UNIVERSITY OF TECHNOLOGY, SYDNEY

Robotic Sound Source Mapping using Microphone Arrays

by

Daobilige Su

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

at the
Centre for Autonomous Systems
Faculty of Engineering and Information Technology
University of Technology, Sydney

August 2017

Declaration of Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signed:

Date:

Abstract

The auditory system constitutes a significant perceptual input for humans and animals. While it is legitimate to say that it ranks behind other senses such as vision or haptics whose understanding has experienced significant advances in the context of computational intelligence and robotics, it is intuitive to assume that service and field robotic systems working closely with humans would benefit from incorporating compelling sound analysis capabilities in the pursuit of accomplishing human-robot collaborative tasks. Within the broad area of robotic audition, one of the most relevant research topics has been identifying and locating multiple sound sources that may be present in the vicinity of the robot at an instant in time. Robotic systems equipped with such ability would gain the faculty to better monitor acoustic events such as a conversation, a ringing alarm or a call for help, for example in a search and rescue scenario, effectively responding to people's needs in a more natural way. Mapping stationary sound sources using a robot equipped with an on-board microphone array is thus the main focus of this thesis.

The first important problem faced when mapping sound sources is the calibration of the auditory sensing unit, which in the scope of robot audition is almost invariably a multichannel microphone array. There are two distinctive cases depending on whether the microphone array is hardware-synchronised or not. If it is, calibration reduces to attaining an accurate estimate of the array geometry of all microphones, whereas for asynchronous arrays a resolution for starting time offsets and clock differences (drift rates) between the various microphones is also required. A novel methodology is hereby proposed using a graph-based Gauss-Newton least square optimisation technique borrowed from the simultaneous localisation and mapping (SLAM) literature. The proposed method starts investigating the calibration problem of a 2D/3D microphone array, and extends the method to the more challenging linear microphone array case.

Having attained a calibrated microphone array, two distinctive contributions are made within the context of a SLAM-based framework to jointly estimate robot poses, positions of

surrounding sound sources and other likely exteroceptive landmarks (e.g. visual features) in 2D/3D scenarios. Solving the SLAM problem purely based on sparse sound observations is quite difficult and often impossible when the number of sound sources is low. The key singularity is whether sound source mapping is carried out with a 2D/3D microphone array, or a linear array. The proposed method invariably adopts a least square optimisation in the form of graph SLAM to jointly optimise the state. This represents an improvement over the conventional work found in the literature in that trajectory estimation and sound source mapping are regarded as uncorrelated, i.e. an update on the robot trajectory does not propagate to the mapping of the sound sources.

While the proposed method is readily able to solve the 2D/3D sound source mapping problem itself, for the case of 2D/3D microphone array geometries, an additional improvement in efficiency is suggested by exploiting the conditional independence property between two maps estimated by two different SLAM algorithms running in parallel. In adopting this approach, the first map has the flexibility that can be built with any SLAM algorithm (filtering or optimisation) of choice to estimate robot poses with an exteroceptive sensor. The second map can then be estimated by using a filtering-based SLAM algorithm with all the stationary sound sources parametrised with Inverse Depth Parametrisation (IDP). Compared to the joint optimisation approach, the improved method is able to save computational cost as the filtering technique is used for the sound source map. Robot locations used during IDP initialisation become the common features shared between the two SLAM maps, which allow to propagate information accordingly. The improved method achieves similar accuracy in mapping sound source when compared to the full joint optimisation approach, while incurring less computational expense and adding significant flexibility in building the localisation map.

The proposed method of mapping sound sources using a 2D/3D microphone array cannot be readily applied to linear microphone arrays given the peculiarity of their sensor observation model, a considerable challenge when initialising a sound source: a linear microphone array can only provide 1 Degree Of Freedom (DOF) observations. Hence, multi-hypotheses tracking combined with a novel sound source parametrisation is proposed in this work to suggest a fitting initial guess for the sound source. Subsequently, a similar graph-based SLAM joint optimisation strategy as that employed for the 2D/3D case can be carried out

to estimate the full 6 DoF robot/sensor poses, 3 DOF landmarks (e.g. visual) and the location of the sound sources. Additionally, a dedicated sensor model is also proposed to more accurately model the noise embedded in the Direction of Arrival (DOA) observation for the specific case of using a linear microphone array. Ultimately, the proposed method provides a generic approach for mapping sound sources in 3D using a linear microphone array with the aid of additional exteroceptive sensing to overcome the prevailing sparsity of sound observations.

Acknowledgements

I would like to take the opportunity to express my gratitude to all the people who have offered me help and support during the past three and a half years of my candidature.

First of all, I would like to give my sincere thanks to my supervisors A/Prof. Jaime Valls Miro and Dr. Teresa Vidal Calleja for their guidance, inspiration, numerous hours of discussion and the opportunities they have presented to me. From them I learned not only the scientific knowledge and technical skills but also the research methodology, which will benefit me throughout my career. The doors to their offices are always open whenever I run into a trouble spot or have a question about my research or writing. They consistently steer me in the right the direction whenever I need it.

I would also like to thank Dr. Keisuke Nakamura and Prof. Kazuhiro Nakadai for their inspiration, continuous support and direct contributions to my research during and after my internship in Honda Research Institute, Japan (HRI-JP). Also I thank my colleges and friends Dr. Randy Gomez, Dr. Anupam Gupta, Mr. Borui Shen and Mr. Severin Bahman for making my internship experience in Japan fruitful and enjoyable.

I am also grateful to A/Prof. Yong Liu, Mr. Jinhong Xu and Ms. Mengmeng Wang for their guidance and help during the work on our collaborative project in both the institute of Cyber-Systems and Control (CSC), Zhejiang University (ZJU) and the Centre for Autonomous Systems (CAS), University of Technology, Sydney (UTS).

I also want to thank A/Prof. Shoudong Huang for organizing weekly SLAM meeting. It really helps me to understand the current state-of-the-art techniques related to SLAM, which is very important in my own research.

Many thanks go to Dr. Leo Shi, Dr. Nalika Ulapane, Dr. Buddhi Wijerathna, Mr. Raphael Falque, Mr. Freek De Bruijin, Mrs. Liye Sun, Mr. Maani Ghaffari jadidi, Mr. Kasra Khosoussi, Mr. Kanzhi Wu, Mr. Teng Zhang and many other colleagues from CAS, UTS. I thank my fellow labmates for the stimulating discussions and for all the fun we have had in the last four years, as well as their hands-on help on setting up the experimental environment to collect and process data; Many thanks to Ms. Camie Jiang and Ms. Katherine Waldron for their help on administrative work. Additional thanks go to my parents and friends, who are always proud of me for my achievements, and supported and cared for me throughout these years. Special thanks to my girlfriend Ms. Rina Dao for her peaceful love and continuous support.

Finally, I extend my gratitude to UTS, Faculty of Engineering and Information Technology (FEIT) and Sydney Water (SW) for the exemption of my tuition fee and assisting with my general living costs through the IRS scholarship and FEIT faculty scholarship; and to the CAS and the FEIT for sponsoring me in attending international conferences.

Contents

List of Figures

List of Tables

Acronyms & Abbreviations

1D, **2D**, and **3D**	1 Dimensional, 2 Dimensional, and 3 Dimensional
ADC	Analogue-to-Digital Converter
ASR	Automatic Speech Recognition
CAS	Centre for Autonomous Systems
CI	Conditional Independent
DSBF	Delay and Sum Beam Forming
DOA	Direction of Arrival
DOF	Degrees of Freedom
FBS	Frequency Band Selection
GCC-PHAT	Generalised Cross-Correlation Phase Transform
GP	Gaussian Process
HRI	Human robot interactions
IDP	Inverse Depth Parametrisation
ML	Maximum-Likelihood
MUSIC	MUltiple SIgnal Classification
PHAT	Phase Transform
RANSAC	Random Sample Consensus
RMS	root mean square
SLAM	Simultaneous Localisation and Mapping
SRP	Steered Response Power
SRP-PHAT	Steered Response Power with Phase Transform
STD	Standard Deviation
TDOA	Time Difference of Arrival

TOF	Time of flights
UKF	Unscented Kalman Filter
USAR	Urban Search and Rescue
USB	Universal Serial Bus
UTS	University of Technology, Sydney

Nomenclature

General Notations

$\overline{bel}(\mathbf{x}_t)$	The belief of the current state vector of the EKF SLAM system after prediction step.
$bel(\mathbf{x}_t)$	The belief of the current state vector of the EKF SLAM system after update step.
c_s	The speed of sound.
d_n^{mic}	The distance between the nth microphone to the origin of the microphone array coordinate.
$d_{i,k}$	The distance between the ith microphone and the sound source at time instance k.
d_k	The distance from the sound source position at the kth time instance to the origin of the global coordinate frame.
$d_k^{m,i}$	The chi-square distance of the ith hypothesis of the mth sound source for a linear array.
$e_{k-1,k}^{p-p}$	The error related to the position-position constraint between the sound source at time instance $k-1$ and k.
e_k^{p-l}	The error related to TDOA observation of microphone array, represented as the position-landmark constraint in graph based optimisation, when the sound source is at time instance k.
$\mathbf{e}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{z}_{ij})$	The error between the node i and the node j in graph based SLAM, which represents a difference between the expected observation $\hat{\mathbf{z}}_{ij}$ and the real observation $\mathbf{z}_{ij}$ gathered by the robot.
f	The signal frequency.

f_{an}	nth landmark observed by an additional sensor.
f_{sn}	nth sound landmark observed by the microphone array.
$\mathbf{F}(\mathbf{x})$	The negative log likelihood of all the observations in the graph SLAM least square optimisation.
$g(u_t, \mathbf{x}_{t-1})$	The robot motion model of EKF SLAM system.
G_t	The Jacobian of robot motion model $g(u_t, \mathbf{x}_{t-1})$.
$h(\bar{\mathbf{x}}_t)$	Observation function of the SLAM system.
$h^s(\mathbf{x}_t^s)$	Observation function in the EKF SLAM based sound map.
H_t	The Jacobian of observation function $h(\bar{\mathbf{x}}_t)$.
H_t^s	The Jacobian of $h^s(\mathbf{x}_t^s)$.
$\mathbf{H}$	The information matrix of the system in the graph SLAM.
$I_{k-1,k}^{p-p}$	The information matrix corresponds to the $\mathbf{z}_{k-1,k}^{p-p}$.
I_k^{p-l}	The information matrix corresponds to the $\mathbf{z}_k^{p-l}$.
K_t	Kalman gain in the EKF SLAM system.
K_t^s	Kalman gain in the EKF SLAM based sound map.
$K(\bullet)$	The pre-defined Kernel function of Gaussian Process.
$Ld_k^{m,i}$	The linearity index of the ith hypothesis of the mth sound source for a linear array.
m	the map of the environment.
$\mathcal{M}()$	The function computing the homogeneous transformation matrix of the a pose.
$\mathbf{n}^{mic}$	Zero-mean Gaussian noise added to each channel of the microphone array with covariance ${\sigma^{mic}}^2\mathbf{I}$.
$p(x_t^R, m\|z_{1:t}, u_{1:t})$	The posterior probability over the robot momentary pose along with the map.
$\mathbf{p}_k$	The position of the sound source at time instance k.
$\mathbf{p}^m$	The Euclidean coordinates of the mth sound source.
$\mathbf{p}_{l,k}^{m,i}$	The Euclidean coordinate of the mth sound source in ith hypothesis under sensor local coordinate at time instance k.
$\mathbf{p}^{j,i}$	The local coordinate of the jth sound source in the ith key frame's reference frame.

$\mathbf{p}_k^m$	The local coordinate of the sound source $\mathbf{p}^m$ in the reference coordinate frame of the sensor/robot pose $\mathbf{x}_{r,k}$.
$P_{MUSIC}(\phi)$	Pseudo-spectrum of MUSIC algorithm corresponds to angle ϕ.
$P(x_{srp})$	SPR-PHAT power of a given candidate point x_{srp}.
$P'(x_{srp})$	Simplified computation of the SPR-PHAT power of a given candidate point x_{srp}.
P_t	The covariance matrix corresponding to the system state vector of EKF SLAM system at time instance t.
$\hat{P}_t$	The covariance matrix corresponding to the system state vector of EKF SLAM system after the prediction step at time instance t.
P_t^s	Covariance matrix at time instance t in the EKF SLAM based sound map.
P_{C_s}	Covariance matrix related to $\mathbf{x}_{C_s}$.
P_S	Covariance matrix related to $\mathbf{x}_S$.
P_{CS}	Cross correlation terms of $\mathbf{x}_{C_s}$ and $\mathbf{x}_S$.
P_{SC}	Cross correlation terms of $\mathbf{x}_S$ and $\mathbf{x}_{C_s}$.
P_{C_a}	Covariance matrix related to $\mathbf{x}_{C_a}$.
P_A	Covariance matrix related to $\mathbf{x}_A$.
P_{CA}	Cross correlation terms of $\mathbf{x}_{C_a}$ and $\mathbf{x}_A$.
P_{AC}	Cross correlation terms of $\mathbf{x}_A$ and $\mathbf{x}_{C_a}$.
$\check{P}^a$	Covariance matrix of the rearranged state vector of the sound map.
$\check{P}^a$	Covariance matrix of the rearranged state vector of the localisation map.
P_S^b	Covariance matrix of state vector of the sound map after back propagation process.
$\mathbf{P}_{ss}^{m,i}$	The covariance matrix associated to $\mathbf{s}^{m,i}$.
$\mathbf{q}_m^{mic}$	One of the eigenvectors of $\mathbf{R}_s^{mic}$ corresponding to the zero eigenvalue.
Q_t	The observation noise variance of the SLAM system.
Q_t^s	Noise level of observation in the EKF SLAM based sound map.
$\mathbf{Q}_n^{mic}$	Matrix of eigenvectors $\mathbf{q}_m^{mic}$ corresponds to the noise. The noise subspace of $\mathbf{Q}^{mic}$.

$\mathbf{Q}_s^{mic}$	Matrix of eigenvectors $\mathbf{q}_m^{mic}$ corresponds to the signal. The signal subspace of $\mathbf{Q}^{mic}$.
$\mathbf{Q}^{mic}$	Matrix of eigenvectors of correlation matrix $\mathbf{R}^{mic}$.
R_t	The robot motion noise variance.
$\mathbf{R}^{mic}$	The correlation matrix of $\mathbf{x}^{mic}$.
$\mathbf{R}_s^{mic}$	Signal covariance matrix of $\mathbf{x}^{mic}$.
$\mathbf{s}^{mic}(\phi_m)$	The steering vector of the signal and ϕ_m is its direction.
$\mathbf{s}^m$	The proposed novel parametrisation of the mth sound source state for a linear array.
$\mathbf{s}^{m,i}$	The state of the mth sound source in ith hypothesis with the proposed novel parametrisation for a linear array.
$\mathbf{S}^{mic}$	Matrix form of steering vectors of the microphone array audio signal $\mathbf{x}^{mic}$.
u_t	Control input to the robot at time instance t.
$\mathbf{v}^{n_v}$	The location of the n_vth visual landmark.
x_{srp}	Candidature point or direction for SRP-PHAT value computation.
x_{srp_s}	The location estimation for a single sound source using SRP-PHAT algorithm.
$\mathbf{x}^{mic}$	The raw received signal of mixture of M sources at each channel of the microphone array.
$\mathbf{x}_{r,t}$	The sensor/robot pose at time t.
$\mathbf{x}_t$	The state vector of the SLAM system at time instance t.
$\bar{\mathbf{x}}_t$	The system state vector of EKF SLAM system after the prediction step at time instance t.
$\mathbf{x}^*$	The best configuration of the nodes, the state vector, in the graph SLAM system.
$\mathbf{x}_{mic}$	The state of the microphone array.
$\mathbf{x}_{mic\_n}$	The state of the nth microphone.
$\mathbf{x}_{lm}^s(i)$	The state of the ith sound source using IDP parametrisation.
$\mathbf{x}^s$	State vector of the sound map.
$\mathbf{x}_r$	State of the current robot pose.

$\mathbf{x}_t^s$	State vector of sound map at time instance t.
$\mathbf{x}_r^s(n_s)$	The past robot pose used to initialise the n_sth sound source.
$\mathbf{x}_{lm}^a(n)$	State of the nth landmark observed by the additional sensor.
$\mathbf{x}_{C_a}$	Part of state vector in localisation map that are shared by both localisation and sound maps.
$\mathbf{x}_A$	Part of state vector in localisation map that are conditionally independent from the sound source map.
$\mathbf{x}_{C_s}$	Part of state vector in sound map that are shared by both localisation and sound maps.
$\mathbf{x}_S$	Part of state vector in sound map that are conditionally independent from the localisation source map.
$\check{\mathbf{x}}^a$	Rearranged state vector of the sound map.
$\check{\mathbf{x}}^a$	Rearranged state vector of the localisation map.
$\mathbf{x}_S^b$	State vector of the sound map after back propagation process.
$\mathbf{x}_{r,k}$	The sensor/robot pose at time instance k.
$\mathbf{x}_{ss,axis}^m$	The anchor axis of the mth sound source location. The state representing the position and direction of the Y axis of the sensor local coordinate.
$\mathbf{x}_{kf}^{n_{kf}}$	The pose of the n_{kf}th key frame.
$\mathbf{x}$	The full state vector of the SLAM system.
$X_k(\omega)$	Audio signal at channel k in frequency domain.
$\bar{X}_l(\omega)$	The complex conjugate of the audio signal at channel l in frequency domain.
z_t	The robot measurement of the environment at time instance t.
z_{an}	Observation of nth landmark using an additional sensor.
z_{sn}	Observation of nth sound landmark by the microphone array.
z_t^s	The observed sound source bearing in the EKF SLAM based sound map.
$\mathbf{z}_{ij}$	The mean of a virtual measurement between the node i and the node j in graph based SLAM.

$\hat{\mathbf{z}}_{ij}(\mathbf{x}_i, \mathbf{x}_j)$	The prediction of a virtual measurement between the node i and the node j in graph based SLAM.
$\mathbf{z}_{k-1,k}^{p-p}$	The observation of the position-position constraint between the sound source at time instance $k-1$ and k.
$\hat{\mathbf{z}}_{k-1,k}^{p-p}$	The observation of the position-position constraint between the sound source at time instance $k-1$ and k.
$\mathbf{z}_{k}^{p-l}$	The TDOA observation of microphone array, represented as the position-landmark constraint in graph based optimisation, when the sound source is at time instance k.
$\hat{\mathbf{z}}_{k}^{p-l}$	The expected TDOA observation of microphone array, represented as the position-landmark constraint in graph based optimisation, when the sound source is at time instance k.
α^m	Complementary angle of β^m, the DOA angle of the mth sound source.
β_k	The DOA angle of the sound source at time instance k for the calibration of a linear array.
β^m	DOA angle of the mth sound source.
$\hat{\beta}_k^{m,DOA}$	The estimated sound DOA angle of the mth sound source from a DOA estimation algorithm.
$\hat{\beta}_{gp*}^{DOA}$	A new test date from the DOA estimation algorithm to the Gaussian Process sensor model.
β_{gp*}	The predicted DOA angle using the Gaussian Process sensor model corresponding to $\hat{\beta}_{gp*}^{DOA}$.
$\hat{\beta}_{gp*,ini}^{m}$	The predicted DOA angle from the Gaussian Process sensor model at the first observation of the mth sound source.
$\hat{\beta}_{gp*}^{j,i}$	The observation of the sound source j from key frame i, which is the predicted DOA angle from the Gaussian Process sensor model for a linear array.
$\hat{\boldsymbol{\beta}}_{gp}^{DOA}$	The set of raw results from the DOA estimation algorithm that are used as function input when training the Gaussian Process sensor model.

$\boldsymbol{\beta}_{gp}$	The set of ground truth DOA angles that are used as function output when training the Gaussian Process sensor model.
γ^m	The circumferential angle of the mth sound source with the proposed novel parametrisation for a linear array.
ρ^m	The inverse depth of the mth sound source with the proposed novel parametrisation for a linear array.
λ_m	The corresponding eigenvalue for $\mathbf{q}_m^{mic}$.
ϕ	The angle between the direction being searched and the direction from the origin of the microphone array to the nth microphone.
ω	The angular frequency.
$\boldsymbol{\Omega}_{ij}$	The information matrix of a virtual measurement between the node i and the node j in graph based SLAM.
τ_{lk}	The TDOA from point x_{srp} to lth channel of the microphone array and kth channel of the microphone array.
τ_k	TOF from point x_{srp} to the kth channel of the microphone array.

Chapter 1

Introduction

With the advance of technologies in mechanics, electronics, control and information technology, robots have started to come out of industrial workshops and operate in almost all fields of service. Application areas include service robots such as Asimo [13], Pepper [14] and PR2 [15], search and rescue like Packbots [16], entertainment like NAO robot [17], surveillance such as all types of micro aerial vehicles (MAVs) [18] and many more. As a result, the perceptual capabilities of a robot have become an important aspect to accomplish the various tasks they have been designed for.

The auditory system is an important perception system for human and animals together with visual, tactile and odour sensing systems. With binaural audition, people can localise sound sources, focus their attention on one or many of them, recognise the speaker and understand who he/she is. It appears intuitive that robots, especially those aimed at accomplishing Human Robot Interaction (HRI) tasks, would benefit from incorporating auditory abilities too. Service robots equipped with some form of sound sensing would be able to interact with people, understand their needs and monitor acoustic events in their surroundings, such as a ringing alarm, a conversation or a call attention, to assist the everyday lives of people in a more natural way. Search and rescue robots with auditory capabilities can help localise victims by sounds they may emit, in particular where visual search might not work due to, for example, dust and smoke on the site, or occlusions introduced by obstacles.

FIGURE 1.1: Asimo robot localises, separates and recognises simultaneous speech signals from three persons.

The research topic of developing a robotic audition system is thus introduced. It is an emerging research field, which incorporates the research field of audio signal processing, artificial intelligence and robotics. Nowadays, there exist many successful robot audition system implementations [19–21], with which robots are able to localise and track multiple sound sources [22–41], separate speech singals coming from several people simultaneously [42–45], recognise human speech [46–49] and automatically recognise speakers [50–53]. An illustrative example is shown in Fig. 1.1, where the robot Asimo can localise and separate speech from three persons talking simultaneously and recognise each separated speech.

There are four broad elements in a robot auditory system, namely sound source localisation, speech separation, speech recognition and speaker recognition, which are detailed as follows:

- **Sound Source Localisation**

  One of the most fundamental processes for robot audition is uncovering the location of the sound source. Moreover, localisation is an important first step as its outcomes

FIGURE 1.2: NAO robot focuses its attention on the person who is speaking.

induce post-signal processes such as sound source separation and speech recognition [54]. Localisation of sound sources can for instance aid a service robot to be able to control its attention to focus on events and changes surrounding itself. As a result, the robot can choose to focus on events such as a phone ringing, a vehicle honking, a person talking. An example is shown in Fig. 1.2 where a NAO robot determines the location of the person speaking to engage in meaningful interaction. Hearing complements well other sensors such as vision because audio sensors are omni-directional, capable of working in the dark, dust, smoke, fog and not limited by physical structure occlusion (such as walls) [55]. As shown in Fig. 1.3, a robot can localise persons by their sound in an area full of smoke, whereas other modalities of sensing such as cameras operating in the visible range would not be able to do so. Another common application of sound source localisation is the robotic tracking of one or more sound emitting targets [56] and the speaker tracking [20] scenarios, in which location or direction of dynamic sound sources can be to estimate the location or direction of the sound emitting targets. Other popular robotic applications of sound source localisation include search and rescue scenario [57] and relative positioning for multiple robots [5].

- **Speech Separation**

  Separation of multiple speech signals is another essential skill for robot auditory perception as it plays a preprocessing step for automatic speech recognition. In many practical scenarios, people often talk at the same time, a situation often referred as "double-talk" in the literature [58]. In HRI, users might speak at the same time while the robot interacts with them through speech. In literature, this situation is

FIGURE 1.3: Robot operating in an area full of smoke.

called "barge-in" [58]. Unfortunately, current speech recognition technology usually assumes a single sound source is present. Therefore, the performance of the current speech recognition technology can be severely degraded in the real-world environment [59], which could include double-talk or barge-in. A robust robot audition in HRI should be capable of handling double-talk and barge-in situations. In order to obtain a double-talk and barge-in free robot audition system, the robot needs to know the original signals, in either digital or analogue format. This introduces the need for separation of multiple speech signals that are spoken simultaneously. The speech separation technique is used to recover every separated signal that represents each original sound source as much as possible, after a robot hears multiple speech signals distorted by spatial transfer functions of microphones including the influence of reflection and echoes by its ears (one or more microphone arrays). Once the system knows original speech signals, it can deal with the problems of speech recognition and thus robot audition performance is expected to improve [58]. An example of separating four simultaneous speech signals coming from three persons and one loud speaker is shown in Fig 1.4. The robot is able to recover its original speech signal of each source due to the speech signals separation function in its audition system. This

FIGURE 1.4: Separation of four simultaneous speech.

separation process is very important since only a near-clean recovered speech signal is more likely to be correctly understood by the robot using the speech recognition technique detailed next.

- **Speech Recognition**

  Robots with HRI ability are increasingly expected to possess perceptual capabilities similar to humans, due to increasing demands for symbiosis of humans and robots. Particularly, hearing capabilities are essential for social interaction since spoken communication is very important for normal hearing people [59]. Therefore, automatic speech recognition (ASR) has been an active research area for over five decades. ASR has always been considered as an important bridge in fostering better human-human and human-machine communication [60]. With ASR, a robot could communicate with users in a natural fashion through speech. This is especially important for robots that are designed to care for elderly people for instance, who are not expected to have special technical skills to interact with robots. An example is shown in Fig. 1.5, where a Pepper robot is talking to its customers to get their orders in a restaurant.

- **Speaker Recognition**

  The last important skill for a robot audition system is the ability to automatically recognise each speaker. Automatic speaker recognition, based on their vocal characteristics implies identifying who is speaking to a machine among a number of persons. This identification process can be done with a closed set or an open set of persons (identifying a known or an unknown speaker, an impostor), and can be text-dependent or independent [61]. Due to the numerous fields of applications it covers,

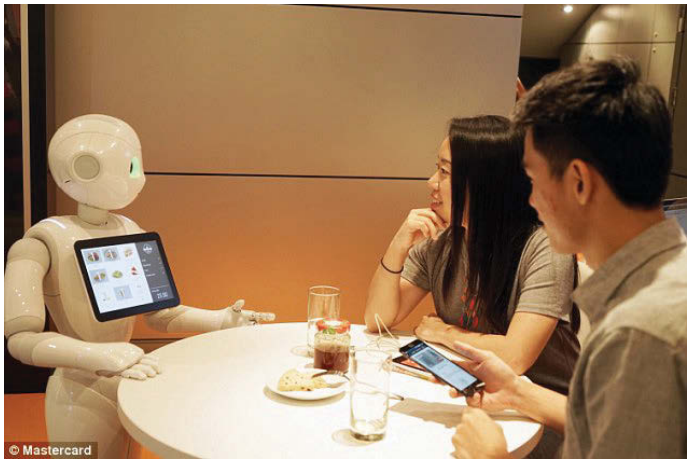


FIGURE 1.5: Pepper robot communication with customers with its ASR module.

research interest in automatic speaker recognition is actually growing. For instance, it can be used for audio surveillance, in which a robot for surveillance purpose can determine if a person is a verified valid person in a particular area or not.

1.1 Motivation and Scope

Among these diverse applications of robot audition systems, localisation and mapping of one or multiple sound sources by a mobile robot which is equipped with one or various microphones is of increasing interest. In Urban Search and Rescue (USAR) scenarios for instance, location information of sound sources in a geometric map created from the inspection of the surrounding environment can be used to locate missing people in disaster sites. In HRI scenarios, localisation results of sound sources can be used to detect and track speakers [62] or discern between multiple people's speech [63].

In order to tackle these issues, this thesis focuses on solving the research question of how to accurately and probabilistically map all sound sources within a geometric map of the environment using a mobile robot. This is a topic that has received significant attention in recent years [2, 4, 10, 11] and the work hereby presented delves further into the examination of mapping stationary sound sources using different configurations of microphone arrays.

The structure of this thesis and its key contributions are pictorially represented in Fig. 1.6. As can be seen from the figure, an audio signal is sampled by a microphone array, be it

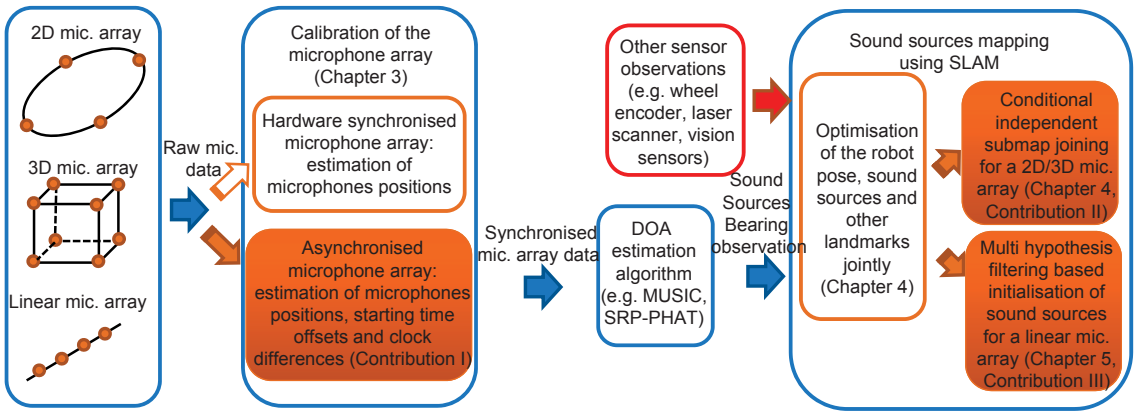


FIGURE 1.6: The structure of the thesis. Solid orange color blocks represent key contributions of the thesis, and orange edge color blocks represent other minor contributions.

from a 2D/3D or linear array. The audio signal needs to be sampled synchronously so that conventional DOA estimation algorithms can be applied to find the bearing information to the sound source. This synchronisation is typically done in hardware by multi-channel Analogue-to-Digital Converter (ADC) boards. However, even when the microphone clocks are synchronised in the hardware, there is a need to estimate the actual geometry of the microphone array for bearing estimation. Therefore, the first important problem of mapping sound sources is the calibration of the synchronous/asynchronous microphone array. The calibration of an asynchronous microphone array estimates positions, starting time offsets and clock differences of microphones, while as indicated above only estimation of the positions of the microphones is required when hardware synchronisation is available. After the calibration, conventional DOA estimation algorithms can be applied directly.

After we obtain bearing information, a SLAM framework to map the sound source is proposed. In order to achieve accurate robot self localisation and sound source mapping, we assume that the robot is also equipped with other sensors such as wheel encoders, a laser scanner or vision sensors, a typical scenario in modern mobile robots. With sound sources bearing estimation from the microphone array audio signal and landmark observation from other sensors, the SLAM framework can estimate robot poses, sound source locations and landmarks locations from other sensors. While a traditional joint optimisation SLAM framework can solve the 2D/3D sound source mapping problem, we propose key improvements:

- For mapping sound source in 2D/3D scenario, we introduced a conditionally independent submap method in order to make the SLAM framework less computationally complex and more flexible so that it can be used together with existing SLAM implementation.

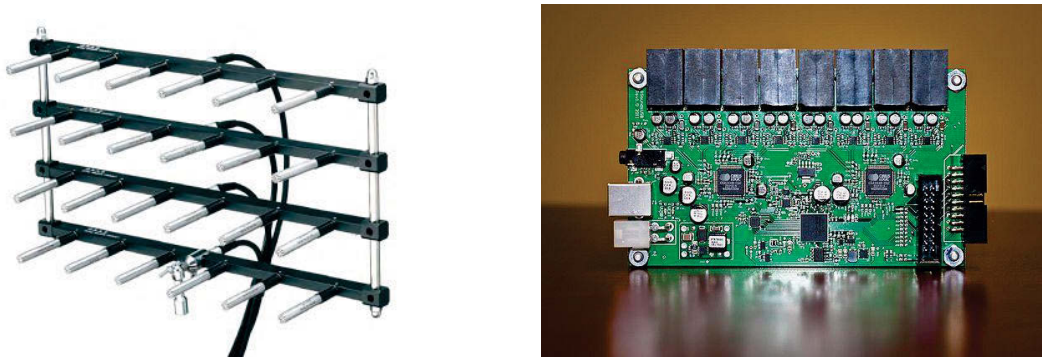
- For the case of linear microphone arrays, the joint optimisation framework can not be applied directly since the initialisation of the sound source is not straightforward, as is the case for 2D and 3D microphone geometries. Therefore, a multi-hypothesis filtering based initialisation strategy is proposed, together with a linear microphone array sensor model based on Gaussian Process.

In summary, the two main motivations for this thesis work are the need for calibration of microphone arrays and sound source mapping using microphone arrays, and are detailed below:

1.1.1 Calibration of a Hardware-Synchronised/Asynchronous Microphone Array

In many robot audition systems [19–21], a microphone array is needed to estimate the bearing information of sound sources. A synchronised microphone array consists of a multi-channel ADC converter and multiple microphones, with each microphone connecting to each channel of the ADC converter. A typical example of a microphone array and ADC converter is shown as in Fig. 1.7. The audio data from all microphones is synchronously sampled by the ADC converter, which is then used to estimate the DOA of sound sources based on Time Difference of Arrival (TDOA) from the sound source to each microphone. Calibration of such a microphone array estimates geometric locations of all microphones which is needed for DOA estimations of the sound source and sometimes can not be easily manually measured, especially when embedded into a robotic system.

Recently, methods [64] [65] [66] [67] have been developed to relax the hardware synchronisation and successfully estimate bearing information of sound sources using an asynchronous microphone array. These methods are capable of computing microphone locations and



a b

FIGURE 1.7: A typical microphone array structure and multi-channel ADC board for the sound source bearing estimation.

starting time-offset between different microphone channels to estimate the bearing information of the sound sources, all of them are based on an assumption that the clock interval for each independent sound card, dedicated to each channel, is identical to that of other channels. This is a strong assumption that disregards errors from fractional differences in clock intervals, which will accumulate over time. A mechanism is proposed in this work to overcome this limitation by calibrating the asynchronous microphone array using graph-based pose SLAM, which estimates the geometric position, time offset and clock difference of each microphone simultaneously using probabilistic Gauss-Newton least square optimisation, thereby making this method suitable for generic sound source localisation from a heterogeneous asynchronous microphone array.

Recent development in MEMS microphone arrays have reduced the size of multichannel ADC board to a certain degree, posing a major advantage for hardware synchronised microphone arrays. However, the need to find the exact locations of the microphones persists even in those case, as that is paramount in the calculation of the TDOA (see more on this below). It is important to note that the process of calibration of an asynchronous microphone provides not only time offsets and clock differences between each pair of microphones (which would indeed not be the case for synchronous systems), but also the exact locations of microphones, which is needed in a synchronised microphone array setting too. The alternative is using asynchronous microphone arrays which are usually significantly more affordable than fully synchronised MEMS microphone arrays, although they can still provide the same functionality if properly calibrated. More relevantly to the area of study

developed in this thesis, synchronisation of an asynchronous microphone array remains a necessary task when it comes to robot audition in general.

In robotic system, given the constraints of embedding microphones somewhere in the robot frame, often alongside other perception devices, it is difficult to measure the exact 3D locations of the microphones accurately. Various alternatives exists for that. Measuring the exact location is an obvious one, although not trivial and prone to error. As an alternative, a transfer function between each microphone and a given sound source can be measured. These measurements, however, can be quite time-consuming since they need to be obtained at multiple directional intervals of the given sound source. For instance, recording the transfer function every 5 degree means 72 measurements when only azimuth angle is sought after, and 2664 measurements when both azimuth and elevation angles are needed. With a conservative assumption of say 1 minute for each transfer function recording, in practice the process is not only fiddly but also time consuming. However, the proposed calibration method only needs to record approx. 1 minute of a chirp signal to accurately obtain the locations of a set of microphones, thereby dramatically reducing the time and effort required, and the accuracy obtained. These factors are also behind the reason why recent works in robot audition literature keep focusing on improving the microphone array location estimation accuracy.

As indicated before, one calibrated, conventional DOA estimation algorithm can be applied directly to obtain sound sources bearing information without using a hardware synchronisation board. We analyse the calibration problem of an 2D/3D and linear asynchronous microphone array separately, due to their different observation models.

1.1.2 Sound Source Mapping by a Robot Embedded Microphone Array

Conventional methods for sound source localisation and mapping rely on a microphone array and either, a proprioceptive sensor (such as wheel odometry), or an additional exteroceptive sensor (such as cameras or lasers) to get the robot locations accurately. Since odometry drifts over time and sound observations are bearing-only, sparse and 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 conventional methods using an additional exteroceptive sensor, 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. Therefore, we propose a least square optimisation based SLAM framework which estimates robot poses and positions of sound source and other landmarks (e.g. visual point features) jointly. Under the proposed framework, the robot can accurately localise itself and its trajectory estimation and sound source mapping are fully correlated.

As previously noted, while the proposed method is able to solve the 2D/3D sound source mapping problem, a further improved method is also presented. In addition, as the joint optimisation framework can not be applied to the linear microphone array case, a multi-hypothesis filtering based sound source initialisation method is proposed. These two methods are detailed as follows:

- The improved method for 2D/3D sound source mapping: In this thesis, an efficient method to correlate robot trajectory with sound source mapping by exploiting the conditional independence property between two maps estimated by two different SLAM algorithms running in parallel is studied in Chapter 4. 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 IDP. Robot locations used during IDP initialisation are the common features shared between the two SLAM maps, which allow the propagation of information accordingly. Since the second map uses a filtering technique, the proposed method has less computational complexity compared to the joint optimisation framework.

- In the literature of sound source mapping in 3D space, conventional approaches use dedicated 3D microphone arrays to map sound sources in 3D scenarios as this type of arrays provide two DOF observations. However, many popular robotic perception devices such as Microsoft Kinect 360, Kinect One, PS3 Eye and PS4 Eye sensors, are equipped with a linear microphone array. Despite easy availability at affordable

price and frequent usage of these sensors, conventional 3D sound source mapping methods in the literature hardly make use of a linear microphone array since a linear microphone array can only provide 1 DOF observation in 3D space. The joint optimisation framework can not be applied to the linear microphone array case directly, since the initialisation of sound sources can not be done as it is done for the 2D/3D microphone array case. Thus a novel method for real-time 3D sound source mapping using an off-shelf robotic perception sensor equipped with a linear microphone array is proposed in Chapter 5. In this framework, multi hypotheses tracking is combined with a new sound source parametrisation to provide a good initial guess for an online optimisation strategy. A joint optimisation is subsequently carried out to estimate 6 DOF sensor poses and 3 DOF landmarks and sound source positions. Additionally, a dedicated sensor model with Gaussian Processes is proposed to model accurately the noise of the DOA observation when using a linear microphone array.

1.1.3 Scope

While there exists a large amount of research related to robot audition, this thesis focuses on a specific area of it, which is sound source mapping with a mobile robot equipped with a microphone array. In particular, we confine the research work in this thesis as follows:

- Sound sources bearing estimation in this thesis is restricted to utilising a robot embedded microphone array. Here, the term robot embedded microphone array means the size of the microphone array is relatively small compared to the distance from any sound source to the center of the microphone array (for instance the microphone array in a Microsoft Kinect sensor). Under this definition, all sound sources are located in the far field of the microphone array, which means the microphone array, based on the TDOA of sound source to each channel of the microphone array, can estimate the bearing (also commonly referred as DOA) but not the 2D/3D coordinates of the sound source. We start investigating the 2D/3D microphone array based robot audition system, then focus on a linear microphone array based system. Note that, in this thesis, hardware synchronisation is not a requirement to the microphone array. When using an asynchronous microphone array, the software synchronisation

method proposed in this work can be used to synchronise acoustic signals sensed by the microphones in the array. This is one of the main contribution of this thesis as detailed in Chapter 3. The near field scenario is only considered when calibrating a 2D/3D microphone array.

- The thesis focuses on sound source mapping using one single robot. Cooperative sound source mapping using multiple robots is beyond the scope of this thesis work. With one single mobile robot equipped with a robot embedded microphone array, at each time instance, since the robot can only obtain a bearing estimation of the sound sources, it is not enough to estimate full 2D/3D coordinates of the sound sources. Therefore, the robot needs to obtain multiple bearing estimation of sound source from multiple locations so that full 2D/3D sound source coordinates can be estimated using probabilistic SLAM frame work. This is substantially different from multi robot cooperative sound source mapping since with multiple bearing information of sound source from the location of multiple robots, full 2D/3D coordinates of sound source can be estimated.

- In order to obtain an accurate sound source map, in this thesis, we assume that the robot is equipped with an additional exteroceptive accurate sensor that can be used to accurately localise itself using SLAM algorithm. Typically, we use a 2D laser scanner for localising the robot in 2D scenario and camera or RGBD sensor for 3D localisation of the robot. With the help of these additional perceptive capability, quite common in today's mobile platforms, the robot can accurately localise itself, draw the environment map and map sound sources on the environmental geometry map.

- In this thesis, the sound source data association problem, significant ego motion noise and an environment with dynamic objects are not considered. For sound source data association, this means the ID of each sound source can be obtained by the consistency of the sound source bearing estimation between likely candidates, and the chi square test. By using these two methods, most sound source data association problem can be solved, as is frequently the case within the SLAM community when it comes to feature associations. However, as only the bearing information can be

obtained for a sound source, this results in a more complex scenario, and a more advanced data association method - such as feature based sound sourced identification methods [61] - should be used. For robot ego motion noise, we assume that it does not affect the sound source bearing estimation considerably. This is mostly true for many sound source mapping scenarios, but in general an advanced ego motion noise cancellation method should be used for situations where ego motion noise cannot be ignored (e.g. the noise of a quadcopter). Finally, we assume that there are no dynamic objects in the environment, which means the negative influence of dynamic objects in localising the sensor when using visual or Lidar based SLAM algorithm can be neglected.

1.2 Contributions

As shown in Fig. 1.6, there are three main contributions in this thesis.

1. Calibration of an asynchronous microphone array: in order to use a microphone array without a hardware synchronisation device, i.e. to tackle the problem in section 1.1.1, a methodology is hereby proposed to calibrate an asynchronous microphone array using a graph-based optimisation method borrowed from the SLAM literature, effectively estimating the array geometry, time offset and clock difference/drift rate of each microphone together with the sound source locations. Simulation and experimental results are presented, which prove the effectiveness of the proposed methodology in achieving accurate estimates of the microphone array characteristics needed to be used on realistic settings with asynchronous sound devices. Once calibrated, the microphone array outputs synchronised recording, which can provide bearing information of sound sources by applying conventional DOA estimation algorithms.

