

# 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

<b>Declaration of Authorship</b>	<b>iii</b>
<b>Abstract</b>	<b>v</b>
<b>Acknowledgements</b>	<b>viii</b>
<b>List of Figures</b>	<b>xv</b>
<b>List of Tables</b>	<b>xix</b>
<b>Acronyms &amp; Abbreviations</b>	<b>xxi</b>
<b>Nomenclature</b>	<b>xxiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation and Scope . . . . .	6
1.1.1 Calibration of a Hardware-Synchronised/Asynchronous Microphone Array . . . . .	8
1.1.2 Sound Source Mapping by a Robot Embedded Microphone Array . .	10
1.1.3 Scope . . . . .	12
1.2 Contributions . . . . .	14
1.3 Publications . . . . .	16
1.3.1 Directly Related Publications . . . . .	16
1.4 Thesis Outline . . . . .	17
<b>2 Review of Related Work</b>	<b>19</b>
2.1 Sound Source Bearing Estimation using A Microphone Array . . . . .	19
2.1.1 Microphone Array . . . . .	19
2.1.2 Sound Sources Bearing Estimation . . . . .	23
2.1.2.1 MUSIC . . . . .	24
2.1.2.2 SRP-PHAT . . . . .	29
2.1.2.3 ESPRIT . . . . .	30
2.2 Simultaneous Localisation and Mapping . . . . .	32
2.2.1 EKF SLAM . . . . .	33

2.2.2	Graph Based SLAM . . . . .	36
2.2.3	SLAM Applications . . . . .	39
2.2.4	Monocular Visual SLAM . . . . .	39
2.2.5	Stereo Visual SLAM . . . . .	40
2.2.6	RGB-D Visual SLAM . . . . .	42
2.2.7	2D/3D Lidar based SLAM . . . . .	42
2.2.8	Visual Inertial SLAM . . . . .	43
2.3	Sound Source Mapping . . . . .	44
2.3.1	Robotic Sound Source Mapping with a Microphone Array Only . . . . .	44
2.3.1.1	Sound Source Mapping by Self-Motion Triangulation . . . . .	44
2.3.1.2	Sound Source Mapping using FastSLAM . . . . .	46
2.3.1.3	Sound Source Mapping using Unscented Kalman Filter (UKF) . . . . .	47
2.3.1.4	Sound Source Mapping using Single-Cluster Probability Hypothesis Density filter . . . . .	48
2.3.2	Robotic Sound Source Mapping with an Additional Exteroceptive Sensor . . . . .	49
2.3.2.1	Sound Source Mapping using Ray Tracing Method . . . . .	49
2.3.2.2	Sound Source Mapping using Auditory Evidence Grid Method . . . . .	52
<b>3</b>	<b>Calibration of a Microphone Array</b> . . . . .	<b>53</b>
3.1	Introduction . . . . .	53
3.2	Calibration of a 2D Asynchronous Microphone Array . . . . .	57
3.2.1	System Model . . . . .	57
3.2.2	Graph-Based Optimisation . . . . .	60
3.3	Calibration of a 3D Asynchronous Microphone Array . . . . .	64
3.4	Calibration of an Asynchronous Linear Microphone Array . . . . .	67
3.4.1	System Model . . . . .	67
3.4.2	Error Functions and Their Jacobians . . . . .	70
3.5	Simulation and Experimental Results . . . . .	71
3.5.1	Application Setup . . . . .	71
3.5.2	Initialisation and Termination Conditions . . . . .	71
3.5.3	Simulation Results . . . . .	72
3.5.3.1	Simulation Results of 2D Microphone Arrays . . . . .	72
3.5.3.2	Simulation Results of 3D Microphone Arrays . . . . .	74
3.5.3.3	Simulations of Calibration of 2D and 3D Asynchronous Microphone Arrays without Estimating Clock Difference . . . . .	78
3.5.3.4	Simulations of Calibration of an Asynchronous Linear Microphone Array . . . . .	81
3.5.4	Experimental Results . . . . .	84
3.5.4.1	Signal Processing . . . . .	85
3.5.4.2	Experimental Results of a 2D Microphone Array . . . . .	87
3.5.4.3	Experiment of Calibration of an Asynchronous Linear Microphone Array . . . . .	87
3.6	Conclusion . . . . .	88

