© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Split Conditional Independent Mapping for Sound Source Localisation with Inverse-Depth Parametrisation

Daobilige Su, Teresa Vidal-Calleja and Jaime Valls Miro

Abstract— **In this paper, we propose a framework to map stationary sound sources while simultaneously localise a moving robot. Conventional methods for localisation and sound source mapping rely on a microphone array and either, a proprioceptive sensor only (such as wheel odometry) or an additional exteroceptive sensor (such as cameras or lasers) to get accurately the robot locations. Since odometry drifts over time and sound observations are bearing-only, sparse and extremely noisy, the former can only deal with relatively short trajectories before the whole map drifts. In comparison, the latter can get more accurate trajectory estimation over long distances and a better estimation of the sound source map as a result. However, in most of the work in the literature, trajectory estimation and sound source mapping are treated as uncorrelated, which means an update on the robot trajectory does not propagate properly to the sound source map. In this paper, we proposed an efficient method to correlate robot trajectory with sound source mapping by exploiting the conditional independence property between two maps estimated by two different Simultaneous Localisation and Mapping (SLAM) algorithms running in parallel. In our approach, the first map has the flexibility that can be built with any SLAM algorithm (filtering or optimisation) to estimate robot poses with an exteroceptive sensor. The second map is built by using a filtering-based SLAM algorithm locating all stationary sound sources parametrised with Inverse Depth Parametrisation (IDP). Robot locations used during IDP initialisation are the common features shared between the two SLAM maps, which allow to propagate information accordingly. Comprehensive simulations and experimental results show the effectiveness of the proposed method.**

I. INTRODUCTION

Robot audition is an emerging research field at the interface of audio signal processing, artificial intelligence and robotics [1]. Recently, mapping of stationary sound sources have gained increasing interest since the ability to localising sound sources has many potential applications in scenarios such as robotic urban search and rescue (USAR) [2]. In these scenarios the position of sound sources can be used to locate missing people in a disastrous sites. Other examples application include human robot interaction (HRI), where location of sound sources can be used to detect and track speakers [3] or discern between multiple people speech [4].

Research literature in simultaneous localisation and mapping (SLAM) provide a sound framework for robot self-localisation and environmental map building. There are many successful implementations based on laser scanners [5] and vision sensors [6]. These sensors can provide range and bearing or bearing-only information of landmarks in the environment with relatively high accuracy.

Despite many important breakthroughs in the field of robot audition during the last decades, precisely simultaneously robot localisation and sound source mapping remains a challenge mainly due to the following reasons. Firstly, in most robot audition systems, robots are equipped with an embedded microphone array, which is used to obtain the direction of arrival (DOA) of a sound source. Therefore, bearing-only information of sound sources from the current robot pose is observed at each time step. Compared to range and bearing information, bearing-only is 1 DOF shorter both in 2D and 3D. Secondly, although robot audition systems are able to estimate directions of multiple sound sources, in a more general scenario, the number of dominant sound sources that can be reliably detected by robots is very limited. In most cases, the number of detected sound sources cannot be compared to number of key image points detected by a vision sensor, making the attempt to solve the SLAM problem purely based in sound sources quite difficult, especially in the 3D case that demands more landmarks to uniquely determine the robot pose. Thirdly, compared to monocular SLAM [6], which also relies on bearing-only landmarks, the bearing information from a sound source is not always available due to the sparseness of audio signal. In other words, the sound source cannot be detected during periods when it does not generate sound. Lastly, in an indoor environment due to the reverberation, the noise of sound source bearing observations can reach up to 10 degrees, while the noise of a calibrated camera is only one or two pixels.

Due to above mentioned reasons, performing SLAM with only sound sources becomes quite difficult or sometimes impossible when the number of sound sources is low, the robot trajectory is large or 3D estimation is required. In most of the examples in the literature for localisation and sound source mapping using only sound source bearing information some considerations need to be imposed. For instance in [7] and [8], the robot moves relatively short distances so the drift in odometry remains small. Also in [7], multiple sound sources are mapped at the same time in order to obtain enough number of observations to constrain the robot pose. In a more general scenario, however, this can not be always guaranteed (e.g. when the robot is moving a along a silent corridor). When the number of landmarks is not enough, estimation of the robot trajectory gets worse and so does the sound sources locations.

In order to overcome the above mentioned drawbacks,

All authors are associated to Centre for Autonomous System (CAS), University of Technology, Sydney (UTS), Australia. daobilige.su@student.uts.edu.au, {teresa.vidalcalleja, jaime.vallsmiro}@uts.edu.au

more recent works tend to include an additional exteroceptive sensor to assist the sound source mapping. With the help of an additional sensor such as a laser range finder, estimation of the robot pose can become accurate and sound sources locations as well. Examples of such help have been shown by Kallakuri *et al.* [9] and Vincent *et al.* [10]. In [10], a mobile robot with laser scanner and microphone array is used to map sound sources producing an occupancy grid sound map. Each occupancy cell is associated to a probability value for being a sound source and expected entropy is used to obtain the optimum robot path for better observation of the sound source. In their work, though both laser scanner and wheel odometry are used, robot pose's uncertainty is not considered and sound source mapping relies on the "known" robot pose that comes after fusing wheel odometry and laser scanner observations. In [9], a Rao-Blackwellized SLAM system is used to localise the robot using laser scan and wheel odometry data. Based on the particle filter, the robot pose's uncertainty is taken into account to estimate sound probability on an occupancy sound map using a ray tracing algorithm. The method has been extend to the 3D case in their later work [11] by replacing 2D occupancy maps with 3D octree map. Although robot pose's uncertainty is considered, after a loop closure the sound map will not be updated accordingly as there is no correlation between robot poses and the sound map once ray tracing took place.