2. 2D/3D sound source mapping by using conditionally independent submap joining: once sound source bearing information is obtained, be it from a synchronised microphone array or an asynchronous microphone array, a robot can map stationary sound sources on a geometric map. Specifically, we propose a least squares optimisation based SLAM framework to map stationary sound sources while simultaneously

localising a moving robot, and other landmarks positions (visual features in this case) jointly. The proposed method correlates robot trajectory with sound source mapping. While the joint optimisation framework is able to solve the 2D/3D sound source mapping problem, an improved method is presented. The proposed method is efficient in correlating robot trajectory with sound source mapping by exploiting the conditional independence property between two maps estimated by two different 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 (e.g. camera or laser) in the traditional sense of SLAM. The sound map can then be constructed in parallel using a filtering-based SLAM algorithm hence making the method computationally less expensive compared to a full joint optimisation framework. Comprehensive simulations and experimental results show the effectiveness of the proposed method.

3. Sound source mapping of a linear microphone array with multi-hypothesis initialisation: finally, we focus on the problem of sound source mapping using a linear microphone array. We present a method for robotic real-time 3D sound source mapping using an off-the-shelf linear microphone array sensor widely used in the robotics community such as the Microsoft Kinect. In the proposed method, multi hypotheses tracking is combined with a new sound source parametrisation to provide an initial guess for an online optimisation strategy. A joint optimisation is carried out to estimate 6 DOF sensors poses and 3 DOF landmarks and sound source locations. Additionally, a dedicated sensor model is proposed to model accurately the noise of the DOA observation when using a liner microphone array. Comprehensive simulation and experimental results show the effectiveness of the proposed method. Furthermore, a real-time implementation of the proposed method has been made available as open source software for the benefit of the community.

1.3 Publications

1.3.1 Directly Related Publications

1. **Daobilige Su**, Teresa Vidal Calleja and Jaime Valls Miro, "Towards Real-Time 3D Sound Sources Mapping with Linear Microphone Arrays", **accepted in** *2017 IEEE International Conference on Robotics and Automation (ICRA 2017)*, 2017.

2. Mengmeng Wang, **Daobilige Su**, Lei Shi, Yong Liu, and Jaime Valls Miro, "Real-time 3D Human Tracking for Mobile Robots with Multisensors", in *2017 IEEE International Conference on Robotics and Automation (ICRA 2017)*, 2017.

3. **Daobilige Su**, Teresa Vidal Calleja and Jaime 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)*, pp. 2000-2006, 2016.

4. **Daobilige Su**, Keisuke Nakamura, Kazuhiro Nakadai and Jaime Valls Miro, "Robust Sound Source Mapping using Three-layered Selective Audio Rays for Mobile Robots", in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2016)*, pp. 2771-2777, 2016.

5. **Daobilige Su**, Teresa Vidal Calleja and Jaime Valls Miro, "Simultaneous asynchronous microphone array calibration and sound source localisation", in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2015)*, pp. 5561-5567, 2015.

6. **Daobilige Su**, Jaime Valls Miro and Teresa Vidal Calleja, "Graph-SLAM Based Calibration of an Embedded Asynchronous Microphone Array for Outdoor Robotic Target Tracking", in *Assistive Robotics: Proceedings of the 18th International Conference on CLAWAR 2015*, pp. 641-648, 2015. (**Best paper award finalist**)

7. **Daobilige Su**, Jaime Valls Miro and Teresa Vidal Calleja, "Real-time sound source localisation for target tracking applications using an asynchronous microphone array", in *2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA 2015)*, pp. 1261-1266, 2015.

8. **Daobilige Su** and Jaime Valls Miro. "An ultrasonic/RF GP-based sensor model robotic solution for indoors/outdoors person tracking", in *2014 13th International Conference on Control Automation Robotics & Vision (ICARCV 2014)*, pp. 1662-1667, 2014.

1.4 Thesis Outline

The structure of the thesis is organised as follows:

In Chapter 2, literature review related to sound source bearing estimation, SLAM and sound source mapping is introduced. Elaboration of the details of typical algorithms of sound source bearing estimation relative to the robot coordinate frame are presented. The sound sources bearing estimation constitutes sound landmark observation for the robot, which can be exploited within the context of standard SLAM techniques to localise itself, as well as map the location of the sound sources. Therefore, some start-of-the-art algorithms related to filtering and optimisation based SLAM are described in section 2.2. Finally, an overview of some conventional methods proposed in the literature for sound source mapping using sound bearing estimation and SLAM techniques are presented.

In Chapter 3, we first present the proposed method for calibrating an asynchronous 2D/3D microphone array. Then, we focus on the key issue of the calibration of a linear microphone array. We elaborate the mathematical formulation of the proposed method. Comprehensive simulation and experimental results are presented, which prove the effectiveness of the proposed methodology in achieving accurate estimates of the microphone array characteristics needed to be used on realistic settings with asynchronous sound devices. The content of this Chapter formed the basis for publications [68, 69]. This part of the work presents the first contribution of this thesis shown orange in Fig. 1.6.

In Chapter 4, we first introduce the proposed method of mapping sound sources using a least square optimisation based SLAM framework. Then, we present our improved method of sound source mapping using a 2D/3D microphone array. Detailed formulation of conditionally independent localisation and sound maps joining, information backward propagation and inverse depth parametrisation of sound source are presented. This part

of the work presents the second contribution of this thesis shown orange in Fig. 1.6. The content of this Chapter is mainly compiled in publications [70].

In Chapter 5, we present sound source mapping using a commonly available linear microphone array. Detailed formulation of the sensor model using Gaussian Process (GP), multi hypotheses sound source initialisation and joint optimisation of sensor poses, visual landmarks and sound sources are presented. This part of the work presents the third contribution of this thesis shown orange in Fig. 1.6. The content of this Chapter is the basis for the publication [71].

Chapter 6 summarises the contributions of the thesis and provides an insight into the future directions of research.

Chapter 2

Review of Related Work

In this Chapter, some existing work related to sound sources bearing estimation with a microphone array, typical methods of solving SLAM problem and conventional approaches for sound sources mapping are presented.

2.1 Sound Source Bearing Estimation using A Microphone Array

2.1.1 Microphone Array

A microphone array is of key importance to a robot audition system. A microphone array is needed for most of the robot audition systems that deal with sound sources bearing estimation [72–80], speech separation [81–84], enhanced speech recognition [85–88] and enhanced speaker verification [50, 51]. In the following part of the section, typical structures of microphone arrays and their application domains are introduced.

Based on the structure of a microphone array, it can be classified into one of the following categories:

- **Linear microphone array**: A linear microphone array is defined as a microphone array in which all microphone channels lie on a straight line. Linear microphone

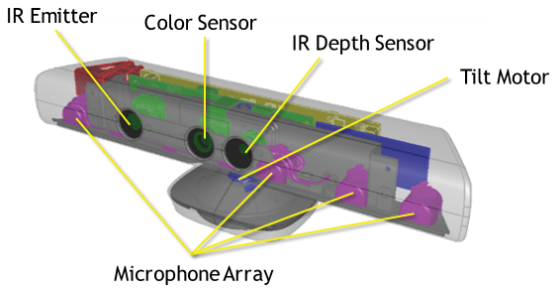


FIGURE 2.1: 4-channel microphone array on Kinect 360.

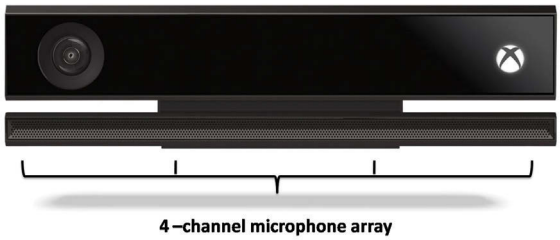


FIGURE 2.2: 4-channel microphone array on Kinect One.

FIGURE 2.3: 4-channel microphone array on PS3 eye.

arrays are the most common type of microphone arrays available. Typical examples of linear microphone arrays are the ones on Microsoft Kinect 360, Kinect One, PS3 eye sensor and PS4 eye sensor as shown in Fig. 2.1, Fig. 2.2, Fig. 2.3 and Fig. 2.4. Linear microphone arrays are mostly used for sound sources bearing estimation in 2D. However, it can not provide full bearing estimation, because of the front and back ambiguity of sound sources bearing estimation.

A special case of linear microphone array is binaural robot audition [89, 90], which is a linear microphone array using only two microphones inspired by human and animal auditory systems. Binaural robot audition is mostly used in humanoid robots. Despite using only two microphones, binaural robot audition has the comparable sound sources bearing estimation capability of any linear microphone array. However,

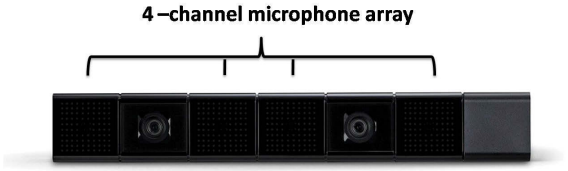


FIGURE 2.4: 4-channel microphone array on PS4 eye.

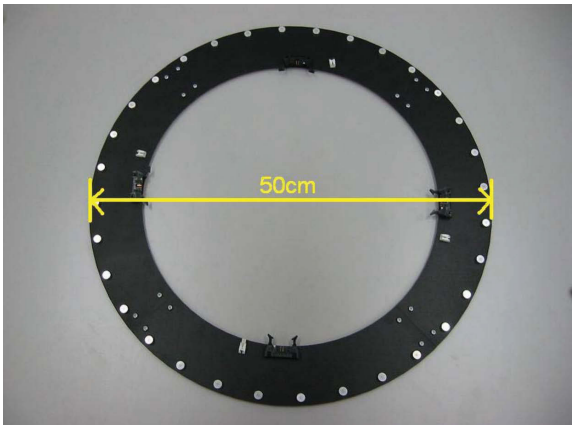


FIGURE 2.5: A circular microphone array [1].

less number of channels degrades its sound sources bearing estimation accuracy. Note that the binaural robot audition additionally involves some sort of structure that works similarly to the human pinna. Using this ear like structure, the sound source bearing estimation goes through a training process, where the robot learns how the ear structure affects sounds coming from different directions.

- **Planner microphone array**: A planner microphone array is defined as a microphone array in which all microphone channels stay on a plane, but can not be connected by one single line. Planner microphone arrays are mostly designed for 2D bearing estimation and acoustic camera application. Typical examples are the circular microphone array design by Tamai [1], the concentric microphone array designed by Sasaki et. al. [2] and the microphone array designed by Perrodin et. al. [3] as shown in Fig. 2.5, Fig. 2.6 and Fig. 2.7.

  When a planner microphone is used for sound sources bearing estimation in 2D, it provides a unique value of sound sources estimated bearing without ambiguity. When it is used for estimating sound sources bearing in 3D, it can also estimate

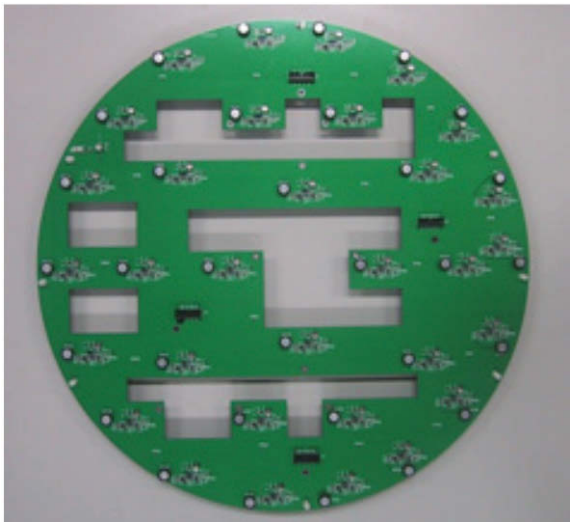


FIGURE 2.6: A concentric microphone array in [2].

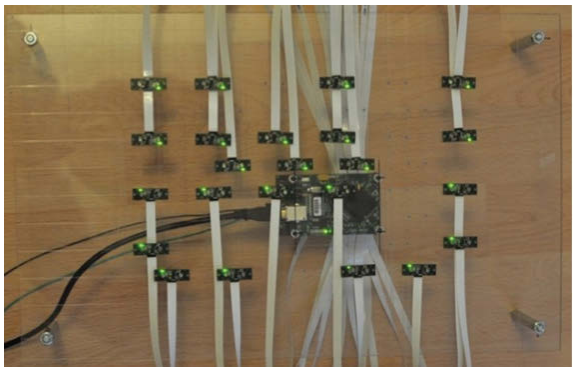


FIGURE 2.7: A large scale 2D microphone array [3].

both azimuth and elevation angle but with ambiguity in the elevation angle. This is due to the fact that two sound sources having the same azimuth angle but the opposite value of elevation angle have the same TDOA to the microphone array, which means after obtaining the TDOA, the sound source can be either on the top part of the space on the plane or on the bottom part.

- **3D microphone array**: A 3D microphone array is defined as a microphone array in which all microphone channels cannot stay on a plane. A 3D microphone array can estimate the azimuth and elevation angle of sound sources without ambiguity. Typical examples are the microphone array used in [4] and in [5] as shown in Fig 2.8 and Fig 2.9.

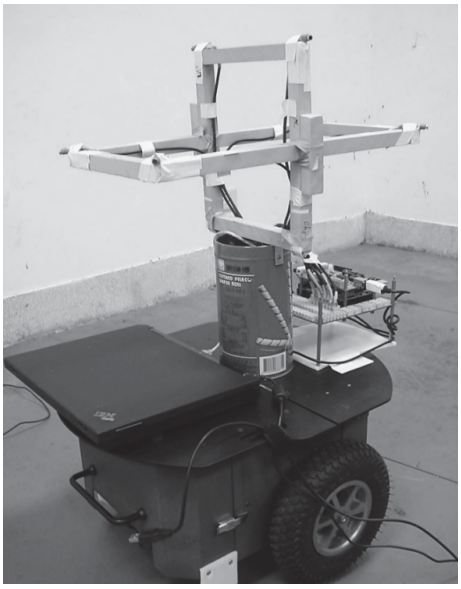


FIGURE 2.8: A 3D microphone array in [4].

FIGURE 2.9: A 3D microphone array in [5].

2.1.2 Sound Sources Bearing Estimation

Bearing estimation (a.k.a. DOA estimation) of sound sources is a natural area of research for array signal processing, which has had a lot of interest over recent decades [91]. As sound source DOA estimation is essential for many diverse robot audition applications, typical methods of DOA estimation are presented here.

In the early years of research in the field of DOA estimation, research literature was mainly focused on scenarios where there was only one single audio source active. The TDOA at different microphone pairs is used by most of the proposed methods. Among them, the Generalised Cross-Correlation PHAse Transform (GCC-PHAT) is the most popular method [33]. Improvements to the TDOA estimation problem were proposed in [92]. In their work, both the multipath and the so-far unexploited information among multiple

microphone pairs were taken into account. An overview of TDOA estimation techniques can be found in [93].

Localising multiple, simultaneously active sources is a more difficult and challenging problem. In such a situation, even the smallest overlap of sources which is caused by a brief interjection as an example, can disrupt the localisation of the original source [33]. A system that is designed to handle the localisation of multiple sources treats the interjection as another source that can be simultaneously captured or rejected as desired. One of the first methods capable of estimating DOAs of multiple sources is the well-known MUltiple SIgnal Classification (MUSIC) algorithm [94] and its wideband variations [54, 95–97]. MUSIC algorithm is one of the classic family of subspace approaches. It depends on the eigen-decomposition of the covariance matrix of the observation vectors. The Steered Response Power with Phase Transform (SRP-PHAT) [98] is another efficient sound localisation algorithm that is suitable for mobile robot applications [99]. Both of MUSIC and SRP-PHAT algorithms (or their modified versions) are frequently used in the robot audition literature for bearing estimation of sound sources [2, 10, 11, 100]. Therefore, a detailed description and mathematical formulation of these two algorithms are provided below. In addition, we also introduce another DOA estimation algorithm for an uniformly distributed linear microphone array: ESPRIT, which stands for Estimation of Signal Parameters via Rotational Invariance Technique. ESPRIT has the advantage of not involving an exhaustive search through all possible steering vectors to estimate DOA. However, it is mainly designed for an uniformly distributed linear microphone array to estimate the sound sources azimuth angles.

2.1.2.1 MUSIC

MUSIC, similar to many adaptive techniques, is based on the correlation matrix of the multi channel audio data. Let us consider a situation of estimation the DOAs of M sound sources using a N-channel microphone array. In this case, the sensor model can be formulated as follows [101],

$$\mathbf{x}^{mic} = \sum_{m=1}^{M} \alpha_m^{mic} \mathbf{s}^{mic}(\phi_m) + \mathbf{n}^{mic}, \tag{2.1}$$

where $m = 1 \cdots M$ denotes the index of sound sources, $\mathbf{x}^{mic}$ is the raw received signal of mixture of M sources at each channel of the microphone array, α_m is the amplitude coefficient of each sound source, $\mathbf{s}^{mic}(\phi_m)$ denotes the steering vector of the signal and its direction ϕ_m is what we are aiming to estimate. $\mathbf{n}^{mic}$ is zero-mean Gaussian noise added to each channel of the microphone array with covariance ${\sigma^{mic}}^2\mathbf{I}$.

Then, we can write Eq. 2.1 in a matrix form as follows [101],

$$\mathbf{x}^{mic} = \mathbf{S}^{mic}\alpha^{mic} + \mathbf{n}^{mic}, \tag{2.2}$$

$$\mathbf{S}^{mic} = [\mathbf{s}^{mic}(\phi_1), \cdots, \mathbf{s}^{mic}(\phi_M)], \tag{2.3}$$

$$\alpha^{mic} = [\alpha_1^{mic}, \cdots, \alpha_M^{mic}]^T, \tag{2.4}$$

where the matrix $\mathbf{S}^{mic}$ is of dimension NxM, which represents the M steering vectors. Here we are assuming that signals from different channels are uncorrelated. Therefore, the correlation matrix of $\mathbf{x}^{mic}$ can be written as follows [101],

$$\begin{aligned}
\mathbf{R}^{mic} &= E[\mathbf{x}^{mic}{\mathbf{x}^{mic}}^H], \\
&= E[\mathbf{S}^{mic}\alpha^{mic}{\alpha^{mic}}^H{\mathbf{S}^{mic}}^H] + E[\mathbf{n}^{mic}{\mathbf{n}^{mic}}^T], \\
&= \mathbf{S}^{mic}\mathbf{A}^{mic}{\mathbf{S}^{mic}}^H + {\sigma^{mic}}^2\mathbf{I}, \\
&= \mathbf{R}_s^{mic} + {\sigma^{mic}}^2\mathbf{I},
\end{aligned} \tag{2.5}$$

where

$$\mathbf{R}_s^{mic} = \mathbf{S}^{mic}\mathbf{A}^{mic}{\mathbf{S}^{mic}}^H, \tag{2.6}$$

$$\mathbf{A}^{mic} = \begin{bmatrix} E[|\alpha_1|^2] & 0 & \cdots & 0 \\ 0 & E[|\alpha_2|^2] & \cdots & 0 \\ 0 & 0 & \cdots & E[|\alpha_M|^2] \end{bmatrix}, \tag{2.7}$$

where $()^H$ stands for "Hermitian" operation, which is a combination of complex conjugate and matrix transpose ($A^H = \bar{A}^T$, where $\bar{A}$ is complex conjugate of A).

The rank of the signal covariance NxN matrix, $\mathbf{R}_s^{mic}$, is clearly M. Therefore, its number of eigenvectors corresponding to the zero eigenvalue is N-M. Let $\mathbf{q}_m^{mic}$ represents one of the eigenvectors corresponding to the zero eigenvalue, then it has the relationship formulated

as follows,

$$\begin{aligned} &\mathbf{R}_s^{mic}\mathbf{q}_m^{mic} = \mathbf{S}^{mic}\mathbf{A}^{mic}\mathbf{S}^{mic^H}\mathbf{q}_m^{mic} = 0, \\ &\implies \mathbf{q}_m^{mic^H}\mathbf{S}^{mic}\mathbf{A}^{mic}\mathbf{S}^{mic^H}\mathbf{q}_m^{mic} = 0, \\ &\implies \mathbf{S}^{mic^H}\mathbf{q}_m^{mic} = 0, \end{aligned} \tag{2.8}$$

where because the matrix $\mathbf{A}^{mic}$ is positive definite, the final equation is valid. From Eq. 2.8, we can conclude that all M signal steering vectors are orthogonal to all N-M eigenvectors ($\mathbf{q}_m^{mic}$) of $\mathbf{R}_s^{mic}$ corresponding to the zero eigenvalue. The above conclusion serves as the basis for MUSIC.

Let $\mathbf{Q}_n^{mic}$ denotes the Nx(N-M) matrix of those eigenvectors mentioned above. Then, pseudo-spectrum of MUSIC is as follows [101],

$$\begin{aligned} P_{MUSIC}(\phi) &= \frac{1}{\sum_{m=1}^{N-M}|\mathbf{s}^{mic^H}(\phi)\mathbf{q}_m^{mic}|^2} \\ &= \frac{1}{\mathbf{s}^{mic^H}(\phi)\mathbf{Q}_n^{mic}\mathbf{Q}_n^{mic^H}\mathbf{s}^{mic}(\phi)} \\ &= \frac{1}{||\mathbf{Q}_n^{mic^H}\mathbf{s}^{mic}(\phi)||^2}. \end{aligned} \tag{2.9}$$

From Eq. 2.9, it can be seen that the denominator becomes zero when ϕ is a signal direction, since the eigenvectors making up $\mathbf{Q}_n^{mic}$ are orthogonal to the signal steering vectors. As a result, the M largest peaks in the pseudo-spectrum corresponds to the estimated signal directions. Unfortunately, the signal covariance matrix $\mathbf{R}_s^{mic}$ would not be available in the practical situation. So, to be able to estimate $\mathbf{R}_s^{mic}$, the signal covariance matrix, is the best we can expect. The key here is that the eigenvectors of $\mathbf{R}^{mic}$ can be used to estimate the eigenvectors in $\mathbf{Q}_n^{mic}$.

For any eigenvector $\mathbf{q}_m^{mic} \in \mathbf{Q}^{mic}$, we can have equations as follows,

$$\begin{aligned} &\mathbf{R}_s^{mic}\mathbf{q}_m^{mic} = \lambda\mathbf{q}_m^{mic} \\ &\implies \mathbf{R}^{mic}\mathbf{q}_m^{mic} = \mathbf{R}_s^{mic}\mathbf{q}_m^{mic} + {\sigma^{mic}}^2\mathbf{I}\mathbf{q}_m^{mic} = (\lambda_m + {\sigma^{mic}}^2)\mathbf{q}_m^{mic}, \end{aligned} \tag{2.10}$$

where λ is the vector of eigenvalues and λ_m is the corresponding eigenvalue for $\mathbf{q}_m^{mic}$. From the Eq. 2.10, it can been seen that any eigenvector of $\mathbf{R}_s^{mic}$ is also an eigenvector of $\mathbf{R}^{mic}$

with corresponding eigenvalue $\lambda + {\sigma^{mic}}^2$. Let $\mathbf{R}_s^{mic}$ represent $\mathbf{Q}^{mic}\mathbf{\Lambda}\mathbf{Q}^{micH}$. Then, we can have the following equation [101],

$$\begin{aligned}\mathbf{R}^{mic} &= \mathbf{Q}^{mic}[\mathbf{\Lambda} + {\sigma^{mic}}^2\mathbf{I}]\mathbf{Q}^{micH} \\ &= \mathbf{Q}^{mic}\begin{bmatrix} \lambda_1 + {\sigma^{mic}}^2 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 + {\sigma^{mic}}^2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_M + {\sigma^{mic}}^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & {\sigma^{mic}}^2 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & {\sigma^{mic}}^2 \end{bmatrix}\mathbf{Q}^{micH}.\end{aligned} \tag{2.11}$$

Following Eq. 2.11, we can partition the eigenvector matrix $\mathbf{Q}^{mic}$ into a signal matrix $\mathbf{Q}_s^{mic}$ with M columns, corresponding to the M signal eigenvalues, and a matrix $\mathbf{Q}_n^{mic}$, with (N-M) columns, corresponding to the noise eigenvalues (${\sigma^{mic}}^2$) based on this eigen-decomposition [101]. Note that $\mathbf{Q}_n^{mic}$ is the Nx(N-M) matrix of eigenvectors corresponding to the noise eigenvalue (${\sigma^{mic}}^2$). $\mathbf{Q}_n^{mic}$ is also equal to the matrix of eigenvectors of $\mathbf{R}_s^{mic}$ corresponding to the zero-eigenvalue. This is the matrix used in Eq. 2.9. The signal subspace is defined by $\mathbf{Q}_s^{mic}$, while $\mathbf{Q}_n^{mic}$ defines the noise subspace.

At this point, there are a few important observations to be made [101]:

- The smallest eigenvalues of $\mathbf{R}^{mic}$ are all equal to ${\sigma^{mic}}^2$ and they are the noise eigenvalues. This character implies that, determining the number of small eigenvalues that are equal can be one way of distinguishing between the signal and noise eigenvalues, which is also equivalently the signal and noise subspaces.
- By orthogonality of $\mathbf{Q}^{mic}$, $\mathbf{Q}_s^{mic} \perp \mathbf{Q}_n^{mic}$.

Based on the above two observations, we can see that all noise eigenvectors are orthogonal to the signal steering vectors, which is the basis for MUSIC algorithm. We can construct

a function of ϕ as follows [101],

$$\begin{aligned} P_{MUSIC}(\phi) &= \frac{1}{\sum_{m=M+1}^{N} |\mathbf{q}_m^{mic\,H} \mathbf{s}^{mic}(\phi)|^2} \\ &= \frac{1}{\mathbf{s}^{mic\,H}(\phi) \mathbf{Q}_n^{mic} \mathbf{Q}_n^{mic\,H} \mathbf{s}^{mic}(\phi)}, \end{aligned} \tag{2.12}$$

where $\mathbf{q}_m^{mic}$ is one of the (N-M) noise eigenvectors. We can see that $\mathbf{s}^{mic}(\phi) \perp \mathbf{q}_m^{mic}$ and the denominator is identically zero if ϕ corresponds to the DOA of one of the signals. Therefore MUSIC algorithm identifies the peaks of the function $P_{MUSIC}(\phi)$ as the DOA of mixed signals.

The correlation matrix $\mathbf{R}^{mic}$ is unknown and must be estimated from the received data in the practical scenario. This estimation requires averaging over several snapshots of data, formulated as follows [101],

$$\mathbf{R}^{mic} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{x}_k^{mic} \mathbf{x}_k^{mic\,H}, \tag{2.13}$$

where $\mathbf{x}_k^{mic}$ is the kth snapshot of audio signal.

Therefore, in practice, the steps of DOA estimation using MUSIC algorithm are [101]:

1. Firstly, estimate the correlation matrix R using Eq. 2.13. Find its eigen-decomposition $\mathbf{R}^{mic} = \mathbf{Q}^{mic} \mathbf{\Lambda} \mathbf{Q}^{mic\,H}$.
2. Secondly, partition $\mathbf{Q}^{mic}$ to obtain $\mathbf{Q}_n^{mic}$, corresponding to the (N-M) smallest eigenvalues of $\mathbf{Q}^{mic}$, which spans the noise subspace.
3. Then, plot the MUSIC function $P_{MUSIC}(\phi)$ in Eq. 2.12 as a function of ϕ.
4. Finally, the M largest peaks of $P_{MUSIC}(\phi)$ are the M signal directions.

The steering vector $\mathbf{s}^{mic}(\phi)$ in Eq. 2.12 of a microphone array is defined as follows,

$$\mathbf{s}^{mic}(\phi) = [e^{-2j\frac{\pi f d_1^{mic} \cos(\bar{\phi})}{c_s}}, \cdots, e^{-2j\frac{\pi f d_N^{mic} \cos(\bar{\phi})}{c_s}}]^T, \tag{2.14}$$

where j is the symbol for the imaginary part, f is the signal frequency, $d_n^{mic}(n = 1 \cdots N)$ is the distance between the nth microphone to the origin of the microphone array coordinate,

c_s is the speed of sound and $\bar{\phi}$ is the angle between the direction being searched and the direction from the origin of the microphone array to the nth microphone. In 2D, $\bar{\phi}$ is the bearing angle to be estimated, and in 3D, $\bar{\phi}$ is a function of azimuth and elevation angle. In 3D, DOA estimation $\bar{\phi}$ in Eq. 2.14 can be formulated as follows,

$$\bar{\phi} = 2\sin^{-1}\left(\frac{\sqrt{(o_n^x - \cos(\phi_e)\cos(\phi_a))^2 + (o_n^y - \cos(\phi_e)\sin(\phi_a))^2 + (o_n^z - \sin(\phi_e)^2}}{2}\right), \quad (2.15)$$

where o_n^x,o_n^y,o_n^z are coordinates of the nth microphone location, and ϕ_a, ϕ_e are the candidate azimuth and elevation angles which we are going to search.

In practice, the steering vector $\mathbf{s}^{mic}(\phi)$ needs to be searched for every possible bearing angle in 2D or every possible combination of azimuth and elevation angle in 3D. Therefore, the problem with MUSIC is its expensive computational cost during the exhaustive search.

2.1.2.2 SRP-PHAT

The core idea of the SRP-PHAT DOA estimation algorithm is similar to the idea of the GCC-PHAT [102]. When computing the likelihood of all possible angles to be DOA angle, we compute their SRP-PHAT value after Phase Transform (PHAT) is applied to the Steered Response Power (SRP). The SRP-PHAT for each point x_{srp} in the space, that is a potential position of a sound source, is defined as follows [103],

$$P(x_{srp}) = \sum_{k=1}^{k=M}\sum_{l=1}^{l=M}\int_{-\infty}^{\infty}\frac{1}{|X_k(\omega)\bar{X}_l(\omega)|}X_k(\omega)\bar{X}_l(\omega)e^{j\omega\tau_{lk}}d\omega, \quad (2.16)$$

where l and k denotes the lth channel and kth channel of the microphone array, ω is the angular frequency, $X_k(\omega)$ is audio signal at channel k in frequency domain, $\bar{X}_l(\omega)$ is the complex conjugate of the audio signal at channel l in frequency domain and $\tau_{lk} = \tau_l - \tau_k$ is the TDOA from point x_{srp} to lth channel of the microphone array and kth channel of the microphone array (τ_l and τ_k are time of flights (TOF) from point x_{srp} to lth and kth channel of the microphone array).

The GCC between microphone k and microphone l and the GCC between microphone l and k is essentially the same. Therefore, the elements, which are summed to form the above SRP-PHAT functional, form a symmetric matrix with fixed energy terms on the

diagonal [103]. As a result, either the upper-part or lower-part of the matrix is the part of the SRP-PHAT that changes with point x_{srp}. This means, for a particular point x_{srp} in the space, the computation of the SRP-PHAT using Eq. 2.16, can be obtained by summing the GCC of only a subset Q^{srp} of the pairs, where $Q^{srp} = [k, l], \forall k \in [1, \ldots, M-1], M \leq l > k$, which can be formulated as follows [103],

$$P'(x_{srp}) = \sum_{k=1}^{k=M} \sum_{l=k+1}^{l=M} \int_{-\infty}^{\infty} \frac{1}{|X_k(\omega)\bar{X}_l(\omega)|} X_k(\omega)\bar{X}_l(\omega) e^{j\omega\tau_{lk}} d\omega, \tag{2.17}$$

In order to find the source locations, the beamformer is steered over all possible points in a focal volume containing the source. The points that give the maximum weighted output power (SRP-PHAT value) of the beamformer are determined as the locations of the sources. The location estimate point x_{srp_s} for a single source is [103],

$$x_{srp_s} = \arg\max_{x_{srp}} P'(x_{srp}), \tag{2.18}$$

where $P'(x_{srp})$ is the SRP-PHAT at point x_{srp}, which is defined in Eq. 2.17. The integration over the angular frequency is implemented as the summation of a range of frequency indexes in practice. The SRP-PHAT value of any particular point of $P'(x_{srp})$ (or 2D/3D bearing estimation if sound sources are in the far field of the microphone array) is called a functional evaluation.

The assumption of the algorithm is that the SRP-PHAT will peak at the actual source location even under very noisy and highly reverberant conditions. However, similar to MUSIC, the problem with SRP-PHAT is its expensive computational cost because the search space has many local maxima, and thus computationally intensive grid-search (in euler space if sound sources are in the near field of the microphone and in all possible bearings otherwise) methods have been required to find the global maximum [103].

2.1.2.3 ESPRIT

ESPRIT stands for Estimation of Signal Parameters via Rotational Invariance Techniques. ESPRIT is another subspace based DOA estimation algorithm which does not involve an exhaustive search through all possible steering vectors to estimate DOA and hence

dramatically reduces the computational and storage requirements compared to MUSIC and SRP-PHAT [104, 105]. As a downside, ESPRIT is mainly designed for the estimation of the azimuth angle using an uniformly distributed linear array.

The goal of the ESPRIT technique is to exploit the rotational invariance in the signal subspace which is created by two arrays with a translational invariance structure [104], i.e. its suits an uniformly distributed linear array. To formulate the sound source bearing estimation using the ESPRIT algorithm, let's assume there two sub arrays: array-1 and array-2. Array-1 consists of microphones 1 to $N-1$ and array-2 consists of microphone 2 to N, and therefore the two sub arrays are displaced by distance d, where d is the distance between two adjacent microphones. The signals induced on each of the arrays are therefore given by

$$\mathbf{x}_1(t) = \mathbf{A}_E\mathbf{s}(t) + \mathbf{n}_1(t), \tag{2.19}$$

and

$$\mathbf{x}_2(t) = \mathbf{A}_E\mathbf{\Lambda}_E\mathbf{s}(t) + \mathbf{n}_2(t), \tag{2.20}$$

where $\mathbf{\Lambda}_E = diag\{e^{-2jk_{wn}d\sin(\theta_1)} \cdots e^{-2jk_{wn}d\sin(\theta_M)}\}$ is a diagonal unitary matrix called the rotation operator with phase shifts between doublets for each DOA angle $e^{-2jk_{wn}d\sin(\theta m)}(m = 1 \cdots M)$, $k_{wn} = \dfrac{\omega}{c_s}$ is the wave number of the signal with angular frequency of ω and speed of sound c_s, $\mathbf{n}_1$ and $\mathbf{n}_2$ are noise terms to the two sub arrays, $\mathbf{A}_E$ is the M by N steering matrix of the two sub arrays [105].

By creating the signal subspace for the two sub arrays, we can obtain two matrices $\mathbf{V}_{E1}$ & $\mathbf{V}_{E2}$. Since the arrays are related in translation, the subspaces of eigenvectors are related by a unique non-singular transformation matrix $\mathbf{\Phi}_E$ such that [104]

$$\mathbf{V}_{E1}\mathbf{\Phi}_E = \mathbf{V}_{E2}. \tag{2.21}$$

In addition, there must also exist a transformation matrix $\mathbf{T}_E$ such that $\mathbf{V}_{E1} = \mathbf{A}_E\mathbf{T}_E$ and $\mathbf{V}_{E2} = \mathbf{A}_E\mathbf{\Lambda}_E\mathbf{T}_E$ [104]. Therefore, it can be obtained that

$$\mathbf{T}_E\mathbf{\Phi}_E\mathbf{T}_E^{-1} = \mathbf{\Lambda}_E. \tag{2.22}$$

Therefore, the eigenvalues of $\mathbf{\Phi}_E$ must be equal to the diagonal elements of $\mathbf{\Lambda}_E$ such that [104]

$$\lambda_1 = e^{-2jk_{wn}d\sin(\theta_1)} \cdots \lambda_M = e^{-2jk_{wn}d\sin(\theta_M)}. \tag{2.23}$$

Finally, when the eigenvalues of $\mathbf{\Phi}_E$, $\lambda_1 \cdots \lambda_M$ are calculated, the DOA angles can be formulated as

$$\theta_m = \sin^{-1}(\frac{arg(\lambda_m)}{k_{wn}d}). \tag{2.24}$$

From the above formulation, it can be seen that the ESPRIT eliminates the search procedure and produces the DOA estimation directly in terms of the eigenvalues without much computational and storage requirements [104].

2.2 Simultaneous Localisation and Mapping

SLAM problems come into play when the robot neither knows its own poses nor has access to a map of the environment. Instead, the robot has access to its controls input $u_{1:t}$ and measurements $z_{1:t}$ of the environment. The terminology "simultaneous localisation and mapping" describes the following problem. In SLAM, the robot needs a map of its environment. At the same time it needs to simultaneously localise itself relative to this map. Since the map is unknown and has to be estimated along the way, SLAM is more difficult than robot localisation given an environmental map. Moreover, since the poses are unknown and have to be estimated along the way, SLAM is also more difficult than mapping with known poses [106].

There are two main forms of the SLAM problem from a probabilistic perspective. Both of them are of equal practical importance. One of them is known as the filtering based SLAM problem. The filtering based SLAM involves estimating the posterior over the momentary pose along with the map [106], described as follows,

$$p(\mathbf{x}_{r,t}, m | z_{1:t}, u_{1:t}), \tag{2.25}$$

where m is the map, $\mathbf{x}_{r,t}$ is the pose at time t, and $z_{1:t}$ and $u_{1:t}$ are the measurements and controls, respectively. Since it only involves the estimation of variables that persist at time t, this problem is called the filtering based SLAM problem. Many filtering based SLAM algorithms are incremental. These algorithms discard past measurements and controls once they have been processed.

The second SLAM problem is called the optimisation based (a.k.a. full) SLAM problem. Instead of just the current pose $\mathbf{x}_{r,t}$ in the filtering based SLAM, a posterior over the entire path $\mathbf{x}_{r,1:t}$ along with the map is to be computed in the optimisation based SLAM as follows [106],

$$p(\mathbf{x}_{r,1:t}, m|z_{1:t}, u_{1:t}). \tag{2.26}$$

The accepted traditional SLAM literature generally refers to filtering when only an estimate of the current robot pose and all the landmark locations are kept in the state vector. Indeed, one can add the previous history of all robot poses into the state vector as well and estimate them all, yet this is what is specifically referred as "smoothing" in the SLAM literature. These two generic filtering techniques are different to optimisation-based SLAM, which forms the backbone of the work in the thesis.

In the following part of this section, Extended Kalman Filter (EKF) based SLAM and Graph based SLAM are introduced as typical filtering based and optimisation based SLAM. Finally, a brief summary state-of-the-art SLAM techniques on different application scenarios will be introduced.

2.2.1 EKF SLAM

Extended Kalman filter based SLAM algorithm is historically the earliest, and perhaps the most influential SLAM algorithm. Briefly speaking, the EKF SLAM algorithm applies the EKF to the filtering based SLAM using maximum likelihood data association. However, EKF SLAM is also subject to a number of approximations and limiting assumptions detailed as follows [106],

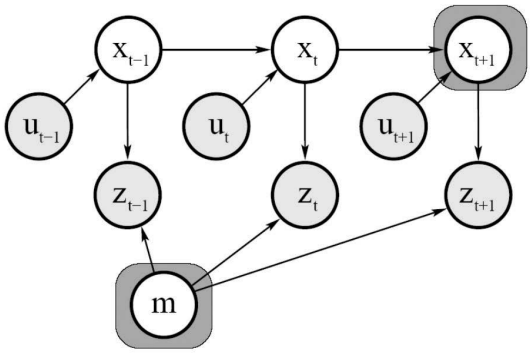


FIGURE 2.10: Bayesian Network for EKF SLAM [6].

- In the EKF SLAM algorithm, feature-based maps are utilised. This means SLAM maps are composed of point landmarks in the EKF. As a result, the number of point landmarks is usually small (e.g., smaller than 1,000) for computational reasons. Furthermore, the EKF SLAM algorithm tends to work better when the landmarks are less ambiguous. Therefore, significant engineering of feature detectors, such as using artificial beacons or landmarks as features, is needed for the EKF SLAM algorithm.

- In the EKF SLAM algorithm, Gaussian noise is assumed. In any EKF algorithm, EKF SLAM makes a Gaussian noise assumption for both the robot motion and the perception. In addition, since the linearisation in EKFs tends to introduce intolerable errors, the amount of uncertainty in the posterior must be relatively small.

- In the EKF SLAM algorithm, all measurements need to be positive measurements. In the EKF SLAM algorithm, only positive sightings of landmarks can be processed. Negative information that arises from the absence of landmarks in a sensor measurements can not be processed.

EKF SLAM problem can be represented as a recursive Bayesian Network as shown in Fig. 2.10. The EKF SLAM only keeps current location of the robot and map information in its state vector, as described in Eq. 2.25. The current estimate of the current robot pose $\mathbf{x}_{r,t}$ and the map m represents the state vector $\mathbf{x}_t$. As a result, only the previous state vector $\mathbf{x}_{t-1}$, the control input u_t and the observation of the map z_t influence the estimation of the state vector. EKF SLAM can be divided into two steps, the prediction

step and the update step. In the prediction step, based on the belief of the state vector in the previous step $bel(\mathbf{x}_{t-1})$, the control input u_t and the robot motion model, the belief of the current state vector after prediction step $\overline{bel}(\mathbf{x}_t)$ can be written as follows [6],

$$\overline{bel}(\mathbf{x}_t) = \int p(\mathbf{x}_t|u_t, \mathbf{x}_{t-1}) bel(\mathbf{x}_{t-1}) d\mathbf{x}_{t-1}. \tag{2.27}$$

During the update step, the robot observes the map and the update its belief of the current state vector based on the sensor observation model. The belief of the current state vector after update step $bel(\mathbf{x}_t)$ can be written as below [6].

$$bel(\mathbf{x}_t) = \eta p(z_t, \mathbf{x}_t) \overline{bel}(\mathbf{x}_t), \tag{2.28}$$

where η is a normalisation coefficient.

In practice, the EKF SLAM algorithm can be summarised as in the following equations [6].

$$\bar{\mathbf{x}}_t = g(u_t, \mathbf{x}_{t-1}), \tag{2.29}$$

$$\bar{P}_t = G_t P_{t-1} G_t^T + R_t, \tag{2.30}$$

$$K_t = \bar{P}_t {H_t}^T (H_t \bar{P}_t {H_t}^T + Q_t)^{-1}, \tag{2.31}$$

$$\mathbf{x}_t = \bar{\mathbf{x}}_t + K_t(z_t - h(\bar{\mathbf{x}}_t)), \tag{2.32}$$

$$P_t = (I - K_t H_t)\bar{P}_t, \tag{2.33}$$

where $\mathbf{x}_{t-1}$, $\bar{\mathbf{x}}_t$ and $\mathbf{x}_t$ are the system state vectors at time instance $t-1$, after the prediction step at time instance t and after the update step at time instance t, and P_{t-1}, $\hat{P}_t$ and P_t represents their corresponding covariance matrix. G_t is the Jacobian of robot motion model $g(u_t, \mathbf{x}_{t-1})$, u_t is the control input and R_t is the robot motion noise variance. H_t is the Jacobian of observation function $h(\bar{\mathbf{x}}_t)$, z_t is the sensor observation and Q_t is the observation noise variance. K_t is called Kalman gain.

