DISTRIBUTED ACOUSTIC SLAM

Łukasz Grzymkowski, Kacper Główczewski, Stanisław A. Raczyński

Gdańsk University of Technology
Faculty of Electronics, Telecommunications and Informatics,
Department of Automatic Control,
ul. G. Narutowicza 11/12
80-233 Gdańsk, Poland
email: stanislaw.raczynski@pg.gda.pl

ABSTRACT

Vision-based methods are very popular for simultaneous localization and environment mapping (SLAM). One can imagine that exploiting the natural acoustic landscape of the robot's environment can prove to be a useful alternative to vision SLAM. Visual SLAM depends on matching local features between images, whereas distributed acoustic SLAM is based on matching acoustic events. Proposed DASLAM is based on distributed microphone arrays, where each microphone is connected to a separate, moving, controllable recording device, which requires compensation for their different clock shifts. We show that this controlled mobility is necessary to deal with underdetermined cases. Estimation is done using particle filtering.

Results show that both tasks can be accomplished with good precision, even for the theoretically underdetermined cases. For example, we were able to achieve mapping error as low as 17.53 cm for sound sources with localization error of 18.61 cm and clock synchronization error of 42 µs for 2 robots and 2 sources.

Index Terms— microphone arrays, distributed microphone arrays, mobile robots, particle filter, robot navigation, source localization

1. INTRODUCTION

Simultaneous localization and mapping (SLAM) is a fundamental problem in robotics, where it solves the problem of navigation in an unknown environment, but is now present in the context of mobile devices as well. Acoustic navigation of robots has been done with non-distributed microphone arrays: either mounted on the walls around the robot to track its position [1] or on the robot itself to track positions of multiple sound sources and determining its own position using triangulation [2]. There is, however, no work done in using distributed microphone arrays for this purpose. On the other hand, vision-based methods are very popular [3]. One can imagine that exploiting the natural acoustic landscape of the robot's environment can prove to be a useful alternative to vision SLAM. Hu *et al.* used a robot-mounted non-distributed microphone array to perform acoustic SLAM in [4]. Just as a typical visual SLAM depends on matching local features between images, *e.g.* SIFT or SURF features [5], distributed acoustic SLAM (DASLAM) would be based on matching acoustic events. There exist many algorithms in the field of audio event detection (AED) [6, 7]. However, in case of acoustic SLAM the event detection and classification accuracy is irrelevant. Instead, it is only important that all devices detect the same events and the AED algorithm needs to be robust against differences between changes in reverberations and signal power, which are different for every microphone in our distributed scenario. To our knowledge there has not been a study like that and in this work we assume that all microphones successfully detect the same events with random time delays and that we detect events in every time step.

Using classical, non-distributed microphone arrays suffers from two problems: it requires costly specialized multichannel (8, 16, 32 and more) A/D converters to ensure sampling synchronization between channels [8]; it also requires that the positions of microphones are known with high precision in order for the system to be able to calculate the TDOA. The first problem can be dealt with by connecting the microphones to much cheaper single-channel A/D converters, however the result is desynchronization of the recorded signals. Lienhart *et al.* proposed to synchronize the recording devices over a network [9]. Another solution has been proposed by Ono *et al.*, who developed a method to jointly estimate the microphone locations, the single source location and the time origins of the recording devices [8, 10]. This method solves both problems at the same time, although it is designed for off-line processing, which limits its applications. Miura *et al.* proposed an on-line algorithm based on the extended Kalman filter [11], which they applied to microphone array calibration, *i.e.*, to estimating the positions of the microphones and a single sound source, as well as the clock shifts. In this paper we extend on this work and introduce microphone mobil-

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

ity and increase the number of sound sources to an arbitrary number. Furthermore, due to high non-linearity of the system, we employ particle filters, which we have found to be much more stable and accurate than the EKF used in [8, 10, 12]. We apply this technique to DASLAM.

2. DASLAM

In DASLAM, we use a distributed mobile microphone array, which is defined as in [12]:

1. There are N sound sources and M microphones, all at arbitrary positions.
2. Positions of the sound sources are fixed, but unknown.
3. Microphones are *distributed* and attached separate recording devices (robots) and their clocks and system times are not synchronized, unknown and the synchronization error is random.
4. Robots are *mobile* and holonomic, i.e., have three degrees of freedom and can freely change their position and orientation.
5. We assume that the speed of the robots is known, since we know their control.
6. The sources emit acoustic events that can be detected and identified by all microphones.

