Towards Real-Time 3D Sound Sources Mapping with Linear Microphone Arrays

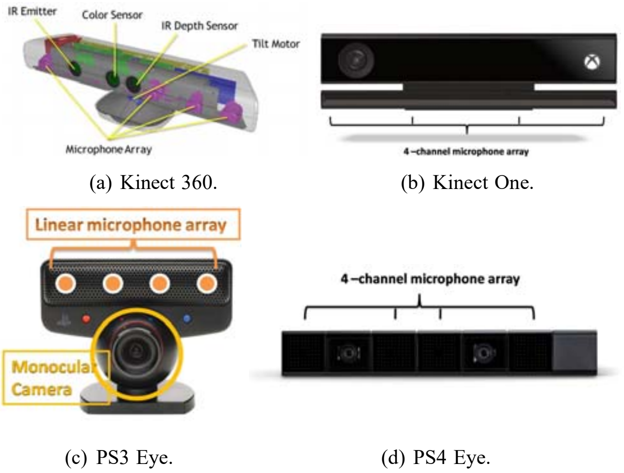
Daobilige Su, Teresa Vidal-Calleja and Jaime Valls Miro

Abstract— **In this paper, we present a method for real-time 3D sound sources mapping using an off-the-shelf robotic perception sensor equipped with a linear microphone array. Conventional approaches to map sound sources in 3D scenarios use dedicated 3D microphone arrays, as this type of arrays provide two degrees of freedom (DOF) observations. Our method addresses the problem of 3D sound sources mapping using a linear microphone array, which only provides one DOF observations making the estimation of the sound sources location more challenging. In the proposed method, multi hypotheses tracking is combined with a new sound source parametrisation to provide with a good initial guess for an online optimisation strategy. A joint optimisation is carried out to estimate 6 DOF sensor poses and 3 DOF landmarks together with the sound sources locations. Additionally, a dedicated sensor model is proposed to accurately model the noise of the Direction of Arrival (DOA) observation when using a linear microphone array. Comprehensive simulation and experimental results show the effectiveness of the proposed method. In addition, a real-time implementation of our method has been made available as open source software for the benefit of the community.**

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 [2], [3], [4] have gained increasing interest since the ability to localise sound sources has many potential applications in scenarios such as robotic urban search and rescue (USAR) [5]. In these scenarios, positions of sound sources can be used to locate missing people in a disastrous sites. Another example application includes human robot interaction (HRI), where location of sound sources can be used to detect and track speakers [6] or discern between multiple people speech [7].

3D cameras such as Microsoft Kinect 360, Kinect One, PS3 Eye and PS4 Eye sensors, as shown in Fig. 1, are more and more used as part of the perception modules of robotic and intelligent systems. A common feature of the microphone arrays on these sensors is that the geometric location of the all microphones are distributed along a straight line, i.e. in a linear array, be it uniformly distributed (in Fig. 1 (b) and (c)) or not (in Fig. 1 (a) and (d)).



(a) Kinect 360.

(b) Kinect One.

(c) PS3 Eye.

(d) PS4 Eye.

Fig. 1. Typical robotic sensors that include a linear microphone array.

Despite easy availability at an affordable price and frequent usage of sensors with a linear microphone array in robotic systems, conventional 3D sound sources mapping methods hardly use this configuration. This is because a linear microphone array only provides 1 DOF estimation (angle between the line connecting a sound source and the origin and the axis of the linear array) out of 3 DOF (2 DOF bearing estimation in terms of azimuth and elevation angles plus 1 DOF estimation of range). This lack of observability makes the 3D mapping of multiple sound sources more challenging.

In recent work of 2D sound sources mapping, Hu et. al. in [2] proposed a FastSLAM based approach to map multiple sound sources using a 3D microphone array. Sasaki et. al. in [3] uses a self motion triangulation method to deal with sound sources mapping using a concentric microphone array. A ray casting based probabilistic 2D sound sources mapping approach is proposed by Kallakuri et. al. in [4]. Conventional approaches such as [8], [9] for mapping stationary sound sources in 3D space usually require a 3D microphone array, which can be used to estimate both azimuth and elevation angles of sound sources. In [8], Even et. al. extend their previous work in [4] to the 3D case by using a 3D microphone array. In [9], Kotus et. al. also use a 3D multi channel acoustic vector sensor to estimate azimuth and elevation angles of sound sources and estimate their 3D location by integrating prior knowledge of the shape of the room. Some other work in 3D sound sources mapping even use multiple microphone arrays. In [10], Ishi et. al. use multiple 3D microphone arrays attached on the ceiling to estimate 3D locations of multiple sound sources. They also exploit the

All authors are with the Centre for Autonomous Systems at the Faculty of Engineering and IT, University of Technology Sydney, Australia. daobilige.su@student.uts.edu.au, {teresa.vidalcalleja, jaime.vallsmiro} @uts.edu.au

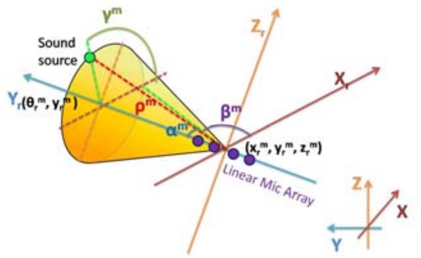


Fig. 2. Linear microphone array notation and parametrisation of a 3D sound source location.

reflection information to improve the localisation accuracy. In [11], Seewald et. al. use two perpendicularly placed Microsoft Kinects to estimate 3D locations of sound sources. Note that in [9], [10], [11], all sensors and sound sources are static.