2.2.2 Graph Based SLAM

A so-called Graph-based formulation is an intuitive way to address the SLAM problem. In a graph based SLAM, solving a SLAM problem has changed to constructing a graph whose nodes represent robot poses or landmarks and edges between two nodes represent sensor measurements that constrain the connected poses. As SLAM is an over determined system, such constraints can be contradictory since observations are always affected by noise. Therefore, the crucial problem is to find a configuration of the nodes that is maximally consistent with the measurements once such a graph of the SLAM problem is constructed. As a result, the graph based SLAM problem has changed to solving a large error minimisation problem [7].

Let $\mathbf{x}_i$ describes the pose of node i and $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)^T$ be the state vector of the SLAM system. Moreover, let $\mathbf{z}_{ij}$ and $\mathbf{\Omega}_{ij}$ be the mean and the information matrix of a virtual measurement between the node i and the node j respectively. This virtual measurement is such a transformation that it makes the observations acquired from i maximally overlap with the observation acquired from j. Given a configuration of the nodes $\mathbf{x}_i$ and $\mathbf{x}_j$, let $\hat{\mathbf{z}}_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ be the prediction of a virtual measurement, which is usually the relative transformation between the two nodes. The log-likelihood l_{ij} of a measurement $\mathbf{z}_{ij}$ therefore can be written as follows [7],

$$l_{ij} \propto [\mathbf{z}_{ij} - \hat{\mathbf{z}}_{ij}(\mathbf{x}_i, \mathbf{x}_j)]^T \mathbf{\Omega}_{ij} [\mathbf{z}_{ij} - \hat{\mathbf{z}}_{ij}(\mathbf{x}_i, \mathbf{x}_j)]. \tag{2.34}$$

A difference between the expected observation $\hat{\mathbf{z}}_{ij}$ and the real observation $\mathbf{z}_{ij}$ gathered by the robot is denoted as a error function $\mathbf{e}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{z}_{ij})$. Here, indices of the measurement are encoded in the indices of the error function for simplicity of notation. Then, the error function can be computed as follows,

$$\mathbf{e}_{ij}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{z}_{ij} - \hat{\mathbf{z}}_{ij}(\mathbf{x}_i, \mathbf{x}_j). \tag{2.35}$$

Figure 2.11 illustrates the functions and the quantities that are very important to definition of an edge of the graph. The actual measurement $\mathbf{z}_{ij}$ introduces the edge connecting the

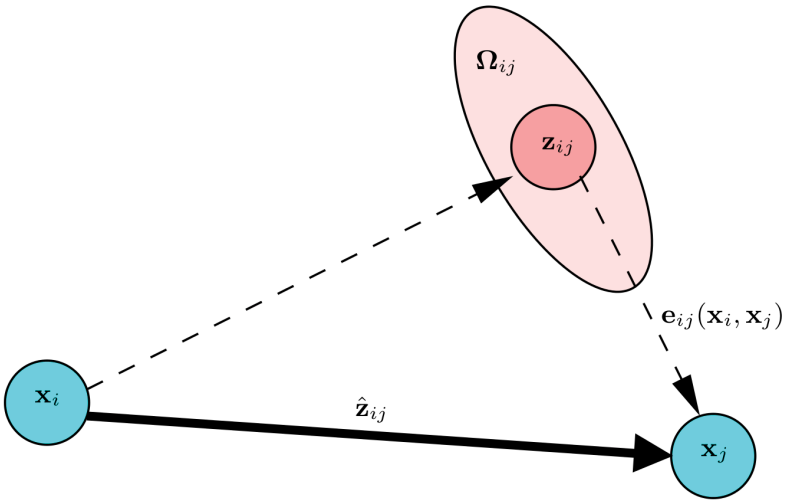


FIGURE 2.11: Aspects of an edge connecting the vertex $\mathbf{x}_i$ and the vertex $\mathbf{x}_i$ [7].

vertex $\mathbf{x}_i$ and the vertex $\mathbf{x}_j$. In addition, it is also possible to compute the expected measurement $\hat{\mathbf{z}}_{ij}$ that represents $\mathbf{x}_j$ seen in the frame of $\mathbf{x}_i$ from the relative position of the two nodes. The error function $\mathbf{e}(\mathbf{x}_i, \mathbf{x}_j)$ represents the difference between the expected and the real measurement. With the error function $\mathbf{e}(\mathbf{x}_i, \mathbf{x}_j)$ and the information matrix $\mathbf{\Omega}_{ij}$ of the measurement that accounts for its uncertainty, an edge is fully characterised.

The set of pairs of indices, for which a constraint (observation) $\mathbf{z}$ exists, is denoted as $\mathcal{C}$. Then, the goal of solving the SLAM problem by a maximum likelihood approach is to find the best configuration of the nodes $\mathbf{x}^*$ in such a way that they minimise the negative log likelihood $\mathbf{F}(\mathbf{x})$ of all the observations as follows [7],

$$\mathbf{F}(\mathbf{x}) = \sum\nolimits_{<i,j>\in\mathcal{C}} \underbrace{\mathbf{e}_{ij}^T \mathbf{\Omega}_{ij} \mathbf{e}_{ij}}_{\mathbf{F}_{ij}}, \tag{2.36}$$

thus, it seeks to solve the following equation:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \mathbf{F}(\mathbf{x}). \tag{2.37}$$

In the next part of this section, an approach to solve Eq. 2.37 and to compute a Gaussian approximation of the posterior over the robot trajectory is detailed. Although the method described in the following section utilises standard optimisation methods, such as the

Gauss-Newton or the Levenberg-Marquardt algorithms, it is much more efficient because the structure of the problem is effectively exploited.

The numerical solution of Eq. 2.37 can be obtained by using the popular Gauss-Newton or Levenberg-Marquardt algorithms if a good initial guess $\breve{\mathbf{x}}$ of the robot poses is given a priori. This is based on the idea that approximation of the error function can be done by its first order Taylor expansion around the current initial guess $\breve{\mathbf{x}}$ as follows [7],

$$\begin{aligned} \mathbf{e}(\breve{\mathbf{x}}_\mathbf{i} + \Delta\mathbf{x}_i, \breve{\mathbf{x}}_\mathbf{j} + \Delta\mathbf{x}_j) &= \mathbf{e}_{ij}(\breve{\mathbf{x}} + \Delta\mathbf{x}) \\ &\simeq \mathbf{e}_{ij} + \mathbf{J}_{ij}\Delta\mathbf{x}, \end{aligned} \tag{2.38}$$

where $\mathbf{e}_{ij} \overset{\text{def.}}{=} \mathbf{e}_{ij}(\breve{\mathbf{x}})$ and $\mathbf{J}_{ij}$ is the Jacobian of $\mathbf{e}_{ij}(\mathbf{x})$ computed in $\breve{\mathbf{x}}$. By combining Eq. 2.38 and the error terms $\mathbf{F}_{ij}$ of Eq. 2.36, we can derive the equation as follows,

$$\begin{aligned} \mathbf{F}_{ij}(\breve{\mathbf{x}} + \Delta\mathbf{x}) &= \mathbf{e}_{ij}(\breve{\mathbf{x}} + \Delta\mathbf{x})^T \mathbf{\Omega}_{ij} \mathbf{e}_{ij}(\breve{\mathbf{x}} + \Delta\mathbf{x}) \\ &\simeq (\mathbf{e}_{ij} + \mathbf{J}_{ij}\Delta\mathbf{x})^T \mathbf{\Omega}_{ij} (\mathbf{e}_{ij} + \mathbf{J}_{ij}\Delta\mathbf{x}) \\ &= \underbrace{\mathbf{e}_{ij}{}^T \mathbf{\Omega}_{ij} \mathbf{e}_{ij}}_{c_{ij}} + 2\underbrace{\mathbf{e}_{ij}{}^T \mathbf{\Omega}_{ij} \mathbf{J}_{ij}}_{\mathbf{b}_{ij}} \Delta\mathbf{x} + \Delta\mathbf{x}^T \underbrace{\mathbf{J}_{ij}{}^T \mathbf{\Omega}_{ij} \mathbf{J}_{ij}}_{\mathbf{H}_{ij}} \Delta\mathbf{x} \\ &= c_{ij} + 2\mathbf{b}_{ij}\Delta\mathbf{x} + \Delta\mathbf{x}^T \mathbf{H}_{ij} \Delta\mathbf{x}. \end{aligned} \tag{2.39}$$

Then, under the local approximation, the function $\mathbf{F}(\mathbf{x})$ can be rewritten in Eq. 2.36 as follows,

$$\mathbf{F}(\breve{\mathbf{x}} + \Delta\mathbf{x}) = \sum_{<i,j>\in\mathcal{C}} \mathbf{F}_{ij}(\breve{\mathbf{x}} + \Delta\mathbf{x}) \tag{2.40}$$

$$\simeq \sum_{<i,j>\in\mathcal{C}} c_{ij} + 2\mathbf{b}_{ij}\Delta\mathbf{x} + \Delta\mathbf{x}^T \mathbf{H}_{ij} \Delta\mathbf{x} \tag{2.41}$$

$$= c + 2\mathbf{b}^T\Delta\mathbf{x} + \Delta\mathbf{x}^T \mathbf{H} \Delta\mathbf{x}. \tag{2.42}$$

After setting $\mathbf{b} = \sum \mathbf{b}_{ij}$, $c = \sum c_{ij}$, and $\mathbf{H} = \sum \mathbf{H}_{ij}$ in Eq. 2.41, the quadratic form in Eq. 2.42 is obtained. By solving the linear system, it can be minimised in $\Delta\mathbf{x}$ as follows,

$$\mathbf{H}\Delta\mathbf{x}^* = -\mathbf{b}. \tag{2.43}$$

Since the matrix $\mathbf{H}$ is obtained by projecting the measurement error in the space of the trajectories via the Jacobians, it is the information matrix of the system. In addition, the matrix $\mathbf{H}$ is also sparse by construction. It has non-zeros between poses connected by a constraint and its number of non-zero blocks is twice the number of constrains plus the number of nodes. The above mentioned characteristic allows us to solve Eq. 2.43 by sparse Cholesky factorisation by using a library such as CSparse [107], which is an efficient yet compact implementation of sparse Cholesky factorisation. The detailed illustration of sparse structure of the $\mathbf{H}_{ij}$ and $\mathbf{b}_{ij}$ can be found in [7].

Then, by adding to the initial guess the computed increments, the linearised solution can be obtained as follows,

$$\mathbf{x}^* = \breve{\mathbf{x}} + \Delta\mathbf{x}^*. \tag{2.44}$$

In the popular Gauss-Newton algorithm, the linearisation is iterated in Eq. 2.42, the solution is iterated in Eq. 2.43, and the update step is iterated in Eq. 2.44. The previous solution is used as the linearisation point and the initial guess in every iteration.

Up to now, we introduced a general approach to multivariate function minimisation which is derived for the special case of the SLAM problem. However, the derived general approach has an assumption, that is the space of parameters $\mathbf{x}$ is Euclidean. This assumption is not valid for SLAM and may lead to suboptimal solutions. Advanced topics related to optimisation on manifold to deal with non-Euclidean spaces can be found in [7].

2.2.3 SLAM Applications

There exist a wide range of SLAM implementations based on different sensors, environments and application scenarios. In this section, we introduce some popular state-of-the-art methods based on visual, Lidar and the combination of visual inertial sensors.

2.2.4 Monocular Visual SLAM

In monocular visual SLAM, one single camera is used to perform the localisation and mapping operation. It estimates 6 DOF pose of the sensor with scale ambiguity.

PTAM [108] proposed by Klein and Murray is the representative keyframe-based monocular SLAM system, which is probably the first work to introduce the idea of splitting camera tracking and mapping in parallel threads. It demonstrates to be successful for real-time augmented reality applications in small environments. Later the authors published an improvement of the original version with edge features, a rotation estimation step during tracking, and a better relocalisation method in [109].

Forster et. al. presents a semidirect visual odometry (SVO), which is in a halfway between direct and feature-based methods. SVO is able to operate at high frame rates and obtain impressive results in a high speed quadcopter, without requiring to extract features in every frame.

Mur-Artal et. al. presents ORB-SLAM [110], a feature-based monocular SLAM system that operates in real time, in small and large indoor and outdoor environments. ORB-SLAM allows full automatic initialisation, wide baseline loop closing and relocalisation. It consists of tracking, mapping, relocalisation and loop closing. When selecting keyframes, a survival of the fittest strategy is applied, which leads to excellent robustness and generates a compact and trackable map that only grows if the scene content changes, allowing lifelong operation.

2.2.5 Stereo Visual SLAM

In stereo visual SLAM, two cameras with known baseline are used to estimate the 6 DOF pose of the sensor without scale ambiguity.

Engel et. al. presents large-scale direct SLAM (LSD-SLAM) [111], which is a semidense direct approach that minimises photometric error in image regions with high gradient. LSD-SLAM is expected to be more robust to motion blur or poorly textured environments since it does not rely on features. As a drawback, the performance of LSD-SLAM can be severely degraded by unmodeled effects like rolling shutter or non-Lambertian reflectance.

ORB-SLAM2 [8] presented by Mur-Artal et. al. for stereo and RGB-D cameras is built upon their monocular feature-based ORB-SLAM [110]. Similar to the ORB-SLAM, the ORB-SLAM2 system also consists of three main parallel threads. The tracking thread is

FIGURE 2.12: ORBSLAM2 with stereo input: Trajectory and sparse reconstruction of an urban environment with multiple loop closures [8].

FIGURE 2.13: ORBSLAM2 with RGB-D input: Keyframes and dense pointcloud of a room scene with one loop closure [8].

to localise the camera with every frame by finding feature matches to the local map and minimising the reprojection error applying motion-only BA. The local mapping manages the local map and optimise it by performing local BA. The loop closing detects large loops and correct the accumulated drift by performing a pose-graph optimisation. It also launches a fourth thread to perform full BA after the pose-graph optimisation to compute the optimal structure and motion solution. Two example results of ORB-SLAM2 with stereo and RGB-D inputs are shown in Fiq. 2.12 and Fig. 2.13.

2.2.6 RGB-D Visual SLAM

In RGB-D visual SLAM, a RGB-D camera is used to estimate the 6 DOF pose of the sensor without scale ambiguity. A RGB-D camera consists of a color image (RGB) camera and a depth image camera.

The RGB-D SLAM proposed by Endres et. al. [112] is probably the first popular open-source SLAM system that uses a RGB-D camera. RGB-D SLAM is a feature-based system and its front end computes frame-to-frame motion by using both feature matching and ICP. Its back end performs pose-graph optimisation with loop closure constraints from a heuristic search.

Similar to RGB-D SLAM, the back end of DVO-SLAM proposed by Kerl et al. [113] optimises a pose graph where keyframe to keyframe constraints are computed from a visual odometry, which minimises both photometric and depth error. Regarding loop closure, DVO-SLAM searches for candidates in a heuristic fashion over all previous frames instead of relying on place recognition.

Whelan et. al. presents ElasticFusion [114], which builds a surfel-based map of the environment. In ElasticFusion, instead of a standard pose-graph optimisation, it forgets poses and performs loop closing applying a nonrigid deformation to the map. ElasticFusion produces the detailed reconstruction and impressive localisation accuracy, while the current implementation is limited to room-size maps as the complexity scales with the number of surfels in the map.

2.2.7 2D/3D Lidar based SLAM

Hess et. al. presents a 2D SLAM called "Cartographer" in [115]. Cartographer combines scan-to-submap matching with loop closure detection and graph optimisation. In Cartographer, individual submap trajectories are created using the local grid-based SLAM approach, while all scans are matched to nearby submaps using pixel-accurate scan matching to create loop closure constraints in the background. The constraint graph of submap and scan poses is periodically optimised. It is also demonstrated that the Cartographer can run the on modest hardware in real-time.

FIGURE 2.14: LOAM results of mapping university campus [9].

Zhang et. al. presents the LOAM method [9], which provides an accurate and effective registration of a 2D sweeping or 3D LiDAR scans. It is further improved by fusion with data provided by a vision sensor in VLOAM [116]. Both LOAM and VLOAM detect edges and planar points in the LiDAR scans for which a set of nonlinear equations constraining the odometry is generated. A non-linear optimisation results in the final 6 DOF poses of the sensor. LOAM and VLOAM methods achieved the best results in the KITTI evaluation benchmark [117]. An example result of mapping a university campus using LOAM is shown in Fig. 2.14.

2.2.8 Visual Inertial SLAM

In visual inertial SLAM, the information from a camera sensor and a inertial sensor is fused to estimate the 6 DOF pose of the sensor.

In [118], Leutenegger et. al. introduces a framework named OKVIS, which is a tightly-coupled fusion of inertial measurements and image keypoints in a nonlinear optimisation problem that applies linearisation and marginalisation in order to achieve keyframing. The OKVIS outputs the sensor poses, velocities, and IMU biases as a time series, together with a 3D map of sparse landmarks. Since the optimisation includes a fixed number of poses, the OKVIS algorithm is bounded in complexity. Its keyframing strategy, in contrast to a fixed-lag smoother, results in high accuracy, while still being able to operate at real-time.

In [119], Forster et. al. proposes a novel preintegration theory, which provides a grounded way to model a large number of IMU measurements as a single motion constraint. As the

proposed method does not commit to a linearisation point during integration, it improves conventional works that perform integration in a global frame. In addition, it brings maturity to the preintegration and uncertainty propagation in SO(3). It also adopts a structureless model for visual measurements which avoids optimising over 3D landmarks.

2.3 Sound Source Mapping

Due to the important application of sound source mapping introduced in 1.1, a number of sound source mapping works have been represented in the literature. In this section, some of the typical sound source mapping methods are introduced.

Based on the sensor being used for robot localisation, these robotic sound source mapping works can be divided into two categories: 1) using a microphone array only [2, 4] and 2) using an additional accurate exteroceptive sensor (such as a camera or a laser range finder) together with a microphone array [10, 120]. In both scenarios, the robot can have an proprioceptive sensor such as wheel odometry.

In the first scenario, due to the fact that sound sources are sparse and noisy and wheel odometry drifts, some considerations need to be imposed to most of the works belong to this category in the literature. These considerations can be: the robot moves relatively short distances so the drift in odometry remains small [2] or multiple sound sources are mapped at the same time in order to obtain a sufficient number of observations to constrain the robot pose [4].

In the second scenario, with the help of an additional exteroceptive sensor, robot self localisation can be much more accurate, so can the sound source mapping accuracy.

2.3.1 Robotic Sound Source Mapping with a Microphone Array Only

2.3.1.1 Sound Source Mapping by Self-Motion Triangulation

In [2, 121], Sasaki et. al. describes a 2D sound source mapping system for a mobile robot. They developed a multiple sound sources localisation method for a mobile robot with a

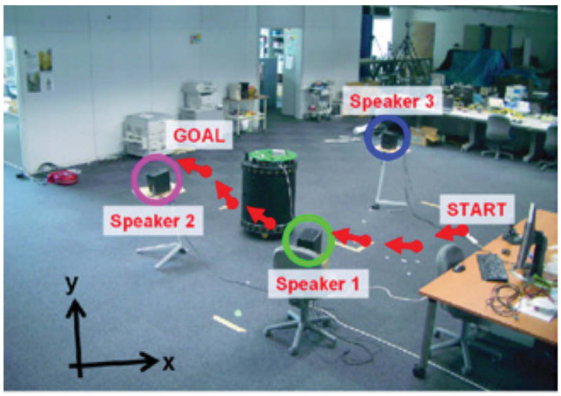


FIGURE 2.15: Experimental setup of the self-motion triangulation method in [2].

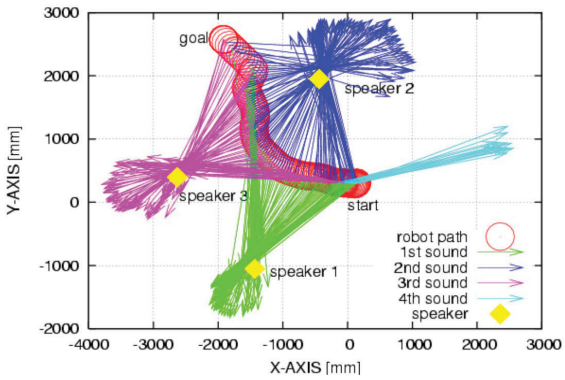


FIGURE 2.16: Experimental result of the self-motion triangulation method in [2].

32-channel concentric microphone array. Directional localisation and separation of different pressure sound sources is achieved using the Delay and Sum Beam Forming(DSBF) and the Frequency Band Selection(FBS) algorithm. Sound sources were mapped by using a wheeled robot equipped with the microphone array. The robot estimates sound sources bearing information on the move and maps sound sources positions using triangulation. RAndom SAmple Consensus(RANSAC) algorithm is used for rejecting outlier triangulation points to improve the sound source mapping accuracy. The system achieved 2D multiple sound source mapping with high accuracy. In addition, moving sound source separation is experimentally demonstrated with segments of the DSBF enhanced signal derived from the sound source mapping process.

The experimental setup and result are shown in Fig. 2.15 and Fig. 2.16. As can be seen from the figures, during the robot motion, the robot obtains bearing estimation of three sound sources from multiple locations. Those rays pointing from robot locations to sound

FIGURE 2.17: Experimental setup of the FastSLAM method in [4].

sources are triangulated to get multiple triangulation points. Outliers of triangulation points will be removed by RANSAC algorithm and the mean locations of triangulation points are returned as estimations of sound sources locations.

There are two limitations with this method. Firstly, since the odometry drifts over longer distance, the accuracy of sound source mapping degrades over the long run. Secondly, there is no uncertainty associated with each sound source location.

2.3.1.2 Sound Source Mapping using FastSLAM

In [4], Hu et. al. proposes a framework that simultaneously localises the mobile robot and multiple sound sources using a microphone array on the robot. An eigenstructure-based GCC method for estimating time delays between microphones under multi-source environment is described. Then, using the estimated time delays, a method to compute the far-field source directions as well as the speed of sound is proposed. The correctness of the sound speed estimate is utilised to eliminate spurious sources, which greatly enhances the robustness of sound source detection. The bearing estimation of the detected sound sources are used as observations in a bearing-only SLAM. The FastSLAM [122] algorithm is used for sound source mapping and sound sources data estimation, since the source signals are not persistent and there is no identification of the signal content.

The experimental setup and result are shown in Fig. 2.17 and Fig. 2.18. As can be seen from the figures, using the FastSLAM algorithm, which is a multi-hypothesis EKF SLAM approach, the robot can simultaneously localise its own position and estimate the locations of sound sources.

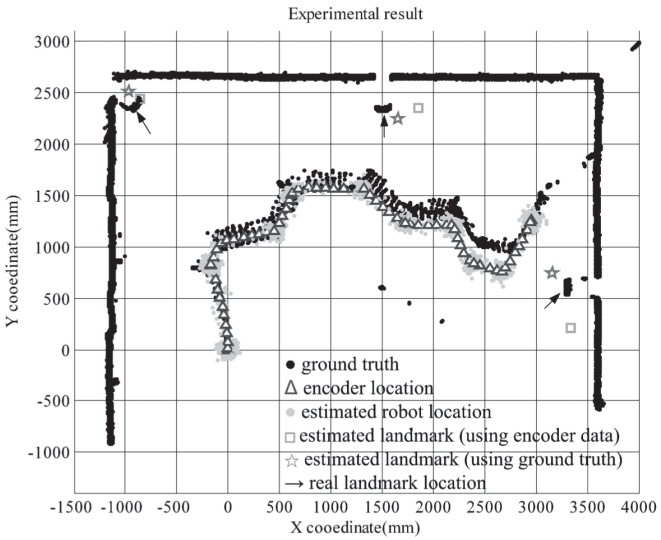


FIGURE 2.18: Experimental result of the FastSLAM method in [4].

There is one limitation with this method. The robot needs to observe multiple sound sources all the time. Otherwise, due to the noisy bearing observation and shortage of number of landmarks, the SLAM accuracy degrades severely. However, reliable observations of multiple sound sources are not likely to occur in most scenarios, so controlled environments like the one in [4] are required.

The authors' previous work in [123], presents a sound source mapping using Structure from Motion (SfM) [124]. Similar to this example, their previous work [123] also needs presence of multiple sound sources to localise the robot.

2.3.1.3 Sound Source Mapping using Unscented Kalman Filter (UKF)

In [90], Portello et. al. presents a method for binaural sound mapping. An UKF is used to provide range and azimuth estimation of a sound source relative to the robot reference frame.

The simulation result is shown in Fig. 2.19. The sensor center follows a circular trajectory, with a constant interaural axis velocity. As the robot moves, the confidence ellipsoid shrinks along all directions during the estimation of range and bearing of the sound source. The estimation of sound sources locations converges to its ground truth locations. The limitation of the method is that it assumes accurate robot location is known.

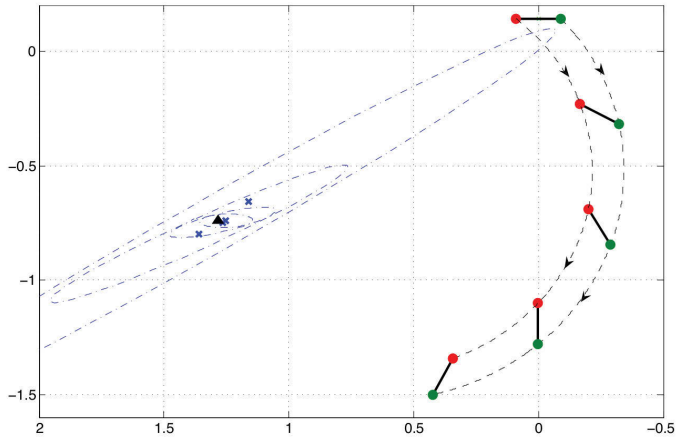


FIGURE 2.19: Experimental result of the FastSLAM method in [4].

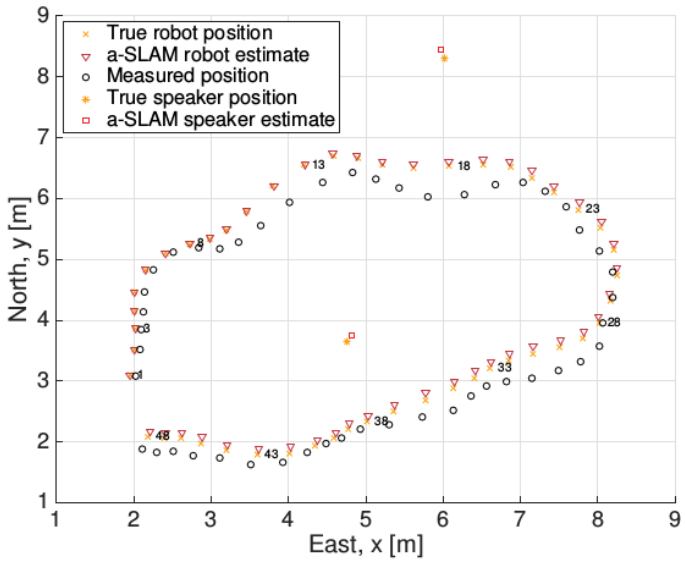


FIGURE 2.20: Experimental result of the Single-Cluster Probability Hypothesis Density filter method in [125].

2.3.1.4 Sound Source Mapping using Single-Cluster Probability Hypothesis Density filter

In [125], Evers et. al. presents a single-cluster probability hypothesis density (SC-PHD) filter based sound source mapping method. They show that localisation of a moving microphone array and mapping of the surrounding sound sources are jointly dependent and can be simultaneously estimated using a bearing-only SC-PHD filter.

The simulation result is shown in Fig. 2.20. The sensor center follows a circular trajectory.

Acoustic scene map in the figure shows the robot pose (red triangles) and the sound sources locations estimation (red squares), the robot pose (orange crosses) and the sound sources locations ground truth (orange asterisks) and the robot trajectory from speed and orientation measurements only (black circles). The robot speed and orientation is assumed to be measurable. From the result, it can be seen that the pose of the robot and the sound sources locations have been successfully estimated with reasonable accuracy. However, similar to the self motion triangulation based sound source mapping method in section 2.3.1.1, there is no uncertainty associated with each sound source location. Moreover, when the robot travels a longer trajectory, the integration of robot speed and orientation measurements drifts significantly. In such a situation, the feedback from the noisy bearing estimation of limited number of sound sources is not too much helpful to correct the robot pose to its ground truth value.

2.3.2 Robotic Sound Source Mapping with an Additional Exteroceptive Sensor

2.3.2.1 Sound Source Mapping using Ray Tracing Method

In [10], Kallakuri et. al. presents a multi-modal sensor approach for mapping sound sources using an omni-directional microphone array on an autonomous mobile robot. A fusion of audio data (from the microphone array), odometry information and the laser range scan data (from the robot) is used to precisely localise and map the audio sources in an environment. An audio map is created while the robot is autonomously navigating through the environment by continuously generating audio scans with a SRP algorithm. Using the poses of the robot, rays are cast in the map in all directions given by the SRP. Then each occupied cell in the geometric map hit by a ray is assigned a likelihood of containing a sound source. This likelihood is derived from the SRP at that particular instant. Since particle filter is used for the localisation of the robot, the uncertainty in the pose of the robot in the geometric map is propagated to the occupied cells hit during the ray casting. This process is repeated while the robot is in motion and the map is updated after every audio scan. The generated sound maps are reused and the changes in the audio environment are updated by the robot as it identifies these changes.

FIGURE 2.21: Experimental setup of the ray tracing method in [10].

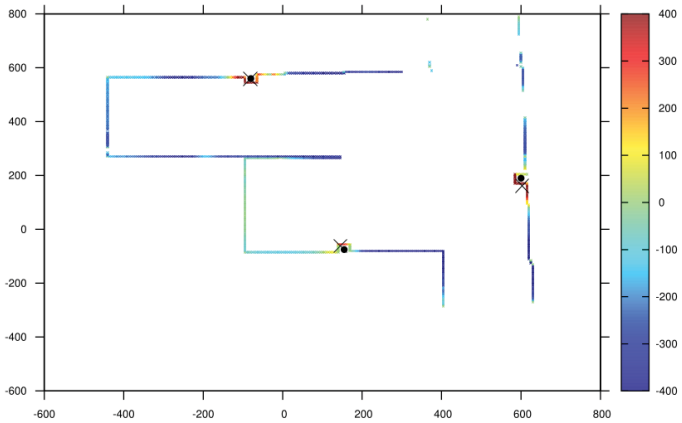


FIGURE 2.22: Experimental result of the ray tracing method in [10].

The experimental setup and result in 2D case are shown in Fig. 2.21 and Fig. 2.22. It can be seen from these figures that the grids that are occupied (hit by the laser scan) and close to the locations of sound sources get a higher likelihood of being a sound source as they are continuously amended with positive log likelihood values. Other occupied grids which are far away from sound sources locations are continuously amended with negative log likelihood values. The obstacle free grids are disregarded. Therefore, a strong assumption of this method in 2D case is that sound sources have to be detectable by laser scanner. Otherwise, those positive log likelihood audio rays would wrongly hit obstacles that are not part of sound sources.

The 3D extension of their work is presented in [11], which presents a framework for creating a 3D map of an environment that contains the probability of a geometric feature to emit

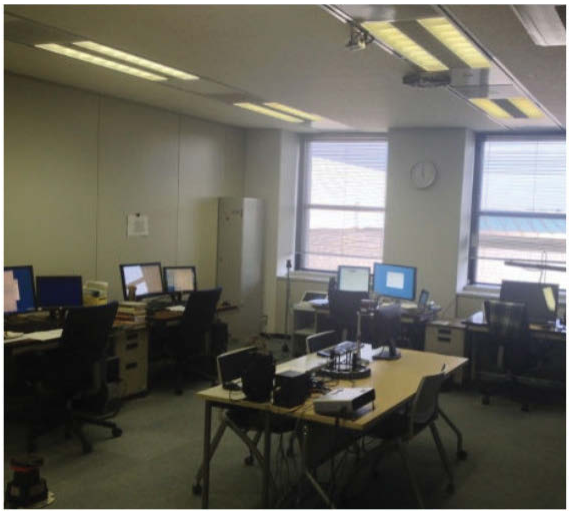


FIGURE 2.23: Experimental setup of the 3D ray tracing method in [11].

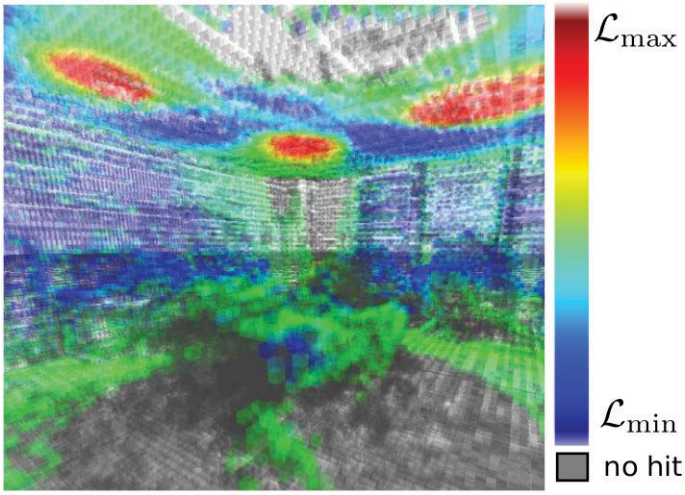


FIGURE 2.24: Experimental result of the 3D ray tracing method in [11].

a sound. The experimental setup and results are shown in Fig 2.23 and Fig 2.24.

The limitation of both 2D and 3D version of ray tracing methods are: firstly, the laser scan needs to hit sound sources first. If, for some reason, the laser scan does not hit the sound sources, these sound sources would not be mapped even if they are sensed by the microphone array. Secondly, some isolated sound sources might lead to false audio ray tracing to the obstacle behind it, as they point out in [10].

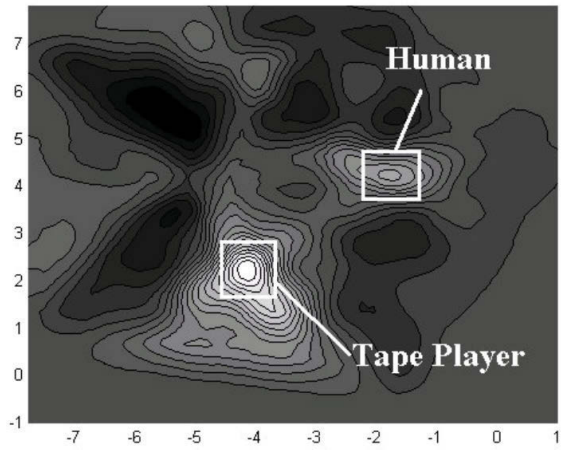


FIGURE 2.25: Experimental result of the auditory occupancy grid method in [12].

2.3.2.2 Sound Source Mapping using Auditory Evidence Grid Method

In [12], Martinson et. al. presents a sound source mapping method based on the auditory evidence grid method. Similar to the occupancy grid mapping, the evidence grid representation uses Bayesian updating to estimate the probability of a sound source being located in a set of predetermined locations (i.e. a grid cell center). Initially, it is assumed that every grid cell has a 50% probability of containing a sound source. Then as each new sensor measurement is added to the evidence grid, those probabilities for each grid cell are adjusted. The robot uses a laser scanner to localise itself. The result is a representation that localises the pertinent objects well over time, can be used to filter poor localisation results, and may also be useful for global re-localisation from sound localisation results.

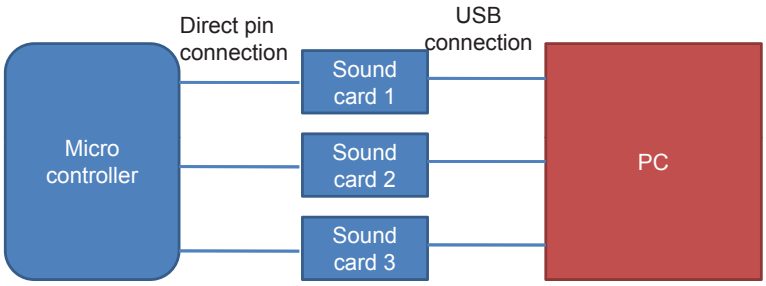
The experimental result is shown in Fig 2.25. As can be seen from the figure, two sound sources from human and tape player are successfully localised. The limitations of this method are as follows. Firstly, as a common limitation of the grid based method, the resolution of the estimation accuracy depends on the size of the grid. Smaller size of grid results in higher accuracy but costs higher memory consumption and vice versa. Secondly, the uncertainty of the robot pose estimation is not considered and the mapping process is essentially a mapping based on "known" pose. This means even if the robot pose estimation is corrected after the loop closure, the mapping result will not be corrected accordingly.

Chapter 3

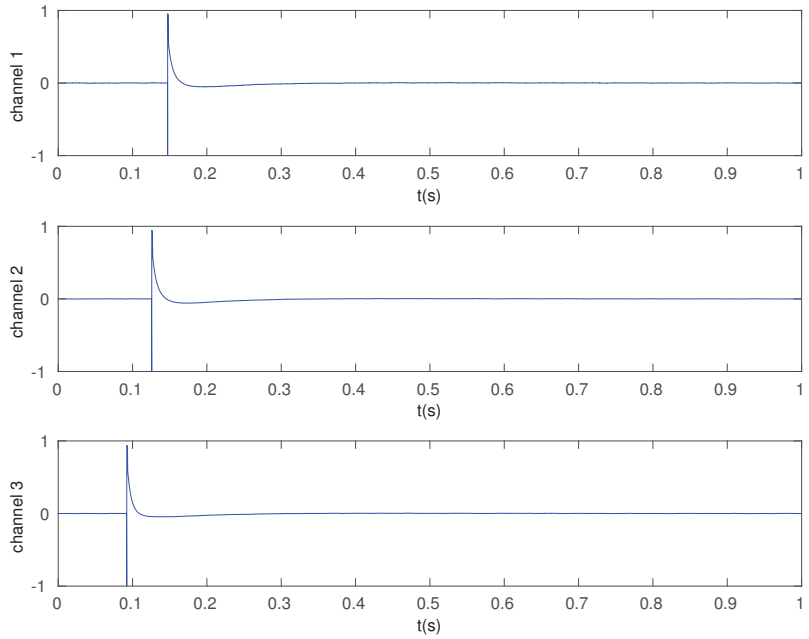
Calibration of a Microphone Array

3.1 Introduction

Processing the signals from a microphone array has proven to be an effective approach to improve robot audition. Many robot audition systems based on microphone arrays have been proposed in the literature [19–21, 59]. By exploiting this technique, robots are able to localise and track different sound sources, separate speech coming from several people simultaneously and automatically recognise each separated speech. Most of these studies utilise a synchronised microphone array, which requires hardware synchronisation of each independent microphone channel. Specifically, synchronisation needs a special sound capturing device such as a multi-channel ADC converter. For a synchronised microphone array, to find the DOA of a sound source, the TDOA from the sound source to each channel of the microphone array is exploited. When the geometric locations of microphones are known, conventional DOA estimation algorithms can be used to estimate bearing information of the sound source from its TDOA information. However, given the constraints of embedding microphones in a robot, often alongside other perception devices, it is difficult to measure the exact location of the microphones accurately. One alternative of measuring the exact location, a transfer function between each microphone and a sound source is measured. These measurements, however, can be quite time-consuming since they need to be obtained at multiple intervals of sound source directions (*e.g.* every 5 degree) [126]. Therefore the proposed strategy, while more suitable for asynchronous microphone arrays



a The hardware connection.



b The pulse signal received on each channel.

FIGURE 3.1: Experimental setup for testing clock differences.

as will become apparent in this Chapter, is also applicable to hardware synchronised arrays where essentially only estimates of the geometric locations of the microphones are required.

While several commercial products of hardware synchronisation boards exist, they are either too expensive or too large in size to be integrated inside robotic platforms [126]. Recent methods have started to relax these assumptions, and use an asynchronous microphone array to localise sound sources. For instance self-localisation approaches for ad-hoc arrays have been proposed in [64] [65] [66] [67]. Most of these approaches can achieve high accuracy in microphone array self-localisation. Both [64] and [67] provide closed-form

estimators, in contrast to the standard iterative solution. The method presented in [64] also considers an acoustically determined, orientation estimate of a device that contains a microphone array. This method has been used in localisation of an asynchronous source in [127]. Raykar et al.'s work [128] on self-localisation formulates a maximum likelihood estimation for unknown parameters of a microphone array (time offsets, microphone positions) and measurements (TDOA or TOF) by utilising active emissions. Ono et al. [129] present a TDOA-based cost function approach, which does not require controlled calibration signal for estimating self-localisation, source localisation and temporal offset estimation. An online approach utilising SLAM is presented by Miura et al. [126], which used extended Kalman filtering and delay-and-sum beamforming to calibrate the stationary array.

While these methods are capable of computing individual microphone locations and the time offsets between different microphone channels, all of them are based on the assumption that the clock interval, in each independent sound card dedicated to each channel, is identical to those of the others. This is a strong assumption that disregards errors from fractional differences in clock intervals, which will accumulate over time. Sound cards, especially those designed for general consumption, have indeed noticeable drifts. An example is shown in Fig. 3.1(a). A microcontroller, connected to the signal line of each of these three microphones, generates a simultaneous pulse after a fixed time interval, remaining at high impedance until the next regular pulse. Fig. 3.1(b) shows a detail of the difference in arrival time for each microphone, whilst Fig. 3.2 represents the evolution in the difference between each pair of channels with respect to the first one (vertical axis indicates the offset as number of samples to normalise the comparison). These signal time offsets in Fig. 3.2 are a combination of the starting time offsets and clock difference rate by the elapsed time. From this simple setup, it can be easily observed how time-offsets between pairs of channels keep increasing over time due to clock drifts from the small variations in the clock intervals of each sound card. This means the calibration of an asynchronous microphone array needs to estimate the starting time offsets and clock differences of microphones in addition to their geometric locations. Estimating the starting time offset and clock difference essentially mean estimating the intersection point of the Y axis and slope of the line in Fig. 3.2, which together determine the line uniquely.

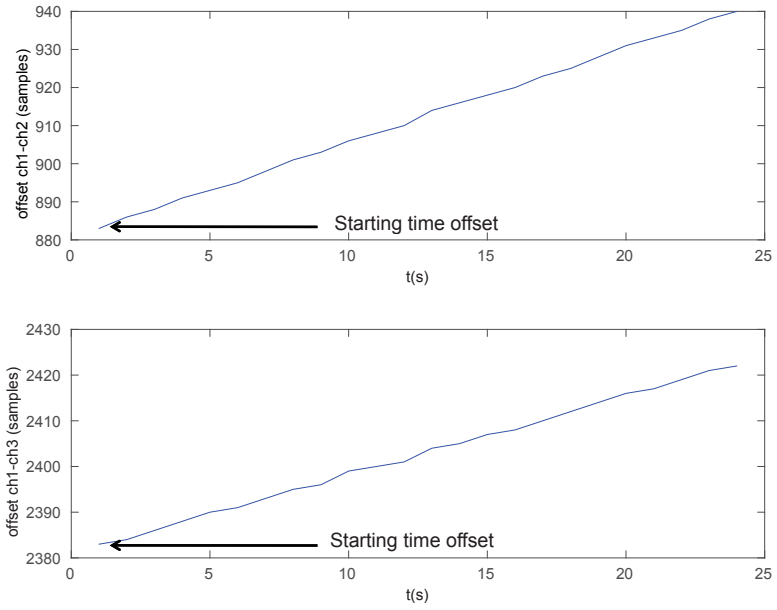


FIGURE 3.2: Detected differences of peak arrival time.

