 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, pp. 017021, 2006.

[9] 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.

[10] 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.

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

[12] 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.

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

[14] Y. Sasaki, S. Kagami, and H. Mizoguchi, "Multiple sound source mapping for a mobile robot by self-motion triangulation," in *Proc. IROS*, 2006, pp. 380–385.

[15] L. Kneip and C. Baumann, "Binaural model for artificial spatial sound localization based on interaural time delays and movements of the interaural axis," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 3108–3119, 2008.

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

[17] 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.

[18] 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.

[19] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46–57, 1989.

[20] S. Thrun, "Learning occupancy grid maps with forward sensor models," *Auton. Robot.*, vol. 15, no. 2, pp. 111–127, 2003.

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

[22] Y.-C. Lu and M. Cooke, "Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners," *Speech Comm.*, vol. 53, no. 5, pp. 622–642, 2011.

[23] E. Berglund and J. Sitte, "Sound source localisation through active audition," in *Proc. IROS*, 2005, pp. 509–514.

[24] K. Song, Q. Liu, and Q. Wang, "Olfaction and hearing based mobile robot navigation for odor/sound source search," *Sensors*, vol. 11, pp. 2129–2154, 2011.

[25] H. Barfuss and W. Kellermann, "An adaptive microphone array topology for target signal extraction with humanoid robots," in *Proc. IWAENC*, 2014.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.