In this paper, we present a method to map 3D sound sources using a robotic perception sensor equipped with a linear microphone array. First, we propose a new parametrisation within a multi-hypotheses tracking framework to obtain a good initial guess for the location of sound sources. Then, an optimisation approach is used to jointly estimate 6 DOF poses of the sensor and 3 DOF locations of sound sources together with visual landmarks.

The contribution of this paper is two-fold: firstly we introduce a framework that allows to map in real-time the location of 3D sound sources using a linear microphone array without any prior knowledge of the sensor hardware as we did in our previous work [12]. Secondly, we propose a new sensor model, which is able to handle the sensor noise in a microphone array. In addition, we release code of real-time implementation open source[1] for the benefit of the community.

II. Sensor Model for A Linear Microphone Array

A graphical representation of the sensor model of a linear microphone array is shown in Fig. 2. The axis of the linear microphone array coincides with the Y axis. The observation of a linear microphone array is the angle β^m, which is the complementary angle of α^m ($\beta^m = \pi - \alpha^m$) that is the angle between the straight line connecting the location of a sound source and the origin of the microphone array and the Y axis. Let $\mathbf{p}^m = [x^m_{ss}, y^m_{ss}, z^m_{ss}]^T$ be the Euclidean coordinates of the mth sound source and $\mathbf{x}_{r,k}$ be sensor pose at time instance k. The observation β^m_k of this sound source $\mathbf{p}^m$ using the linear microphone array from the sensor pose $\mathbf{x}_{r,k}$ is

$$\begin{bmatrix} {}^m\mathbf{p}_k \\ 1 \end{bmatrix} = \mathrm{M}^{-1}(\mathbf{x}_{r,k}) \begin{bmatrix} \mathbf{p}^m \\ 1 \end{bmatrix}, \tag{1}$$

$$\beta^m_k = atan2({}^m\mathbf{p}_k(2), \sqrt{{}^m\mathbf{p}_k(1)^2 + {}^m\mathbf{p}_k(3)^2}), \tag{2}$$

where $\mathrm{M}(\mathbf{x}_{r,k})$ is the homogeneous transformation of the sensor pose $\mathbf{x}_{r,k}$, ${}^m\mathbf{p}_k$ is the local coordinate of the sound source $\mathbf{p}^m$ in the reference coordinate frame of the sensor pose $\mathbf{x}_{r,k}$ and function $atan2(\cdot)$ returns the four-quadrant inverse tangent angle.

The observation β^m_k, in practice, is obtained by processing a multi channel audio signal. The Time Difference of Arrival (TDOA) from a sound source to all channels of the microphone array is commonly exploited to estimate the DOA observation β^m_k. Typical methods for estimation of DOA from multi channel audio signal include MUSIC [13] and SRP-PHAT [14]. These algorithms search all possible DOA angles and assign likelihood values to them, and angles with local maximum likelihoods are treated as the estimation of DOAs corresponding to the sound sources.

Due to the presence of noise in audio signal, the estimated angle from a DOA estimation algorithm $\hat{\beta}^{m,DOA}_k$ is affected by noise. Unlike azimuth angle estimation using a circular microphone array, whose DOA estimation noise can be approximated by a constant value, the noise level of the estimated DOA observation $\hat{\beta}^{m,DOA}_k$ of a linear microphone array depends on true value of DOA observation β^m_k. This is because the sensitivity of a linear microphone array is different at different DOA angle. When β^m_k is close to 0 rad, the sensitivity is higher and the uncertainty of the estimation is lower, while when β^m_k is close to the two limits ($\pm\pi/2$ rad), the sensitivity degrades and uncertainty of estimation becomes larger. Moreover, the mean value of DOA estimation $\hat{\beta}^{m,DOA}_k$ does not necessarily coincide with the true value, especially around $\pm\pi/2$. Therefore the raw result of the DOA estimation is not suitable for being used as a noisy observation of a linear array directly, which will be later used to map sound sources in the 3D space.

In order to obtain a reliable observation of the DOA, we introduced a sensor model based on a Gaussian Process [15]. This kind of sensor model aims to transfer the raw biased estimation result into a normally distributed function, whose mean values locate near the true values and uncertainty values change according to different DOA angles. The GP sensor model is formulated as follows,

$$\boldsymbol{\beta}_{gp} \sim \mathrm{N}(\mathbf{o}, K(\hat{\boldsymbol{\beta}}^{DOA}_{gp}, \hat{\boldsymbol{\beta}}^{DOA}_{gp}) + \sigma^2_n \mathbf{I}), \tag{3}$$

where $\hat{\boldsymbol{\beta}}^{DOA}_{gp}$ is a set of raw results from the DOA estimation algorithm, $\boldsymbol{\beta}_{gp}$ is the corresponding set of ground truth values, $K(\cdot)$ is pre-defined Kernel function and σ_n is the variance of the noise. $\hat{\boldsymbol{\beta}}^{DOA}_{gp}$ and $\boldsymbol{\beta}_{gp}$ are used to train the GP sensor model.