The method proposed in this thesis overcomes this issue with clock drifts by calibrating the asynchronous microphone array. The approach is based on a graph-based SLAM method to calibrate the array, which implies estimating the position, the starting time offset and clock difference of each microphone simultaneously. As an additional advantage, the trajectory of the sound source can also be recovered at the same time. In the same way as the SLAM problem, the problem in hand is formulated as a sparse, least-squares minimisation problem, where the minimum is found iteratively using Gauss-Newton algorithm. This is equivalent to finding the Maximum-Likelihood (ML) estimate of the sequence of sound sources locations and the array calibration under the assumption of Gaussian noise. Simulation and experimental results for a random walk sound source and arrays of variable number of microphones show the viability of the approach. Note that the proposed method works for both an asynchronous microphone array and a hardware synchronised microphone array.

For the hardware synchronised microphone array, we simply do not estimate the starting time offsets and clock differences parameters of microphones by removing them from the state vector, since these parameters are all the same and equal to zero. Therefore, in the following part of the Chapter, only the case of an asynchronous microphone array is presented. With a synchronised array, one can simply remove the stating time offsets and

clock differences parameters from the state vector and only focus on estimating geometric positions of microphones as stated before.

The rest of this Chapter is organised as follows. We first start investigating the calibration of a 2D/3D microphone array and extend the method to a linear microphone array case. In section 3.2, the detailed explanation of the proposed method in 2D scenario is presented. In section 3.3, the proposed method in 2D scenario has been extended to 3D scenario. In section 3.4, the calibration method of a 2D/3D microphone array is extended to a linear microphone array scenario. In section 3.5, comprehensive simulations and experimental results are presented. Section 3.6 presents the conclusion and discussion about further work.

3.2 Calibration of a 2D Asynchronous Microphone Array

In our system, the nonlinear least squares minimisation aims at recovering a sequence of sound source positions and the static configuration of the microphone array given a set of relative measurements and the number of microphones of the array. As the sound source moves around, the microphone array produces bearing observations of it from multiple viewpoints. Enforcing consistency between the different views gives rise to the location constraints (of sound source and microphone array).

3.2.1 System Model

Let $\mathbf{x}_{mic}$ be the state of the microphone array and $\mathbf{p}_k$ be the position of the sound source at the time $t_k = t_1 ... t_K$. Thus the full state is given by,

$$\mathbf{x}^T = (\mathbf{x}_{mic}^T \mathbf{p}_1^T \dots \mathbf{p}_K^T), \tag{3.1}$$

where

$$\mathbf{x}_{mic}^T = (\mathbf{x}_{mic\_1}^T \dots \mathbf{x}_{mic\_N}^T), \tag{3.2}$$

and N is the total number of microphones, which is known.

Note that in this case the part of the state of each microphone contains

$$\mathbf{x}_{mic\_n}^T = (x_{mic\_n}^x x_{mic\_n}^y x_{mic\_n}^\tau x_{mic\_n}^\delta) \tag{3.3}$$

for $n = 1...N$, where the location is given by the variables with superscripts x and y and the variables with subscripts τ and δ represent the starting time offset and the clock difference per second of each microphone respectively.

In a similar way, the part of the state of the sound source is

$$\mathbf{p}_k^T = (p_k^x p_k^y) \tag{3.4}$$

which contains only two variables that represent x and y position, as the orientation is not estimated in this case.

Let the microphone 1 be used as the reference, then time offsets and clock differences of other microphones are computed relative to the microphone 1. Hence, $x_{mic\_1}^\tau = 0$ and $x_{mic\_1}^\delta = 0$. Moreover, in order to define the position and orientation of the reference frame, the origin and, x and y axes need to be defined. As microphone 1 is the reference its position is set as $(0, 0)$. Let also another microphone (for instance the 2nd) define the positive direction of the x axis. However, if the microphone array is bi-dimensional there will be two possible solutions for the position of the microphone array $\pm y$. This will fully define our reference frame. In practice if the structure of the array is known, it can be exploited to remove the ambiguity in the y direction, *e.g.* another microphone can define the positive direction of the Y axis. This will fully define our reference frame. Note that it is assumed in any case that the number of microphones N is known and fixed.

To make the analogy to a standard SLAM framework as shown in Fig. 3.3, the sound source locations are treated as robot poses and the microphone array is treated as a single landmark. This landmark has the particularity of being observable at all sound source positions. Note that the main difference with a standard landmark-pose SLAM system is that here all the microphones are "observed" at all the time. In a standard SLAM system only part of the landmarks are observed at any time. This fact allows the microphone array to be treated as a single landmark with a large state that contains all microphones

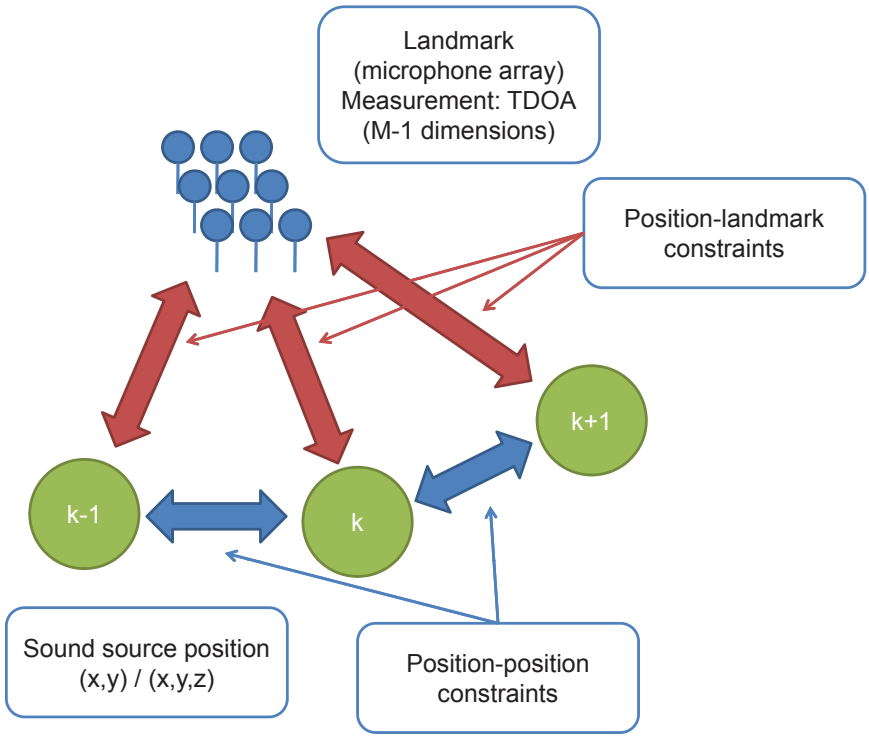


FIGURE 3.3: Description of the poses, landmarks and constraints in the SLAM framework.

locations. The same solution, however, can be achieved if the microphones are considered independently. The microphone array becomes the first node in the graph-based SLAM framework and each sound source position becomes one node.

Continuing the analogy with the standard graphical SLAM framework, for the position-position constraints, in this work the trajectory of the sound source is assumed to be arbitrary (no-odometry prior is considered), with the only constraint of two adjacent locations set to be not too distant from each other. Therefore, we use a random walk model[1] in which the sound source position of the next time instance is expected to be at the same location as the previous time with a large uncertainty associated as

$$\mathbf{z}_{k-1,k}^{p-p} = \mathbf{0} \tag{3.5}$$

$$I_{k-1,k}^{p-p} = \frac{1}{\sigma_{p-p}^2} I\,, \tag{3.6}$$

[1]Note that other motion models can also be used to describe the constraint between two adjacent positions as long as the motion model represents the fact the two adjacent positions are close to each other.

where $\mathbf{z}_{k-1,k}^{p-p}$ and $I_{k-1,k}^{p-p}$ denote the measurement and information matrices between positions $k-1$ and k for $k = 1 \dots K$. σ_{p-p} is the standard deviation of the random walk model within which the location of the next sound source should fall. I denotes the identity matrix.

Regarding position-landmark constraints, the measurement represents TDOA values at each position of sound source. Specifically, the measurement is defined as

$$\mathbf{z}_k^{p-l^T} = (TDOA_{mic\_2,mic\_1} \dots TDOA_{mic\_N,mic\_1}) \tag{3.7}$$

where $TDOA_{mic\_n,mic\_1}$ for $n = 2 \dots N$ is the TDOA between microphone n and microphone 1, which is used as the reference as mentioned before. The information matrices for this position-landmark constraint is given by

$$I_k^{p-l} = \frac{1}{\sigma_{p-l}^2} I\,, \tag{3.8}$$

where σ_{p-l} is the standard deviation of Gaussian distribution within which the error of each TDOA measurement should be.

3.2.2 Graph-Based Optimisation

In our case, as in the least square problem of the graph-based SLAM, the optimal state vector is found by minimising the error over all position-position constraints and position-landmarks constraints [130]

$$\mathbf{x}^* = argmin \sum_{ij} e_{ij}^T \Omega_{ij} e_{ij} \tag{3.9}$$

where i and j mean ith and jth nodes in the graph.

This estimated $\mathbf{x}^*$ can be obtained by iterative Gauss-Newton optimisation [130].

$$\mathbf{x} = \mathbf{x} + \Delta\mathbf{x} \tag{3.10}$$

where

$$H\Delta\mathbf{x} = -b \tag{3.11}$$

where H and b are called coefficient matrices and coefficient vector respectively. These two coefficients are defined as follows[130],

$$\begin{aligned} \bar{b}_i^T &= \sum e_{ij}^T \Omega_{ij} A_{ij} \\ \bar{b}_j^T &= \sum e_{ij}^T \Omega_{ij} B_{ij} \end{aligned} \tag{3.12}$$

and

$$\begin{aligned} \bar{H}_{ii} &= \sum A_{ij}^T \Omega_{ij} A_{ij} \\ \bar{H}_{ij} &= \sum A_{ij}^T \Omega_{ij} B_{ij} \\ \bar{H}_{ji} &= \sum B_{ij}^T \Omega_{ij} A_{ij} \\ \bar{H}_{jj} &= \sum B_{ij}^T \Omega_{ij} B_{ij} \end{aligned} \tag{3.13}$$

where $\bar{b}_i$ and $\bar{b}_j$ are ith and jth element of the coefficient vector b. $\bar{H}_{ii}$,$\bar{H}_{ij}$,$\bar{H}_{ji}$ and $\bar{H}_{jj}$ are sub block matrices parts of the coefficient matrices H. A_{ij} and B_{ij} are the Jacobian matrices of e_{ij} over the graph node $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively

$$\begin{aligned} A_{ij} &= \frac{\partial e(\mathbf{x}_i, \mathbf{x}_j)}{\partial \mathbf{x}_i} \\ B_{ij} &= \frac{\partial e(\mathbf{x}_i, \mathbf{x}_j)}{\partial \mathbf{x}_j} . \end{aligned} \tag{3.14}$$

In particular for the asynchronous microphone array, the calibration problem is computed as shown below. For each position-position constraint of the sound source at t_{k-1} to t_k, the error function is computed as

$$\begin{aligned} e_{k-1,k}^{p-p} &= (\mathbf{p}_k - \mathbf{p}_{k-1}) - \mathbf{z}_{k-1,k}^{p-p} \\ &= \begin{bmatrix} p_k^x - p_{k-1}^x \\ p_k^y - p_{k-1}^y \end{bmatrix} . \end{aligned} \tag{3.15}$$

Then, Jacobian matrices for this position-position constraint are

$$A_{k-1,k}^{p-p} = \frac{\partial e_{k-1,k}^{p-p}}{\partial \mathbf{p}_{k-1}} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \tag{3.16}$$

and

$$B_{k-1,k}^{p-p} = \frac{\partial e_{k-1,k}^{p-p}}{\partial \mathbf{p}_{k}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{3.17}$$

Regarding the position-landmark constraint of the microphone array to the sound source location at t_k, the error function is defined as

$$\begin{aligned} e_k^{p-l} &= \hat{\mathbf{z}}_k^{p-l} - \mathbf{z}_k^{p-l} \\ &= \begin{bmatrix} \frac{\sqrt{(x_{mic\_2}^x - p_k^x)^2 + (x_{mic\_2}^y - p_k^y)^2}}{c} \\ \vdots \\ \frac{\sqrt{(x_{mic\_n}^x - p_k^x)^2 + (x_{mic\_n}^y - p_k^y)^2}}{c} \end{bmatrix} \\ &- \begin{bmatrix} \frac{\sqrt{p_k^{x2} + p_k^{y2}}}{c} \\ \vdots \\ \frac{\sqrt{p_k^{x2} + p_k^{y2}}}{c} \end{bmatrix} + \begin{bmatrix} x_{mic\_2}^{\tau} \\ \vdots \\ x_{mic\_n}^{\tau} \end{bmatrix} \\ &+ k\Delta t \begin{bmatrix} x_{mic\_2}^{\delta} \\ \vdots \\ x_{mic\_n}^{\delta} \end{bmatrix} - \mathbf{z}_k^{p-l} \end{aligned} \tag{3.18}$$

where $\hat{\mathbf{z}}_k^{p-l}$ denotes the predicted TDOA from the current state vector, c refers to speed of sound and Δt means the time interval between each sound source position.

Then, the Jacobian matrices for this error function result in

$$A_k^{p-l} = \frac{\partial e_k^{p-l}}{\partial \mathbf{x}_{mic}} = [\mathbf{0} J_{mic\_2}^{p-l} \dots J_{mic\_N}^{p-l}] \tag{3.19}$$

and

$$B_k^{p-l} = \frac{\partial e_k^{p-l}}{\partial \mathbf{p}_k} = [J_{k\_x}^{p-l} J_{k\_y}^{p-l}] \tag{3.20}$$

where $J_{mic\_n}^{p-l}$ for $n = 1 \dots N$ is the partial derivative of e_k^{p-l} with respect to the state of the microphone n. Since the microphone 1 is used as a reference, its state is a constant value. Thus, the Jacobian is equal to zero. $J_{mic\_n}^{p-l}$ is only nonzero at row n and this nonzero row is computed as

$$J_{mic\_n}^{p-l}(n,:) = \begin{bmatrix} \frac{x_{mic\_2}^x - p_k^x}{c\sqrt{(x_{mic\_2}^x - p_k^x)^2 + (x_{mic\_2}^y - p_k^y)^2}} \\ \frac{x_{mic\_2}^y - p_k^y}{c\sqrt{(x_{mic\_2}^x - p_k^x)^2 + (x_{mic\_2}^y - p_k^y)^2}} \\ 1 \\ k\Delta t \end{bmatrix}^T \tag{3.21}$$

$J_{k\_x}^{p-l}$ and $J_{k\_y}^{p-l}$ in Eq. 3.20 represent the Jacobian matrices of e_k^{p-l} with respect to the state of x and y locations of the sound source at time t_k respectively. These matrices are computed as

$$\begin{aligned} J_{k\_x}^{p-l} = & \begin{bmatrix} \frac{p_k^x - x_{mic\_2}^x}{c\sqrt{(x_{mic\_2}^x - p_k^x)^2 + (x_{mic\_2}^y - p_k^y)^2}} \\ \vdots \\ \frac{p_k^x - x_{mic\_N}^x}{c\sqrt{(x_{mic\_N}^x - p_k^x)^2 + (x_{mic\_N}^y - p_k^y)^2}} \end{bmatrix} \\ & - \begin{bmatrix} \frac{p_k^x}{c\sqrt{p_k^{x2} + p_k^{y2}}} \\ \vdots \\ \frac{p_k^x}{c\sqrt{p_k^{x2} + p_k^{y2}}} \end{bmatrix} \end{aligned} \tag{3.22}$$

$$J_{k\_y}^{p-l} = \begin{bmatrix} \frac{p_k^y - x_{mic\_2}^y}{c\sqrt{(x_{mic\_2}^x - p_k^x)^2 + (x_{mic\_2}^y - p_k^y)^2}} \\ \vdots \\ \frac{p_k^y - x_{mic\_N}^y}{c\sqrt{(x_{mic\_N}^x - p_k^x)^2 + (x_{mic\_N}^y - p_k^y)^2}} \end{bmatrix} - \begin{bmatrix} \frac{p_k^y}{c\sqrt{p_k^{x2} + p_k^{y2}}} \\ \vdots \\ \frac{p_k^y}{c\sqrt{p_k^{x2} + p_k^{y2}}} \end{bmatrix} \tag{3.23}$$

Finally, the block corresponding to the 1st microphone in H is set to identity matrices,

$$H(1:4, 1:4) = I\,. \tag{3.24}$$

3.3 Calibration of a 3D Asynchronous Microphone Array

When a 3D microphone is considered, the system described above for a 2D microphone array is subjected to the following changes.

In order to fix the reference frame in 3D scenario, in addition to the constraints of microphones positions defined in 2D case, the z coordinate of a third microphone needs be set to zero.

Firstly, the state of the microphone array in Eq. 3.3 needs to include the z coordinate of the microphone as follows,

$$\mathbf{x}_{mic\_n}^T = (x_{mic\_n}^x x_{mic\_n}^y x_{mic\_n}^z x_{mic\_n}^\tau x_{mic\_n}^\delta). \tag{3.25}$$

Similarly, the state of the sound source at time instance k in Eq. 3.4 also includes its z coordinate as follows,

$$\mathbf{p}_k^T = (p_k^x p_k^y p_k^z). \tag{3.26}$$

The z coordinates of microphone positions and sound sources come into play when computing sound sources positions and the TDOA. Therefore, for each position-position constraint of the sound source, the error function in Eq. 3.15 becomes as follows,

$$\begin{aligned} e^{p-p}_{k-1,k} &= (\mathbf{p}_k - \mathbf{p}_{k-1}) - \mathbf{z}^{p-p}_{k-1,k} \\ &= \begin{bmatrix} p^x_k - p^x_{k-1} \\ p^y_k - p^y_{k-1} \\ p^z_k - p^z_{k-1} \end{bmatrix}. \end{aligned} \tag{3.27}$$

and its corresponding Jacobian matrix in Eq. 3.16 and Eq. 3.17 can be rewritten as follows,

$$A^{p-p}_{k-1,k} = \frac{\partial e^{p-p}_{k-1,k}}{\partial \mathbf{p}_{k-1}} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \tag{3.28}$$

$$B^{p-p}_{k-1,k} = \frac{\partial e^{p-p}_{k-1,k}}{\partial \mathbf{p}_{k}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.29}$$

Regarding the position-landmark constraint of the 3D microphone array to the sound source location at t_k, the error function in Eq. 3.18 can be rewritten as follows,

$$\begin{aligned} e^{p-l}_k &= \hat{\mathbf{z}}^{p-l}_k - \mathbf{z}^{p-l}_k \\ &= \begin{bmatrix} \frac{d_{2,k}}{c} \\ \vdots \\ \frac{d_{n,k}}{c} \end{bmatrix} - \begin{bmatrix} \frac{\sqrt{p^{x2}_k + p^{y2}_k}}{c} \\ \vdots \\ \frac{\sqrt{p^{x2}_k + p^{y2}_k}}{c} \end{bmatrix} + \begin{bmatrix} x^{\tau}_{mic\_2} \\ \vdots \\ x^{\tau}_{mic\_n} \end{bmatrix} \\ &\quad + k\Delta t \begin{bmatrix} x^{\delta}_{mic\_2} \\ \vdots \\ x^{\delta}_{mic\_n} \end{bmatrix} - \mathbf{z}^{p-l}_k, \end{aligned} \tag{3.30}$$

where $d_{i,k}$, $i \in (1 \ldots n)$ is the distance between the ith microphone and the sound source at time instance k, which can be formulated as follows,

$$\begin{aligned} d_{i,k} =&((x^x_{mic\_2} - p^x_k)^2 + (x^y_{mic\_2} - p^y_k)^2 \\ &+ (x^z_{mic\_2} - p^z_k)^2)^{1/2}. \end{aligned} \tag{3.31}$$

The Jacobian matrices A^{p-l}_k for this error function has the same structure as in Eq. 3.19 while B^{p-l}_k in Eq. 3.20 can be rewritten as follows,

$$B^{p-l}_k = \frac{\partial e^{p-l}_k}{\partial \mathbf{p}_k} = [J^{p-l}_{k\_x} J^{p-l}_{k\_y} J^{p-l}_{k\_z}]. \tag{3.32}$$

$J^{p-l}_{mic\_n}$ is Eq. 3.21 is rewritten as

$$J^{p-l}_{mic\_n}(n,:) = \begin{bmatrix} \frac{x^x_{mic\_2} - p^x_k}{cd_{n,k}} \\ \frac{x^y_{mic\_2} - p^y_k}{cd_{n,k}} \\ \frac{x^z_{mic\_2} - p^z_k}{cd_{n,k}} \\ 1 \\ k\Delta t \end{bmatrix}^T . \tag{3.33}$$

$J^{p-l}_{k\_x}$, $J^{p-l}_{k\_y}$ and $J^{p-l}_{k\_z}$ in Eq. 3.32 can be formulated as follows,

$$J^{p-l}_{k\_x} = \begin{bmatrix} \frac{p^x_k - x^x_{mic\_2}}{cd_{n,k}} \\ \vdots \\ \frac{p^x_k - x^x_{mic\_N}}{cd_{n,k}} \end{bmatrix} - \begin{bmatrix} \frac{p^x_k}{cd_k} \\ \vdots \\ \frac{p^x_k}{cd_k} \end{bmatrix}, \tag{3.34}$$

$$J^{p-l}_{k\_y} = \begin{bmatrix} \frac{p^y_k - x^y_{mic\_2}}{cd_{n,k}} \\ \vdots \\ \frac{p^y_k - x^y_{mic\_N}}{cd_{n,k}} \end{bmatrix} - \begin{bmatrix} \frac{p^y_k}{cd_k} \\ \vdots \\ \frac{p^y_k}{cd_k} \end{bmatrix}, \tag{3.35}$$

$$J_{k\_z}^{p-l} = \begin{bmatrix} \frac{p_k^z - x_{mic\_2}^z}{cd_{n,k}} \\ \vdots \\ \frac{p_k^z - x_{mic\_N}^z}{cd_{n,k}} \end{bmatrix} - \begin{bmatrix} \frac{p_k^z}{cd_k} \\ \vdots \\ \frac{p_k^z}{cd_k} \end{bmatrix}, \tag{3.36}$$

where d_k is the distance from the sound source position at the kth time instance to the origin of the global coordinate frame, which is formulated as follows,

$$d_k = \sqrt{p_k^{x2} + p_k^{y2} + p_k^{z2}} \tag{3.37}$$

3.4 Calibration of an Asynchronous Linear Microphone Array

In this section, details about the calibration of an asynchronous linear microphone array is presented[2].

For an embedded linear microphone array, since the location information of each microphone channel (distance between each microphone) can be easily obtained by the fabrication data or measured by the user, it is assumed to be known here. Then, the only information to be estimated for a microphone is the starting time offset and clock difference.

3.4.1 System Model

Let $\mathbf{x}_{mic}$ is the state of the microphone array and β is the state of the sound source. Then, the state vector of the SLAM framework is as follow,

$$\mathbf{x}^T = (\mathbf{x}_{mic}^T \beta^T). \tag{3.38}$$

[2]For those who have already got a commercial 3D camera with an embedded synchronised linear microphone array as shown if Fig. 5.1, this section can be skipped.

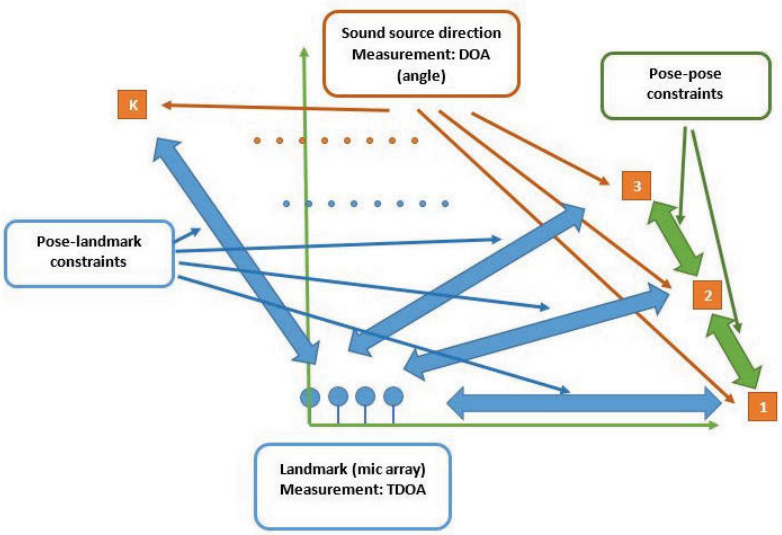


FIGURE 3.4: Description of the poses, landmarks and constraints.

The state of the microphone array $\mathbf{x}_{mic}$ and the state vector of the sound source β have the following structures,

$$\mathbf{x}_{mic}^T = (\mathbf{x}_{mic\_1}^T \dots \mathbf{x}_{mic\_N}^T), \tag{3.39}$$

$$\beta^T = (\beta_1^T \dots \beta_K^T), \tag{3.40}$$

where N is the total number of microphones and K is total number of sound source positions used in the calibration while it moves around the microphones. $\beta_k(k = 1 \dots K)$ is the DOA angle of the sound source at time instance k. $\mathbf{x}_{mic\_n}(n = 1 \dots N)$ is the individual state of each microphone and it is parametrised as follows,

$$\mathbf{x}_{mic\_n}^T = (x_{mic\_n}^{\tau} x_{mic\_n}^{\delta})(n = 1 \dots N), \tag{3.41}$$

where $x_{mic\_n}^{\tau}$ is the starting time offset and $x_{mic\_n}^{\delta}$ is the clock difference per second of microphone n. The first microphone is used as the reference and time offsets and clock differences of other microphones are computed relative to this microphone. Therefore, $x_{mic\_1}^{\tau} = 0$ and $x_{mic\_1}^{\delta} = 0$.

Similar to calibration of a 2D/3D microphone array, to make the analogy to a standard SLAM framework as shown in Fig. 3.4, DOA angles of the sound source are treated as robot poses and the microphone array is treated as a single landmark that is observed at all

sound source directions.. The microphone array becomes the first node in the graph-based SLAM framework and each sound source DOA angle becomes one node.

For the position-position constraints, similar to calibration of a 2D/3D microphone array, the trajectory of the sound source is assumed to be arbitrary (no-odometry prior is considered), and we use a random walk like model in which the sound source DOA angle of the next time instance is expected to be the same as that of the previous time with the standard deviation of reasonably big value as follows,

$$z_{k-1,k}^{p-p} = 0(k = 1 \dots K), \tag{3.42}$$

$$I_{k-1,k}^{p-p} = \frac{1}{\sigma_{p-p}^2}(k = 1 \dots K), \tag{3.43}$$

where $z_{k-1,k}^{p-p}$ and $I_{k-1,k}^{p-p}$ denote the measurement and information matrix between poses $k-1$ and k. σ_{p-p} is the standard deviation of the Gaussian distribution within which the DOA angle of next sound source should fall.

For the pose-landmark constraints, the measurement is the TDOA values of the sound source to the microphone array at each angle of sound source. The measurement of this pose-landmark constraint is defined as follows,

$$\mathbf{z}_k^{p-l^T} = (TDOA_{mic\_2,mic\_1} \dots TDOA_{mic\_N,mic\_1}), \tag{3.44}$$

where $TDOA_{mic\_n,mic\_1}(n = 2 \dots N)$ is the TDOA between microphone n and microphone 1 as microphone 1 is used as reference. The information matrix for this pose-landmark constraint is

$$I_k^{p-l} = \frac{1}{\sigma_{p-l}^2}\mathbf{I}, \tag{3.45}$$

where σ_{p-l} is the standard deviation of Gaussian distribution within which the error of TDOA observation error should be.

3.4.2 Error Functions and Their Jacobians

For each pose-pose constraint from sound source position $k-1$ to k, the error function is defined as follows,

$$e^{p-p}_{k-1,k} = (\beta_k - \beta_{k-1}) - z^{p-p}_{k-1,k} = \beta_k - \beta_{k-1}. \tag{3.46}$$

Therefore, the corresponding Jacobian matrix for this pose-pose constraint is as follow.

$$A^{p-p}_{k-1,k} = \frac{\partial e^{p-p}_{k-1,k}}{\partial \beta_{k-1}} = -1, \tag{3.47}$$

$$B^{p-p}_{k-1,k} = \frac{\partial e^{p-p}_{k-1,k}}{\partial \beta_{k}} = 1, \tag{3.48}$$

where $A^{p-p}_{k-1,k}$ and $B^{p-p}_{k-1,k}$ are Jacobian matrix of $e^{p-p}_{k-1,k}$ over the graph node β_{k-1} and β_k respectively.

For the pose-landmark constraint between microphone array and sound source location at time instance k, the error function is computed as follows,

$$\begin{aligned} e^{l-p}_k &= TDOA_{pre} - \mathbf{z}^{p-l}_k \\ &= -\frac{1}{c}\begin{bmatrix} d_{mic} \\ \vdots \\ (N-1) * d_{mic} \end{bmatrix} cos(\beta_k) + \begin{bmatrix} x^{\tau}_{mic\_2} \\ \vdots \\ x^{\tau}_{mic\_n} \end{bmatrix} + k\Delta t \begin{bmatrix} x^{\delta}_{mic\_2} \\ \vdots \\ x^{\delta}_{mic\_n} \end{bmatrix} - \mathbf{z}^{p-l}_k, \end{aligned} \tag{3.49}$$

where d_{mic} is the distance between two adjacent microphones, $TDOA_{pre}$ is the predicted TDOA value from the current state vector, c denotes the speed of sound and Δt refers to time interval between time instance k and $k+1$.

The Jacobian matrices correspond to this error function are computed as follows,

$$A^{l-p}_k = \frac{\partial e^{l-p}_k}{\partial \mathbf{x}_{mic}} = [\mathbf{0}\mathbf{J}^{l-p}_{mic\_2} \dots \mathbf{J}^{l-p}_{mic\_N}], \tag{3.50}$$

$$B^{l-p}_k = \frac{\partial e^{l-p}_k}{\partial \beta_k} = \frac{1}{c}\begin{bmatrix} d_{mic}\sin(\beta_2) \\ \vdots \\ (N-1)d_{mic}\sin(\beta_K) \end{bmatrix}, \tag{3.51}$$

where $\mathbf{J}_{mic\_n}^{l-p}, (n = 1 \dots N)$ is the Jacobian of e_k^{l-p} over state vector of microphone n. Since microphone 1 is used as reference, its state vector is a constant value and the Jacobian of it is all zero. $\mathbf{J}_{mic\_n}^{l-p}$ is only nonzero at rule n and this nonzero rule is computed as

$$\mathbf{J}_{mic\_n}^{l-p}(n,:) = \begin{bmatrix} 1 \\ k\Delta t \end{bmatrix}^T . \tag{3.52}$$

3.5 Simulation and Experimental Results

The validation of the proposed methodology is studied first in a simulation environment, where the performance of the proposed algorithm is tested under a variety of conditions with known ground truth. An experiment with a set of ordinary microphones was then conducted to show the effectiveness under realistic conditions.

3.5.1 Application Setup

For all the scenarios of this Chapter, a sound-source is considered to be moving randomly or following a pre-defined path around a room, where an array of microphones is fixed and recording. Then, recorded audio signals from all microphone channels are processed using the proposed graph-based optimisation method, and sound source positions and the locations, starting time offsets and clock differences of all microphones are estimated. Note that once the array is calibrated, popular synchronous microphone array processing techniques can be applied for separation or localisation of multiple sound sources as described previously in section 2.1.2.

3.5.2 Initialisation and Termination Conditions

No prior knowledge of the microphone positions or sound source trajectories are assumed as initial values for the optimisation. Therefore, the initial values corresponding to those variables are randomly selected within the workspace. Moreover, zero starting time offsets and zero clock differences are assumed as part of the state. Like any other optimisation problem, least square optimisation based graph SLAM also suffers from non-convergence

from a bad initial value, a situation that can be minimised if approximate priors can be supplied.

The termination condition is based on the maximum number of iterations and change of the state vector $\Delta\mathbf{x}$. If the algorithm reaches a predefined maximum number of iterations, or the change of the state vector is smaller than a predefined threshold ϵ, the algorithm stops.

3.5.3 Simulation Results

3.5.3.1 Simulation Results of 2D Microphone Arrays

The parameters used in 2D simulations are summarised in Table 3.1. The realistic observation noise and random walk model noise are multiplied by conservative factor to deal with the worst case scenarios. In this section, three different types of microphone arrays are simulated. They are an array of 9 microphones with 3×3 structure as shown in Fig. 3.5(b), an array of 6 microphones with 3×2 structure as shown in Fig. 3.7(a) and an array of 16 microphones with 4×4 structure as shown in Fig. 3.7(b).

In all simulations, in order to obtain unique solutions, the position of microphone 1 is fixed at the origin of the coordinate system, microphone 3 in the 3×3 microphone array (microphone 2 in the 3×2 microphone array and microphone 4 in the 4×4 microphone array) is fixed at positive x axis and the y coordinate of microphone 4 in the 3×3 microphone array (microphone 3 in the 3×2 microphone array and microphone 5 in the 4×4 microphone array) is set to be positive.

Firstly, we performed a 10-runs Monte Carlo simulation, which considers an array of 9 (3×3) microphones and sound source moving as (Eq. 3.5). The results of the 1st Monte Carlo run are shown in Fig. 3.5 and Fig. 3.6. From the figure, it can be seen that, despite random initialisation, the proposed method is able to converge with good accuracy to the simulated values.

Secondly, in order to show the influence of the number of microphones over the estimation accuracy, another two 10-runs Monte Carlo simulations for a 3×2 and a 4×4 arrays

TABLE 3.1: Parameters setting in simulation

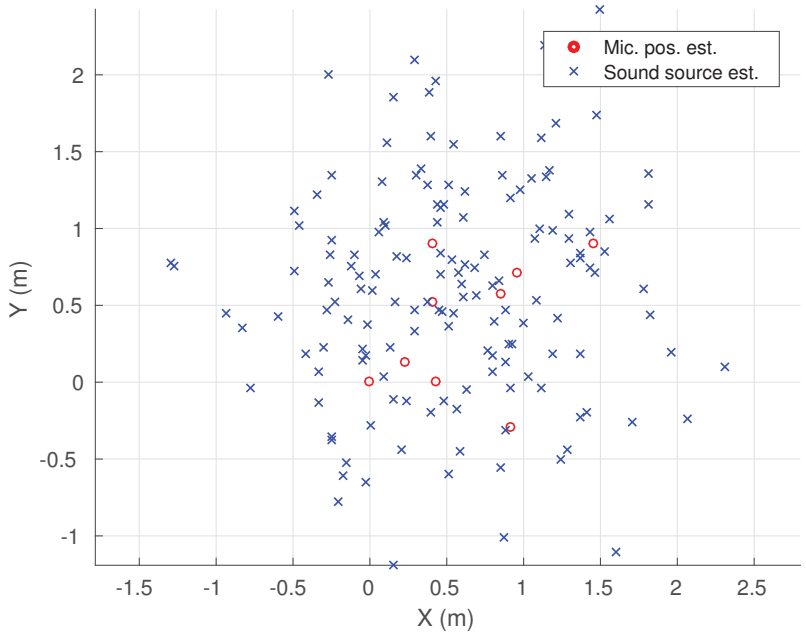
Parameters	Values
Number of microphones	9
Distance between microphones	0.5m
Maximum starting time offset	0.1s
Maximum clock difference	0.1ms
Observation (TDOA) noise STD	0.333ms
Sampling frequency	44.1 KHz
random walk STD	0.333m
Adjacent sound source distance	0.05m
Maximum iterations	50
ϵ for $\Delta\mathbf{x}$	0.0001

TABLE 3.2: RMS errors over 10-run Monte Carlo simulations

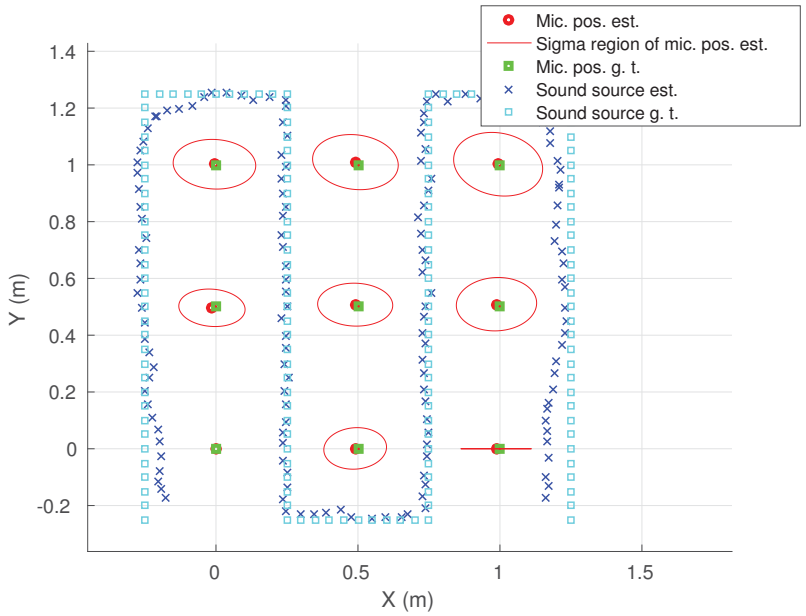
Arrangement	3×2	3×3	4×4
mean RMS error of mic. pos.(m)	0.1207	0.0335	0.0103
mean RMS error of τ(ms)	0.2486	0.0812	0.0276
mean RMS error of δ(micro s)	1.6446	0.6952	0.1936

are performed. The comparison of the root mean square (RMS) errors for microphone positions in the 3×2, 3×3 and 4×4 arrangements is given in Table 3.2. From the RMS errors, it can be seen that increasing the number of microphones results in better estimation accuracy of the microphone position and the usage of 9 microphones is sufficient, under the simulated TDOA observation noise, to recover the trajectory of the sound source with low RMS error.

Finally, to test the influence of the number of sound source positions over the estimation accuracy, another two 10-runs Monte Carlo simulations with half of the sound source positions and twice the number of the sound source positions are performed using the 3×3 microphone array. The results are shown in Fig. 3.8. The comparison of RMS errors of the microphone positions is given in Table 3.3. The figure and table show that an increased number of sound source positions can lead to better estimation accuracy.



a Initialisation of the state vector.

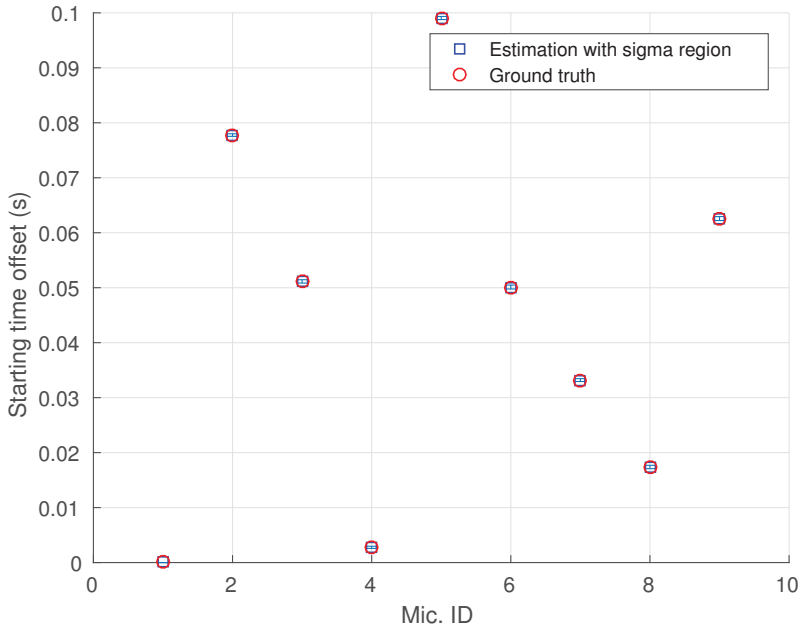


b Final estimation results for microphone and sound source positions after convergence over 17 iterations.

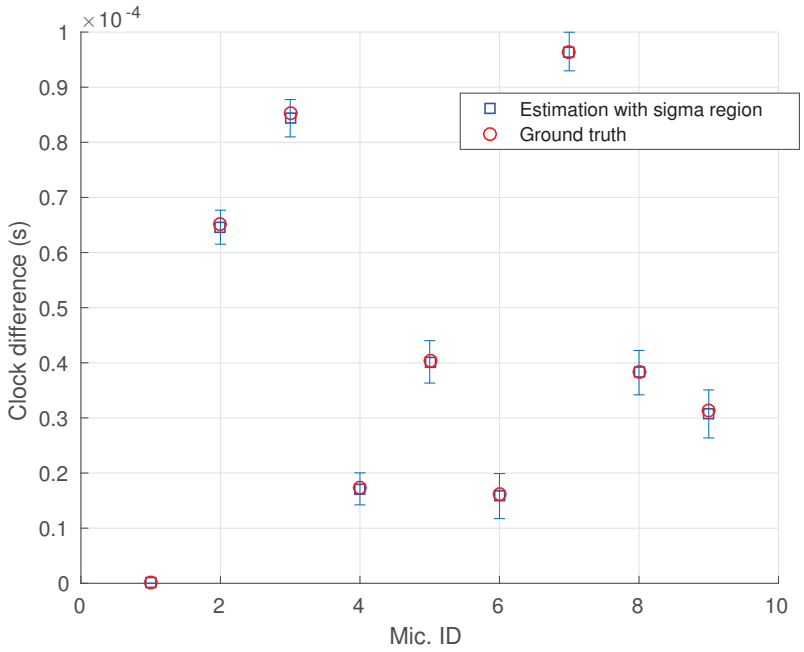
FIGURE 3.5: Initialisation and final estimation results for a 3×3 array.

3.5.3.2 Simulation Results of 3D Microphone Arrays

In this section, calibration of a 3D microphone array simulation scenario is considered. Key simulation parameters are the same as those in 2D simulation as shown in Table 3.1. Two



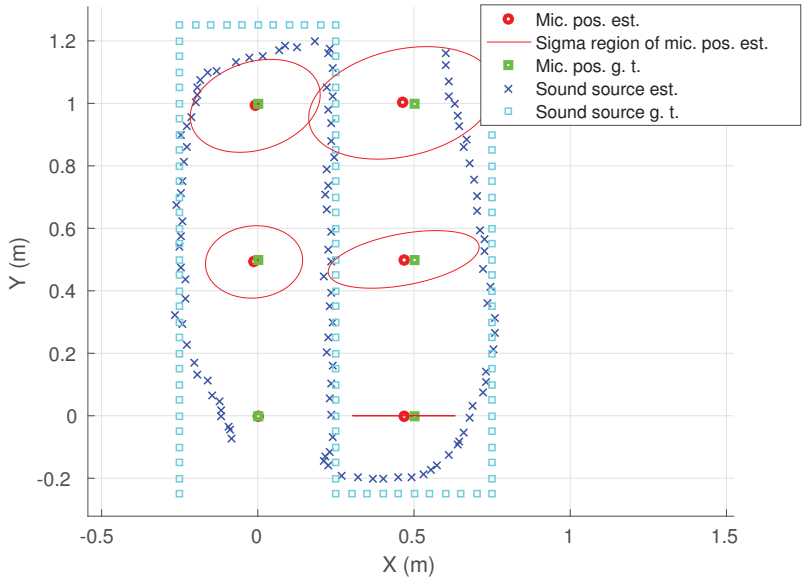
a Final estimation results for starting time offset.