<b>4</b>	<b>Sound Source Mapping using a 2D/3D Microphone Array</b>	<b>93</b>
4.1	Introduction . . . . .	93
4.2	Sound Source Mapping using a Least Squares Optimisation based SLAM Framework . . . . .	97
4.3	Sound Source Mapping by CI Submap Joining using a 2D/3D Microphone Array . . . . .	98
4.3.1	Structure of the Split CI Maps . . . . .	98
4.3.2	The Localisation Map . . . . .	101
4.3.3	The Sound Source Map . . . . .	101
4.3.4	Correlation Propagation . . . . .	103
4.4	Simulation and Experimental Results . . . . .	106
4.4.1	Simulation Results . . . . .	106
4.4.1.1	Sound Source Mapping with only Odometry Information . . . . .	106
4.4.1.2	Sound Source Mapping by a Least Squares Optimisation based SLAM Framework with Odometer and Range-Bearing Observations of Environment Landmarks . . . . .	107
4.4.1.3	Sound Source Mapping by CI Submap Joining Method with Odometer and Range-Bearing Observations of Environment Landmarks . . . . .	109
4.4.2	Experimental Results . . . . .	115
4.4.2.1	2D Sound Source Mapping by a Mobile Robot with a Microphone Array and a Laser Scanner . . . . .	116
4.4.2.2	3D Sound Source Mapping using a Hand Held PS3-eye (Monocular Camera with a Linear Microphone Array) . . . . .	117
4.5	Conclusion . . . . .	121
<b>5</b>	<b>Sound Source Mapping using a Linear Microphone Array</b>	<b>123</b>
5.1	Introduction . . . . .	123
5.2	Gaussian Processes to Model Linear Microphone Arrays Sensors . . . . .	125
5.3	Initialisation of Sound Source using Multi Hypotheses . . . . .	128
5.4	Joint Optimisation of Sensor Poses, Visual Landmarks and Sound Sources Locations . . . . .	134
5.5	Simulation and Experimental Results . . . . .	136
5.5.1	Simulations of Sound Source Mapping with a Linear Microphone Array . . . . .	136
5.5.2	Experiments of Sound Source Mapping with a Linear Microphone Array . . . . .	140
5.6	Conclusion . . . . .	143
<b>6</b>	<b>Conclusion</b>	<b>147</b>
6.1	Summary of the Thesis . . . . .	147
6.2	Potential Future Work . . . . .	148

**Bibliography**

# List of Figures

1.1	Asimo robot localises, separates and recognises simultaneous speech signals from three persons. . . . .	2
1.2	NAO robot focuses its attention on the person who is speaking. . . . .	3
1.3	Robot operating in an area full of smoke. . . . .	4
1.4	Separation of four simultaneous speech. . . . .	5
1.5	Pepper robot communication with customers with its ASR module. . . . .	6
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. . . . .	7
1.7	A typical microphone array structure and multi-channel ADC board for the sound source bearing estimation. . . . .	9
2.1	4-channel microphone array on Kinect 360. . . . .	20
2.2	4-channel microphone array on Kinect One. . . . .	20
2.3	4-channel microphone array on PS3 eye. . . . .	20
2.4	4-channel microphone array on PS4 eye. . . . .	21
2.5	A circular microphone array [1]. . . . .	21
2.6	A concentric microphone array in [2]. . . . .	22
2.7	A large scale 2D microphone array [3]. . . . .	22
2.8	A 3D microphone array in [4]. . . . .	23
2.9	A 3D microphone array in [5]. . . . .	23
2.10	Bayesian Network for EKF SLAM [6]. . . . .	34
2.11	Aspects of an edge connecting the vertex $\mathbf{x}_i$ and the vertex $\mathbf{x}_j$ [7]. . . . .	37
2.12	ORB_SLAM2 with stereo input: Trajectory and sparse reconstruction of an urban environment with multiple loop closures [8]. . . . .	41
2.13	ORB_SLAM2 with RGB-D input: Keyframes and dense pointcloud of a room scene with one loop closure [8]. . . . .	41
2.14	LOAM results of mapping university campus [9]. . . . .	43
2.15	Experimental setup of the self-motion triangulation method in [2]. . . . .	45
2.16	Experimental result of the self-motion triangulation method in [2]. . . . .	45
2.17	Experimental setup of the FastSLAM method in [4]. . . . .	46
2.18	Experimental result of the FastSLAM method in [4]. . . . .	47
2.19	Experimental result of the FastSLAM method in [4]. . . . .	48
2.20	Experimental result of the Single-Cluster Probability Hypothesis Density filter method in [125]. . . . .	48