An SLAM algorithm (optimisation or filtering), which contains robot poses and/or environmental landmarks, and sound sources locations (using an appropriate parametrisation) will be the ideal framework to tackle the above issues. However, we argue that bearing-only, sparse and extremely noisy observations, such as sound ones, will be of little help to improve robot trajectory and/or environmental landmarks. This case is acute in filtered-based SLAM methods when large linearisation errors can cause major failures in the estimation process. Thus in this paper we present an algorithm that still utilises robot pose's uncertainty and allows to update a sound source map after closing a loop in a sound manner. However, it decouples the sound source locations from the rest of the state-vector.

The key idea of the proposed approach is to split the full SLAM map into two independent maps given some common part of the state-vector, *i.e.* Conditional Independent (CI) maps. The first map (the localisation map) contains the robot poses and/or the landmarks observed by a relatively accurate exteroceptive sensor. The second map (the sound source map) contains the robot locations from which the sound sources are observed together with the sound sources encoded as Inverse-Depth Parametrisation (IDP) [6]. The only consideration is that the first map needs to contain in the state-vector the robot locations at the instant when the sound source locations were first observed. By exploiting the conditional independence property, the sound source map can be updated efficiently right after the first map gets updated, producing more accurate sound source mapping results after long periods with loop closures.

The contributions of the paper are two-fold; the novel use of IDP to map sound sources and an efficient algorithm that exploits the CI property to propagate information from a map use for localisation to a sound sources map.

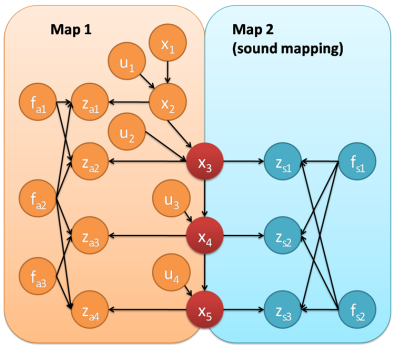


Fig. 1. Bayesian network that describes probabilistic dependency between two CI maps.

The rest of the paper is organised as follows. In section II, the details of the proposed method is illustrated. In section III, various simulation and experimental results are presented to show effectiveness of the proposed method. Section IV presents the conclusion and discussion about further work.

II. The Proposed Method

In this section we present the details to generate two conditionally independent maps split from a full SLAM map, maintain and update by two different SLAM algorithms for simultaneous trajectory estimation and sound source mapping.

A. Structure of the Split CI maps

Let us first examine the Bayesian network in Fig. 1, in which a robot observes different modality landmarks with two sensors, an exteroceptive sensor and a microphone array, during its navigation process. We will use this example, without loss of generality, to illustrate the development of the approach. As shown in Fig. 1, the robot starts from pose x_1, then it moves to x_2 after control input u_1. At x_2, it gets an observation z_{a1} from an additional exteroceptive sensor. z_{a1} is the observation of the landmarks f_{a1} and f_{a2}. Next, the robot moves to x_3 after control input u_2. From x_3, it gets the observations z_{a2} from landmarks f_{a1} and f_{a2} using the additional sensor and z_{s1} from the sound sources f_{s1} and f_{s2} respectively. Then it moves to x_4 after control input u_3 and observes f_{a2} and f_{a3} through z_{a3} and f_{s1} and f_{s2} through z_{s2}. Similarly it moves to x_5 and obtains corresponding observations. From this network, it can be seen that landmarks f_{a1}, f_{a2} and f_{a3} observed using the additional exteroceptive sensor are conditionally independent of the sound sources f_{s1} and f_{s2}. Thus in this example the map generated with the exteroceptive sensor is independent of the map generated with the microphone array given the robot poses x_3, x_4 and x_5. Then, the full map can be optimally split into two CI map as shown in Fig. 1.

Note that the situation in Fig. 1 is a special case of the structure of conditionally independent submaps method

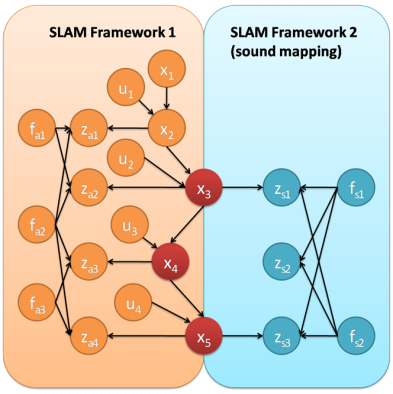


Fig. 2. Modified Bayesian network that describes probabilistic dependency between SLAM variables in two maps.

presented in [12]. It can be seen as a situation where the robot frequently revisit two maps continuously. Robot locations which have observations from both, the additional sensor and the microphone array, are the common elements of the state-vector in both maps. As pointed out in [12], in a frequently revisiting scenario, keeping all robot poses which are common in two submaps in the state vectors of both SLAM maps increases the length of both state-vectors, which leads to a significant increase of the computational complexity. In [12] is suggested to approximate the solution by disregarding the odometry information of the re-visited poses (in our example, x_4 and x_5 would be marginalised out). However, we opted instead to approximate the solution by duplicating the part of the state that contains robot poses that have not been used to initialise sound sources with IDP (see Fig. 2). Although at first glance seems different, in the proposed framework results in an equivalent approximation. The main reason will become apparent when the framework to build and maintain the sound source map is explained. In short, as this latter map is built using a filtered-based framework, all these poses are marginalised eventually leading to a similar simplification to the one proposed in [12].