The following variables need to be estimated in the system: positions of the sources (sound source mapping), positions of the robots (robot localization), system clock shifts of individual robots.

For each sound source, its state can be represented by vector

$$\xi_i^{(S)}(t) = \left[x_i^{(S)} \quad y_i^{(S)}\right]^T, \tag{1}$$

where $x_i^{(S)}$ and $y_i^{(S)}$ are the coordinates of the i-th sound source, and for each robot by

$$\xi_j^{(R)}(t) = \left[x_j^{(R)}(t) \quad y_j^{(R)}(t) \quad \varphi_j(t) \quad \tau_j(t)\right]^T, \tag{2}$$

where $x_j^{(R)}$ and $y_j^{(R)}$ are j-th robot's position, φ_j its orientation and τ_j its clock difference with the real time.

2.1. State transition

We assume that the state changes according to

$$\xi_i^{(s)}(t+1) = \xi_i^{(s)}(t), \tag{3}$$

$$\xi_j^{(r)}(t+1) = \xi_j^{(r)}(t) + \begin{bmatrix} v_j(t)\sin(\varphi_j) \\ v_j(t)\cos(\varphi_j) \\ 0 \end{bmatrix} + \mathbf{z}(t), \tag{4}$$

where v_j is the robot's speed, φ_j the direction of movement and $\mathbf{z}(t)$ is the process noise and represents the imprecision of robot movement control. Since the noise comes from many sources (errors for no wheel position/speed feedback, odometry measurement errors, wheel slippage, *etc.*), we assume it is a zero-mean Gaussian noise:

$$\mathbf{z}(t) \sim \mathcal{N}(\mathbf{0}, \Sigma_v^2), \tag{5}$$

where $\Sigma_z = \operatorname{diag}(\sigma_{xy}, \sigma_{xy}, \sigma_\varphi, 0, 0)$.

2.2. Observed variables

The time of an audio event emitted by the i-th source recorded by the j-th robot at time $t_{i,j}$ is equal to

$$t_{i,j} = t_i + \frac{D_{i,j}(t)}{c} + \tau_j, \tag{6}$$

where $D_{i,j}$ is the distance between them and c is the speed of sound. This distance for time t can be calculated as

$$D_{i,j}^2(t) = \left(x_j^{(r)}(t) - x_i^{(s)}\right)^2 + \left(y_j^{(r)}(t) - y_i^{(s)}\right)^2. \tag{7}$$

Since we do not know the actual emission time of each audio event t_i, we follow [11] and use difference of arrival times between each microphone and a reference microphone $\Delta t_{i,j} = t_j - t_1$. There are therefore $N \times (M-1)$ observed variables in the system:

$$\zeta(t) = \begin{bmatrix} \vdots \\ \Delta t_{i,j} \\ \vdots \end{bmatrix} + \mathbf{w}(t) \tag{8}$$

where $i = 1, \dots, N$, and $j = 2, \dots, M$ and

$$\Delta t_{i,j} = \frac{D_{i,j}(t) - D_{i,1}(t)}{c} + \tau_j - \tau_1 \tag{9}$$

where $\mathbf{w}(t)$ is the measurement noise, also assumed to be Gaussian:

$$\mathbf{w}(t) \sim \mathcal{N}(\mathbf{0}, \sigma_w^2 \mathbf{I}), \tag{10}$$

where $\mathbf{I}$ is the identity matrix.

2.3. Determinedness

As pointed out in [8], the system is not always determined, *i.e.*, there are fewer independent observed variables as is the dimensionality of the state vector. The above system is determined only if we have at least 3 microphones and 3 sound sources and the number of observed variables $N(M-1)$ is at least as large as the number of the unknown variables $3M + 2N$, so the following inequality needs to be satisfied:

$$NM \geq 3(N + M). \tag{11}$$

However, this constraint does not hold in case there is correlation between subsequent states. Fig. 1 shows the theoretical error of estimating the state of one of the robots (x_1, y_1)

