AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

Contact : ddoc-theses-contact@univ-lorraine.fr

LIENS

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10
http://www.cfcopies.com/V2/leg/leg_droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Cartographie d'un environnement sonore par un robot mobile

∴ ∵ ∴

Mapping of a sound environment by a mobile robot

THÈSE

présentée et soutenue publiquement le 3 novembre 2017

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

NGUYEN Van Quan

Composition du jury

Rapporteurs :	Patrick Danès	Professeur, Université Paul Sabatier, Toulouse
	David Filliat	Professeur, ENSTA ParisTech, Palaiseau
Examinateurs :	Nancy Bertin	Chargée de recherche CNRS, Irisa, Rennes
	Dominique Martinez	Chargé de recherche CNRS, Loria, Nancy
Directeurs de thèse :	Emmanuel Vincent	Directeur de recherche, Inria, Nancy
	Francis Colas	Chargé de recherche, Inria, Nancy

Institut National de Recherche en Informatique et en Automatique
Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Inria INVENTEURS DU MONDE NUMÉRIQUE

Loria Laboratoire lorrain de recherche en informatique et ses applications

Mis en page avec la classe thesul.

Acknowledgements

First and foremost I wish to express my sincere gratitude to my advisors Emmanuel Vincent and Francis Colas for giving me an excellent opportunity to conduct research, for continuous support, and for patient correction of my research work during these past three years. Their great inspiration, precise guidance and constant encouragement have helped me a lot to improve important skills of a researcher. I have learnt a great deal of problem exploration and solving from them. They have taught me how to develop half-baked ideas into complete and practical contributions which, in the end, need to be presented in convincing arguments. I could not have imagined having better advisors for my PhD study.

My sincere thanks also go to François Charpillet who let me join LARSEN team, and who gave access to the laboratory and research facilities that made this research possible.

I also want to acknowledge the jury members of this thesis who have accepted to evaluate my work, Patrick Danès, David Filliat, Nancy Bertin, and Dominique Martinez, for their reviewing efforts and insightful remarks.

I thank all the current and former members of LARSEN and MULTISPEECH team that I have had chance to interact, discuss, and share with in a wonderful environment: Iñaki Fernandez, Matthieu Zimmer, Nassim Kaldé, Mihai Andries, Abdallah Dib, Adrien Malaisé, Vassilis Vassiliades, Konstantinos Chatzilygeroudis, Adrian Bourgaud, Oriane Dermy, Rituraj Kaushik, Yassine El-Khadiri, Thomas Moinel, Dorian Goepp, Mélanie Lelaure, Nicolas Beaufort, Maxime Rio, Théo Biasutto-Lervat, Xuan-Son Nguyen, Serena Ivaldi, Olivier Buffet, Vincent Thomas, Aditya Arie Nugraha, Sunit Sivasankaran, Dung Tran, Antoine Liutkus, Yann Salaün, Imran Sheikh, Nathan Souviraà-Labastie, and many others. People here are genuinely nice and want to help you out and I'm glad to have interacted with many.

I thank our assistants, Véronique Constant and Sophie Drouot, who helped organise all the missions to conferences and summer schools.

At last, I would like to thank all my family, especially my parents, brother and sister, to continuously keep watch on me and offer me the state of mind for running up this thesis. I thank my lovely daughter Mélodie (Nhat An Thu Nguyen) who always brings me pure joys and happiness. I want to thank my best friend, soul-mate, and wife Cuc Dang, for her patience, support and understanding during my thesis. You and Mélodie have been the main motivation of all the efforts I have realized during this thesis. These past several years have not been an easy ride, both academically and personally. And there is not enough word here to describe how grateful I am to have both of you by my side.

I dedicate this thesis to my family, my wife Cuc, and my little daughter Mélodie
for their constant support and unconditional love.
I love you all dearly.

Abstract

Robot audition provides a hearing capability for robot and helps it to explore and understand a sound environment. In this thesis, we focus on the task of sound source localization using a mobile robot equipped with a microphone array. We consider source localization for a single or multiple, intermittent, and possibly moving sources in a reverberant environment and exploiting robot motion to improve the source localization.

We first propose a Bayesian filtering framework to localize the position of a single, intermittent and possibly moving sound source. This framework jointly estimates the source location and its activity over time and can apply to any microphone array geometry. Thanks to the movement of the robot, it can estimate the distance to the source and also can solve the front-back ambiguity which appears in the case of a linear microphone array. We conduct a number of experiments to show the robustness of the extended mixture Kalman filter (MKF) framework to false measurements of the angle of arrival (AoA) or the source activity detection (SAD) when localizing an intermittent, moving source. Experiments and statistical results also show that our proposed MKF method outperforms the method that does not consider the source activity in the model. In addition, we propose a particle filtering technique for jointly estimating the source location and its activity and comparing with the performance of the extended MKF in term of localization performance, and computational time.

After the contribution on a single source, we extend it to the context of multiple intermittent, possibly moving sources. In the context of multiple sound sources, the uncertainty is not only increased when the sources are inactive but also when they are active because we have no information to tell an AoA measurement is originated from which target. The peaks of the multiple signal classification (MUSIC) spectrum may be occluded and false alarms may occur. By implementing an extended MKF with joint probabilistic data association filter (JPDAF), we can jointly estimate the two source locations and their activities over time. The experimental evaluation shows the effectiveness of the proposed method in localizing and tracking two target sources in reverberant environment.

Lastly, we make a contribution on long-term robot motion planning to optimally reduce the uncertainty on estimating the location of a single, intermittent, and moving source. We define a cost function for long-term planning with two alternative criteria: the Shannon entropy or the standard deviation of the estimation belief on the source location. These entropies or standard deviations are integrated over time with a discount factor. We represent the belief about the source location at each time step by a mixture of Gaussians, and propagate this belief by the proposed extended MKF framework. Then, we adapt the Monte Carlo tree search (MCTS) method for efficiently finding the optimal robot motion that will minimize the above cost function. Experiments show that the proposed method outperforms other robot motion planning methods for robot audition in the long run, especially compared to a greedy planning method. The results of the MCTS with standard deviation highlight the effect of the discount factor on the performance of MCTS over time. Moreover, the analysis of all the results show a coherent estimation error result when optimizing the standard deviation of the estimated belief instead of the entropy.

Résumé

L'audition est une modalité utile pour aider un robot à explorer et comprendre son environnement sonore. Dans cette thèse, nous nous intéressons à la tâche de localiser des sources sonores à l'aide d'un robot mobile équipé d'une antenne de microphones. Nous considérons la localisation d'une ou plusieurs sources intermittentes et mobiles dans un environnement réverbérant en exploitant les mouvement du robot pour améliorer la localisation.

Nous proposons d'abord un modèle bayésien pour localiser une seule source intermittente mobile. Ce modèle estime conjointement la position et l'activité de la source au cours du temps et s'applique à tout type d'antennes. Grâce au mouvement du robot, il peut estimer la distance de la source et résoudre l'ambigüité avant-arrière qui apparait dans le cas des antennes linéaires. Nous conduisons des expériences pour montrer la robustesse d'un modèle de filtre de Kalman basé sur des mixtures (MKF) à de mauvaises observations de l'angle d'arrivée (AoA) ou de l'activité (SAD). Nous montrons en particulier que notre méthode surpasse les méthodes qui ne considèrent pas explicitement l'activité de la source. De plus, nous proposons une implémentation de notre modèle sous forme de filtre à particule que nous comparons au MKF en termes de performance et de temps de calcul.

Après ces contributions pour une seule source, nous étendons note modèle à plusieurs sources intermittentes et mobiles. Dans ce cas, l'incertitude n'advient pas seulement à cause des périodes d'inactivité mais aussi à cause de l'association entre une observation d'un angle d'arrivée et d'une source. De plus l'analyse des angles d'arrivée est plus difficile et des fausses alarmes peuvent avoir lieu. En combinant notre MKF avec un *joint probability data association filter* (JPDAF), nous pouvons estimer conjointement les positions et activités de deux sources sonore dans un environnement réverbérant.

Enfin, nous faisons une contribution à la planification de mouvement pour réduire l'incertitude sur l'estimation de la position d'une source sonore mobile et intermittente. Nous définissons une fonction de cout avec l'alternative entre deux critères : l'entropy de Shannon ou l'écart-type sur l'estimation de la position. Ces deux critères sont intégrés dans le temps avec un facteur d'actualisation. Nous représentons la distribution sur la position par une mixture de gaussiennes et réalisons l'estimation avec notre modèle de MKF. Nous adaptons alors l'algorithme de *Monte-Carlo tree search* (MCTS) pour trouver, efficacement, le mouvement du robot qui minimise notre fonction de cout, Nos expériences montrent que notre méthode surpasse, sur le long terme, d'autres méthodes de planification pour l'audition robotique, notamment une approche gloutonne. Nous montrons en particulier l'effet du facteur d'actualisation sur la performance au cours du temps. Enfin l'analyse des résultats montre une estimation cohérente en optimisant l'écart-type plutôt que l'entropie.

Contents

Chapter 1
Introduction **1**

1.1 Motivation . . . 1
1.1.1 Audio for robots, robots for audio . . . 1
1.1.2 Audio source localization is essential . . . 2
1.1.3 Conquering uncertainty . . . 2
1.2 Problem . . . 3
1.2.1 Problem formulation . . . 3
1.2.2 General framework of source localization for robot audition . . . 4
1.3 Contributions . . . 4
1.4 Outline . . . 5

Chapter 2
State of the art **7**

2.1 Angle of arrival measurement . . . 7
2.1.1 General concepts . . . 7
2.1.1.1 Source localization cues . . . 7
2.1.1.2 Far-field source . . . 8
2.1.1.3 Source signal model . . . 8
2.1.2 Overview of source localization methods . . . 11
2.1.2.1 Generalized cross-correlation with phase transform . . . 12
2.1.2.2 Time difference of arrival based methods . . . 12
2.1.2.3 Steered response power based methods . . . 12
2.1.2.4 Multiple signal classification based methods . . . 13
2.1.3 MUSIC-GSVD algorithm . . . 15
2.2 Source activity detection . . . 16
2.3 Sequential filtering for a single source . . . 16
2.3.1 State vector . . . 17

2.3.2 Observation vector . . . 17
2.3.3 Recursive Bayesian estimation . . . 18
2.3.4 Nonlinear mixture Kalman filtering . . . 19
2.3.5 Particle filtering . . . 21
2.3.6 Occupancy grids . . . 22
2.4 Sequential filtering for multiple sources . . . 25
2.4.1 State vector . . . 26
2.4.2 Observation vector . . . 26
2.4.3 Joint probabilistic data association filter . . . 27
2.4.3.1 Prediction step . . . 27
2.4.3.2 Update step . . . 27
2.5 Motion planning for robot audition . . . 27
2.5.1 General robot motion planning . . . 27
2.5.2 Motion planning for robot audition . . . 30

Chapter 3
Source localization in a reverberant environment **33**

3.1 Proposed Bayesian filtering framework . . . 33
3.1.1 State vector . . . 33
3.1.2 Dynamical model . . . 34
3.1.2.1 Dynamical model of the robot . . . 34
3.1.2.2 Dynamical model of the sound source . . . 34
3.1.2.3 Full dynamical model . . . 34
3.1.3 Observation vector . . . 35
3.1.4 Recursive Bayesian estimation . . . 35
3.2 Extended mixture Kalman filtering . . . 36
3.2.1 Prediction step . . . 36
3.2.2 Update step . . . 37
3.2.3 Hypothesis pruning . . . 38
3.2.4 Experimental evaluation . . . 38
3.2.4.1 Data . . . 39
3.2.4.2 Algorithm settings . . . 40
3.2.4.3 Example run - Visualization . . . 40
3.2.4.4 Example run – Estimated trajectories . . . 40
3.2.4.5 Error rate of source location estimation . . . 42
3.2.4.6 Error rate of source activity estimation . . . 43
3.2.4.7 Statistical analysis . . . 46

3.3 Particle filtering . . . 47
3.3.1 Prediction step . . . 47
3.3.2 Update step . . . 47
3.3.3 Particle resampling step . . . 48
3.3.4 Example run . . . 48
3.4 Comparison of the extended MKF with the particle filtering . . . 49
3.4.1 Data . . . 49
3.4.2 Algorithm settings . . . 50
3.4.3 Experimental results . . . 50
3.5 Summary . . . 52

Chapter 4
Multiple source localization **55**

4.1 Learning the sensor model for multiple source localization . . . 55
4.2 Proposed extended MKF with joint probabilistic data association filter . . . 57
4.2.1 State and observation vectors . . . 57
4.2.1.1 State vector . . . 57
4.2.1.2 Observation vector . . . 57
4.2.1.3 Joint associations . . . 58
4.2.2 Prediction step . . . 58
4.2.3 Update step . . . 59
4.2.3.1 Joint association events . . . 60
4.2.3.2 Update step . . . 61
4.3 Experimental evaluation . . . 63
4.3.1 Data . . . 63
4.3.2 Algorithm settings . . . 64
4.3.3 Example run . . . 64
4.3.4 Statistical result . . . 67
4.4 Summary . . . 67

Chapter 5
Optimal motion control for robot audition **69**

5.1 Cost function . . . 69
5.1.1 Shannon entropy criterion . . . 69
5.1.2 Standard deviation criterion . . . 71
5.2 Monte Carlo tree search . . . 71
5.2.1 Algorithm outline . . . 71

5.2.2 Optimism in the face of uncertainty . . . 72
5.3 Adapting MCTS for robot audition . . . 73
5.3.1 Formulation . . . 73
5.3.2 Selection . . . 74
5.3.2.1 Bounded entropy . . . 74
5.3.2.2 Bounded standard deviation . . . 75
5.3.3 Expansion . . . 75
5.3.4 Simulation . . . 75
5.3.5 Backpropagation . . . 76
5.4 Evaluation . . . 76
5.4.1 Experimental protocol . . . 76
5.4.2 Example trajectory . . . 77
5.4.3 MCTS vs other motion planning approaches . . . 79
5.4.3.1 Entropy criterion . . . 81
5.4.3.2 Standard deviation criterion . . . 84
5.4.4 Relation of both criteria with estimation error . . . 84
5.4.5 Effect of the discount factor . . . 87
5.4.5.1 Entropy criterion . . . 87
5.4.5.2 Standard deviation criterion . . . 90
5.5 Summary . . . 90

Chapter 6
Conclusion and perspectives **93**

6.1 Conclusion . . . 93
6.2 Perspectives . . . 94

Appendix A
Résumé en français **97**

A.1 Introduction . . . 97
A.2 État de l'art . . . 98
A.3 Localisation d'une source en environnement réverbérant . . . 99
A.4 Localisation de plusieurs sources . . . 103
A.5 Planification de mouvement pour l'audition . . . 104
A.6 Conclusion et perspectives . . . 106

Bibliography **107**

List of Figures

1.1 Robot audition in a typical audio scene. . . . 3
1.2 General framework of source localization for robot audition. . . . 4

2.1 The acoustic shadow is significant only for high frequencies. . . . 9
2.2 The ITD cue due to the TDOA between the two ears. . . . 10
2.3 AoA computation in the near-field. . . . 10
2.4 AoA computation in the far-field. . . . 11
2.5 Illustration of delay-and-sum beamforming based SRP. . . . 13
2.6 MUSIC method for estimating the AoA. . . . 14
2.7 MUSIC-SEVD and MUSIC-GSVD spectra for the localization of $N = 4$ speakers in the presence of noise. . . . 16
2.8 Distribution of the measured AoA when the actual source is at different angles 156° (top) and 30° (bottom) and different distances from the microphone array. . . . 18
2.9 Visualization of an extended MKF in an example scenario of mapping a source location using a mobile robot. . . . 20
2.10 Illustration of the particle filter with importance sampling and resampling. . . . 21
2.11 Visualization of a particle filter in an example scenario of mapping a source location using a mobile robot. . . . 23
2.12 Visualization of an occupancy grid map in an example scenario of mapping a source location using a mobile robot. . . . 24
2.13 General framework of graph-based robot motion planning. . . . 29

3.1 Distribution of the measured AoA when the actual source is at 96° and 0.5 m from the microphone array. . . . 35
3.2 Turtlebot equipped with a Kinect sensor. . . . 39
3.3 Linear array of 4 microphones inside the Kinect sensor. . . . 39
3.4 Visualization of our extended MKF in an example run. . . . 41
3.5 Estimated trajectories of the two extended MKF methods in the example run. . . . 42
3.6 Top: Estimation error over time of our extended MKF with activity model vs the extended MKF without activity model. Bottom: Ground truth source activity over time. . . . 43
3.7 Median and 95% confidence interval of the estimation error on the source location as a function of the error rate in the SAD observations. . . . 44
3.8 Median and 95% confidence interval of the estimation error on the source location as a function of the error rate in the extended MKF model. . . . 44

3.9 The average probability of incorrectly estimating the source activity over all experiments and 95% confidence interval as a function of the error rate in the SAD observation. . . . 45
3.10 The average probability of incorrectly estimating the source activity over all experiments and 95% confidence interval as a function of the error rate in the extended MKF. . . . 45
3.11 Estimation error distribution of our extended MKF method and of the extended MKF without activity model. . . . 46
3.12 An example run of the particle filtering algorithm to estimate the source location and its activity. . . . 49
3.13 Estimation error of the particle filtering algorithm and the ground truth source activity over time in the example run. . . . 50
3.14 Average estimation error and 95% confidence interval of the particle filtering algorithm with different number of particles N. . . . 51
3.15 Average estimation error and 95% confidence interval of the extended MKF algorithm with different maximum number of components $N_{\max}$. . . . 52

4.1 Visualization of our extended MKF with the JPDAF in an example of localizing two sources. . . . 65
4.2 Top: Estimation error of the two source locations over time. Bottom: Ground truth source activities over time. . . . 66
4.3 Average estimation error and 95% confidence interval over time of two source locations over all 200 experiments. . . . 67

5.1 The four main steps in an iteration of the MCTS algorithm. . . . 72
5.2 An iteration of the MCTS algorithm. . . . 73
5.3 Initial position of the robot and the source and estimated belief before running motion planning strategies. . . . 77
5.4 Top: Example robot motion sequence obtained from the MCTS algorithm with entropy criterion. Bottom: Example robot motion sequence obtained from the greedy algorithm with entropy criterion. . . . 78
5.5 Estimation of the source location when the robot follows the trajectory from the MCTS algorithm. . . . 80
5.6 Estimation of the source location when the robot follows the trajectory from the greedy algorithm. . . . 81
5.7 Entropy over time with the MCTS algorithm and the greedy algorithm. . . . 82
5.8 Estimation error over time with the MCTS algorithm and the greedy algorithm. 82
5.9 Average entropy and 95% confidence interval over time of the 4 algorithms with the entropy criterion over all 200 experiments. . . . 83
5.10 Average estimation error and 95% confidence interval over time of the 4 algorithms with the entropy criterion over all 200 experiments. . . . 83
5.11 Average standard deviation and 95% confidence interval over time with the standard deviation criterion of the 4 algorithms over all 200 experiments. . . . 85
5.12 Average estimation error and 95% confidence interval over time of the 4 algorithms with the standard deviation criterion over all 200 experiments. . . . 85
5.13 Correlation between the entropy and the estimation error. . . . 86
5.14 Correlation between the standard deviation and the estimation error. . . . 86

5.15 Average entropy and 95% confidence interval over time with different discount factor values for the entropy criterion. 87
5.16 Average estimation error and 95% confidence interval over time with different discount factor values for the entropy criterion. 88
5.17 Error bars with the entropy criterion at $t = 1$ s and $t = 10$ s with different discount factor values. 88
5.18 Average standard deviation and 95% confidence interval over time with different discount factor values for the standard deviation criterion. 89
5.19 Average estimation error and 95% confidence interval over time with different discount factor values for the standard deviation criterion. 89
5.20 Error bars with the standard deviation criterion at $t = 1$ s and $t = 10$ s with different discount factor values. 90

A.1 Schéma général de localisation de sources sonores. 98
A.2 Distribution de l'AoA mesuré quand la source réelle est à 96° et 0,5 m de l'antenne. 100
A.3 Visualisation de notre MKF étendu sur un exemple. 102
A.4 Une itération de l'algorithme MCTS. 104
A.5 Erreur d'estimation moyenne et intervalle de confiance à 95% en fonction du temps pour les quatre algorithmes avec l'écart-type comme critère calculés à l'aide de 200 expériences. 105

Chapter 1

Introduction

1.1 Motivation

1.1.1 Audio for robots, robots for audio

Robots that can understand and naturally interact with humans are a long-lasting vision of the scientists. As the development of artificial intelligence and robotics keeps on growing, robots will be more friendly and more helpful to human beings in everyday life. To achieve this goal, the crucial requirement for robots is to be fully aware of the surrounding environment. Imagine that you have an assistive robot at home. You are in a living room watching a movie and asking your robot to make coffee and then bring it to your place. At that time, the robot is in the kitchen that is hidden from the living room. Your robot cannot see you but it can listen to your commands through its embedded microphones. Before executing the commands, the robot has to recognize and understand your speech. When robot has understood what you just said, it issues a confirmation by talking back to you and then starts making coffee. To efficiently manipulate object for making coffee, the robot must perceive the environment in the kitchen based on a camera. After your coffee is ready, the robot brings it to your place which was localized from your speech signal and also from the visual signal but only when you are in the line of sight of the robot camera. This is one example task for a home service robot. Such service robots can give most benefit to people with disabilities, as well as elder people, by helping them live independently. However, building a robot like this is a huge challenge which requires achievements in many fields. The robot needs not only humanlike functional mechanical parts but also a cognitive system that is fully aware of the environment. Besides a visual sensing system, the robot should also have an auditory sensing system.

For an autonomous assistive robot, understanding speech is a necessary and important function. Speech is considered as the most natural and convenient way of communicating and giving a command to the robot. Having a hearing ability, the robot can perceive the contents of human voices and gather information about the sound environment such as location and the activity of the sound sources. This hearing capability is called robot audition [Nakadai et al., 2000, Nakadai et al., 2010, Okuno and Nakadai, 2015]. The auditory knowledge from robot audition completes the information delivered by other sensors like cameras or laser range-finders. When the robot gathers more knowledge about the environment, it will better know what to do next. Human-robot interaction will be more efficient and the robot will be more friendly to humans.

Although still recent compared to static microphones, robot audition has brought new advancements to audio signal processing. By exploiting robot motion, i.e. active audition [Nakadai

et al., 2000], more information can be obtained such as the range of the sound source. Before that, source localization in the far field could only estimate the source angle of arrival (AoA) and even suffered from the front-back ambiguity in the case of a linear microphone array [Nakadai et al., 2000, Kim et al., 2008, Nakamura et al., 2012]. With active audition, the front-back ambiguity can be eliminated by simply turning towards the target sound source [Nakadai et al., 2003, Berglund and Sitte, 2005]. Using body movements of the robot, the distance to the sound source [Portello et al., 2014, Vincent et al., 2015] can also be estimated. Finally, robot audition provides the means of getting closer to the target source for increasing the signal-to-noise ratio [Martinson and Schultz, 2009, Song et al., 2011] and relocating the robot in order to avoid noisy areas. Overall, this improves the quality of signal processing and the communication.

In order to achieve the full hearing function for robot audition, the robot needs to solve many problems in signal processing such as sound source localization, sound source separation, automatic speech recognition, sound classification and identification. With robot audition, the latter problems can be better solved by getting closer to the sources and then doing signal processing methods as for static microphones. In this dissertation, we will focus on applying the problem of sound source localization for robot audition.

1.1.2 Audio source localization is essential

Equipped with microphones, robots are capable of detecting the sound sources and localizing their origin. Using this source location information, they can separate a mixture of sounds, increasing the performance of speech recognition. Based on that, they can process the data to extract other useful information, e.g., source identity, emotion, and better understand their environment. In addition, by knowing the sound source location, they can move closer to the target source or move away from the noisy space to increase signal-to-noise ratio and enhance the audio signal. As a result, this improves the interaction between robots and humans.

Source localization benefits not only human-robot interaction in robotics but it is also important in many diverse applications such as tracking of marine animals, military surveillance and reconnaissance, search and rescue in earthquake areas, seismic remote sensing, and wildlife monitoring. Recently, source localization has been utilized in many media applications such as target tracking, smart video conference, and audiovisual sensor fusion.

In nature, animals with better sound localization ability have a clear evolutionary advantage. They can avoid dangerous predators when hearing them from the distance. Vice versa, hearing helps predators detect their preys without seeing them.

1.1.3 Conquering uncertainty

Detecting and precisely localizing a sound source is not an easy task due to a noisy measurements. Fig. 1.1 shows a typical sound environment around the robot. Besides the direct signal from the target sources, the robot also receives reflected signals and the noises, e.g., fan noise. Such an environment is called reverberant and noisy environment. Reverberation and background noise severely degrade the performance of source localization [Vermaak and Blake, 2001, Blandin et al., 2012] and also source activity detection (SAD) [Ramírez et al., 2007, Zhang and Wu, 2013].

In addition, the sound sources in the real world, e.g. speech, are not always active, making the estimation even more difficult. The silence intervals and the transitions between activity and silence intervals increase the uncertainty in detecting and localizing the source and induce false measurements. Most studies try to avoid this by assuming a priori knowledge of the source activity (i.e., whether the speaker is active or inactive in a given time frame) or by evaluating

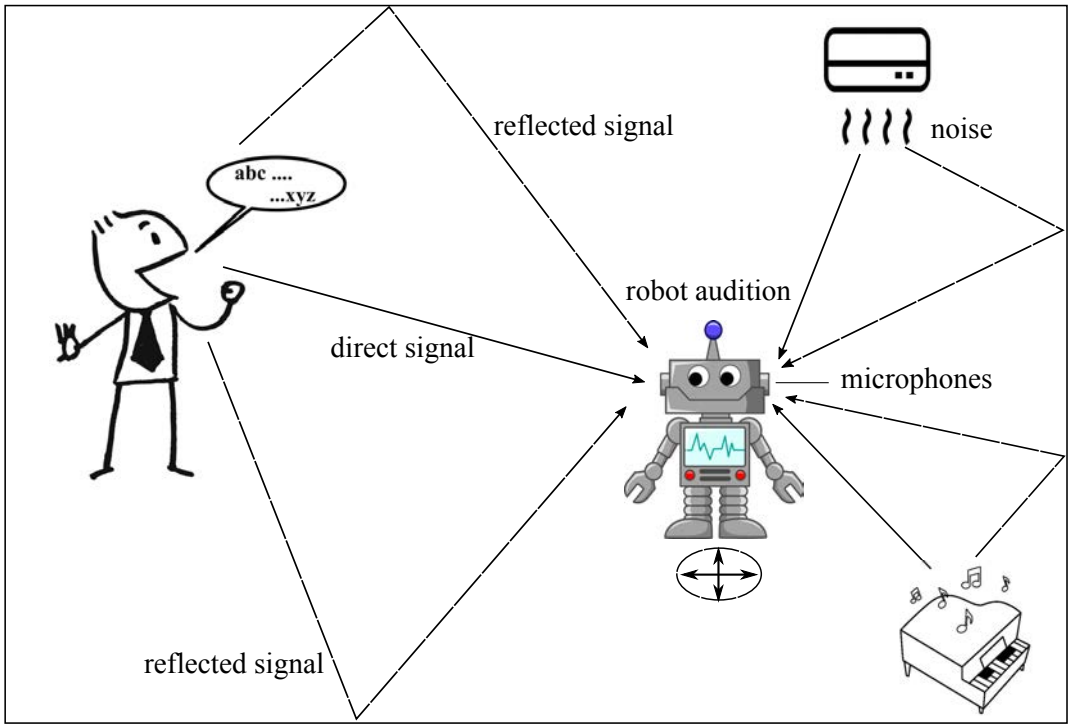

Figure 1.1: Robot audition in a typical audio scene.

in an anechoic environment [Portello et al., 2014]. Obviously, these solutions do not perform well in a real-world situation where false measurements will happen. Moreover, when the source is not static or in the context of multiple sound sources, the uncertainty significantly increases. Therefore, conquering uncertainty is the key challenge for inferring precise information about the source location.

The main target of this thesis is to address the uncertainty in reverberant environments, in the context of intermittent, moving and multiple sources. In addition, we exploit robot motion to quickly minimize the uncertainty in source localization.

1.2 Problem

1.2.1 Problem formulation

We consider three problems.

The first problem consists of locating the position of a single sound source using a mobile robot equipped with a microphone array. The target sound source is intermittent and possibly moving. The noise level and the reverberation time of the room are assumed to be stationary over time. The robot can be equipped with any kind of microphone array. The microphone positions are known and fixed with respect to the robot. The robot can easily move on its wheels in a flat room area.

At each time t, the robot receives an audio signal. From that signal, the robot detects the activity of the target source. For a far-field source, i.e., when the distance from the microphone array to the source is larger than the array size, the robot can only measure the AoA but not the distance to the source. For a linear microphone array, there also exists a front-back ambiguity on the source AoA. The source AoA and SAD measurements contain uncertainty. The problem

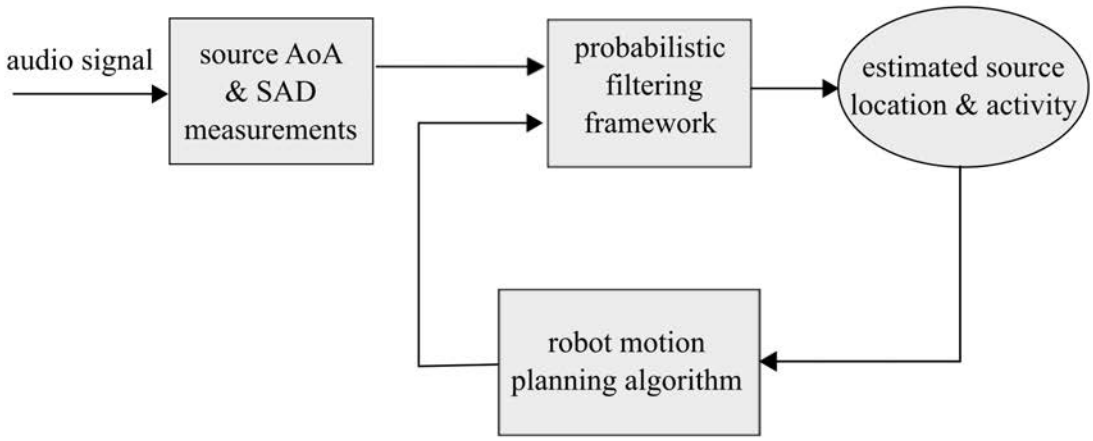

Figure 1.2: General framework of source localization for robot audition.

is to estimate the belief distribution on the source location after each measurement by fusing the robot motion information with the measurements.

The second problem is to extend the above work to the context of multiple intermittent, possibly moving sources.

The third problem is to reduce the uncertainty on the source location by controlling the robot movements. The theory is applicable with ego noise due to the robot moving, but in practice we only consider the situation when the robot is not moving while taking the source AoA and SAD measurements.

In this thesis, we focus on estimating the source locations in a noisy and reverberant environment using the robot audition and improving the estimation result by finding the optimal robot motion.

1.2.2 General framework of source localization for robot audition

The general framework of source localization for robot audition is depicted in Fig. 1.2. The audio signal is captured by the microphone array on the robot. This audio signal could be any source that is active for a sufficient amount of time, e.g., a human speaker. The source AoA and SAD measurements are uncertain due to reverberation and the noisy background. By fusing these measurements with the robot motion information using a probabilistic filtering framework, we can deal with uncertainties in the observations and estimate the source location and activity over time.

The estimated source location can be improved by implementing a robot motion planning algorithm. The optimal actions are selected and executed to minimize the estimation uncertainty.

1.3 Contributions

We made the following key contributions regarding the three problems stated above.

Our first contribution lies in the development of an extended mixture Kalman filter (MKF) framework that can apply to any microphone array geometry for jointly estimating the location and activity of an intermittent and moving source in a reverberant environment [Nguyen et al., 2016]. The experimental results show that our filtering framework has better performance

compared to a method that does not track the source activity. The source location and activity estimation in our framework is robust to false measurements in both the SAD and the AoA.

The second contribution concern the extension of the extended MKF algorithm to multiple source localization. We implement the extended MKF with the joint probabilistic data association filter (JPDAF) for dealing with the association problem and jointly estimating the location of the two sources and their activities over time. We build a sensor model for the case of two sound sources by simulating two sources with background noise and reverberation and learn the probability of correctly observing each AoA measurement. The experimental evaluation shows the ability of the proposed framework to handle uncertainty in the observations when localizing and tracking two intermittent, moving sources in a noisy, reverberant environment.

The third main contribution concerns the adaptation of the Monte Carlo tree search (MCTS) method for long-term robot motion planning to improve source localization [Nguyen et al., 2017]. In this contribution, we have defined an exact cost function for long-term robot motion planning. We introduce two criteria as cost functions: the Shannon entropy and the standard deviation. We show and compare the correlation between the estimation error and these two criteria. In the cost function, the effect of the discount factor, i.e. the factor for tuning the tradeoff between short vs long term, on the performance of planning algorithm is also investigated. In addition, we compare our method with other motion planning methods.

The first and third contributions have been published in the following conferences, respectively:

- Nguyen, Q. V., Colas, F., Vincent, E., & Charpillet, F.. Localizing an intermittent and moving sound source using a mobile robot. in Proc. IROS 2016 [Nguyen et al., 2016].

- Nguyen, Q. V., Colas, F., Vincent, E., & Charpillet, F.. Long-Term Robot Motion Planning for Active Sound Source Localization with Monte Carlo Tree Search. in Proc. HSCMA 2017 [Nguyen et al., 2017].

1.4 Outline

This thesis consists of six chapters. After this introductory chapter, the rest of the thesis is organized as follows.

Chapter 2 provides a literature review of prior work related to source localization and motion planning for robot audition. We present the existing work on source localization for static microphone arrays. Later on, we write about source localization for robot audition. Then, we summarize the existing work on multiple source localization. The last section surveys the literature related to motion planning algorithms for robots in general and for robot audition. We describe the technical differences between our methods and conventional ones.

Chapter 3 presents our method for localizing a sound source in a reverberant environment. We describe the proposed extended MKF framework for localizing an intermittent, moving source. In our method, we explicitly estimate both the source activity and the source location. Experiments show that our filter is robust to false measurements of the AoA and the SAD. The performance of this method is compared with a method that does not consider the source activity in the model. In the rest of this chapter, we introduce the particle filtering method for jointly estimating the source location and its activity and comparing with the extended MKF in term of localization performance, and computational time.

Chapter 4 extends the work on source localization to the context of multiple sound sources. We present a sequential filtering framework for localizing multiple sources. In the context of

multiple sound sources, the uncertainty is not only increased when the sources are inactive but also when they are active. The peaks of the multiple signal classification (MUSIC) spectrum may be occluded and false alarms may occur. By implementing an extended MKF with the JPDAF, we can jointly estimate the two source locations and their activities over time. The experimental evaluation shows the effectiveness of the proposed method in localizing and tracking two target sources in reverberant environment.

Chapter 5 focuses on optimal motion control to improve the source localization result. We define the cost function for long-term robot motion planning with different criteria: the Shannon entropy and the standard deviation. In order to investigate the tradeoff between short vs long term, we also integrate a discount factor in the cost function. Later on, we introduce an approach that adapts the MCTS method for efficiently finding the optimal robot trajectory. This method can balance between exploration and exploitation in the motion planning. The experimental results show better performance compared to other robot motion planning methods in the long term.

Chapter 6 summarizes this dissertation, discusses our achievements as well as limitations, and presents perspectives of future research directions.

Chapter 2

State of the art

For addressing the problem of building a map of sound source locations over time, we first need AoA measurements at each time step. Then, a sequential filtering algorithm combines these measurements with the knowledge about robot motion to estimate the position of the target sources. A motion planning technique can be implemented to find a robot trajectory that improves the estimation result.

Following the above workflow, the first section reviews state-of-the-art techniques for AoA measurement. Sequential filtering algorithms for a single source and for multiple sources are presented in the subsequent sections. In the final section, motion planning methods for mobile robots and specifically for robot audition are described.

2.1 Angle of arrival measurement

Source localization for static microphones has been investigated since the beginning of research on signal processing. Using static microphones, classical source localization methods [DiBiase et al., 2001] often estimate only the source AoA. That is the case for a far field source, when the distance from the microphone array to the source is larger than the array size. Generally speaking, these techniques aim to derive an accurate localization method that is robust to various acoustic environments, e.g., noisy or reverberant environments, and operates in real time. When the target sources and the microphones are both static or move slowly compared to the speed of sound, the Doppler effect can be neglected. With that assumption, we can also apply these source localization methods to moving sources or moving microphones. This section will present the general principles of source localization and summarize state-of-the-art source localization techniques.

2.1.1 General concepts

In this section, we present the general concepts about source localization cues, the notion of far-field source and the source signal model.

2.1.1.1 Source localization cues

From the duplex theory proposed by Lord Rayleigh [Rayleigh, 1907], the two main cues for humans to determine the source AoA are the interaural intensity difference (IID) and the interaural time difference (ITD).

The IID refers to the fact that the intensity of sound emitted from a single source will differ at the two ears. The sound received at the far ear is shadowed by the human head as in Fig. 2.1(a), resulting in a difference of level (or intensity) of sound between the two ears. However, the intensity difference is large only for sounds whose frequency is above 1500 Hz, as in Fig. 2.1.

The second cue for humans to estimate the source AoA is related to the time difference of arrival (TDOA) (or ITD) between the two ears. This can be explained by the example illustrated in Fig. 2.2. The original sound is placed to the right of the listener in a horizontal plane. The left ear is farther from the source compared to the right ear. As a result, the sound wave emitted by the source takes more time to reach the left ear. The TDOA between the two ears is a fraction of a millisecond. The time difference is dominant in human perception for sounds whose frequency is below 1500 Hz, that is when the wavelength of the sound is similar to, or larger than, the distance between the ears.

In signal processing, artificial audition systems are often equipped with two or more microphones. The microphones are commonly embedded in a simple structure with small dimensions, e.g., a Kinect sensor. Therefore, the shadow effect is negligible. This is why most source localization techniques in the literature are based entirely on the ITD.

2.1.1.2 Far-field source

Source localization differs in far-field and near-field conditions. Fig. 2.3 illustrates a near-field condition where the distance from the source to the array is comparable to the array size. The curvature of the wavefronts is significant compared to the array size. In this condition, the distance to the target source can be estimated from the signals received at the microphones, if there are three or more microphones.

A source is considered to be in the far-field when the distance to the array center is significantly larger than the array size. In that case, the received signals at each microphone can be approximated as a plane wave and they have the same AoA. This is depicted in Fig. 2.4. Therefore, we can only estimate the AoA but not the distance to the source. In this thesis, we assume that the target sources are in the far-field. In that condition, the AoA can be computed as

$$\alpha = \arcsin\left(\frac{\Delta r}{d}\right) = \arcsin\left(\frac{\Delta t \cdot c}{d}\right), \tag{2.1}$$

where c is the speed of sound which equals 343 m/s, Δr is the difference of distances between the source and two microphones, Δt is the corresponding TDOA, and d is the distance between the two microphones. From the above equation, we can obtain the value of the AoA α by computing the time delay Δt.

2.1.1.3 Source signal model

In the following, we introduce the source signal model that will be used for all the source localization techniques presented below.