The most interesting part of splitting the full SLAM map into two CI map is that they can be maintained independently as long as the back propagation algorithm proposed in [12] is applied to propagate information between the maps after an update (in any of the maps) takes place. Note, this algorithm does not contain any approximation and it will produce the same solution as the full SLAM map. In our particular case, we deliberatively avoid propagating the information once the sound map has been update. However, we applied the back propagation algorithm after each update of the localisation map.

B. The localisation map

The aim of this map is to obtain an accurate estimation of the trajectory of the robot and/or landmark map at all times. Any given standard SLAM algorithm (filtering or optimisation, landmark or pose based) to estimate robot poses with a relatively accurate exteroceptive sensor can be used to built and maintain the localisation map. The only requirement is that it has to be amendable to incorporate as part of the state-vector multiple robot poses from where the sound sources are initialised. There are many SLAM implementations available for the common exteroceptive sensors that meet our requirements. For example, Pose SLAM [13] can be used for laser scanner based SLAM, RGB-D SLAM [14] can be used for RGB-D sensors and ORB-SLAM [15] can be used for monocular or stereo camera. In last two cases, poses from key frames can be used for sound landmarks initialisation and parametrisation so that after each optimisation step, poses of key frames are updated and so do sound landmarks.

C. The sound source map

The objective of this map is to accurately localise stationary sound sources utilising the current robot pose estimate (mean and uncertainty). We propose to use an Extended Kalman Filter (EKF)-based SLAM approach and parametrise sound sources locations using IDP. The main advantages of using IDP for bearing only observations are that it models correctly the uncertainty from faraway landmarks and it is less prone to linearisation errors [6]. Under IDP parametrization, the state of each sound source in 2D is,

$$\mathbf{x}^s_{lm}(i) = (x_i y_i \theta_i \rho_i)^T \tag{1}$$

and in 3D case is

$$\mathbf{x}^s_{lm}(i) = (x_i y_i z_i \theta_i \phi_i \rho_i)^T \tag{2}$$

where x_i, y_i and z_i are the Euclidean coordinates of the robot position, which is used for initialising the i-th sound source. θ_i and ϕ_i are the azimuth and elevation angle of sound source respectively. ρ_i is the inverse of distance from the initial robot position to the sound source. Then the full state-vector of the system is

$$\mathbf{x}^s = (\mathbf{x}_r, \mathbf{x}^s_{lm}(1), \mathbf{x}^s_{lm}(2), \ldots, \mathbf{x}^s_{lm}(n))^T \tag{3}$$

where $\mathbf{x}_r$ represent the state of robot pose, being

$$\mathbf{x}_r = (x_r, y_r, \theta_r)^T \tag{4}$$

in the 2D case and

$$\mathbf{x}_r = (x_r, y_r, z_r, qw_r, qx_r, qy_r, qz_r)^T \tag{5}$$

in the 3D case. Variables x_r, y_r and z_r are the Euclidean coordinates, θ_r is the robot yaw angle in 2D, and in 3D we chose quaternions $(qw_r, qx_r, qy_r, qz_r)^T$ to represent the orientation of the robot.

At each iteration of the EKF SLAM, the current robot pose $\mathbf{x}_r$ is either copied with cross-correlations from the localisation map to the sound source map. In the EKF correction step, the sound sources are either initialised if they are observed for the first time or updated with standard EKF update as follows,

$$K^s_t = \Sigma^s_{t-1} H^{sT}_t (H^s_t \Sigma^s_{t-1} H^{sT}_t + Q^s_t)^{-1} \tag{6}$$

$$\mathbf{x}^s_t = \mathbf{x}^s_t + K^s_{t-1}(z^s_t - h^s(\mathbf{x}^s_{t-1})) \tag{7}$$

$$\Sigma^s_t = (I - K^s_t H^s_t)\Sigma^s_{t-1} \tag{8}$$

where Σ^s_{t-1} and Σ^s_t are previous and current estimate of covariance matrix, H^s_t is Jacobian of observation function $h^s(\cdot)$, Q^s_t is the observation noise variance of sound bearing observation and z^s_t is the observed sound source bearing. A detailed discussion of bearing only landmark initialisation under IDP can be found in [16].

Note that with IDP parametrisation of sound sources locations in Eq.1 or Eq.2, only the robot position (x_i and y_i in 2D and x_i, y_i and z_i in 3D) during IDP initialisation is common in both maps and the rest of the state-vector (θ_i, ρ_i in 2D and θ_i, ϕ_i, ρ_i in 3D) is conditionally independent of the localisation map.

D. Back propagation

As mentioned above every time any of the two maps gets updated, a back propagation is needed to update the other map, but we propose to do it only unidirectional. Before describing equations of back propagation, let us first summarise the structure of the state vectors and covariance matrix of the localisation and sound source maps.

The localisation map in terms of its state vector and covariance can be written as

$$p(\mathbf{x}^a|\mathbf{u}_{1:n}, \mathbf{z}_{a1:an}) = \mathcal{N}(\mathbf{x}^a, P^a) \tag{9}$$

where $\mathbf{x}^a$ is the full state vector, $\mathbf{u}_{1:n}$ are control inputs and $\mathbf{z}_{a1:an}$ are landmark observations. The full state vector $\mathbf{x}^a$ is

$$\mathbf{x}^a = (\mathbf{x}_r, \mathbf{x}^s_r(1), \ldots, \mathbf{x}^s_r(n_s), \mathbf{x}^a_{lm}(1), \ldots, \mathbf{x}^a_{lm}(n))^T \tag{10}$$