2.21	Experimental setup of the ray tracing method in [10]. . . . .	50
2.22	Experimental result of the ray tracing method in [10]. . . . .	50
2.23	Experimental setup of the 3D ray tracing method in [11]. . . . .	51
2.24	Experimental result of the 3D ray tracing method in [11]. . . . .	51
2.25	Experimental result of the auditory occupancy grid method in [12]. . . . .	52
3.1	Experimental setup for testing clock differences. . . . .	54
3.2	Detected differences of peak arrival time. . . . .	56
3.3	Description of the poses, landmarks and constraints in the SLAM framework. . . . .	59
3.4	Description of the poses, landmarks and constraints. . . . .	68
3.5	Initialisation and final estimation results for a $3 \times 3$ array. . . . .	74
3.6	Final estimation results for a $3 \times 3$ array. . . . .	75
3.7	Estimation results of $3 \times 2$ and $4 \times 4$ arrays. . . . .	76
3.8	Estimation results of various number of sound source positions. . . . .	77
3.9	Initialisation and final estimation results for a $3 \times 3 \times 2$ array. . . . .	79
3.10	Final estimation results for a $3 \times 3 \times 2$ array. . . . .	80
3.11	Initialisation and final estimation results for a $3 \times 3 \times 2$ array. . . . .	81
3.12	Final estimation results for a $3 \times 3 \times 2$ array. . . . .	82
3.13	Calibration of 2D and 3D microphone arrays diverges when ignoring the clock difference. . . . .	83
3.14	Simulation results compared to the ground truth values. . . . .	84
3.15	Mean RMS error w.r.t. number of calibration data. . . . .	85
3.16	Experimental setup of the asynchronous microphone array. Each channel of the array is sampled independently using individual USB sound cards. . . . .	85
3.17	Pre signal processing and detection of signal arrival (plot below) for raw audio data (plot above). . . . .	86
3.18	Experiments results of a $2 \times 3$ array. . . . .	88
3.19	Experimental results of a $2 \times 3$ array. . . . .	89
3.20	Experimental setup of the asynchronous microphone array. . . . .	90
3.21	DOA estimation results after the calibration. . . . .	90
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. . . . .	99
4.2	Modified Bayesian network that describes probabilistic dependency between SLAM variables in two maps. . . . .	100
4.3	EKF parametrised by IDP with highly accurate odometry information. . . . .	108
4.4	Least square optimisation with highly accurate odometry information. . . . .	109
4.5	RMS errors and convergence rates under different odometry noise. . . . .	110
4.6	2D sound source mapping by the least square optimisation based SLAM framework. . . . .	111
4.7	3D sound source mapping by the least square optimisation based SLAM framework. Initialisation of the system. . . . .	112



4.8	3D sound source mapping by the least square optimisation based SLAM framework. Final estimation results. . . . .	113
4.9	sound source mapping with additional range-bearing observations before loop closure. . . . .	114
4.10	sound source mapping with additional range-bearing observations after loop closure. . . . .	115
4.11	Mean RMS errors with STD of 10 Monte Carlo runs under various length of robot trajectories. . . . .	116
4.12	Turtelbot equipped with a laser scanner and a Microcone (circular microphone array). . . . .	117
4.13	2D sound source mapping results using a mobile with laser scanner. . . . .	117
4.14	PS3-eye configuration (a) and experimental setup (b). . . . .	119
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. . . . .	119
4.16	3D sound source mapping results using a hand hold PS3-eye (monocular camera with linear microphone array). . . . .	120
5.1	Typical robotic sensors that include a linear microphone array. . . . .	124
5.2	Linear microphone array notation and parametrisation of a 3D sound source location. . . . .	126
5.3	Intersection of two 3D bearings (a) and cone surfaces (b). . . . .	129
5.4	Multi hypotheses using (a) Euclidean (c) IDP and (b),(d) the proposed parametrisation. . . . .	132
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. . . . .	139
5.6	Final result of joint optimisation. . . . .	140
5.7	Final result of joint optimisation in the second trajectory. . . . .	141
5.8	Mean convergence rate and RMS error over 20 Monte Carlo runs under various number of hypotheses. . . . .	142
5.9	Experimental setup. . . . .	143
5.10	Mapping of two sound sources using Kinect (RGBD sensor) and PS3 Eye (monocular camera). . . . .	144
5.11	sound source mapping result in a computer lab. . . . .	145