Let us assume that there are N active sources and we have an array of M microphones. The signal received at each microphone can be modeled as

$$x_i(t) = \sum_{j=1}^{N} a_{i,j} * s_j(t) + n_i(t), \tag{2.2}$$

where $i = 1, \ldots, M$ and $j = 1, \ldots, N$ are respectively the indices of microphones and sound sources, $*$ is the convolution operator, $x_i(t)$ is the received signal at the ith microphone, $s_j(t)$

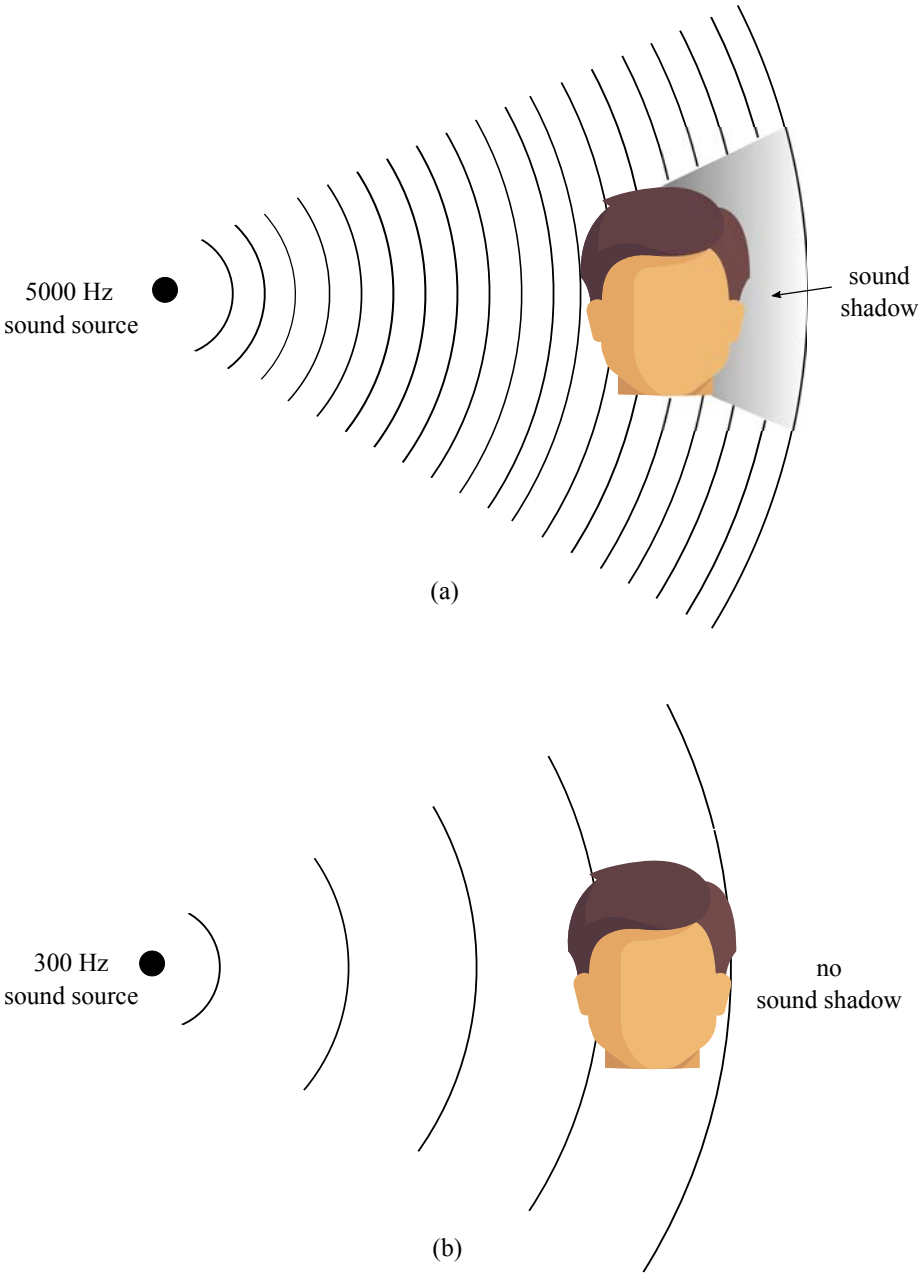

Figure 2.1: The acoustic shadow is significant only for high frequencies.

is the jth emitted signal, $n_i(t)$ is the background noise which includes reverberation, and $a_{i,j}$ is the acoustic impulse response which models the direct path from source j to microphone i.

The frequency contents of the source signals are changing over time. Therefore, to efficiently analyze the signals we use the short-time Fourier transform (STFT) to transform the received signals into the time-frequency domain.

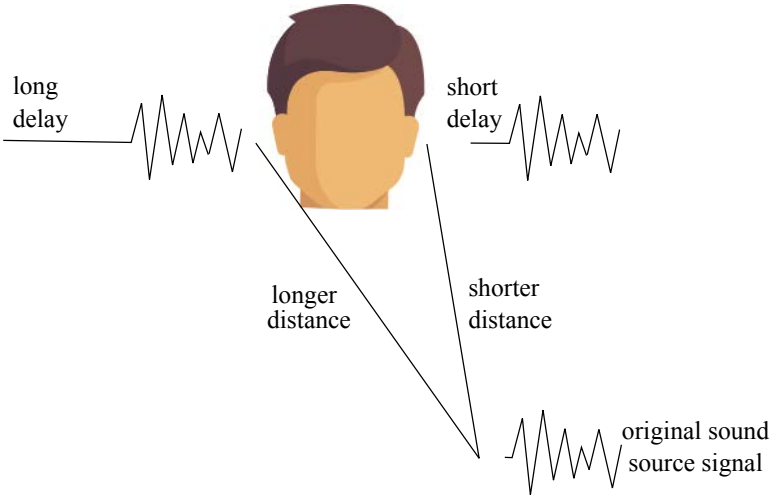

Figure 2.2: The ITD cue due to the TDOA between the two ears.

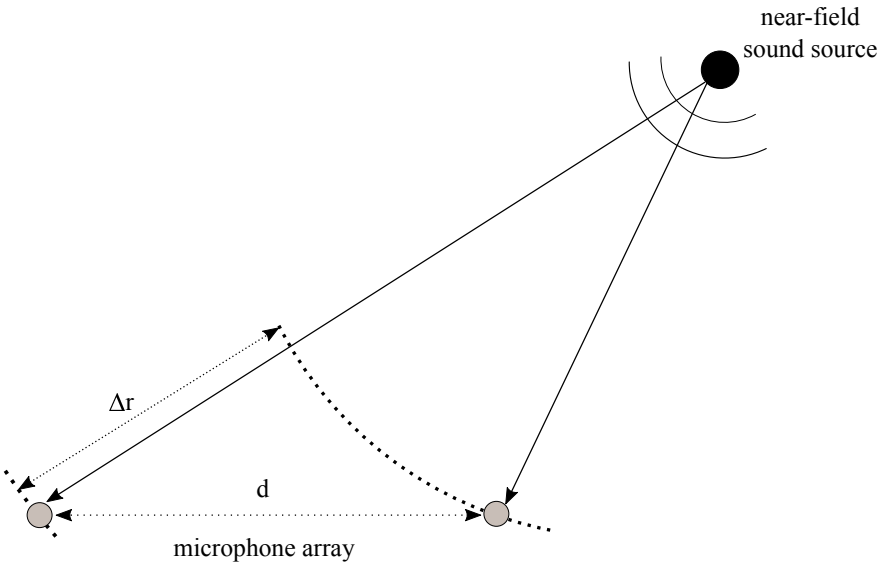

Figure 2.3: AoA computation in the near-field.

For the discrete-time STFT, we have

$$X(k,f) = \sum_{t=0}^{T-1} x\left(\frac{kT}{2}+t\right) w(t) e^{\frac{-2i\pi ft}{T}}, \tag{2.3}$$

where $w(t)$ is a window function, T is the window length, $k = 0, ..., K-1$ and $f = 0, ..., F-1$ with $F = T$ are, respectively the time frame and frequency bin indices. We assume half-overlapping windows.

We can model the received signals in the time-frequency domain as

$$\mathbf{x}(k,f) = \sum_{j=1}^{N} \mathbf{a}(\alpha_j, f) S_j(k,f) + \mathbf{n}(k,f), \tag{2.4}$$

where $\mathbf{x}(k,f) = [X_1(k,f), \ldots, X_M(k,f)]^T$, $S_j(k,f)$, $\mathbf{n}(k,f) = [N_1(k,f), \ldots, N_M(k,f)]^T$ are the

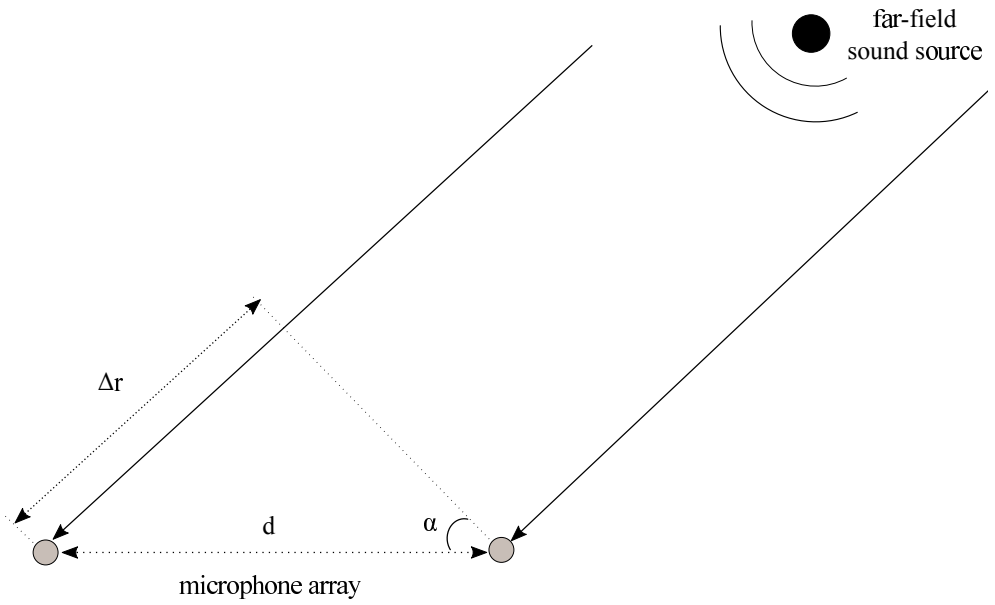

Figure 2.4: AoA computation in the far-field.

STFTs of the observed signals at microphones, the source signals and the noise, respectively. $\mathbf{a}(\alpha_j, f) = [A_1(\alpha_j, f), ..., A_M(\alpha_j, f)]^T$ where $A_i(\alpha_j, f)$ is the Fourier transform of $a_{i,j}(t)$: it depends only on frequency f and on the source AoA α_j and it is called steering vector. This steering vector $\mathbf{a}(\alpha_j, f)$ can be expressed up to a multiplicative factor as

$$\mathbf{a}(\alpha_j, f) = \begin{bmatrix} 1 \\ e^{\frac{-2i\pi f \Delta t_1(\alpha_j)}{F}} \\ \vdots \\ e^{\frac{-2i\pi f \Delta t_{M-1}(\alpha_j)}{F}} \end{bmatrix}, \tag{2.5}$$

where $\Delta t_{i-1}(\alpha_j)$ denotes the TDOA in samples between microphone 1 and microphone i for source j.

2.1.2 Overview of source localization methods

Existing source localization techniques can be classified into three main classes [DiBiase et al., 2001]. The first class includes the techniques that exploit TDOA information [Knapp and Carter, 1976]. In the second class are the techniques which are based upon maximizing the steered response power (SRP) of a beamformer [Hahn and Tretter, 1973, Van Veen and Buckley, 1988, Johnson and Dudgeon, 1992]. The localization methods that are adapted from the field of high resolution spectral analysis are in the third class [Schmidt, 1986, Wang and Kaveh, 1985]. Experimental comparisons of these algorithms are detailed in the literature [DiBiase et al., 2001, Badali et al., 2009, Blandin et al., 2012].

In this section, we provide an overview of prominent source localization methods in each of the three above classes.

2.1.2.1 Generalized cross-correlation with phase transform

The generalized cross-correlation (GCC) [Knapp and Carter, 1976] method is the most popular method for estimating the TDOA information for a microphone pair. The type of filtering, or weighting, used with GCC is crucial to the performance of TDOA estimation. Maximum likelihood weighting is theoretically optimal for single-path propagation in the presence of uncorrelated noise, however its performance degrades significantly with increasing reverberation [Champagne et al., 1996]. The phase transform (PHAT) weighting is more robust against reverberation, even though it is suboptimal under reverberation-free conditions. The generalized cross-correlation with phase transform (GCC-PHAT) has been shown to perform well in realistic environments [Omologo and Svaizer, 1996, Svaizer et al., 1997, Brandstein and Silverman, 1997].

Given two signals $x_i(k)$ and $x_{i'}(k)$, the GCC-PHAT is defined as

$$P_{ii'}(\Delta t, k) = \sum_{f=f_{\min}}^{f_{\max}} \frac{\Re(X_i(k,f)\bar{X}_{i'}(k,f)e^{\frac{2i\pi f \Delta t}{F}})}{|X_i(k,f)\bar{X}_{i'}(k,f)|}, \tag{2.6}$$

where $X_i(k,f)$ and $X_{i'}(k,f)$ are the STFTs of the two signals, . The TDOA for a single source can be computed as:

$$\Delta t_{\text{PHAT}ii'}(k) = \arg\max_{\Delta t} P_{ii'}(\Delta t, k). \tag{2.7}$$

2.1.2.2 Time difference of arrival based methods

In the case of three or more microphones, after the TDOA between each pair of microphones has been estimated, the geometric relationship between the sound source and the microphone array can be utilized to estimate the source AoA. By applying a triangulation method to different microphone pairs, the source location can be estimated [Brutti and Nesta, 2013]. The accuracy of AoA measurement of the TDOA based methods depends on the accuracy of the TDOA estimation. The geometry of microphone array can also affect the performance of TDOA based methods. Such methods are well suited to AoA measurement over a limited spatial range when there are sufficient microphone data available.

2.1.2.3 Steered response power based methods

The general idea of SRP based methods is to steer a spatial filter (also known as a beamformer) to a predefined spatial region or direction [Johnson and Dudgeon, 1992] by adjusting its steering parameters and then search for maximal output. The output of the beamformer is known as the steered response. The maximal SRP is obtained when the direction of the beamformer matches the location of the target source.

One of the simplest and conventional approaches for SRP is the delay-and-sum beamformer [Johnson and Dudgeon, 1992, Flanagan et al., 1985]. Its principle is illustrated in Fig. 2.5. Delay-and-sum beamformers apply time shifts to the signals received at the microphones to compensate for the TDOA of the target source signal at each microphone. The delay-and-sum beamformer output is defined as

$$y(k,f) = \mathbf{a}^H(\alpha, f)\mathbf{x}(k,f), \tag{2.8}$$

where $\mathbf{a}$ is the steering vector whose value depends on the hypothesized AoA α, H is the Hermitian transpose operator. Therefore, signal power is enhanced in the look direction α and attenuated in all other directions.

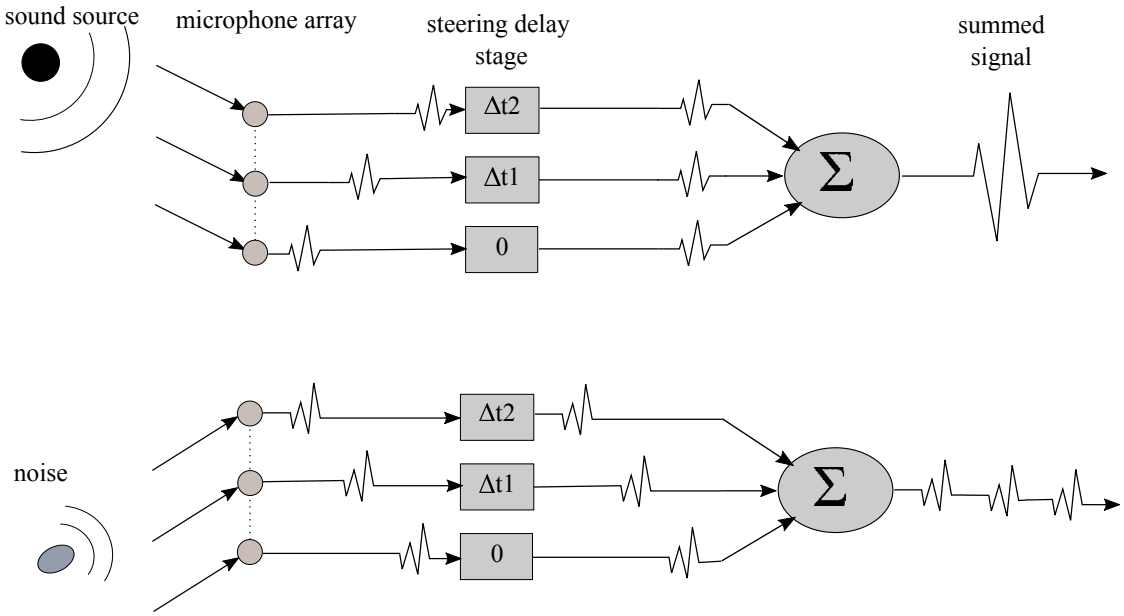

Figure 2.5: Illustration of delay-and-sum beamforming based SRP.

The output power is computed as

$$P(\alpha, k) = \sum_{f=f_{\min}}^{f_{\max}} |y(k, f)|^2. \tag{2.9}$$

The source AoA at time k is then determined by maximizing the output power over all possible look directions α.

The disadvantage of the conventional SRP approach is that the beamformed maps contain interference patterns referred to as side lobes besides a main lobe. In addition, the angular width of the main lobe is proportional to the wavelength of the target signal. Therefore the spatial resolution of the beamformer for lower frequencies is reduced.

More advanced SRP approaches apply filters to the array signals [DiBiase, 2000, Dmochowski et al., 2007]. Among them, the steered response power with phase transform (SRP-PHAT) provides robust source localization in noisy and reverberant environments and more precise estimation than the delay-and-sum beamformer [DiBiase, 2000]. The SRP-PHAT can be expressed as the sum of all possible pairwise GCC-PHAT time shifted for the set of steering delays. The SRP-PHAT of a 2-element array is equal to the GCC-PHAT of those two microphones.

2.1.2.4 Multiple signal classification based methods

The multiple signal classification (MUSIC) algorithm is commonly used for source localization using microphone arrays. It was first introduced by [Schmidt, 1986] and is technically based on the standard eigenvalue decomposition (SEVD). Fig. 2.6 describes the framework of the MUSIC approach.

The covariance matrix between channels of the incoming signal can be defined as follows

$$\mathbf{R}(k, f) = \mathbb{E}(\mathbf{x}(k, f)\mathbf{x}^H(k, f)), \tag{2.10}$$

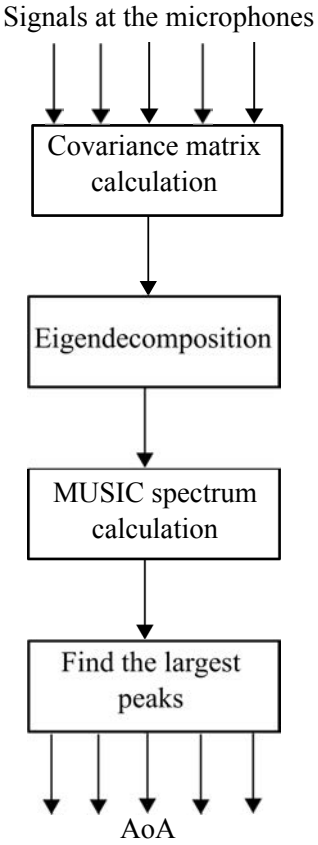

Figure 2.6: MUSIC method for estimating the AoA.

where $\mathbb{E}$ is the expectation. In practice, we compute the covariance matrix by averaging over the time frames for an interval Δk:

$$\mathbf{R}(k,f) = \frac{1}{\Delta k} \sum_{\tau=k-\Delta k+1}^{k} \mathbf{x}(\tau,f)\mathbf{x}^H(\tau,f). \tag{2.11}$$

In the MUSIC-SEVD method, we decompose $\mathbf{R}(k,f)$ into the signal subspace and the noise subspace as follows:

$$\mathbf{R}(k,f) = \mathbf{E}(k,f)\mathbf{\Lambda}(k,f)\mathbf{E}^{-1}(k,f), \tag{2.12}$$

where $\mathbf{E}(k,f) = [\mathbf{e}_1(k,f),\ldots,\mathbf{e}_M(k,f)]$ contains singular eigenvectors which are perpendicular to each other, and $\mathbf{\Lambda}(k,f) = \mathrm{diag}(\lambda_1(k,f),\ldots,\lambda_M(k,f))$ is the diagonal matrix with the eigenvalues as the diagonal elements. In addition, these eigenvalues are arranged in decreasing order.

The value of λ represents the power of received sound. For N received sound sources, where N is strictly smaller than the number of microphones, eigenvalues from λ_1 to λ_N have larger values corresponding to the power of the sound sources and the remaining eigenvalues correspond to the power of noise. We take into account the fact that the noise subspace corresponding to small eigenvalues is perpendicular to the signal subspace corresponding to larger eigenvalues.

The MUSIC spatial spectrum is obtained as

$$P(\alpha,k,f) = \frac{|\mathbf{a}^H(\alpha,f)\mathbf{a}(\alpha,f)|^2}{\sum_{i=N+1}^{M} |\mathbf{a}^H(\alpha,f)\mathbf{e}_i(k,f)|^2}, \tag{2.13}$$

where $\mathbf{a}(\alpha, f)$ is the M-dimensional steering vector corresponding to the AoA α and frequency f. The eigenvectors that correspond to noises ($\mathbf{e}_{N+1}(k, f)$ through $\mathbf{e}_M(k, f)$) are assumed to be perpendicular to the steering vector $\mathbf{a}(\alpha, f)$ with the AoA α. Therefore, their inner product in the denominator of (2.13) will be approximately equal to 0 at the AoA α and $P(\alpha, k, f)$ will be large. This corresponds to a sharp peak in the MUSIC spectrum at the AoA α.

The broadband spatial MUSIC spectrum is computed by accumulating $P(\alpha, k, f)$ in (2.13) over frequency bin indices

$$P(\alpha, k) = \frac{1}{f_{\max} - f_{\min} + 1} \sum_{f=f_{\min}}^{f_{\max}} P(\alpha, k, f), \tag{2.14}$$

where $f_{\min}$ and $f_{\max}$ represent the minimum and maximum frequency bins, respectively.

The MUSIC-SEVD method works properly when the target sources have stronger power than the noise sources. In the situation when there exists a noise with high power, e.g., motor noise of the robot, some of the eigenvectors for the target signals are wrongly picked from noises. As a result, undesired peaks occur in the MUSIC spectrum.

In order to solve this problem, the generalized eigenvalue decomposition (GEVD) method [Asano et al., 2008] was proposed for MUSIC [Nakamura et al., 2009]. MUSIC-GEVD takes the noise covariance matrix into account for suppressing noise sources. Firstly, the noise covariance matrix is determined as follows

$$\mathbf{K}(k, f) = \mathbb{E}(\mathbf{n}(k, f)\mathbf{n}^H(k, f)). \tag{2.15}$$

Then, MUSIC-GEVD is performed by extending equation (2.12) as

$$\mathbf{K}^{-\frac{1}{2}}(k, f)\mathbf{R}(k, f)\mathbf{K}^{-\frac{1}{2}}(k, f) = \mathbf{E}(k, f)\mathbf{\Lambda}(k, f)\mathbf{E}^{-1}(k, f), \tag{2.16}$$

which whitens the noise-related eigenvalues.

The drawback of this algorithm is that it has a high computational cost due to the subspace decomposition. Robot audition needs to achieve both high resolution and real-time processing simultaneously. An extension of MUSIC-GEVD called MUSIC-generalized singular value decomposition (GSVD) [Nakamura et al., 2012] has been proposed to solve that problem.

2.1.3 MUSIC-GSVD algorithm

Compared to previous decomposition strategies for MUSIC, GSVD has efficient computational cost while maintaining noise robustness in source localization [Nakamura et al., 2012]. In the MUSIC-GSVD method, to reduce the computational cost, equation (2.16) is modified as

$$\mathbf{K}^{-1}(k, f)\mathbf{R}(k, f) = \mathbf{E}_l(k, f)\mathbf{\Lambda}(k, f)\mathbf{E}_r^{-1}(k, f), \tag{2.17}$$

where $\mathbf{E}_l(k, f)$ and $\mathbf{E}_r(k, f)$ are left and right singular vectors, respectively. They are unitary and mutually orthogonal.

Fig. 2.7 shows the MUSIC spectrum obtained by MUSIC-SEVD and MUSIC-GSVD for the localization of $N = 4$ speakers speaking simultaneously in the presence of noise. The diffuse noise and the directional noise observed are presented in Fig. 2.7(a). As shown in Fig. 2.7(c), when using MUSIC-GSVD, the noise signal is suppressed correctly and the strong peaks correspond to the direction of the four speakers. However, when using MUSIC-SEVD in Fig. 2.7(b), there is a wrong peak corresponding to the noise. This shows that MUSIC-GSVD is more robust to noise compared to MUSIC-SEVD. Throughout the thesis, the MUSIC-GSVD method will be employed for estimating the source AoA.

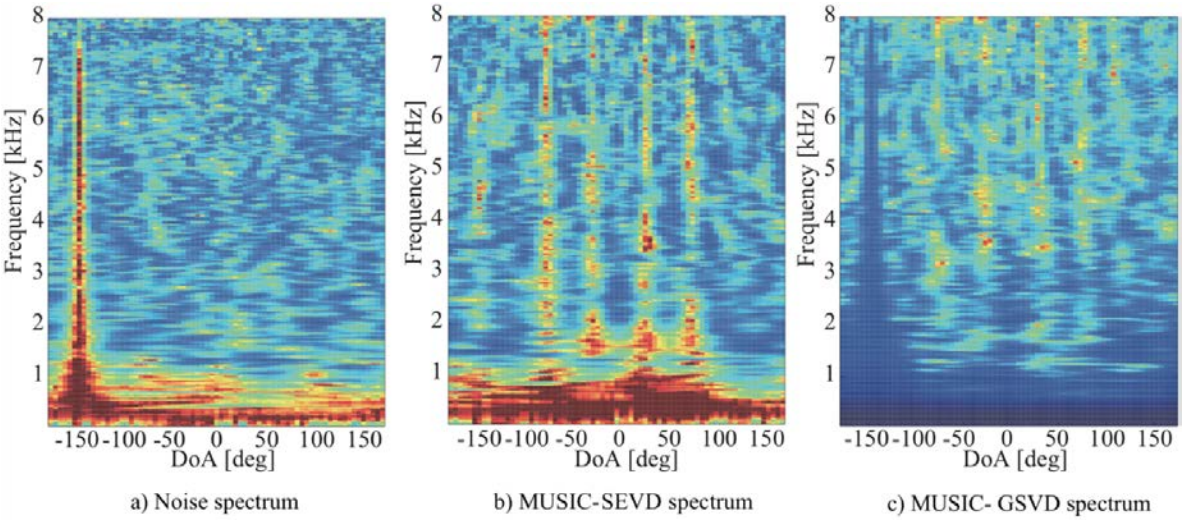

Figure 2.7: MUSIC-SEVD and MUSIC-GSVD spectra for the localization of $N = 4$ speakers in the presence of noise (from http://www.hark.jp/document/2.3.0/hark-document-en/).

2.2 Source activity detection

In the real world, sound sources often contain silent intervals which can degrade the performance of the source localization. Therefore, to achieve a robust source localization, the audio system should acquire information about the activity of the target sources. SAD refers to the problem of identifying the active or silent state of a single or multiple target sources in an audio signal. SAD is often used as an important front-end technique for many speech and audio processing systems [Karray and Martin, 2003, Ramirez et al., 2003].

The different detection principles include those based on signal-to-noise ratio (SNR), energy thresholds [Woo et al., 2000], pitch detection [Chengalvarayan, 1999], spectrum analysis [Marzinzik and Kollmeier, 2002], periodicity, dynamics of speech, zero-crossing rate or combinations of different features [Tanyer and Ozer, 2000]. Some SAD techniques use statistical models to detect the activity of the source based on the average of the log-likelihood ratios between the observed signal and background noise in individual frequency bins [Sohn et al., 1999].

Research in recent years has focused on developing robust SAD systems [Ramirez et al., 2003, Germain et al., 2013, Alam et al., 2014]. However, in a noise environment that has low SNR, SAD is still a serious challenge. In this thesis, we will not go into detail about the SAD technique. We assume that the SAD is not always perfect, and there can be 0% to 10% false detections, corresponding to false positive or false negative.

2.3 Sequential filtering for a single source

Robots equipped with microphones for sensing sound signals have been introduced quite a long time ago [Kato et al., 1974,Kato et al., 1987]. However, they were mostly used to recognize simple voice commands from humans. Over the last two decades, robot audition started getting more attention in robotics and signal processing [Hashimoto et al., 1997, Nakadai et al., 2000, Valin et al., 2007a, Nakadai et al., 2002, Okuno and Nakadai, 2015]. Most research still utilized robot audition in a similar way as with a static microphone array [Nakadai et al., 2000,Kim et al., 2008, Nakamura et al., 2012]. Some research has focused on exploiting robot motion for improving

signal processing results, especially in source localization [Valin et al., 2007b, Martinson and Schultz, 2009, Lu and Cooke, 2011, Portello et al., 2014, Vincent et al., 2015]. Robots provide a mobile platform to continuously take AoA measurements at different locations and orientations of the microphone array. By using a sequential filtering algorithm for integrating robot motion with AoA measurements over time, we can reduce the uncertainty in the source AoA. In addition, the distance to the source, which is unavailable for a static microphone array, can be attained with robot audition.

In this section, we provide an overview of sequential filtering techniques for estimating the location of a single source. In those techniques, the information about the pose of the robot can be obtained from a laser sensor or a motion capture system with high accuracy and therefore is often assumed to be known.

In the following, we will use notations which are common in robotics and may be different from the notations used in signal processing.

2.3.1 State vector

Most filtering techniques for robot audition consider only the case of a continuous sound source when there is no silence in the sound signal [Portello et al., 2011, Martinson and Schultz, 2006, Vincent et al., 2015]. Commonly, the state vector can be defined as follows:

$$X = \begin{bmatrix} X_{\mathrm{r}} \\ X_{\mathrm{s}} \end{bmatrix} = \begin{bmatrix} x_{\mathrm{r}} \\ y_{\mathrm{r}} \\ \theta_{\mathrm{r}} \\ x_{\mathrm{s}} \\ y_{\mathrm{s}} \\ \theta_{\mathrm{s}} \\ v_{\mathrm{s}} \\ w_{\mathrm{s}} \end{bmatrix}, \tag{2.18}$$

where X_{r} is the pose of the robot, i.e., its absolute position $[x_{\mathrm{r}}, y_{\mathrm{r}}]$ and its orientation θ_{r} w.r.t. the x-axis; X_{s} is the state of the sound source, i.e., its absolute position $[x_{\mathrm{s}}, y_{\mathrm{s}}]$, its orientation θ_{s} w.r.t. the x-axis, and its linear and angular velocities $[v_{\mathrm{s}}, w_{\mathrm{s}}]$.

2.3.2 Observation vector

As mentioned in the previous section, a far-field audio source localization technique can estimate the source AoA but not its distance. We assume that at a certain time step k, we obtain an AoA measurement Z_k via a localization technique. The observation model $P(Z_k|X_k)$ represents the likelihood that a given localization technique applied to one signal frame provides the AoA measurement Z_k given the source position X_{sk} and the robot pose X_{rk}. The distribution of the observation model differs for different microphone array geometries. For a linear microphone array, the distribution is bimodal. This is due to the front-back ambiguity that happens when the TDOA is the same for the front and back of the microphone array [Wightman and Kistler, 1999]. As a result, the probability density concentrates around the true AoA and its symmetric w.r.t. the microphone axis, while the probability density for other AoAs is nonzero, but much smaller. Fig. 2.8 shows an example of the observation model $P(Z_k|X_k)$ for a linear microphone array at different AoAs and distances. The observation model is built by applying the MUSIC-GSVD method explained in Section 2.1.2. At large distances, spurious peaks appear at 0° and 180° due to lower SNR.

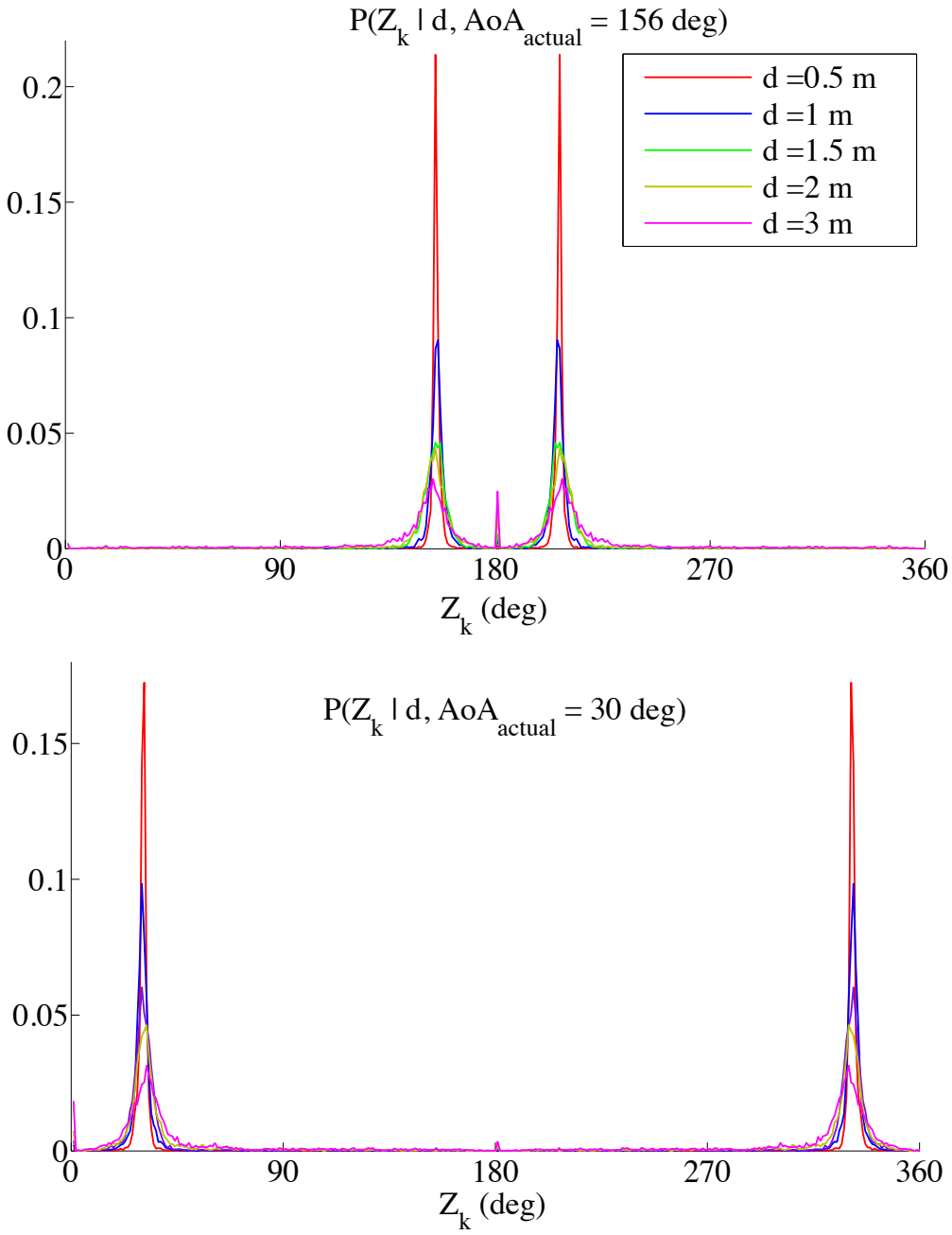

Figure 2.8: Distribution of the measured AoA when the actual source is at different angles 156° (top) and 30° (bottom) and different distances from the microphone array.

For a planar but not linear microphone array, no front-back confusion would occur in the plane anymore, and the observation model would become unimodal.

2.3.3 Recursive Bayesian estimation

Recursive Bayesian estimation is a general probabilistic approach for recursively estimating the belief over the state X_k based on a dynamical model and on all measurements up to the current

time step $Z_{1:k}$. In the problem of estimating the location of a sound source, X_k includes the source location and Z_k is the AoA measurement at time step k. The recursive estimation contains two main steps: the prediction step and the update step.

In the prediction step, the predicted belief $P(X_k|Z_{1:k-1})$ is obtained based on the state transition density $P(X_k|X_{k-1})$ and the previous belief $P(X_{k-1}|Z_{1:k-1})$ as follows:

$$P(X_k|Z_{1:k-1}) = \int P(X_k|X_{k-1})P(X_{k-1}|Z_{1:k-1})dX_{k-1}. \tag{2.19}$$

The state transition probability $P(X_k|X_{k-1})$ is given by the Gaussian distribution $\mathcal{N}(f(X_{k-1}, u_k), Q)$ which can equivalently be defined by the following dynamical model:

$$X_k = f(X_{k-1}, u_k) + d_k, \tag{2.20}$$

where f is the state transition function, u are the robot commands and d_k is the process noise of the robot and the sound source with zero mean and covariance matrix Q. The previous belief $P(X_{k-1}|Z_{1:k-1})$ is the estimated belief at the previous time step $k-1$. In the beginning, when there is no measurement available yet, the prior belief can be uniformly distributed over the state space.

In the update step, when a new measurement Z_k is available, we obtain the posterior belief $P(X_k|Z_{1:k})$ as

$$P(X_k|Z_{1:k}) = \eta P(Z_k|X_k)P(X_k|Z_{1:k-1}), \tag{2.21}$$

where η is a normalizing constant.

Except in the Gaussian case, the solution for this recursive propagation cannot be determined analytically. The source localization model is nonlinear and nongaussian. In that case, nonlinear Kalman filtering, particle filtering or grid-based filtering can be used to approximate the optimal Bayesian solution. In the following, we detail how those filters are implemented to estimate the posterior belief over time.

2.3.4 Nonlinear mixture Kalman filtering

To deal with the nongaussian distribution of the observation model, the MKF models the estimated belief by a mixture of Gaussians:

$$P(X_k|Z_{1:k}) = \sum_{i=1}^{N_k} \omega_{k|k}^i \mathcal{N}(\hat{X}_{k|k}^i, \hat{P}_{k|k}^i), \tag{2.22}$$

where $\omega_{k|k}^i$, $\hat{X}_{k|k}^i$ and $\hat{P}_{k|k}^i$ are the weight, mean and covariance matrix of each Gaussian component in the mixture, respectively. For handling nonlinear state transition and observation models, two nonlinear extensions of the Kalman filter, the extended Kalman filter (EKF) and the unscented Kalman filter (UKF), are mostly used.