where $\mathbf{x}_r$ is the current robot pose, $\mathbf{x}^s_r(1), \ldots, \mathbf{x}^s_r(n_s)$ are past robot poses used to initialise sound source IDPs and $\mathbf{x}^a_{lm}(1), \ldots, \mathbf{x}^a_{lm}(n)$ are landmarks observed by the additional sensor. We can rearrange the state vector by grouping elements that are shared by the two maps and those which are not. First, we split $\mathbf{x}^s_r(i)$ as

$$\mathbf{x}^s_r(i) = (\mathbf{x}^{s\text{-}p}_r(i), \mathbf{x}^{s\text{-}o}_r(i))^T, \tag{11}$$

where $\mathbf{x}^{s\text{-}p}_r(i)$ and $\mathbf{x}^{s\text{-}o}_r(i))^T$ represent position and orientation of the robot pose that is used to initialize ith sound source. Then, the full state vector can also be written as

$$\begin{aligned}\mathbf{x}^a =&(\mathbf{x}_r, \mathbf{x}^{s\text{-}p}_r(1), \mathbf{x}^{s\text{-}o}_r(1), \ldots,\\ &\mathbf{x}^{s\text{-}p}_r(n_s), \mathbf{x}^{s\text{-}o}_r(n_s), \mathbf{x}^a_{lm}(1), \ldots, \mathbf{x}^a_{lm}(n))^T.\end{aligned} \tag{12}$$

grouping the localisation map as

$$\begin{aligned}\check{\mathbf{x}}^a =&(\mathbf{x}_r, \mathbf{x}^{s\text{-}p}_r(1), \ldots, \mathbf{x}^{s\text{-}p}_r(n_s),\\ &\mathbf{x}^{s\text{-}o}_r(1), \ldots, \mathbf{x}^{s\text{-}o}_r(n_s), \mathbf{x}^a_{lm}(1), \ldots, \mathbf{x}^a_{lm}(n))^T.\end{aligned} \tag{13}$$

Let $\mathbf{x}_{C_a} = (\mathbf{x}_r, \mathbf{x}^{s\text{-}p}_r(1), \ldots, \mathbf{x}^{s\text{-}p}_r(n_s))^T$ represents elements that are shared by both maps and $\mathbf{x}_A = (\mathbf{x}^{s\text{-}o}_r(1), \ldots, \mathbf{x}^{s\text{-}o}_r(n_s), \mathbf{x}^a_{lm}(1), \ldots, \mathbf{x}^a_{lm}(n))^T$ represents elements that are conditionally independent from the sound source map, then the rearranged full state vector can be written as

$$\check{\mathbf{x}}^a = (\mathbf{x}_{C_a}, \mathbf{x}_A)^T. \tag{14}$$

Similarly, we can rearrange and group covariance matrix of the localisation map as

$$\check{P}^a = \begin{bmatrix} P_{C_a} & P_{CA} \\ P_{AC} & P_A \end{bmatrix}, \tag{15}$$

where P_{C_a}, P_A, P_{CA} andP_{AC} are covariance matrix related to $\mathbf{x}_{C_a}$ and $\mathbf{x}_A$ and their cross correlation terms.

We can apply a similar rearrangement to the state vector and covariance matrix of the sound source map,

$$\check{\mathbf{x}}^s = (\mathbf{x}_{C_s}, \mathbf{x}_S)^T, \tag{16}$$

$$\check{P}^s = \begin{bmatrix} P_{C_s} & P_{CS} \\ P_{SC} & P_S \end{bmatrix}, \tag{17}$$

where $\mathbf{x}_{C_s} = (\mathbf{x}_r, \mathbf{x}^{s\text{-}p}_{lm}(1), \ldots, \mathbf{x}^{s\text{-}p}_{lm}(n))^T$, in which $\mathbf{x}^{s\text{-}p}_{lm}(i)$ represents position of ith robot pose that can be used for sound source initialisation ($(x_i y_i)^T$ of Eq.1 in 2D case and $(x_i y_i z_i)^T$ of Eq.2 in 3D case). $\mathbf{x}_{C_s}$ corresponds to $\mathbf{x}_{C_a}$ in Eq.14 and they are shared part of state vectors of two maps. $\mathbf{x}_S = (\mathbf{x}^{s\text{-}o}_{lm}(1), \ldots, \mathbf{x}^{s\text{-}o}_{lm}(n))^T$, where $\mathbf{x}^{s\text{-}o}_{lm}(i)$ represents bearing and inverse distance of ith sound source ($(\theta_i \rho_i)^T$ of Eq.1 in 2D case and $(\theta_i \phi_i \rho_i)^T$ of Eq.2 in 3D case), is other elements of the state vector in the second map which is conditionally independent from the first map. P_{C_s}, P_S, P_CS and P_SC in Eq.17 are covariance matrix of $\mathbf{x}_{C_s}$ and $\mathbf{x}_S$ and their cross correlation terms.

Once state vectors and covariance matrix of the localisation and sound source maps are rearranged, back propagation can be performed following the algorithm in [12]. Notice that the only information used to back-propagate is the difference in the robot locations at the IDPs initialisation. Each time the localisation map gets internally updated, the state vector and covariance matrix in Eq.16 and Eq.17 of the sound source map are updated as

$$\mathbf{x}^b_{C_s} = \mathbf{x}_{C_a} \tag{18}$$

$$P^b_{C_s} = P_{C_a} \tag{19}$$

$$K^b_{12} = P_{SC} P^{-1}_{C_s} \tag{20}$$

$$P^b_{SC} = K^b_{12} P^b_{C_s} \tag{21}$$

$$P^b_S = P_S + K^b_{12}(P^b_{CS} - P_{CS}) \tag{22}$$

$$\mathbf{x}^b_S = \mathbf{x}_S + K^b_{12}(\mathbf{x}^b_{C_s} - \mathbf{x}_{C_s}), \tag{23}$$

where $\mathbf{x}^b_{C_s}$, $\mathbf{x}^b_S$, $P^b_{C_s}$, P^b_{SC}, P^b_{CS} and P^b_S are updated estimates of $\mathbf{x}_{C_s}$, $\mathbf{x}_S$, P_{C_s}, P_{SC}, P_{CS} and P_S after back propagation. Note that P^b_{CS} is transpose of P^b_{SC} due to the symmetry of covariance matrix.