# List of Tables

3.1	Parameters setting in simulation . . . . .	73
3.2	RMS errors over 10-run Monte Carlo simulations . . . . .	73
3.3	RMS errors over 10-run Monte Carlo simulations . . . . .	78
3.4	Parameters setting in simulation . . . . .	84
3.5	Experimental set-up parameters . . . . .	86
3.6	Parameters setting in experiment . . . . .	90
4.1	Parameters in simulation . . . . .	107
5.1	Parameters in simulation . . . . .	138



# Acronyms & Abbreviations

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

<b>TOF</b>	Time of flights
<b>UKF</b>	Unscented Kalman Filter
<b>USAR</b>	Urban Search and Rescue
<b>USB</b>	Universal Serial Bus
<b>UTS</b>	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 $n$ th microphone to the origin of the microphone array coordinate.
$d_{i,k}$	The distance between the $i$ th microphone and the sound source at time instance $k$ .
$d_k$	The distance from the sound source position at the $k$ th time instance to the origin of the global coordinate frame.
$d_k^{m,i}$	The chi-square distance of the $i$ th hypothesis of the $m$ th 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.

$f_{an}$	$n$ th landmark observed by an additional sensor.
$f_{sn}$	$n$ th 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 $i$ th hypothesis of the $m$ th 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 $m$ th sound source.
$\mathbf{p}_{l,k}^{m,i}$	The Euclidean coordinate of the $m$ th sound source in $i$ th hypothesis under sensor local coordinate at time instance $k$ .
$\mathbf{p}^{j,i}$	The local coordinate of the $j$ th sound source in the $i$ th 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 $\phi$ .
$P(x_{srp})$	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 $\mathbf{x}_A$ .
$P_{C_A}$	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 $\phi_m$ is its direction.
$\mathbf{s}^m$	The proposed novel parametrisation of the $m$ th sound source state for a linear array.
$\mathbf{s}^{m,i}$	The state of the $m$ th sound source in $i$ th 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 $n_v$ th 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 $n$ th microphone.
$\mathbf{x}_{lm}^s(i)$	The state of the $i$ th 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 $n_s$ th sound source.
$\mathbf{x}_{lm}^a(n)$	State of the $n$ th 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.
$\tilde{\mathbf{x}}^a$	Rearranged state vector of the sound map.
$\tilde{\mathbf{x}}^a$	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 $m$ th 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 $n$ th landmark using an additional sensor.
$z_{sn}$	Observation of $n$ th 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 $k$ .
$\alpha^m$	Complementary angle of $\beta^m$ , the DOA angle of the $m$ th sound source.
$\beta_k$	The DOA angle of the sound source at time instance $k$ for the calibration of a linear array.
$\beta^m$	DOA angle of the $m$ th sound source.
$\hat{\beta}_k^{m,DOA}$	The estimated sound DOA angle of the $m$ th 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.
$\beta_{gp*}$	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 $m$ th 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.
$\hat{\beta}_{gp}^{DOA}$	The set of raw results from the DOA estimation algorithm that are used as function input when training the Gaussian Process sensor model.

---

$\beta_{gp}$	The set of ground truth DOA angles that are used as function output when training the Gaussian Process sensor model.
$\gamma^m$	The circumferential angle of the $m$ th sound source with the proposed novel parametrisation for a linear array.
$\rho^m$	The inverse depth of the $m$ th sound source with the proposed novel parametrisation for a linear array.
$\lambda_m$	The corresponding eigenvalue for $\mathbf{q}_m^{mic}$ .
$\phi$	The angle between the direction being searched and the direction from the origin of the microphone array to the $n$ th microphone.
$\omega$	The angular frequency.
$\mathbf{\Omega}_{ij}$	The information matrix of a virtual measurement between the node $i$ and the node $j$ in graph based SLAM.
$\tau_{lk}$	The TDOA from point $x_{srp}$ to $l$ th channel of the microphone array and $k$ th channel of the microphone array.
$\tau_k$	TOF from point $x_{srp}$ to the $k$ th channel of the microphone array.