In the EKF, a local linearization of the models is obtained by utilising a Taylor expansion [Jazwinski, 1970]. The state transition and observation matrices are defined to be the following Jacobians

$$F_{k-1}^i = \left.\frac{\partial f(X,u)}{\partial X}\right|_{X=\hat{X}_{k-1|k-1}^i} \tag{2.23}$$

$$H_k^i = \left.\frac{\partial h(X)}{\partial X}\right|_{X=\hat{X}_{k|k-1}^i}, \tag{2.24}$$

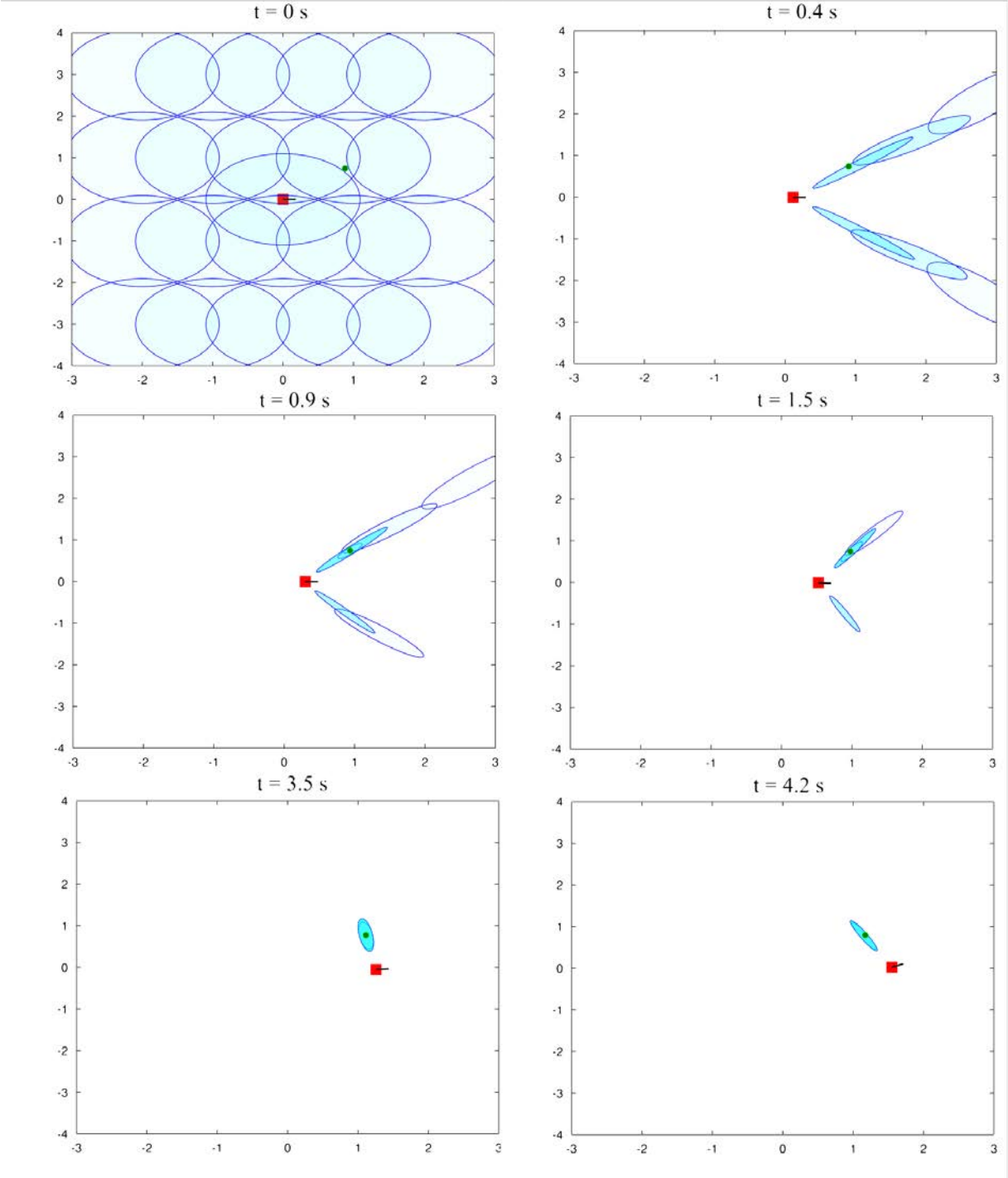

Figure 2.9: Visualization of an extended MKF in an example scenario of mapping a source location using a mobile robot. Robot positions are shown as red squares, and the actual source position as a green circle. Blue ellipses represent 95% confidence regions of source location estimation of various hypotheses in the mixture with a transparency proportional to the weight of the components.

where h is the observation likelihood function. After linearization, the procedure to estimate the posterior belief is the same as in the linear MKF which involves a prediction and an update step. Fig. 2.9 presents an example of tracking a sound source using an extended MKF. The blue ellipses represent the estimation belief of the source location. After several iterations, the front-back ambiguity is eliminated and the estimated belief of source location is close to the actual source position. In this method, the number of components in the mixture increases exponentially after each iteration due to the bimodality of the observation model. Therefore, it is necessary to prune the hypotheses when their number is above a threshold. In the example above, the threshold is set to 300 components.

The UKF extends the EKF framework [Julier and Uhlmann, 2004] for dealing with nonlinearity in the models. In the UKF, a set of points (sigma points) are sampled deterministically around the current estimation based on its covariance is taken to approximate the probability density. These points are then propagated through the true nonlinear distribution to get more accurate estimation of the mean and covariance of the estimated belief. The UKF avoids the need to calculate the Jacobian as in the EKF. A mixture of UKF has been implemented to track a continuous source as well as an intermittent source in [Portello et al., 2011, Portello et al., 2012, Portello et al., 2014].

2.3.5 Particle filtering

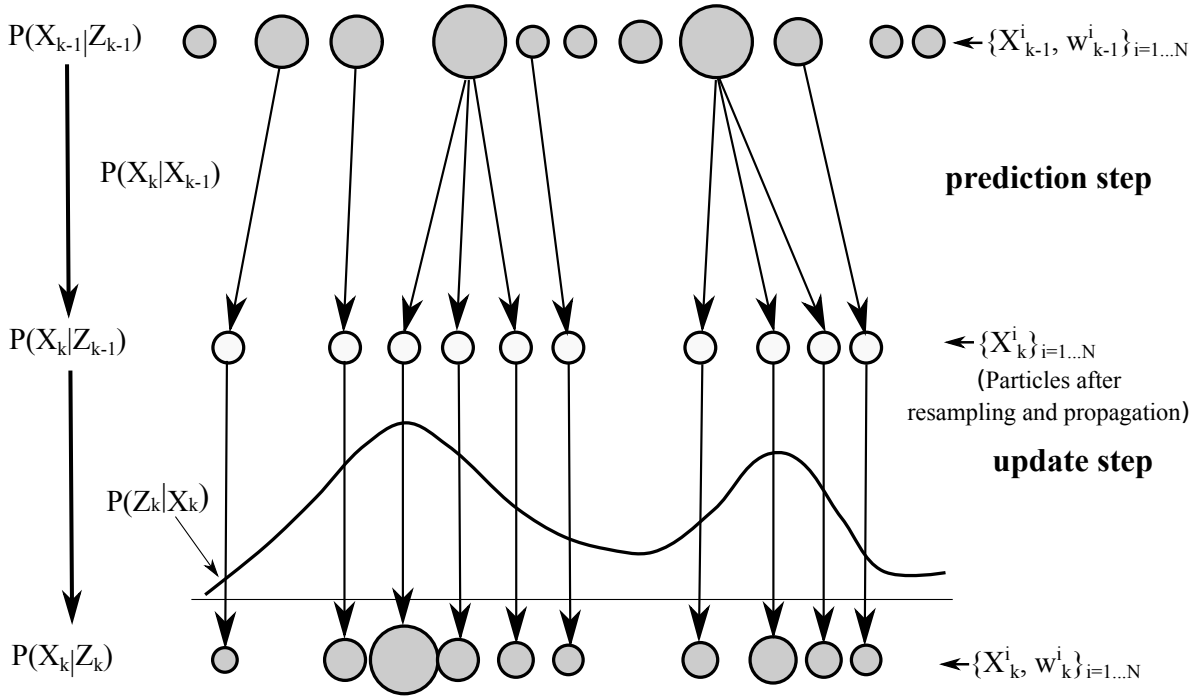

Figure 2.10: Illustration of the particle filter with importance sampling and resampling. The size of each dot is proportional to its weight.

In the particle filtering method, we represent the posterior belief at time step $k-1$ by a set of N state samples (particles) and associated weights $\{X^i_{k-1}, w^i_{k-1}\}_{i=1}^N$. The sequential importance sampling method is the fundamental principle behind most particle filters [Doucet et al., 2000, Arulampalam et al., 2002, Kitagawa, 1996]. The key idea is to approximate the

posterior probability of the source position by drawing several samples (or particles) X^i via importance sampling. The weight of each particle w^i represents the probability of that particle being sampled from the probability density function.

At the beginning, suppose that we have no knowledge of the source location. For initialization, we scatter the particles uniformly over the entire space. So, all the particles have an equal weight

$$w_0^i = 1/N. \tag{2.25}$$

In the prediction step, due to the process noise in the dynamical model, we need to add random noise to the predicted state to have a reasonable chance of capturing the actual movement of the source. The particles are relocated in the state space according to the transition equation

$$X_k^i = f(X_{k-1}^i) + d_k^i, \tag{2.26}$$

where d_k^i is zero mean random noise with covariance matrix Q.

In the update step, when a new measurement Z_k comes in, we assign a weight or probability to each particle based on how well it matches the measurement. This new probability will be multiplied with the current weight of the particle. After normalizing the weights of the particles, those which are matching better the observation will have a higher weight than the others. The weights are updated as

$$w_k^i = \frac{w_{k-1}^i P(Z_k|X_k^i)}{\sum_j w_{k-1}^j P(Z_k|X_k^j)}. \tag{2.27}$$

Fig. 2.10 shows the principle of particle filtering with importance sampling and resampling.

The sequential importance sampling particle filter commonly has a problem called degeneracy. After several iterations, the particles which are far from the true source position will have extremely low weight. On the contrary, only few particles that are near the source have significant weight. A simple method to measure the degeneracy of the particle filter is to determine the effective sample size N_{eff}

$$N_{\text{eff}} = \frac{1}{\sum_i {w_k^i}^2}. \tag{2.28}$$

A small value of N_{eff} indicates severe degeneracy. A common method for solving the degeneracy problem is to use particle resampling. When N_{eff} falls below some threshold, we resample the particles by eliminating particles with very low weights and replacing them with particles with higher weights. The various types of particle filters proposed in the literature can be derived from the sequential importance sampling algorithm by different choices of importance sampling or resampling strategies. The most frequently encountered algorithms are multinomial resampling [Gordon et al., 1993], stratified resampling [Doucet et al., 2001, Kitagawa, 1996], systematic resampling [Kitagawa, 1996, Arulampalam et al., 2002] and residual resampling [Liu and Chen, 1998].

With the advantage in handling any state-space model, particle filtering had been widely used to track acoustic sources [Ward et al., 2003, Lehmann and Williamson, 2006, Valin et al., 2007b, Fallon and Godsill, 2012, Marković et al., 2013]. Fig. 2.11 shows an example of using a particle filter to track the source location over time.

2.3.6 Occupancy grids

Another approach for mapping the location of a sound source is based on occupancy grids. The basic idea is to represent the environment as a discrete grid whose cells are binary random

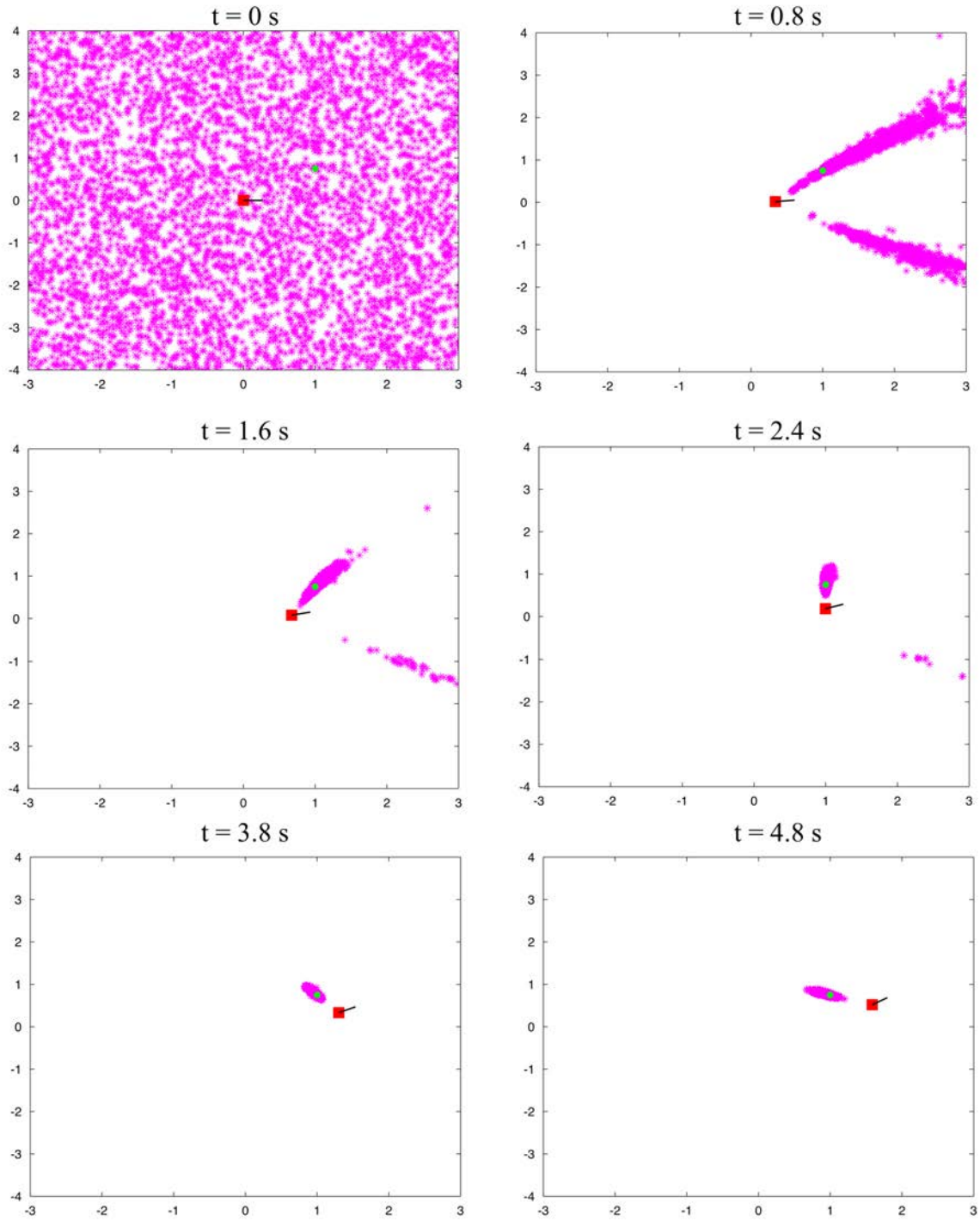

Figure 2.11: Visualization of a particle filter in an example scenario of mapping a source location using a mobile robot. Robot positions are shown as red squares, the actual source position as a green circle. There are 5000 particles which are represented by stars.

variables. Each variable represents the presence or absence of an audio source at that location in the environment. The goal of grid mapping is to estimate the posterior probability of these

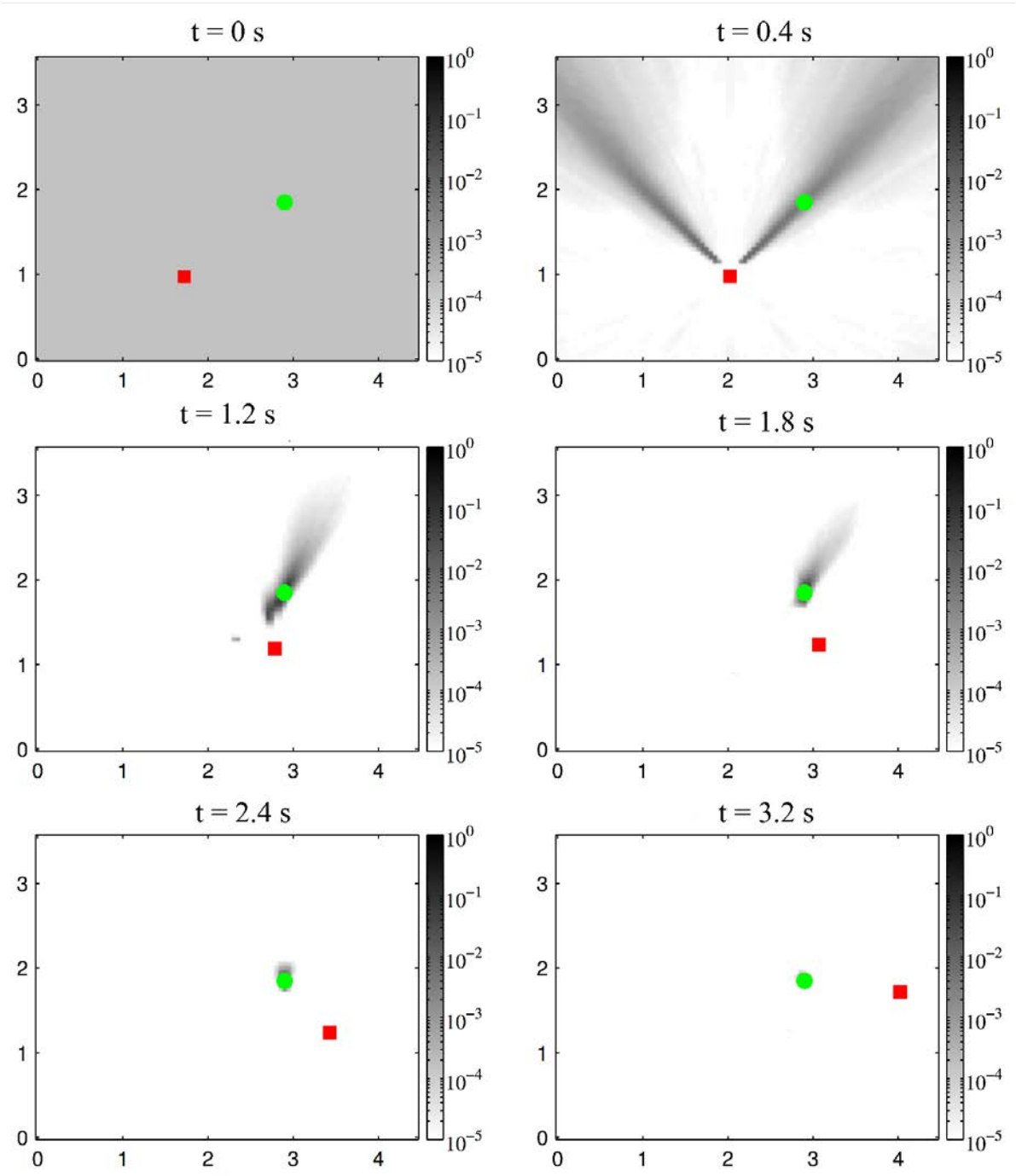

Figure 2.12: Visualization of an occupancy grid map in an example scenario of mapping a source location using a mobile robot [Vincent et al., 2015]. Robot positions are shown as red squares, the actual source position as a green circle.

variables given a set of measurements up to the current time step. One problem with estimating the posterior is that the number of possible maps grows exponentially with the size of the grid, hence approximations are required.

A common approximation method is to estimate the occupancy of each cell separately and

independently [Martinson and Schultz, 2006, DeJong, 2012]. Then, a recursive Bayesian filter can be used to estimate the occupancy probability for each grid cell. However, this approach may lead to inconsistent maps. When the robot does not move, the probabilities will converge to 1 for all cells lying in the direction from the source to the microphone array, while the source is actually located in one of those cells only. In addition, the probabilities do not sum up to 1 which can cause a problem for quantifying the amount of information carried by the grid.

A rigorous approach [Vincent et al., 2015] based on a probabilistic forward sensor model [Thrun, 2003] has been proposed for solving the above problem when localizing a static source. The occupancy grid map is updated over time based on this sensor model by recursive Bayesian estimation. This approach overcomes the critical independence assumption between neighboring grid cells and generates a more accurate map. Fig. 2.12 shows an example of using an occupancy grid to map the source location over time. The method is applicable for tracking a moving source by multiple grids [Duan et al., 2016]. However, it is more complicated and requires greater computational cost.

2.4 Sequential filtering for multiple sources

In a normal situation, there could be two or more sources present in the environment. The problem of estimating the positions of multiple target sound sources is more challenging than for a single target source. In the context of multiple target sources, the observation contains more uncertainty since the peaks of the MUSIC spectrum may be occluded and false alarms may occur. As a result, the observation at each time step is a set of measurements which are not all generated by real targets and we have no clue about which targets generated which measurements. In the observation, in general, it is assumed that each target generates measurements independently from the other targets and from false alarms. Moreover, each measurement can only be generated from at most one target.

The assignment of a given measurement to a target will make the main difference between multiple target tracking methods. Let's take a quick look at sequential filtering algorithms for multiple targets.

The simplest sequential filtering for multiple targets is the global nearest neighbor (GNN) tracker [Blackman and Popoli, 1999]. The GNN approach evaluates each association of measurements to targets and select the one that optimizes a cost, e.g., distance to a predicted location. Then, we can apply Bayesian filtering for each target based on these associations. Although the GNN is simple, it could easily lead to conflicting associations, e.g., when more than one target is associated with the same measurement.

One of the most popular algorithms for tracking multiple targets is the JPDAF [Bar-Shalom, 1990, Cox, 1993]. In the JPDAF, joint association events and joint association probabilities are used to avoid conflicting measurements and to track assignments in the presence of multiple targets. The basic JPDAF assumes a fixed and known number of targets. The complexity of computing joint association probabilities grows exponentially with the number of targets and the number of measurements.

The multiple hypothesis tracking (MHT) method [Reid, 1979, Blackman, 2004] is a delayed decision approach to data association based multiple target tracking. After each iteration, the MHT algorithm will only maintain a set of association hypotheses with high posterior probability. Based on those hypotheses, the estimation of individual targets will be updated by Bayesian filtering. In the MHT algorithm, the total number of possible hypotheses increases exponentially with time. Therefore, we need to consider heuristic pruning/merging of hypotheses after each

iteration to reduce computational requirements.

Recently, the random finite set (RFS) approach is an emerging paradigm for multiple target tracking [Vo and Ma, 2006, Mahler, 2007, Goutsias et al., 2012]. In the RFS approach, the states of objects are represented as random sets. Random sets are random elements whose values are sets. Then, a probability hypothesis density (PHD) filter based on RFS is used to approximate the first moment of the multiple target posterior density. The PHD filter operates on a single target state space and can avoid the problem that arises from data association.

These multiple targets tracking methods have been applied to track multiple sound sources [Gehrig and McDonough, 2006, Chakrabarty et al., 2014, Lee et al., 2010, Levy et al., 2011, Oualil et al., 2012, Evers et al., 2016b]. [Gehrig and McDonough, 2006, Chakrabarty et al., 2014] used the JPDAF algorithm to estimate the locations of sound sources. The MHT is implemented in [Levy et al., 2011, Oualil et al., 2012]. Currently, the RFS approach is used by [Evers et al., 2016b]. The belief about the source locations over time is often estimated by an EKF [Chakrabarty et al., 2014, Oualil et al., 2012] or particle filtering [Levy et al., 2011, Valin et al., 2007b, Pertilä and Hämäläinen, 2010] However, most of the above approaches only consider tracking multiple sources when the microphone array is not moving and the microphones are often distributed around the room area [Gehrig and McDonough, 2006, Levy et al., 2011, Oualil et al., 2012]. In addition, the estimation of the source activity is not included in these frameworks.

In this section, we present an overview of the sequential filtering method using the JPDAF for estimating the location of multiple target sources.

2.4.1 State vector

Considering the problem of tracking n target sources. Assuming that the sources move independently from each other, we can define the state vector of each source as:

$$X^i = \begin{bmatrix} x^i \\ y^i \\ \theta^i \\ v^i \\ w^i \end{bmatrix}, \tag{2.29}$$

where $i = \{1, 2, .., n\}$ is the index of the source, X^i includes the absolute position $[x^i, y^i]$ of the source, its orientation θ^i w.r.t. the x-axis, and its linear and angular velocities $[v^i, w^i]$.

The robot pose is assumed to be known over time. The notations for multiple source tracking are more complex than for a single source. So, to simplify, we omitted the state of the robot in the state vector and only denote X^i as the state of each source.

2.4.2 Observation vector

In the observation vector, we assume that there are m AoA measurements at each time step. Depending on the microphone array geometry, the distribution of the observation model corresponding to each AoA measurement Z^j, where $j = 1, 2, .., m$, will be bimodal (for a linear microphone array) or unimodal (for a planar but not linear microphone array). The key question in tracking multiple target sources is how to assign the observed AoAs to the individual target sources.

2.4.3 Joint probabilistic data association filter

In the following, we present a JPDAF framework with the Bayesian filtering framework which has been used in [Gehrig and McDonough, 2006, Chakrabarty et al., 2014] to estimate the source locations. In this framework, all target sources are assumed to be active. The source locations are estimated recursively with the prediction step and update step as follows.

2.4.3.1 Prediction step

Given the belief $P(X_{k-1}^i|Z_{1:k-1})$ about the location of target source i at the previous time step $k-1$, the predicted belief $P(X_k^i|Z_{1:k-1})$ is computed based on the state transition model:

$$P(X_k^i|Z_{1:k-1}) = \int P(X_k^i|X_{k-1}^i)P(X_{k-1}^i|Z_{1:k-1})dX_{k-1}^i. \tag{2.30}$$

2.4.3.2 Update step

Whenever new measurements arrive, the updated belief $P(X_k^i|Z_k)$ is estimated according to:

$$P(X_k^i|Z_{1:k}) = \eta P(Z_k|X_k^i)P(X_k^i|Z_{1:k-1}), \tag{2.31}$$

where η is a normalization factor.

In the JPDAF framework, we define a joint association event β_{ij} which determines the assignment of measurement j to target i. The situation when a measurement j does not correspond to any target is also considered. Since we do not know which of the measurements in Z_k is originated by target i, we integrate the single measurements according to the assignment probabilities $P(\beta_{ij})$ with:

$$\sum_{j=1}^{m}\sum_{i=0}^{n} P(\beta_{ij}) = 1. \tag{2.32}$$

Then, the updated belief for each target can be expressed as:

$$P(X_k^i|Z_{1:k}) = \eta \sum_{j=1}^{m} P(\beta_{ij})P(Z_k^j|X_k^i)P(X_k^i|Z_{1:k-1}). \tag{2.33}$$

2.5 Motion planning for robot audition

One of the key abilities of autonomous robots is to do motion planning for effectively achieving a specific objective. For robot audition, the objective of motion planning can be to quickly estimate the location of a source with minimum uncertainty.

In this section, we first provide an overview of robot motion planning techniques in general. Then, we discuss state-of-the-art motion planning algorithms that have been applied to robot audition.

2.5.1 General robot motion planning

Motion planning for mobile robot platforms has received extensive interest in robotics [Latombe, 1991, Torras, 1992, Meyer and Filliat, 2003, LaValle, 2006, Colas et al., 2013]. Conventionally, motion planning is considered as the problem of searching a feasible trajectory which guides the robot from its current position to a target position [Latombe, 1991]. The development of

robot motion planning includes planning with different constraints or objectives, planning under uncertainty, etc.

In classical motion planning, avoiding collision is the most common goal [LaValle, 2006]. A complete description of the geometry of a robot and a static environment are given. The final position for the robot to reach is also provided. The robot needs to find a collision-free path to move from the current position and orientation to the final position and orientation. A framework of classical motion planning is illustrated in Fig. 2.13. In order to simplify the planning problem, in the first step, we define a configuration space in which the complex geometry of the robot is considered as a point. Normally, the configuration of a mobile robot is its position and orientation. A configuration space is free if the robot placed in that space does not collide with the obstacles. By defining the configuration space, we can transform the original problem to that of motion planning for a point. In the next step, the free configuration space is discretized and a graph that represents the connectivity of the points in the space is constructed. Finally we search this graph to find a suitable path for the robot to reach the goal. In the situation that no path is found, we can try again by refining the discretization.

Different approaches in classical motion planning often differ in the construction of the connectivity graph and graph search. For the graph search, a desired path can be found by standard graph search techniques, such as Dijkstra's algorithm or the A* algorithm [Dijkstra, 1959, Hart et al., 1968]. For the first criterion, there are several general approaches for path planning such as cell decomposition or roadmaps [Latombe, 1991, Nilsson, 1969, Khatib, 1986]. In their original forms, all the above approaches can effectively solve motion planning problems in 2-D configuration spaces. However, these approaches face computational cost issues when scaled up to configuration spaces of higher dimensions.

In the recent years, random sampling has emerged as a powerful approach for motion planning [Boor et al., 1999, Siméon et al., 2000, Sun et al., 2005, Panchea et al., 2017]. Although the random sampling method is first targeted for robot manipulators with many degrees of freedom, the configuration space framework allows us to use it for mobile robots equally well. Algorithms based on random sampling, e.g., the probabilistic roadmap planner, are both efficient and simple to implement. Informally speaking, sampling-based methods provide large amounts of computational savings by avoiding explicit construction of obstacles in the state space, as opposed to classical motion planning algorithms that require explicit representation of the obstacles in the configuration space.

Normally, uncertainty often exists at both planning and execution time of the robot, especially in the dynamical model of the robot and the robot sensors. Even the most complex models or expensive sensors cannot be perfectly accurate. In addition, increasing model complexity often reduces the effectiveness of motion planning. Therefore, a simplified model is often selected instead. Classical motion planning methods, which assume null uncertainty, are clearly insufficient due to errors in robot sensing and motion. Effective planning must take these uncertainties into account. With imperfect state information, a robot cannot decide the best action based on a single known state. Instead, the expected best action will depend on the set of all possible states consistent with the available information. This will result in much higher computational complexity for motion planning to decide the best actions. Partially observable Markov decision processes (POMDPs) provide a principled mathematical framework for such planning tasks [Kaelbling et al., 1998, Hsiao et al., 2007, Araya-López et al., 2010]. A POMDP models a robot taking a sequence of actions under uncertainty to maximize its expected total reward. The state in POMDPs is partially observable and not known exactly. Solutions to POMDPs are known to be extremely complex and have only been successfully applied to very small and low-dimensional state spaces.

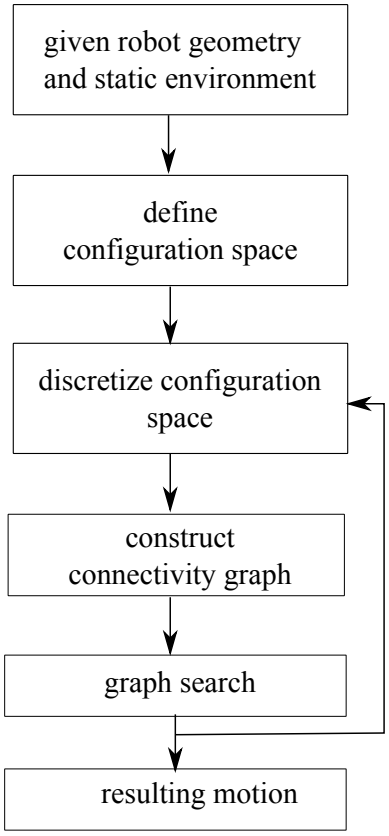

Figure 2.13: General framework of graph-based robot motion planning.

Besides the objective of avoiding obstacles when moving, there are other objectives such as minimizing the time for reaching the goal, minimizing the consumed energy, or acquiring information about the environment. Gathering information about the environment through sensing is an important task for surveillance and mapping of unknown environments. This task is also called as robot exploration. Motion planning plays a key role in improving exploration tasks. The purpose of sensing is to acquire an understanding of the surrounding environment from sensor data. The basic act of sensing is passive. It becomes active when the robot motion directs the sensor to make sensing more effective. Robot motion can help the robot keep a target within the sensor range or gain new information about an unexplored environment. More generally, a robot motion is executed to maintain a set of constraints on the state of the world or to achieve a certain state of knowledge about the world.

Given the map of the environment known so far, where should the robot move next to observe the unexplored regions is a question that an autonomous robot needs to answer. In typical robot exploration, the general answer is moving to the direction that can maximize the explored area. One strategy is to move the robot to the nearest frontier [Yamauchi, 1997], a border between empty and unknown environment. A next move is decided based on next-best view approach [Gonzalez-Banos and Latombe, 2002]. A multi-objective strategy to find the optimal exploration path under multiple constraints and objectives is introduced in [Amanatiadis et al., 2013]. Most of these approaches consider a static environment in which the robot finds a destination for gathering information and then moves there without additional planning during the travel. For the case of dynamic environments with moving objects, continuous planning

during movement is necessary.

Another branch of robot motion control method includes vision-based robot control, a.k.a the visual servo [Agin, 1979, Espiau et al., 1992, Hutchinson et al., 1996, Chaumette and Hutchinson, 2006]. This technique use feedback information extracted from a vision sensor to control the robot motion instead of planning ahead. The robot motion is based on the error between current and desired features in the observations, and does not involve any estimation of the pose of the target.

2.5.2 Motion planning for robot audition

Motion planning for robot audition has started attracting attention. There exists uncertainty in the source AoA measurement. The main objective of motion planning for robot audition is to find a robot trajectory that will maximize the information about the source location or minimize the uncertainty on the source location.

In a simple strategy, the localization accuracy can be improved by following a fixed patrol loop and covering a potential maneuver area [Martinson and Schultz, 2009]. This approach is suboptimal because it often takes a lot of time for finding a potential area. In addition, if the target source is moving, the detected area will change. Therefore, it is not an efficient planning method and not posed as a formal optimization problem.

More sophisticated motion strategies based on information-theoretic criteria, i.e., Shannon entropy, have been proposed [Sommerlade and Reid, 2008, Bustamante et al., 2017, Kumon et al., 2010]. The target is to maximize the estimation accuracy. However, this information is not available for robots since it requires the ground-truth source location. Instead of that, the general idea is to drive the robot in the direction which leads to minimum uncertainty on the source location. Let us assume that the robot has taken measurements up to a certain time step k. All the knowledge about the source location and robot pose at time k is represented by the belief $P(X_k|Z_{1:k})$. The uncertainty on the source location at time k can be quantified as the entropy of the belief

$$H(P(X_k|Z_{1:k})) = -\int P(X_k|Z_{1:k}) \log P(X_k|Z_{1:k}) dX_k. \tag{2.34}$$

The lower the entropy, the greater the amount of information. In [Bustamante et al., 2017], a gradient descent method was proposed to find the robot movement u_{k+1} that minimizes the entropy of the belief on the position of a static sound source one time step ahead. Since the measurement Z_{k+1} is unavailable at current time step, we can only compute its expectation. Therefore, the next optimal motion would be selected as below:

$$u_{k+1} = \arg\min_{u_{k+1}} \mathbb{E}_{Z_{k+1}}[H(P(X_{k+1}|Z_{1:k+1}))]. \tag{2.35}$$

This method yields a locally optimal robot motion but, in the long run, this sequence of local optima is generally not globally optimal. In addition, in [Bustamante et al., 2017] the belief about the source and robot positions is represented by a mixture of Gaussians but the determination of the best possible move is based on the single Gaussian with largest weight.

An approach using Monte Carlo exploration to sample and select the next optimal action was proposed by [Schymura et al., 2017]. In this approach, the final goal is predefined. The optimal sequence motion will minimize both the estimation uncertainty and the distance to the goal. This motion planning technique with a predefined goal is suboptimal when the target source is far from the robot's final destination.

A long-term motion planning method was introduced by using a dynamic programming algorithm to find the optimal trajectory for localizing a static source by [Vincent et al., 2015]. In order to find the optimal motion at time $k+1$, we need to compute the entropy conditionally to the motion sequence $u_{k+1:k+T}$ for all possible motion sequences up to a fixed horizon $k+T$. This entropy cannot be deterministically computed, since future measurements $Z_{k+1:k+T}$ are unavailable, but its expectation can be expressed as

$$H_T = \sum_{i=1}^{T} \mathbb{E}_{Z_{k+1:k+i}}[H(P(X_{k+i}|Z_{1:k+i}))], \tag{2.36}$$

where $\mathbb{E}$ stands for the expectation. The optimal motion is selected so that the value of H_T is minimum

$$u_{k+1} = \underset{u_{k+1:k+T}}{\arg\min}\, H_T. \tag{2.37}$$

This method [Vincent et al., 2015] approximates the sum of entropies over a finite time horizon by assuming that the entropy at each future pose does not depend on the trajectory used to reach that pose. This assumption is required for dynamic programming, but the entropy actually depends on the followed trajectory in practice. Therefore, it is also potentially suboptimal. Another long-term robot motion planning algorithm for a binaural sensor is proposed by [Bustamante and Danès, 2017] recently. An optimal motion of the robot is obtained by solving a constrained optimization problem involving the gradient of the reward function.

An alternative motion control approach, the so-called aural servo, is recently introduced [Magassouba, 2016]. This approach actually does not localize sound sources nor plan ahead. It selects and executes a robot motion that satisfies given conditions characterized by a set of auditory measurements.

Chapter 3

Source localization in a reverberant environment

In this chapter, we first present our contribution on developing a nonlinear MKF framework for estimating the location and activity of an intermittent and possibly moving source. By explicitly estimating the source activity in addition to the source location, the filtering framework gains more robustness to false measurements in both the SAD and the AoA than the state-of-the-art methods. We evaluate this sequential filtering method with a linear microphone array where the front-back ambiguity exists.

Second, we present our contribution on using particle filtering for localizing an intermittent and possibly moving source. In the particle filtering method, we also jointly estimate the location and activity of the target source. We then compare particle filtering with the extended MKF above in term of localization performance and computational time.

3.1 Proposed Bayesian filtering framework

3.1.1 State vector

Most of the methods in the literature consider only the state of source location when localizing a continuous source as well as an intermittent source [Portello et al., 2012, Portello et al., 2014]. By contrast with them, our proposed method also takes the source activity into account in the state vector. Therefore, we define the state vector as follows:

$$\begin{bmatrix} X \\ a \end{bmatrix} = \begin{bmatrix} X_{\mathrm{r}} \\ X_{\mathrm{s}} \\ a \end{bmatrix} = \begin{bmatrix} x_{\mathrm{r}} \\ y_{\mathrm{r}} \\ \theta_{\mathrm{r}} \\ x_{\mathrm{s}} \\ y_{\mathrm{s}} \\ \theta_{\mathrm{s}} \\ v_{\mathrm{s}} \\ w_{\mathrm{s}} \\ a \end{bmatrix}, \tag{3.1}$$

where X_{r} is the pose of the robot, i.e., its absolute position $[x_{\mathrm{r}}, y_{\mathrm{r}}]$ and its orientation θ_{r} w.r.t. the x-axis; X_{s} is the continuous state of the sound source, i.e., its absolute position $[x_{\mathrm{s}}, y_{\mathrm{s}}]$, its orientation θ_{s} w.r.t. the x-axis, and its linear and angular velocities $[v_{\mathrm{s}}, w_{\mathrm{s}}]$; a is the source

activity which is a discrete variable, where $a = 1$ indicates that the source is active, otherwise $a = 0$. So, we have both the continuous variable X and discrete variable a in the state vector.

We assume that the pose of the robot X_r is known and we need to estimate the continuous variable X_s and the activity a of the source.

3.1.2 Dynamical model

3.1.2.1 Dynamical model of the robot

The dynamical model of the robot at time k can be written as

$$X_{\mathrm{r}k} = f_\mathrm{r}(X_{\mathrm{r}k-1}, u_k) + d_{\mathrm{r}k}, \tag{3.2}$$

where u are the robot commands which consist of the angular speeds v_l and v_r of the left and right wheel of the robot, and d_r is the process noise of the robot that has a Gaussian distribution with zero mean and covariance matrix Q_r. Due to the above assumption that we know the pose of the robot, the process noise d_r is set to zero. We model the state transition function f_r as

$$f_\mathrm{r}(X_\mathrm{r}, u) = X_\mathrm{r} + \begin{bmatrix} \frac{r}{2}(v_\mathrm{r} + v_\mathrm{l})\cos(\theta_\mathrm{r}) \\ \frac{r}{2}(v_\mathrm{r} + v_\mathrm{l})\sin(\theta_\mathrm{r}) \\ \frac{r}{l}(v_\mathrm{r} - v_\mathrm{l}) \end{bmatrix} dt, \tag{3.3}$$

where l is the distance between the two robot wheels, r is the wheel radius, and dt is the time step or the sampling period [Siegwart et al., 2011].

Note that the control input u for the robot is given during the estimation, so for simplicity it is later omitted from the equations.