Differently to CI submaps scenario, back propagation processes in our special case is simplified as back-propagation is not applied in both directions. The consideration here is that the two maps are obtained using different sensors (one accurate, the other not). As the shared mean estimate ($\mathbf{x}_{C_a}$ and $\mathbf{x}_{C_s}$) and covariance (P_{C_a} and P_{C_s}) of two maps represents robot positions used for sound landmarks initialisation, they are mainly estimated by the localisation map anyway. A minor contribution from the sound source map to this robot locations ($\mathbf{x}_{C_a}$ and $\mathbf{x}_{C_s}$) is disregarded due to the following reasons. Firstly, sound sources are sparse in time axis and in most cases total number of sound sources that are reliably detected at each robot pose are a lot less than visual or laser features. Secondly, in reverberating

TABLE I

PARAMETERS IN SIMULATION

Parameters	Values
Part I	
Distance per odometry step	0.2m
Odometry noise (Trans. and Orient.)	0.001m and 0.001 deg
Sound bearing noise (Azimuth & Elevation)	10 deg
Least square optimizer	Levenberg-Marquardt
Part II	
Noise of range bearing sensor	0.01m and 1 deg
Odometry noise (Trans. and Orient.)	0.02m and 5 deg

indoor environments, accuracy of the bearing observations of sound sources cannot be compared to that of visual or laser landmarks so uncertainties of sound sources locations are higher. As a result, when the sound source map gets updated, robot positions used to initialise sound source locations $\mathbf{x}_{C_s}$ and its covariance P_{C_s}, which are copied from $\mathbf{x}_{C_a}$ and P_{C_a} during last back propagation step from the localisation map, only have negligible change. Therefore we assume,

$$\mathbf{x}_{C_s} \approx \mathbf{x}_{C_a} \tag{24}$$

$$P_{C_s} \approx P_{C_a}, \tag{25}$$

losing only a small part of the information and avoiding the back propagation step from the sound source map to the localisation map, which incurs in extra time complexity.

III. SIMULATION AND EXPERIMENTAL RESULTS

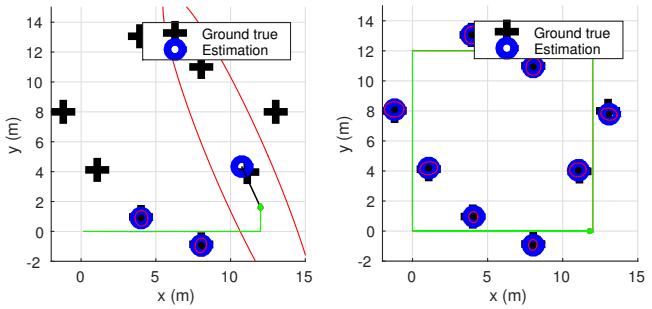
In this section, comprehensive simulation and experimental results are presented to evaluate and compare the method described in section II to the optimal and other possible solutions.

A. Simulation Results

1) Sound sources mapping with only odometry information: In the simulation scenario shown in Fig. 3(a), the robot follows a square trajectory using only information from odometry and sound sources. When it reaches its original position, it continues to travel along X axis for loop closure. First, we set the wheel odometry to be very accurate to allow accurate sound mapping. Later, we increase odometry noise gradually to see the effect in sound mapping. At each time step, the robot moves a fixed distance and random Gaussian noise is linearly added. The parameters used in the simulation are shown if Table I part I. Bearing estimation noise is set to ± 10 degrees as in typical indoor environments, where sound reverberation is present.

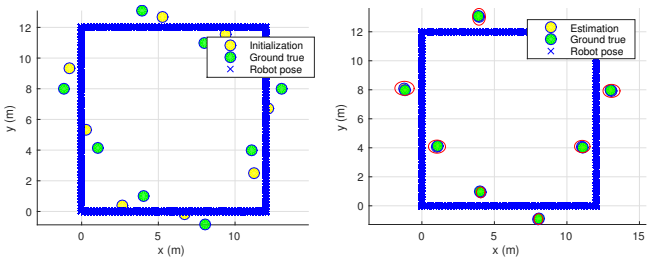
In this simulation scenario, we studied both EKF with IDP parametrisation method and least square optimisation method in 2D and 3D cases. Initialisation and final estimation results of 2D case are shown in Fig. 3 and Fig. 4. Similar results are obtained from 3D simulation. From those figures, it can be seen that sound mapping works well under very accurate odometry.

A 20 runs Monte Carlo simulation show that by increasing odometry noise, the sound mapping estimation fails even



(a) Initial state of IDP sound sources. (b) Final estimation results.

Fig. 3. Initial and final estimation using EKF and IDP with highly accurate odometry information. Simulation is in 2D.