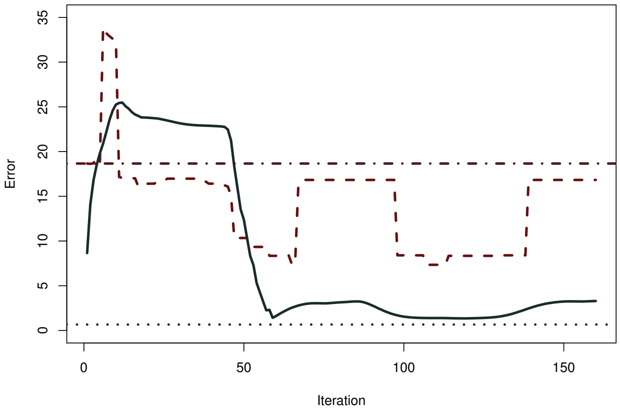

Fig. 1. Theoretical MAP estimation errors for $\mathrm{P}(x_1, y_1|x_2 = x_2, y_2 = y_2, \zeta_1)$ (dotted line), $\mathrm{P}(x_1, y_1|\zeta_1)$ (dashed-dotted line), $\mathrm{P}(x_1, y_2|\zeta_1, \ldots, \zeta_T)$ (dashed line) and $\mathrm{P}(x_1, y_1|\zeta_1, \ldots, \zeta_T, u_1, \ldots, u_T)$ (solid line).

in a 3-source, 2-robot scenario with known sound source positions, *i.e.*, when we observe 3 variables and need to estimate 4. The position is found as the maximum a posteriori estimate. If we know the position of the other robot, the mode of the posterior $\mathrm{P}(x_1, y_1|x_2 = x_2, y_2 = y_2, \zeta_1)$ points to the correct solution even if we have only one measurement, because the system is overdetermined. However, if we do not know it and need to integrate it out, the pdf $\mathrm{P}(x_1, y_1|\zeta_1)$ becomes blurred and multimodal, with the maximum at a wrong place. Multiple measurements, *i.e.* $\mathrm{P}(x_1, y_2|\zeta_1, \ldots, \zeta_T)$, result in the estimate jumping between local minima and never reaching the true solution. However, the known dependence between subsequent robot positions, *i.e.*, knowledge of the model Λ and the control of the system u_t (control signal of the robot, which we have access to), *i.e.* $\mathrm{P}(x_1, y_1|\zeta_1, \ldots, \zeta_T, u_1, \ldots, u_T)$, removes the bias of the estimate.

3. INFERENCE

Due to the non-linear nature of the observation, the inference in our system can be performed using the particle filter. The original particle filter algorithm, called sequential importance resampling (SIR) [13], is used. It consists of the following steps:

1. Uniformly draw P particles $\hat{\xi}_k(0)$ and assign each of them equal weights: $w_k(t) = 1/P$, $k \in \{1, \ldots, P\}$.
2. Estimate the next state $\hat{\xi}_k(t)$ of each particle based on the state model from (3) and 4.
3. Add random noise to each particle.
4. Update the weight $w_k(t)$ of each of the particles based on the Gaussian measurement distribution from (8): $w_k(t) = w_k(t-1)p(\zeta(t)|\hat{\xi}_k(t))$.
5. If the effective number of particles is less than the desired threshold, $P_{\text{eff}}(t) = \left(\sum_{k=1}^{P} w_k^2(t)\right)^{-1} < P_{\text{thr}}$, perform resampling. This is done by randomly selecting new particles with probabilities $w_k(t)$ and using them to replace the old particles, and then setting the new weights to $w_k(t) = 1/P$, $k \in \{1, \ldots, P\}$.
6. Go to point 2.

The point estimate of the system's state can be found as the weighted (using particle weights) mean of the particles.

4. EXPERIMENTAL RESULTS

4.1. Set-up

The proposed approach was tested through simulations. The simulation included a group of robots that moved with a known, constant velocity and stationary sound sources. The robots were moving independently and each robot could change its orientation at random time intervals by a random degree. The orientation, as the robots were holonomic, was known, but affected by process noise. The movements of all robots and positions of sound sources were constrained to a room 80 m by 80 m with point $(0, 0)$ in its center. When robots reached the wall, they were re-oriented towards the center of the room. The initial positions of both the robots and sound sources were unknown and randomly selected from zero-mean normal distribution. Each robot was equipped with a single microphone. The speed of sound was assumed to be 343 m/s. The tests were run multiple times for $(2, 4, 8, 12)$ sound sources and $(2, 5, 8, 10, 12, 14)$ robots. Process noise for the robot state transition was $\mathcal{N}(\mathbf{0}, \sigma_{xy}^2)$ for position and $\mathcal{N}(\mathbf{0}, \sigma_\varphi^2)$ for orientation, where σ_{xy} was $10\ cm$ and σ_φ was $1\deg$. The clock synchronization error is due to, apart from errors in synchronizing system clocks of different robots, varying hardware and ADC sampling rates, and was assumed to be constant, but different among the robots and unknown. The measurement noise σ_ζ was set to $1.7\ cm$. The effective number of particles threshold N_{thr} was initially large in each experiment to improve the convergence speed and then decreased to assure that resampling was done less frequently. All tests were run with 2500 particles and stopped after 10000 iterations.