When a new data $\hat{\beta}^{DOA}_{gp*}$ from the DOA estimation algorithm is available, the joint Gaussian distribution is

$$\begin{bmatrix} \boldsymbol{\beta}_{gp} \\ \beta_{gp*} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \begin{bmatrix} K(\hat{\boldsymbol{\beta}}^{DOA}_{gp}, \hat{\boldsymbol{\beta}}^{DOA}_{gp}) + \sigma^2_n \mathbf{I} & K(\hat{\boldsymbol{\beta}}^{DOA}_{gp}, \hat{\beta}^{DOA}_{gp*}) \\ K(\hat{\beta}^{DOA}_{gp*}, \hat{\boldsymbol{\beta}}^{DOA}_{gp}) & K(\hat{\beta}^{DOA}_{gp*}, \hat{\beta}^{DOA}_{gp*}) \end{bmatrix} \right), \tag{4}$$

where β_{gp*} is the predicted DOA estimation from GP. Then, the mean and covariance of the predicted DOA from the GP

[1] Open source implementation and experimental data is available on https://github.com/daobilige-su/SSM LinearArray.

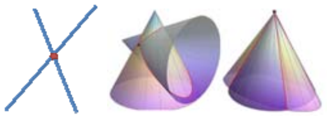

(a) (b)

Fig. 3. Intersection of two 3D bearings (a) and cone surfaces (b).

sensor model can be computed as follows,

$$\hat{\beta}_{gp*} = K(\hat{\beta}^{DOA}_{gp*}, \hat{\beta}^{DOA}_{gp})(K(\hat{\beta}^{DOA}_{gp*}, \hat{\boldsymbol{\beta}}^{DOA}_{gp}) + \sigma_n^2\mathbf{I})^{-1}\boldsymbol{\beta}_{gp}, \quad (5)$$

$$P^{\beta}_{gp*} = K(\hat{\beta}^{DOA}_{gp*}, \hat{\beta}^{DOA}_{gp*}) - K(\hat{\beta}^{DOA}_{gp*}, \hat{\boldsymbol{\beta}}^{DOA}_{gp})(K(\hat{\boldsymbol{\beta}}^{DOA}_{gp}, \hat{\boldsymbol{\beta}}^{DOA}_{gp}) + \sigma_n^2\mathbf{I})^{-1}K(\hat{\boldsymbol{\beta}}^{DOA}_{gp}, \hat{\beta}^{DOA}_{gp*}). \quad (6)$$

A squared exponential kernel function

$$k_{i,j} = \sigma_f^2 exp(-\frac{1}{2\pounds^2}(\hat{\boldsymbol{\beta}}^{DOA}_{gp}(i) - \hat{\boldsymbol{\beta}}^{DOA}_{gp}(j))^2) \quad (7)$$

is used in our GP sensor model. In Eq. 7, $k_{i,j}$ denotes the ith row and jth column of covariance K, and $\hat{\boldsymbol{\beta}}^{DOA}_{gp}(i)$ and $\hat{\boldsymbol{\beta}}^{DOA}_{gp}(j)$ are ith and jth data of $\hat{\boldsymbol{\beta}}^{DOA}_{gp}$ or $\hat{\beta}^{DOA}_{gp*}$. The maximum of the marginal likelihood is used to train the set of hyper-parameters σ_f,£ and σ_n as described in [15].

III. Initialisation of Sound Sources using Multi Hypotheses

As mentioned above, a linear microphone array provides 1 DOF observation out of 3 DOF of the sound source position. This means that given an angle observation α^m, the sound source can be located anywhere on a cone surface, which extends from the sensor location to infinity, as shown by the yellow surface in the Fig. 2. This produces a partial observability which introduces an great difficulty in the initialisation of the sound sources in the map. This issue is similar to the one on point feature initialisation in monocular SLAM [16]. In monocular SLAM, visual point features parametrised by their Euclidean coordinates can be initialised after triangulating two 3D bearing observations as shown in Fig. 3 (a). However, intersection of two cone surfaces is more complicated to model with simple Gaussian distribution as shown in Fig. 3 (b).

In order to initialise the sound source location, a multi-hypotheses strategy is required, which will allow us to model correctly the uncertainty. Tracking these hypotheses until they have converged would allow us to use a joint optimisation algorithm to estimate sensor poses, other landmarks and sound sources together.

Firstly, we parametrise the state of mth sound source as follows,

$$\mathbf{s}^m = (\beta^m, \gamma^m, \rho^m)^T. \quad (8)$$

Note that in Eq. 8, we use symbol $\mathbf{s}$ to represent the proposed parametrisation of the sound source state instead of the Euclidean coordinates parametrisation of $\mathbf{p}$. In Eq. 8, $\beta^m, \gamma^m, \rho^m$ are axis angle, circumferential angle and inverse depth of the sound source as shown in Fig. 2. As can be seen from the figure, the origin of the sensor coordinate frame is (x^m_r, y^m_r, z^m_r), the azimuth and elevation angle of positive Y axis are $(\theta^m_r, \varphi^m_r)$. These five parameters come from the sensor pose at the first observation of the sound source, and once they are fixed, the axis and direction of the linear microphone array on global coordinate is determined. The remaining DOF, the roll angle along Y axis, is not required, since the cone surface is the same with different roll angles. The anchor axis of the sound source location is therefore parametrised as follows,

$$\mathbf{x}^m_{ss,axis} = (x^m_r, y^m_r, z^m_r, \theta^m_r, \varphi^m_r)^T. \quad (9)$$

Note that $\mathbf{x}^m_{ss,axis}$ needs to be stored to recover the sound source locations when multi hypotheses initialisations have converged. The axis angle β^m determines the angle of the cone, and its initial value comes from the predicted DOA angle $\beta^m_{gp*,ini}$ obtained by the GP sensor model at the first observation of the sound source. The circumferential angle γ^m is the angle between the positive X axis and the direction pointing from the origin of the sensor coordinate frame to the projected point of sound source on X,Z plane of the sensor coordinate frame. The inverse depth ρ^m is the inverse of the distance as defined for the inverse depth parametrisation (IDP) in the visual SLAM algorithm in [17] [18].