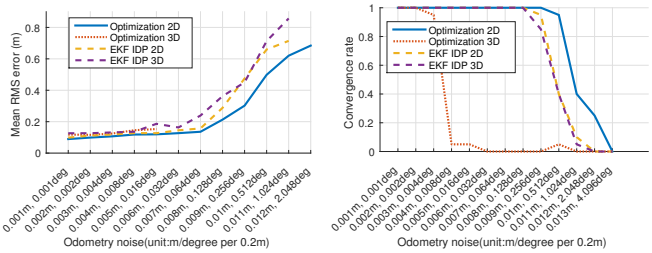


(a) Initial state of IDP sound sources. (b) Final estimation results.

Fig. 4. Initial and final estimation using least square optimisation with highly accurate odometry information. Simulation is in 2D.

with very reasonable noise values of less than 5% the displacement. As shown in Fig. 5, for a range of odometry noises the mean RMS errors of estimated sound source locations grows exponentially with time. The figure also present the convergence rate for all algorithms.

2) Sound sources mapping with odometer and range-bearing observations of environment landmarks: We also simulated a scenario adding an exteroceptive sensor (e.g. a laser scanner), which observes range and bearing information of point landmarks in the environment (e.g. corner points). In this simulation scenario, the proposed method utilises an EKF-SLAM algorithm for the localisation map fusing this additional range and bearing observations. Robot locations at sound sources initialisation instants are used as common elements of two maps as explained before. In the simulation, the robot follows the same trajectory as before and comes



(a) RMS errors. (b) Convergence rates.

Fig. 5. RMS errors and convergence rates of sound sources mapping under different odometry noise using EKF IDP parametrization and least square optimization in 2D and 3D cases with 20 Monte Carlo runs for each case.

back to its original point for loop closure. The parameters used for the additional sensors are shown in Table I part II and other parameters are in Table I part I. The odometry noise is set at typical levels of a real mobile platform ($\sim 10\%$ of the displacement) to reflect a more general scenario.

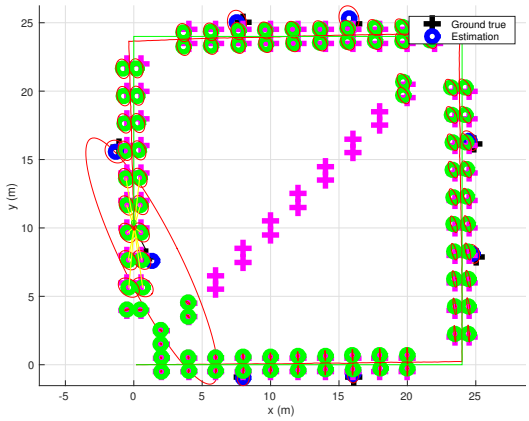
Simulation results are shown in Fig. 6. From the figure, it is clear that an additional sensor allows accurate sound mapping under typical odometry noise. From sub figure (a) and (b), it can be seen that before loop closure happens, environment landmarks and sound sources mean estimation are drifting (although the filter is still consistent). From sub figure (c) and (d), we can see that after the loop closure, drifted landmarks are corrected in Y axis of the localisation map. Since some robot locations are shared between the two maps, the estimated positions of sound landmarks are also updated in Y axis after the back propagation process.

Next, we compared the proposed method with the optimal SLAM solution of sound source mapping using a single map, whose state vector contains both landmarks with range and bearing observations and sound sources (we refer to it as full SLAM). In full SLAM method, we use EKF SLAM algorithm and parametrise sound sources using IDP. We compared the proposed method with the full SLAM method in terms of sound mapping accuracy with various trajectory lengths. For each trajectory, A 10 runs Monte Carlo simulation is used to compute the Mean RMS errors. The results are shown in Fig.7. From the figure, we can see that our proposed method has a comparable accuracy with the full SLAM method, which means that the approximation made (back propagation from the second map to the first can be neglected) is reasonable. In addition, the overall execution time of the proposed method is slightly smaller than the full SLAM method (e.g. 0.0142s with the proposed method and 0.0161s with full SLAM method for 185m trajectory at one EKF step). In the full SLAM method, when the robot trajectory is relatively long, in some runs the localisation error is large. A reason might be related to linearisation errors due the extremely noisy sound bearing-only observations becoming high and negatively impact on the robot trajectory estimation. Our method avoids this issue by semi-decoupling the two sensors observations so the noisy information of sound sources sensor does not propagate back to the localisation, not affecting the robot pose estimation of the localisation map and as result producing more accurate results the full EKF SLAM. Note that in an optimisation SLAM framework this issue will not be present producing better results than our proposed method, but at the cost of execution time (e.g. 99.365s for 185m trajectory).

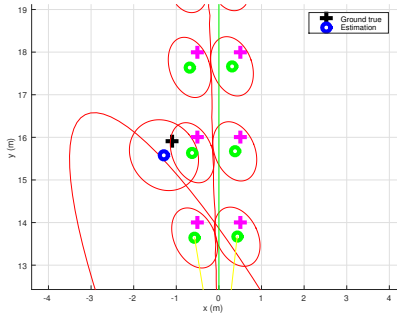
B. Experimental Results

In this section, two different experimental scenarios are used to show the effectiveness and flexibility of the proposed method.