3.1.2.2 Dynamical model of the sound source

The dynamical model of the target sound source at time k is defined as follows:

$$X_{\mathrm{s}k} = f_\mathrm{s}(X_{\mathrm{s}k-1}) + d_{\mathrm{s}k}, \tag{3.4}$$

where d_s denotes the process noise of the sound source that has a Gaussian distribution with zero mean and covariance matrix Q_s, and f_s is modeled as

$$f_\mathrm{s}(X_\mathrm{s}) = X_\mathrm{s} + \begin{bmatrix} v_\mathrm{s}\cos(\theta_\mathrm{s}) \\ v_\mathrm{s}\sin(\theta_\mathrm{s}) \\ w_\mathrm{s} \\ 0 \\ 0 \end{bmatrix} dt. \tag{3.5}$$

3.1.2.3 Full dynamical model

We can write the full dynamical model of the system at time k as follows:

$$X_k = f(X_{k-1}) + d_k, \tag{3.6}$$

where $f(X) = \begin{bmatrix} f_\mathrm{r}(X_\mathrm{r}) \\ f_\mathrm{s}(X_\mathrm{s}) \end{bmatrix}$ and $d = \begin{bmatrix} d_\mathrm{r} \\ d_\mathrm{s} \end{bmatrix}$ is the process noise with covariance matrix $Q = \begin{bmatrix} Q_\mathrm{r} & 0 \\ 0 & Q_\mathrm{s} \end{bmatrix}$.

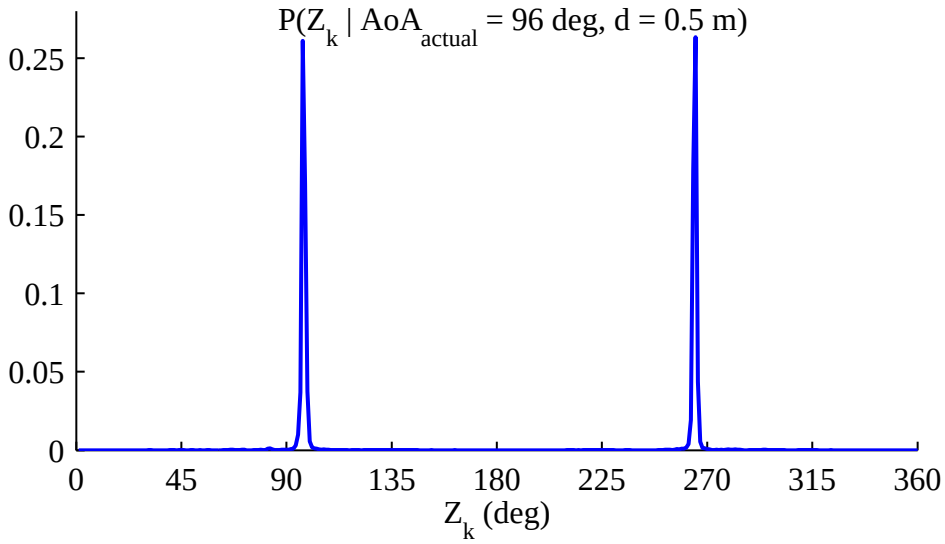

Figure 3.1: Distribution of the measured AoA when the actual source is at 96° and 0.5 m from the microphone array.

3.1.3 Observation vector

As mentioned in the previous chapter, audio source localization techniques can estimate the source AoA but not its distance. Therefore, we assume that the observation vector Z_k in a given time frame k consists of one AoA measurement Z_k^{l} (obtained via a localization technique) and one source activity measurement Z_k^{a} (obtained via an SAD technique).

The likelihood of the state vector w.r.t. this observation can be expressed as

$$P(Z_k|X_k, a_k) = \begin{cases} P_{\mathrm{sn}}(Z_k^{\mathrm{l}}|X_k)P(Z_k^{\mathrm{a}}|a_k) & \text{for } a_k = 1 \\ P_{\mathrm{n}}(Z_k^{\mathrm{l}})P(Z_k^{\mathrm{a}}|a_k) & \text{for } a_k = 0, \end{cases} \tag{3.7}$$

with P_{sn} and P_{n} denoting the distribution of the measured AoA when the source is active or inactive, respectively. In the latter case, it is supposed that the recorded signal consists of spatially diffuse noise, so P_{n} will have a uniform distribution and does not depend on X_k.

An example of P_{sn} is shown in Fig. 3.1 for our linear microphone array and the localization technique MUSIC-GSVD. The probability density concentrates around the true AoA and its symmetric w.r.t. the microphone axis: this phenomenon is known as front-back confusion. The probability density for other AoAs is nonzero, but much smaller. Therefore we can approximate the observation model by a mixture of two Gaussians:

$$P_{\mathrm{sn}}(Z_k^{\mathrm{l}}|X_k) = \sum_{j=1}^{2} \frac{1}{2} \mathcal{N}(Z_k^{\mathrm{l}}; \hat{Z}_k^{\mathrm{l},j}, R_k^j). \tag{3.8}$$

For a nonlinear microphone array, the observation model could be represented by a single Gaussian instead.

3.1.4 Recursive Bayesian estimation

The transition probability between time steps is given by:

$$P(X_k, a_k|X_{k-1}, a_{k-1}) = P(a_k|X_{k-1}, a_{k-1})P(X_k|X_{k-1}, a_{k-1}, a_k). \tag{3.9}$$

The source activity a_k and the state X_k are conditionally independent so we can rewrite the above equation as

$$P(X_k, a_k|X_{k-1}, a_{k-1}) = P(a_k|a_{k-1})P(X_k|X_{k-1}). \tag{3.10}$$

The state transition probability $P(X_k|X_{k-1})$ follows the dynamical model of the robot (3.2) and the source (3.4).

The source activity transition probability $P(a_k|a_{k-1})$ is defined by $P_{\text{appear}} = P(a_k = 1|a_{k-1} = 0)$ which is the source appearance probability and $P_{\text{disappear}} = P(a_k = 0|a_{k-1} = 1)$ which is the source disappearance probability.

The posterior probability of the state vector can be recursively computed by alternating these two steps:

- **prediction step**:

  compute $P(X_k, a_k|Z_{1:k-1})$ given the previous belief $P(X_{k-1}, a_{k-1}|Z_{1:k-1})$ and the state transition model:

$$\begin{aligned} P(X_k, a_k|Z_{1:k-1}) &= \sum_{a_{k-1}} \int P(X_k, a_k|X_{k-1}, a_{k-1})P(X_{k-1}, a_{k-1}|Z_{1:k-1})dX_{k-1} \\ &= \sum_{a_{k-1}} \int P(a_k|a_{k-1})P(X_k|X_{k-1})P(X_{k-1}, a_{k-1}|Z_{1:k-1})dX_{k-1}. \end{aligned} \tag{3.11}$$

- **update step**:

  recompute the belief $P(X_k, a_k|Z_{1:k})$ given the prediction and the new measurement Z_k:

$$P(X_k, a_k|Z_{1:k}) = \eta P(Z_k|X_k, a_k)P(X_k, a_k|Z_{1:k-1}), \tag{3.12}$$

  where η is a normalizing constant.

3.2 Extended mixture Kalman filtering

Since the state vector includes both continuous and discrete variables and the observation model is a mixture of Gaussians, we can propose an extended MKF to address these two issues. We present in detail the extended MKF in the following.

3.2.1 Prediction step

We assume that at the previous time step $k-1$ the belief about the state (X_{k-1}, a_{k-1}) is given by the mixture of $N_{k-1|k-1}$ Gaussians:

$$P(X_{k-1}, a_{k-1}|Z_{1:k-1}) = \sum_{i=1}^{N_{k-1|k-1}} \omega^i_{k-1|k-1} \mathcal{N}(X_{k-1}; \hat{X}^i_{k-1|k-1}, P^i_{k-1|k-1})\delta(a^i_{k-1}), \tag{3.13}$$

with weights $\omega^i_{k-1|k-1}$ such that $\sum_i \omega^i_{k-1|k-1} = 1$.

Applying the prediction rule (3.11) to this density yields the predicted density:

$$\begin{aligned} P(X_k, a_k|Z_{1:k-1}) = \sum_{i=1}^{N_{k-1|k-1}} \omega^i_{k-1|k-1} [&P(a^i_k = 0|a^i_{k-1})\mathcal{N}(X_k; \hat{X}^i_{k|k-1}, P^i_{k|k-1})\delta(a^i_k = 0) \\ &+ P(a^i_k = 1|a^i_{k-1})\mathcal{N}(X_k; \hat{X}^i_{k|k-1}, P^i_{k|k-1})\delta(a^i_k = 1)]. \end{aligned} \tag{3.14}$$

The predicted density $P(X_k, a_k|Z_{1:k-1})$ is thus also a mixture of Gaussians, with the number of components $N_{k|k-1} = 2N_{k-1|k-1}$ and can be expressed as follows:

$$P(X_k, a_k|Z_{1:k-1}) = \sum_{i=1}^{N_{k|k-1}} \omega^i_{k|k-1}\mathcal{N}(X_k; \hat{X}^i_{k|k-1}, P^i_{k|k-1})\delta(a^i_k). \tag{3.15}$$

Its means, variances and weights are given by

$$\hat{X}^i_{k|k-1} = f(\hat{X}^i_{k-1|k-1}, u_k), \tag{3.16}$$

$$F^i_{k-1} = \frac{\partial f(X, u_k)}{\partial X}|_{X=\hat{X}^i_{k-1|k-1}}, \tag{3.17}$$

$$P^i_{k|k-1} = F^i_{k-1}P^i_{k-1|k-1}F^{iT}_{k-1} + Q^i_{k-1}, \tag{3.18}$$

$$\omega^i_{k|k-1} = P(a^i_k|a^i_{k-1})\omega^i_{k-1|k-1}. \tag{3.19}$$

3.2.2 Update step

By applying the update rule (3.12) to the predicted density and replacing the observation model by the mixture of two Gaussians in (3.8), we obtain the new belief:

$$\begin{aligned}
P(X_k, a_k|Z_{1:k}) =& \eta \sum_{i=1}^{N_{k|k-1}} \omega^i_{k|k-1}P(Z_k|X_k, a_k)\mathcal{N}(X_k; \hat{X}^i_{k|k-1}, P^i_{k|k-1})\delta(a^i_k) \\
=& \eta \sum_{i=1}^{N_{k-1|k-1}} \omega^i_{k|k-1}\Big[P(Z^{\mathrm{a}}_k|a_k)P_{\mathrm{n}}(Z^{\mathrm{l}}_k)\mathcal{N}(X_k; \hat{X}^i_{k|k-1}, P^i_{k|k-1})\delta(a^i_k = 0) \\
&+ P(Z^{\mathrm{a}}_k|a_k)P_{\mathrm{sn}}(Z^{\mathrm{l}}_k|X_k)\mathcal{N}(X_k; \hat{X}^i_{k|k-1}, P^i_{k|k-1})\delta(a^i_k = 1)\Big] \\
=& \eta \sum_{i=1}^{N_{k-1|k-1}} \omega^i_{k|k-1}\Big[P(Z^{\mathrm{a}}_k|a_k)P_{\mathrm{n}}(Z^{\mathrm{l}}_k)\mathcal{N}(X_k; \hat{X}^i_{k|k-1}, P^i_{k|k-1})\delta(a^i_k = 0) \\
&+ P(Z^{\mathrm{a}}_k|a_k)\sum_{j=1}^{2}\frac{1}{2}\mathcal{N}(Z^{\mathrm{l}}_k; \hat{Z}^{\mathrm{l},j}_k, R^{i,j}_k)\mathcal{N}(X_k; \hat{X}^i_{k|k-1}, P^i_{k|k-1})\delta(a^i_k = 1)\Big] \\
=& \eta \sum_{i=1}^{N_{k-1|k-1}} \omega^i_{k|k-1}\Big[P(Z^{\mathrm{a}}_k|a_k)P_{\mathrm{n}}(Z^{\mathrm{l}}_k)\mathcal{N}(X_k; \hat{X}^i_{k|k-1}, P^i_{k|k-1})\delta(a^i_k = 0) \\
&+ P(Z^{\mathrm{a}}_k|a_k)\sum_{j=1}^{2}\frac{1}{2}\mathcal{N}(Z^{\mathrm{l}}_k; h^j(\hat{X}^i_{k|k-1}), R^{i,j}_k)\mathcal{N}(X_k; \hat{X}^i_{k|k-1}, P^i_{k|k-1})\delta(a^i_k = 1)\Big]
\end{aligned} \tag{3.20}$$

The product of every two Gaussians $\mathcal{N}(Z^{\mathrm{l}}_k; h^j(\hat{X}^i_{k|k-1}), R^{i,j}_k)\mathcal{N}(X_k; \hat{X}^i_{k|k-1}, P^i_{k|k-1})$ can be computed in closed form as $\lambda^{i,j}\mathcal{N}(X_k; \hat{X}^i_{k|k}, P^i_{k|k})$ where $\lambda^{i,j}$ is a constant defined as:

$$\lambda^{i,j} = \frac{1}{\sqrt{\left|2\pi(HP^i_{k|k-1}H^T + R^{i,j}_k)\right|}} e^{-\frac{1}{2}\left[Z^{l,j}_k - h^j(\hat{X}^i_{k|k-1})\right]^T \left[HP^i_{k|k-1}H^T + R^{i,j}_k\right]^{-1} \left[Z^{l,j}_k - h^j(\hat{X}^i_{k|k-1})\right]}. \tag{3.21}$$

Therefore, the new belief can be expressed as a mixture of Gaussians:

$$\begin{aligned} P(X_k, a_k|Z_{1:k}) =& \eta \sum_{i=1}^{N_{k-1|k-1}} \omega_{k|k-1}^i \Big[P(Z_k^{\mathrm{a}}|a_k) P_{\mathrm{n}}(Z_k^{\mathrm{l}}) \mathcal{N}(X_k; \hat{X}_{k|k-1}^i, P_{k|k-1}^i) \delta(a_k^i = 0) \\ &+ P(Z_k^{\mathrm{a}}|a_k) \sum_{j=1}^{2} \frac{1}{2} \mathcal{N}(Z_k^{\mathrm{l}}; h^j(\hat{X}_{k|k-1}^i), R_k^{i,j}) \mathcal{N}(X_k; \hat{X}_{k|k-1}^i, P_{k|k-1}^i) \delta(a_k^i = 1) \Big] \\ =& \sum_{i=1}^{N_{k|k}} \omega_{k|k}^i \mathcal{N}(X_k; \hat{X}_{k|k}^i, P_{k|k}^i) \delta(a_k^i), \end{aligned} \tag{3.22}$$

where its number of components is $N_{k|k} = 3N_{k-1|k-1}$ and its weights:

$$\omega_{k|k}^i = \begin{cases} \omega_{k|k-1}^i P_{\mathrm{n}}(Z_k^{\mathrm{l}}) P(Z_k^{\mathrm{a}}|a_k) \eta & \text{if} \quad a_k = 0 \\ \frac{1}{2} \omega_{k|k-1}^{i'} P(Z_k^{\mathrm{a}}|a_k) \lambda^{i',j} \eta & \text{if} \quad a_k = 1. \end{cases} \tag{3.23}$$

Its means and variances are computed as follows.
If $a_k = 0$, then $\hat{X}_{k|k}^i = \hat{X}_{k|k-1}^i$ and $P_{k|k}^i = P_{k|k-1}^i$.
If $a_k = 1$, then

$$\hat{X}_{k|k}^i = \hat{X}_{k|k-1}^{i'} + G_k^{i'} [Z_k^{l,j} - h^j(\hat{X}_{k|k-1}^{i'})], \tag{3.24}$$

$$H_k^i = \frac{\partial h^j(X)}{\partial X}|_{X=\hat{X}_{k|k-1}^{i'}}, \tag{3.25}$$

$$P_{k|k}^i = P_{k|k-1}^{i'} - G_k^{i'} H_k^{i'} P_{k|k-1}^{i'}, \tag{3.26}$$

$$S_k^i = H_k^{i'} P_{k|k-1}^{i'} H_k^{i'T} + R_k^{i',j}, \tag{3.27}$$

$$G_k^i = P_{k|k-1}^{i'} H_k^{i'T} (S_k^{i'})^{-1}. \tag{3.28}$$

3.2.3 Hypothesis pruning

From (3.20), we can realize that the number of hypotheses in the extended MKF increases over time, which will consume a lot of memory and computational time. To deal with this problem, when the number of hypotheses is larger than $N_{\max}$ we simply keep the $N_{\max}$ hypotheses with the highest weights and prune the other hypotheses which have lower weights.

3.2.4 Experimental evaluation

We conducted numerical experiments to evaluate our extended MKF algorithm for tracking one intermittent and moving speech source with room reverberation and noise.

Due to the statistical nature of false measurements, a large number of experiments is needed to obtain statistically meaningful results. Such a large number of experiments can hardly be conducted with a real robot. Therefore, we resort to simulation of the robot movements, the source movements, and the resulting location and activity measurements.

Our experimental settings mimic the *smart room* at Inria Nancy, where the robot is a Turtlebot equipped with a Kinect sensor as in Fig. 3.2. In this work, we are only using the linear array of 4 microphones included in the Kinect. Fig. 3.3 shows the geospatial distribution of 4 microphones in the Kinect with 3 microphones on the left, and 1 microphone on the right.

Figure 3.2: Turtlebot equipped with a Kinect sensor (http://www.turtlebot.com).

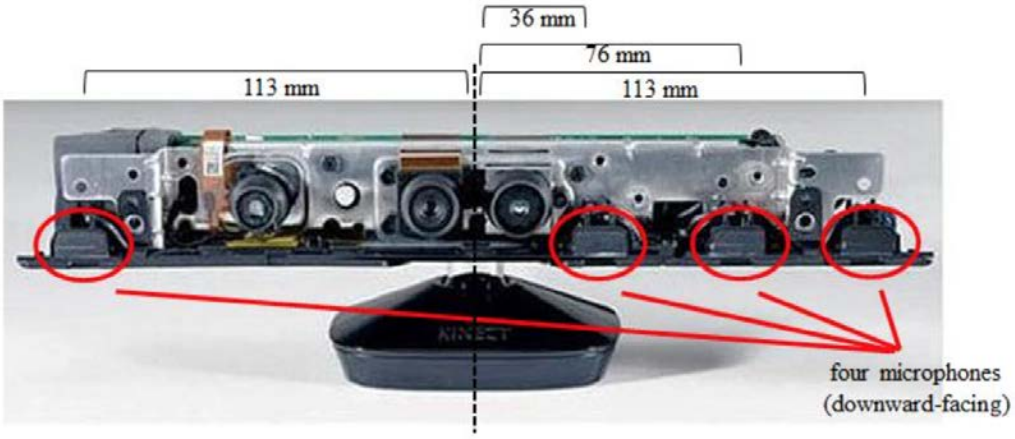

Figure 3.3: Linear array of 4 microphones inside the Kinect sensor (https://www.microsoft.com).

3.2.4.1 Data

We employed state-of-the-art techniques for the simulation of reverberation and acoustic noise, whose parameters are fixed as in [Vincent et al., 2015] and closely match the real conditions in that room. More specifically, the reverberation time (250 ms), the intensity of speech and noise, and the noise spectrum match those of the real environment.

The source AoA was estimated by MUSIC-GSVD. The probability distribution of AoAs estimated by MUSIC-GSVD for each of 360 true AoAs (from 0° to 359°) and 5 distances (from 0.5 to 3 m) was constructed and used to simulate the observed source location. We considered an SAD error rate of 5% and it could be a false negative or a false positive.

The target source is a speech. We considered four different scenarios, depending whether the sound source is static or mobile ($v_s = 0.07\,\mathrm{m/s}$, $w_s = 8°/\mathrm{s}$) and inactive for several short time intervals (0.5 s) or a long time interval (2 s). For each scenario, we randomly generated 100 source trajectories for a duration of 10 s. The robot trajectory was fixed in all experiments with a maximum speed of 0.38 m/s.

3.2.4.2 Algorithm settings

We set the parameter values of the extended MKF as follows.
The time step was $dt = 0.1\,\text{s}$.
The covariance matrix Q was set as

$$Q = \text{diag}(0, 0, 0, 0.00095\,\text{m}^2, 0.00062\,\text{m}^2, (6.2°)^2, 0, 0). \tag{3.29}$$

The initial position of the robot and the initial covariance of the source estimation can be visualized in Fig. 3.4. The variance $R^{i,j}$ varied as a function of the source distance between $(0.8°)^2$ at 0.3 m and $(4.5°)^2$ at 3 m.
The source appearance/disappearance probabilities were set to $P_{\text{appear}} = 0.5$ and $P_{\text{disappear}} = 0.2$. We set the number of hypotheses in the extended MKF to $N_{\text{max}} = 50$. We evaluate the performance of our extended MKF method with vs. without tracking of the source activity presented in Chapter 2. The extended MKF method without tracking of the source activity is also similar to Portello et al.'s [Portello et al., 2014][1].

3.2.4.3 Example run - Visualization

Fig. 3.4 shows the first few seconds of tracking one intermittent, moving source. At time $t = 0\,\text{s}$, the mixture is initialized with several components evenly distributed over the room in order to approximate a uniform prior. After 1 s, half of the hypotheses for the source position are distributed along the direction from the source to the robot and the rest are symmetric w.r.t. the microphone axis. This symmetrical uncertainty is due to the front-back confusion phenomenon illustrated in Fig. 3.1. These symmetrical hypotheses become smaller and disappear after 3 s, thanks to the robot motion. More precisely, the motion of the source for the symmetric components is bigger, less coherent and therefore less probable than of the correct components. For a nonlinear microphone array, the transitory phase with two directions would be shorter or perhaps nonexistent.

3.2.4.4 Example run – Estimated trajectories

Fig. 3.5 compares the source trajectory with the estimations of our extended MKF and the extended MKF without activity model. As both models are mixture models, the posterior distribution is usually not unimodal. In order to generate a single point estimate, we simply compute the mean of the distribution. After the first few seconds discussed above, both trajectories follow the sound source. We can also observe that there are some moments when the estimated source location of the extended MKF without activity model is far away from the actual source location, however our extended MKF still can track the source location with lower estimation error.

Fig. 3.6 shows the estimation error, that is the distance between the estimated source position and the true position. During the first 3 s, both extended MKFs have high estimation error due to the front-back ambiguity. Between 3.5 and 3.9 s, the source is inactive and the SAD is correct. During this period, the uncertainty about the source position increases because of the lack of measurements. The uncertainty gets smaller when the source becomes active again.

At time $t = 2.8\,\text{s}$, a false measurement of the source activity occurs: the source is active but SAD detects it as inactive. The estimation error of the MKF without activity model becomes

[1] The difference with their method lies in the number of Gaussians used to approximate the observation likelihood.

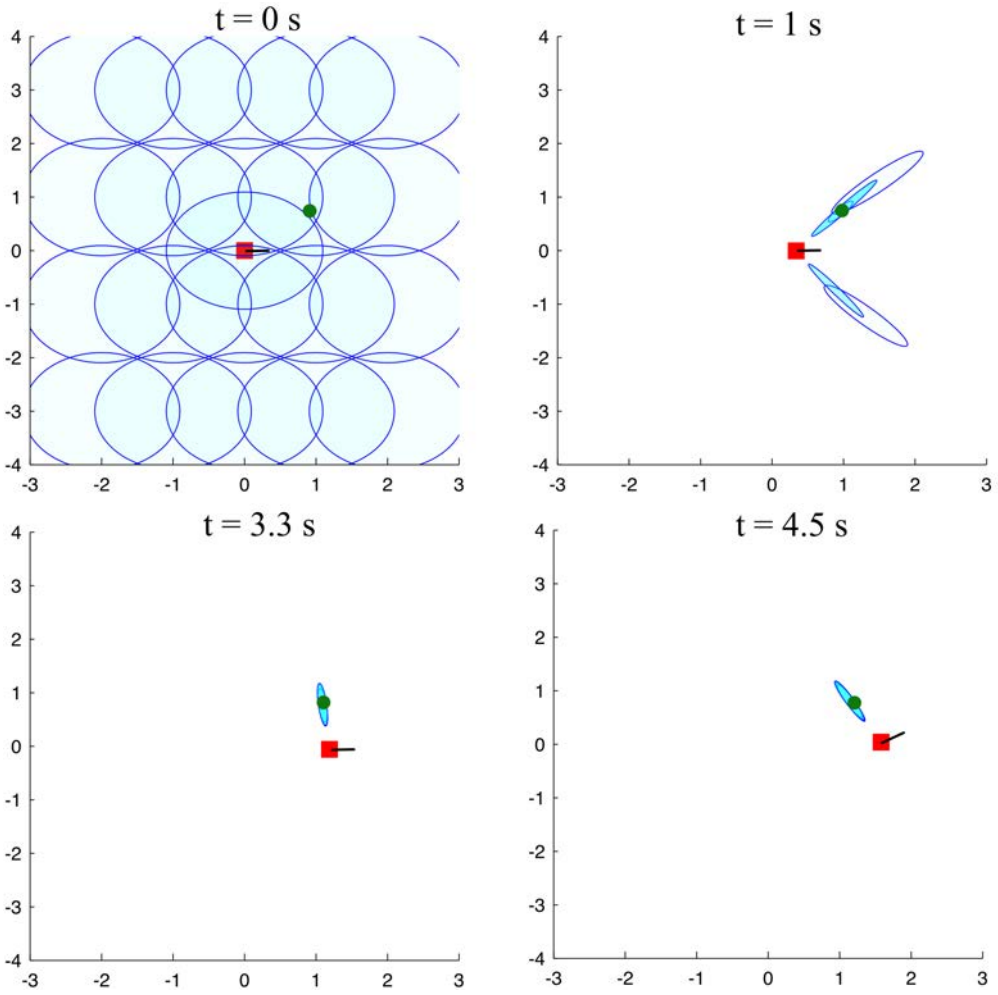

Figure 3.4: Visualization of our extended MKF in an example run. Robot positions are shown as red squares, and the actual source position as a green circle. Blue ellipses represent 95% confidence regions of source location estimation of various hypotheses in the mixture with a transparency proportional to the weight of the components.

larger than ours but only for one time step. Conversely, from $t = 3.2\,\mathrm{s}$ to $7\,\mathrm{s}$, the estimation error of the MKF without activity model is lower than ours but by $2\,\mathrm{cm}$ only.

A false measurement of AoA occurs at time $t = 4.8\,\mathrm{s}$. The AoA difference between the observation and the ground truth is 9°, which is not a large value. As a result, both the estimated error of our extended MKF and the extended MKF without activity model increase slightly.

At time $t = 7\,\mathrm{s}$, a false AoA measurement occurs: the ground truth AoA is 81°, but the measured AoA is 62°. Although such a false measurement can occur with very low probability, it can have a major impact. Indeed, the estimation error of the extended MKF without activity model increases drastically and remains large. By contrast, the estimation error of our extended MKF does not change much. This is an unexpected benefit of the proposed approach: when a false AoA measurement occurs, the weight of the hypotheses corresponding to an inactive source increases, so that the belief is little affected.

At time $t = 8.2\,\mathrm{s}$, a false measurement of the source activity occurs: the source is inactive but

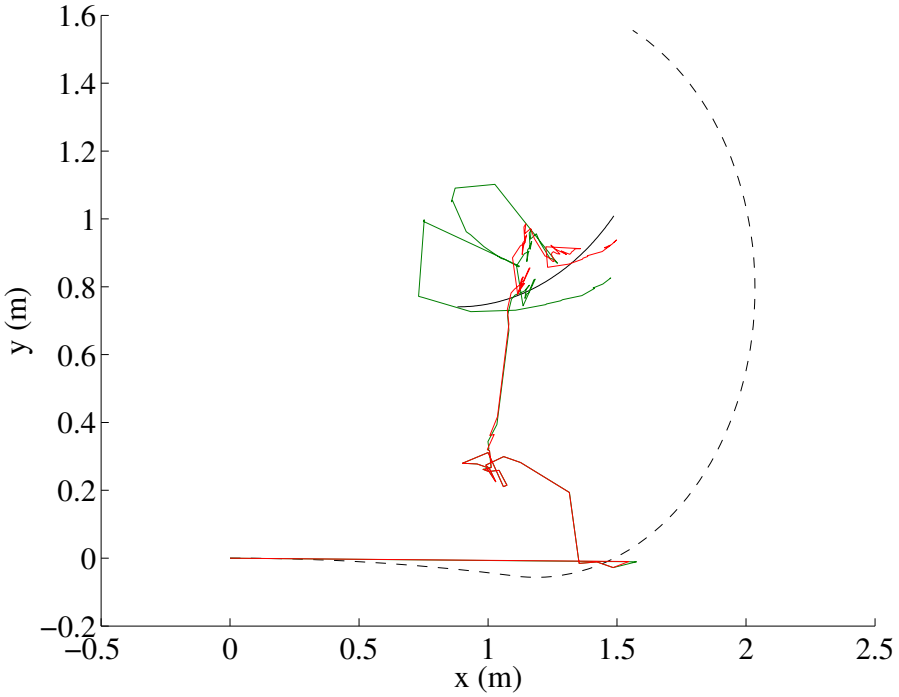

Figure 3.5: Estimated trajectories of the two extended MKF methods in the example run. Dashed black line: trajectory of the robot, Solid black line: ground truth trajectory of the sound source, Green line: estimate of the extended MKF without activity model, Red line: estimate of our extended MKF with activity model.

SAD detects it as active. Again, it appears that our extended MKF can handle this situation but the extended MKF without activity model severely suffers from it. This was expected, since the tracking of the source activity implemented by our method was designed precisely to address this issue.

3.2.4.5 Error rate of source location estimation

To evaluate the robustness of our proposed extended MKF framework to false measurements in the SAD, we change the error rate in the SAD observations from 0% to 10% and we run 100 experiments for each SAD error rate value as stated above. The parameter values in the extended MKF are still the same as in 3.2.4.2.

Fig. 3.7 shows the median of the estimation error on the source location as a function of the SAD error rate. The median value is around 0.45 m and does not change so much when the SAD error rate varies. This verifies that our proposed extended MKF framework is robust to false SAD measurements.

To see how the error rate set in the extended MKF model can affect the performance of the source localization, we tune the error rate in the extended MKF model between 0% and 10% and fix the error rate in the SAD observations to 5%. Similarly, we conduct 100 experiments for each error rate value.

Fig. 3.8 shows the median of the estimation error in the source location as a function of the error rate in the extended MKF model. As expected, the estimation error is much higher when the error rate in the extended MKF model is equal to 0%: this corresponds to the extended

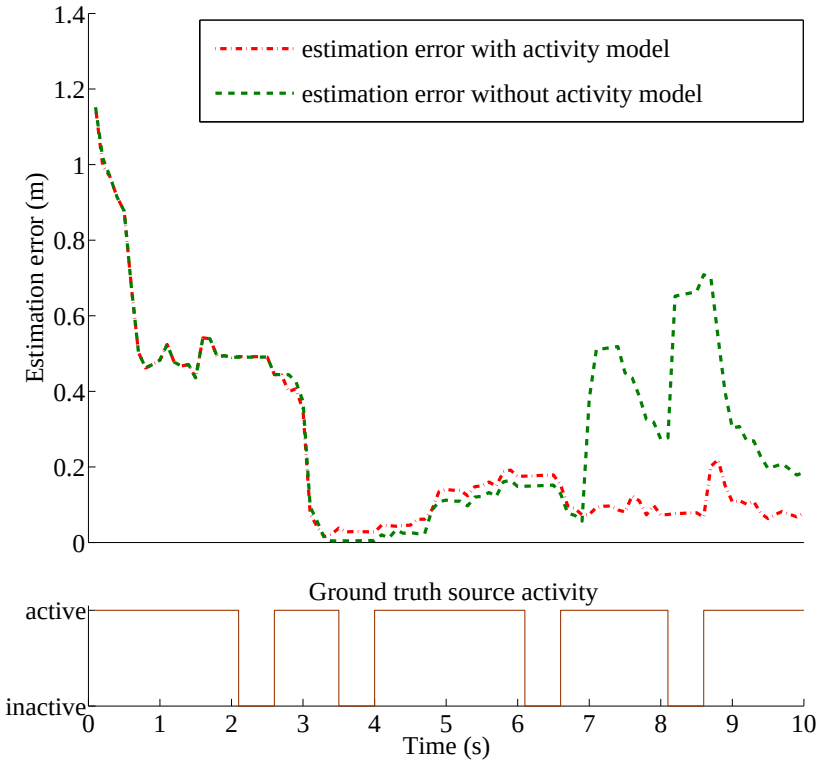

Figure 3.6: Top: Estimation error over time of our extended MKF with activity model vs the extended MKF without activity model. Bottom: Ground truth source activity over time.

MKF without activity model. When the error rate in the extended MKF model is equal or greater than 1%, the estimation error does not change much and is around 0.45 m.

3.2.4.6 Error rate of source activity estimation

Similar to the previous evaluation on the estimation of source location, in this section, we evaluate the impact of SAD error rate on the activity estimation.

At each time step, we compute the estimated source activity by summing the weights of all components for which source is active. It is a real value between 0 and 1, and we can call it as the estimated probability of being active of the source. The probability of incorrectly estimating the source activity is the value difference between the estimated source activity value and the ground truth source activity.

Fig. 3.9 shows the average probability of incorrectly estimating the source activity over all experiments when we change the error rate of the SAD in the observation from 0% to 10%. We can see that this probability does not change so much as the SAD error rate increases.

The probability of incorrectly estimating the source activity when we change the SAD error rate in the extended MKF model from 0% to 10% but keep the SAD error rate in the observation at 5% is presented in Fig. 3.10. This probability is steady when the error rate in the extended MKF model is from 1% to 10%. However, it is smaller when we set the error rate in the extended MKF model at 0%. The reason is when we set the error rate in the extended MKF greater than 0%, we have weights on the hypothesis that the source activity is not the same as in the SAD. When the false AoA measurement happens, these weights will increase, as the result, the estimation of the source location is not affected but the estimation of the source activity.

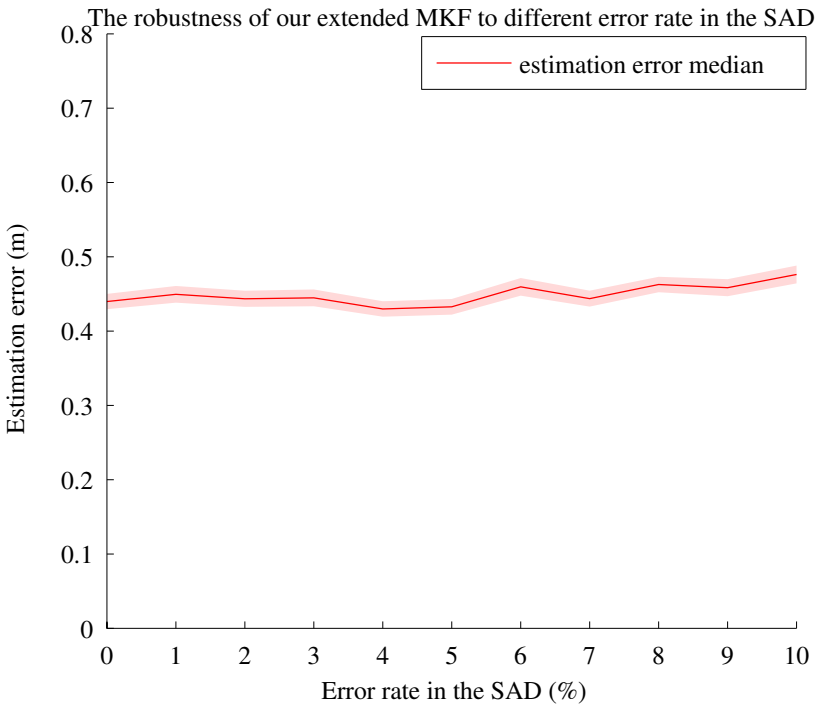

Figure 3.7: Median and 95% confidence interval of the estimation error on the source location as a function of the error rate in the SAD observations.

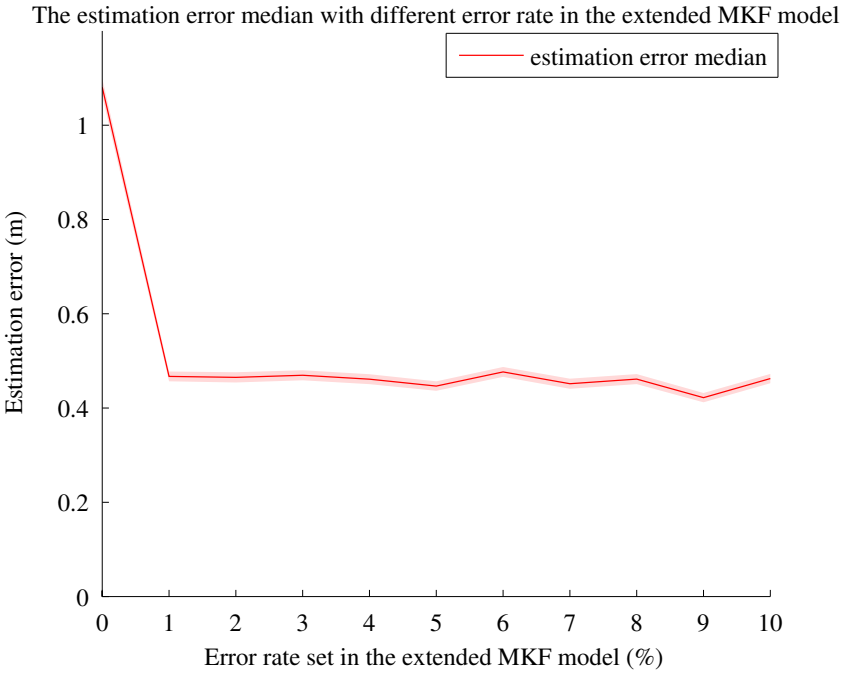

Figure 3.8: Median and 95% confidence interval of the estimation error on the source location as a function of the error rate in the extended MKF model.

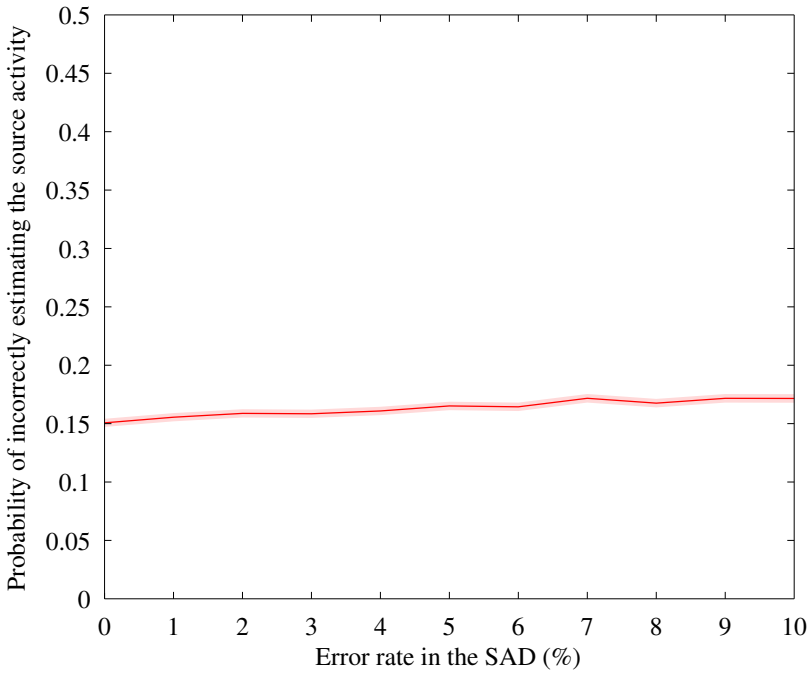

Figure 3.9: The average probability of incorrectly estimating the source activity over all experiments and 95% confidence interval as a function of the error rate in the SAD observation.