Among three parameters determining the state of the sound source, two of them, the circumferential angle γ^m and the inverse depth ρ^m, are unobservable at the first observation of the sound source. We can initialise the inverse depth $\rho^m = 1/3d_{min}$ the same way as visual SLAM [17] [18], where d_{min} is the minimum possible distance from the sound source to the sensor coordinates origin. ρ^m will converge after observing the same sound source with some parallax. To initialise the circumferential angle γ^m, we introduce a multi hypotheses framework. Specifically, we divide the range of the possible circumferential angles into N_h spaces and each hypothesis covers one region. Let the state of the sound source m in the ith hypothesis be

$$\mathbf{s}^{m,i} = (\beta^{m,i}, \gamma^{m,i}, \rho^{m,i})^T, \quad (10)$$

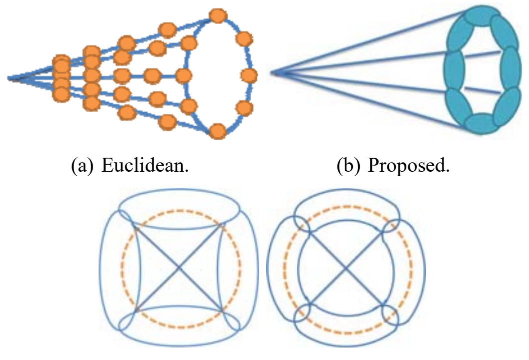
where the circumferential angle $\gamma^{m,i}$ is uniformly distributed along the range $-\pi$ to π as follows,

$$\gamma^{m,i} = \frac{2\pi}{N_h}i - \pi, i \in (1 \cdots N_h). \quad (11)$$

The covariance of the mth sound source in the ith hypothesis can be initialised as follows,

$$\mathbf{P}^{m,i}_{ss} = \begin{bmatrix} P^{\beta,m}_{gp*,ini} & 0 & 0 \\ 0 & \frac{\pi}{N_h}^2 & 0 \\ 0 & 0 & \frac{1}{3d_{min}^2} \end{bmatrix}, \quad (12)$$

where $P^{\beta,m}_{gp*,ini}$ is the predicted variance of DOA angle $\hat{\beta}^m_{gp*,ini}$ using the GP sensor model. The covariance of the inverse depth is the same as suggested in [18]. The covariance of the circumferential angle is set to $(\pi/N_h)^2$ so


(a) Euclidean. (b) Proposed.

(c) IDP. (d) Proposed.

Fig. 4. Multi hypotheses using (a) Euclidean (c) IDP and (b),(d) the proposed parametrisation.

that one sigma region of all hypotheses covers all possible range.

The advantage of the proposed parametrisation is shown in Fig. 4. When using the Euclidean parametrisation for multi hypotheses as shown in the subfigure (a), infinite Euclidean points, hence infinite hypotheses, are needed to represent the cone surface extending to infinity, while the proposed parametrisation only needs several hypotheses to represent the cone surface thanks to the inverse depth as shown in subfigure (b). When IDP [17] is used, there exists a polygon effect when looking from the right side of the cone as shown in subfigure (c), especially less number of hypotheses are used. With the proposed parametrisation, the polygon effect does not exist and the cone surface is represented better as shown in the subfigure (d).

As the sensor gets more observations of the sound source, the state of the sound source can be updated as follows by using an extended Kalman filtering strategy,

$$\hat{z}_k^{m,i} = atan2(\mathbf{p}_{l,k}^{m,i}(2), \sqrt{\mathbf{p}_{l,k}^{m,i}(1)^2 + \mathbf{p}_{l,k}^{m,i}(3)^2}), \quad (13)$$

$$z_k^{m,i} = \beta_{gp*,k}^{m}, \quad (14)$$

$$Q_k^{m,i} = P_{gp*,k}^{\beta,m}, \quad (15)$$

$$\mathbf{K}_k^{m,i} = \mathbf{P}_{ss,k-1}^{m,i}(\mathbf{H}_k^{m,i})^T / (\mathbf{H}_k^{m,i}\mathbf{P}_{ss,k-1}^{m,i} (\mathbf{H}_k^{m,i})^T + Q_k^{m,i}), \quad (16)$$

$$\mathbf{s}_k^{m,i} = \mathbf{s}_{k-1}^{m,i} + \mathbf{K}_k^{m,i} f_{na}(z_k^{m,i} - \hat{z}_k^{m,i}), \quad (17)$$

$$\mathbf{P}_{ss,k}^{m,i} = (\mathbf{I} - \mathbf{K}_k^{m,i}\mathbf{H}_k^{m,i})\mathbf{P}_{ss,k-1}^{m,i}, \quad (18)$$