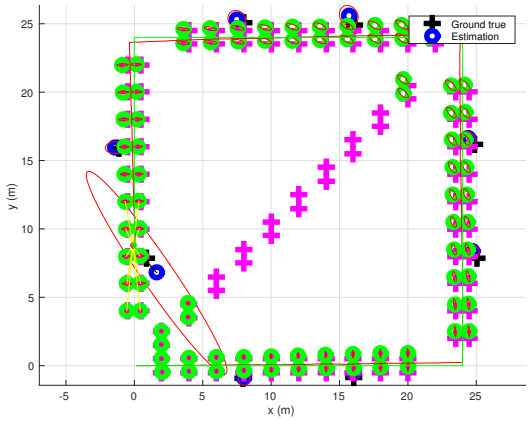
1) 2D sound sources mapping by a mobile robot with a microphone array and a laser scanner: A turtlebot with Hokuyo laser range finder and Microcone (6-microphone circular array (top one is not used)) is used to localise two



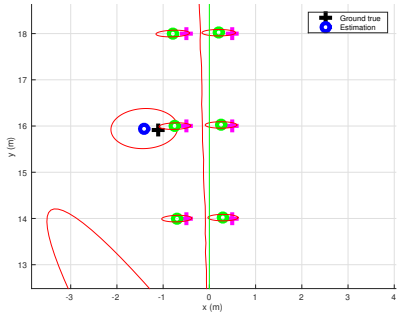
(a) Sound source mapping before loop closure.



(b) Zoomed in view of top left sound source before loop closure.



(c) Sound source mapping after loop closure.



(d) Zoomed view of the top left sound source after loop closure.

Fig. 6. Sound sources mapping with additional range-bearing observations of environment landmarks before and after loop closure. In all figures, green circular markers represent estimated environment landmarks, pink plus markers represent ground true locations of range-bearing landmarks, read eclipses represent 3 σ region, green line represents ground true robot trajectory and red line represents estimated robot trajectory.

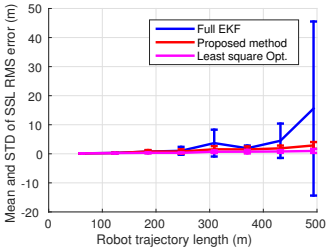


Fig. 7. Mean RMS errors with STD under various length of robot trajectories for 10 runs Monte Carlo simulation each case.

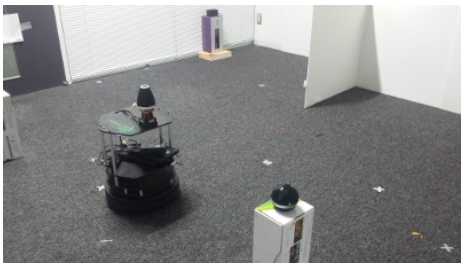
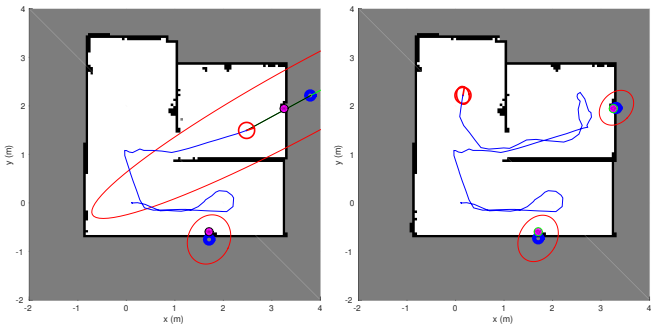


Fig. 8. Turtelbot equipped with lase scanner and Microcone (circular microphone array).

sound sources generating white noise (see Fig. 8). We use the EKF-SLAM describe above for the sound map and the pose SLAM implementation in [13] as SLAM framework to estimate the localisation map. In our case robot poses that are used for sound source initialisation, their covariance and cross correlations are share at each SLAM step with the sound source map. Then the shared part of the state-vector allow us to back propagated the information to the sound source map after each update in the localisation map. Sound bearing observation noise is set to ± 10 deg. HARK [1] is used for sound source bearing estimation using MUSIC algorithm.

The results are shown in Fig. 9. It can be seen that the proposed method has successfully estimated two sound sources with reasonable good accuracy given the noisy nature of the audio observations.



(a) Sound landmark initialization with IDP parametrization. (b) Final estimation results.

Fig. 9. 2D sound sources mapping results using a mobile with laser scanner. In all figures, blue markers represent estimated sound landmarks, pink markers represent ground true locations, read eclipses represent 3σ region and blue line represents estimated robot trajectory.

(a) (b)

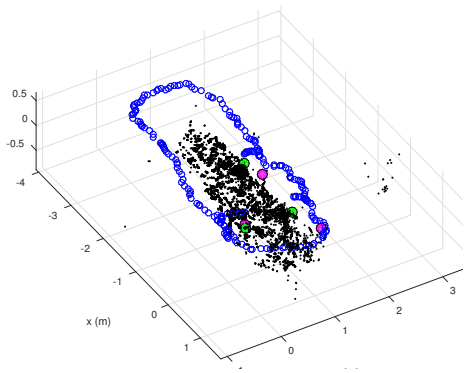
Fig. 10. PS3-eye configuration (a) and experimental setup (b).

2) 3D sound sources mapping using a hand held PS3-eye (monocular camera with linear microphone array): And off-the-shelf visual SLAM implementation without any modification is used in this experiment. ORB-SLAM [15] is used at first to estimate the localisation map. Estimated sensor poses on keyframes are used to initialise sound sources so that these poses can be updated at each time the ORB-SLAM runs a local bundle adjustment. Current sensor pose is also obtained from the newest keyframe pose so that pose covariance can be available. Sound bearing observation noise is measured at different azimuth angle since the linear array has different sensitivity at different azimuth directions. As the linear microphone array cannot provide elevation angle observations, the observation noise is set quite large (± 60 deg) to hint that the sound source is in front of the sensor (due to the casing for PS3-eye, it mostly detect sound sources in front). The sound map is the same in our previous experiment, in this case with three sound sources from two mobile phones and one pad playing music and speech.