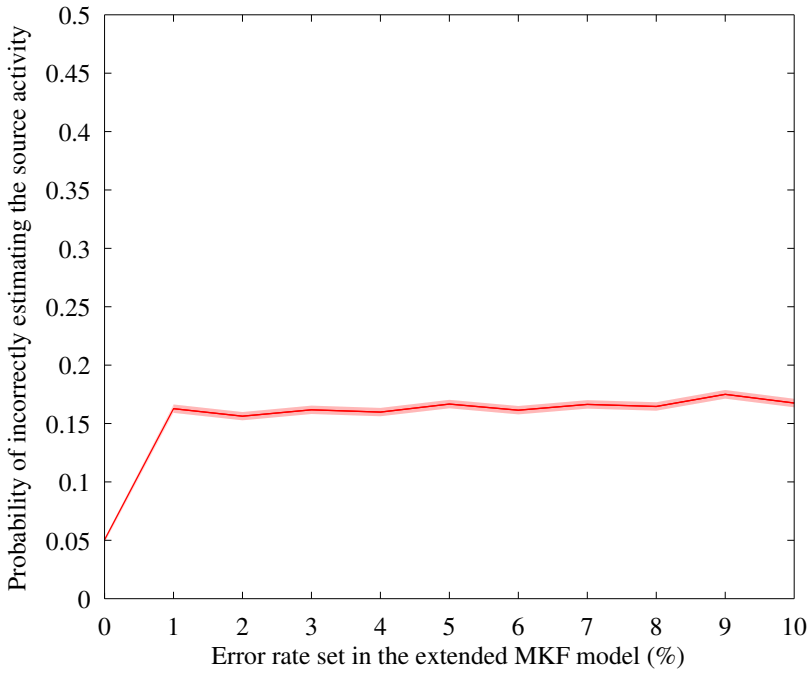

Figure 3.10: The average probability of incorrectly estimating the source activity over all experiments and 95% confidence interval as a function of the error rate in the extended MKF.

3.2.4.7 Statistical analysis

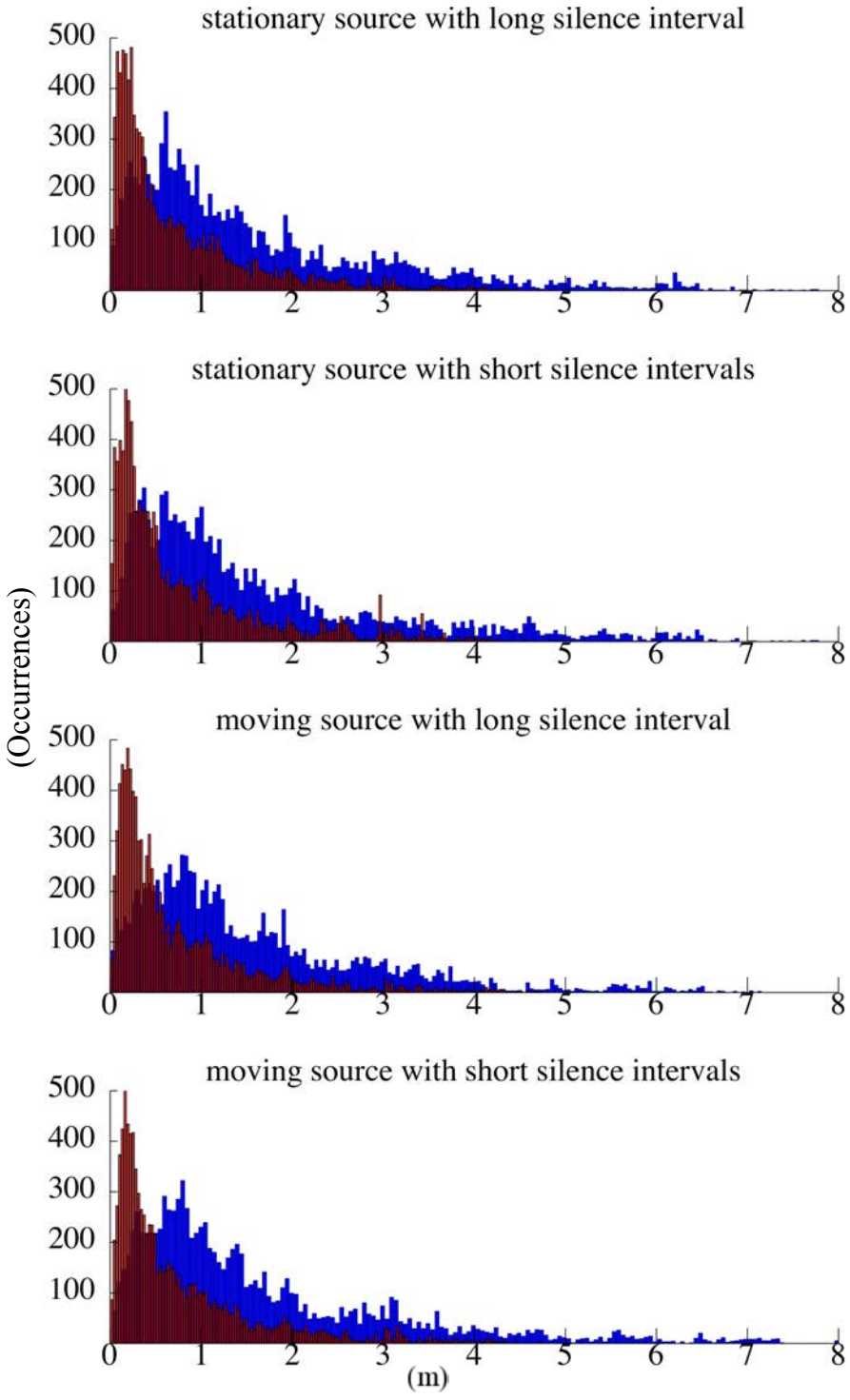

Figure 3.11: Estimation error distribution of our extended MKF method in red and of the extended MKF without activity model in blue.

In order to conduct a statistical comparison of our extended MKF method and the extended

MKF without activity model, we compare the distributions of errors between the estimate and the ground truth. These distributions, shown in Fig. 3.11, summarize the error for all time steps and all experimental runs. We can see that the error for our extended MKF is mostly concentrated below 1 m while the extended MKF without activity model shows a bigger error distribution with a heavy tail. This is confirmed with a Wilcoxon signed-rank test that assesses that our extended MKF has a significantly smaller error than the extended MKF without activity model with $p < 0.01$. Interestingly, our extended MKF outperforms the extended MKF without activity model in all four scenarios, whether the source is static or moving and includes short or long silences.

3.3 Particle filtering

In our contribution with the particle filter, we represent the posterior density by a set of samples of the state space (X^i, a^i) (particles) with corresponding weights w^i, $i = 1, ..., N$. Different from the particle filter presented in Chapter 2, we have the source activity in these particles. All these particles will be propagated through a recursive particle filtering process to estimate the belief on the source location. The three main steps of the recursive filtering process are presented in the following.

3.3.1 Prediction step

The belief $P(X_{k-1}, a_{k-1}|Z_{1:k-1})$ at time step $k-1$ can be expressed by the set of particles and weights $\{X^i_{k-1}, a^i_{k-1}, w^i_{k-1}\}^N_{i=1}$. These particles are relocated in the state space according to the transition equation (3.11) to obtain particles (X^i_k, a^i_k) from the previous belief $P(X_{k-1}, a_{k-1}|Z_{1:k-1})$. Each state sample X^i_{k-1} of the particle follows the transition probability $P(X^i_k|X^i_{k-1})$ which is computed from the dynamical model of the robot (3.2) and the source (3.4)

$$X^i_k = f(X^i_{k-1}) + d^i_k, \tag{3.30}$$

where d is random noise.

Each source activity state sample a^i_{k-1} of the particle follows the source activity transition probability $P(a^i_k|a^i_{k-1})$ which is defined by the source appearance probability $P_{\text{appear}} = P(a_k = 1|a_{k-1} = 0)$ and by the source disappearance probability $P_{\text{disappear}} = P(a_k = 0|a_{k-1} = 1)$.

3.3.2 Update step

In the update step, we observe a new measurement Z_k consisting of one AoA measurement Z^{l}_k and one source activity measurement Z^{a}_k. Given this new measurement Z_k, we update the weights of all particles according to the likelihood of each prior sample

$$w^i_k = \frac{w^i_{k-1} P(Z_k|X^i_k, a^i_k)}{\sum_j w^j_{k-1} P(Z_k|X^j_k, a^j_k)} \tag{3.31}$$

In more detail, if the particle corresponds to an active source, its weight can be updated as

$$w^i_k = \eta w^i_{k-1} P_{\text{sn}}(Z^{\text{l}}_k|X^i_k) P(Z^{\text{a}}_k|a^i_k) \tag{3.32}$$

if the particle corresponds to an inactive source, its weight can be updated as

$$w^i_k = \eta w^i_{k-1} P_{\text{n}}(Z^{\text{l}}_k) P(Z^{\text{a}}_k|a^i_k) \tag{3.33}$$

3.3.3 Particle resampling step

To reduce the effect of degeneracy, we resample the particles whenever a significant degeneracy is observed (i.e., when the effective sample size N_{eff} which defined in Section 2.3.5 falls below a threshold N_{T}). The particle resampling step involves generating a new set of particles $\{X_k^{i*}, a_k^{i*}\}_{i=1}^N$ from the set $\{X_k^i, a_k^i\}_{i=1}^N$ according to the weights $\{w_k^i\}_{i=1}^N$.

We use residual resampling for improving the run time, and ensures that the sampling is uniform across the population of particles. After multiplying the normalized weights by N, the integer value of each weight is used to define how many samples of that particle will be taken. This ensures that all higher weight particles are chosen at least once, and has O(N) running time. To select the remained particles, we compute the residual: the weights minus the integer part, which leaves the fractional part of the number. Based on the residual, we can use a simpler sampling scheme such as multinomial to select the rest of the particles.

These new particles after resampling are approximately distributed as the posterior density $P(X_k, a_k|Z_{1:k})$. After the particle resampling step, the weights of the particles are now reset to $w_k^i = \frac{1}{N}$.

3.3.4 Example run

We show an example run of the above particle filtering method to estimate the source location and its activity over time in Fig. 3.12. In this example, we set the source appearance/disappearance probabilities, the inactive intervals of the source and the source velocity as in the example run of the extended MKF filter. We use 800 particles to estimate the state of the source.

At time $t = 0\,\text{s}$, we initialize with 800 particles evenly distributed over the room to approximate a uniform prior. Some particles represent an active source (magenta particles) and some represent an inactive sources (black particles). At time $t = 0.4\,\text{s}$, due to the front-back ambiguity, the particles are distributed symmetrically w.r.t. the microphone axis with half of them along the direction from the source to the robot. After $t = 2\,\text{s}$, the front-back ambiguity disappears thanks to the robot motion. At time $t = 2.1\,\text{s}$, more black particles appear which show that the source is estimated as inactive. At time $t = 5.8\,\text{s}$, all the particles are close to the actual source position and in magenta. This means that the source is estimated as active at that time.

Fig. 3.13 shows the estimation error and the ground truth source activity over time for the above example run. We compute the estimated source position by the weighted sum of all particle positions. The estimation error decreases drastically after 1 s, because the front-back ambiguity starts to disappear. When the ambiguity has totally disappeared after 1.2 s, the estimation error is more steady and slightly decreases. There is a small rise at time $t = 4\,\text{s}$ due to a false AoA measurement happening at the same time when the source changes from inactive to active. However, this has only a minor impact which is limited in time.

Due to the similarity between this particle filtering method and the proposed extended MKF method, we don't analyze its robustness to the SAD error rate and the error rate in the model as in Section 3.2.4.5 and 3.2.4.6. Instead of that, in the next section, we compare the performance of the proposed extended MKF with this particle filtering method in a large number of experiments.

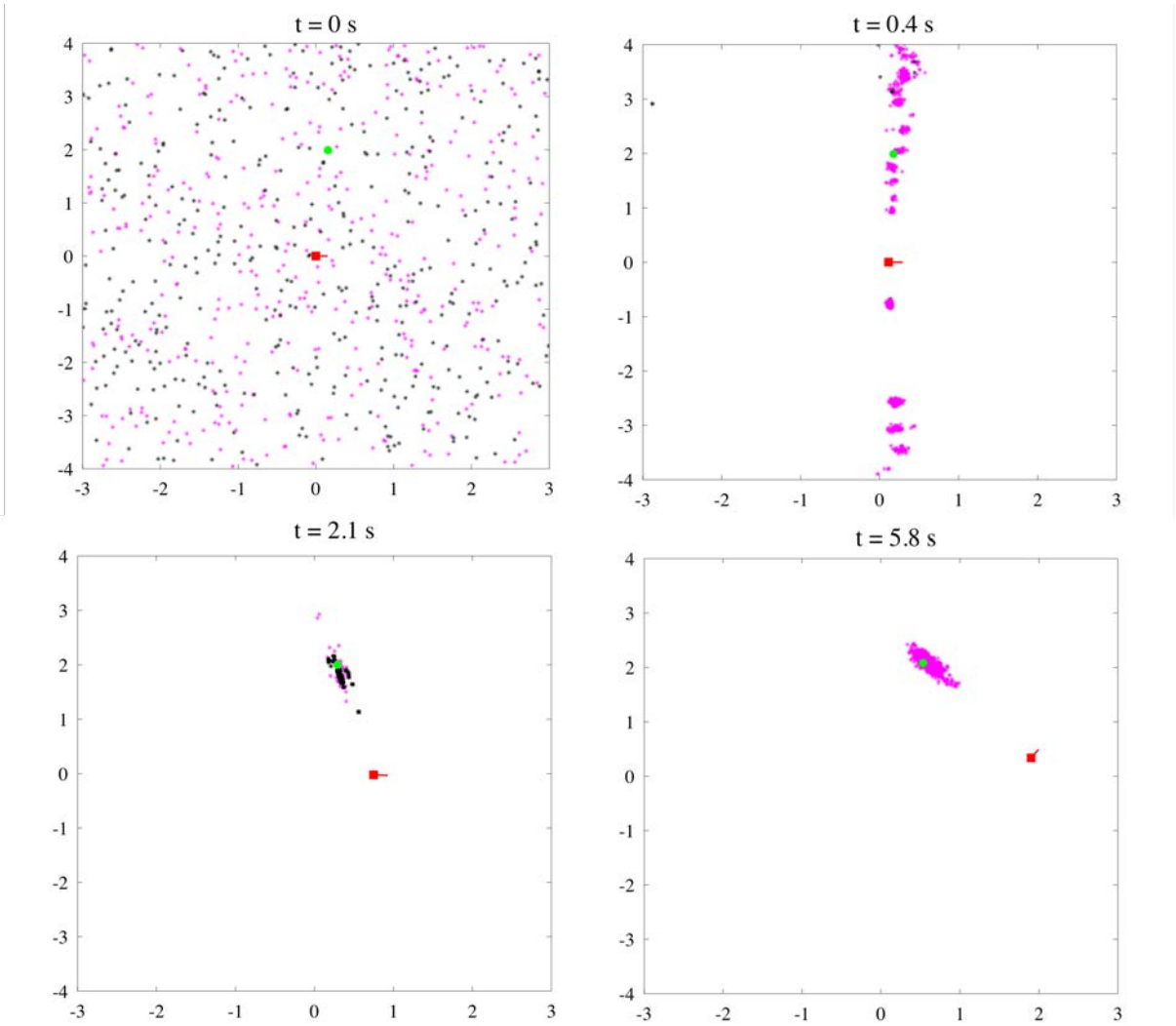

Figure 3.12: An example run of the particle filtering algorithm to estimate the source location and its activity. Robot positions are shown as red squares, and the actual source position as a green circle. The particles in magenta represent the hypotheses corresponding to active sources. The particles in black represent the hypotheses corresponding to inactive sources

3.4 Comparison of the extended MKF with the particle filtering

In this section, we compare the extended MKF with the particle filter for tracking an intermittent, moving source in a large number of experiments. We compare the performance of the two proposed filtering methods in term of estimation error, computational time and number of components or particles used in the estimation.

3.4.1 Data

For both approaches, we use the same data as in Section 3.2. The source AoA is estimated by MUSIC-GSVD and the SAD error rate is set to 5%. In all experiments, the robot trajectory was fixed.

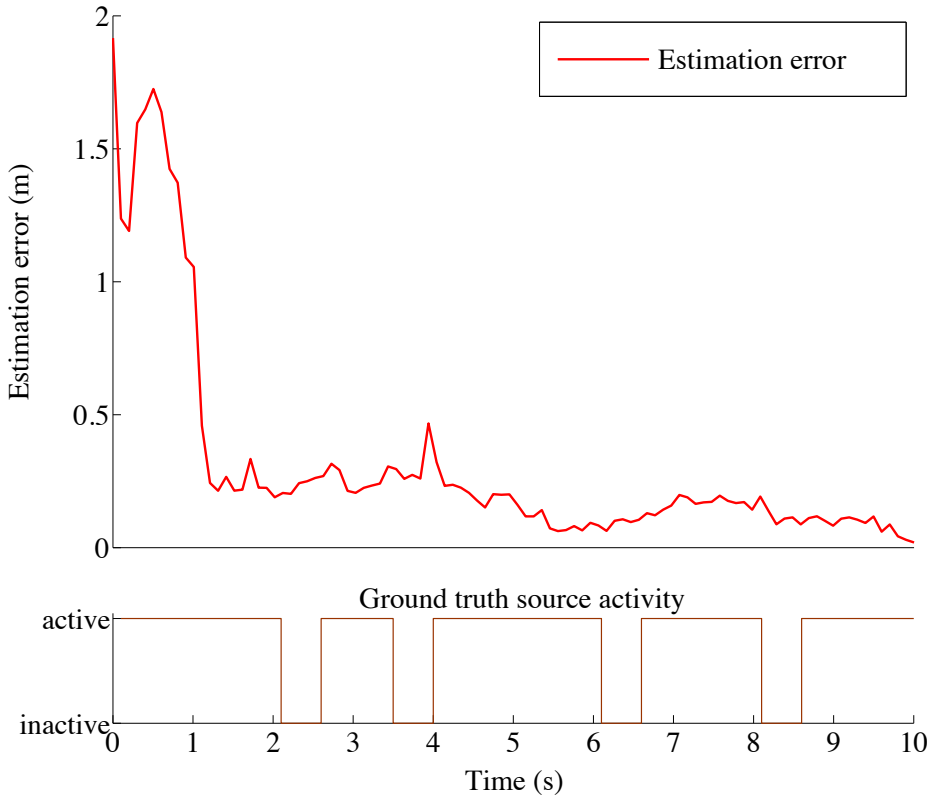

Figure 3.13: Estimation error of the particle filtering algorithm and the ground truth source activity over time in the example run.

3.4.2 Algorithm settings

The settings for the extended MKF method are the same as in the Section 3.2. We set the maximum number of components $N_{\max}$ in the extended MKF to one of the following: 2, 5, 10, 20, 50, or 100 for each set of 100 experiments. After each iteration, we only keep the $N_{\max}$ components with highest weights in the extended MKF, and the other components are pruned.

For the particle filtering method, the number of particles N for each set of 100 experiments is set to: 100, 300, 600, 800, 1000, or 1200. In the particle filter, the number of particles will not change after each iteration.

3.4.3 Experimental results

Fig. 3.14 depicts the average estimation error over time and the 95% confidence interval achieved by the particle filtering algorithm with different numbers of particles. The particle filter with 100 particles does not yield good estimation: the average estimation error is around 1.4 m and does not decrease over time. With 300 particles and more, the estimation is significantly improved. For the particle filter with 300 particles, the average estimation error decreases steadily until time $t = 3.5$ s. After that, its average estimation error reaches a stable value around 0.5 m. The particle filters with 600 to 1200 particles have significantly lower estimation error after 2.5 s compared to the particle filter with 300 particles. However, they are not significantly different from each other. After 3.5 s, the average estimation error of these four particle filters becomes steady around 0.2 m. At the end, at $t = 2$ s, the minimum average estimation error is 0.14 m for

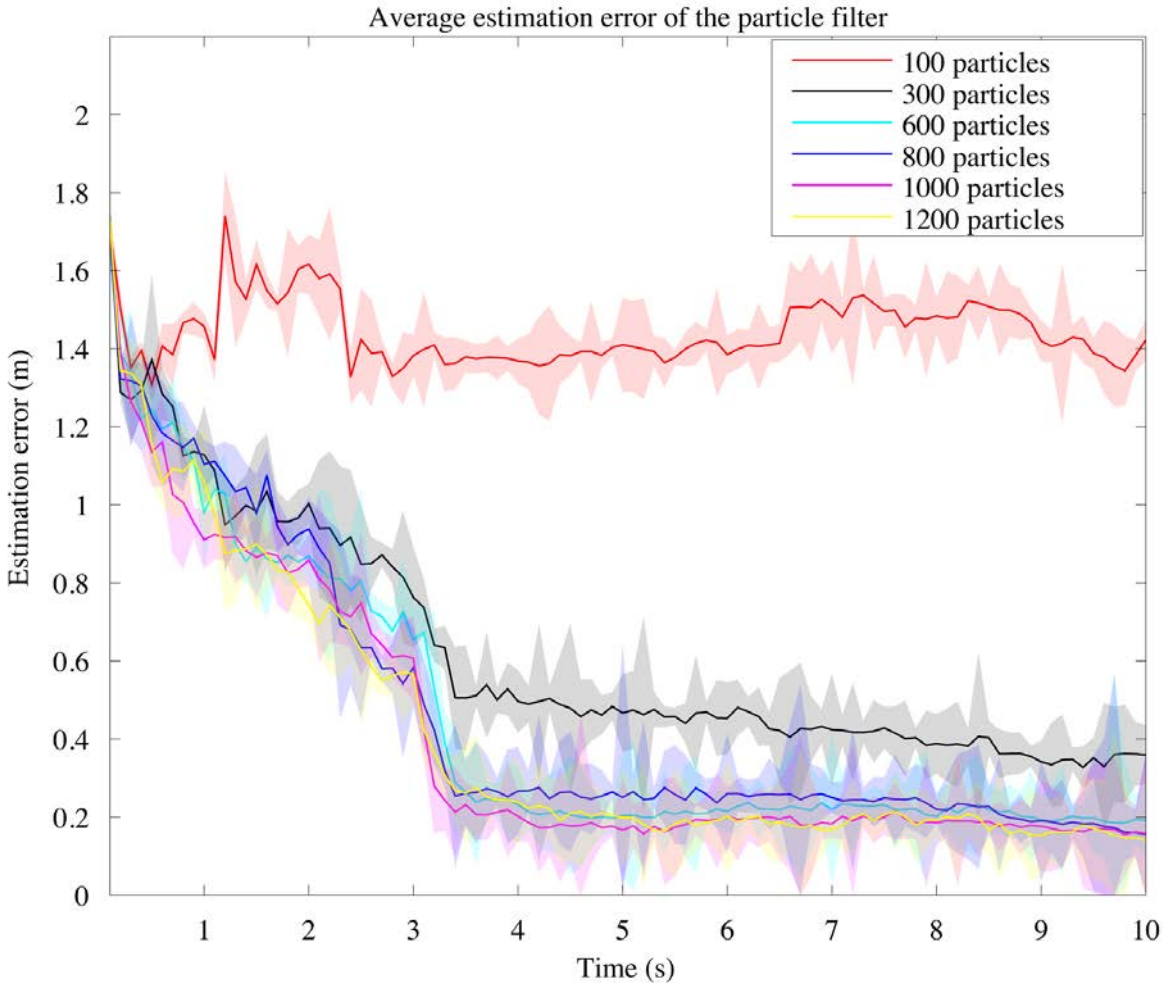

Figure 3.14: Average estimation error and 95% confidence interval of the particle filtering algorithm with different number of particles N.

1200 particles, 0.15 m for 800 and 1000 particles and is 0.20 m for 600 particles. We can see that the particle filter with 600 to 800 particles could be good enough for source localization.

Fig. 3.15 shows the average estimation error over time and the 95% confidence interval achieved by the extended MKF algorithms with different numbers of components. In general, the average estimation error of the extended MKF decreases over time and is lower when the maximum number of components increases. With 2 to 10 components in the extended MKF, although the average estimation error decreases over time it remains significantly higher than for the extended MKF with 20 components. The source estimation is significantly improved when the maximum number of components is equal or greater than 50. From the beginning until time $t = 3.5$ s, the corresponding average estimation error decrease steadily. After that, it is stable around 0.25 m and slowly decreases to 0.22 m at the end. We can see that for the extended MKF, 50 components are good enough for estimation, i.e, performance does not improve further with more components.

From these two figures, the minimum average estimation error value at the final time step is

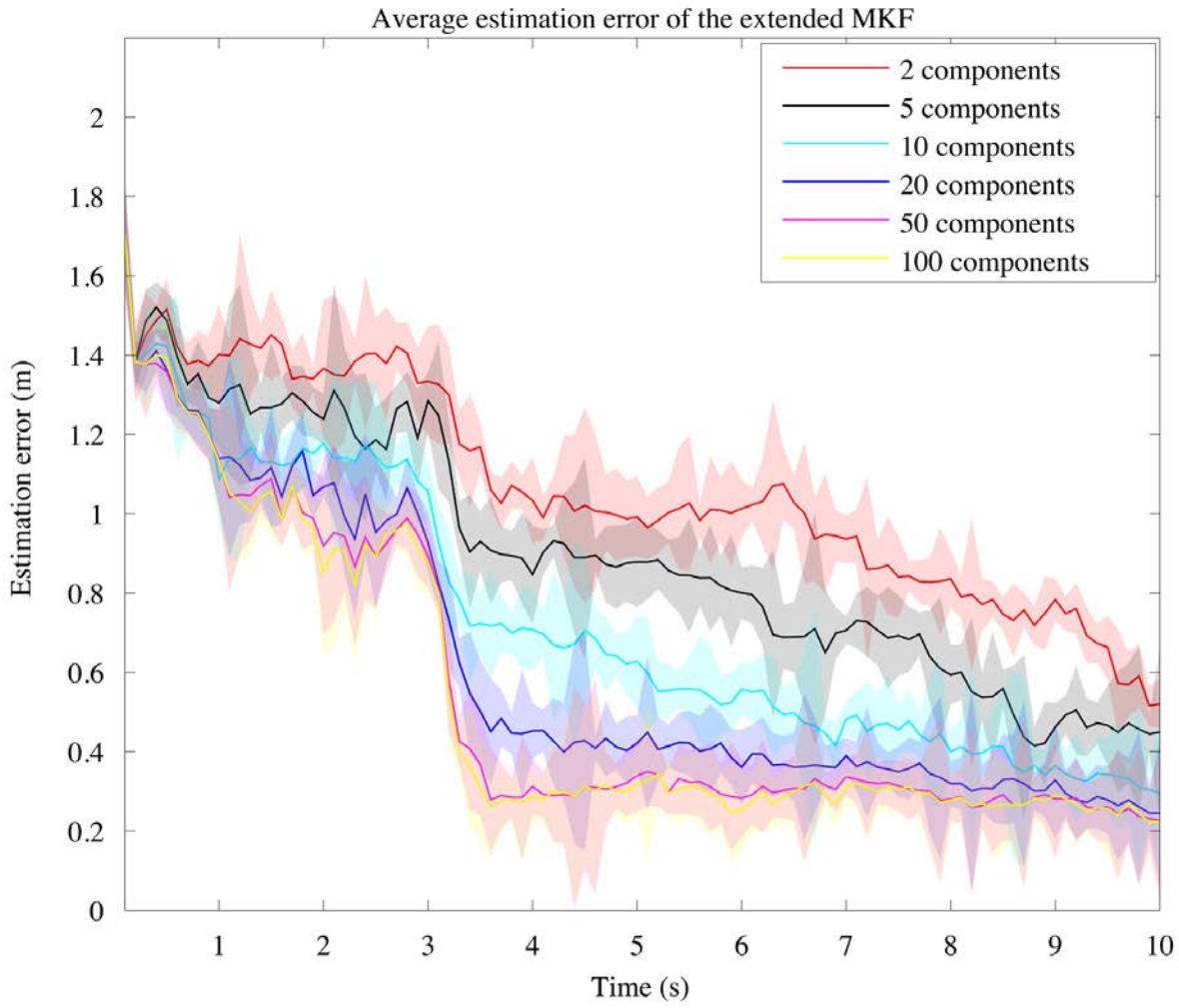

Figure 3.15: Average estimation error and 95% confidence interval of the extended MKF algorithm with different maximum number of components $N_{\max}$.

slightly smaller for the particle filter than for the extended MKF, however it is not significantly so.

In addition to comparing in the estimation error, we also compare the computational time of the two filtering methods. We implement these methods in Matlab R2012a version on a 2.50GHz CPU, no multithreading. The average time required for one iteration of the particle filtering method with 600 particles is 5.8 s. However, the computational time required for one iteration of the extended MKF with 50 components is only 1.7 s.

3.5 Summary

We proposed a Bayesian filtering framework that can be applied to any microphone array geometry for tracking one intermittent, possibly moving sound source in a reverberant environment using a mobile robot. The main theoretical contribution of the method is the explicit estimation

of the source activity, which allows us to cope with false SAD and/or AoA measurements.

We presented an extended MKF and a particle filtering method for jointly estimating the source location and its activity. Experimental results with the extended MKF have demonstrated the significantly better performance of our algorithm compared to recent work in the literature. They also showed the additional advantage of our algorithm in tracking the location of the source in the presence of false AoA measurements.

We conducted a number of experiments to compare the extended MKF and the particle filtering method in terms of estimation error and computational time with different numbers of components or particles. Both methods obtained good estimation with a proper number of components in the filter: 50 components for the extended MKF and 600 to 800 particles for the particle filter.

Chapter 4

Multiple source localization

In this chapter, we present an extension of our previous work on the localization of a single source to the context of multiple sources using a mobile robot. The AoA measurements would have much lower accuracy for three or more sources, for that reason, in this chapter we will only address two sources.

We propose an extended MKF with the JPDAF for dealing with the association problem and jointly estimating the location of the two sources and their activities over time. We first introduce a technique to build a sensor model for the case of two sound sources by simulating two sources with background noise and reverberation and learn the probability of correctly observing each AoA measurement. After that, the extended MKF with the JPDAF for localizing two sources is presented. The experimental evaluation shows the ability of the proposed framework to handle uncertainty in the observations when localizing and tracking two intermittent, moving sources in a noisy, reverberant environment.

4.1 Learning the sensor model for multiple source localization

The sensor model exposed in Chapter 3 was built using a simulation of a source in a reverberant room and varying the parameters of the source. Coupled with a model of noise we could describe the cases when a single was either active or inactive. With a second source, we need to update the sensor model to include the case when two sources are simultaneously active. We train the sensor model by simulating two active sound sources with background noise and a microphone array (Kinect sensor) in a closed room area.

For each of 360 true angles (from 0° to 359°) and 5 distances (from 0.5 to 3 m) from the first source to the robot, we generate randomly the angle of the second source and its distance so that the ratio of distances between the two sources and the microphone array is between 1.2 and 2. For each generated signal, we obtain the MUSIC spectrum estimated by MUSIC-GSVD over 10 STFT frames corresponding to 100 ms. We compute 5 disjoint MUSIC spectra for each signal. From each MUSIC spectrum, we detect the two highest peaks. Based on those, from t training samples we learn the sensor model which is expressed by the probability distribution $P(\hat{\phi}_1, \hat{\phi}_2 | \alpha_1, \alpha_2, \phi_1, \phi_2, d_1, d_2)$ where $\hat{\phi}_1, \hat{\phi}_2$ are the two peaks, α_1, α_2 are the probabilities of correctly detecting peak $\hat{\phi}_1$ and peak $\hat{\phi}_2$, ϕ_1, ϕ_2 are the true AoAs, and d_1, d_2 are the true distances from each source to the microphones.

In order to simplify the inference we want to decouple both sources. We assume that the probability of detecting the peak $\hat{\phi}_1$ does not depend on the probability of detecting the peak $\hat{\phi}_2$ as well as the true AoA ϕ_2 and d_2, α_1, α_2. The same goes for the probability of detecting the

peak $\hat{\phi}_2$. If all the peaks were correctly observed in the right order and there were no missing peak we would have:

$$\begin{aligned} P(\hat{\phi}_1, \hat{\phi}_2|\alpha_1, \alpha_2, \phi_1, \phi_2, d_1, d_2) =& P(\hat{\phi}_1|\alpha_1, \alpha_2, \phi_1, \phi_2, d_1, d_2) P(\hat{\phi}_2|\hat{\phi}_1, \alpha_1, \alpha_2, \phi_1, \phi_2, d_1, d_2) \\ =& P_{\text{sn}}(\hat{\phi}_1|\phi_1, d_1) P_{\text{sn}}(\hat{\phi}_2|\phi_2, d_2). \end{aligned} \tag{4.1}$$

However, we know that it is not always the case. There is a situation that the peaks are correctly ordered but we do not know whether they are missing or not. It can happen that no peak is missing, or one of the two peaks is missing, or both peaks are missing. With the assumption that these events are independent from each other, we would have:

$$\begin{aligned} &P(\hat{\phi}_1, \hat{\phi}_2|\alpha_1, \alpha_2, \phi_1, \phi_2, d_1, d_2) \\ &\quad =\alpha_1\alpha_2 P_{\text{sn}}(\hat{\phi}_1|\phi_1, d_1) P_{\text{sn}}(\hat{\phi}_2|\phi_2, d_2) + (1-\alpha_1)\alpha_2 P_{\text{n}}(\hat{\phi}_1) P_{\text{sn}}(\hat{\phi}_2|\phi_2, d_2) \\ &\qquad + \alpha_1(1-\alpha_2) P_{\text{sn}}(\hat{\phi}_1|\phi_1, d_1) P_{\text{n}}(\hat{\phi}_2) + (1-\alpha_1)(1-\alpha_2) P_{\text{n}}(\hat{\phi}_1) P_{\text{n}}(\hat{\phi}_2) \\ &\quad =\Big[\alpha_1 P_{\text{sn}}(\hat{\phi}_1|\phi_1, d_1) + (1-\alpha_1) P_{\text{n}}(\hat{\phi}_1)\Big]\Big[\alpha_2 P_{\text{sn}}(\hat{\phi}_2|\phi_2, d_2) + (1-\alpha_2) P_{\text{n}}(\hat{\phi}_2)\Big]. \end{aligned} \tag{4.2}$$

However, in practice, we do not know whether the peaks are correctly ordered or not and whether there is a peak missing or not. Therefore, we can write:

$$\begin{aligned} &P(\hat{\phi}_1, \hat{\phi}_2|\alpha_1, \alpha_2, \phi_1, \phi_2, d_1, d_2) \\ &\quad =\frac{1}{2}\Big[\alpha_1 P_{\text{sn}}(\hat{\phi}_1|\phi_1, d_1) + (1-\alpha_1) P_{\text{n}}(\hat{\phi}_1)\Big]\Big[\alpha_2 P_{\text{sn}}(\hat{\phi}_2|\phi_2, d_2) + (1-\alpha_2) P_{\text{n}}(\hat{\phi}_2)\Big] \\ &\qquad + \frac{1}{2}\Big[\alpha_1 P_{\text{sn}}(\hat{\phi}_1|\phi_2, d_2) + (1-\alpha_1) P_{\text{n}}(\hat{\phi}_1)\Big]\Big[\alpha_2 P_{\text{sn}}(\hat{\phi}_2|\phi_1, d_1) + (1-\alpha_2) P_{\text{n}}(\hat{\phi}_2)\Big]. \end{aligned} \tag{4.3}$$

From this equation, we can learn $\alpha = [\alpha_1; \alpha_2]$ in the maximum likelihood sense by trying different values of α then computing the log of the probability $P(\hat{\phi}_1, \hat{\phi}_2|\alpha_1, \alpha_2, \phi_1, \phi_2, d_1, d_2)$ and maximizing the sum over t training samples:

$$\hat{\alpha} = \arg\max_{\alpha} \sum_{t} \log P(\hat{\phi}_1, \hat{\phi}_2|\alpha_1, \alpha_2, \phi_1, \phi_2, d_1, d_2). \tag{4.4}$$

That is the technique we should use to learn the values of α_1 and α_2. However, for time reasons, we did not have sufficient time to learn these values. Therefore, based on our experience, we set the value of α_1 to 1, and the value of α_2 to 0.8 in the rest of this chapter. With these value of α_1 and α_2, we can simplify and rewrite (4.3) as

$$\begin{aligned} &P(\hat{\phi}_1, \hat{\phi}_2|\alpha_1, \alpha_2, \phi_1, \phi_2, d_1, d_2) \\ &\quad =\frac{1}{2}\alpha_2 P_{\text{sn}}(\hat{\phi}_1|\phi_1, d_1) P_{\text{sn}}(\hat{\phi}_2|\phi_2, d_2) + \frac{1}{2}(1-\alpha_2) P_{\text{sn}}(\hat{\phi}_1|\phi_1, d_1) P_{\text{n}}(\hat{\phi}_2) \\ &\qquad + \frac{1}{2}\alpha_2 P_{\text{sn}}(\hat{\phi}_1|\phi_2, d_2) P_{\text{sn}}(\hat{\phi}_2|\phi_1, d_1) + \frac{1}{2}(1-\alpha_2) P_{\text{sn}}(\hat{\phi}_1|\phi_2, d_2) P_{\text{n}}(\hat{\phi}_2). \end{aligned} \tag{4.5}$$

This expression is the sum of four terms that correspond to different association between the peaks of the MUSIC spectrum and actual sound sources. It can be used to define the probabilities of the joint association events as shown in the following section.

4.2 Proposed extended MKF with joint probabilistic data association filter

We consider the problem of tracking two sound sources using a mobile robot in a noisy, reverberant environment. These sources are intermittent and possibly moving. The robot is equipped with a linear microphone array. By contrast with other methods [Gehrig and McDonough, 2006, Chakrabarty et al., 2014, Evers et al., 2016b] for tracking multiple sound sources, we jointly estimate the location and activity of all the sources over time.

In this framework, we make several assumptions. First, we assume that the sources move independently from each other and two AoA measurements are made, which are conditionally independent from each other. Second, we assume that the SAD will provide only one source activity value for all the sources. Third, each measurement can only be generated from at most one target source. Fourth, we make the assumption that the first observed AoA corresponds to one of the active sources, and the second one might correspond to the other source or can be a false alarm. Fifth, similarly to Chapter 3, we assume that the position of the robot is known. For simplicity, we will omit the state of the robot in the state vector and denote X as the state of the source.

In this section, we present an extended MKF with the JPDAF to localize two intermittent and moving sound sources over time. We first present the variables of our model before detailing the prediction and update step of the inference.

4.2.1 State and observation vectors

4.2.1.1 State vector

For each source, we define the state vector with the source activity, as follows:

$$\begin{bmatrix} X^t \\ a^t \end{bmatrix} = \begin{bmatrix} x^t \\ y^t \\ \theta^t \\ v^t \\ w^t \\ a^t \end{bmatrix}, \tag{4.6}$$

where $t \in \{1, 2\}$ is the index of the source, X^t contains the absolute position $[x^t, y^t]$ of the source, its orientation θ^t w.r.t. the x-axis, and its linear and angular velocities $[v^t, w^t]$; a^t is the source activity which is a discrete variable, where $a^t = 1$ indicates that the source is active, otherwise $a^t = 0$.

4.2.1.2 Observation vector

The observation vector Z_k in a given time frame k consists of two AoA measurements $Z_k^{\mathrm{l},1}$ and $Z_k^{\mathrm{l},2}$ which are obtained via a localization technique and the source activity measurement Z_k^{a} which is obtained via an SAD technique.

For the SAD, at each time step the SAD will provide only one source activity measurement which indicates whether there is at least one active source or not. The likelihood of the source activity w.r.t. the source activity measurement is represented by $P(Z_k^{\mathrm{a}}|a_k^1, a_k^2)$. To simplify the notation, we denote this likelihood by $P(Z_k^{\mathrm{a}}|a_k)$. If at least one source is active and there is no false detection, then the SAD will indicate that there is an active source. However, we do

not know exactly which source is currently active. If no source is active and there is no false detection, then the SAD will return that all sources are inactive.