The measurements of sounds emitted from the sound sources do not carry bearing information as the robots are equipped with a single microphone. What is measured is the distance difference between the reference robot and other robots to each of the sound sources. This means that we estimate the positions in a coordinate system anchored in the reference robot. The relation between the nodes of the system in the estimated coordinate system, i.e. the distances and orientations between robots and sound sources, are estimated correctly, but the entire estimated coordinate system is

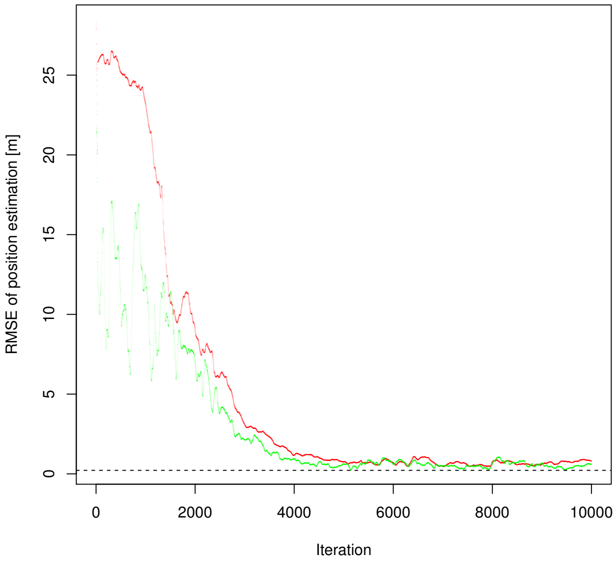

Fig. 2. RMSE of position estimation of 5 robots (green) and 8 sound sources (red) its asymptote is plotted at 22 cm.

translated. In order to remove this translation bias from the results, we apply an affine transformation to the estimated values. The parameters of the affine transform are found by means of least squares estimate. The estimates obtained with a reverse transform are used to calculate the errors.

4.2. Results

The system was found to be convergent, even for the underdetermined cases (*cf.* Fig. 3), although in some cases the estimates would fall into local minima, and then high N_{thr} and levels of particle noise would slow down the descent from a locally to the globally optimal solution. The results are presented in Table 1 and visualized in Fig. 3. These were computed as root mean squared errors of position estimation of robots and sound sources and as an arithmetic mean for clock synchronization estimation errors, with outliers removed from averaging due to slow convergence. Figure 2 shows an example of algorithm's convergence. Reduced particle noise provides more stability and lower estimation error once the global solution is reached.

The results indicate that with more nodes present in the system, the estimation errors are greater. The reason for this is the fact that with more robots and sound sources there are more possible solutions and the time of convergence is higher. On the other hand, when the complexity of the system was lower, the global solution was found very quickly and performance of the estimation was below 20 cm for both robots and sound sources. This suggests that the number of particles and the number of iterations should be adjusted according to the number of estimated state variables.

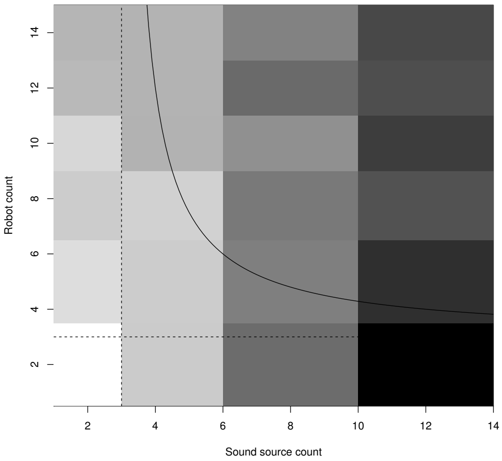

Fig. 3. Results from Table 1, whereas darker colors represent higher error. The solid curve marks the determinedness threshold of the system with its asymptotes marked with dashed lines, *i.e.* situations below this curve from Eq. 11 have less independent observed variables than the dimensionality of the state vector.