where $\hat{z}_k^{m,i}$ is the expected observation of the mth sound source in the ith hypothesis at time instant k, $z_k^{m,i}$ is the actual observation from GP sensor model, $Q_k^{m,i}$ is the observation noise coming from GP sensor model, $\mathbf{K}_k^{m,i}$ is the Kalman gain, $\mathbf{H}_k^{m,i}$ is Jacobian of the sensor observation under the proposed parametrisation, $\mathbf{s}_{k-1}^{m,i}$, $\mathbf{P}_{ss,k-1}^{m,i}$, $\mathbf{s}_k^{m,i}$, $\mathbf{P}_{ss,k}^{m,i}$ are the mth sound source state and the associated covariance in the ith hypothesis at time instance $k-1$ and k. $f_{na}(\cdot)$ is the function to normalise an angle between $-\pi$ to π and $\mathbf{p}_{l,k}^{m,i}$ is the Euclidean coordinate of the mth sound source in ith hypothesis under sensor local coordinate. $\mathbf{p}_{l,k}^{m,i}$ can be computed from the sound source state $\mathbf{s}_{k-1}^{m,i}$ and the current sensor pose $\mathbf{x}_{r,k}$ as

$$\mathbf{p}_{k-1}^{m,i} = f_{eul\,mat}(\mathbf{x}_{ss,axis}^{m}(1), \mathbf{x}_{ss,axis}^{m}(2), \mathbf{x}_{ss,axis}^{m}(3), \mathbf{x}_{ss,axis}^{m}(4) - \pi/2, 0, \mathbf{x}_{ss,axis}^{m}(5)) \begin{bmatrix} \frac{cos(\mathbf{s}_{k-1}^{m,i}(1))cos(\mathbf{s}_{k-1}^{m,i}(2))}{\mathbf{s}_{k-1}^{m,i}(3)} \\ \frac{sin(\mathbf{s}_{k-1}^{m,i}(1))}{\mathbf{s}_{k-1}^{m,i}(3)} \\ \frac{cos(\mathbf{s}_{k-1}^{m,i}(1))sin(\mathbf{s}_{k-1}^{m,i}(2))}{\mathbf{s}_{k-1}^{m,i}(3)} \\ 1 \end{bmatrix}, \quad (19)$$

$$\mathbf{p}_{l,k}^{m,i} = [\mathbf{I}_3\mathbf{0}]M^{-1}(\mathbf{x}_{r,k})\mathbf{p}_{k-1}^{m,i}, \quad (20)$$

where function $f_{eul\,mat}(x_t, y_t, z_t, yaw_r, pitch_r, roll_r)$ transform translational XYZ and rotational yaw pitch roll angle into a homogeneous transformation matrix and $\mathbf{I}_3$ is a 3x3 identity matrix.

After a sound source is initialised, we use a chi-square test to validate each hypothesis at the time a new observation is available. The chi-square distance $d_k^{m,i}$ is formulated as follows,

$$P_{z_k^{m,i}} = \mathbf{H}_k^{m,i}\mathbf{P}_{ss,k-1}^{m,i}(\mathbf{H}_k^{m,i})^T, \quad (21)$$

$$d_k^{m,i} = (f_{na}(\hat{z}_k^{m,i} - z_k^{m,i}))^T P_{z_k^{m,i}} f_{na}(\hat{z}_k^{m,i} - z_k^{m,i}). \quad (22)$$

We invalidate a hypothesis when the mean value of the chi-square distance $d_k^{m,i}$ is larger than a predefined value. This hypothesis pruning process will continue until all remaining hypotheses (usually one or two) converge.

The convergence of a hypothesis is determined by the linearity index $Ld_k^{m,i}$ of the inverse depth of the hypothesis according to [17] as follows,

$$\mathbf{h}_{XYZ,k}^{W,m,i} = \mathbf{x}_{ss,axis}^{m}(1:3) - \mathbf{p}_k^{m,i}, \quad (23)$$

$$\sigma_{\rho,k}^{m,i} = \sqrt{\mathbf{P}_{ss,k}^{m,i}(3,3)}, \quad (24)$$

$$\mathbf{m}_k^{m,i} = \frac{\mathbf{p}_k^{m,i} - \mathbf{x}_{ss,axis}^{m}(1:3)}{||\mathbf{p}_k^{m,i} - \mathbf{x}_{ss,axis}^{m}(1:3)||}, \quad (25)$$

$$\sigma_{d,k}^{m,i} = \frac{\sigma_{\rho,k}^{m,i}}{\mathbf{s}_k^{m,i}(3)}, \quad (26)$$

$$Ld_k^{m,i} = \frac{4 * \sigma_{d,k}^{m,i}||(\mathbf{m}_k^{m,i})^T\mathbf{h}_{XYZ,k}^{W,m,i}||\mathbf{h}_{XYZ,k}^{W,m,i}||^{-1}||}{||\mathbf{h}_{XYZ,k}^{W,m,i}||}, \quad (27)$$

where $\mathbf{p}_k^{m,i}$ can be computed in the same way as $\mathbf{p}_{k-1}^{m,i}$ in Eq. 19. When the linearity index $Ld_k^{m,i}$ is small enough, convergence of the hypothesis is determined. When all remaining valid hypotheses converge, we take the mean value of sound source states in Euclidean coordinates of all valid hypotheses, and the mean value will be fed as the initial guess for sound sources in the joint optimisation process detailed in the next Section.