With the linear microphone array, the distribution of the observation model corresponding to each AoA measurement $Z^{\mathrm{l},m}$, where $m \in \{1, 2\}$, given the state X_k^t of source t is bimodal when the source is active. As in Chapter 3, we can represent that observation model by a mixture of two Gaussians with equal weights as follows:

$$P_{\mathrm{sn}}(Z_k^{\mathrm{l},m}|X_k^t) = \sum_{j=1}^{2} \frac{1}{2}\mathcal{N}(Z_k^{\mathrm{l},m}; \hat{Z}_k^{\mathrm{l},m,j}, R_k^{m,j}), \tag{4.7}$$

where $\hat{Z}_k^{\mathrm{l},m,j}$ and $R_k^{m,j}$ are respectively the mean and variance of each Gaussian. When the source is not active, the observation model will have a uniform distribution and will be denoted as $P_{\mathrm{n}}(Z_k^{\mathrm{l},m})$.

4.2.1.3 Joint associations

In the context of multiple sound sources, we still need to deal with the association problem since we have no information about which AoA measurement originated from which target. Therefore, we define a joint association event β as the set of pairs of a measurement m and a target t, e.g., a joint association event could be "the first measurement corresponds to the second target and the second measurement corresponds to the first target". We also consider the situation when there is a false alarm: the measurement $Z_k^{l,m}$ does not correspond to a real target source. The probabilities of all joint association events sum to one and can be computed based on (4.5).

4.2.2 Prediction step

For each source, we compute the predicted belief $P(X_k^t, a_k^t|Z_{1:k-1})$ given the previous belief $P(X_{k-1}^t, a_{k-1}^t|Z_{1:k-1})$ and the state transition model $P(X_k^t, a_k^t|X_{k-1}^t, a_{k-1}^t)$ as follows:

$$\begin{aligned} P(X_k^t, a_k^t|Z_{1:k-1}) &= \sum_{a_{k-1}^t} \int P(X_k^t, a_k^t|X_{k-1}^t, a_{k-1}^t) P(X_{k-1}^t, a_{k-1}^t|Z_{1:k-1}) dX_{k-1}^t \\ &= \sum_{a_{k-1}^t} \int P(a_k^t|a_{k-1}^t) P(X_k^t|X_{k-1}^t) P(X_{k-1}^t, a_{k-1}^t|Z_{1:k-1}) dX_{k-1}^t, \end{aligned} \tag{4.8}$$

where $P(a_k^t|a_{k-1}^t)$ and $P(X_k^t|X_{k-1}^t)$ are respectively the source activity transition probability and the continuous state transition probability of each source t, which have been defined in Chapter 3.

Assume that at the previous time step $k-1$ the belief about the state (X_{k-1}, a_{k-1}) is given by a mixture of $N_{k-1|k-1}$ Gaussians:

$$\begin{aligned} &P(X_{k-1}, a_{k-1}|Z_{1:k-1}) \\ &= \sum_{i=1}^{N_{k-1|k-1}} \omega_{k-1|k-1}^i \mathcal{N}(X_{k-1}^1; \hat{X}_{k-1|k-1}^{1,i}, P_{k-1|k-1}^{1,i}) \mathcal{N}(X_{k-1}^2; \hat{X}_{k-1|k-1}^{2,i}, P_{k-1|k-1}^{2,i}) \delta(a_{k-1}^{1,i}) \delta(a_{k-1}^{2,i}), \end{aligned} \tag{4.9}$$

with weights $\omega_{k-1|k-1}^i$ such that $\sum_i \omega_{k-1|k-1}^i = 1$.

After the prediction step, we obtain the predicted density which contains four hypotheses. These hypotheses are at the same place which have the same mean and covariance but have different weights and activities:

$$
\begin{aligned}
&P(X_k, a_k|Z_{1:k-1}) \\
&= \sum_{i=1}^{N_{k-1|k-1}} \omega^i_{k-1|k-1} \Big[\\
&P(a_k^{1,i}=0|a_{k-1}^{1,i})P(a_k^{2,i}=0|a_{k-1}^{2,i})\mathcal{N}(X_k^1;\hat{X}^{1,i}_{k|k-1},P^{1,i}_{k|k-1})\mathcal{N}(X_k^2;\hat{X}^{2,i}_{k|k-1},P^{2,i}_{k|k-1})\delta(a_k^{1,i}=0)\delta(a_k^{2,i}=0)+ \\
&P(a_k^{1,i}=0|a_{k-1}^{1,i})P(a_k^{2,i}=1|a_{k-1}^{2,i})\mathcal{N}(X_k^1;\hat{X}^{1,i}_{k|k-1},P^{1,i}_{k|k-1})\mathcal{N}(X_k^2;\hat{X}^{2,i}_{k|k-1},P^{2,i}_{k|k-1})\delta(a_k^{1,i}=0)\delta(a_k^{2,i}=1)+ \\
&P(a_k^{1,i}=1|a_{k-1}^{1,i})P(a_k^{2,i}=0|a_{k-1}^{2,i})\mathcal{N}(X_k^1;\hat{X}^{1,i}_{k|k-1},P^{1,i}_{k|k-1})\mathcal{N}(X_k^2;\hat{X}^{2,i}_{k|k-1},P^{2,i}_{k|k-1})\delta(a_k^{1,i}=1)\delta(a_k^{2,i}=0)+ \\
&P(a_k^{1,i}=1|a_{k-1}^{1,i})P(a_k^{2,i}=1|a_{k-1}^{2,i})\mathcal{N}(X_k^1;\hat{X}^{1,i}_{k|k-1},P^{1,i}_{k|k-1})\mathcal{N}(X_k^2;\hat{X}^{2,i}_{k|k-1},P^{2,i}_{k|k-1})\delta(a_k^{1,i}=1)\delta(a_k^{2,i}=1)\Big],
\end{aligned}
\tag{4.10}
$$

where the means and variances corresponding to each source are given by

$$\hat{X}^i_{k|k-1} = f(\hat{X}^i_{k-1|k-1}, u_k), \tag{4.11}$$

$$F^i_{k-1} = \frac{\partial f(X,u)}{\partial X}\Big|_{X=\hat{X}^i_{k-1|k-1}}, \tag{4.12}$$

$$P^i_{k|k-1} = F^i_{k-1}P^i_{k-1|k-1}F^{iT}_{k-1} + Q^i_{k-1}. \tag{4.13}$$

The predicted density $P(X_k, a_k|Z_{1:k-1})$ is thus also a mixture of Gaussians, with the number of components $N_{k|k-1} = 4N_{k-1|k-1}$ and can be expressed as follows:

$$
\begin{aligned}
&P(X_k,a_k|Z_{1:k-1}) \\
&\quad= \sum_{i=1}^{N_{k|k-1}} \omega^i_{k|k-1}\mathcal{N}(X_k^1;\hat{X}^{1,i}_{k|k-1},P^{1,i}_{k|k-1})\mathcal{N}(X_k^2;\hat{X}^{2,i}_{k|k-1},P^{2,i}_{k|k-1})\delta(a_k^{1,i})\delta(a_k^{2,i}).
\end{aligned}
\tag{4.14}
$$

4.2.3 Update step

In the update step, similarly to Chapter 3 we compute the posterior belief $P(X_k, a_k|Z_{1:k})$ given the prediction and the new measurement Z_k:

$$
\begin{aligned}
P(X_k, a_k|Z_{1:k}) &= \eta P(Z_k|X_k, a_k)P(X_k, a_k|Z_{1:k-1}) \\
&= \eta P(Z^l_k|X_k)P(Z^a_k|a_k)P(X_k, a_k|Z_{1:k-1}).
\end{aligned}
\tag{4.15}
$$

However, as stated earlier, in the context of two sound sources, we must deal with the association problem when both sources are active. Considering the situation when the two sources are active, the posterior belief will be updated based on the association probability:

$$
\begin{aligned}
P(X_k, a_k|Z_{1:k}) =& \eta \sum_\beta P(X_k, a_k|Z_{1:k-1})P(Z_k|\beta, X_k, a_k)P(\beta) \\
=& \eta \sum_\beta \prod_{t=1}^{2} P(X^t_k, a^t_k|Z_{1:k-1}) \prod_{m=1}^{2} P(Z^m_k|\beta, X_k, a_k)P(\beta),
\end{aligned}
\tag{4.16}
$$

where η is a normalizing constant.

So, contrary to Chapter 3, the update step for the case of two sources requires computing the probabilities of the joint association events $P(\beta)$ and marginalizing over these associations.

In the following, we first present how the probabilities of the joint association events are defined and computed, and then present the update step in detail.

4.2.3.1 Joint association events

At time k, we obtain two AoA measurements but do not know their origin and whether they are false alarms. With two active sources, two AoA measurements and also considering a false alarm, we can have seven association events in total. As explained above, our choice of probability for detecting the first peak is 1, which reduces the possible associations to only four. Their probability $P(\beta)$, which are computed based on the training samples of the sensor model in 4.1, can be set as follows.

• $\beta = \beta_1$ corresponds to the case when the measurement Z^1 corresponds to the target X^1 but the measurement Z^2 does not correspond to a real target. Then, we have $P(\beta_1) = \frac{\alpha_1(1-\alpha_2)}{2} = \frac{1-\alpha_2}{2}$. This can be represented by the association table:

	FA	X_k^1	X_k^2
$Z^{l,1}$	0	1	0
$Z^{l,2}$	1	0	0

, where FA denotes a false alarm.

• $\beta = \beta_2$ corresponds to the case when the measurement Z^1 corresponds to the target X^2 but the measurement Z^2 does not correspond to a real target. Then, we have $P(\beta_2) = \frac{\alpha_1(1-\alpha_2)}{2} = \frac{1-\alpha_2}{2}$. This can be represented by the association table:

	FA	X_k^1	X_k^2
$Z^{l,1}$	0	0	1
$Z^{l,2}$	1	0	0

.

• $\beta = \beta_3$ corresponds to the case when the measurement Z^1 corresponds to the target X^1 and the measurement Z^2 corresponds to the target X^2. Then, we have $P(\beta_3) = \frac{\alpha_1\alpha_2}{2} = \frac{\alpha_2}{2}$. This can be represented by the association table:

	FA	X_k^1	X_k^2
$Z^{l,1}$	0	1	0
$Z^{l,2}$	0	0	1

.

• $\beta = \beta_4$ corresponds to the case when the measurement Z^1 corresponds to the target X^2 and the measurement Z^2 corresponds to the target X^1. Then, we have $P(\beta_4) = \frac{\alpha_1\alpha_2}{2} = \frac{\alpha_2}{2}$. This can be represented by the association table:

	FA	X_k^1	X_k^2
$Z^{l,1}$	0	0	1
$Z^{l,2}$	0	1	0

.

Those probabilities sum to 1.

By considering the four association events above, we can update the belief based on (4.16)

as:

$$\begin{aligned}P(X_k, a_k|Z_{1:k}) =&\eta \sum_{\beta} \prod_{t=1}^{2} P(X_k^t, a_k^t|Z_{1:k-1}) \prod_{m=1}^{2} P(Z_k^m|\beta, X_k, a_k)P(\beta)\\ =&\eta P(\beta_1)P(X_k^1, a_k^1|Z_{1:k-1})P_{\mathrm{sn}}(Z_k^{l,1}|X_k^1)P(X_k^2, a_k^2|Z_{1:k-1})P_{\mathrm{n}}(Z_k^{l,2})P(Z_k^a|a_k)\\ &+ \eta P(\beta_2)P(X_k^1, a_k^1|Z_{1:k-1})P_{\mathrm{n}}(Z_k^{l,2})P(X_k^2, a_k^2|Z_{1:k-1})P_{\mathrm{sn}}(Z_k^{l,1}|X_k^2)P(Z_k^a|a_k)\\ &+ \eta P(\beta_3)P(X_k^1, a_k^1|Z_{1:k-1})P_{\mathrm{sn}}(Z_k^{l,1}|X_k^1)P(X_k^2, a_k^2|Z_{1:k-1})P_{\mathrm{sn}}(Z_k^{l,2}|X_k^2)P(Z_k^a|a_k)\\ &+ \eta P(\beta_4)P(X_k^1, a_k^1|Z_{1:k-1})P_{\mathrm{sn}}(Z_k^{l,2}|X_k^1)P(X_k^2, a_k^2|Z_{1:k-1})P_{\mathrm{sn}}(Z_k^{l,1}|X_k^2)P(Z_k^a|a_k).\end{aligned} \tag{4.17}$$

So, the updated belief, which corresponds to the hypotheses that both source are active, contains four components corresponding to four joint association events.

4.2.3.2 Update step

By applying the update rule (4.15) to the predicted density $P(X_k, a_k|Z_{1:k-1})$, we obtain the new belief:

$$\begin{aligned}&P(X_k, a_k|Z_{1:k})\\ =&\eta \sum_{i=1}^{N_{k|k-1}} \omega_{k|k-1}^i P(Z_k^a|a_k^i)P(Z_k^l|X_k)\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i})\delta(a_k^{1,i})\delta(a_k^{2,i}).\end{aligned} \tag{4.18}$$

Let $P_c = P(Z_k^l|X_k)\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i})$. The probability distribution P_c will depend on the value of a_k^1 and a_k^2 as follows.

• If the activity hypothesis of the given component state that both sources are inactive, which means $a_k^1 = a_k^2 = 0$, then the two observations are uniform. In this case, the mean and covariance of the hypotheses will not be updated but only their weights. We have:

$$\begin{aligned}P_c =&P_{\mathrm{n}}(Z_k^{l,1})\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i})P_{\mathrm{n}}(Z_k^{l,2})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i})\\ =&P_{\mathrm{n}}(Z_k^{l,1})P_{\mathrm{n}}(Z_k^{l,2})\mathcal{N}(X_k^1; \hat{X}_{k|k}^{1,i}, P_{k|k}^{1,i})\mathcal{N}(X_k^2; \hat{X}_{k|k}^{2,i}, P_{k|k}^{2,i}),\end{aligned} \tag{4.19}$$

• If the activity hypothesis of the given component state that the first source is inactive but the second source is active, which means $a_k^1 = 0$ and $a_k^2 = 1$. Due to the assumption that the first observed AoA corresponds to an active source, so the first observation, which is a mixture of two Gaussians, corresponds to the second source and the second observation is uniform corresponds to a false alarm. In this case, only the mean and covariance of the Gaussians correspond to the second source will be updated. We have:

$$\begin{aligned}P_c =&P_{\mathrm{n}}(Z_k^{l,2})\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i})P_{\mathrm{sn}}(Z_k^{l,1}|X_k^{2,i})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i})\\ =&\sum_{j=1}^{2} \frac{1}{2} P_{\mathrm{n}}(Z_k^{l,2})\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i})\mathcal{N}(Z_k^{l,1}; h^j(X_k^{2,i}), R_k^{i,j})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i})\\ =&\sum_{j=1}^{2} \frac{1}{2} P_{\mathrm{n}}(Z_k^{l,2})\lambda_{Z^1}^{2,i,j}\mathcal{N}(X_k^1; \hat{X}_{k|k}^{1,i}, P_{k|k}^{1,i})\mathcal{N}(X_k^2; \hat{X}_{k|k}^{2,i}, P_{k|k}^{2,i}).\end{aligned} \tag{4.20}$$

• If the activity hypothesis of the given component state that the first source is active but the second source is inactive, which means $a_k^1 = 1$ and $a_k^2 = 0$. Then similarly, the first observation, which is a mixture of two Gaussians, corresponds to the first source and the second observation is uniform corresponds to a false alarm. In this case, only the mean and covariance of the Gaussians correspond to the first source will be updated. We have:

$$\begin{aligned} P_c =& P_{\text{sn}}(Z_k^{l,1}|X_k^{1,i})\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i}) P_{\text{n}}(Z_k^{l,2})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i}) \\ =& \sum_{j=1}^{2} \frac{1}{2}\mathcal{N}(Z_k^{l,1}; h^j(X_k^{1,i}), R_k^{1,i,j})\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i}) P_{\text{n}}(Z_k^{l,2})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i}) \\ =& \sum_{j=1}^{2} \frac{1}{2} P_{\text{n}}(Z_k^{l,2}) \lambda_{Z^1}^{1,i,j} \mathcal{N}(X_k^1; \hat{X}_{k|k}^{1,i}, P_{k|k}^{1,i})\mathcal{N}(X_k^2; \hat{X}_{k|k}^{2,i}, P_{k|k}^{2,i}). \end{aligned} \tag{4.21}$$

• If the activity hypothesis of the given component state that both source are active, which means $a_k^1 = 1$ and $a_k^2 = 1$. In this case, as we discussed earlier, the AoA measurement may not correspond to a source and we can have four association events. Based on (4.17) we have:

$$\begin{aligned} P_c =& \sum_{\beta} P(\beta) P(Z_k|\beta, X_k)\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i}) \\ =& P(\beta_1) P_{\text{sn}}(Z_k^{l,1}|X_k^{1,i})\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i}) P_{\text{n}}(Z_k^{l,2})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i}) \\ &+ P(\beta_2) P_{\text{n}}(Z_k^{l,2})\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i}) P_{\text{sn}}(Z_k^{l,1}|X_k^{2,i})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i}) \\ &+ P(\beta_3) P_{\text{sn}}(Z_k^{l,1}|X_k^{1,i})\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i}) P_{\text{sn}}(Z_k^{l,2}|X_k^{2,i})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i}) \\ &+ P(\beta_4) P_{\text{sn}}(Z_k^{l,2}|X_k^{1,i})\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i}) P_{\text{sn}}(Z_k^{l,1}|X_k^{2,i})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i}) \\ =& P(\beta_1) \sum_{j=1}^{2} \frac{1}{2}\mathcal{N}(Z_k^{l,1}; h^j(X_k^{1,i}), R_k^{i,j})\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i}) P_{\text{n}}(Z_k^{l,2})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i}) \\ &+ P(\beta_2) \sum_{j=1}^{2} P_{\text{n}}(Z_k^{l,2})\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i}) \frac{1}{2}\mathcal{N}(Z_k^{l,1}; h^j(X_k^{2,i}), R_k^{i,j})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i}) \\ &+ P(\beta_3) \sum_{j=1}^{2} \frac{1}{2}\mathcal{N}(Z_k^{l,1}; h^j(X_k^{1,i}), R_k^{i,j})\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i}) \sum_{q=1}^{2} \frac{1}{2}\mathcal{N}(Z_k^{l,2}; h^q(X_k^{2,i}), R_k^{i,q})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i}) \\ &+ P(\beta_4) \sum_{j=1}^{2} \frac{1}{2}\mathcal{N}(Z_k^{l,2}; h^j(X_k^{1,i}), R_k^{i,j})\mathcal{N}(X_k^1; \hat{X}_{k|k-1}^{1,i}, P_{k|k-1}^{1,i}) \sum_{q=1}^{2} \frac{1}{2}\mathcal{N}(Z_k^{l,1}; h^q(X_k^{2,i}), R_k^{i,q})\mathcal{N}(X_k^2; \hat{X}_{k|k-1}^{2,i}, P_{k|k-1}^{2,i}) \\ =& P(\beta_1) \sum_{j=1}^{2} \frac{1}{2}\lambda_{Z^1}^{1,i,j} P_{\text{n}}(Z_k^{l,2})\mathcal{N}(X_k^1; \hat{X}_{k|k}^{1,i}, P_{k|k}^{1,i})\mathcal{N}(X_k^2; \hat{X}_{k|k}^{2,i}, P_{k|k}^{2,i}) \\ &+ P(\beta_2) \sum_{j=1}^{2} \frac{1}{2}\lambda_{Z^1}^{2,i,j} P_{\text{n}}(Z_k^{l,2})\mathcal{N}(X_k^1; \hat{X}_{k|k}^{1,i}, P_{k|k}^{1,i})\mathcal{N}(X_k^2; \hat{X}_{k|k}^{2,i}, P_{k|k}^{2,i}) \\ &+ P(\beta_3) \sum_{j=1}^{2} \sum_{q=1}^{2} \frac{1}{4}\lambda_{Z^1}^{1,i,j} \lambda_{Z^2}^{2,i,q} \mathcal{N}(X_k^1; \hat{X}_{k|k}^{1,i}, P_{k|k}^{1,i})\mathcal{N}(X_k^2; \hat{X}_{k|k}^{2,i}, P_{k|k}^{2,i}) \\ &+ P(\beta_4) \sum_{j=1}^{2} \sum_{q=1}^{2} \frac{1}{4}\lambda_{Z^2}^{1,i,j} \lambda_{Z^1}^{2,i,q} \mathcal{N}(X_k^1; \hat{X}_{k|k}^{1,i}, P_{k|k}^{1,i})\mathcal{N}(X_k^2; \hat{X}_{k|k}^{2,i}, P_{k|k}^{2,i}). \end{aligned} \tag{4.22}$$

In the above equations, $\lambda_{Z^m}^{t,i,j}$ is a constant, where $t \in \{1,2\}$ is the target index, i is the index of the component, $j \in \{1,2\}$ is the index of the Gaussian in the observation model and $m \in \{1,2\}$ is the measurement index. Its value is computed as

$$\lambda_{Z^m}^{t,i,j} = \frac{1}{\sqrt{\left|2\pi(HP_{k|k-1}^{t,i}H^T + R_{Z^m}^{i,j})\right|}} e^{-\frac{1}{2}\left[Z_k^{1,m,j} - h^j(X_k^{t,i})\right]^T \left[HP_{k|k-1}^{t,i}H^T + R_{Z^m}^{i,j}\right]^{-1} \left[Z_k^{1,m,j} - h^j(X_k^{t,i})\right]}. \tag{4.23}$$

Therefore, the posterior belief $P(X_k, a_k|Z_{1:k})$ is also a mixture of Gaussians and can be expressed as:

$$P(X_k, a_k|Z_{1:k}) = \sum_{i=1}^{N_{k|k}} \omega_{k|k}^i \mathcal{N}(X_k^1; \hat{X}_{k|k}^{1,i}, P_{k|k}^{1,i}) \mathcal{N}(X_k^2; \hat{X}_{k|k}^{2,i}, P_{k|k}^{2,i}) \delta(a_k^{1,i}) \delta(a_k^{2,i}), \tag{4.24}$$

with $N_{k|k} = 17N_{k-1|k-1}$.

We update the state of each source as in Chapter 3. If $a_k = 0$, then $\hat{X}_{k|k}^i = \hat{X}_{k|k-1}^{i'}$. If $a_k = 1$, then

$$\hat{X}_{k|k}^i = \hat{X}_{k|k-1}^{i'} + G_k^{i'}\left[Z_k^j - h^j(\hat{X}_{k|k-1}^{i'})\right] \tag{4.25}$$

$$H_k^i = \frac{\partial h^j(X)}{\partial X}|_{X=\hat{X}_{k|k-1}^{i'}} \tag{4.26}$$

$$P_{k|k}^i = P_{k|k-1}^{i'} - G_k^{i'} H_k^{i'} P_{k|k-1}^{i'} \tag{4.27}$$

$$S_k^i = H_k^{i'} P_{k|k-1}^{i'} H_k^{i'T} + R_k^{i',j} \tag{4.28}$$

$$G_k^i = P_{k|k-1}^{i'} H_k^{i'T} (S_k^{i'})^{-1}, \tag{4.29}$$

where i' is the index of the component at prediction step on which component i is based, and it may not have the same value as index i.

With these equations we are now able to (i) predict the next belief on the state and activities of both sources and (ii) update this belief based on the two AoA and the SAD observation.

4.3 Experimental evaluation

We conducted numerical experiments to evaluate our extended MKF with the JPDAF technique for tracking two intermittent and possibly moving sources using a mobile robot in a noisy, reverberant environment.

Similarly to Chapter 3, we simulate experiments in an environment that mimics the *smart room* at Inria Nancy, with the same reverberation and acoustic noise conditions as in that room. The robot is a Turtlebot equipped with a Kinect sensor and the two sources are intermittent, and possibly moving.

4.3.1 Data

We randomly generated 200 trajectories of two sources for a duration of 10 s each. Each sound source is either static or moving with $v_s = 0.07$ m/s, $w_s = 8°$/s. One source is inactive from $t = 4.6$ s to $t = 6$ s and the other source is inactive from $t = 7.6$ s to $t = 8.5$ s. The robot trajectory was fixed like in Chapter 3.

The source AoA measurements were obtained from MUSIC-GSVD. The probability of correctly detecting the first peak is 1, and that for the second peak is 0.8. The SAD error rate is 5% and could be a false negative or a false positive.

4.3.2 Algorithm settings

We set the parameter values of the extended MKF with JPDAF to the same values as in Chapter 3.

We keep the number of hypotheses in the extended MKF to $N_{\max} = 50$. The time step duration, the covariance matrix Q, the initial position of the robot, the variance $R^{i,j}$, and the appearance/disappearance probabilities for the sound source are identical to the algorithm settings in Chapter 3.

We don't know the source locations at the beginning. Therefore, in all experiments, we initialize 21 hypotheses for each source estimation and distribute them evenly over the room area. These hypotheses can be visualized in Fig. 4.1 at time $t = 0\,\text{s}$.

4.3.3 Example run

Fig. 4.1 shows an example of localizing two intermittent, moving sources over time using the proposed framework. The two estimated sources are represented by different colors, namely blue and pink.

At time $t = 0\,\text{s}$, the mixture is initialized with evenly distributed components to approximate a uniform prior. The violet ellipses are due to the overlap between blue and pink ellipses which represent 95% confidence regions of each source location estimation. At time $t = 0.4\,\text{s}$, the hypotheses are mainly distributed along the directions from the robot to each source and the rest are symmetric w.r.t. the microphone axis due to the front-back ambiguity. Thanks to the robot motion, these symmetrical hypotheses become smaller at time $t = 2.5\,\text{s}$ and totally disappear after $t = 3.2\,\text{s}$. At time $t = 3.2\,\text{s}$, we also see that the colors of the two remaining hypotheses close to two sources can be differentiated. It is even easier to recognize at later time steps since the uncertainty in associating the source to target is reduced. The hypotheses corresponding to each target all move closer to the actual source positions. At time $t = 9\,\text{s}$, although the two sources are at a similar AoA, the filter can still localize accurately the two sources.

Fig. 4.2 depicts the estimation error in localizing the two sources and the ground truth source activities over time. In order to compute the estimation error, we must find a correct permutation to assign each estimated source to a true source. General speaking, there are two possible metrics to solve the permutation issue and compute the estimation error. The first metric is to compute the estimation error at each time step for all permutations and select the permutation with lowest estimation error. The second metric is to look at the overall trajectories and compute the estimation error between the real and the estimated trajectory of each source. This result in a single permutation over the whole trajectory. In this section and the following, we use the first metric to compute the estimation error.

In Fig. 4.2, from the beginning until $t = 3.2\,\text{s}$, the estimation error of the two sources decreases drastically due to the elimination of the front-back ambiguity over time. Between 4.6 and 6 s and between 7.6 and 8.5 s, one of two sources is inactive, however, this does not affect much the source localization and the estimation error of the two sources just slightly changes. From 9 s, the estimation error of two sources increase. The reason, as can be seen in Fig. 4.2 at

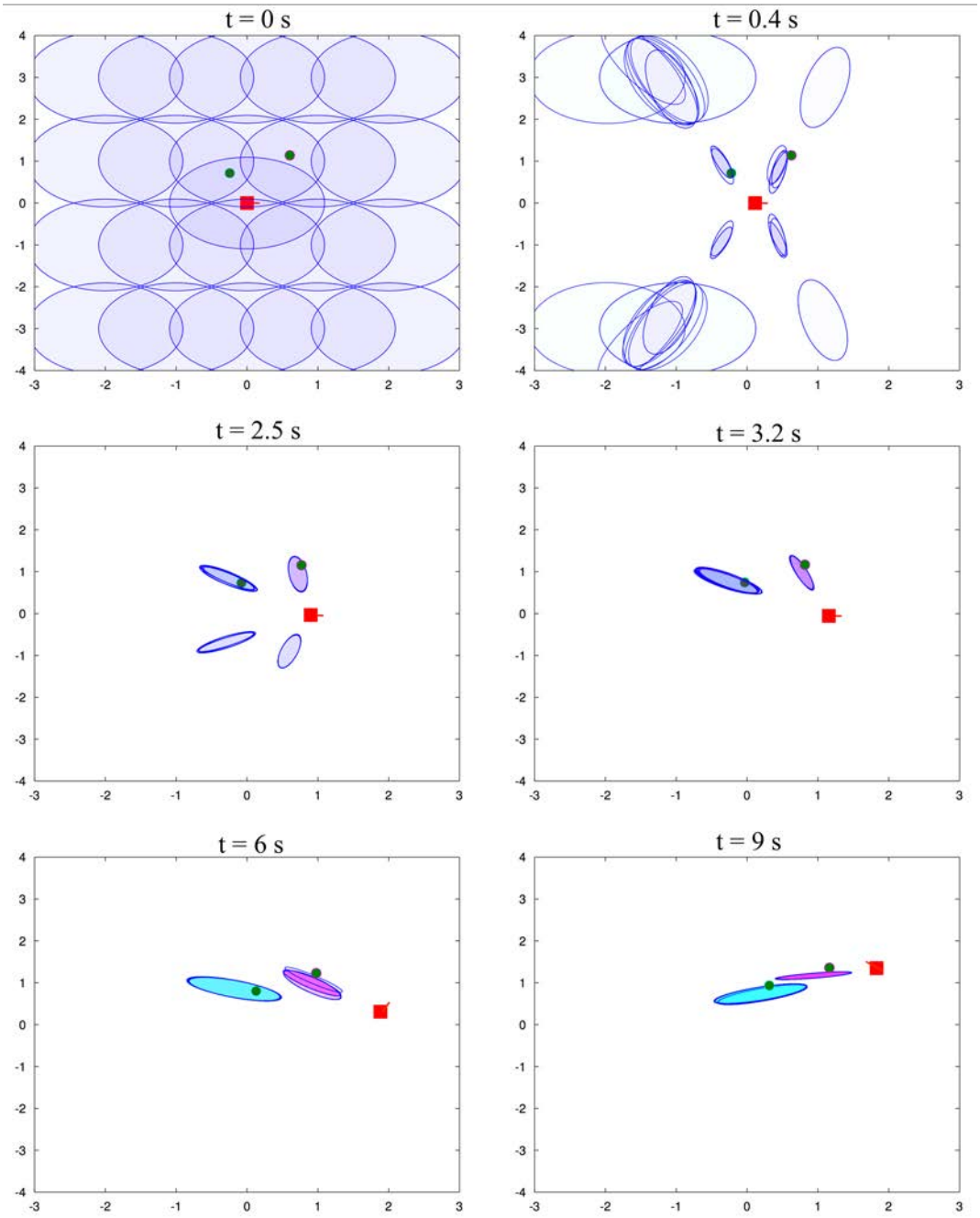

Figure 4.1: Visualization of our extended MKF with the JPDAF in an example of localizing two sources. Robot positions are shown as red squares, and the actual source positions as green circles. Blue and pink ellipses represent 95% confidence regions of source location estimation of various hypotheses in the mixture with a transparency proportional to the weight of the components.

time $t = 9\,\mathrm{s}$, is that the robot and the two sources start moving in the opposite directions. This actually makes it more difficult to correctly estimate the two source positions.

In general, over 10 s of localization, the estimation error of one of the sources is always lower

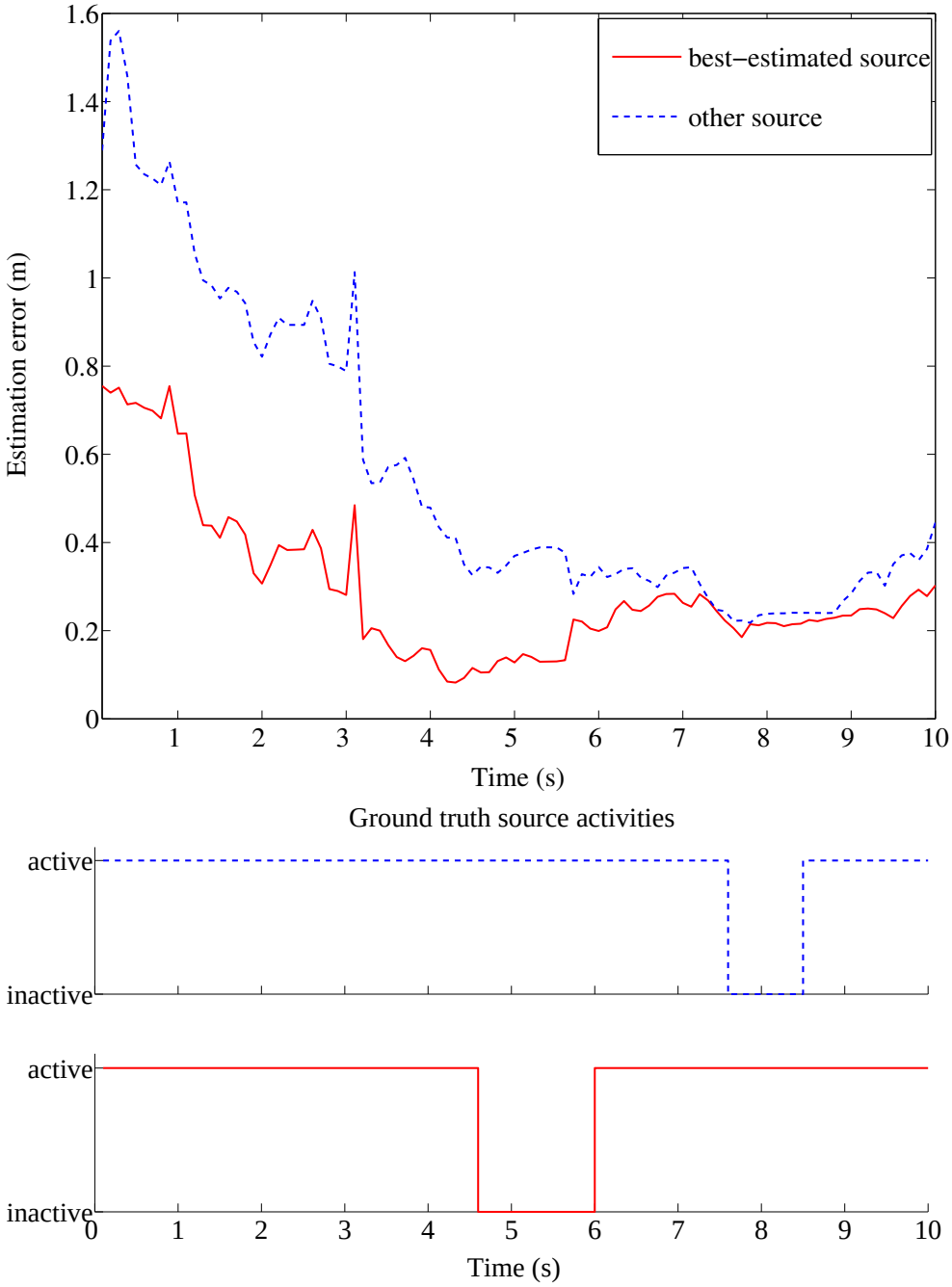

Figure 4.2: Top: Estimation error of the two source locations over time. Bottom: Ground truth source activities over time.

compared to the estimation error of the other. This could be explained by looking again at Fig. 4.1, in each time step, there is one source closer to the auditory fovea, where the localization accuracy is higher, compared to the other source. As a result, the estimation error of the source close to the auditory fovea is lower.

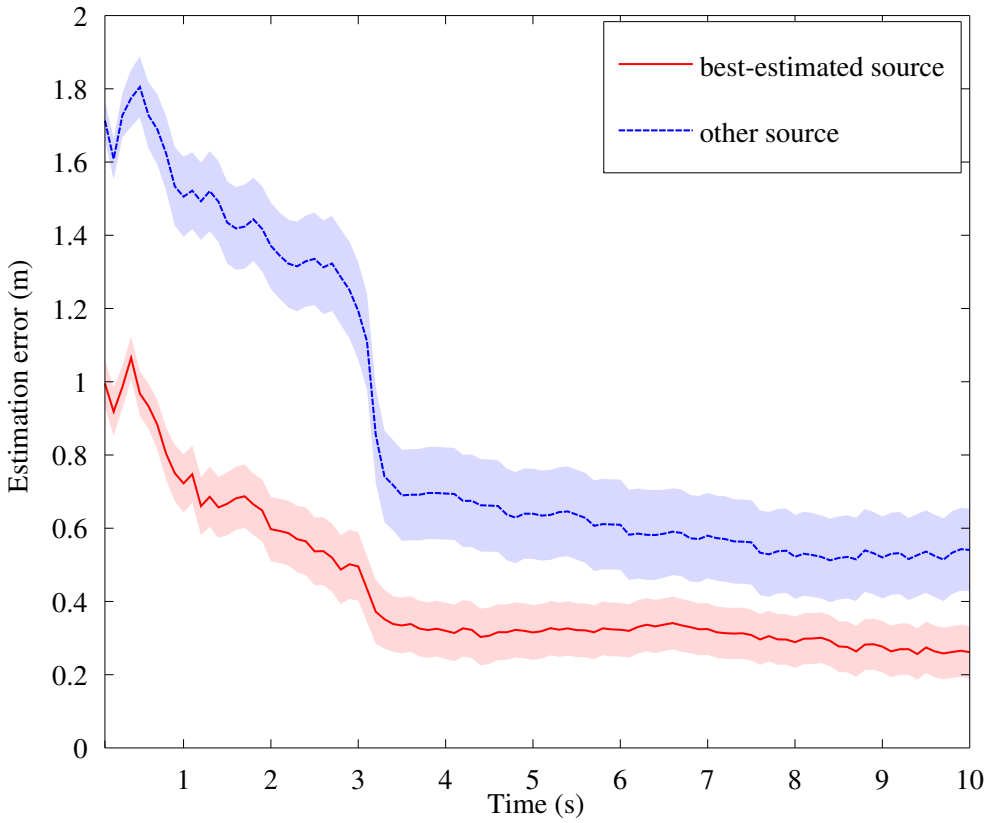

Figure 4.3: Average estimation error and 95% confidence interval over time of two source locations over all 200 experiments.

4.3.4 Statistical result

In this section, we investigate the performance of the proposed extended MKF with the JPDAF in tracking two intermittent, moving sources on a large experimental dataset from 200 experiments.

Fig. 4.3 shows the average estimation error over time and the 95% confidence interval over the two source locations over all experiments. Overall, one of the sources is significantly better localized with lower average estimation error than the other. Both estimation errors decrease drastically in the first 3.5 s. After 3.5 s, the average estimation error is more steady around 0.35 m for best-estimated source, and around 0.6 m for other source. The estimation error of the best-estimated source is slightly higher compared to the estimation error for a single source in Chapter 3. And it is significantly higher in the estimation error of the other source. The reason is that there is more uncertainty in the context of multiple sources compared to the context of a single source.

4.4 Summary

We proposed an extended MKF with the JPDAF for tracking two intermittent, moving sources in a noisy, reverberant environment. The main contribution of the proposed method is to jointly

estimate the location of two sources and their activities over time. A method for building a sensor model for the situation of two sound sources has been proposed.

We conducted a number of experiments with the proposed framework. Both the example run and the statistical analysis showed the ability of the method of handling uncertainty when tracking two intermittent, moving sources.

In theory, this framework can be extended to the context of more than two intermittent, moving sound sources. However, in practice, the computational complexity of the extended MKF with the JPDAF grows quickly with an increasing number of sources. In addition, the AoA measurements for three or more sources would have much lower accuracy.

Chapter 5

Optimal motion control for robot audition

In this chapter, we focus on the problem of finding the optimal robot motion to minimize the uncertainty on the source location. The target source is a single, intermittent, and possibly moving sound source. The belief about the source location is represented by a mixture of Gaussians as in Chapter 3. We present our contribution on defining the cost function for long-term motion planning with two alternative criteria: the Shannon entropy or the standard deviation of the estimated belief on the source location. The two criteria are integrated over time using a discount factor. After that, we present our contribution on adapting the MCTS method for efficiently finding the optimal robot motion which optimizes the cost function. Finally, we conduct a number of experiments to validate the proposed method and compare it with other robot motion planning methods. We show and compare the correlation between the estimation error and the two criteria. We also investigate the effect of the discount factor on the performance of the motion planning algorithm.