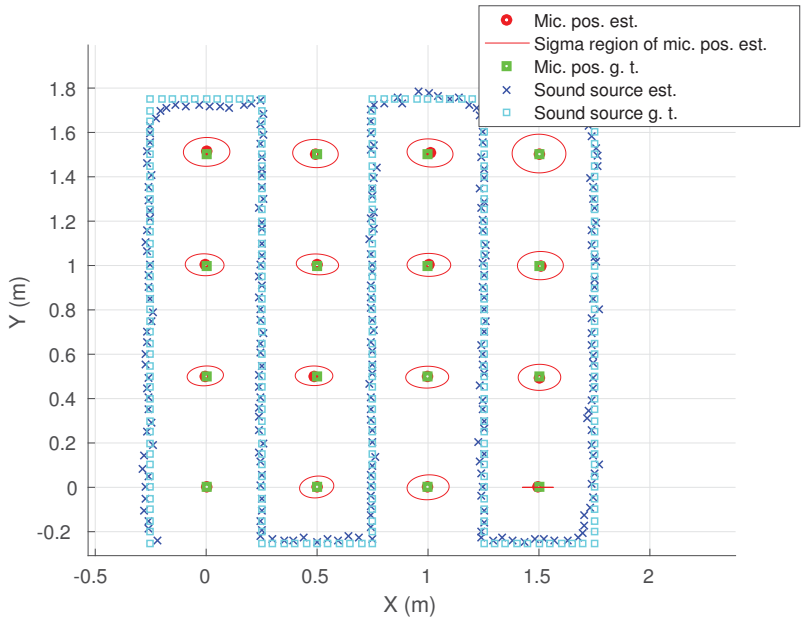
b Final estimation results for clock difference.

FIGURE 3.6: Final estimation results for a 3×3 array.

sets of simulations of a $3 \times 3 \times 2$ microphone array, as shown in Fig. 3.9, are performed. In all simulations of the 3D microphone array, in order to obtain unique solutions, the position of microphone 1 is fixed at the origin of the coordinate system, microphone 3 is fixed at positive x axis, the z coordinate of microphone 7 is set to be zero and z coordinate



a Estimation results for one of 3×2 microphone array.

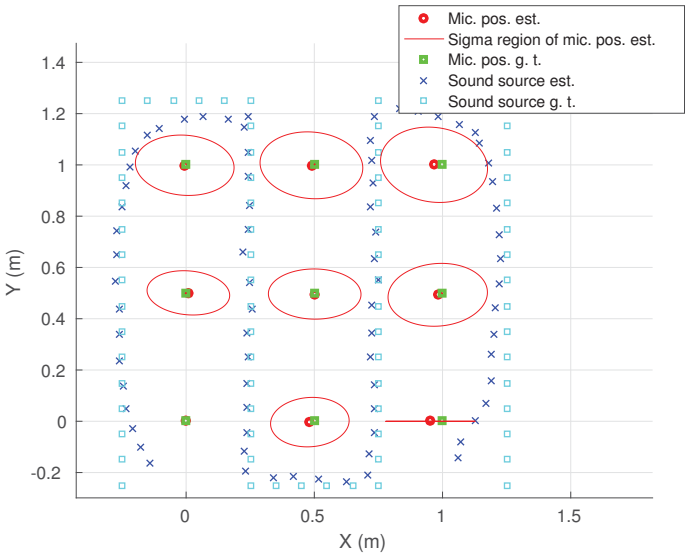


b Estimation results for one of 4×4 microphone array.

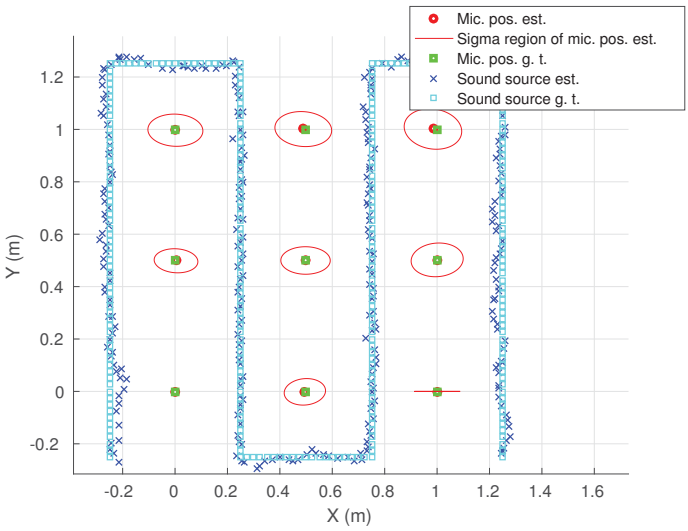
FIGURE 3.7: Estimation results of 3×2 and 4×4 arrays.

of the microphone 10 is set to positive.

Firstly, we simulated a situation in which the sound source can move through the internal space of the microphone array. The microphone array state and locations of the sound source are randomly initialised as shown in Fig. 3.9(a), final estimation of the sound source positions and microphones locations are shown in Fig. 3.9(b), and starting time offsets and



a Estimation results of one of a 3×3 microphone array with half of the sound source positions.



b Estimation results of one of a 3×3 microphone array with twice of the sound source positions.

FIGURE 3.8: Estimation results of various number of sound source positions.

clock differences are shown in Fig. 3.10(a) and Fig. 3.10(b). From these results, it can be seen that, the proposed method can successfully calibrate a 3D microphone array and localise a sound source in 3D space.

Next, we simulated another situation in which the sound source can only access the exterior space of the microphone array as shown in Fig. 3.11 and Fig. 3.12. These simulations reflect the scenario in which the casing of the microphone array does not allow the sound source

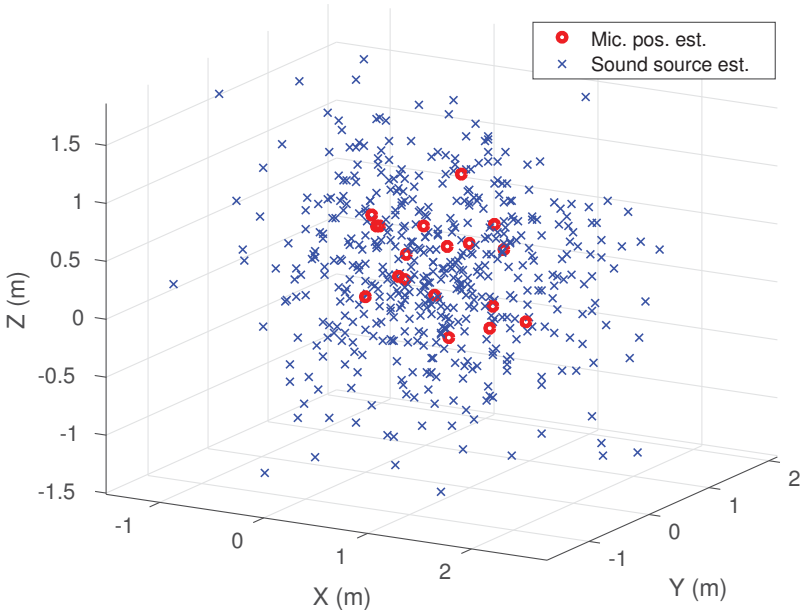
TABLE 3.3: RMS errors over 10-run Monte Carlo simulations

number of sound source positions	half	original	twice
mean RMS error of mic. pos.(m)	0.1480	0.0335	0.0106
mean RMS error of τ(ms)	0.3038	0.0812	0.0239
mean RMS error of δ(micro s)	2.4248	0.6952	0.1450

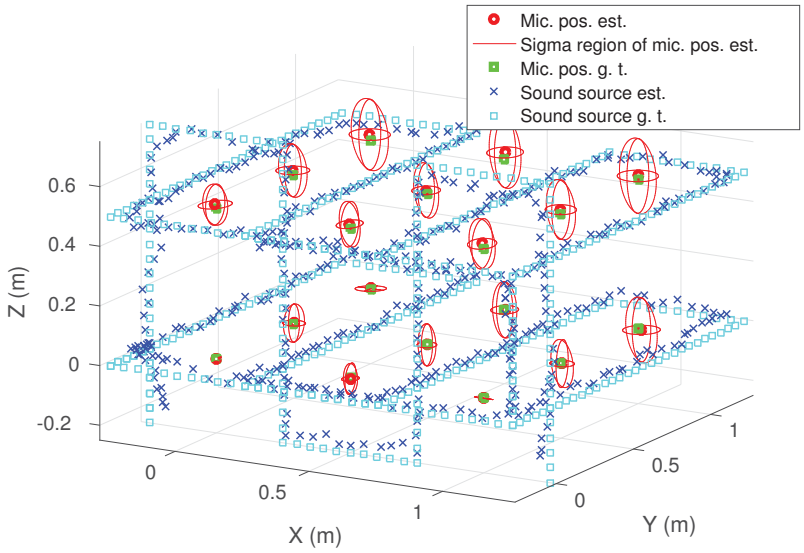
to be inside the space of the microphone array. The results prove the sound performance of the method in this simulation scenario. From the experimental result it also can be seen that the uncertainty associated to the z coordinates of the microphones are smaller than those of the previous simulation. This is due to the fact that the sound source travels a longer distance along the z axis (-0.75m to 1.25m) compared to that in the previous simulation (-0.25m to 0.75m).

3.5.3.3 Simulations of Calibration of 2D and 3D Asynchronous Microphone Arrays without Estimating Clock Difference

The conventional method [126] for calibration of an asynchronous microphone array disregards the clock differences between the microphone sound cards. However, according to our own experiments (shown before in Fig. 3.1 and Fig. 3.2), clock timing in different sound cards does present marginal differences. Therefore, in order to test the relevance of the proposed clock difference estimation, which is one of the main novelties in the thesis, in this section the simulation results of a 2D and 3D asynchronous microphone array calibration without the clock difference estimation are presented. We use the same 2D and 3D microphone arrays as simulated before in Fig. 3.5 and Fig. 3.11, and the same simulation parameters and initialisation values. Unsurprisingly, when difference in clock timings are ignored, the optimisation results of both 2D and 3D microphone arrays start to diverge after only a few iterations - as depicted in Fig. 3.13. This is due to the fact that after a certain time period (more than 2 minutes in both simulation scenarios), the time offsets between different channels of microphones have changed considerably given the clock differences. Therefore, assuming a constant time offsets by disregarding clock differences have resulted in the divergence from the optimal solution. The reason why in [126] the calibration problem converged even when the clock differences were ignored is



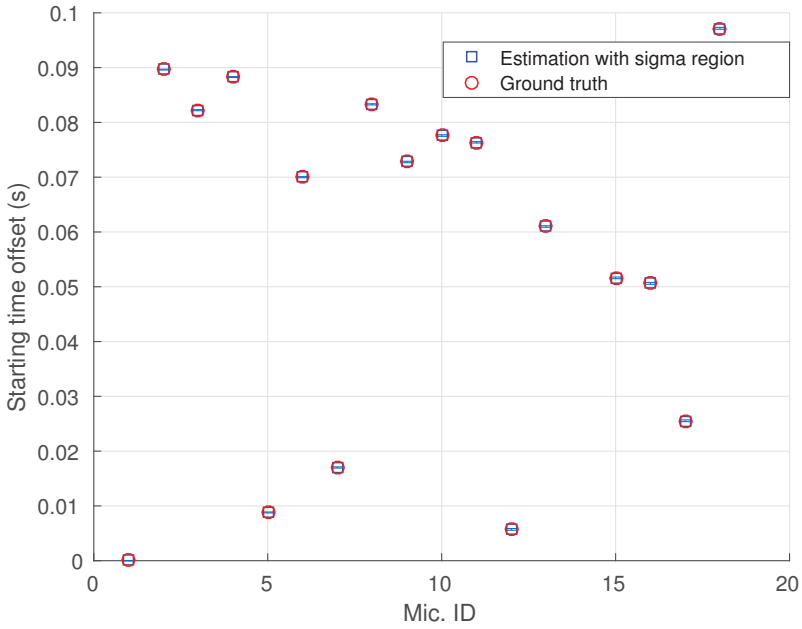
a Initialisation of the state vector.



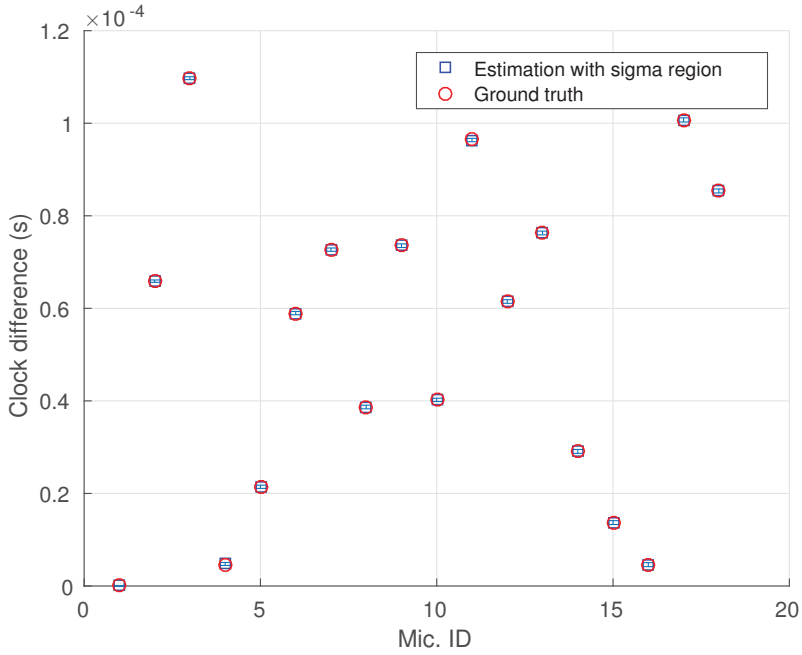
b Final estimation results for microphone and sound source positions after convergence over 17 iterations.

FIGURE 3.9: Initialisation and final estimation results for a $3 \times 3 \times 2$ array.

explained by the fact that they used a synchronised microphone array for the experiment, and manually added temporal offsets. Therefore, there is no clock difference in their data.

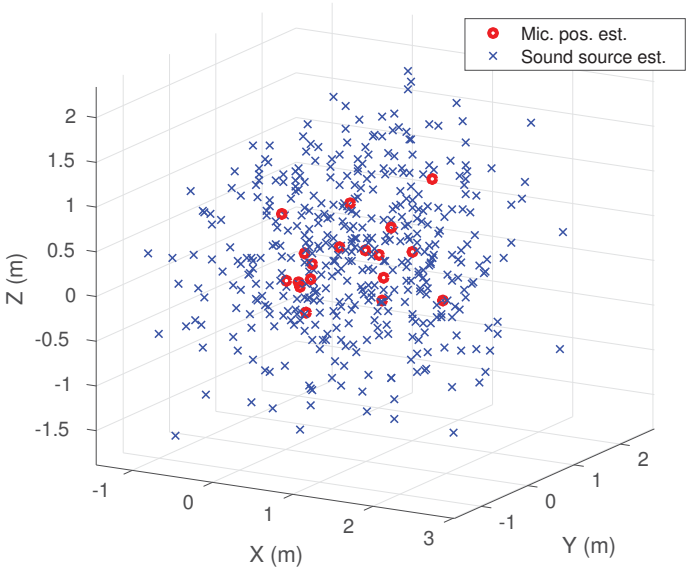


a Final estimation results for starting time offset.

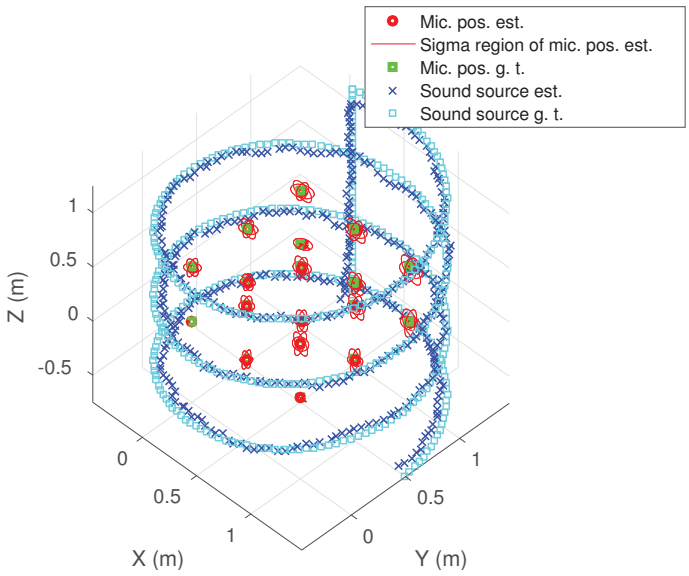


b Final estimation results for clock difference.

FIGURE 3.10: Final estimation results for a $3 \times 3 \times 2$ array.



a Initialisation of the state vector.

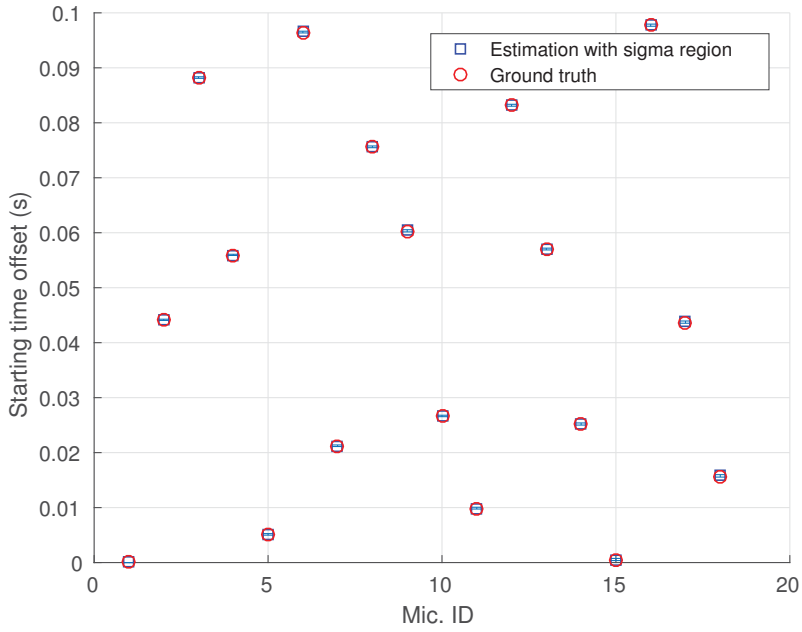


b Final estimation results for microphone and sound source positions after convergence over 17 iterations.

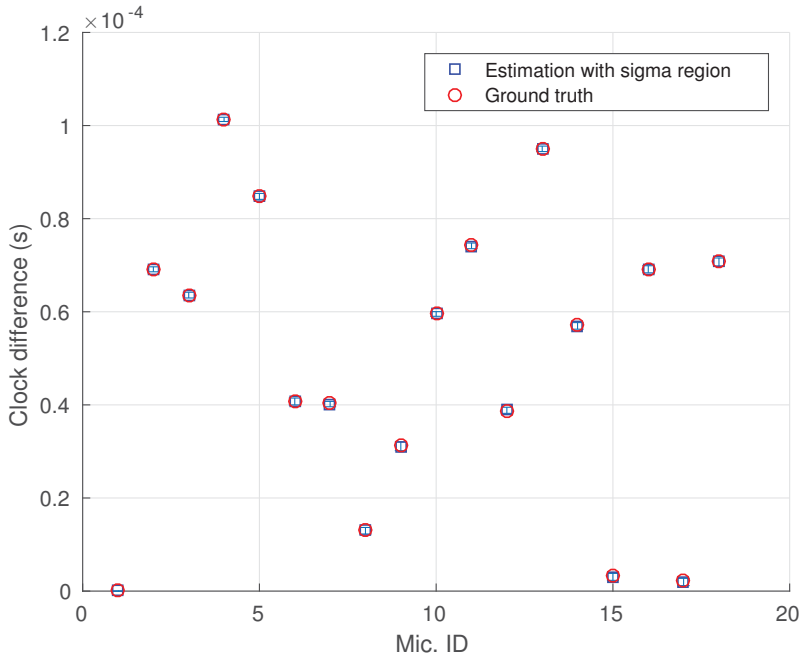
FIGURE 3.11: Initialisation and final estimation results for a $3 \times 3 \times 2$ array.

3.5.3.4 Simulations of Calibration of an Asynchronous Linear Microphone Array

A simulation of the proposed method with an array of 4 microphones is studied and the parameters used in the simulation are summarised in Table 3.4. The estimations of the starting time offset and the clock difference of each microphone channel are shown in Fig.



a Final estimation results for starting time offset.

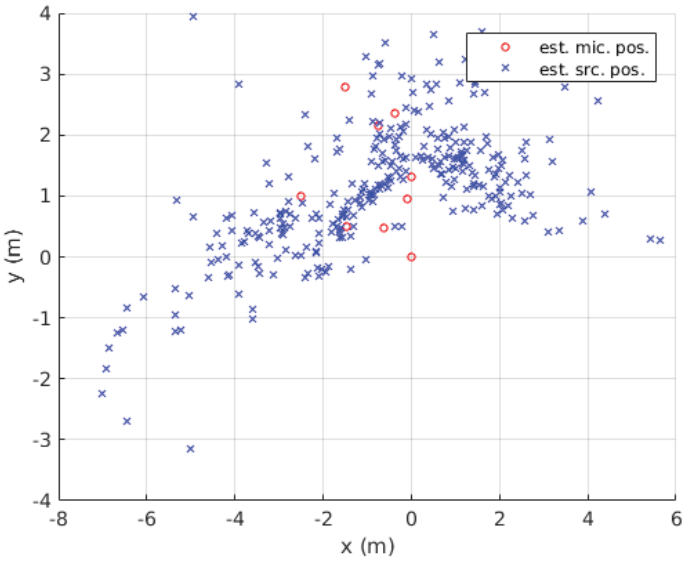


b Final estimation results for clock difference.

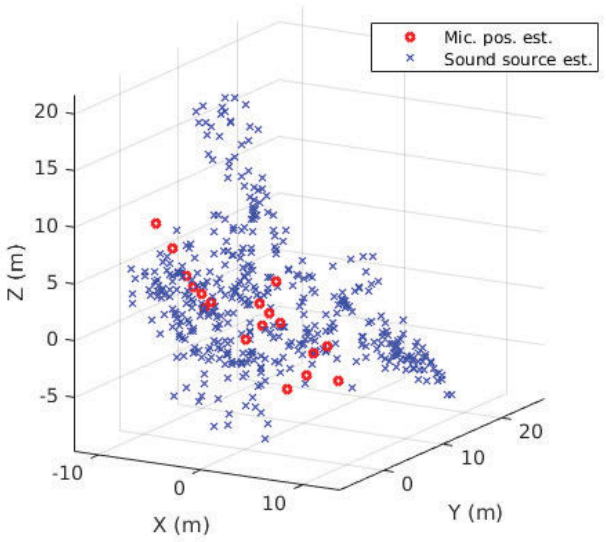
FIGURE 3.12: Final estimation results for a $3 \times 3 \times 2$ array.

3.14. It can be seen from the figure that our method is able to estimate the values for these two variables with good accuracy.

Secondly, the influence of the number of calibration data over the estimation accuracy is



a Calibration of the 2D microphone array in Fig. 3.5 diverges when ignoring the clock difference.



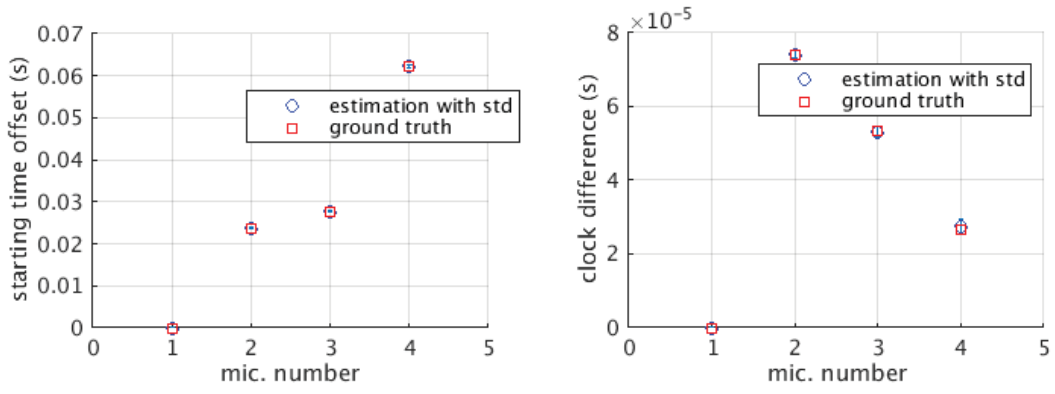
b Calibration of the 3D microphone array in Fig. 3.11 diverges when ignoring the clock difference.

FIGURE 3.13: Calibration of 2D and 3D microphone arrays diverges when ignoring the clock difference.

examined by simulations of various numbers of it with 20 Monte Carlo runs for each case. The results are shown in the Fig. 3.15. From the comparison of mean RMS errors plot, it is clear that a lower number of calibration data (smaller than 180) can dramatically degrade the estimation accuracy of both starting time offsets and clock differences.

TABLE 3.4: Parameters setting in simulation

Parameters	Values
Number of microphones	4
Distance between microphones	0.05m
Maximum starting time offset	0.1s
Maximum clock difference	0.1ms
Observation (TDOA) noise STD	3.33ms
Sampling frequency	44.1 KHz
random walk STD	50 (degree)
Adjacent sound source angle	1 (degree)
Number of calibration data	180
Speed of sound	340m/s
Maximum iterations	20
ϵ for ΔX	10



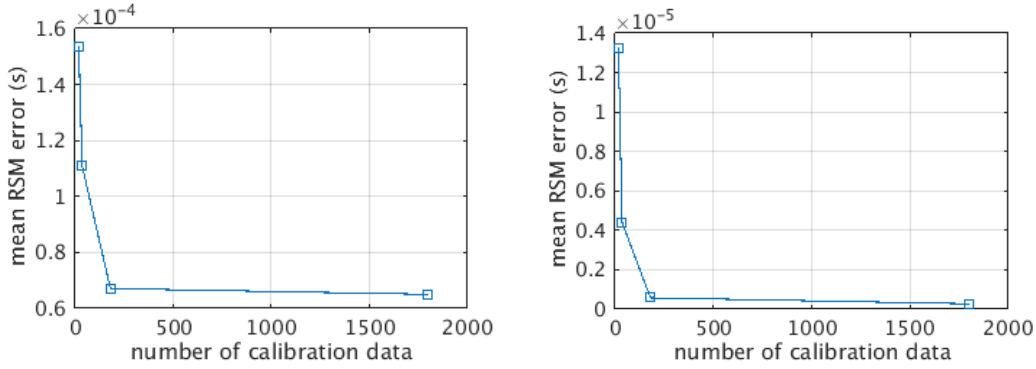
a starting time offset

b clock difference

FIGURE 3.14: Simulation results compared to the ground truth values.

3.5.4 Experimental Results

To validate the proposed methodology, the following experimental set-up in an indoor setting was devised: an array of 6 microphones was fixed at a known location as shown in Fig. 3.16. These microphones were individually sampled by independent USB sound cards. Relevant parameters of the experimental set-up are summarised in Table 3.5. The observation noise and random walk model noise were empirically obtained. Again, in order to obtain a unique solution, the position of microphone 1 is fixed at the origin of the coordinate system, microphone 3 is fixed at positive x axis and the y coordinate of microphone 4 is set to be positive.



a starting time offset b clock difference

FIGURE 3.15: Mean RMS error w.r.t. number of calibration data.

FIGURE 3.16: Experimental setup of the asynchronous microphone array. Each channel of the array is sampled independently using individual USB sound cards.

When recording of an incoming sound signal (a short time chirp) commences, a hand-held sound emitter (a smart phone producing a known sound wave) moves around the microphone array following one similar and one different trajectories to those in simulations.

3.5.4.1 Signal Processing

The raw audio recording contains background noise and reverberation as shown in Fig. 3.17. An Equiripple high pass filter was used to clean the low frequency noise with a frequency lower than the lowest frequency of the emitted chirp signal. The first distinctive peak of the filtered wave was chosen as the arrival time of the signal. Note that any other sound signals (*e.g.* hand clapping and continuous speech) can be used as the sound source.

TABLE 3.5: Experimental set-up parameters

Parameters	Values
Number of microphones	6
Distance between microphones	0.5m
Observation (TDOA) noise STD assumed	0.167ms
Sampling frequency	44.1 KHz
random walk STD	0.167m
sound source	Samsung Galaxy S4 phone
sound wave	short time chirp
time interval of sound	0.5s
total duration of recording	1min
Maximum iterations	50
ϵ for $\mathbf{\Delta x}$	0.0001

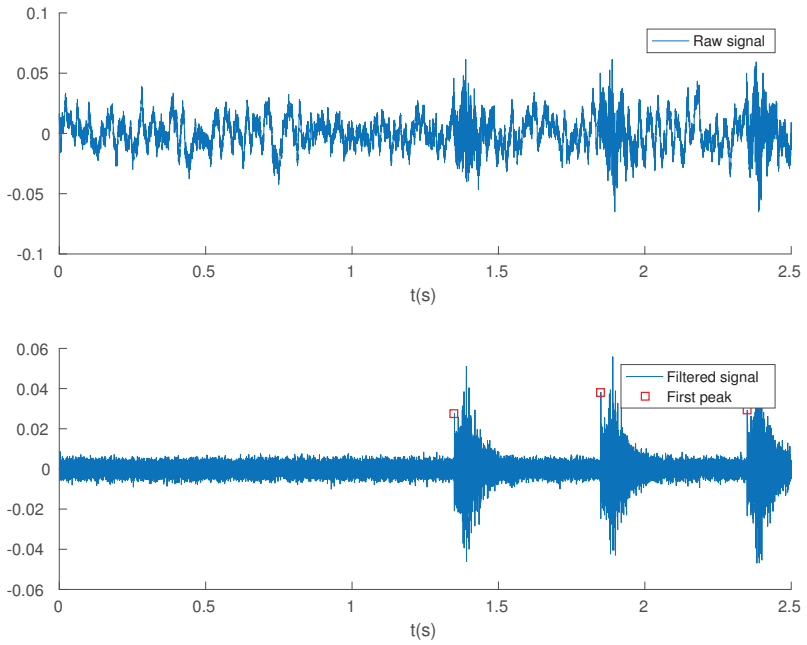


FIGURE 3.17: Pre signal processing and detection of signal arrival (plot below) for raw audio data (plot above).

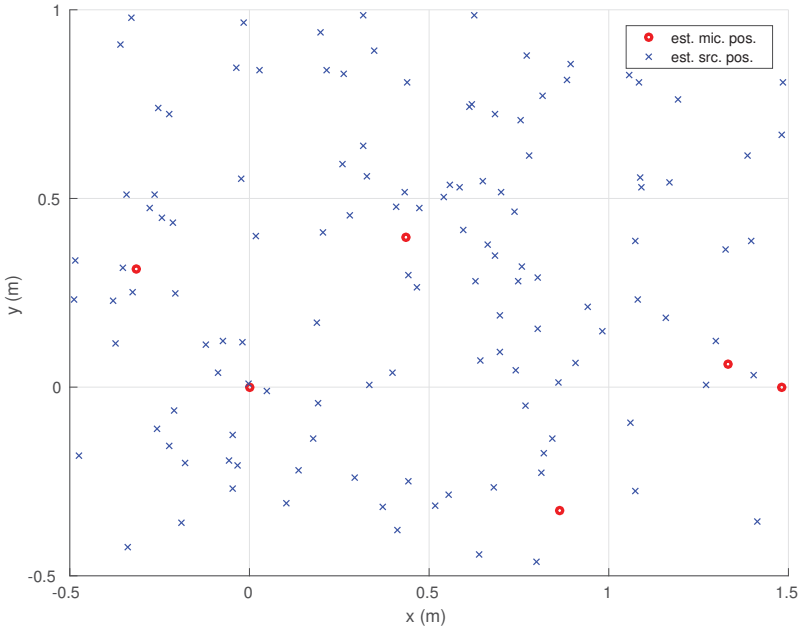
Moreover, any other signal processing method (like GCC-PHAT [33]) for obtaining TDOA can be also adopted as long as its noise is properly characterised.

3.5.4.2 Experimental Results of a 2D Microphone Array

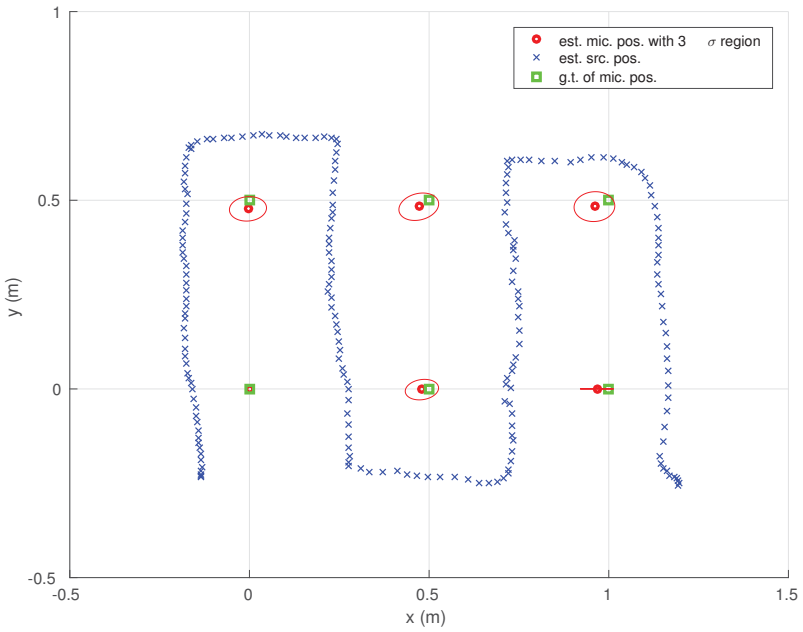
The algorithm described in the section 3.2 is used to calibrate the microphone array described above. The final estimation results are shown in Fig. 3.18 and Fig. 3.19. Since we only have 6 microphones in total, the accuracy of the estimation is expected to be similar to the 3×2 array and not as good as those with 9 or 16 microphones. However, final RMS errors for microphone positions is 0.0288m and 0.0204m. These errors are much better than the simulation result (RMS error of 3×2 in Table 3.2). The reason for this is that we have more sound source positions in the experimental setup than the simulation of 6 microphones. Moreover, the TDOA observation noise in the experiment is smaller than the one used in the simulation, which is conservatively assumed to be 0.333ms and this can be easily achieved under 44.1 KHz sampling rate. The uncertainty associated with x and y positions of microphone 1 and y position of microphone 3 are zero since microphone 1 is fixed at origin and microphone 3 is fixed at positive x axis. The error of the estimation result can also come from non-precise measurements of the speed of the sound in the current experimental setup in addition to the observation noise. Using more microphones, such as 9 or 16, or moving the sound source slower to have more sound source positions can improve estimation results further.

3.5.4.3 Experiment of Calibration of an Asynchronous Linear Microphone Array

Our experiment of calibrating an asynchronous microphone array is performed in an indoor environment. In the experiment, we first collect the calibration data for estimation of starting time offsets and clock differences. After the calibration, the standard ESPRIT algorithm is used to find DOA estimation for sound source. The experimental setup of the microphone array is shown in Fig. 3.20. In the experiment, four microphones are individually sampled by their own USB sound cards. The extra parameters in the experiment are summarised in Table. 3.6. Once calibrated, the DOA estimation results using ESPRIT algorithms w.r.t. ground truth values are shown in Fig. 3.21. The RMS error is 8.7961 degree.



a Initialisation of the state vector for the 1st trajectory.

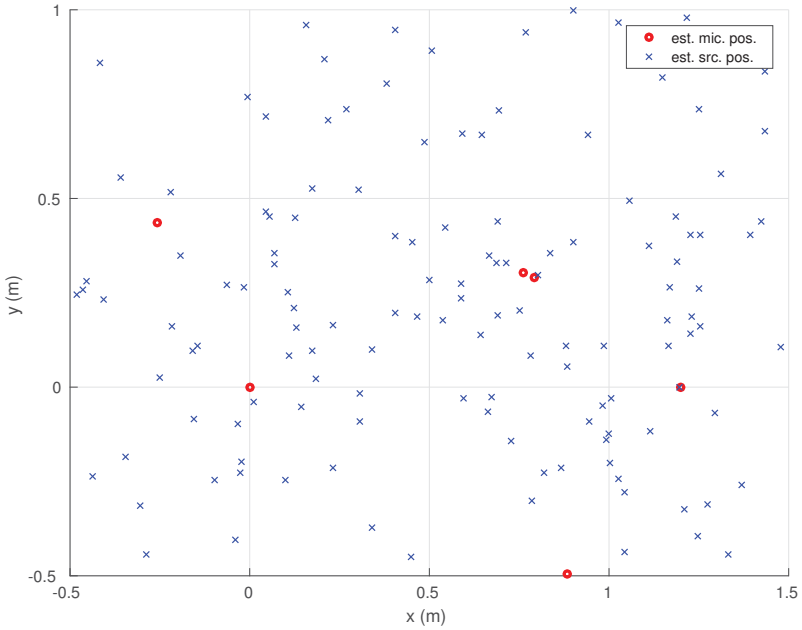


b Final estimation results for the 1st trajectory. RMS error of microphone positions is 0.0288m.

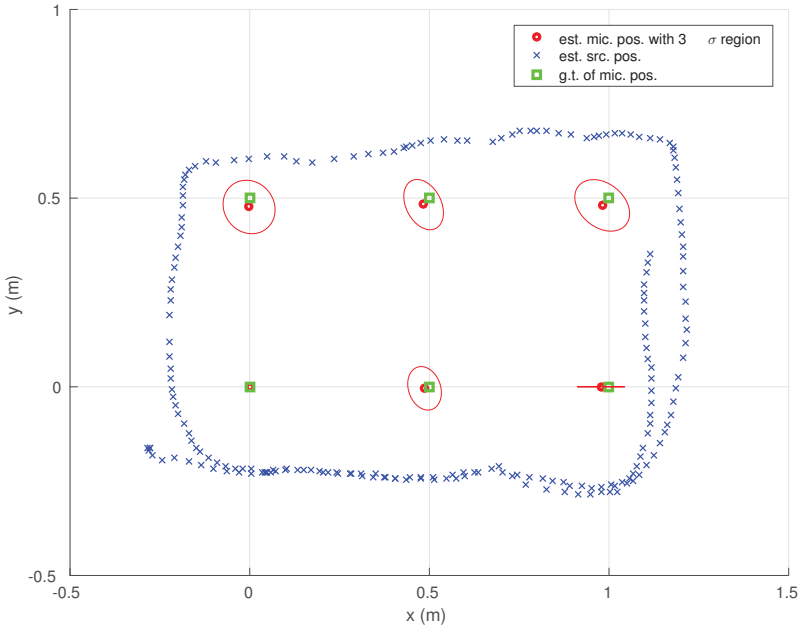
FIGURE 3.18: Experiments results of a 2×3 array.

3.6 Conclusion

In this Chapter, we presented our proposed method of calibrating a microphone array sensor, be it a synchronised or an asynchronous microphone array, although results are only



a Initialisation of the state vector



b Final estimation results. RMS error of microphone positions is 0.0204m.

FIGURE 3.19: Experimental results of a 2×3 array.

presented for the more challenging asynchronous case. The technique relies on observations from a moving sound source and a pose-graph filtering framework. Since the proposed method estimates geometric positions, the starting time offsets and clock differences of microphones, it relaxes two key constraints imposed by traditional techniques employed

FIGURE 3.20: Experimental setup of the asynchronous microphone array.

TABLE 3.6: Parameters setting in experiment

Parameters	Values
Number of microphones	4
Distance between microphones	0.05m
sound source	Samsung Galaxy S4 phone
sound wave	short time chirp
time interval of sound	0.5s
total duration of recording	1 min 14 s
ϵ for ΔX	10

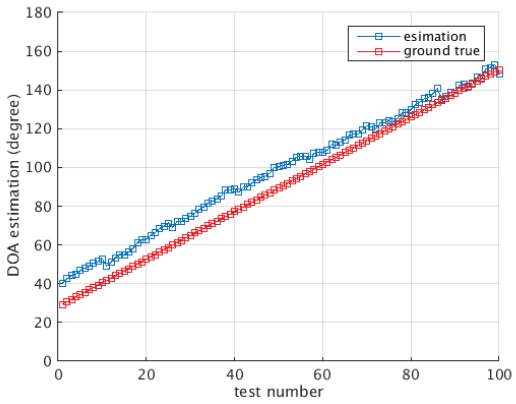


FIGURE 3.21: DOA estimation results after the calibration.

for microphone array based sound source localisation and separation to obtain synchronous readings of an audio signal: knowledge of accurate geometry information of the microphone array, as well as availability of a multichannel analog-to-digital converter. In comparison with relevant techniques of calibrating an asynchronous microphone array, the proposed methodology estimates the clock difference of each independent sound card in addition to the starting time offset, thereby making it more suitable for generic applications with standard audio devices.The proposed method can be used to calibrate a 2D/3D/linear

asynchronous microphone array. Once it has been calibrated, an asynchronous microphone array can be used for estimating DOA of sound sources just like a hardware synchronised microphone array, and thus it can be used for mapping sound sources as illustrated in Chapter 4 and 5 .

Chapter 4

Sound Source Mapping using a 2D/3D Microphone Array

4.1 Introduction

Due to the important application of sound source mapping in USAR and HRI scenarios [131] [62] [63], as explained in Chapter 1, recently, mapping of stationary sound sources has gained increased interest. With a microphone array, the robot can estimate bearing information of sound sources using DOA estimation algorithms once it has been calibrated. With multiple sound sources bearing observations from different robot poses, the sound source locations can be recovered.

Research literature in SLAM provide a sound framework for robot self-localisation and environmental map building. There are many successful implementations based on laser scanners [132] and vision sensors [133]. 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, precise simultaneous 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 DOA

of a sound source. Therefore, bearing-only information of sound source 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 the number of key image points detected by a vision sensor, making the attempt to solve the SLAM problem purely based on sound source quite difficult, especially in the 3D case that demands more landmarks to uniquely determine the robot pose. Thirdly, compared to monocular SLAM [133], which also relies on bearing-only landmarks, the bearing information from a sound source is not always available due to the sparseness of the audio signals. 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 the above mentioned reasons, performing SLAM with only sound source 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 [4] and [2], the robot moves relatively short distances so the drift in odometry remains small. Also in [4], multiple sound sources are mapped at the same time in order to obtain a sufficient 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 along a silent corridor). When the number of landmarks is not adequate, estimation of the robot trajectory becomes less accurate and so does the sound source location estimation.