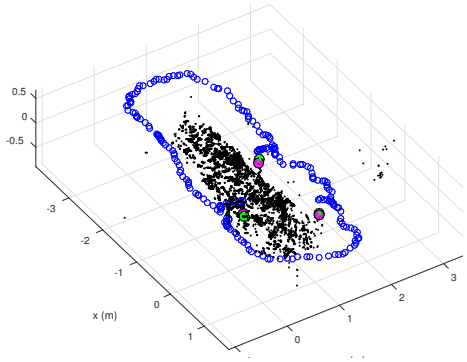
The final estimation results are shown in Fig. 10. Note that the SLAM from a monocular camera can only provide robot poses and feature points locations up to scale. So the scale factor is recovered by manually marking three locations of the sensor trajectory to align estimation results with ground true locations. From sub figure (b), we can see that the sensor trajectory is drifted before loop closure. Therefore, the estimated sound sources locations are also drifted. From sub figure (c), we can see that after loop closure is detected, sensor trajectory is corrected and so do position estimates of sound sources. This is again thanks to the split CI maps. From the experiment, we also can see that although the linear microphone array only provides azimuth angle (which means 3D estimation lacks 1DOF), with the help of the mono camera observations, it is sufficient to obtain an accurate sound sources mapp in 3D with the proposed method.

IV. Conclusion

In this paper, we proposed a split CI mapping method for sound source mapping and robot localisation. Our method utilises two SLAM algorithm algorithms running in parallel with some common information used to propagate information unidirectionally. One SLAM algorithm is in charge of estimating accurately the location of the sensor, while the other is used for sound sources mapping parameterised as inverse-depth points. As sound sources observation are



(a) Estimation results (before loop closure).



(b) Final estimation results (after loop closure).

Fig. 11. 3D sound sources mapping results using a hand hold PS3-eye (monocular camera with linear microphone array). In all figures, green markers represent estimated sound sources, pink markers represent ground true locations, blue markers represent key frames' locations and black dots represent final feature points from ORB-SLAM.

bearing-only, extremely noisy and sparse, they are not use for localisation. However, any update in the localisation reflects back to the sound source mapping by exploiting the conditional independence between split maps.

Moreover, we propose to use inverse-depth parametrisation to represent the sound sources locations. The key advantage of using IDP is that models accurately uncertainty of faraway points, utilises all information contained in bearing-only sound observations and linearisation errors are small than with Euclidean points.

The proposed method is flexible enough to allow the use of off-the-shelf SLAM implementations (optimisation or filter-based) to estimate the localisation map. It is also flexible to be used with any relatively accurate exteroceptive sensor such as lasers or cameras.

Although some approximation are made to the otherwise optimal solution, the extensive simulation and experimental results show that our method produces consistent and bounded estimation quite close to the maximum aposteri solution produced by least-square optimisation or EKF approaches.

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," *Advanced Robotics*, vol. 24(5-6), pp. 739–761, 2010.

[2] J. Scholtz, J. Young, J. L. Drury, and H. A. Yanco, "Evaluation of human-robot interaction awareness in search and rescue," in *2004 IEEE International Conference on Robotics and Automation (ICRA 2004)*, 2004, pp. 2327–2332.

[3] H. G. Okuno, K. Nakadai, K. ichi Hidai, H. Mizoguchi, and H. Kitano, "Human-robot interaction through real-time auditory and visual multiple-talker tracking," in *2001 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2001)*, 2001, pp. 1402–1409.

[4] K. Nakadai, H. G. Okuno, and H. Kitano, "Real-time sound source localization and separation for robot audition," in *INTERSPEECH*, 2002.

[5] D. M. Cole and P. M. Newman, "Using laser range data for 3D SLAM in outdoor environments," in *2006 IEEE International Conference on Robotics and Automation (ICRA 2006)*, 2006, pp. 1556–1563.

[6] J. Civera, A. J. Davison, and J. M. Montiel, "Inverse depth parametrization for monocular SLAM," *IEEE Transactions on Robotics*, vol. 24(5), pp. 932–945, 2008.

[7] 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(1-2), pp. 135–152, 2011.

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

[9] N. Kallakuri, J. Even, Y. Morales, C. Ishi, and N. Hagita, "Probabilistic approach for building auditory maps with a mobile microphone array," in *2013 IEEE International Conference on Robotics and Automation (ICRA 2013)*, 2013, pp. 2270–2275.

[10] E. Vincent, A. Sini, and F. Charpillet, "Audio source localization by optimal control of a mobile robot," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, 2015, pp. 5630–5634.

[11] J. Even, Y. Morales, N. Kallakuri, J. Furrer, C. T. Ishi, and N. Hagita, "Mapping sound emitting structures in 3D," in *2014 IEEE International Conference on Robotics and Automation (ICRA 2014)*, 2014, pp. 677–682.

[12] P. Pinis and J. D. Tards, "Large-scale slam building conditionally independent local maps: Application to monocular vision," *IEEE Transactions on Robotics*, vol. 24(5), pp. 1094–1106, 2008.

[13] V. Ila, J. M. Porta, and J. Andrade-Cetto, "Information-based compact Pose SLAM," *IEEE Transactions on Robotics*, vol. 26(1), pp. 78–93, 2010.

[14] F. Endres, J. Hess, J. Sturm, D. Cremers, and W. Burgard, "3-D mapping with an RGB-D camera," *IEEE Transactions on Robotics*, vol. 30(1), pp. 177–1873, 2014.

[15] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 30(5), pp. 1147–1163, 2015.

[16] J. Sola, T. Vidal-Calleja, J. Civera, and J. M. M. Montiel, "Impact of landmark parametrization on monocular EKF-SLAM with points and lines," *International journal of computer vision*, vol. 97(3), pp. 339–3683, 2012.