5.1 Cost function

The goal of optimal motion planning for robot audition is to find a sequence of robot motions that will minimize the estimation error on the source location. In reality, the ground truth source location is unknown. Therefore, we do not know the actual estimation error. Instead, we quantify the uncertainty about the estimated source location by the Shannon entropy or the standard deviation of the estimated belief distribution. By minimizing the uncertainty, we indirectly minimize the estimation error on the source location.

Let us assume that the robot has taken measurements up to a certain time step k. All the knowledge about the source location, source activity and robot pose at time k is represented by the belief $P(X_k, a_k|Z_{1:k})$. Now, we consider moving the robot to a new pose at time $k+1$. In the following, we use the terms "motion" and "action" interchangeably.

5.1.1 Shannon entropy criterion

For every possible motion sequence $u_{k+1:k+T}$ up to horizon $k+T$, we can quantify the uncertainty about the estimated source location by the entropy of the belief at each future time step $k+i$, $1 \leq i \leq T$. We sum these entropies with a forgetting factor $\gamma \in \mathbb{R}^+$ so that we can investigate the tradeoff between short vs long term. These entropies depend on future observations which

are unknown at current time k. By considering their expectation, the cost function can be expressed as

$$J_T = \sum_{i=1}^{T} \gamma^{i-1} \mathbb{E}_{Z_{k+1:k+i}}[H(P(X_{k+i}|Z_{1:k+i}))], \tag{5.1}$$

where $\mathbb{E}$ and H respectively stand for the expectation and the entropy. In practice, we approximate the expectation over $Z_{k+1:k+T}$ by drawing N_s random samples Z^s from the distribution $P(Z_{k+1:k+T}|Z_{1:k})$:

$$J_T \approx \sum_{i=1}^{T} \gamma^{i-1} \frac{1}{N_s} \sum_{Z^s} H(P(X_{k+i}|Z_{1:k}, Z^s_{k+1:k+i})). \tag{5.2}$$

One sample from this distribution is obtained by first drawing a sample Z^s_{k+1} from $P(Z_{k+1}|Z_{1:k})$, followed by a sample Z^s_{k+2} from $P(Z_{k+2}|Z_{1:k}, Z^s_{k+1})$, and so on. Then, we have:

$$P(Z_{k+1:k+T}|Z_{1:k}) = \prod_{i=1}^{T} P(Z_{k+i}|Z_{1:k+i-1}), \tag{5.3}$$

with

$$P(Z_{k+i}|Z_{1:k+i-1}) = \sum_{a_{k+i}} \int P(Z_{k+i}|X_{k+i}, a_{k+i}) P(X_{k+i}, a_{k+i}|Z_{1:k+i-1}) dX_{k+i}. \tag{5.4}$$

The optimal pose at time $k+1$ is achieved by performing the first action u_{k+1} in that sequence and the optimization is done iteratively after each new observation.

The entropies of the estimated beliefs cannot be computed in closed form for mixtures of Gaussians. An approximation based on a second-order Taylor series was proposed in [Huber et al., 2008] which we adopt hereafter. Assume that we have a probability density function $f(X)$ which is a mixture of Gaussians:

$$f(X) = \sum_{i=1}^{N} \omega_i \mathcal{N}(X; \mu_i, P_i). \tag{5.5}$$

The entropy of $f(X)$ can be computed as

$$\begin{aligned} H(f(X)) &\approx H_0(f(X)) - \sum_{i=1}^{N} \frac{\omega_i}{2} \int \mathcal{N}(X; \mu_i, P_i) F(\mu_i) \circ (X - \mu_i)(X - \mu_i)^T dX \\ &= H_0(f(X)) - \sum_{i=1}^{N} \frac{\omega_i}{2} F(\mu_i) \circ P_i, \end{aligned} \tag{5.6}$$

where $\circ$ is the so-called matrix contradiction operator which consists of an element-wise matrix multiplication and a subsequent summation of all matrix elements, $H_0(f(X))$ is computed based on the zeroth-order Taylor-series expansion:

$$\begin{aligned} H_0(f(X)) &\approx -\sum_{i=1}^{N} \int \omega_i \mathcal{N}(X; \mu_i, P_i) \log f(\mu_i) dX \\ &= -\sum_{i=1}^{N} \omega_i \log f(\mu_i), \end{aligned} \tag{5.7}$$

and

$$F(X) = \frac{1}{f(X)} \sum_{j=1}^{N} \omega_j P_j^{-1} \left(\frac{1}{f(X)} (X - \mu_j)(\nabla f(X))^T + (X - \mu_j)(P_j^{-1}(X - \mu_j))^T - I \right) \mathcal{N}(X; \mu_j, P_j). \tag{5.8}$$

5.1.2 Standard deviation criterion

As an alternative criterion, for all the possible motion sequences $u_{k+1:k+T}$ up to horizon $k+T$, we quantify the uncertainty by the standard deviation of the belief at each future time step $k+i$, $1 \leq i \leq T$. The standard deviation of the belief is the square root of its variance. We also sum these standard deviation terms with a forgetting factor. The future observations are unknown at current time k but by considering the expectation, the cost function can be computed as

$$J_T = \sum_{i=1}^{T} \gamma^{i-1} \mathbb{E}_{Z_{k+1:k+i}}[\sigma(P(X_{k+i}|Z_{1:k+i}))], \tag{5.9}$$

where σ is the standard deviation of the estimated source location belief. Similarly, we can approximate the expectation over $Z_{k+1:k+T}$ by drawing N_s random samples Z^s from the distribution $P(Z_{k+1:k+T}|Z_{1:k})$:

$$J_T \approx \sum_{i=1}^{T} \gamma^{i-1} \frac{1}{N_s} \sum_{Z^s} \sigma(P(X_{k+i}|Z_{1:k}, Z^s_{k+1:k+i})). \tag{5.10}$$

The method to draw a sample from this distribution is the same as in Section 5.1.1.

5.2 Monte Carlo tree search

Monte Carlo tree search (MCTS) [Chaslot et al., 2008, Browne et al., 2012] is a best-first search algorithm guided by the results of Monte-Carlo simulations for finding the optimal action from some root state. It is mostly employed in game play and has revolutionised Computer Go [Silver et al., 2016]. MCTS is rapidly replacing traditional search algorithms as the method of choice in challenging domains.

The algorithm builds and uses an adaptive tree of possible future states. It evaluates each state in the tree by the average outcome of simulations from that state. At each step, a new action is selected to create a new node and improve the evaluation of all parents of the tree. MCTS enables planning many time steps ahead and is often effective even with little or no prior domain knowledge. In addition, MCTS with the upper confidence bound for trees (UCT) algorithm [Kocsis et al., 2006] as the selection criterion can address the exploration-exploitation dilemma. In the following, we briefly describe the MCTS algorithm.

5.2.1 Algorithm outline

In MCTS, each node in the tree presents a possible state. A node contains at least two pieces of information: the current cumulated reward of the node (usually the average of the results of the simulations that visited this node), and the visit count of this node. An iteration of MCTS always starts from the root node and each iteration consists of four steps illustrated in Fig. 5.1, which are repeated as long as computational budget is left. The four sequential steps are as follows:

- Selection: The tree is traversed from the root node to a leaf node, using a selection strategy.
- Expansion: From this leaf node, an expansion strategy is called to create a new child node.
- Simulation: A simulation strategy is run until an end node is reached. In a game, this end node is usually when the game is over. The result of the simulation is encoded into a reward value.
- Backpropagation: The reward obtained above is propagated back through the tree according

to a backpropagation strategy.

Finally, when finishing building the tree, the selected action is the child of the root with the highest reward or highest visit count.

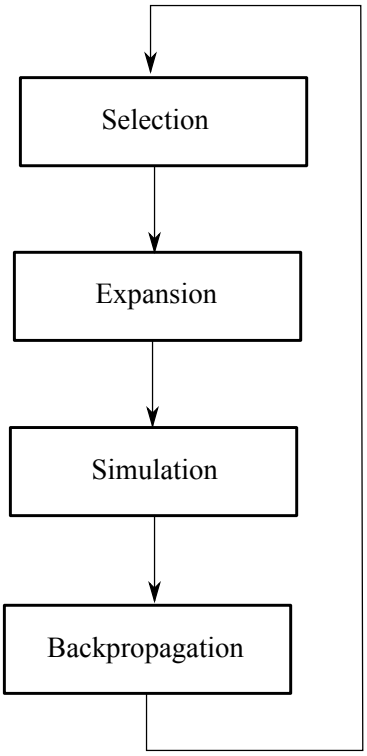

Figure 5.1: The four main steps in an iteration of the MCTS algorithm.

5.2.2 Optimism in the face of uncertainty

In the selection step of the MCTS algorithm, we select one of the children of a given node based on a strategy. The building of the tree mainly depends on the selection strategy. It controls the balance between exploitation and exploration. Exploitation is when we select the action that can lead to the best result according to the current knowledge. On the contrary, exploration deals with less promising actions that still need to be examined, due to the uncertainty in the execution and observation. In addition, sometimes a less promising action at the beginning can give a better result at a later time. The problem of balancing exploitation and exploration has been studied in the literature, in particular with respect to the multi-armed bandit problem [Katehakis and Veinott Jr, 1987, Auer et al., 2002]. The selection in MCTS could be viewed as a multi-armed bandit problem for a given node: selecting the next action which will give an unpredictable reward (similar to the outcome of a single random game). Based on the knowledge about the past results (past outcomes), we find the optimal action. However, the difference with the multi-armed bandit problem is that in MCTS we have sequential selections: the outcome at a time step depends on past actions.

For MCTS, a popular algorithm to balance between exploitation and exploration is the upper confidence bound applied to trees (UCT) algorithm [Kocsis et al., 2006]. UCT selects the child

of a node that satisfies:

$$UCT = \underset{n' \in \text{ children of } n}{\arg\max} \frac{\bar{Q}(n')}{N(n')} + C_p \sqrt{\frac{2 \log N(n)}{N(n')}}, \tag{5.11}$$

where $\bar{Q}(n')$ is the accumulated reward of the child node, $N(n)$ is the number of times the current (parent) node has been visited, $N(n')$ the number of times child n' has been visited, and $C_p > 0$ is a constant.

5.3 Adapting MCTS for robot audition

In practice, considering all possible motion sequences $u_{k+1:k+T}$ is intractable. We propose a method to adapt the MCTS algorithm to solve that problem. In the standard MCTS for games, the outcome of each simulation is often a binary value which represents "win" or "lose". We adapt MCTS for robot audition by defining the outcome or the reward as a function of the entropy or the standard deviation of the estimated belief. In addition, each node in the tree will contain the expected belief on the source location.

In this section, we briefly describe how the MCTS algorithm [Browne et al., 2012] can be adapted to efficiently search the tree of possible sequences to obtain the optimal action and minimize the uncertainty on the source location.

5.3.1 Formulation

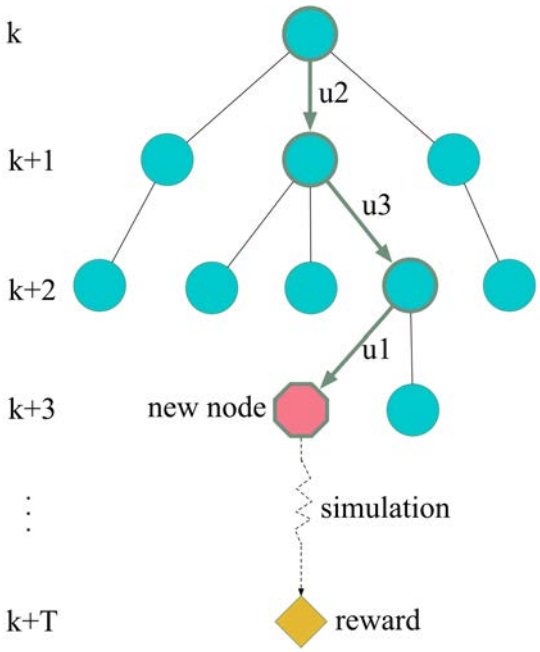

Figure 5.2: An iteration of the MCTS algorithm.

Fig. 5.2 shows an iteration of the MCTS algorithm at time k. Each level of the tree corresponds to one future time step. Each node n contains the information about: the pose of the robot, the estimated belief $b(n)$ on the source location, the untried actions among a finite set of possible actions, the accumulated reward $\bar{Q}(n)$ (see below), and the visit count $N(n)$.

The root node n_0 represents the pose of the robot at time k and carries the estimated belief $P(X_k, a_k|Z_{1:k})$. The links between a node and its child nodes represent different actions.

Starting from the root, a tree is built iteratively by selecting a node, adding a child corresponding to an untried action from this node, and following a random robot trajectory from this new node up to time $k+T$. The negative entropy or standard deviation Q corresponding to this trajectory is propagated upwards the tree to update the accumulated reward $\bar{Q}$ and the visit count N.

5.3.2 Selection

We select a node in the tree by applying the UCT criterion in (5.11). This criterion derives from the Chernoff-Hoeffding inequality which is valid for a bounded reward function [Hoeffding, 1963]. The Chernoff-Hoeffding inequality is stated in the theorem below.

Theorem: Let $Y_1, Y_2, ..., Y_n$ be independent random variables whose values are within the range $[a, b]$. Denote $\mu_i = \mathbb{E}(Y_i)$ as their expected values, $Y = \frac{1}{n}\sum_i Y_i$ and $\mu = \mathbb{E}(Y) = \frac{1}{n}\sum_i \mu_i$. Then for all $\epsilon > 0$, we have:

$$P(|Y - \mu| > \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}. \tag{5.12}$$

For us, Y_i is the reward Q at the end of each simulation and Y is the average reward $\frac{\bar{Q}(n')}{N(n')}$ at each child node in the tree. So, to satisfy the Chernoff-Hoeffding inequality, the two criteria must be bounded. We show that the two criteria are bounded in the following subsections.

5.3.2.1 Bounded entropy

For a scalar random variable X in the range $[a, b]$ with no other constraints, the maximum entropy distribution of X is the uniform distribution over this range. In that case the formula for calculating the maximum entropy is expressed as follows:

$$\begin{aligned} H_{\max} &= -\int p(X)\log p(X)dX \\ &= -\log p(X)\int p(X)dX \\ &= -\log p(X) \\ &= -\log\frac{1}{|b-a|} \\ &= \log|b-a|\,. \end{aligned} \tag{5.13}$$

With the state vector defined in (2.18), the formula for computing the maximum entropy of the belief is:

$$\begin{aligned} H_{\max} &= \log|\text{range}_{x_\text{r}}| + \log\left|\text{range}_{y_\text{r}}\right| + \log|\text{range}_{\theta_\text{r}}| + \log|\text{range}_{x_\text{s}}| + \log\left|\text{range}_{y_\text{s}}\right| \\ &\quad + \log|\text{range}_{\theta_\text{s}}| + \log|\text{range}_{v_\text{s}}| + \log|\text{range}_{w_\text{s}}| \\ &= 20.3027. \end{aligned} \tag{5.14}$$

In theory, the minimum entropy with perfect knowledge is $-\infty$ but it is not achievable in practice. So the lower bound can be computed as follows. In order to find the minimum entropy of the estimated belief, we begin the belief propagation with perfect knowledge about the source position. In the nonlinear MKF, it is represented by one hypothesis of active source

whose variance for the source position is equal to 0. After the prediction step, there will be one hypothesis of active source with a higher weight and one hypothesis of inactive source with a lower weight. The uncertainty will appear due to the process noise Q in the dynamic model. We then evaluate the uncertainty of estimation of the source location after the update step. We find the angle from the robot to the source, the distance from the robot to the source in the range from 0.18 m to 8 m, and the AoA observation such that the entropy of the belief after the update step above is minimum. As a result, the minimum entropy $H_{\min}$ is -38.7824 obtained for an AoA of 176°, which does not suffer from the front-back ambiguity, and a distance of 0.18 m from the robot to the source.

So, the entropy is bounded upwards by the entropy of the uniform distribution on the state vector, and downwards by the entropy of the probability distribution in the case when there is no front-back ambiguity and the sound source is as close as possible to the robot (0.18 m due to the size of the robot).

5.3.2.2 Bounded standard deviation

Let a and b be lower and upper bounds on the values of any random variable with a particular probability distribution. Then, according to Popoviciu's inequality on variances [Popoviciu, 1935], its variance satisfies:

$$\sigma^2 \leq \frac{1}{4}(b-a)^2, \tag{5.15}$$

or its standard deviation is bounded as follows:

$$-\frac{1}{2}|b-a| \leq \sigma \leq \frac{1}{2}|b-a|. \tag{5.16}$$

5.3.3 Expansion

In the example depicted in Fig. 5.2, the UCT criterion is iteratively used at each level to select a node until depth $t = k+2$ where an untried action is chosen to create a new node. The belief at this new node is computed as follows.

The predicted belief distribution $P(X_{t+1}, a_{t+1}|Z_{1:t})$ is first obtained:

$$P(X_{t+1}, a_{t+1}|Z_{1:t}) = \sum_{a_t} \int P(X_{t+1}, a_{t+1}|X_t, a_t) P(X_t, a_t|Z_{1:t}) dX_t \tag{5.17}$$

$$= \sum_{a_t} \int P(a_{t+1}|a_t) P(X_{t+1}|X_t) P(X_t, a_t|Z_{1:t}) dX_t. \tag{5.18}$$

Given one observation Z^s_{t+1} sampled as in (5.4), the updated belief at time step $t+1$ is expressed as:

$$P(X_{t+1}, a_{t+1}|Z_{1:k}, Z^s_{k+1:t+1}) = \eta P(Z^s_{t+1}|X_{t+1}, a_{t+1}) P(X_{t+1}, a_{t+1}|Z_{1:k}, Z^s_{k+1:t}), \tag{5.19}$$

where η is a normalizing constant.

5.3.4 Simulation

From the new expanded node, we perform a simulation until time step $k+T$ by selecting an action at random at each time step. The procedure to update the estimated belief after selecting one action is identical to that in the expansion step. At the end of the simulation, we evaluate

the reward by summing the negative entropy or standard deviation of the expected future belief from time step $k+1$ until time step $k+T$ with a corresponding forgetting factor γ:

$$Q = -\sum_{i=1}^{T} \gamma^{i-1} H(P(X_{k+i}|Z_{1:k}, Z^s_{k+1:k+i}), \tag{5.20}$$

or

$$Q = -\sum_{i=1}^{T} \gamma^{i-1} \sigma(P(X_{k+i}|Z_{1:k}, Z^s_{k+1:k+i})). \tag{5.21}$$

5.3.5 Backpropagation

After finishing the simulation, the number of times a node has been visited and the accumulated reward value are updated for all ancestor nodes, up to the root node. For each ancestor node n, $N(n)$ is incremented by 1 and $\bar{Q}(n)$ is incremented by Q.

The iterations of building the tree terminate when the computational budget has been used up. The optimal pose n at time $k+1$ is then chosen based on the average reward $\frac{\bar{Q}(n)}{N(n)}$. The robot performs the corresponding action in order to move to this pose, takes a new real measurement Z_{k+1}, builds a new tree to find the next optimal pose, and so on.

5.4 Evaluation

In order to obtain statistically meaningful results, a large number of experiments is needed that can hardly be conducted with a real robot. Therefore, like in Chapter 3, we conducted simulated experiments mimicking the *smart room* at Inria Nancy, where the robot is a Turtlebot equipped with a 4-microphone Kinect sensor forming a linear array. The target source is an intermittent and possibly moving sound source. We simulated the robot and source movements and the resulting location and activity measurements. The source is silent from $k = 1.2\,\text{s}$ to $k = 2\,\text{s}$. It is static or mobile ($v_s = 0.07\,\text{m/s}$, $w_s = 8°/\text{s}$). Instead of using an approximate Gaussian sensor model as in [Bustamante et al., 2017], we generated the location measurements using an accurate sensor model for GSVD-based localization trained on simulated data. The training data were simulated using state-of-the-art methods for the simulation of reverberation and acoustic noise, whose parameters closely match the real reverberation and acoustic noise conditions in the smart room [Vincent et al., 2015]. For each motion planning approach, we generated 200 experiments with different random initial robot locations and source locations.

5.4.1 Experimental protocol

To start from an informative belief, the robot first follows a fixed trajectory while updating the belief using the extended MKF for 3 s. After this, it follows the proposed MCTS algorithm with $T = 20$, and 700 tree nodes. The action set contains 13 discretized actions as listed in the following table. They are basically moving forward, moving backward, turning left or right while moving forward or backward, turning with different speed and radius.

v_l	0.6	0.6	0.6	0.6	0.6	0.5	0.4	0.3	0.2	0.4	-0.6	0.6	-0.4
v_r	0.6	0.5	0.4	0.3	0.2	0.6	0.6	0.6	0.6	-0.6	-0.6	-0.6	0.6

The optimal selected action is applied 5 times in a row, with 0.2 s time step. For comparison, we applied the same procedure to three other motion planning algorithms: a greedy algorithm

inspired from [Bustamante et al., 2016, Bustamante et al., 2017, Schymura et al., 2017], MCTS with an approximate cost function inspired from [Vincent et al., 2015], and random motion. The greedy algorithm finds the optimal action that minimizes the expected entropy or the expected standard deviation of the belief one step ahead. The MCTS algorithm with approximate cost function computes the expected entropy or standard deviation for each future pose in (5.20) or (5.21) by recursively predicting the belief via (5.18) at all time steps, but updating it via (5.19) with an observation only for the last step. The random method simply chooses the next pose of the robot at random.

5.4.2 Example trajectory

In this section, we show an example scenario and compare the robot trajectory from the MCTS algorithm with the greedy algorithm.

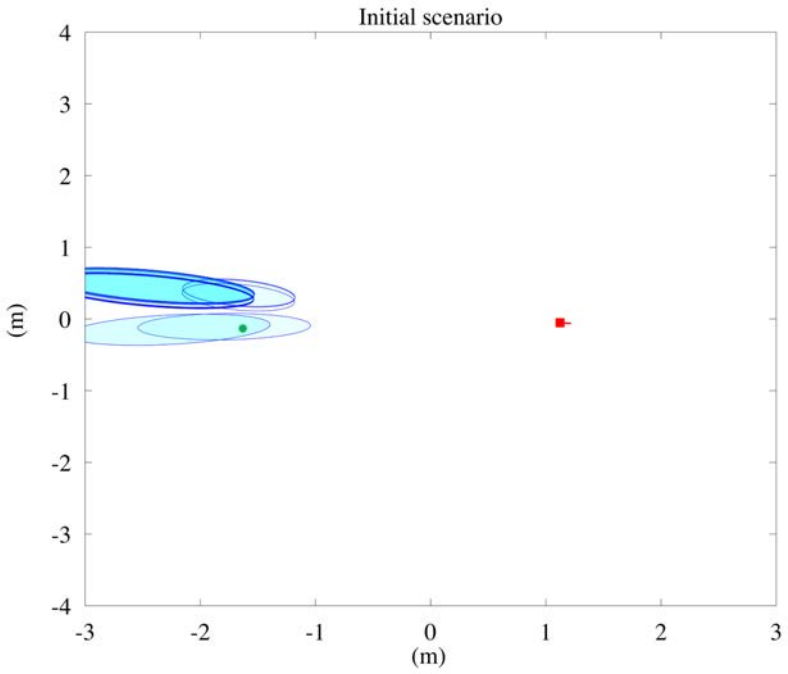

Figure 5.3: Initial position of the robot and the source and estimated belief before running motion planning strategies. Blue ellipses represent 95% confidence regions of source location estimation of various hypotheses in the mixture with a transparency proportional to the weight of the components. The robot position is shown as a red square with a red line indicating the direction of the microphone array, and the actual source position is shown as a green circle.

The initial example scenario is depicted in Fig. 5.3. In this situation, the front-back ambiguity still exists in the source location belief. The distance from the center of the microphone array to the true robot position is around 3 m. At this point, the microphone array is in the endfire position, i.e., the microphones are arranged in line with the source position. In this endfire position, the localization accuracy will be lower compared to the broadside position in which the microphones are arranged in a perpendicular line to the direction of sound propagation.

From this initial robot-source position and estimate, we implement the MCTS algorithm and the greedy algorithm separately to find the subsequent optimal robot trajectory. The optimal

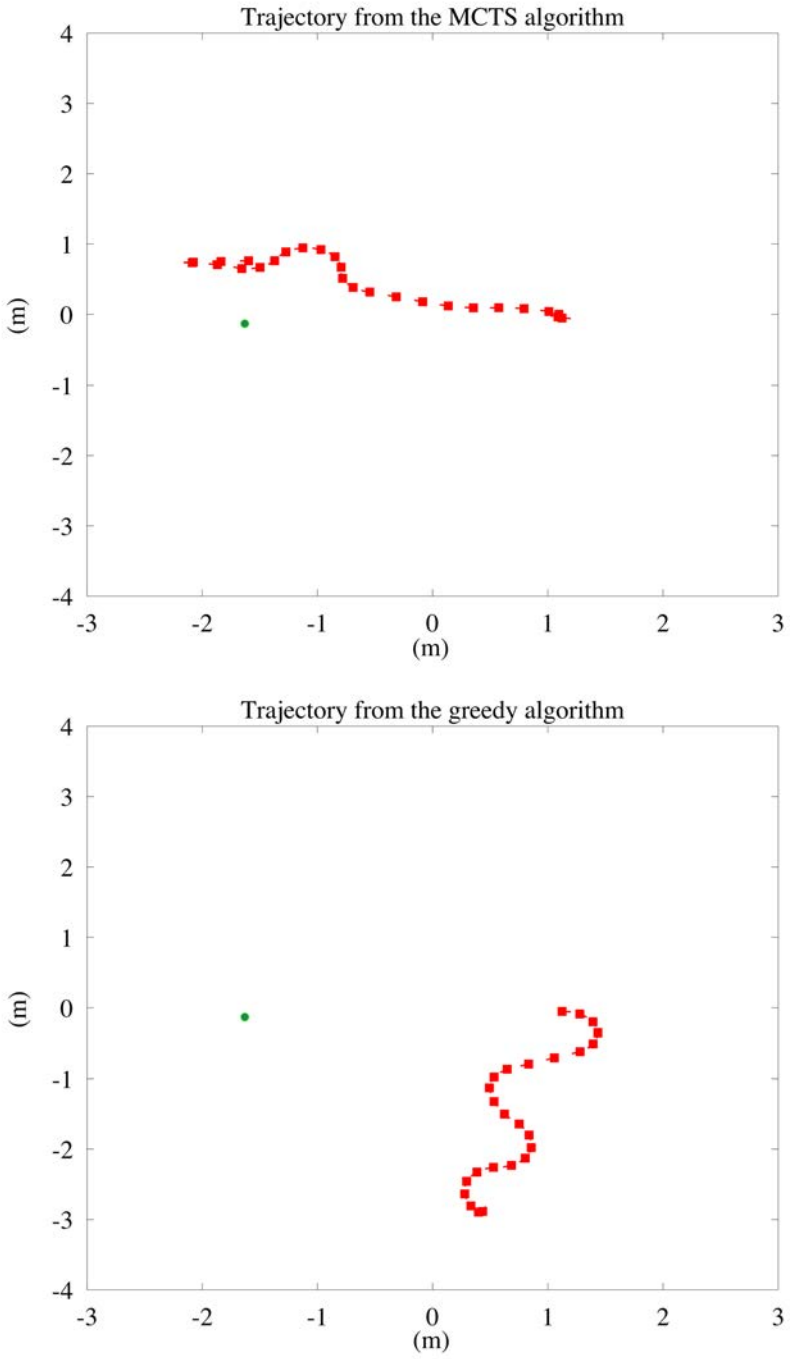

Figure 5.4: Top: Example robot motion sequence obtained from the MCTS algorithm with entropy criterion. Bottom: Example robot motion sequence obtained from the greedy algorithm with entropy criterion. Robot positions are shown as red squares with a red line indicating the direction of the microphone array, and the actual source position is shown as a green circle.

robot trajectories from the MCTS algorithm (top) and from the greedy algorithm (bottom) are showed in Fig. 5.4. In the robot trajectory obtained from the MCTS method, the robot rotates 180° around the current position first. This rotation does not minimize the entropy of the

estimated belief in the short term because the microphones are still in endfire position. However, it helps to eliminate the front-back confusion in the estimated belief. After the rotation, the robot moves toward to the direction of the source in order to increase the SNR. Actually, from the figure, we see that the robot does not move in a straight line towards the direction of the source but it makes a slight turn when approaching target source. Gradually, the microphones rotates to the broadside position, a.k.a the auditory fovea, where the localization accuracy is higher.

In contrast, the robot trajectory attained from the greedy method in Fig. 5.4 (bottom) shows that the robot makes a turn down first. The purpose of this movement is to put the microphone array in the auditory fovea and minimize the entropy of the estimated belief. It then makes a slight turn but still in the downward direction, and as the result, the robot is still far from the source and the SNR does not change, or even decreases.

Fig. 5.5 and Fig. 5.6 show the estimated source location when the robot follows the trajectory from the MCTS and greedy algorithm respectively. Blue ellipses represent 95% confidence regions of source location estimation of various hypotheses in the mixture with a transparency proportional to the weight of the components. Robot positions are shown as red squares, and the actual source position as a green circle. From these two figures, we can clearly see that the robot motion from the MCTS algorithm leads to better localization result compared to the robot motion from the greedy algorithm, although they started from the same initial position.

Fig. 5.7 shows the entropy value of the estimated belief over time for both the MCTS and the greedy method. During the first 2 s, the greedy method yields a lower entropy compared to MCTS. This result is due to the greedy motion taken to minimize the entropy in the short term. However the entropy does not change much over time afterwards. By contrast, from 2.5 s to 4 s, the entropy value obtained by MCTS decreases drastically. This corresponds to the time when the robot gets closer to the source and makes a turn to put the microphone array in broadside position. The entropy value is more steady after 4 s. This could be the minimum entropy value.

Fig. 5.8 presents the estimation error over time for both the MCTS and the greedy method, that is the distance between the estimated source position and the true position, which is available only in simulation. After 2 s, it is not surprising that the estimation error of MCTS decreases and becomes much lower compared to the greedy method.

From this example scenario, we can see that the greedy method tends to choose the action that will immediately reduce the entropy, however, in the long term this does not yield the optimal trajectory. In contrast, with the long-term motion planning method using MCTS, at the beginning, the motion may not result in an optimal value, but in the long run, we can obtain a better result.

In the following, we investigate the performance of the MCTS method for improving the estimated source location over time on a large experimental dataset.

5.4.3 MCTS vs other motion planning approaches

We compare the performance of the MCTS approach with other motion planning approaches for two criteria: entropy and standard deviation. In the MCTS approach, to ensure a fair comparison with other motion planning approaches, we set the value of the discount factor to $\gamma = 1$.

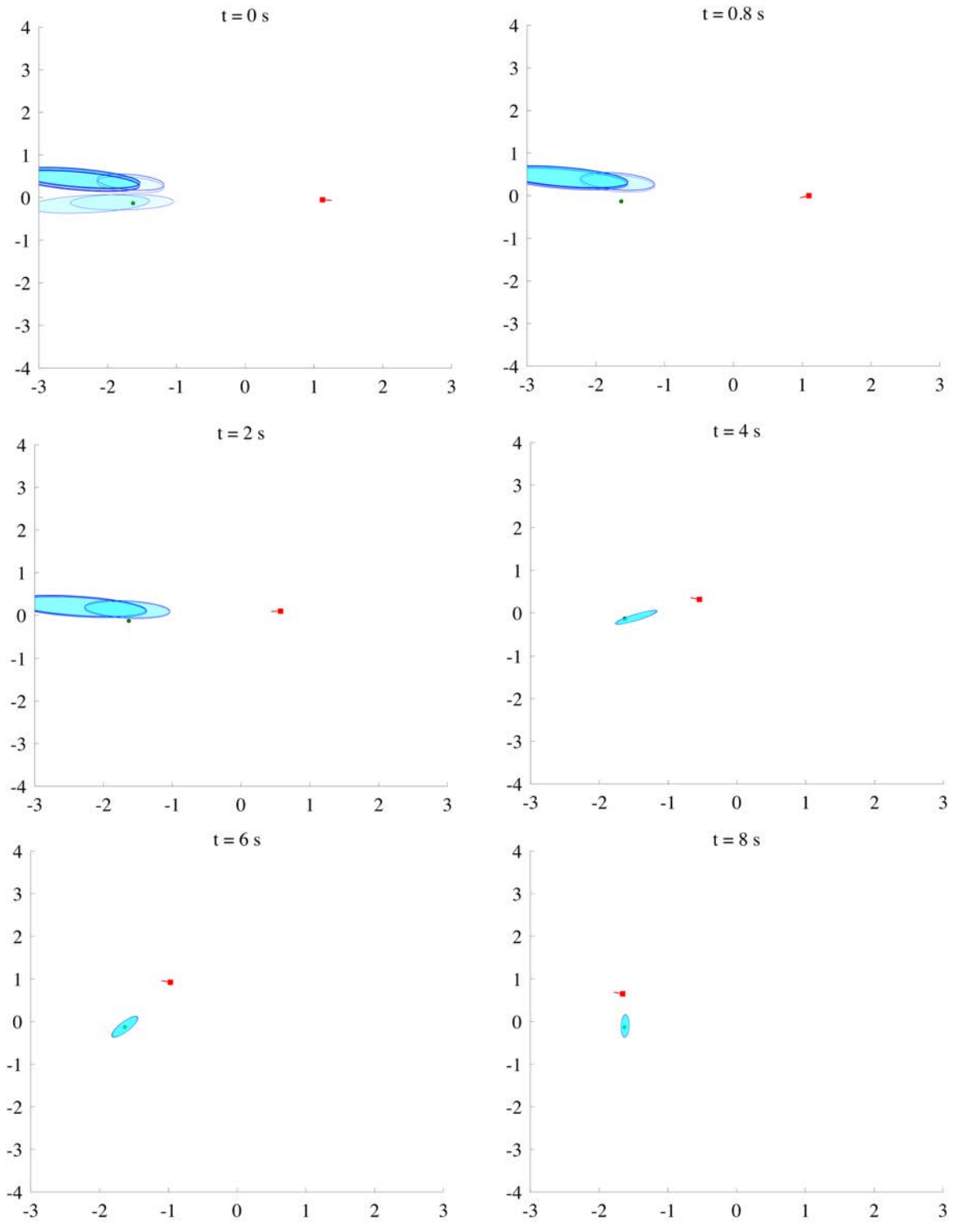

Figure 5.5: Estimation of the source location when the robot follows the trajectory from the MCTS algorithm.

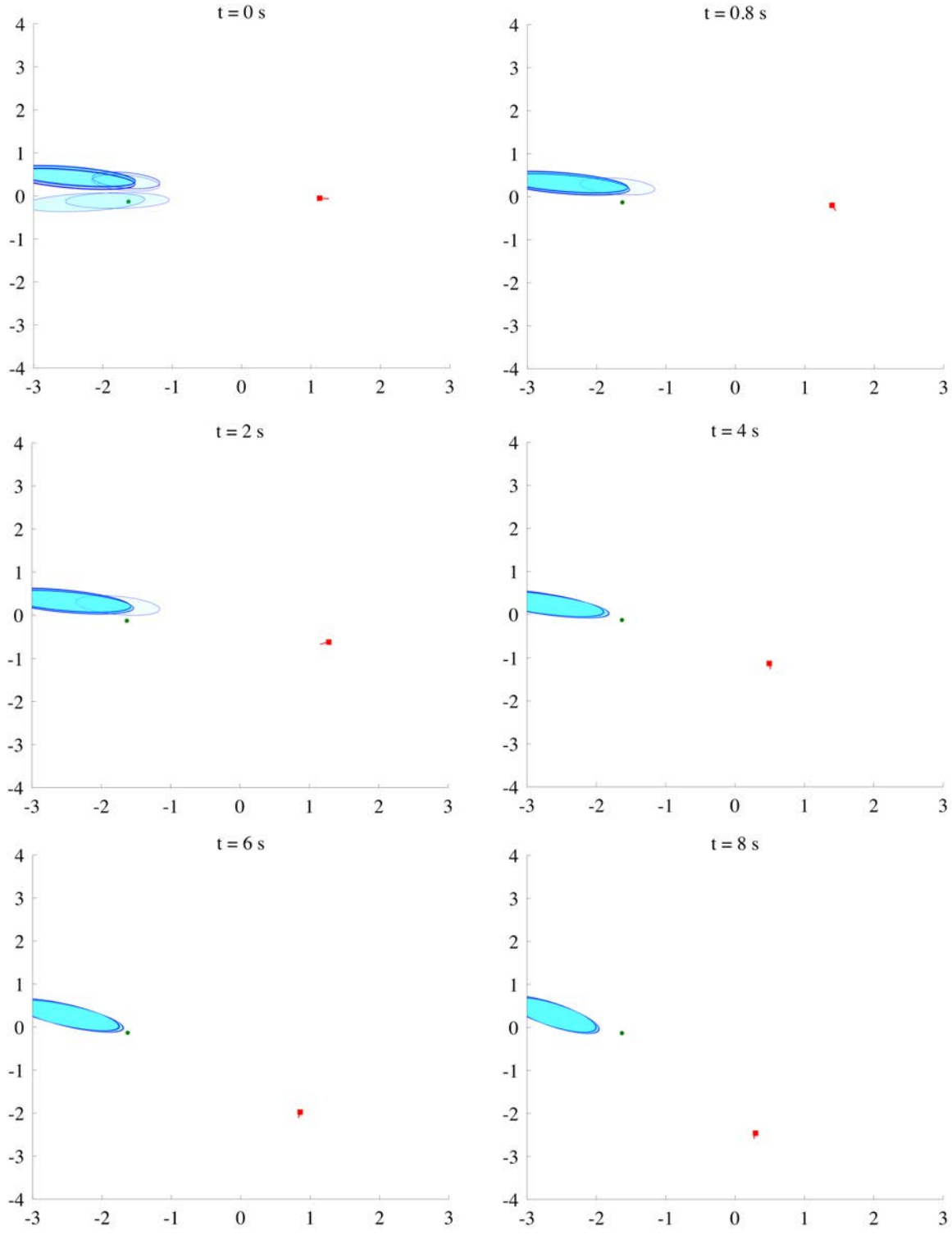

Figure 5.6: Estimation of the source location when the robot follows the trajectory from the greedy algorithm.

5.4.3.1 Entropy criterion

Fig. 5.9 shows the entropy of the estimated belief over time for all algorithms, on average over all experiments. The entropy values of all methods decrease until the time $k = 1.2$ s. From

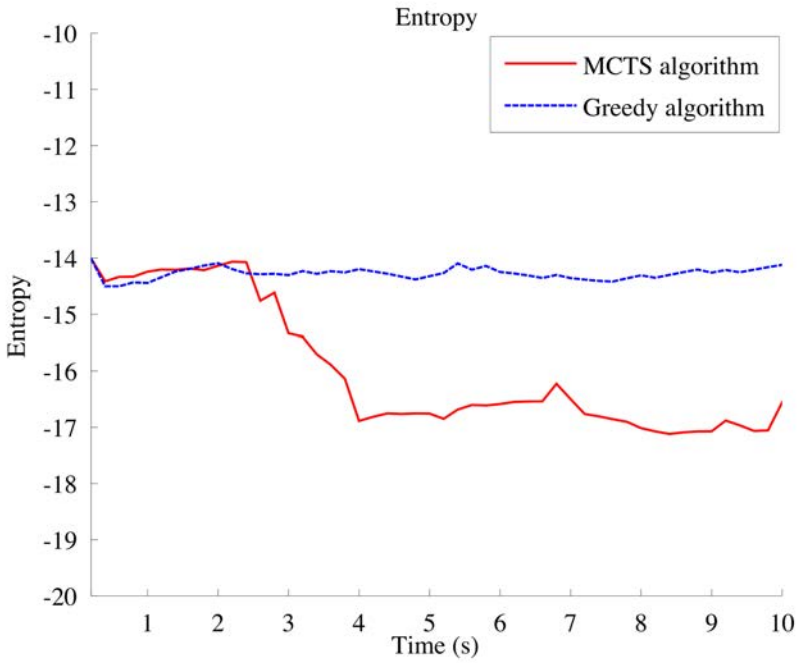

Figure 5.7: Entropy over time with the MCTS algorithm and the greedy algorithm.