In order to overcome these drawbacks, more recent work tends 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 source locations as well. Examples of such aiding have been shown by Kallakuri *et*

al. [10] and Vincent *et al.* [120]. In [120], a mobile robot with laser scanner and microphone array is used to map sound source 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, although both laser scanner and wheel odometry are used, robot pose 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 [10], a Rao-Blackwellised 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 extended to the 3D case in their later work [11] by replacing 2D occupancy maps with 3D octree map. Although the robot pose 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 has taken place.

A SLAM algorithm based on least square optimisation, which contains robot poses and environmental landmarks, and sound source locations will be the ideal framework to tackle the above issues. By keeping all robot poses, sound sources and other landmarks in a state vector of a least square optimisation based graph SLAM, the robot trajectory and sound source states are fully correlated. This guarantees that an update on the robot trajectory leads to the update on sound source positions, which makes the method more consistent. Therefore, firstly, we will present this general least square optimisation based SLAM framework to estimate robot poses, sound sources and other landmarks positions jointly. The proposed framework is able to map sound sources with either a 2D/3D or linear microphone array, though the initialisation of sound sources needs special treatment when applied to a linear microphone array as explained in Chapter 5.

Then, we come up with an improved method which has more flexibility and less computational cost. The improved method exploits the fact 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 Chapter we present an algorithm that still utilises robot pose uncertainty and allows

the updating of 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. It has the flexibility of using either a filtering or an optimisation based SLAM mapping. The second map (the sound source map) contains the robot locations from which the sound sources are observed together with the sound source encoded as IDP [133]. 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 are 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. As the second map uses a filtering technique, it is computationally less expensive than the joint optimisation framework.

The contributions of the Chapter are two-fold; a least square optimisation based SLAM framework, which estimates robot poses, sound sources and other landmarks positions jointly, is proposed to map sound sources using a 2D/3D or linear microphone array (the case of using a linear microphone array is detailed in Chapter 5 as initialisation of sound sources needs special treatment). Secondly, an efficient method of mapping sound sources using a 2D/3D microphone array is presented. This method shows the novel use of IDP to map sound sources and is a computationally efficient and flexible algorithm that exploits the CI property to propagate information from a map used for localisation to a sound source map.

The rest of the Chapter is organised as follows. In section 4.2, the least square optimisation based SLAM framework for mapping sound source is presented. In section 4.3, the details of the improved method are illustrated. In section 4.4, various simulation and experimental results are presented to show effectiveness of two proposed methods. Section 4.5 presents the conclusion and discussion about further work.

4.2 Sound Source Mapping using a Least Squares Optimisation based SLAM Framework

In this section, the least squares optimisation based SLAM framework for mapping sound source is presented. Here, we assume the robot observes visual landmarks by a camera as an example. Note that in general, the landmarks for localising the robot are not restricted to the visual landmarks (e.g. can be feature points from 3D laser scans).

Let $\mathbf{x}$ be the state vector of the graph SLAM, it contains robot poses, visual landmarks and sound sources locations as follows,

$$\mathbf{x} = [\mathbf{x}_r^1, \cdots, \mathbf{x}_r^K, \mathbf{x}_{lm}^1, \cdots, \mathbf{x}_{lm}^{N_v}, \mathbf{x}_{ss}^1, \cdots, \mathbf{x}_{ss}^{N_s}]^T, \tag{4.1}$$

where $\mathbf{x}_r^k (kf = 1 \cdots K)$ is the robot pose at time instance k, $\mathbf{x}_{lm}^{n_v} (n_v = 1 \cdots N_v)$ is the location of the n_vth visual landmark and $\mathbf{x}_{ss}^m (m = 1 \cdots N_s)$ is the location of the mth sound source. The 3D pose of the robot pose is $\in$ SE3 space. Location of a visual landmark or a sound source is parametrised as 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 square problem of the graph-based SLAM, the estimated state vector is found by minimising the error over all pose-pose constraints and pose-landmark constraints [7],

$$\hat{\mathbf{x}} = argmin \sum_{ij} \mathbf{e}_{ij}^T \mathbf{\Omega}_{ij} \mathbf{e}_{ij}, \tag{4.2}$$

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 observation of a sound source from a robot pose, the error function is the difference of the expected bearing information of the sound source, which is the azimuth and elevation angles, to the real observed values. When an edge represents

observation of a visual landmark from a robot pose, it depends on the nature of the sensor (Monocular, Stereo or RGBD) and details regarding them can be found in [110]. After all nodes and edges are defined, Eq. 4.2 can be solved by Gauss-Newton or Levenberg-Marquardt optimisation.

4.3 Sound Source Mapping by CI Submap Joining using a 2D/3D Microphone Array

While the least squares optimisation based SLAM framework is able to map sound sources by itself, it is not convenient to exploit existing SLAM implementations. One needs to modify existing SLAM implementations, which have to be based on least squares optimisation, by adding sound source states and constraints to make it work. In addition, the optimisation process takes substantial computational time.

In this section, we present the improved method of mapping sound sources using a 2D/3D microphone array by CI submap joining. By splitting the full map into the localisation map and the sound map, the localisation map has the flexibility of using any existing SLAM implementation. The sound map uses the EKF to estimate sound source locations, hence consumes less computational time. We present the details to generate two conditionally independent maps split from a full SLAM map. These maps are maintained and updated by two different SLAM algorithms, one for simultaneous trajectory estimation and the other for sound source mapping.

4.3.1 Structure of the Split CI Maps

Let us first examine the Bayesian network in Fig. 4.1, 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. 4.1, 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}.

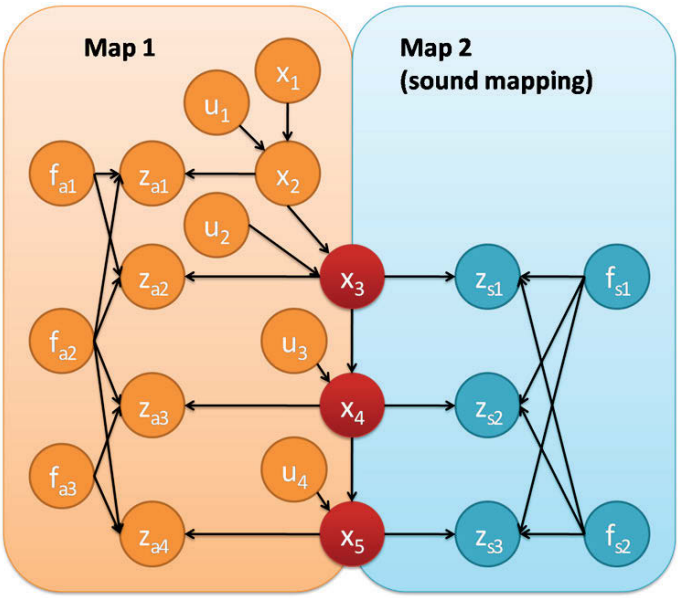


FIGURE 4.1: Bayesian network that describes probabilistic dependency between two CI maps. Map 1 represents the localisation map which estimates the robot pose and landmarks locations of the exteroceptive sensor, whereas map 2 represents the sound map which estimates locations of sound sources.

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. 4.1.

Note that the situation in Fig. 4.1 is a special case of the structure of conditionally independent submaps method presented in [134]. It can be seen as a situation where the robot frequently revisits 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 [134], in a frequently revisiting scenario, keeping all robot poses which are common in two submaps in the state vectors of both

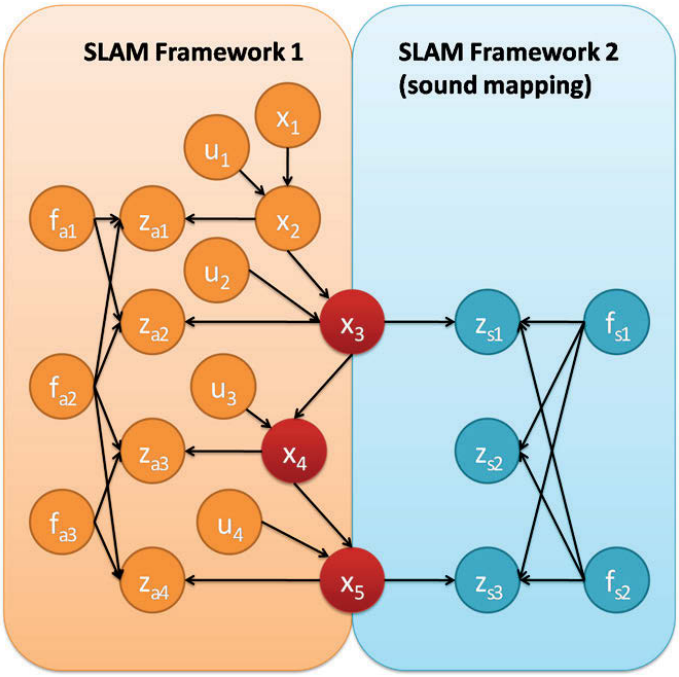


FIGURE 4.2: Modified Bayesian network that describes probabilistic dependency between SLAM variables in two maps.

SLAM maps increases the length of both state-vectors, which leads to a significant increase of the computational complexity. In [134], it 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. 4.2). Although at first glance it seems different, 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 [134].

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 [134] 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 updated. However, we applied the back propagation algorithm after each update of the localisation map.

4.3.2 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 [135] can be used for laser scanner based SLAM, RGB-D SLAM [112] can be used for RGB-D sensors and ORB-SLAM [110] can be used for monocular or stereo camera. In the 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.

4.3.3 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 source 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 [133]. Under IDP parametrisation, the state of each sound source in 2D is,

$$\mathbf{x}_{lm}^{s}(i) = (x_i y_i \theta_i \rho_i)^T \tag{4.3}$$

and in 3D case is

$$\mathbf{x}_{lm}^{s}(i) = (x_i y_i z_i \theta_i \phi_i \rho_i)^T \tag{4.4}$$

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 angles of the 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

$$\mathbf{x}^s = (\mathbf{x}_r, \mathbf{x}^s_{lm}(1), \mathbf{x}^s_{lm}(2), \ldots, \mathbf{x}^s_{lm}(n))^T \tag{4.5}$$

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

$$\mathbf{x}_r = (x_r, y_r, \theta_r)^T \tag{4.6}$$

in the 2D case and

$$\mathbf{x}_r = (x_r, y_r, z_r, qw_r, qx_r, qy_r, qz_r)^T \tag{4.7}$$

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 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,

$$K^s_t = P^s_{t-1} H^{sT}_t (H^s_t P^s_{t-1} H^{sT}_t + Q^s_t)^{-1} \tag{4.8}$$

$$\mathbf{x}^s_t = \mathbf{x}^s_t + K^s_{t-1}(z^s_t - h^s(\mathbf{x}^s_{t-1})) \tag{4.9}$$

$$P^s_t = (I - K^s_t H^s_t) P^s_{t-1} \tag{4.10}$$

where P^s_{t-1} and P^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 [136].

Note that with IDP parametrisation of sound source locations in Eq.4.3 or Eq.4.4, 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.

4.3.4 Correlation 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 unidirectionally. 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

$$p(\mathbf{x}^a|\mathbf{u}_{1:n}, \mathbf{z}_{a1:an}) = \mathcal{N}(\mathbf{x}^a, P^a) \tag{4.11}$$

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

$$\mathbf{x}^a = (\mathbf{x}_r, \mathbf{x}^s_r(1), \ldots, \mathbf{x}^s_r(n_s), \mathbf{x}^a_{lm}(1), \ldots, \mathbf{x}^a_{lm}(n))^T \tag{4.12}$$

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

$$\mathbf{x}^s_r(i) = (\mathbf{x}^{s\text{-}p}_r(i), \mathbf{x}^{s\text{-}o}_r(i))^T, \tag{4.13}$$

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 initialise ith sound source. Then, the full state vector can also be written as

$$\begin{aligned} \mathbf{x}^a = &(\mathbf{x}_r, \mathbf{x}^{s\text{-}p}_r(1), \mathbf{x}^{s\text{-}o}_r(1), \ldots, \\ &\mathbf{x}^{s\text{-}p}_r(n_s), \mathbf{x}^{s\text{-}o}_r(n_s), \mathbf{x}^a_{lm}(1), \ldots, \mathbf{x}^a_{lm}(n))^T. \end{aligned} \tag{4.14}$$

grouping the localisation map as

$$\begin{aligned}\check{\mathbf{x}}^a =&(\mathbf{x}_r, \mathbf{x}_r^{s\text{-}p}(1), \ldots, \mathbf{x}_r^{s\text{-}p}(n_s),\\ &\mathbf{x}_r^{s\text{-}o}(1), \ldots, \mathbf{x}_r^{s\text{-}o}(n_s), \mathbf{x}_{lm}^a(1), \ldots, \mathbf{x}_{lm}^a(n))^T.\end{aligned} \tag{4.15}$$

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

$$\check{\mathbf{x}}^a = (\mathbf{x}_{C_a}, \mathbf{x}_A)^T. \tag{4.16}$$

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

$$\check{P}^a = \begin{bmatrix} P_{C_a} & P_{CA} \\ P_{AC} & P_A \end{bmatrix}, \tag{4.17}$$

where P_{C_a}, P_A, P_{CA} andP_{AC} are covariance matrices 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,

$$\check{\mathbf{x}}^s = (\mathbf{x}_{C_s}, \mathbf{x}_S)^T, \tag{4.18}$$

$$\check{P}^s = \begin{bmatrix} P_{C_s} & P_{CS} \\ P_{SC} & P_S \end{bmatrix}, \tag{4.19}$$

where $\mathbf{x}_{C_s} = (\mathbf{x}_r, \mathbf{x}_{lm}^{s\text{-}p}(1), \ldots, \mathbf{x}_{lm}^{s\text{-}p}(n))^T$, in which $\mathbf{x}_{lm}^{s\text{-}p}(i)$ represents position of ith robot pose that can be used for sound source initialisation ($(x_i y_i)^T$ of Eq.4.3 in 2D case and $(x_i y_i z_i)^T$ of Eq.4.4 in 3D case). $\mathbf{x}_{C_s}$ corresponds to $\mathbf{x}_{C_a}$ in Eq.4.16 and they are the shared part of state vectors of the two maps. $\mathbf{x}_S = (\mathbf{x}_{lm}^{s\text{-}o}(1), \ldots, \mathbf{x}_{lm}^{s\text{-}o}(n))^T$, where $\mathbf{x}_{lm}^{s\text{-}o}(i)$ represents bearing and inverse distance of ith sound source ($(\theta_i \rho_i)^T$ of Eq.4.3 in 2D case and $(\theta_i \phi_i \rho_i)^T$ of Eq.4.4 in 3D case), and it is other element of the state vector in the second map which is conditionally independent from the first map. P_{C_s}, P_S, $P_C S$ and $P_S C$ in Eq.4.19 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 [134]. 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.4.18 and Eq.4.19 of the sound source map are updated as

$$\mathbf{x}^b_{C_s} = \mathbf{x}_{C_a} \tag{4.20}$$

$$P^b_{C_s} = P_{C_a} \tag{4.21}$$

$$K^b_{12} = P_{SC} P^{-1}_{C_s} \tag{4.22}$$

$$P^b_{SC} = K^b_{12} P^b_{C_s} \tag{4.23}$$

$$P^b_S = P_S + K^b_{12}(P^b_{CS} - P_{CS}) \tag{4.24}$$

$$\mathbf{x}^b_S = \mathbf{x}_S + K^b_{12}(\mathbf{x}^b_{C_s} - \mathbf{x}_{C_s}), \tag{4.25}$$

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 the transpose of P^b_{SC} due to the symmetry of the covariance matrix.

Differently to CI submaps scenario, the back propagation process 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 these robot locations ($\mathbf{x}_{C_a}$ and $\mathbf{x}_{C_s}$) is disregarded due to the following reasons. Firstly, sound sources are sparse in the time axis and in most cases the 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 indoor environments, accuracy of the bearing observations of sound source 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,

$$\mathbf{x}_{C_s} \approx \mathbf{x}_{C_a} \tag{4.26}$$

$$P_{C_s} \approx P_{C_a}, \tag{4.27}$$

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 extra time complexity.

4.4 Simulation and Experimental Results

In this section, comprehensive simulation and experimental results are presented to evaluate and compare the method described in section 4.2 and section 4.3 to the optimal and other possible solutions.

4.4.1 Simulation Results

4.4.1.1 Sound Source Mapping with only Odometry Information

In the simulation scenario shown in Fig. 4.3(a), the robot follows a square trajectory using only information from odometry and sound source. 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 in Table 4.1 part I. Bearing estimation noise is set to be a Gaussian noise with standard deviation of ± 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. 4.3 for EKF with IDP parametrisation and Fig. 4.4 for

TABLE 4.1: Parameters in simulation

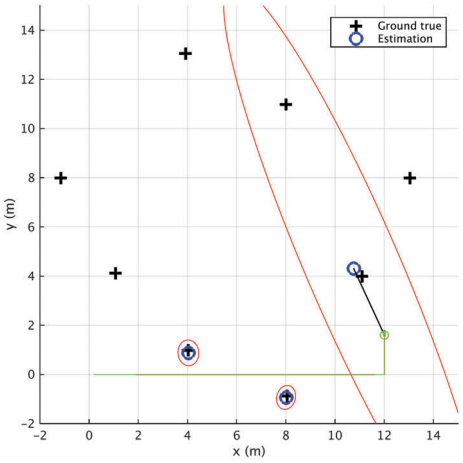
Parameters	Values
Part I	
Distance per odometry step	0.2m
Odometry noise (Trans. and Orient.)	0.001m and 0.001 deg
Sound bearing noise (Azimuth & Elevation)	10 deg
Least square optimiser	Levenberg-Marquardt
Part II	
Noise of range bearing sensor	0.01m and 1 deg
Odometry noise (Trans. and Orient.)	0.02m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

least square optimisation. Similar results are obtained from 3D simulation. From those figures, it can be seen that sound mapping works well under very accurate odometry.

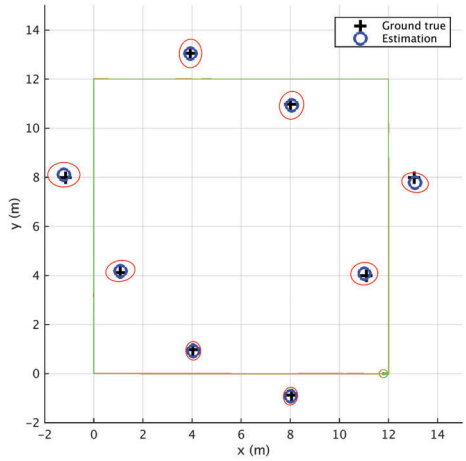
A 20 runs Monte Carlo simulation shows that by increasing odometry noise, the sound mapping estimation fails even with very reasonable noise values of less than 5% of the displacement. The RMS errors and convergence rates of sound source mapping under different odometry noise using EKF IDP parametrisation and least square optimisation in 2D and 3D cases with 20 Monte Carlo runs for each case is shown in Fig. 4.5. As can be seen from the figure, for a range of odometry noises the mean RMS errors of estimated sound source locations grows exponentially with time. The figure also presents the convergence rate for all algorithms. It can be seen from the figure that, in order to get sound mapping with reasonably good accuracy (e.g. 0.2m) in the simulated situation, odometry needs to be highly accurate (0.008m in translation and 0.128 deg in orientation for each step(0.2m)). Most mobile robotic platforms (e.g. Turtlebot) can not provide such a high accuracy odometry. Therefore, including other landmarks for accurate robot self localisation is necessary.

4.4.1.2 Sound Source Mapping by a Least Squares Optimisation based SLAM Framework with Odometer and Range-Bearing Observations of Environment Landmarks

We simulated a scenario adding an exteroceptive sensor (e.g. a laser scanner for a 2D scenario and RGBD camera for a 3D scenario), which observes range and bearing information



a Initial state of IDP sound sources.

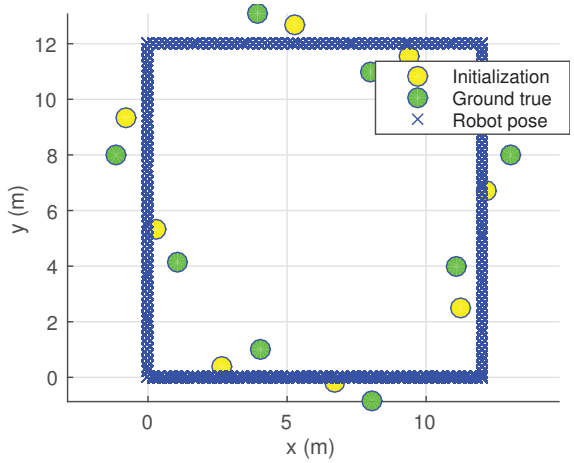


b Final estimation results.

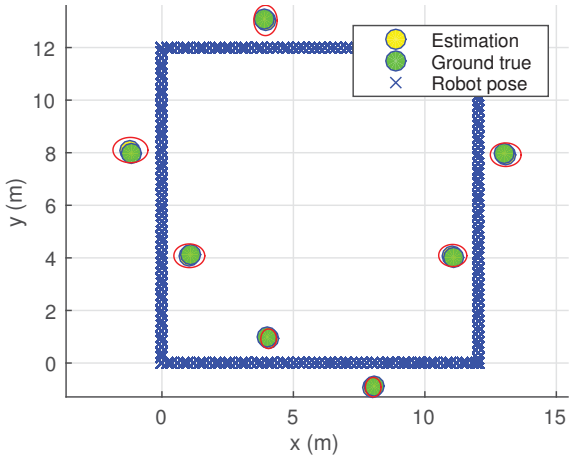
FIGURE 4.3: EKF parametrised by IDP with highly accurate odometry information.

of point landmarks in the environment (e.g. corner points of laser scans or visual point features). In the simulation, the robot follows the same trajectory as before and comes back to its original point for loop closure. The parameters used for the additional sensors are shown in Table 4.1 part II and other parameters are in Table 4.1 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.

The results of the 2D/3D simulations are shown in Fig. 4.6, Fig. 4.7 and Fig. 4.8. In all figures, green, red and blue unit lines denote the X,Y,Z axis of the robot local coordinate



a Initial state of IDP sound sources.



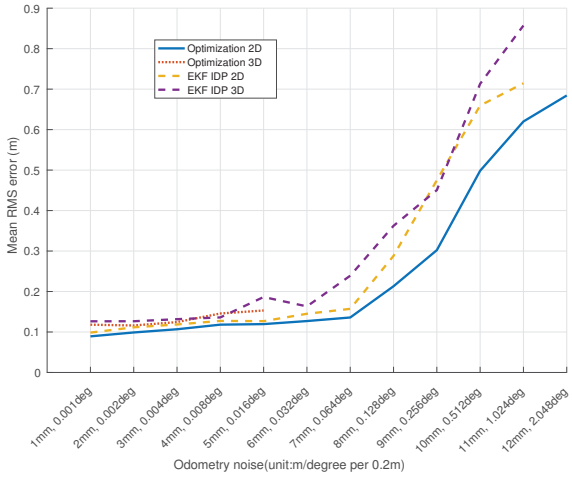
b Final estimation results.

FIGURE 4.4: Least square optimisation with highly accurate odometry information.

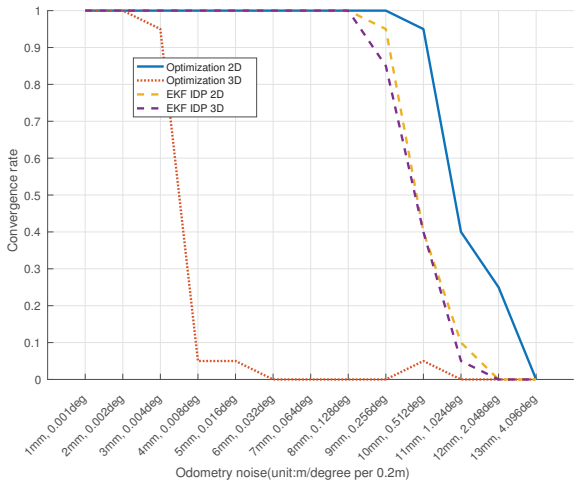
frame. It can be seen from the figure that the proposed least square optimisation based SLAM framework can accurately map all sound sources even in a long trajectory.

4.4.1.3 Sound Source Mapping by CI Submap Joining Method with Odometer and Range-Bearing Observations of Environment Landmarks

In the next step, we simulated a 2D scenario of the previous simulation using our improved method of sound source mapping by CI submap joining. In this simulation scenario, the proposed method utilises an EKF-SLAM algorithm for the localisation map fusing these additional range and bearing observations. Robot locations at sound source initialisation instants are used as common elements of two maps as explained before. In the simulation,



a RMS errors.

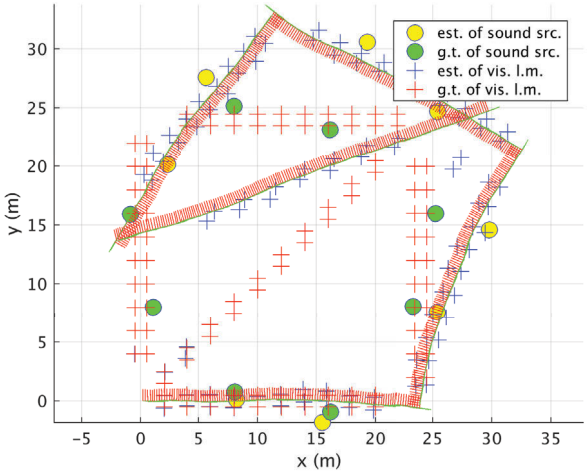


b Convergence rates.

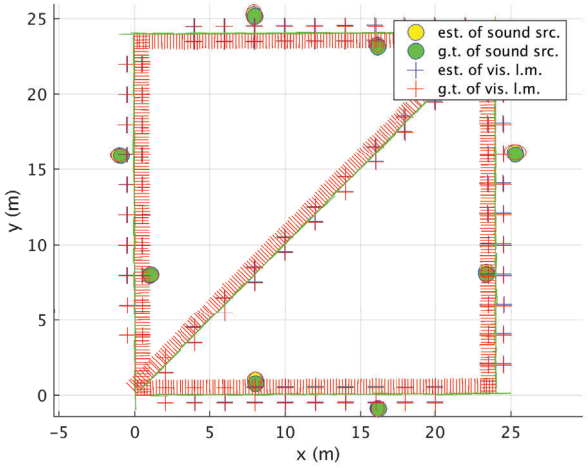
FIGURE 4.5: RMS errors and convergence rates under different odometry noise.

again the robot follows the same trajectory as before and comes back to its original point for loop closure. The parameters used for the additional sensors are the same as shown in Table 4.1 part II and other parameters are in Table 4.1 part I. The odometry noise is again 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. 4.9 and Fig. 4.10. In all figures, green circular markers represent estimated environment landmarks, pink plus markers represent ground true locations of range-bearing landmarks, red eclipses represent 3 σ region, green line represents ground truth robot trajectory and red line represents estimated robot trajectory. The



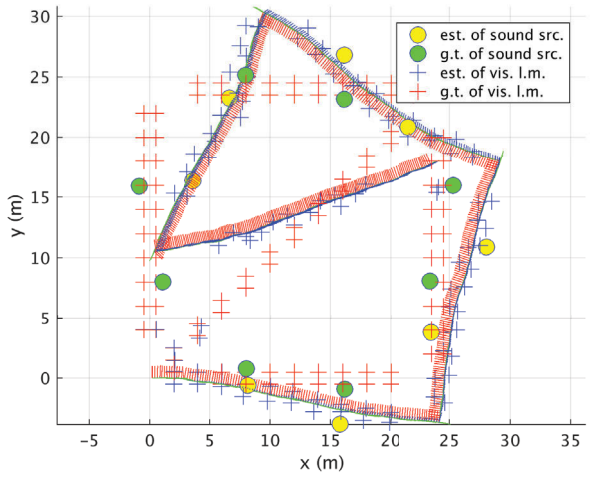
a Initialisation.



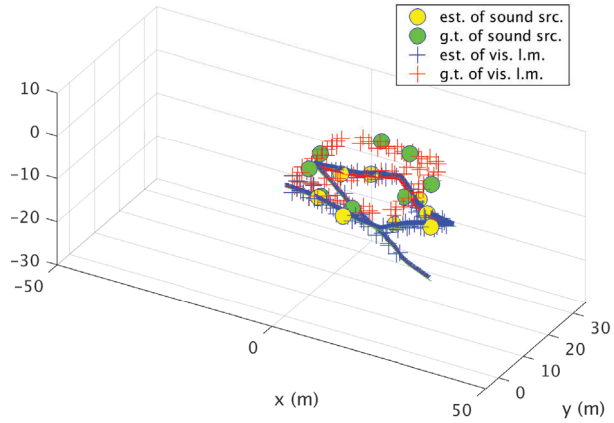
b Final estimation result.

FIGURE 4.6: 2D sound source mapping by the least square optimisation based SLAM framework.

meaning of the 3 σ region for a two-dimensional variable is the 3 σ Gaussian probability region of the landmark location in the 2D plane. 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



a Top view.

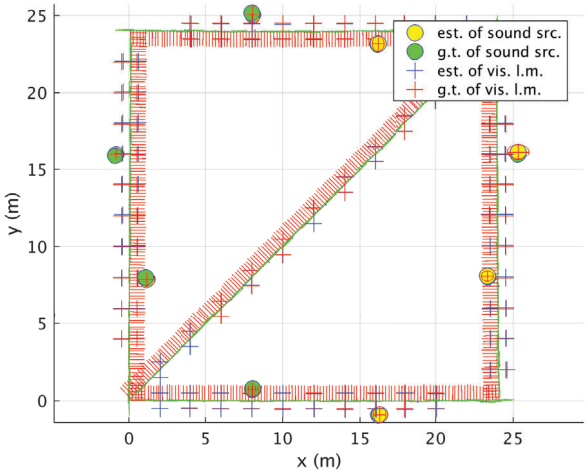


b Side view.

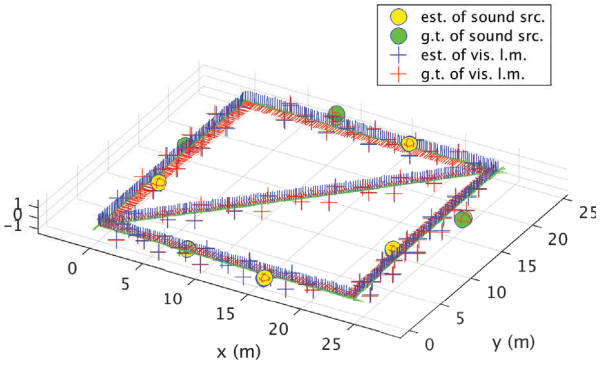
FIGURE 4.7: 3D sound source mapping by the least square optimisation based SLAM framework. Initialisation of the system.

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 both EKF SLAM algorithm with sound source parametrised by IDP and least squares optimisation as stated in section 4.2. We compared the proposed method with the full SLAM method in terms of sound mapping accuracy with various trajectory



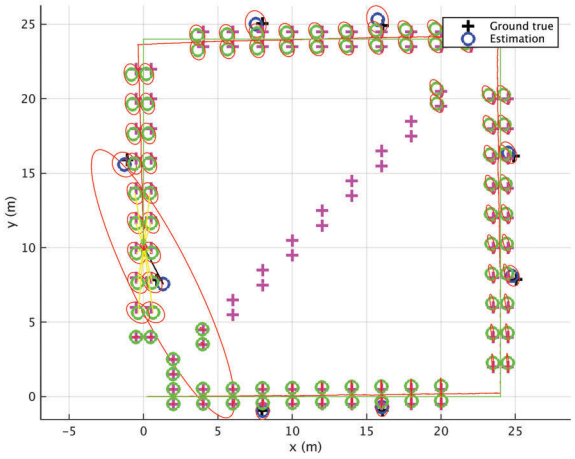
a Top view.



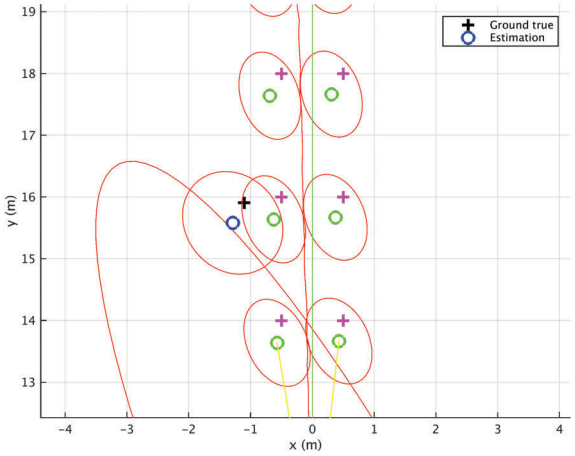
b Side view.

FIGURE 4.8: 3D sound source mapping by the least square optimisation based SLAM framework. Final estimation results.

lengths. For each trajectory, a 10 runs Monte Carlo simulation is used to compute the Mean RMS errors. The results are shown in Fig.4.11. 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



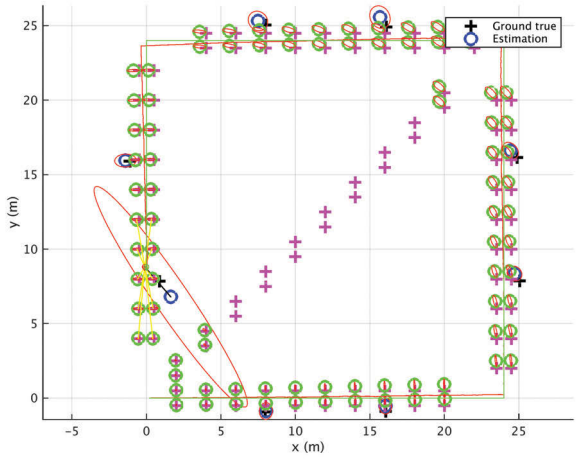
a Sound source mapping before loop closure.



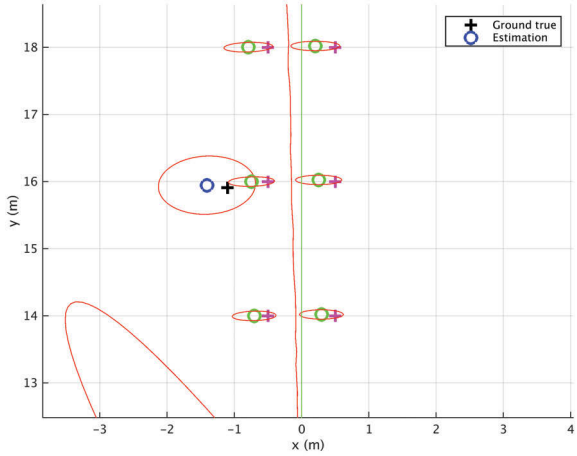
b Zoomed in view of top left sound source before loop closure.

FIGURE 4.9: sound source mapping with additional range-bearing observations before loop closure.

error of EKF filtering 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 a result producing more accurate results than 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).



a Sound source mapping after loop closure.



b Zoomed view of the top left sound source after loop closure.

FIGURE 4.10: sound source mapping with additional range-bearing observations after loop closure.

4.4.2 Experimental Results

In this section, two different experimental scenarios are used to show the effectiveness and flexibility of the improved sound source mapping method by CI submap joining.

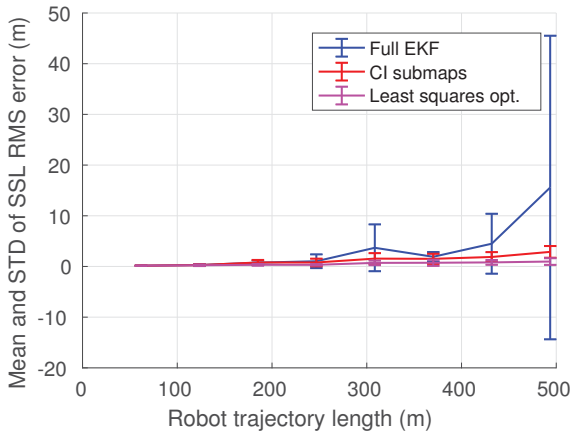


FIGURE 4.11: Mean RMS errors with STD of 10 Monte Carlo runs under various length of robot trajectories.

4.4.2.1 2D Sound Source Mapping by a Mobile Robot with a Microphone Array and a Laser Scanner

A turtlebot [137] with Hokuyo laser range finder and Microcone (6-microphone circular array) is used to localise two sound sources generating white noise (see Fig. 4.12). Turtlebot is a differential drive mobile robot made by Clearpath Robotics Inc [138]. It is also equipped with a wheel encoder for odometry estimation in addition to the laser scanner and the microphone array that we installed on it. We use the EKF-SLAM described above for the sound map and the pose SLAM implementation in [135] 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 shared at each SLAM step with the sound source map. Then the shared part of the state-vector allow us to back propagate the information to the sound source map after each update in the localisation map. Sound bearing observation noise is set to ± 10 deg. HARK [100] is used for sound source bearing estimation using MUSIC algorithm.

The results are shown in Fig. 4.13. In all figures, blue markers represent estimated sound landmarks, pink markers represent ground truth locations, red eclipses represent 3σ region and blue line represents estimated robot trajectory. It can be seen that the proposed method has successfully estimated two sound sources with reasonably good accuracy given the noisy nature of the audio observations.

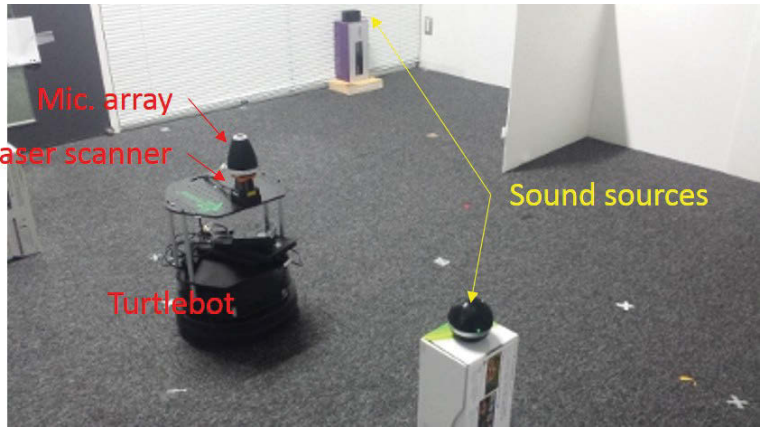
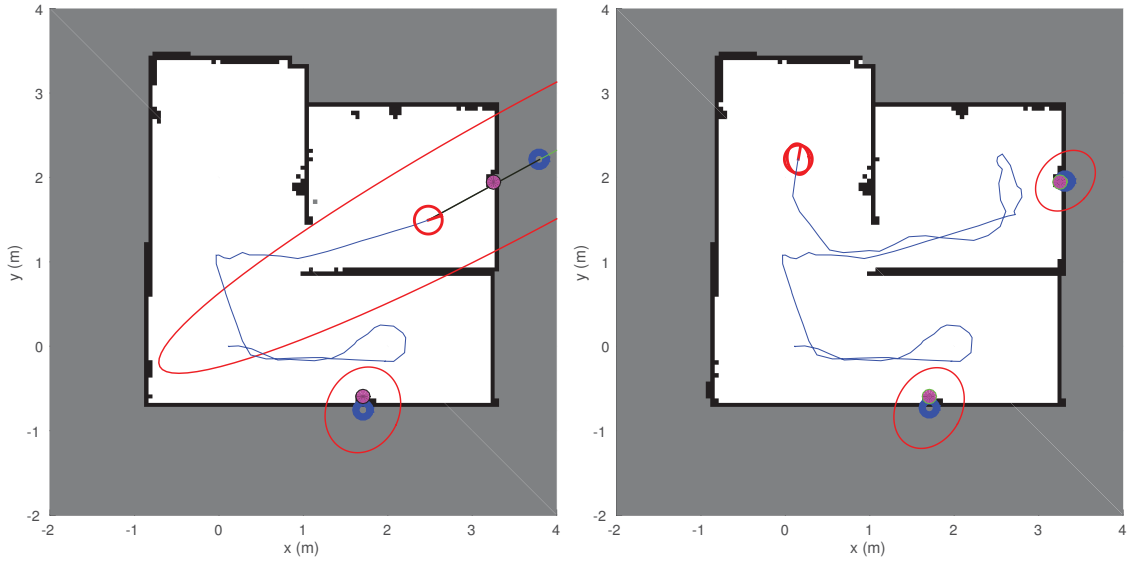


FIGURE 4.12: Turtelbot equipped with a laser scanner and a Microcone (circular microphone array).



a Sound landmark initialisation with IDP parametrisation.

b Final estimation results.

FIGURE 4.13: 2D sound source mapping results using a mobile with laser scanner.

4.4.2.2 3D Sound Source Mapping using a Hand Held PS3-eye (Monocular Camera with a Linear Microphone Array)

The configuration of the sensor and the experimental setup is shown in Fig. 4.14. As can be seen from the Fig. 4.14(a), the PS3-eye sensor consists of a monocular camera and a 4-channel uniformly distributed linear microphone array. During the experiment, the sensor follows a certain trajectory around three sound sources in an office environment as shown in Fig. 4.14(b). An off-the-shelf visual SLAM implementation without any modification

is used in this experiment. ORB-SLAM [110] 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 ($\pm$60 deg) to hint that the sound source is in front of the sensor (due to the casing for PS3-eye, it is more sensitive when detecting sound sources in front of it). The sound map is the same as 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. 4.15 and Fig. 4.16. In all figures, green markers represent estimated sound sources, pink markers represent ground truth locations, blue markers represent key frames' locations and black dots represent final feature points from ORB-SLAM. 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 truth locations. From Fig. 4.15, we can see that the sound source is initialised with IDP when it is first observed by the sensor. The green ellipses represent the one sigma uncertainty region of the sound source locations along X, Y and Z axes. We can see that the uncertainty is higher along the elevation angle and the depth since these two parameters are unobservable at the first observation of the sound source . From Fig. 4.15(a), we can see that the sensor trajectory has drifted before loop closure. Therefore, the estimated sound sources locations have also drifted. From Fig. 4.15(a), we can see that after a loop closure is detected, the sensor trajectory is corrected and so do the position estimates of the sound sources. This is again thanks to the split CI maps. From the experiment, we can also 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 source map in 3D with the proposed method. A video showing the performance of the proposed system in this experiment is publicly

a b

FIGURE 4.14: PS3-eye configuration (a) and experimental setup (b).

