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 ${\sigma^{mic}}^2\mathbf{I}$.
$p(x_t^R, m \mid z_{1:t}, u_{1:t})$	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.
$\mathbf{z}^{p-p}_{k-1,k}$	The observation of the position-position constraint between the sound source at time instance $k-1$ and k.
$\hat{\mathbf{z}}^{p-p}_{k-1,k}$	The observation of the position-position constraint between the sound source at time instance $k-1$ and k.
$\mathbf{z}^{p-l}_{k}$	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.
$\hat{\mathbf{z}}^{p-l}_{k}$	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.
$\hat{\beta}^{m,DOA}_{k}$	The estimated sound DOA angle of the mth sound source from a DOA estimation algorithm.
$\hat{\beta}^{DOA}_{gp*}$	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 $\hat{\beta}^{DOA}_{gp*}$.
$\hat{\beta}^{m}_{gp*,ini}$	The predicted DOA angle from the Gaussian Process sensor model at the first observation of the mth sound source.
$\hat{\beta}^{j,i}_{gp*}$	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.
$\hat{\boldsymbol{\beta}}^{DOA}_{gp}$	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.