IV. Joint Optimisation of Sensor Poses, Visual Landmarks and Sound Sources Locations

A graph based SLAM [19] is used for optimisation to estimate jointly sensor poses, landmarks and sound sources. Note we will particularise this algorithm for an online implementation using key frames and visual landmarks, but any offline and other landmark-type can be utilised in a similar way.

Let $\mathbf{x}$ be the state vector of the graph SLAM,

$$\mathbf{x} = [\mathbf{x}^1_{kf}, \cdots, \mathbf{x}^{N_{kf}}_{kf}, \mathbf{v}^1, \cdots, \mathbf{v}^{N_v}, \mathbf{p}^1, \cdots, \mathbf{p}^{N_s}]^T, \quad (28)$$

where $\mathbf{x}^{n_{kf}}_{kf}(n_{kf} = 1 \cdots N_{kf})$ is the pose of the n_{kf}th key frame, $\mathbf{v}^{n_v}(n_v = 1 \cdots N_v)$ is the location of the n_vth visual landmark parametrised as Euclidean point and $\mathbf{p}^m(m = 1 \cdots N_s)$ is the location of the mth sound source. In the optimisation, since the sound source state is converged after the multi hypotheses initialisation, it is also parametrised by an Euclidean point. Any state of a key frame pose, a visual landmark or a sound source location is represented as a node and the measurement of a visual landmark or a sound source from a key frame pose, which is a constraint between two nodes, is represented by an edge in the graph SLAM.

In the least squares problem of the graph-based SLAM, the estimated state vector is found by minimising the error over all pose-pose constraints and pose-landmarks constraints [19],

$$\hat{\mathbf{x}} = argmin \sum_{ij} \mathbf{e}_{ij}^T \mathbf{\Omega}_{ij} \mathbf{e}_{ij}, \quad (29)$$

where $\mathbf{e}_{ij}$ denotes the error in the constraint between ith and jth nodes, and $\mathbf{\Omega}_{ij}$ is the associated information matrix.

When an edge represents an observation of a sound source of node j from a key frame of node i, the e_{ij} can be computed as follows,

$$\mathbf{p}^{j,i} = [\mathbf{I}_3 \mathbf{0}] M^{-1}(\mathbf{x}^i_{kf}) \begin{bmatrix} \mathbf{p}^j \\ 1 \end{bmatrix}, \quad (30)$$

$$e_{ij} = atan2(\mathbf{p}^{j,i}(2), \sqrt{\mathbf{p}^{j,i}(1)^2 + \mathbf{p}^{j,i}(3)^2}) - \hat{\beta}^{j,i}_{gp*} \quad (31)$$

where $\mathbf{p}^{j,i}$ is the local coordinate of the jth sound source in the ith key frame's reference frame and $\hat{\beta}^{j,i}_{gp*}$ is the observation of the sound source j from key frame i, which is the predicted DOA angle from GP sensor model. The associated information matrix is

$$\Omega_{ij} = (P^{\beta,j,i}_{gp*})^{-1}. \quad (32)$$

The observations of visual landmarks from the key frame poses depend on the nature of the sensor (monocular, stereo or RGBD) and details regarding them can be found in [16]. After all nodes and edges are defined, Eq. 29 can be solved by Gauss-Newton or Levenberg-Marquardt optimisation.

Regarding the real time implementation, following ORB-SLAM implementation [16], only the last key frames, either a fixed number or the co-visible key frames of the current key frame, and their related visual landmarks and sound sources are optimised. A full optimisation is performed only when a loop closure is detected. Any intermediate frame, which is not a key frame, is disregarded due to the real-time constraint. ORB features are used for visual landmarks and parallel tracking, optimisation and loop closure detection is performed as done in [16].

There are two limitations in the proposed method. Firstly, all sound sources are assumed to be static to be jointly optimised with other landmarks and poses. Note that if the sound sources are moving, once the hypotheses have converged to one, they could be tracked independently outside the joint optimisation. Secondly, the sensor is required to observe sound sources from different sensor poses. This is to compensate the partial angle observation of a linear microphone array. Without observing from several different poses, sound sources are not guaranteed to converge.

V. Simulation and Experimental Results

In this section, comprehensive simulation and experimental results are presented to evaluate the performance of the proposed method.