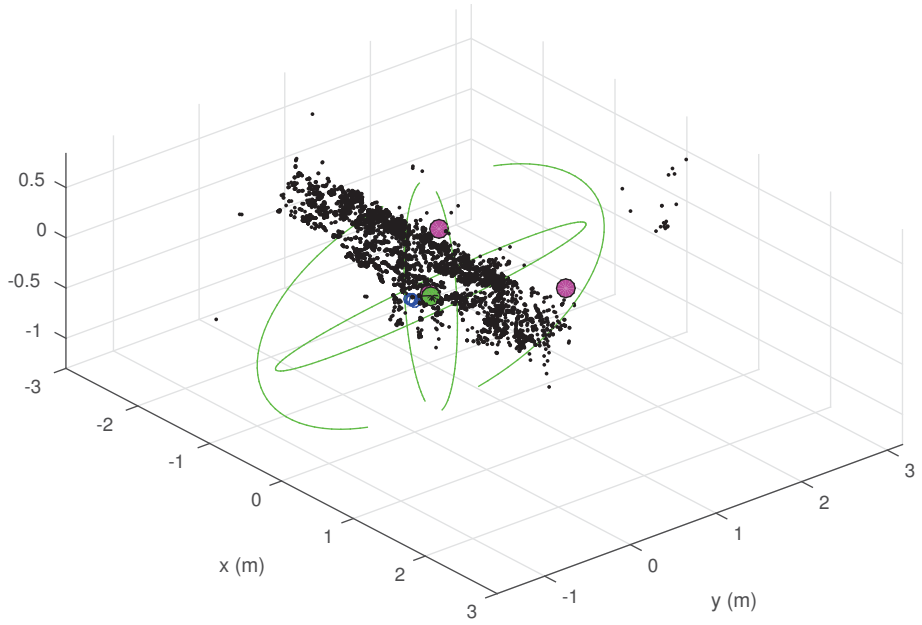
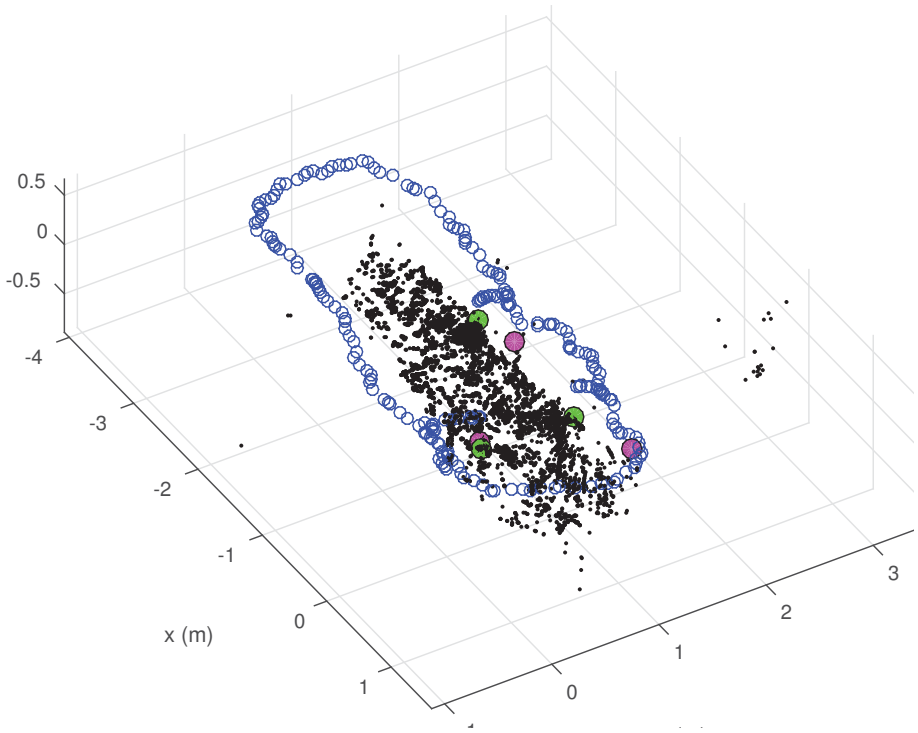


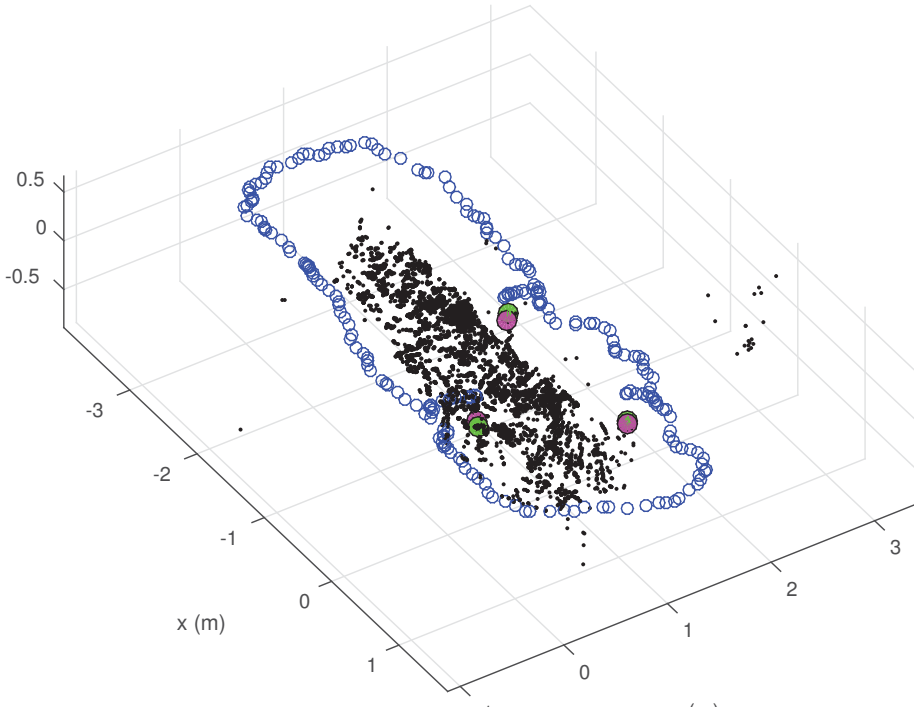
FIGURE 4.15: Sound landmark initialisation with IDP parametrisation in 3D sound source mapping using a hand hold PS3-eye experiment (monocular camera with linear microphone array). The green ellipses represent the one sigma uncertainty region of the sound source locations along X, Y and Z axes. The uncertainty is higher along the elevation angle and the depth from the sensor since these two parameters are unobservable during initialisation.

available online[1].

[1]https://youtu.be/QyqY2eIX1wk



a Estimation results (before loop closure). Estimation of sound sources locations (represented by green markers) has considerable error w.r.t. the ground truth locations (represented by pink markers) due to the drifted sensor poses.



b Final estimation results (after loop closure). The sensor poses are corrected after the loop closure in the localisation map, which leads to the update of sound sources locations estimation towards their ground truth locations.

FIGURE 4.16: 3D sound source mapping results using a hand hold PS3-eye (monocular camera with linear microphone array).

4.5 Conclusion

In this Chapter, we first presented a least squares optimisation framework to jointly estimate robot poses, positions of sound sources and other landmarks using a 2D/3D or linear microphone array (the case of using a linear microphone array is detailed in Chapter 5 as initialisation of sound sources needs special treatment). Secondly, an improved method of mapping sound source using a 2D/3D microphone array was presented. Specifically, we proposed a split CI mapping method for sound source mapping and robot localisation. Our efficient method utilises two SLAM 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 source mapping parametrised as inverse-depth points. As sound source observations are bearing-only, extremely noisy and sparse, they are not used 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 source locations. The key advantage of using IDP is that it models accurately uncertainty of faraway points, utilises all information contained in bearing-only sound observations and linearisation errors are small compared with Euclidean points. The improved 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 approximations 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.

Chapter 5

Sound Source Mapping using a Linear Microphone Array

5.1 Introduction

3D cameras such as Microsoft Kinect 360, Kinect One, PS3 Eye and PS4 Eye sensors, as shown in Fig. 5.1, are becoming an integral part of the perception modules of robotic and intelligent systems. A common feature of these microphone arrays is that the geometric location of all microphones are distributed along a straight line, i.e. in a linear array, be it uniformly distributed (in Fig. 5.1 (b) and (c)) or not (in Fig. 5.1 (a) and (d)).

Despite easy availability at an affordable price and frequent usage of sensors with a linear microphone array in robotic systems, conventional 3D sound source 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, which we will discuss in Section 5.3. The least square optimisation based SLAM framework proposed in Chapter 4 can not be applied directly, since the initialisation of sound sources is not as straightforward as it is in the 2D/3D microphone array case.

a Kinect 360.

b Kinect One.

c PS3 Eye.

d PS4 Eye.

FIGURE 5.1: Typical robotic sensors that include a linear microphone array.

In recent work of 2D sound source mapping, Hu et. al. in [4] proposed a FastSLAM based approach to map multiple sound sources using a 3D microphone array. Sasaki et. al. in [2] uses a self motion triangulation method to deal with sound source mapping using a concentric microphone array. A ray casting based probabilistic 2D sound source mapping approach is proposed by Kallakuri et. al. in [10]. Conventional approaches such as [11, 139] 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 [11], Even et. al. extend their previous work in [10] to the 3D case by using a 3D microphone array. In [139], 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 works in 3D sound source mapping even use multiple microphone arrays. In [31], Ishi et. al. use multiple 3D microphone arrays attached on the ceiling to estimate 3D locations of multiple sound sources. They also exploit the reflection information to improve the localisation accuracy. In [140], Seewald et. al. use two perpendicularly placed Microsoft

Kinects to estimate 3D locations of sound sources. Note that in [31, 139, 140], all sensors and sound sources are static.

In this Chapter, 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 Chapter is two-fold: firstly we introduce a framework that allows the mapping in real-time of the location of 3D sound sources using a linear microphone array without any prior knowledge of the sensor hardware as was done in our previous work [70]. 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.

The rest of the Chapter is organised as follows. In Section 5.2, sensor model for a linear microphone array using Gaussian Process is presented. In Section 5.3, initialisation of sound sources using multi-hypotheses filters is presented. In Section 5.4, details about jointly optimising sensor poses, visual landmarks and sound sources are presented. In Section 5.5, various simulation and experimental results are presented to show the effectiveness of the proposed method. Section 5.6 presents the conclusion and discussion about further work.

5.2 Gaussian Processes to Model Linear Microphone Arrays Sensors

A graphical representation of the sensor model of a linear microphone array is shown in Fig. 5.2. 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

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

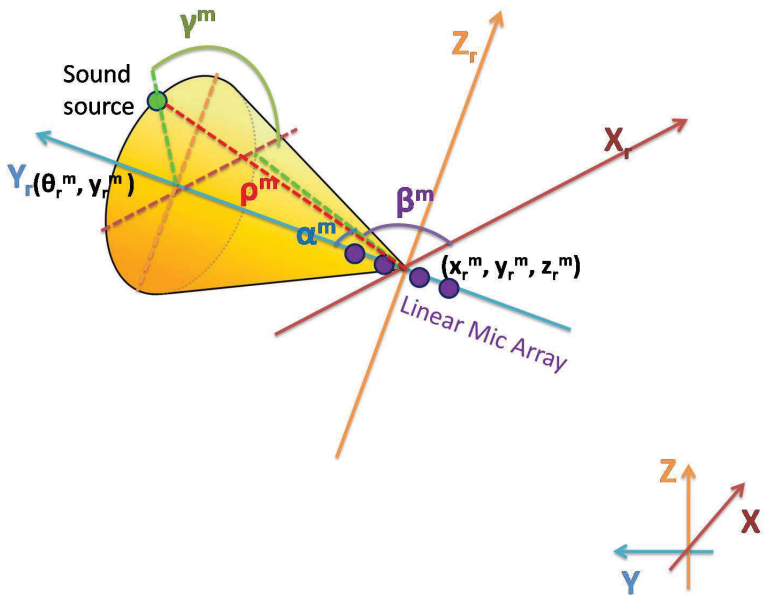


FIGURE 5.2: Linear microphone array notation and parametrisation of a 3D sound source location.

$\mathbf{p}^m = [x_{ss}^m, y_{ss}^m, z_{ss}^m]^T$ be the Euclidean coordinates of the mth sound source and $\mathbf{x}_{r,k}$ be sensor pose at time instance k. The observation β_k^m of this sound source $\mathbf{p}^m$ using the linear microphone array from the sensor pose $\mathbf{x}_{r,k}$ is

$$\begin{bmatrix} \mathbf{p}_k^m \\ 1 \end{bmatrix} = \mathcal{M}^{-1}(\mathbf{x}_{r,k}) \begin{bmatrix} \mathbf{p}^m \\ 1 \end{bmatrix}, \tag{5.1}$$

$$\beta_k^m = atan2(\mathbf{p}_k^m(2), \sqrt{\mathbf{p}_k^m(1)^2 + \mathbf{p}_k^m(3)^2}), \tag{5.2}$$

where $\mathcal{M}(\mathbf{x}_{r,k})$ is the homogeneous transformation of the sensor pose $\mathbf{x}_{r,k}$, $\mathbf{p}_k^m$ 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 $atan2(\bullet)$ returns the four-quadrant inverse tangent angle.

The observation β_k^m, in practice, is obtained by processing a multi channel audio signal. The TDOA from a sound source to all channels of the microphone array is commonly exploited to estimate the DOA observation β_k^m. Typical methods for estimation of DOA from multi channel audio signal include MUSIC [94] and SRP-PHAT [141]. 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 an audio signal, the estimated angle from a DOA estimation algorithm $\hat{\beta}_k^{m,DOA}$ is affected by noise. The DOA estimation ($\hat{\beta}_k^{m,DOA}$) accuracy for a linear microphone array varies according to the true DOA estimation angle (β_k^m). When the true DOA angle is close to 0 rad, which means the sound source is in front of the linear array, DOA estimation accuracy is best. On the other hand, when the true DOA angle is $\pm\pi/2$ degree, DOA estimation accuracy is at its worst.

In addition, there is a bias in the mean estimation values around a true DOA angle, particularly at the limits ($\pm\pi/2$). Therefore, the sensor model of a linear microphone array cannot be constructed simply by the DOA estimation plus a constant noise term - the sensor models applicable to most other microphone arrays (e.g. a circular microphone array).

Since there is no obvious parametric model that can describe this bias in the mean value and how the noise term increase around $\pm\pi/2$ degree, a machine learning model via non-parametric Gaussian Process [142] model is adopted to capture this behavior using real experimental dataset (with the estimated and ground truth DOA angles). A Gaussian Process is a generalisation of the Gaussian probability distribution. In Gaussian Process, a probability distribution describes random variables and a stochastic process governs the properties of functions [142]. 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,

$$\boldsymbol{\beta}_{gp} \sim \mathcal{N}(\mathbf{0}, K(\hat{\boldsymbol{\beta}}_{gp}^{DOA}, \hat{\boldsymbol{\beta}}_{gp}^{DOA}) + \sigma_n^2 \mathbf{I}), \tag{5.3}$$

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

$$\begin{bmatrix} \boldsymbol{\beta}_{gp} \\ \beta_{gp*} \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \\ \begin{bmatrix} K(\hat{\boldsymbol{\beta}}^{DOA}_{gp}, \hat{\boldsymbol{\beta}}^{DOA}_{gp}) + \sigma_n^2 \mathbf{I} & K(\hat{\boldsymbol{\beta}}^{DOA}_{gp}, \hat{\beta}^{DOA}_{gp*}) \\ K(\hat{\beta}^{DOA}_{gp*}, \hat{\boldsymbol{\beta}}^{DOA}_{gp}) & K(\hat{\beta}^{DOA}_{gp*}, \hat{\beta}^{DOA}_{gp*}) \end{bmatrix}), \tag{5.4}$$

where β_{gp*} is the predicted DOA estimation from GP. Then, the mean and covariance of the predicted DOA from the GP sensor model can be computed as follows,

$$\hat{\beta}_{gp*} = K(\hat{\beta}^{DOA}_{gp*}, \hat{\boldsymbol{\beta}}^{DOA}_{gp})(K(\hat{\beta}^{DOA}_{gp*}, \hat{\boldsymbol{\beta}}^{DOA}_{gp}) + \sigma_n^2 \mathbf{I})^{-1} \boldsymbol{\beta}_{gp}, \tag{5.5}$$

$$\begin{aligned} P^{\beta}_{gp*} =& K(\hat{\beta}^{DOA}_{gp*}, \hat{\beta}^{DOA}_{gp*}) - K(\hat{\beta}^{DOA}_{gp*}, \hat{\boldsymbol{\beta}}^{DOA}_{gp})(K(\\ & \hat{\boldsymbol{\beta}}^{DOA}_{gp}, \hat{\boldsymbol{\beta}}^{DOA}_{gp}) + \sigma_n^2 \mathbf{I})^{-1} K(\hat{\boldsymbol{\beta}}^{DOA}_{gp}, \hat{\beta}^{DOA}_{gp*}). \end{aligned} \tag{5.6}$$

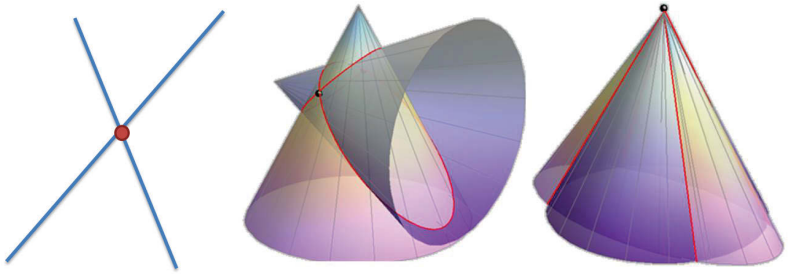
A squared exponential kernel function

$$k_{i,j} = \sigma_f^2 exp(-\frac{1}{2\ell^2}(\hat{\boldsymbol{\beta}}^{DOA}_{gp}(i) - \hat{\boldsymbol{\beta}}^{DOA}_{gp}(j))^2) \tag{5.7}$$

is used in our GP sensor model. In Eq. 5.7, $k_{i,j}$ 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,ℓ and σ_n as described in [142].

5.3 Initialisation of Sound Source 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. 5.3. This produces a partial observability which introduces a 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



a b

FIGURE 5.3: Intersection of two 3D bearings (a) and cone surfaces (b).

SLAM [110]. In monocular SLAM, visual point features parametrised by their Euclidean coordinates can be initialised after triangulating two 3D bearing observations as shown in Fig. 5.3 (a). However, intersection of two cone surfaces is not possible to model with simple Gaussian distribution as shown in Fig. 5.3 (b).

In order to initialise the sound source location, a multi-hypotheses strategy is required, which will allow us to model the uncertainty correctly. 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,

$$\mathbf{s}^m = (\beta^m, \gamma^m, \rho^m)^T. \tag{5.8}$$

Note that in Eq. 5.8, 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. 5.8, $\beta^m, \gamma^m, \rho^m$ are axis angle, circumferential angle and inverse depth of the sound source as shown in Fig. 5.2. As can be seen from the figure, the origin of the sensor coordinate frame is (x_r^m, y_r^m, z_r^m), the azimuth and elevation angle of positive Y axis are (θ_r^m, ϕ_r^m). 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 coordinates 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,

$$\mathbf{x}^m_{ss,axis} = (x^m_r, y^m_r, z^m_r, \theta^m_r, \phi^m_r)^T. \tag{5.9}$$

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 $\hat{\beta}^m_{gp*,ini}$ 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 IDP in the visual SLAM algorithm in [133] [136].

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 [133] [136], 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

$$\mathbf{s}^{m,i} = (\beta^{m,i}, \gamma^{m,i}, \rho^{m,i})^T, \tag{5.10}$$

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

$$\gamma^{m,i} = \frac{2\pi}{N_h} i - \pi, i \in (1 \cdots N_h). \tag{5.11}$$

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

$$\mathbf{P}_{ss}^{m,i} = \begin{bmatrix} P_{gp*,ini}^{\beta,m} & 0 & 0 \\ 0 & \left(\frac{\pi}{N_h}\right)^2 & 0 \\ 0 & 0 & \frac{1}{3d_{min}^2} \end{bmatrix}, \tag{5.12}$$

where $P_{gp*,ini}^{\beta,m}$ is the predicted variance of DOA angle $\hat{\beta}_{gp*,ini}^{m}$ using the GP sensor model. The covariance of the inverse depth is the same as suggested in [136]. The covariance of the circumferential angle is set to $(\pi/N_h)^2$ so that one sigma region of all hypotheses covers all possible ranges.

The advantage of the proposed parametrisation is shown in Fig. 5.4. 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 [133] is used, there exists a polygon effect when looking from the right side of the cone as shown in subfigure (c), especially when 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,

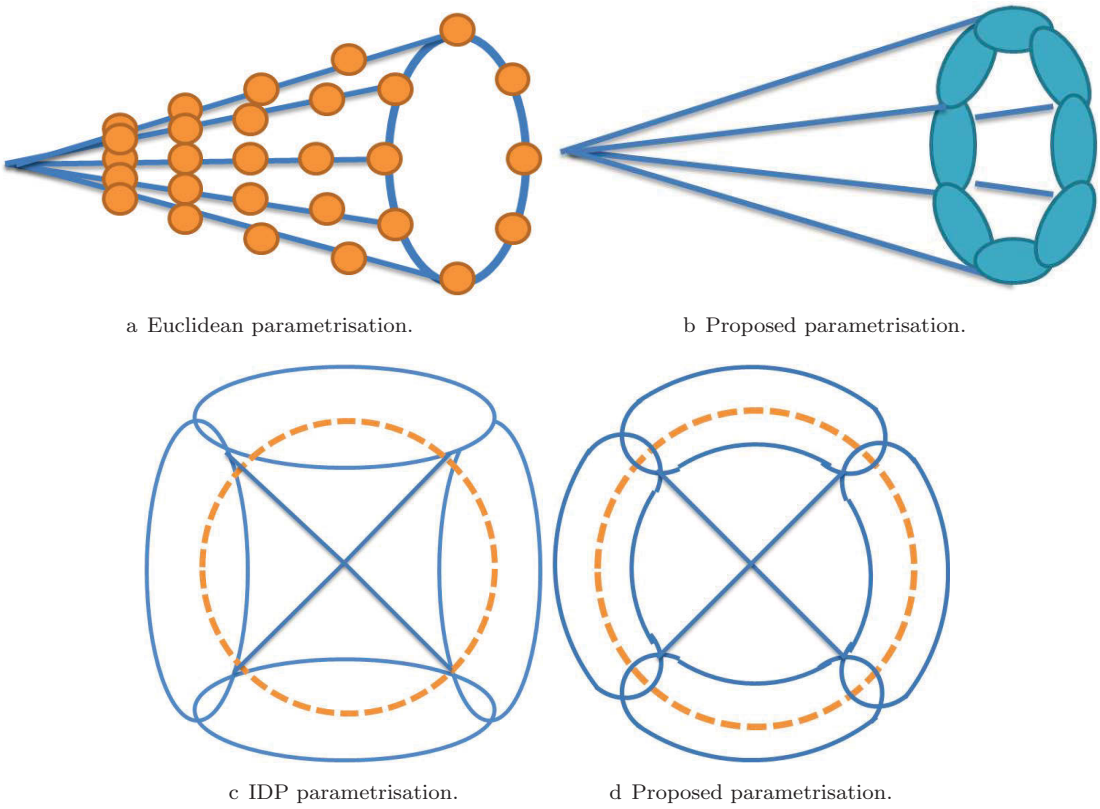
$$\hat{z}_k^{m,i} = atan2(\mathbf{p}_{l,k}^{m,i}(2), \sqrt{\mathbf{p}_{l,k}^{m,i}(1)^2 + \mathbf{p}_{l,k}^{m,i}(3)^2}), \tag{5.13}$$

$$z_k^{m,i} = \hat{\beta}_{gp*,k}^{m}, \tag{5.14}$$

$$Q_k^{m,i} = P_{gp*,k}^{\beta,m}, \tag{5.15}$$

$$\begin{aligned} \mathbf{K}_k^{m,i} =& \mathbf{P}_{ss,k-1}^{m,i}(\mathbf{H}_k^{m,i})^T / (\mathbf{H}_k^{m,i}\mathbf{P}_{ss,k-1}^{m,i} \\ & (\mathbf{H}_k^{m,i})^T + Q_k^{m,i}), \end{aligned} \tag{5.16}$$

$$\mathbf{s}_k^{m,i} = \mathbf{s}_{k-1}^{m,i} + \mathbf{K}_k^{m,i} f_{na}(z_k^{m,i} - \hat{z}_k^{m,i}), \tag{5.17}$$



a Euclidean parametrisation. b Proposed parametrisation.

c IDP parametrisation. d Proposed parametrisation.

FIGURE 5.4: Multi hypotheses using (a) Euclidean (c) IDP and (b),(d) the proposed parametrisation.

$$\mathbf{P}_{ss,k}^{m,i} = (\mathbf{I} - \mathbf{K}_k^{m,i}\mathbf{H}_k^{m,i})\mathbf{P}_{ss,k-1}^{m,i}, \tag{5.18}$$

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. $f_{na}(\bullet)$ 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

$$\begin{aligned}\mathbf{p}_{k-1}^{m,i} =& f_{eul\_mat}(\mathbf{x}_{ss,axis}^{m}(1), \mathbf{x}_{ss,axis}^{m}(2), \mathbf{x}_{ss,axis}^{m}(3),\\ & \mathbf{x}_{ss,axis}^{m}(4) - \pi/2, 0, \mathbf{x}_{ss,axis}^{m}(5))\\ & \begin{bmatrix} \frac{cos(\mathbf{s}_{k-1}^{m,i}(1))cos(\mathbf{s}_{k-1}^{m,i}(2))}{\mathbf{s}_{k-1}^{m,i}(3)} \\ \frac{sin(\mathbf{s}_{k-1}^{m,i}(1))}{\mathbf{s}_{k-1}^{m,i}(3)} \\ \frac{cos(\mathbf{s}_{k-1}^{m,i}(1))sin(\mathbf{s}_{k-1}^{m,i}(2))}{\mathbf{s}_{k-1}^{m,i}(3)} \\ 1 \end{bmatrix},\end{aligned} \tag{5.19}$$

$$\mathbf{p}_{l,k}^{m,i} = [\mathbf{I}_3\mathbf{0}]\mathcal{M}^{-1}(\mathbf{x}_{r,k})\mathbf{p}_{k-1}^{m,i}, \tag{5.20}$$

where function $f_{eul\_mat}(x_t, y_t, z_t, yaw_r, pitch_r, roll_r)$ transforms 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,

$$P_{z_k^{m,i}} = \mathbf{H}_k^{m,i}\mathbf{P}_{ss,k-1}^{m,i}(\mathbf{H}_k^{m,i})^T, \tag{5.21}$$

$$d_k^{m,i} = (f_{na}(\hat{z}_k^{m,i} - z_k^{m,i}))^T P_{z_k^{m,i}} f_{na}(\hat{z}_k^{m,i} - z_k^{m,i}). \tag{5.22}$$

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 [133] as follows,

$$\mathbf{h}_{XYZ,k}^{W,m,i} = \mathbf{x}_{ss,axis}^{m}(1:3) - \mathbf{p}_k^{m,i}, \tag{5.23}$$

$$\sigma_{\rho,k}^{m,i} = \sqrt{\mathbf{P}_{ss,k}^{m,i}(3,3)}, \tag{5.24}$$

$$\mathbf{m}_k^{m,i} = \frac{\mathbf{p}_k^{m,i} - \mathbf{x}_{ss,axis}^m(1:3)}{||\mathbf{p}_k^{m,i} - \mathbf{x}_{ss,axis}^m(1:3)||}, \tag{5.25}$$

$$\sigma_{d,k}^{m,i} = \frac{\sigma_{\rho,k}^{m,i}}{\mathbf{s}_k^{m,i}(3)}, \tag{5.26}$$

$$Ld_k^{m,i} = \frac{4 * \sigma_{d,k}^{m,i} ||(\mathbf{m}_k^{m,i})^T \mathbf{h}_{XYZ,k}^{W,m,i} ||\mathbf{h}_{XYZ,k}^{W,m,i}||^{-1}||}{||\mathbf{h}_{XYZ,k}^{W,m,i}||}, \tag{5.27}$$

where $\mathbf{p}_k^{m,i}$ can be computed in the same way as $\mathbf{p}_{k-1}^{m,i}$ in Eq. 5.19. 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.

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

A graph based SLAM [7] 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,

$$\mathbf{x} = [\mathbf{x}_{kf}^1, \cdots, \mathbf{x}_{kf}^{N_{kf}}, \mathbf{v}^1, \cdots, \mathbf{v}^{N_v}, \mathbf{p}^1, \cdots, \mathbf{p}^{N_s}]^T, \tag{5.28}$$

where $\mathbf{x}_{kf}^{n_{kf}}(n_{kf} = 1 \cdots N_{kf})$ 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 a 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 [7],

$$\hat{\mathbf{x}} = argmin \sum_{ij} \mathbf{e}_{ij}^T \mathbf{\Omega}_{ij} \mathbf{e}_{ij}, \tag{5.29}$$

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,

$$\mathbf{p}^{j,i} = [\mathbf{I}_3 \mathbf{0}] \mathcal{M}^{-1}(\mathbf{x}_{kf}^i) \begin{bmatrix} \mathbf{p}^j \\ 1 \end{bmatrix}, \tag{5.30}$$

$$e_{ij} = atan2(\mathbf{p}^{j,i}(2), \sqrt{\mathbf{p}^{j,i}(1)^2 + \mathbf{p}^{j,i}(3)^2}) - \hat{\beta}_{gp*}^{j,i}, \tag{5.31}$$

where $\mathbf{p}^{j,i}$ is the local coordinate of the jth sound source in the ith key frame's reference frame and $\hat{\beta}_{gp*}^{j,i}$ 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

$$\Omega_{ij} = (P_{gp*}^{\beta,j,i})^{-1}. \tag{5.32}$$

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 [110]. After all nodes and edges are defined, Eq. 5.29 can be solved by Gauss-Newton or Levenberg-Marquardt optimisation.

Regarding the real time implementation, following ORB-SLAM implementation [110], 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 [110].

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 for the partial angle observation of a linear microphone array. Without observing from several different poses, sound source location estimation is not guaranteed to converge.

5.5 Simulation and Experimental Results

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

5.5.1 Simulations of Sound Source Mapping with a Linear Microphone Array

In the simulation scenario shown in Fig. 5.5 and Fig. 5.6, a sensor with a RGBD camera and a linear microphone array for sound source mapping is simulated. 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 Fig. 5.5(a) and Fig. 5.5(b), blue circle markers represent initial multi hypotheses of sound sources. The sensor follows a 3D trajectory as shown in figure (d) and (e). It starts from the origin and travels along positive X axis direction. After 2m, it follows a 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 movement is repeated 4 times until it goes back to the origin. There are 8 sound sources in the simulation. The ground truth locations of them along with other simulation parameters can be found in Table 5.1. The sound source bearing observation noise is set to be a

Gaussian noise with a standard deviation of 10-20 degree. The linear microphone array is distributed along Y axis (represented by the red unit length line) of the sensor local coordinate frame. The sound bearing observation noise at different DOA angles is obtained empirically, and it is added to the ground truth value to be treated as a noisy observation. Specifically, the estimated the DOA angle using MUSIC algorithm and the ground truth DOA angle data is collected at each 5 degree interval of all possible range of DOA angle using a real 4 channel linear microphone array in the PS3-eye sensor. Then, the data is fitted into the Gaussian Process machine learning model to model the noise terms in the DOA estimation w.r.t. different true DOA angles. As can be seen from figure (a) and (b), 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 the 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. 5.6 (a) and (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.

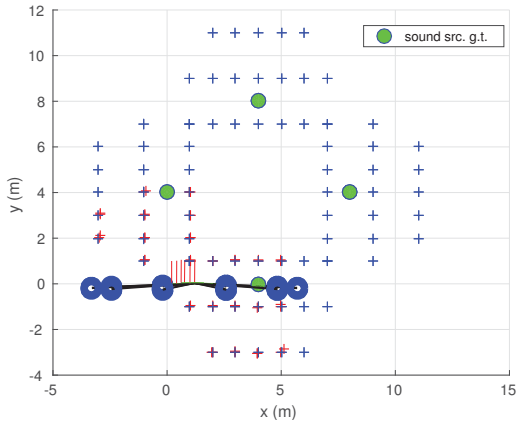
In the second simulation scenario, we validate the system performance when sound sources are mostly observed by the least sensitive region of a linear microphone array, which is at the two sides of the linear microphone array of DOA angle of ± 90 degrees. Locations of sound sources and the sensor trajectory is shown in Fig. 5.7. 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. As shown in the figure, the sensor starts from the origin, moves along the positive X axis direction. After 4m, it moves up and down, followed by another 4m along positive X axis and up and down movement. Then it moves another 4m along X axis and reaches the point (12,0,0).

TABLE 5.1: Parameters in simulation

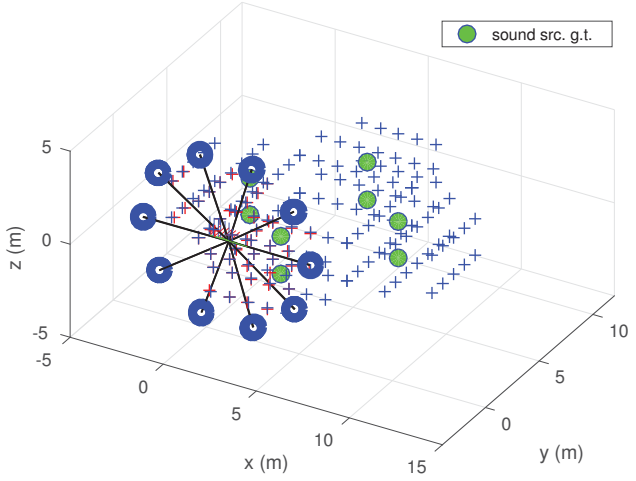
Parameters	Values
Number of initial hypotheses	10
Sound source ground truth	(4,0,1), (4,0,-1), (4,8,1), (4,8,-1), (8,4,1), (8,4,-1), (0,4,1), (0,4,-1)
Sound bearing estimation noise	10-20 deg
Mic. array max sensing distance	3m
Distance per odometry step	0.2m
Least square optimiser	Levenberg-Marquardt
RGBD landmark observation noise	1 deg and 0.01m
Sound source ground truth (2nd sim)	(4,1,0.5), (8,1,-0.5), (11,4,0.5), (11,8,-0.5), (8,11,0.5), (4,11,-0.5), (1,4,0.5), (1,8,-0.5)

This pattern is repeated 4 times until the sensor reaches the origin. Finally, it moves diagonally to generate another loop closure that better constrains the system. Positions of sound sources are shown in Table 5.1, other simulation parameters are the same as in the previous simulation. 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 are mostly observed by the highly sensitive region around 0 degree.

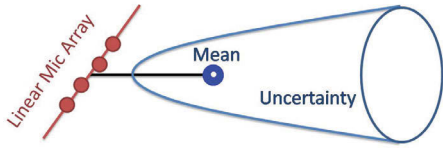
In the third set of simulation, we test the influence of the number of hypotheses over the final convergence of sound sources. 20 Monte Carlo runs of the first and second set of simulations are performed under various number of hypotheses. Mean convergence rate of the multi hypotheses filters, which is determined by the linearity index $Ld_k^{m,i}$ in Eq. 5.27, and RMS error of converged sound sources are shown in Fig. 5.8. From the figure, it can be seen that the number of hypotheses mainly affect the mean convergence rate and 6 or more hypotheses are suggested for better convergence. Regarding both convergence rate and sound source mapping accuracy, in terms of RMS error, the first set of simulations is



a Initialisation of multi hypotheses (top view).



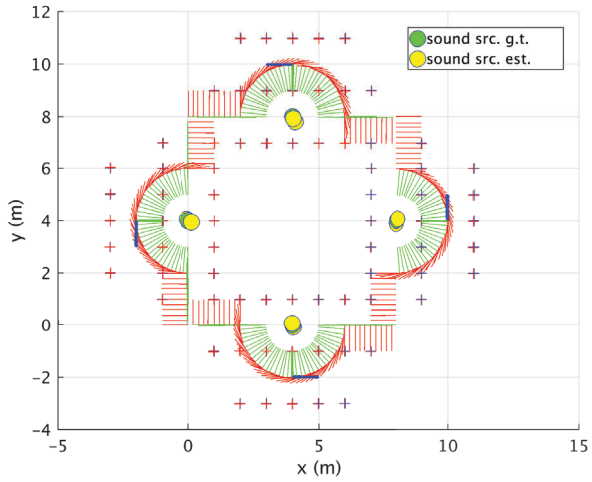
b Initialisation of multi hypotheses (side view).



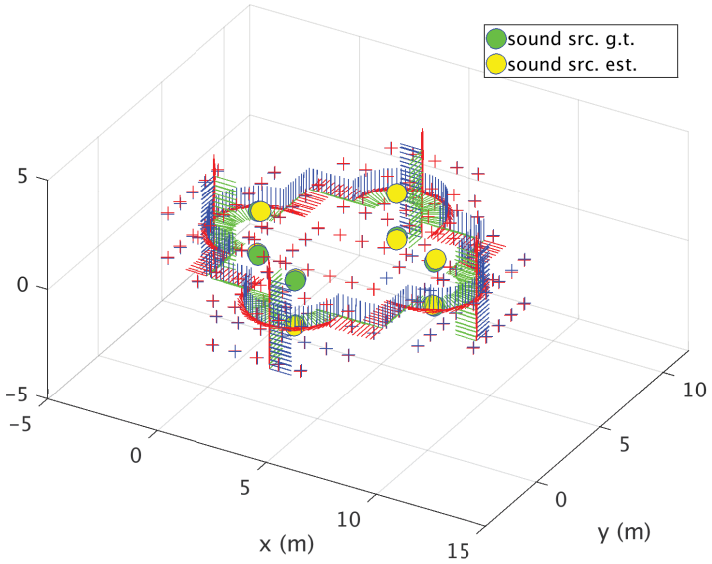
c Uncertainty value associated to each hypothesis during initialisation. The uncertainty of the sound source in each hypothesis extends to infinity when using inverse depth.

FIGURE 5.5: Initialisation of multi hypotheses. When the sensor first observes the sound source around 0 degree DOA angle, the cone surface approximates a plane and 10 hypotheses are uniformly distributed along the cone surface.

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 Final result of joint optimisation.

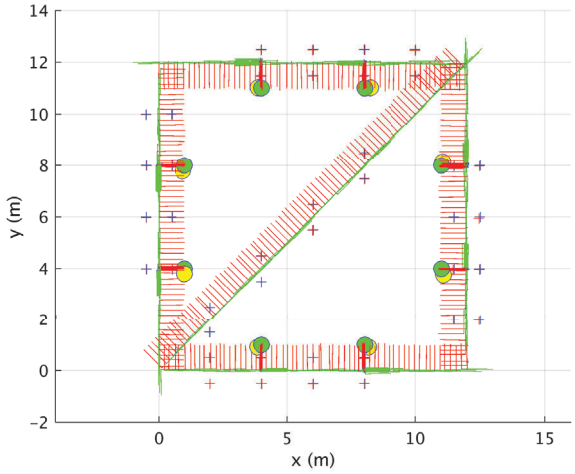


b Final result of joint optimisation.

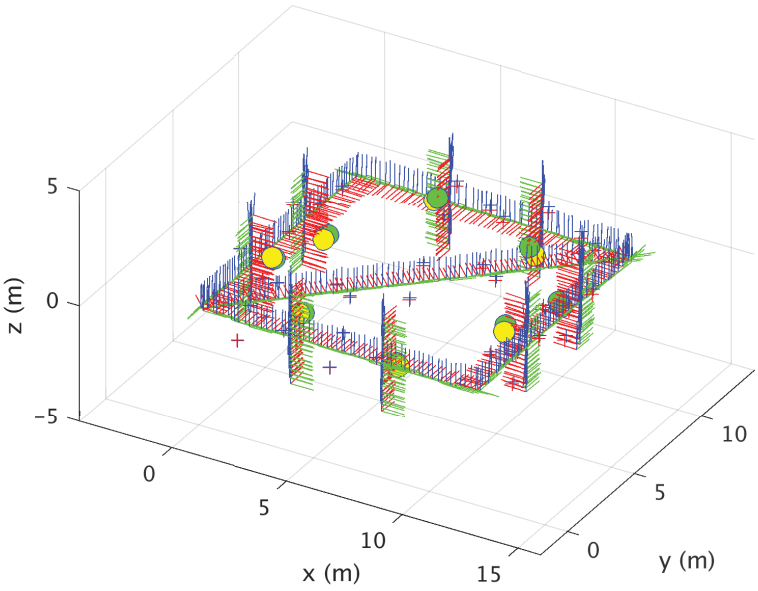
FIGURE 5.6: Final result of joint optimisation.

5.5.2 Experiments of Sound Source Mapping with a Linear Microphone Array

In this section, experimental results of sound source mapping using Kinect 360 and PS3 Eye, as shown in Fig 5.1 (a) and (c), are presented as examples of monocular and RGBD vision sensors respectively.

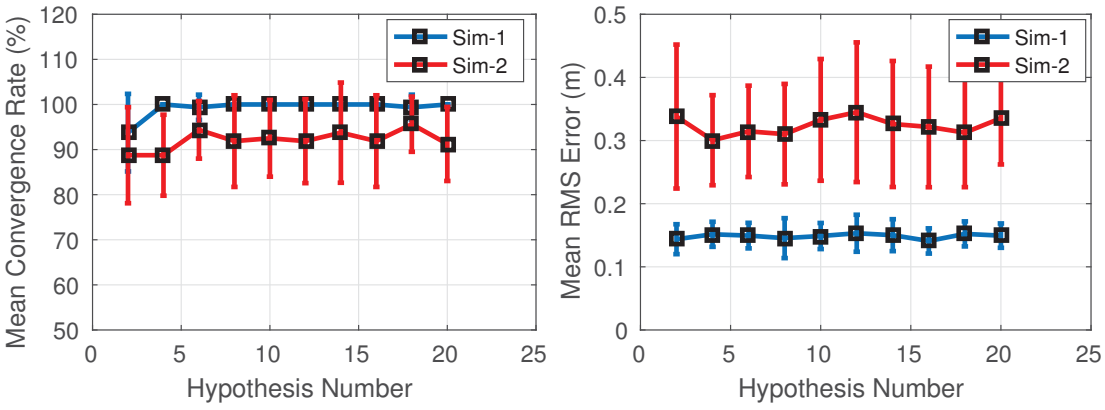


a Final result of joint optimisation.



b Final result of joint optimisation.

FIGURE 5.7: Final result of joint optimisation in the second trajectory.



a Mean convergence rate.

b Mean RMS error.

FIGURE 5.8: Mean convergence rate and RMS error over 20 Monte Carlo runs under various number of hypotheses.

Two experiments are conducted in a small office room and a computer lab as shown in Fig. 5.9 (a) and (b). 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 5.2. 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 music or a continuous human speech. The sampling frequency of the microphone array is at 16KHz. Sound source bearing estimation is performed at 5Hz. The sensors are handheld following a random trajectory around the sound sources. In Fig. 5.10 and Fig. 5.11, 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.