5. CONCLUSION AND FUTURE WORK

In this article we have proposed a new idea of distributed acoustic SLAM (DASLAM), which is analogous to the very common visual SLAM. We use distributed mobile microphone arrays, and test our idea on the tasks of robot navigation and localization of multiple sound sources. Results show that both tasks can be accomplished with good precision, even for the theoretically underdetermined cases.

In future work we plan to test existing acoustic event detection methods for robustness against reverb and sound level differences between microphones, and combine them with our DASLAM approach, using real audio for navigation and mapping. We also plan to explore the possibilities of applying this technique to mobile devices, combined with their built-in accelerometer and gyroscope data.

REFERENCES

[1] Q. H. Wang, T. Ivanov, and P. Aarabi, "Acoustic robot navigation using distributed microphone arrays," *Information Fusion*, vol. 5, no. 2, pp. 131–140, 2004.

[2] Y. Sasaki, S. Kagami, and H. Mizoguchi, "Multiple sound source mapping for a mobile robot by self-motion triangulation," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2006, pp. 380–385.

RC	SC	SP [cm]	RP [cm]	C [µs]
2	2	17.53	18.61	42.26
5	2	35.33	40.60	122.52
8	2	44.16	61.73	115.26
10	2	38.17	66.67	130.37
12	2	53.95	90.90	125.86
14	2	55.68	105.18	119.53
2	4	44.83	45.14	63.92
5	4	44.17	47.16	118.11
8	4	41.33	54.82	138.61
10	4	57.28	70.06	126.82
12	4	56.94	82.01	155.23
14	4	57.01	93.62	151.64
2	8	93.81	74.71	118.89
5	8	83.52	59.60	148.68
8	8	86.71	68.70	109.10
10	8	75.05	76.18	118.76
12	8	94.54	98.02	155.76
14	8	82.32	91.70	120.70
2	12	149.51	95.66	65.60
5	12	125.15	86.58	95.17
8	12	107.00	82.93	140.82
10	12	117.69	89.97	118.46
12	12	108.96	93.19	141.92
14	12	112.23	102.96	178.45

Table 1. Estimation errors for varying number of robots (RC) and sound sources (SC). Abbreviations: SP – sound source position errors, RP – robot position errors, C – clock sync. errors.

[3] C. Kerl, J. Sturm, and D. Cremers, "Dense visual SLAM for RGB-D cameras," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov 2013, pp. 2100–2106.

[4] J.-S. Hu, C.-Y. Chan, C.-K. Wang, M.-T. Lee, and C.-Y. Kuo, "Simultaneous localization of a mobile robot and multiple sound sources using a microphone array," *Advanced Robotics*, vol. 25, no. 1-2, pp. 135–152, 2011.

[5] N. Engelhard, F. Endres, J. Hess, J. Sturm, and W. Burgard, "Real-time 3D visual SLAM with a hand-held RGB-D camera," in *Proc. of the RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum, Vasteras, Sweden*, 2011.

[6] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro, "Non-speech audio event detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2009, pp. 1973–1976.

[7] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2005, pp. 1306–1309.

[8] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2009, pp. 161–164.

[9] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, "On the importance of exact synchronization for distributed audio signal processing," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASS)*, 2003, vol. 4, pp. 840–843.

[10] K. Hasegawa, N. Ono, S. Miyabe, and S. Sagayama, "Blind estimation of locations and time offsets for distributed recording devices," in *Proc. Latent Variable Analysis and Signal Separation (LCA/ICA)*, 2010, pp. 57–64.

[11] H. Miura, T. Yoshida, K. Nakamura, and K. Nakadai, "SLAM-based online calibration of asynchronous microphone array for robot audition," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2011, pp. 524–529.

[12] S. Raczyński, Ł. Grzymkowski, and K. Główczewski, "Distributed mobile microphone arrays for robot navigation and acoustic source localization," in *Proc. Control, Automation, Robotics & Vision, International Conference on (ICARCV)*, 2014, pp. 1039–1044.

[13] N. Gordon, D. Salmond, and A. Smith, "Novel approach to nonlinear/non-gaussian bayesian state estimation," in *Proc. IEE Proceedings F (Radar and Signal Processing)*. IET, 1993, vol. 140, pp. 107–113.