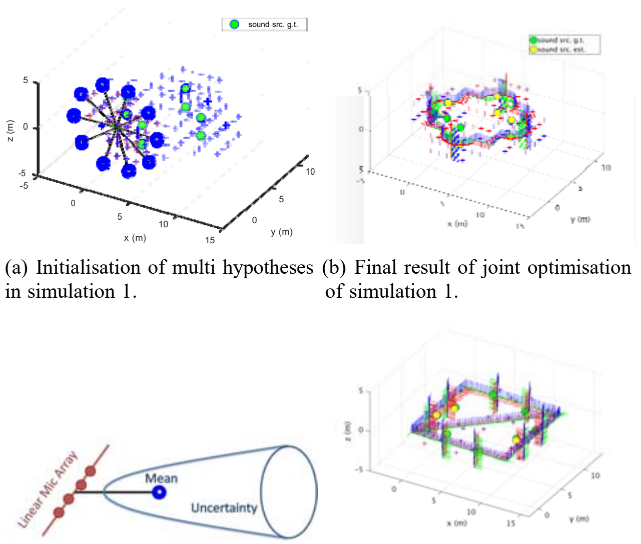
A. Simulation Results

In the simulation scenario shown in Fig. 5, a sensor with a RGBD camera and a linear microphone array for sound sources mapping is simulated. The sensor follows a 3D trajectory as shown in Fig. 5 (b). It starts from the origin and travels along positive X axis direction. After 2m, it follows an 1/4 arc. Then it moves vertically up and down, followed by another 1/4 arc returning to the positive X axis and travels along it for another 2m. This pattern of moment is repeated 4 times until it goes back to the origin. There are 8 sound sources in the simulation. The simulation parameters can be found in Table I. The sound bearing observation noise at different DOA angle was obtain empirically, and it is added to the ground truth value to be treated as noisy observation. As can be seen from figure (a), when the sensor first observes a sound source, it initialises 10 hypotheses along its instantaneously unobservable circumferential angle. The covariance value associated to each hypothesis is shown in (c). As the sensor keeps observing sound sources from different angles, most of the hypotheses are invalidated and only one of them will converge. From the time of convergence, the converged sound source is added to the joint optimisation process, where last 5 poses of the sensor, their associated visual feature points and sound sources are optimised. During the joint optimisation process, the error of the sound source location estimation continuously decreases. When a loop closure is encountered, the full graph is optimised. The final result is shown in Fig. 5 (b). We can see that all sound sources are converged to their ground truth locations. The RMS error of sound sources locations w.r.t. the absolute positions is 0.1302m. This result is quite reasonable, given the lack of DOF in observation and the large sound source observation noise.

In the second simulation scenario, we validate the system performance when sound sources are mostly observed by the least sensitive region of a linear microphone array, which is at the two sides of the linear microphone array of DOA

Parameters in simulation

Parameters	Values
Number of initial hypotheses	10
Sound bearing estimation noise	10-20 deg
Mic. array max sensing distance	3m
Distance per odometry step	0.2m
Least square optimiser	Levenberg-Marquardt
RGBD landmark observation noise	1 deg and 0.01m



(a) Initialisation of multi hypotheses in simulation 1.

(b) Final result of joint optimisation of simulation 1.

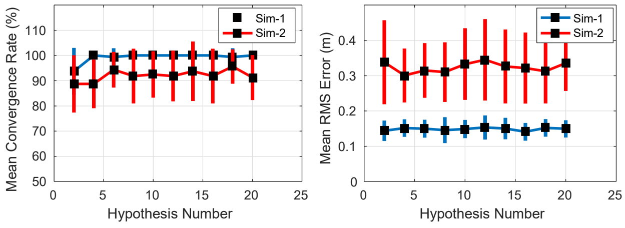
(c) Uncertainty value associated to each hypothesis during initialisation.

(d) Final result of joint optimisation of simulation 2.

Fig. 5. Initialisation of multi hypotheses and final result of joint optimisation. In all figures, red and blue (+) markers represent estimation and ground truth of RGBD landmarks. Green, red and blue unit lines denote the X,Y,Z axis of sensor local coordinate frame. In figure (a), blue circle markers represent initial multi hypotheses of sound sources.

angle of ±90 degrees. Locations of sound sources and the sensor trajectory is shown in Fig. 5(d). It can be seen from the figure that, most of the time, sound sources are around 90 degree DOA angle, which is the least sensitive region for a linear array. Despite the noisy observation around 90 degree DOA angle, sound sources are converged in the end with mean RMS error of 0.2688m. The error, as expected, is larger than the previous one, in which sound sources mostly are observed by highly sensitive region around 0 degree.

In the third set of simulation, we test the influence of the number of hypotheses over the final convergence of sound sources. 20 Monte Carlo runs of the first and second set of simulations are performed under various number of hypotheses. Mean convergence rate of the multi hypotheses filters, which is determined by the linearity index $L\alpha_{k}^{m,i}$ in Eq. 27, and RMS error of converged sound sources are shown in Fig. 6. From the figure, it can be seen that the number of hypotheses mainly affect the mean convergence rate and 6 or more number of hypotheses are suggested for better convergence. Regarding both convergence rate and sound sources mapping accuracy, in terms of RMS error, the first set of simulation is always better than the second set due to its observation of sound sources mostly in the sensitive region of the linear microphone array and from wide parallax angle.



(a) Mean convergence rate.

(b) Mean RMS error.

Fig. 6. Mean convergence rate and mean RMS error over 20 Monte Carlo runs for simulation 1 (better case) and 2 (worse case) when various number of hypotheses are used for initialisation.

B. *Experimental Results*

In this section, experimental results of sound sources mapping using Kinect 360 and PS3 Eye are presented as examples of monocular and RGBD vision sensors respectively.

Two experiments are conducted in a small office room and a computer lab. In the small office setup, mapping of two sound sources using both Kinect RGBD sensor and PS3 Eye Monocular camera, both with a linear microphone array inside, are performed. In the computer lab setup, mapping of five sound sources using the Kinect RGBD sensor is performed. Before performing the experiment, a set of sound source DOA estimation results using the SRP-PHAT algorithm and ground truth DOA angles are collected using both sensors in order to build the sensor model using GP as explained in Section II. Sound sources are emitted from a phone and a loud speaker for mapping two sound sources and fives phones for mapping fives sound sources. These devices are playing either a music or a continuous human speech. The sampling frequency of the microphone array is at 16KHz. Sound sources bearing estimation is performed at 5Hz. The sensors are handheld following a random trajectory around the sound sources. In Fig. 7, Fig. 8 and Fig. 8, yellow cubes represent estimated positions of sound sources and red hollow rectangles represent the manually measured ground truth positions of sound sources from the dense (using Kinect RGBD sensor) or sparse (using the PS3 Eye Camera) map.