Results of mapping two sound sources are shown in Fig. 5.10. In (a) and (b), sound source map using Kinect RGBD sensor are presented, while in (c) and (d) mapping results using PS3 Eye are presented. Using Kinect RGBD sensor, a dense map of the environment can be obtained, whereas only a sparse map of the environment is obtained when using a

a Small office room.

b Computer lab.

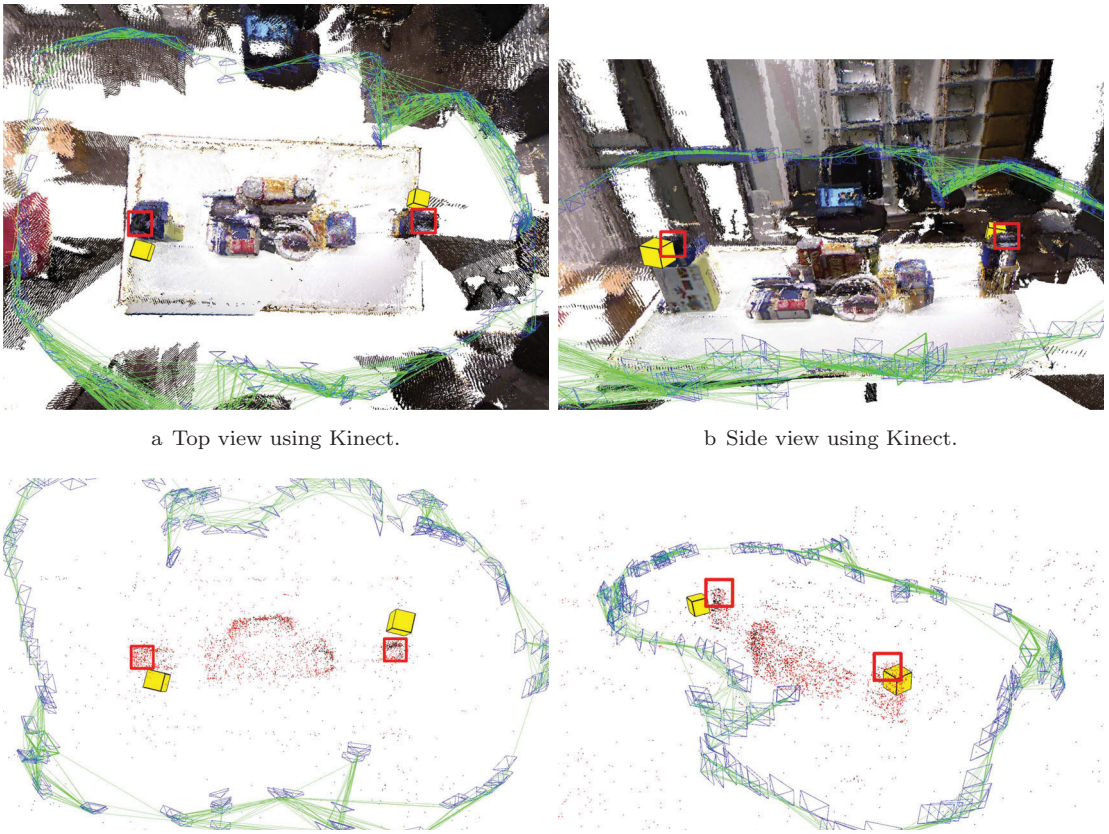
FIGURE 5.9: Experimental setup.

monocular camera. From these results, it can be seen that the proposed method can map sound sources with reasonably good accuracy. Mapping of five sound sources using Kinect RGBD sensor in a computer lab is shown in Fig. 5.11. The large image on top is the top view of the mapping result, while the five images at the bottom are the corresponding side view of five sound sources that are on top of each image. The result shows the proposed method performs well in a larger area. Covariances of sound sources are not shown in the figures for clarity, but they are consistent with the estimation errors. Note that a 3D microphone array provides 2 DOF measurement (azimuth and elevation angles), while a linear array can only provide 1 DOF measurement (axis angle). Most of the results in the literature use 3D microphone arrays and those results are not comparable with the results attained with a linear array. A video showing the performance of the proposed system during these two experiments is publicly available online[2].

5.6 Conclusion

In this Chapter, we presented a method for real-time 3D sound source 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 source locations for an online

[2]https://youtu.be/Ry_i3kmvlHM



a Top view using Kinect.

b Side view using Kinect.

c Top view using PS3 Eye.

d Side view using PS3 Eye.

FIGURE 5.10: Mapping of two sound sources using Kinect (RGBD sensor) and PS3 Eye (monocular camera).

optimisation 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 includes robust sound source data association and optimal active path planning to achieve better sound source mapping performance.

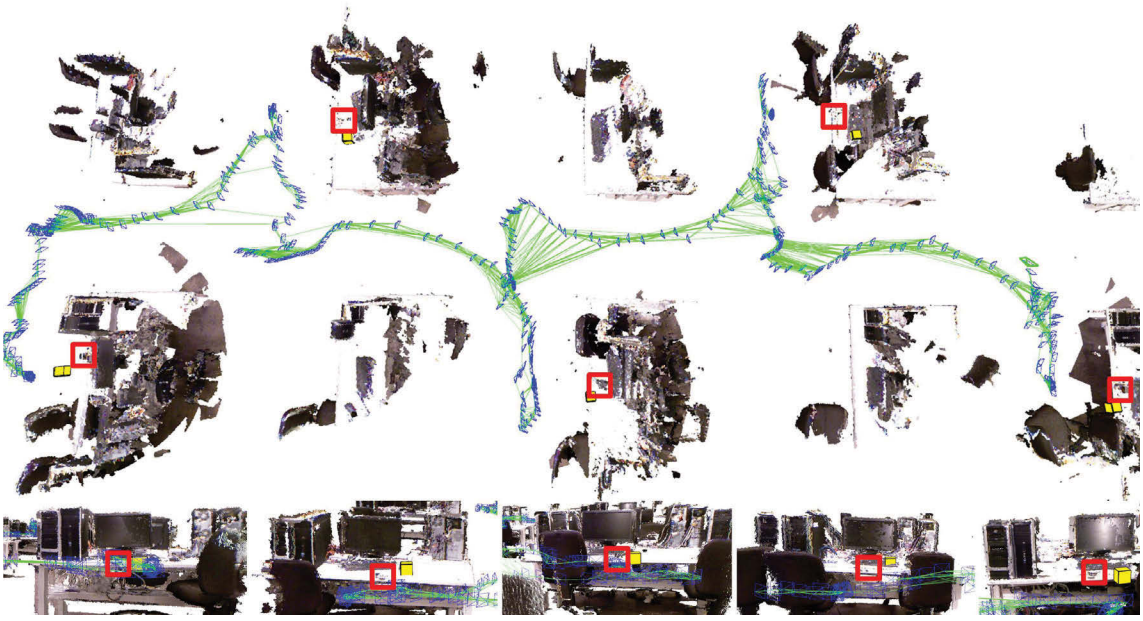


FIGURE 5.11: sound source mapping result in a computer lab.

Chapter 6

Conclusion

6.1 Summary of the Thesis

In this thesis, sound source mapping using a microphone array and calibration of a microphone array are studied. This thesis proposes three novel approaches in the field of sound source mapping. Ordered by their occurrence in this thesis, these contributions are as follows,

- **Calibration of a 2D/3D/linear microphone array**: A methodology is hereby proposed to calibrate a 2D/3D/linear microphone array using a graph-based optimisation method borrowed from the SLAM literature, effectively estimating the array geometry, time offset and clock difference/drift rate of each microphone (only the array structure for a hardware synchronised microphone array) together with the sound source locations. Simulation and experimental results are presented, which prove the effectiveness of the proposed methodology in achieving accurate estimates of the microphone array characteristics needed to be used on realistic settings with asynchronous sound devices.

- **Sound source mapping using a 2D/3D microphone array**: Firstly, we propose a least squares optimisation based SLAM framework to map stationary sound sources while simultaneously localising a moving robot. The proposed method jointly

estimates robot poses, positions of sound source and other landmarks, and hence is efficient in correlating robot trajectory with sound source mapping. Then an flexible and efficient method based on CI submap joining has been proposed to map sound source using a 2D/3D microphone array. This method exploits the conditional independence property between two maps estimated by two different SLAM algorithms running in parallel. The first map has the flexibility that it can be built with any off-the-shelf 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 IDP. Robot locations used during IDP initialisation become common features shared between the two SLAM maps, which allow the propagation of information accordingly. Since filtering techniques are proposed to build the second (sound) map, the suggested methodology has less computational complexity compared to the full joint optimisation method.

- **Sound source mapping using a linear microphone array**: We present a method for real-time 3D sound source mapping using an off-the-shelf robotic perception sensor equipped with a linear microphone array such as Kinect and PS3-Eye. In the proposed method, multi hypotheses filters are combined with a new sound source parametrisation to provide good initial guesses of sound source locations for an online optimisation strategy. A joint optimisation is carried out to estimate 6 DOF sensor poses and 3 DOF visual landmarks and sound source locations. In addition, a dedicated sensor model for a linear microphone array is proposed to model accurately the noise of the DOA observation.

6.2 Potential Future Work

All of the research topics covered in this thesis, including sound source mapping, bearing estimation and asynchronous microphone array calibration, are rapidly evolving research areas. It is possible that other methods and techniques may improve the results reported in this thesis. Particularly, the following topics are recommended for future research and development.

- **Data association**: The data association problem in the sound source mapping method presented in this thesis and many other works in the literature, assume either the IDs of sound sources to be known, or obtained by consistency of sound source bearing estimation or chi square test, as frequently used in the SLAM data association. However, there are problems with both of these methods in sound source mapping with a robot embedded microphone array scenario. Firstly, the consistency of sound sources bearing estimation only happens if the sound sources are continuously emitting sound. This is not likely to be always true in a real world scenario. If a sound source is active for a long period, becomes silent for a period, and then is activated again, using the the consistency of sound source bearing estimation, this sound source could be treated as two sound sources since the bearing information of this sound source has jumped a certain angle. Secondly, as a robot embedded microphone array typically estimates sound source bearing only, if two active sound sources stay on the same line, chi square test cannot distinguish between the two of them. We believe that using audio features of the sound source, this data association problem can be improved. Some sound features, such as the Mel Frequency Cepstrum Coefficient (MFCC) feature [143], can be used as a characteristic of a sound emitter, which is frequently used in speaker verification [143]. Therefore, it can help to determine the ID of a sound source when both of the methods mentioned above fail, especially in a situation in which all sound sources are human speech. For example, in [144], Alexandridis et. al. present a data association algorithm that finds the correct DOA association to the sources based on features extracted for each source. They propose the use of a feature that describes how the frequencies of the captured signals in each array are distributed to the sources. Their method results in high association and localization accuracy in scenarios with missed detections, reverberation, and noise and outperforms other recently proposed methods.

- **Ego motion noise suppression**: Throughout this thesis work, all simulations and experiments do not include or include only a little ego motion noise. However, this is not always true in many other application scenarios. One of the typical examples is an Unmanned Aerial Vehicle (UAV), such as a quadcopter. Due to the huge noise from a quadcopter's propellers, the sound source mapping result proposed in this

thesis could be severely influenced. Besides UAV, some large humanoid robots and some outdoor mobile robots, due to the inherently large noise from their motors and fans, also have considerable ego motion noise. Therefore, dealing with this ego motion noise is another challenge and potential future work. Some existing methods in [73, 145] can be tested.

- **Sound source mapping in a dynamic environment**: A static environment is assumed in this thesis. In a real world scenario, this assumption can hardly hold true. In most situations there are people walking around, doors opening and closing and objects being displaced. How to deal with these challenging non-static environments is a full research field in itself. Vision based mapping in dynamic environments [146, 147] has been studied in the past. These methods can be incorporated into the sound source mapping to deal with dynamic moving objects.

- **Multi-robot cooperative sound source mapping**: In this thesis, sound source mapping by a single robot scenario is considered. As a potential future work, sound source mapping using multiple robots is an interesting topic, specially for USAR scenarios, where in order to localise victims in a disaster area, it is much better to employ multiple robots to map the environment and position of victims rather than using only one robot. Therefore, how to effectively and cooperatively map an environment and all sound sources has become a challenging potential research area as these robots need to share information on an environmental and sound source map. Existing work on multi-robot SLAM [148, 149] and submap joining [134] can be a good starting point for this.

- **Active sound source mapping and exploration**: In this thesis, the robot (or sensor) motion are controlled by humans in all experiments. In a fully autonomous robotic application, a robot needs to plan its own path in addition to the sound source mapping. In this scenario, how to plan a path that minimises the uncertainty of sound source locations (exploitation) and maximises the exploration of an unknown area (exploration) is of key importance. Existing work in [120, 150] can be used to address this problem.

Bibliography

[1] Yuki Tamai, Satoshi Kagami, Yutaka Amemiya, Yoko Sasaki, Hiroshi Mizoguchi, and Tachio Takano. Circular microphone array for robot's audition. In *Proceedings of 2004 IEEE Sensors*, pages 565–570, 2004.

[2] Yoko Sasaki, Satoshi Kagami, and Hiroshi 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)*, pages 380–385, 2006.

[3] Florian Perrodin, Janosch Nikolic, Jol Busset, and Roland Siegwart. Design and calibration of large microphone arrays for robotic applications. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012)*, pages 4596–4601, 2012.

[4] Jwu-Sheng Hu, Chen-Yu Chan, Cheng-Kang Wang, Ming-Tang Lee, and Ching-Yi Kuo. Simultaneous localization of a mobile robot and multiple sound sources using a microphone array. *Advanced Robotics*, 25(1-2):135–152, 2011.

[5] Meysam Basiri, Felix Schill, Dario Floreano, and Pedro Lima. Audio-based relative positioning system for multiple micro air vehicle systems. In *Robotics: Science and Systems 2013 (RSS 2013)*, 2013.

[6] Cyrill Stachniss. Lecture notes in robot mapping, 2014.

[7] Giorgio Grisetti, Rainer Kummerle, Cyrill Stachniss, and Wolfram Burgard. A tutorial on graph-based SLAM. *IEEE Intelligent Transportation Systems Magazine*, 2(4):31–43, 2010.

[8] Raul Mur-Artal and Juan D. Tardos. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Transactions on Robotics*, PP(99):1–8, 2017.

[9] Ji Zhang and Sanjiv Singh. Low-drift and real-time lidar odometry and mapping. *Autonomous Robots*, 41(2):401–416, 2017.

[10] Nagasrikanth Kallakuri, Jani Even, Yoichi Morales, Carlos Ishi, and Norihiro Hagita. Probabilistic approach for building auditory maps with a mobile microphone array. In *2013 IEEE International Conference on Robotics and Automation (ICRA 2013)*, pages 2270–2275, 2013.

[11] Jani Even, Yaileth Morales, Nagasrikanth Kallakuri, Jonas Furrer, Carlos Toshinori Ishi, and Norihiro Hagita. Mapping sound emitting structures in 3D. In *2014 IEEE International Conference on Robotics and Automation (ICRA 2014)*, pages 677–682, 2014.

[12] Eric Martinson and Alan Schultz. Auditory evidence grids. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*, pages 1139–1144, 2006.

[13] Yoshiaki Sakagami, Ryujin Watanabe, Chiaki Aoyama, Shinichi Matsunaga, Nobuo Higaki, and Kikuo Fujimura. The intelligent ASIMO: System overview and integration. In *2002 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2002)*, pages 2478–2483, 2002.

[14] Stephen Brown. Meet Pepper, The Emotion Reading Robot. 2014.

[15] Jonathan Bohren, Radu Bogdan Rusu, E. Gil Jones, Eitan Marder-Eppstein, Caroline Pantofaru, Melonee Wise, Lorenz Mosenlechner, Wim Meeussen, and Stefan Holzer. Towards autonomous robotic butlers: Lessons learned with the pr2. In *2011 International Conference on Robotics and Automation (ICRA 2011)*, pages 5568–5575, 2011.

[16] Brian M Yamauchi. PackBot: a versatile platform for military robotics. In *Defense and Security*, pages 228–237. International Society for Optics and Photonics, 2004.

[17] Syamimi Shamsuddin, Hanafiah Yussof, Luthffi Ismail, Fazah Akhtar Hanapiah, Salina Mohamed, Hanizah Ali Piah, and Nur Ismarrubie Zahari. Initial response of autistic children in human-robot interaction therapy with humanoid robot NAO. In *2012 IEEE 8th International Colloquium on Signal Processing and its Applications (CSPA 2012)*, pages 188–193, 2012.

[18] Friedrich Fraundorfer, Lionel Heng, Dominik Honegger, Gim Hee Lee, Lorenz Meier, Petri Tanskanen, and Marc Pollefeys. Vision-based autonomous mapping and exploration using a quadrotor MAV. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012)*, pages 4557–4564, 2012.

[19] Jean-Marc Valin, Jean Rouat, and Franois Michaud. Enhanced robot audition based on microphone array source separation with post-filter. In *IEEE/RSJ International Conference on Intelligent Robots and Systems 2014 (IROS 2014)*, pages 2123–2128, 2004.

[20] K. Nakamura, K. Nakadai, F. Asano, and G. Ince. Intelligent Sound Source Localization and its application to multimodal human tracking. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011)*, pages 143–148, 2011.

[21] Kazuhiro Nakadai, Shunichi Yamamoto, Hiroshi G. Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. A robot referee for rock-paper-scissors sound games. In *The 2008 IEEE International Conference on Robotics and Automation (ICRA 2008)*, pages 3469–3474, 2008.

[22] Keisuke Nakamura, Kazuhiro Nakadai, and Gkhan Ince. Real-time super-resolution sound source localization for robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012)*, pages 694–699, 2012.

[23] Keisuke Nakamura, Kazuhiro Nakadai, Hirofumi Nakajima, and Gkhan Ince. Correlation matrix interpolation in sound source localization for a robot. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, pages 4324–4327, 2011.

[24] Gkhan Ince, Kazuhiro Nakadai, Tobias Rodemann, Hiroshi Tsujino, and Jun-Ichi Imura. Whole body motion noise cancellation of a robot for improved automatic speech recognition. *Advanced Robotics*, 25(11-12):1405–1426, 2011.

[25] M. B. Dehkordi. Sound Source Localization with CS Based Compressed Neural Network. *American Journal of Signal Processing*, 1(1):1–5, 2011.

[26] Ali Pourmohammad and Seyed Mohammad Ahadi. Real time high accuracy 3-D PHAT-based sound source localization using a Simple 4-Microphone Arrangement. *IEEE Systems Journal*, 6(3):455–468, 2012.

[27] Ana M. Torres, Maximo Cobos, Basilio Pueo, and Jose J. Lopez. Robust acoustic source localization based on modal beamforming and timefrequency processing using circular microphone arrays. *The Journal of the Acoustical Society of America*, 132(3): 1511–1520, 2012.

[28] Keonwook Kim and Anthony Choi. Binaural sound localizer for azimuthal movement detection based on diffraction. *Sensors*, 12(8):10584–10603, 2012.

[29] Yoko Sasaki, Naotaka Hatao, Kazuyoshi Yoshii, and Satoshi Kagami. Nested igmm recognition and multiple hypothesis tracking of moving sound sources for mobile robot audition. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013)*, pages 3930–3936, 2013.

[30] Koutarou Furukawa, Keita Okutani, Kohei Nagira, Takuma Otsuka, Katsutoshi Itoyama, Kazuhiro Nakadai, , and Hiroshi G. Okuno. Noise correlation matrix estimation for improving sound source localization by multirotor UAV. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013)*, pages 3943–3948, 2013.

[31] Carlos T. Ishi, Jani Even, and Norihiro 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)*, pages 3937–3942, 2013.

[32] Yilu Zhao, Xiong Chen, and Bin Wang. Real-time sound source localization using hybrid framework. *Applied Acoustics*, 74(12):1367–1373, 2013.

[33] Despoina Pavlidi, Anthony Griffin, Matthieu Puigt, and Athanasios Mouchtaris. Real-time multiple sound source localization and counting using a circular microphone array. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10): 2193–2206, 2013.

[34] Xiaofei Li and Hong Liu. Sound source localization for HRI using FOC-based time difference feature and spatial grid matching. *IEEE transactions on cybernetics*, 43(4): 1199–1212, 2013.

[35] Scott Kaghaz-Garan, Anurag Umbarkar, and Alex Doboli. Joint localization and fingerprinting of sound sources for auditory scene analysis. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014)*, pages 49–54, 2014.

[36] Xin Zhang, Enliang Song, JingChang Huang, Huawei Liu, YuePeng Wang, Baoqing Li, and Xiaobing Yuan. Acoustic Source Localization via Subspace Based Method Using Small Aperture MEMS Arrays. *Journal of Sensors*, 2014.

[37] Ui-Hyun Kim, Kazuhiro Nakadai, and Hiroshi G. Okuno. Improved sound source localization in horizontal plane for binaural robot audition. *Applied Intelligence*, 42(1):63–74, 2015.

[38] Nozomu Hamada and Ning Ding. Source separation and DOA estimation for underdetermined auditory scene. *Soundscape SemioticsLocalization and Categorization*, 1, 2014.

[39] Sylvain Argentieri, Patrick Dans, and Philippe Soures. A survey on sound source localization in robotics: From binaural to array processing methods. *Computer Speech & Language*, 34(1):87–112, 2015.

[40] Martin Rothbucher, Christian Denk, Martin Reverchon, Hao Shen, and Klaus Diepold. Robotic Sound Source Separation using Independent Vector Analysis. 2014.

[41] Gautam Narang, Keisuke Nakamura, and Kazuhiro Nakadai. Auditory-aware navigation for mobile robots based on reflection-robust sound source localization and visual slam. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2014)*, pages 4021–4026, 2014.

[42] Emmanuel Vincent, Rmi Gribonval, and Cdric Fvotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006.

[43] Kazuhiro Nakadai, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. Sound source separation of moving speakers for robot audition. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, pages 3685–3688, 2009.

[44] Ryu Takeda, Kazuhiro Nakadai, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Upper-limit evaluation of robot audition based on ICA-BSS in multi-source, barge-in and highly reverberant conditions. In *2010 IEEE International Conference on Robotics and Automation (ICRA 2010)*, pages 4366–4371, 2010.

[45] Gkhan Ince, Kazuhiro Nakadai, Tobias Rodemann, Hiroshi Tsujino, and Jun ichi Imura. Multi-talker speech recognition under ego-motion noise using Missing Feature Theory. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010)*, pages 982–987, 2010.

[46] Toru Takahashi, Kazuhiro Nakadai, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Missing-feature-theory-based robust simultaneous speech recognition system with non-clean speech acoustic model. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009)*, pages 2730–2735, 2009.

[47] Ryu Takeda, Kazuhiro Nakadai, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Step-size parameter adaptation of multi-channel semi-blind ICA with piecewise linear model for barge-in-able robot audition. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009)*, pages 2277–2282, 2009.

[48] Kazunori Komatani, Naoyuki Kanda, Mikio Nakano, Kazuhiro Nakadai, Hiroshi Tsujino, Tetsuya Ogata, and Hiroshi G. Okuno. Multi-domain spoken dialogue system with extensibility and robustness against speech recognition errors. In *the 7th SIGdial Workshop on Discourse and Dialogue*, pages 9–17, 2009.

[49] Randy Gomez, Tatsuya Kawahara, Keisuke Nakamura, and Kazuhiro Nakadai. Multi-party human-robot interaction with distant-talking speech recognition. In *The seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 439–446, 2012.

[50] Qiguang Lin, Ea-Ee Jan, and James Flanagan. Microphone arrays and speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(4):622–629, 1994.

[51] Darren C. Moore and Iain A. McCowan. Microphone array speech recognition: Experiments on overlapping speech in meetings. In *2003 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2003)*, pages 497–500, 2003.

[52] Takeshi Yamada, Satoshi Nakamura, and Kiyohiro Shikano. Robust speech recognition with speaker localization by a microphone array. In *1996 Fourth International Conference on Spoken Language*, pages 1317–1320, 1996.

[53] Eduardo Lleida, Julian Fernndez, and Enrique Masgrau. Robust continuous speech recognition system based on a microphone array. In *1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1998)*, pages 241–244, 1998.

[54] Keisuke Nakamura, Kazuhiro Nakadai, Futoshi Asano, Yuji Hasegawa, and Hiroshi Tsujino. Intelligent sound source localization for dynamic environments. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009)*, pages 664–669, 2009.

[55] Jean-Marc Valin, Franois Michaud, and Jean Rouat. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems*, 55(3):216–228, 2007.

[56] Ren C. Luo, Chien H. Huang, and Chun Y. Huang. Search and track power charge docking station based on sound source for autonomous mobile robot applications. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010)*, pages 1347–1352, 2010.

[57] Meysam Basiri, Felix Schill, Pedro U. Lima, and Dario Floreano. Robust acoustic source localization of emergency signals from micro air vehicles. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012)*, pages 4737–4742, 2012.

[58] Ryu Takeda, Kazuhiro Nakada, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Exploiting known sound source signals to improve ICA-based robot audition in speech separation and recognition. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2007)*, pages 1757–1762, 2007.

[59] Shunichi Yamamoto, Jean-Marc Valin, Kazuhiro Nakadai, Jean Rouat, Franois Michaud, Tetsuya Ogata, and Hiroshi G. Okuno. Enhanced robot speech recognition based on microphone array source separation and missing feature theory. In *The 2005 IEEE International Conference on Robotics and Automation (ICRA 2005)*, pages 1477–1482, 2005.

[60] Dong Yu and Li Deng. *Automatic Speech Recognition*. Springer, 2012.

[61] Karim Youssef, Sylvain Argentieri, and Jean-Luc Zarader. Binaural speaker recognition for humanoid robots. In *2010 11th International Conference on Control Automation Robotics & Vision (ICARCV 2010)*, pages 2295–2300, 2010.

[62] Hiroshi G. Okuno, Kazuhiro Nakadai, Ken ichi Hidai, Hiroshi Mizoguchi, and Hiroaki 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)*, pages 1402–1409, 2001.

[63] Kazuhiro Nakadai, Hiroshi G. Okuno, and Hiroaki Kitano. Real-time sound source localization and separation for robot audition. In *INTERSPEECH*, 2002.

[64] P. Pertila, Mikael Mieskolainen, and M. S. Hamalainen. Closed-form self-localization of asynchronous microphone arrays. In *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pages 139–144, 2011.

[65] Marius H. Hennecke and Gernot A. Fink. Towards acoustic self-localization of ad hoc smartphone arrays. In *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pages 127–132, 2011.

[66] H. Howard Fan and Chunpeng Yan. Asynchronous differential TDOA for sensor self-localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing 2007 (ICASSP 2007)*, pages II–1109–II–1112, 2007.

[67] Jr. Bove, V. Michael, and Ben Dalton. Audio-based self-localization for ubiquitous sensor networks. In *Audio Engineering Society Convention 118*, 2005.

[68] Daobilige Su, Teresa Vidal Calleja, and Jaime Valls Miro. Simultaneous asynchronous microphone array calibration and sound source localisation. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2015)*, pages 5561–5567, 2015.

[69] Daobilige Su, Jaime Valls Miro, and Teresa Vidal Calleja. Graph-SLAM Based Calibration of an Embedded Asynchronous Microphone Array for Outdoor Robotic Target Tracking. In *Assistive Robotics: Proceedings of the 18th International Conference on CLAWAR 2015*, pages 641–648, 2015.

[70] Daobilige Su, Teresa Vidal Calleja, and Jaime 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)*, pages 2000–2006, 2016.

[71] Daobilige Su, Teresa Vidal Calleja, and Jaime Valls Miro. Towards Real-Time 3D Sound Sources Mapping with Linear Microphone Arrays. In *2017 International Conference on Robotics and Automation (ICRA 2017)*, 2017.

[72] Tao Wu, Longji Sun, Qi Cheng, and Pramod K. Varshney. Fusion of multiple microphone arrays for blind source separation and localization. In *2012 IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM 2012)*, pages 173–176, 2012.

[73] Gkhan Ince, Kazuhiro Nakadai, and Keisuke Nakamura. Online learning for template-based multi-channel ego noise estimation. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012)*, pages 3282–3287, 2012.

[74] Keita Okutani, Takami Yoshida, Keisuke Nakamura, and Kazuhiro Nakadai. Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter. In *2012 IEEE/RSJ International Conference on on Intelligent Robots and Systems (IROS 2012)*, pages 3288–3293, 2012.

[75] Caleb Rascon and Luis A. Pineda. Lightweight multidirection-of-arrival estimation on a mobile robotic platform. In *the 2012 International Conference on Signal Processing and Imaging Processing*, 2012.

[76] Despoina Pavlidi, Anthony Griffin, Matthieu Puigt, and Athanasios Mouchtaris. Source counting in real-time sound source localization using a circular microphone array. In *2012 IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM 2012)*, pages 521–524, 2012.

[77] Ai Kijima, Yusuke Hioka, and Nozomu Hamada. Tracking of multiple moving sound sources using particle filter for arbitrary microphone array configurations. In *2012 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS 2012)*, pages 108–113, 2012.

[78] Gabriel Bustamante, Alban Portello, and Patrick Danes. A three-stage framework to active source localization from a binaural head. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, pages 5620–5624, 2015.

[79] Aly Magassouba, Nancy Bertin, and Franois Chaumette. Sound-based control with two microphones. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2015)*, pages 5568–5573, 2015.

[80] Franois Grondin and Franois Michaud. Time difference of arrival estimation based on binary frequency mask for sound source localization on mobile robots. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2015)*, pages 6149–6154, 2015.

[81] Kazuhiro Nakadai, Hirofumi Nakajima, Gkhan Ince, and Yuji Hasegawa. Sound source separation and automatic speech recognition for moving sources. In *2010*

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010), pages 976–981, 2010.

[82] Ryu Takeda, Kazuhiro Nakadai, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Speedup and performance improvement of ica-based robot audition by parallel and resampling-based block-wise processing. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010)*, pages 1949–1956, 2010.

[83] Hirofumi Nakajima, Kazuhiro Nakadai, Yuji Hasegawa, and Hiroshi Tsujino. Blind source separation with parameter-free adaptive step-size method for robot audition. *IEEE transactions on audio, speech, and language processing*, 18(6):1476–1485, 2010.

[84] Takeshi Mizumoto, Kazuhiro Nakadai, Takami Yoshida, Ryu Takeda, Takuma Otsuka, Toru Takahashi, and Hiroshi G. Okuno. Design and implementation of selectable sound separation on the Texai telepresence system using HARK. In *2011 IEEE International Conference on Robotics and Automation (ICRA 2011)*, pages 2130–2137, 2011.

[85] Toru Takahashi, Kazuhiro Nakadai, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Improvement in listening capability for humanoid robot HRP-2. In *2010 IEEE International Conference on Robotics and Automation (ICRA 2010)*, pages 470–475, 2010.

[86] Hirofumi Nakajima, Gkhan Ince, Kazuhiro Nakadai, and Yuji Hasegawa. An easily-configurable robot audition system using histogram-based recursive level estimation. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010)*, pages 958–963, 2010.

[87] Toru Takahashi, Kazuhiro Nakadai, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. An improvement in automatic speech recognition using soft missing feature masks for robot audition. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010)*, pages 964–969, 2010.

[88] Takami Yoshida, Kazuhiro Nakadai, and Hiroshi G. Okuno. Two-layered audio-visual speech recognition for robots in noisy environments. In *2010 IEEE/RSJ*

International Conference on Intelligent Robots and Systems (IROS 2010), pages 988–993, 2010.

[89] Ui-Hyun Kim, Takeshi Mizumoto, Tetsuya Ogata, and Hiroshi G. Okuno. Improvement of speaker localization by considering multipath interference of sound wave for binaural robot audition. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011)*, pages 2910–2915, 2011.

[90] Alban Portello, Patrick Danes, and Sylvain Argentieri. Acoustic models and Kalman filtering strategies for active binaural sound localization. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011)*, pages 137–142, 2011.

[91] Hamid Krim and Mats Viberg. Two decades of array signal processing researchThe parametric approach. *IEEE Signal processing magazine*, 13(4):67–94, 1996.

[92] Jacob Benesty, Jingdong Chen, and Yiteng Huang. Time-delay estimation via linear interpolation and cross correlation. *IEEE Transactions on speech and audio processing*, 12(5):509–519, 2004.

[93] Jingdong Chen, Jacob Benesty, and Yiteng Huang. Time delay estimation in room acoustic environments: an overview. *EURASIP Journal on applied signal processing*, pages 1–19, 2006.

[94] Sylvain Argentieri and Patrick 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)*, pages 2009–2014, 2007.

[95] Fabio Belloni and Visa Koivunen. Unitary root-MUSIC technique for uniform circular array. In *2003 Signal Processing and Information Technology (ISSPIT 2003)*, pages 451–454, 2003.

[96] Johan Xi Zhang, Mads Graesboll Christensen, Joachim Dahl, Soren Holdt Jensen, and Marc Moonen. Robust implementation of the MUSIC algorithm. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, pages 3037–3040, 2009.

[97] Carlos T. Ishi, Olivier Chatot, Hiroshi Ishiguro, and Norihiro Hagita. Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009)*, pages 2027–2032, 2009.

[98] Michael S. Brandstein and Harvey F. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1997)*, pages 375–378, 1997.

[99] Anthony Badali, Jean-Marc Valin, Franois Michaud, and Parham Aarabi. Evaluating real-time audio localization algorithms for artificial audition in robotics. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009)*, pages 2033–2038, 2009.

[100] Kazuhiro Nakadai, Toru Takahashi, Hiroshi G. Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. Design and Implementation of Robot Audition System'HARK'-Open Source Software for Listening to Three Simultaneous Speakers. *Advanced Robotics*, 24(5-6):739–761, 2010.

[101] Raviraj Adve. Lecture notes in smart antennas, 2007.

[102] Charles Knapp and Glifford Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976.

[103] Hoang Tran Huy Do. *Real-time SRP-PHAT source location implementations on a large-aperture microphone array.* PhD thesis, BROWN UNIVERSITY, 2009.

[104] Tukaram Baburao Lavate, V. K. Kokate, and A. M. Sapkal. Performance analysis of MUSIC and ESPRIT DOA estimation algorithms for adaptive array smart antenna in mobile communication. In *2010 Second International Conference on Computer and Network Technology (ICCNT 2010)*, pages 308–311, 2010.

[105] N. P. Waweru, D. B. O. Konditi, and P. K. Langat. Performance analysis of MUSIC, root-MUSIC and ESPRIT DOA estimation algorithm. *International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering*, 8(1): 209–216, 2014.

[106] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. MIT press, 2005.

[107] Timothy A. Davis. *Direct methods for sparse linear systems*. SIAM, 2006.

[108] Georg Klein and David Murray. Parallel tracking and mapping for small AR workspaces. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2007)*, pages 225–234, 2007.

[109] Georg Klein and David Murray. Improving the agility of keyframe-based SLAM. In *2008 European Conference on Computer Vision (ECCV 2008)*, pages 802–815, 2008.

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

[111] Jakob Engel, Thomas Schops, and Daniel Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *2014 European Conference on Computer Vision (ECCV 2014)*, pages 834–849, 2014.

[112] Felix Endres, Jurgen Hess, Jurgen Sturm, Daniel Cremers, and Wolfram Burgard. 3-D mapping with an RGB-D camera. *IEEE Transactions on Robotics*, 30(1):177–1873, 2014.

[113] Christian Kerl, Jurgen Sturm, and Daniel Cremers. Dense visual SLAM for RGB-D cameras. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (ICRA 2013)*, pages 2100–2106, 2013.

[114] Thomas Whelan, Renato F. Salas-Moreno, Ben Glocker, Andrew J. Davison, and Stefan Leutenegger. ElasticFusion: Real-time dense SLAM and light source estimation. *International Journal of Robotics Research*, 35(14):1697–1716, 2016.

[115] Wolfgang Hess, Damon Kohler, Holger Rapp, and Daniel Andor. Real-time loop closure in 2D LIDAR SLAM. In *2016 IEEE International Conference on Robotics and Automation (ICRA 2016)*, pages 1271–1278, 2016.

[116] Ji Zhang and Sanjiv Singh. Visual-lidar odometry and mapping: Low-drift, robust, and fast. In *2015 IEEE International Conference on Robotics and Automation (ICRA 2015)*, pages 2174–2181, 2015.

[117] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, 32(11): 1231–1237, 2013.

[118] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visualinertial odometry using nonlinear optimization. *International Journal of Robotics Research*, 34(3):314–334, 2015.

[119] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. On-Manifold Preintegration for Real-Time Visual–Inertial Odometry. *IEEE Transactions on Robotics*, 33(1):1–21, 2017.

[120] Emmanuel Vincent, Aghilas Sini, and Franois Charpillet. Audio source localization by optimal control of a mobile robot. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, pages 5630–5634, 2015.

[121] Yoko Sasaki, Satoshi Kagami, and Hiroshi Mizoguchi. Online short-term multiple sound source mapping for a mobile robot by robust motion triangulation. *Advanced Robotics*, 23(1-2):145–164, 2009.

[122] Michael Montemerlo, Sebastian Thrun, Daphne Koller, and Ben Wegbreit. FastSLAM: A factored solution to the simultaneous localization and mapping problem. In *AAAI Innovative Applications of Artificial Intelligence Conferences (AAAI/IAAI 2002)*, pages 593–598, 2002.

[123] Chieh-Chih Wang, Chi-Hao Lin, and Jwu-Sheng Hu. Probabilistic structure from sound. *Advanced Robotics*, 23(12-13):1687–1702, 2009.

[124] David A. Forsyth and Jean Ponce. *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference, 2002.

[125] Christine Evers, Alastair H. Moore, and Patrick A. Naylor. Acoustic simultaneous localization and mapping (a-SLAM) of a moving microphone array and its surrounding

speakers. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, pages 6–10, 2016.

[126] Hiroaki Miura, Takami Yoshida, Keisuke Nakamura, and Kazuhiro Nakadai. SLAM-based online calibration of asynchronous microphone array for robot audition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011)*, pages 524–529, 2011.

[127] Antonio Canclini, E. Antonacci, Augusto Sarti, and Stefano Tubaro. Acoustic source localization with distributed asynchronous microphone networks. *IEEE Transactions on Audio, Speech and Signal Processing*, 21(2):439–443, 2013.

[128] Vikas C. Raykar, B. Yegnanarayana, S. Prasanna, and Ramani Duraiswami. Speaker localization using excitation source information in speech. *IEEE Transactions on Speech and Audio Processing*, 13(5):751–761, 2005.

[129] Keisuke Hasegawa, Nobutaka Ono, Shigeki Miyabe, and Shigeki Sagayama. Blind estimation of locations and time offsets for distributed recording devices. *Latent Variable Analysis and Signal Separation*, pages 57–64, 2010.

[130] Giorgio Grisetti, Rainer Kummerle, Cyrill Stachniss, and Wolfram Burgard. A tutorial on graph-based SLAM. *IEEE Intelligent Transportation Systems Magazine*, 2(4):31–43, 2010.

[131] Jean Scholtz, Jeff Young, Jill L. Drury, and Holly A. Yanco. Evaluation of human-robot interaction awareness in search and rescue. In *2004 IEEE International Conference on Robotics and Automation (ICRA 2004)*, pages 2327–2332, 2004.

[132] David M. Cole and Paul M. Newman. Using laser range data for 3D SLAM in outdoor environments. In *2006 IEEE International Conference on Robotics and Automation (ICRA 2006)*, pages 1556–1563, 2006.

[133] Javier Civera, Andrew J. Davison, and JM Martinez Montiel. Inverse depth parametrization for monocular SLAM. *IEEE Transactions on Robotics*, 24(5):932–945, 2008.

[134] Pedro Pinis and Juan D. Tards. Large-scale slam building conditionally independent local maps: Application to monocular vision. *IEEE Transactions on Robotics*, 24(5): 1094–1106, 2008.

[135] Viorela Ila, Josep M. Porta, and Juan Andrade-Cetto. Information-based compact Pose SLAM. *IEEE Transactions on Robotics*, 26(1):78–93, 2010.

[136] Joan Sola, Teresa Vidal-Calleja, Javier Civera, and Jos Mara Martnez Montiel. Impact of landmark parametrization on monocular EKF-SLAM with points and lines. *International journal of computer vision*, 97(3):339–3683, 2012.

[137] Turtlebot. `https://www.clearpathrobotics.com/turtlebot-2-open-source-robot/`. [Online; accessed 19-02-2017].

[138] Clearpath robotics Clearpath Robotics Inc. `https://www.clearpathrobotics.com/`. [Online; accessed 19-02-2017].

[139] Jozef Kotus, Kuba Lopatka, and Andrzej Czyzewski. Detection and localization of selected acoustic events in acoustic field for smart surveillance applications. *Multimedia Tools and Applications*, 68(1):5–21, 2014.

[140] Lucas Adams Seewald, Luiz Gonzaga, Mauricio Roberto Veronez, Vicente Peruffo Minotto, and Claudio Rosito Jung. Combining SRP-PHAT and two Kinects for 3D Sound Source Localization. *Expert Systems with Applications*, 41(16):7106–7113, 2014.

[141] Hoang Do and Harvey F. Silverman. *SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data.* PhD thesis, 2010.

[142] Carl Edward Rasmussen. *Gaussian processes for machine learning.* 2006.

[143] Md Sahidullah and Goutam Saha. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech Communication*, 54(4):543–565, 2012.

[144] Anastasios Alexandridis, Giorgos Borboudakis, and Athanasios Mouchtaris. Addressing the data-association problem for multiple sound source localization using

doa estimates. In *2015 23rd European Signal Processing Conference (EUSIPCO 2015)*, pages 1551–1555, 2015.

[145] Gkhan Ince, Keisuke Nakamura, Futoshi Asano, Hirofumi Nakajima, and Kazuhiro Nakadai. Assessment of general applicability of ego noise estimation - Applications to Automatic Speech Recognition and Sound Source Localization -. In *2011 IEEE International Conference on Robotics and Automation (ICRA 2011)*, pages 3517–3522, 2011.

[146] Danping Zou and Ping Tan. Coslam: Collaborative visual slam in dynamic environments. *IEEE transactions on pattern analysis and machine intelligence*, 35(2): 354–366, 2013.

[147] Wei Tan, Haomin Liu, Zilong Dong, Guofeng Zhang, and Hujun Bao. Robust monocular SLAM in dynamic environments. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR 2013)*, pages 209–218, 2013.

[148] Maria Teresa Lazaro, Lina Mara Paz, Pedro Pinies, Jos A. Castellanos, and Giorgio Grisetti. Multi-robot SLAM using condensed measurements. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013)*, pages 1069–1076, 2013.

[149] Alexander Cunningham, Kai M. Wurm, Wolfram Burgard, and Frank Dellaert. Fully distributed scalable smoothing and mapping with robust multi-robot data association. In *2012 International Conference on Robotics and Automation (ICRA 2012)*, pages 1093–1100, 2012.

[150] Maani Ghaffari Jadidi, Jaime Valls Miro, and Gamini Dissanayake. Mutual information-based exploration on continuous occupancy maps. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2015)*, pages 6086–6092, 2015.