The results show the proposed method performs well in small and larger areas. Covariances of sound sources are not shown in the figures for clarity, but they are consistent with the estimation errors.

VI. Conclusion

In this paper, we present a method for real-time 3D sound sources mapping using an off-the-shelf robotic perception sensor equipped with a linear microphone array. In the proposed method, multi hypotheses filters are combined with a new sound sources parametrisation to provide good initial guesses of sound sources locations for an online optimisation

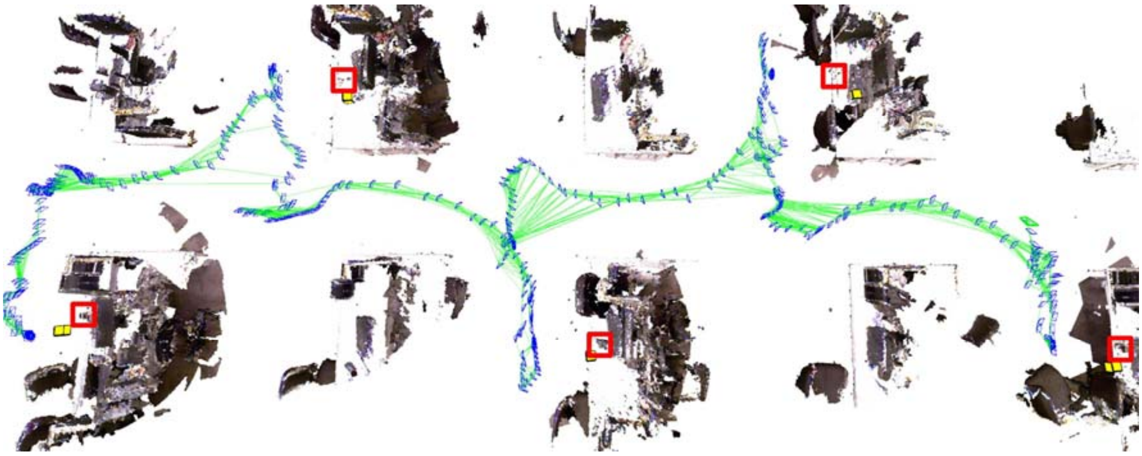
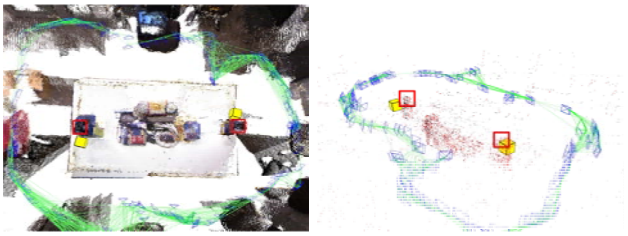


Fig. 8. Sound sources mapping result in a computer lab.



(a) Top view using Kinect.

(b) Side view using PS3 Eye.

Fig. 7. Mapping of two sound sources using Kinect (RGBD sensor) and PS3 Eye (monocular camera) with a embedded linear array inside.

strategy. A joint optimisation is carried out to estimate 6 DOF sensors poses and 3 DOF visual landmarks and sound sources locations. In addition, a dedicated sensor model for a linear microphone array is proposed to model accurately the noise of the DOA observation. Future work include robust sound sources data association and optimal active path planning to achieve better sound sources mapping performance.

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.-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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] J. Kotus, K. Lopatka, and A. Czyzewski, "Detection and localization of selected acoustic events in acoustic field for smart surveillance applications," *Multimedia Tools and Applications*, vol. 68(1), pp. 5–21, 2014.

[10] C. T. Ishi, J. Even, and N. Hagita, "Using multiple microphone arrays and reflections for 3D localization of sound sources," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013)*, 2013, pp. 3937–3942.

[11] L. A. Seewald, L. Gonzaga, M. R. Veronez, V. P. Minotto, and C. R. Jung, "Combining SRP-PHAT and two Kinects for 3D Sound Source Localization," *Expert Systems with Applications*, vol. 41(16), pp. 7106–7113, 2014.

[12] D. Su, T. Vidal-Calleja, and J. Valls Miro, "Split Conditional Independent Mapping for Sound Source Localisation with Inverse-Depth Parametrisation," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2016)*, 2016.

[13] S. Argentieri and P. Danes, "Broadband variations of the MUSIC high-resolution method for sound source localization in robotics," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2007)*, 2007, pp. 2009–2014.

[14] H. Do and H. F. Silverman, "SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data," Ph.D. dissertation, 2010.

[15] C. E. Rasmussen, *Gaussian processes for machine learning*, 2006.

[16] 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.

[17] 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.

[18] 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.

[19] G. Grisetti, R. Kummerle, C. Stachniss, and W. Burgard, "A tutorial on graph-based SLAM," *IEEE Intelligent Transportation Systems Magazine*, vol. 2(4), pp. 31–43, 2010.