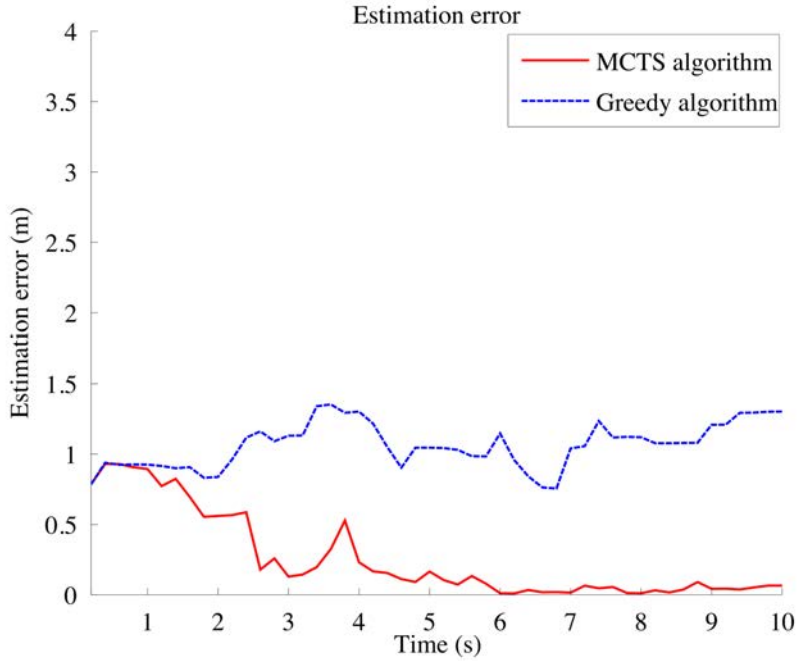

Figure 5.8: Estimation error over time with the MCTS algorithm and the greedy algorithm.

that to time $k = 2$ s, they rise up. This is due to the silent interval of the sound source. However, after this period, when the source is active again and we have better information from the AoA measurement, the entropy of all methods decreases over time except for the random

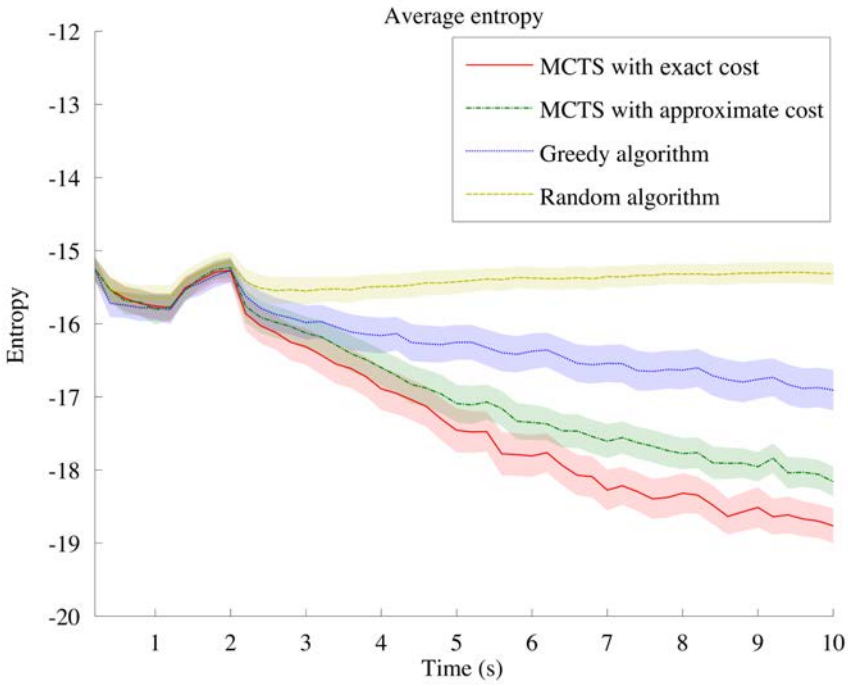

Figure 5.9: Average entropy and 95% confidence interval over time of the 4 algorithms with the entropy criterion over all 200 experiments.

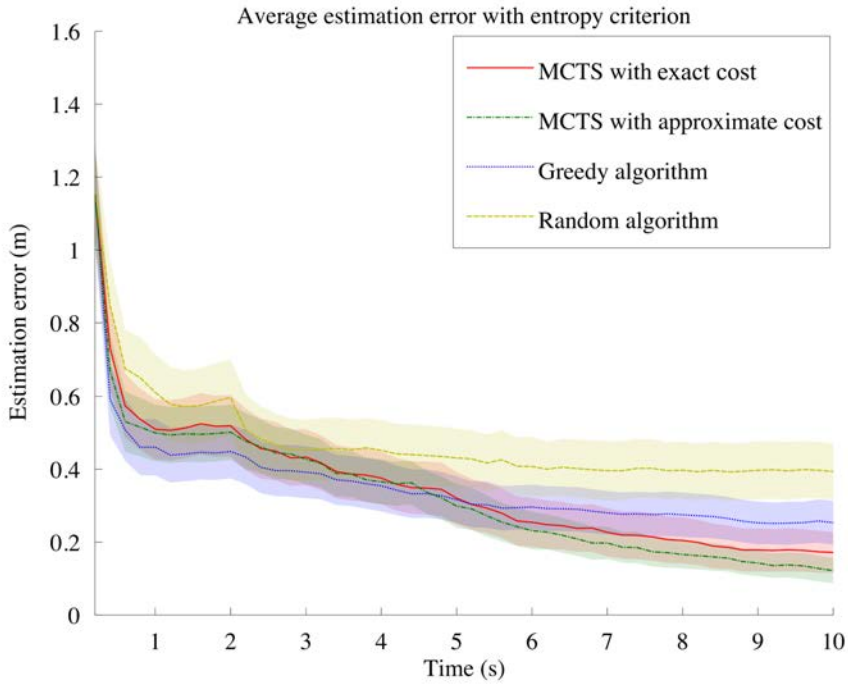

Figure 5.10: Average estimation error and 95% confidence interval over time of the 4 algorithms with the entropy criterion over all 200 experiments.

algorithm. So, after the silent interval, the random motion approach cannot help to improve the performance.

From time $k = 4$ s, the two flavors of MCTS already yield significantly lower entropy compared to the greedy algorithm and the random algorithm. The entropy decreases drastically at further time steps. This is because MCTS optimizes the entropy in the long run. In contrast to that, random motion barely reduces it and the greedy algorithm only optimizes it one time step ahead.

This result is expected as our method is actually optimizing the entropy. However, the objective of this work is to use robots to better localize sound sources. Therefore the evaluation of the algorithms should be done according to the estimation error. The average estimation error over time is presented for all algorithms in Fig. 5.10. Again, the average estimation error of both MCTS methods is smaller compared to greedy and random motion in the long run, although the estimation error from the greedy algorithm is smaller but not significantly so in the first four seconds compared to both MCTS methods. During the silent interval of the source, the estimation error of all methods does not change, because the MKF method that we use for estimating the source location can deal well with the intermittent source.

With a Wilcoxon signed-rank test [Wilcoxon, 1945], we can show that both MCTS variants yield a significantly smaller entropy and estimation error than the greedy and random algorithms ($p < 0.01$). The MCTS variant with exact cost has significantly smaller entropy value than the MCTS variant with approximate cost ($p < 0.01$). However, the two MCTS methods are not significantly different in terms of estimation error.

5.4.3.2 Standard deviation criterion

Fig. 5.11 shows the standard deviation of the estimated belief over time for all algorithms, on average over all experiments. The average standard deviation for all methods fall drastically until $k = 1.2$ s. There is slight rise in standard deviation for all methods during the time when the source is not active (from $k = 1.2$ to 2 s). Similarly with the entropy criterion, after time $k = 4$ s, the two MCTS approaches yield significantly lower standard deviation compared to the greedy algorithm and the random algorithm. The standard deviation decreases drastically at further time steps.

The average estimation error over time with standard deviation criterion is presented for all algorithms in Fig. 5.12. In the long run, the average estimation error of both MCTS methods is smaller compared to greedy and random motion. In addition, the estimation error of all methods does not change during the silent interval of the source when we still use the MKF method for estimation.

With a Wilcoxon signed-rank test, we can show that both MCTS approaches yield a significantly smaller entropy and estimation error than the greedy and random algorithms ($p < 0.01$). However, contrary to the entropy criterion, the two MCTS approaches with standard deviation criterion are not significantly different from each other in terms of both standard deviation and estimation error.

5.4.4 Relation of both criteria with estimation error

We investigate the relation between the entropy criterion and the estimation error as well as between the standard deviation criterion and the estimation error in the MCTS approach. We plot the entropy as a function of the estimation error during the 2 s interval from $k = 8$ s to

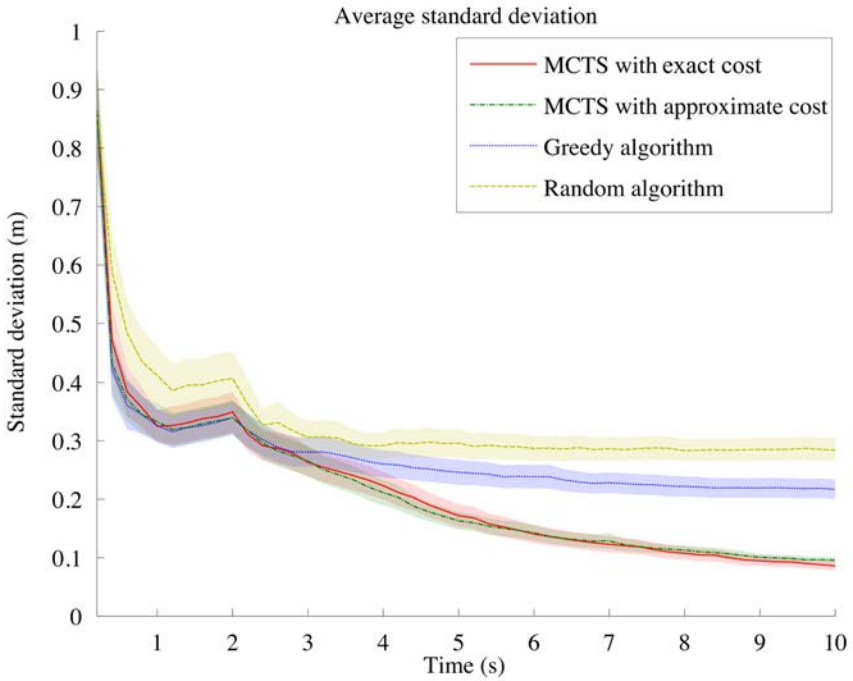

Figure 5.11: Average standard deviation and 95% confidence interval over time with the standard deviation criterion of the 4 algorithms over all 200 experiments.

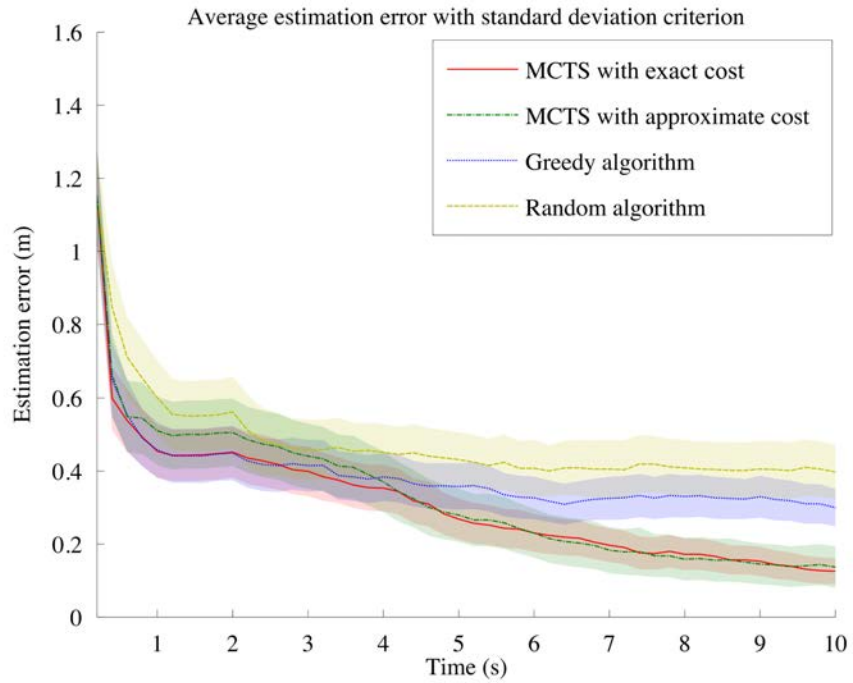

Figure 5.12: Average estimation error and 95% confidence interval over time of the 4 algorithms with the standard deviation criterion over all 200 experiments.

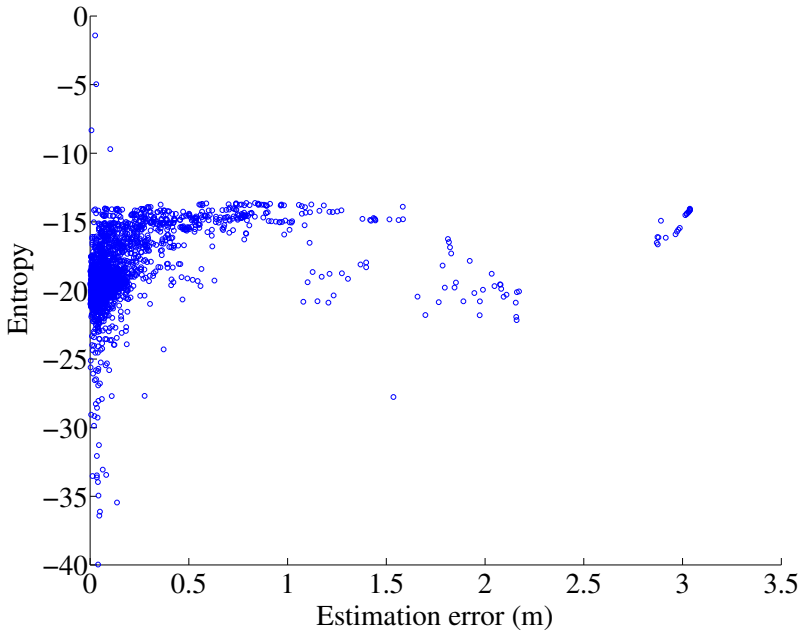

Figure 5.13: Correlation between the entropy and the estimation error.

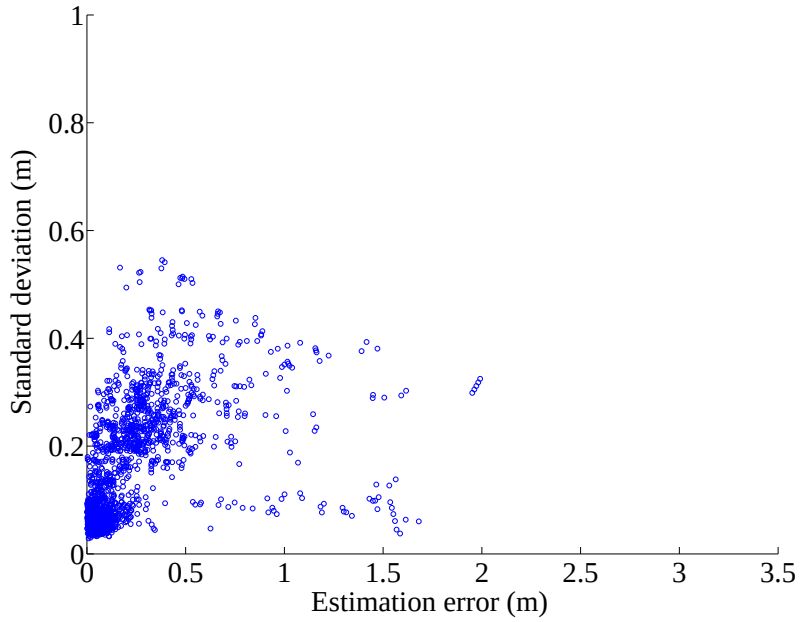

Figure 5.14: Correlation between the standard deviation and the estimation error.

$k = 10\,\mathrm{s}$. Similarly, we plot the standard deviation as a function of the estimation error in the same interval of time.

From Fig. 5.13, we can see that not only small values of entropy correspond with a small estimation error but the larger values of entropy also sometimes correspond with a small estimation error. This could explain the reason why the average entropy of MCTS with exact cost is significantly smaller than the average entropy of MCTS with approximate cost in the Fig. 5.9 but it is not so when we compare the estimation errors in Fig. 5.10.

In Fig. 5.14, most of the smaller values of standard deviation correspond with smaller values of estimation error and most of the larger values of standard deviation correspond with larger values of estimation error. We can also see this relation presented in Fig. 5.11 and Fig. 5.12.

From these two figures, we can conclude that the standard deviation is a better criterion due to the fact that optimizing the standard deviation would minimize the estimation error.

5.4.5 Effect of the discount factor

We also analyze the impact of different discount factor values in the entropy and the standard deviation criteria. Six discount factor values are tested. For each discount factor value, we run 200 experiments for different random initial robot locations and source locations.

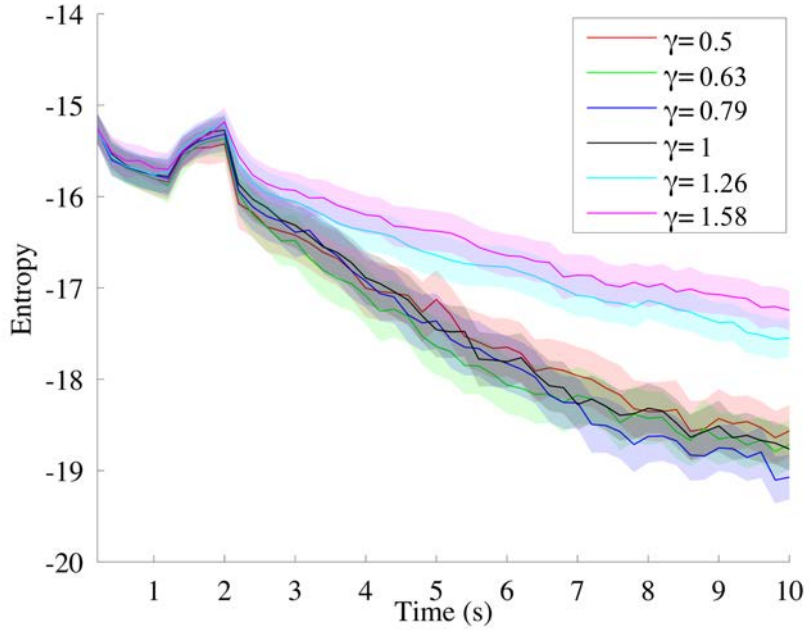

Figure 5.15: Average entropy and 95% confidence interval over time with different discount factor values for the entropy criterion.

5.4.5.1 Entropy criterion

Fig. 5.15 shows the average entropy over time with different discount factor values. From 2 s to 10 s, we see a significant difference between the entropy value of the two discount factors $\gamma > 1$ and the others. For the discount factors which are equal or smaller than one, the entropy value decreases drastically after 2 s. For the discount factors that are greater than one, the entropy value decreases more slowly.

However, we don't see those significant differences in Fig. 5.16 which shows the average estimation error over all experiments with different discount factors. This is due to the relation between the entropy criterion and the estimation error as presented in the previous section: a smaller entropy does not always result in a smaller estimation error.

Fig. 5.17 presents the error bars which show the average estimation error and 95% confidence interval for different discount factors at 1 s and 10 s with the entropy criterion. This figure is

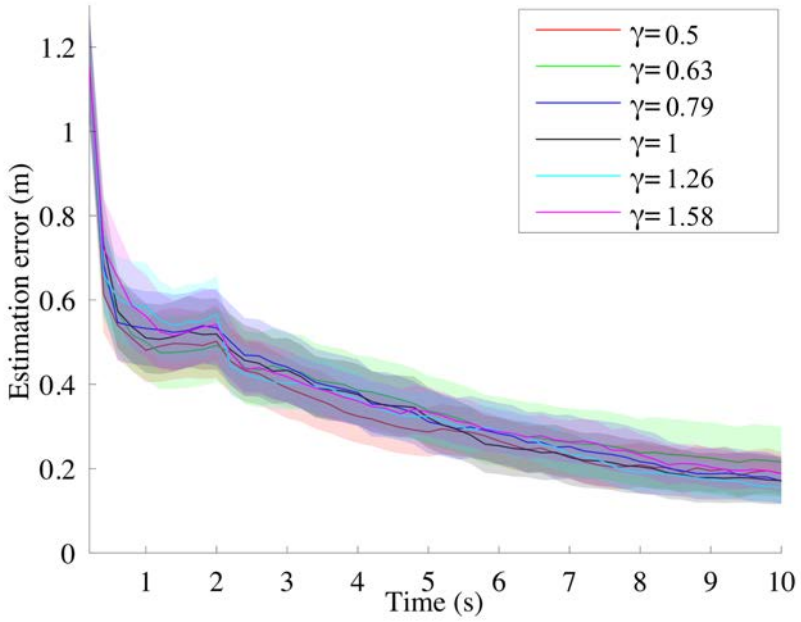

Figure 5.16: Average estimation error and 95% confidence interval over time with different discount factor values for the entropy criterion.

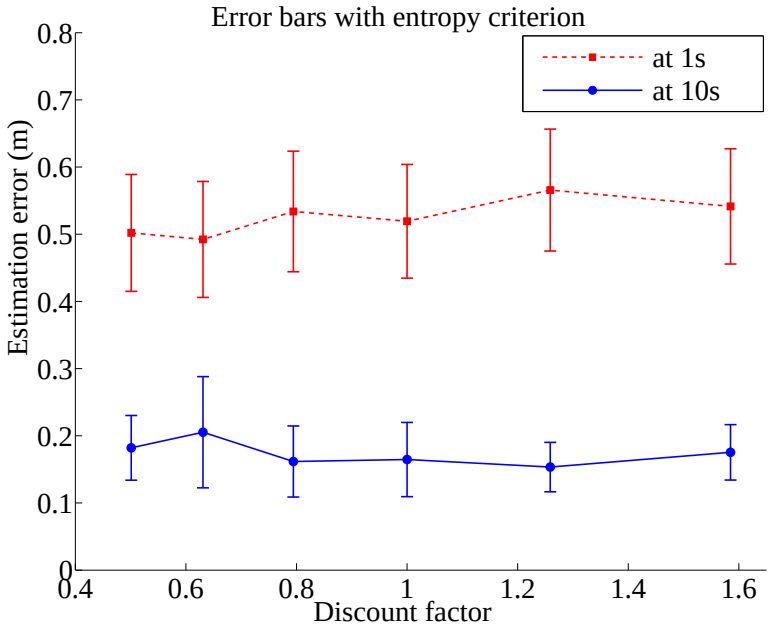

Figure 5.17: Error bars with the entropy criterion at $t = 1$ s and $t = 10$ s with different discount factor values.

extracted from Fig. 5.16, with a closer look at two time steps: 1 s and 10 s. The error bars for all discount factors at 10 s are smaller compared to those at 1 s. That is as desired with the MCTS approach. However, the effect of the discount factor on the estimation is not obvious because

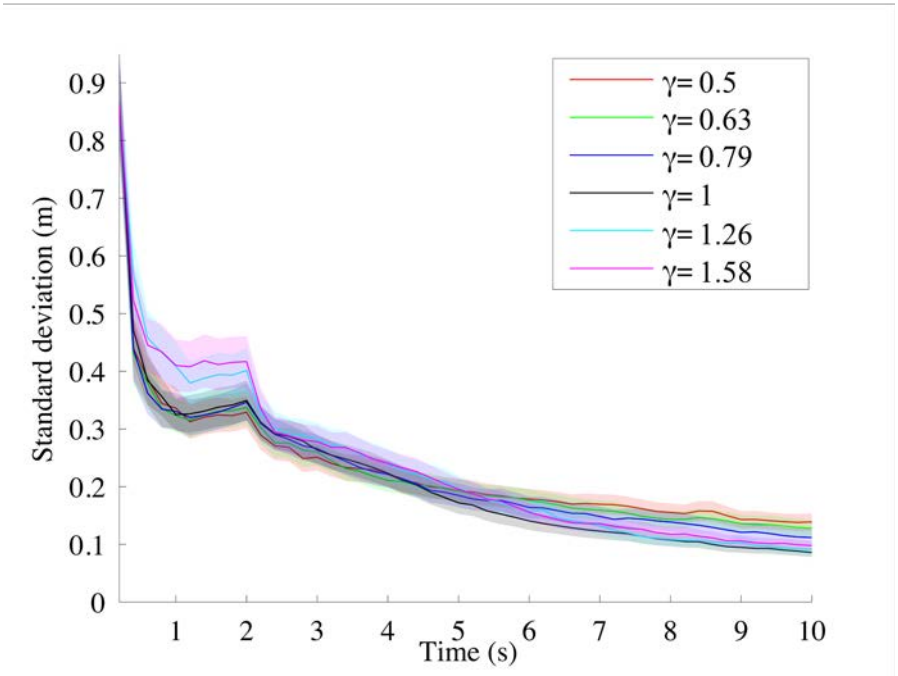

Figure 5.18: Average standard deviation and 95% confidence interval over time with different discount factor values for the standard deviation criterion.

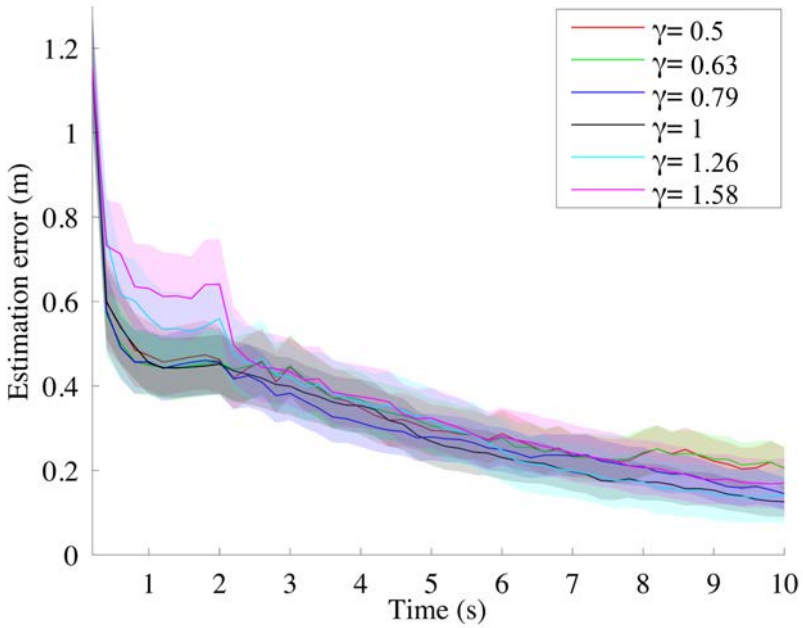

Figure 5.19: Average estimation error and 95% confidence interval over time with different discount factor values for the standard deviation criterion.

the error bars are not so different at a given time step.

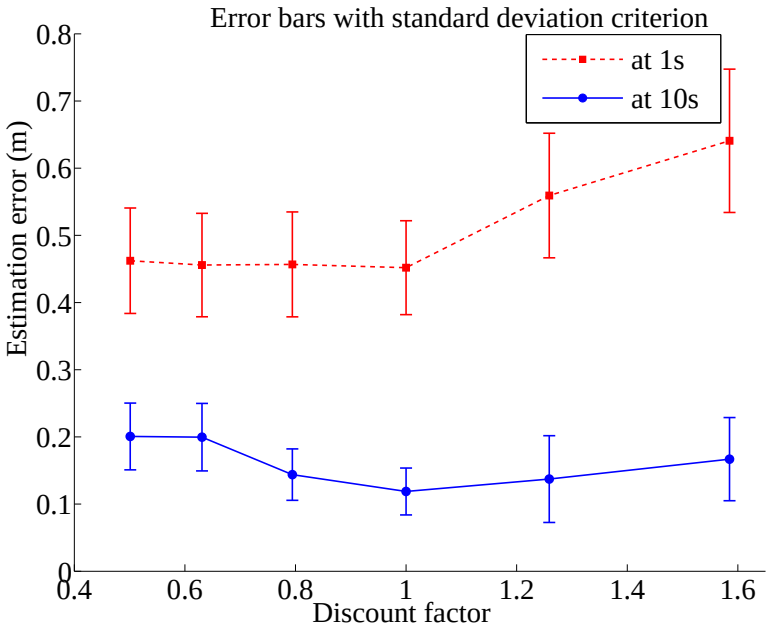

Figure 5.20: Error bars with the standard deviation criterion at $t = 1$ s and $t = 10$ s with different discount factor values.

5.4.5.2 Standard deviation criterion

Fig. 5.18 shows the average standard deviation over time and Fig. 5.19 shows the corresponding average estimation error over time in all experiments with different discount factors. We can see a similar shape between the two figures. After 2 s, both the values of standard deviation and estimation error decrease. For the discount factors that are greater than one, before 2 s, both the standard deviation and the estimation error are significant larger than those corresponding to the discount factors that are equal or smaller than one. However, after 2 s, their values decrease faster compared to those corresponding to the discount factors that are equal or smaller than one.

Fig. 5.20 depicts the error bars which show the average estimation error and 95% confidence interval for different discount factors at 1 s and 10 s with the standard deviation criterion. This figure is extracted from Fig. 5.19, with a closer look at two time steps: 1 s and 10 s. At 1 s, the two discount factors with $\gamma > 1$ have larger error bars at higher positions. This is reasonable because with smaller discount factor, the faster the decrease of the error. At 10 s, we have smaller error bars and lower error values compared to those at 1 s. We see that the two lowest discount values: $\gamma = 10^{-0.3}$ and $\gamma = 10^{-0.2}$ have higher error bars compared to the others. The smallest and lowest error bar is for the discount factor $\gamma = 10^{0}$ but it is not significantly different from the discount factors $\gamma = 10^{-0.1}$ and $\gamma > 1$.

5.5 Summary

We presented a long-term robot motion planning algorithm for estimating the location of an intermittent and possibly moving sound source. Our main theoretical contributions concern the cost function with two criteria: the standard deviation and the entropy of the estimated belief. In addition, we adapted a practical MCTS algorithm for finding an optimal sequence of robot

movements that will minimize the estimation uncertainty in the long run.

The experiments showed that our long-term planning algorithm achieves better performance compared to greedy or random motion. We also evaluated the performance of the MCTS approach with different discount factors. With the standard deviation criterion, the discount factor has a clear effect on the performance of MCTS over time. The analysis of all the results showed a coherent estimation error result when optimizing the standard deviation of the estimated belief. By contrast, optimizing the entropy of the estimated belief does not always guarantee minimum estimation error.

Chapter 6

Conclusion and perspectives

In this thesis, we presented several contributions related to the problem of mapping of a sound environment by using a mobile robot. We summarize our contributions and discuss the future perspectives of the thesis in the following sections.

6.1 Conclusion

This thesis addressed the problem of estimating the location of one or more sound sources using a mobile robot equipped with a microphone array. We proposed solutions specifically for the context of intermittent, moving or multiple sources.

We first considered the problem of locating a single intermittent and possibly moving source in Chapter 3. There is uncertainty in both the source AoA and SAD measurements. We proposed an extended MKF framework which jointly estimates the source location and its activity over time. This framework is applicable to any microphone array geometry. Thanks to the movement of the robot, it can solve the front-back ambiguity which appears in the case of a linear microphone array. We conducted a number of experiments to show the robustness of the extended MKF framework to false measurements of the AoA or the SAD when localizing an intermittent, moving source. When a false measurement of the source activity occurs, e.g., the source is inactive but SAD detects it as active, our extended MKF method can handle this situation but the extended MKF without activity model severely suffers from it. When a false measurement of the source AoA occurs, it has a major impact on the MKF without activity model, although the probability of false AoA measurements is low. By contrast, the estimation error of our MKF does not change much. The reason is that, when a false AoA measurement occurs, the weight of the hypotheses corresponding to an inactive source increases, so that the belief is little affected. In general, experiments and statistical results showed that our proposed MKF method outperforms the method that does not consider the source activity in the model.

In the last section of Chapter 3, we proposed a particle filtering technique for locating an intermittent source. Similarly, we jointly estimate the source location and its activity over time. We then compared the performance of the extended MKF with the particle filter in term of estimation error over time, number of components or particles used, and computational time. The average estimation error at the final time step is slightly smaller for the particle filter than for the extended MKF, however it is not significantly so. For the extended MKF, 50 components are good enough for estimation and the average computational time for each iteration is 1.7 s. To achieve similar performance, the particle filter will need about 600 particles and, as the result, the average computational time for each iteration would be 5.8 s.

Chapter 4 presented the extension of the work on source localization to the context of multiple sources. We proposed a filtering framework which jointly estimates the locations of two sound sources and their activities. These sources are intermittent and possibly moving. We make the assumption that the first observed AoA is corresponds to one of the two sound sources and the second observed AoA might correspond to the remaining source but can be a false alarm. In addition, we assume that the SAD will provide one source activity value for both sources. In the hypothesis that both sources are active, we do not know which of the two AoA measurements is caused by which source. There could be also the case when the false alarm happens, that is when one of the AoA measurements does not correspond to any source. Therefore, we defined joint association events of a measurement to a target source. By implementing an extended MKF with joint probabilistic data association, we can jointly estimate the two source locations and their activities over time. The experimental evaluation shows the effectiveness of the proposed method in localizing and tracking two target sources in a noisy and reverberant environment.

In Chapter 5, we focused on finding the optimal robot motion to improve the source localization result. The target source is intermittent and possibly moving. Our first contribution in this chapter was to define the cost function for long-term robot motion planning with two alternative criteria: the Shannon entropy or the standard deviation of the estimation belief on the source location. In the cost function, in order to investigate the tradeoff between short vs long term planning, we integrate these entropies or standard deviations over time with a discount factor. We represent the belief about the source location at each time step by a mixture of Gaussians, and this belief is propagated by the extended MKF framework proposed in Chapter 3. For long-term motion planning, the future observations are unknown at current time step, hence, we compute the cost function by considering the expectation over those future observations.

In the second contribution, we adapted the MCTS method for efficiently finding the optimal robot motion that will minimize the above cost function. We built a tree that contains possible future states which are associated with the tree's nodes. In the selection step of MCTS, we choose the UCT as the selection criterion to address the exploration-exploitation dilemma. We verified that the entropy and the standard deviation of the belief about the source location both satisfy the applicability condition of the UCT. Experiments showed that the proposed method outperforms other robot motion planning methods in the long run, especially compared to a greedy planning method. We evaluated the performance of the MCTS approach with different discount factors. The results of the MCTS with standard deviation highlighted the effect of the discount factor on the performance of MCTS over time. The analysis of all the results showed a coherent estimation error result when optimizing the standard deviation of the estimated belief instead of the entropy.

6.2 Perspectives

A number of directions built upon the proposed framework could be explored in future work. In the following, we present some future perspectives of this thesis.

Dealing with uncertainty on the robot's position

In this thesis, we assumed that the robot position is known. We could extend our extended MKF framework by considering uncertainty in the robot's position. The problem would become that of simultaneous localization and mapping (SLAM) for robot audition: We must to update the map of source locations and simultaneously keep track of the robot's location [Su et al., 2015a, Evers et al., 2016a, Su et al., 2015b]. We could use the state vector presented in Chapter 3. With the

state of the robot in the state vector, this model has the potential to deal with non-zero process noise of the robot motion. In addition, we would need to integrate additional observations informing about the robot's position in the observation vector besides the AoA observation and the SAD. These observations would also contain uncertainty in the measurement. A laser sensor could be used for this purpose, for instance.

Improving the motion planning technique for localizing a single source

We could improve the computational efficiency of the motion planning algorithm proposed in Chapter 5 by providing prior knowledge to the simulation and the selection step. This prior knowledge could be learned by running a number of simulations. From each simulation, we would compute the entropy or the standard deviation difference of the belief before running the simulation and after finishing the simulation. We would find the mean and variance of the distribution of this difference. When running the MCTS algorithm, in the simulation step, instead of doing the simulation as before (this takes time), we would sample a reward difference based on the distribution we learned. From that, we would reduce the time required to run the simulation in MCTS. In the selection step, we could evaluate and select an action by combining the prior knowledge with the current statistical value of the reward and the visit count.

Implementing on a real robot

The proposed algorithms in Chapters 3, 4, and 5 could be implemented on a real robot and assessed in a real environment. To implement them on a real robot, we would need to address ego noise generated from the robot motion and from the laser sensor. In the situation when the target source is far from the microphones, the SNR would be very low, so the AoA measurement would be severely affected by this ego noise. As the result, the estimation performance of the Bayesian filtering methods would also be affected. One solution would be to use a signal processing method to learn the features of the ego noise and suppress it from the recorded signal [Ince et al., 2010, Schmidt et al., 2016]. Another practical solution would be to use another device to obtain the position of robot instead of using a laser sensor, e.g., a depth camera and take AoA measurements only when the robot is not moving.

Robot motion planning for multiple sources

The motion planning algorithm in Chapter 5 could be extended to the context of multiple, intermittent, moving sound sources. We could estimate the belief on the source locations at each time step based on the sequential filtering technique presented in Chapter 4. In the context of multiple sound sources, the optimal motion of the robot must minimize the estimation error for not only one source but for all sources. Therefore, one solution could be to define the cost function for each source, and to solve the resulting multiobjective optimization problem. In this case, we could use Pareto optimality to find the solution that would minimize the cost function for each source [Tesch et al., 2013, Giagkiozis and Fleming, 2014]. Another solution could be to build only one cost function for joint estimation of the source locations. Then we could use the MCTS algorithm to find the optimal robot trajectory that would minimize the cost function.

Audio-visual tracking

We could track moving sound objects in a scene by performing audio-visual sensor fusion. The information from the visual sensor can complement the audio signal: It does not suffer from

noise and reverberation and can handle the situation when one or more sources are silent. The audio signal would provide useful information when the object is out of the field of view. The MKF framework could be used to fuse those two pieces of information to improve the localization performance [Zotkin et al., 2002, Gebru et al., 2017]. By also estimating the source activity, this framework could possibly deal with the situation when both sources are silent and out of the field of view.

Appendix A

Résumé en français

A.1 Introduction

Pour un robot d'assistance autonome, l'audition est une modalité nécessaire et importante ; on parle alors d'audition robotique [Nakadai et al., 2000, Nakadai et al., 2010, Okuno and Nakadai, 2015]. Ce terme regroupe de nombreux problèmes de traitement du signal, tels que la localisation de sources sonores, la séparation de sources, la reconnaissance automatique de la parole, la classification et l'identification de sons. À l'aide d'un robot mobile, ces derniers problèmes peuvent être mieux résolus en s'approchant de la source qu'en utilisant des techniques de traitement du signal pour les microphones fixes. Dans cette thèse, nous nous intéressons au problème de localisation de sources sonores par un robot mobile.

En raison du caractère bruité des mesures, détecter et localiser précisément une source de parole n'est pas facile. En plus du signal provenant en ligne droite de la source, le robot reçoit aussi le signal réverbéré par l'environnement ainsi que des bruits provenant d'autres sources. La réverbération et le bruit de fond dégradent grandement la performance de la localisation de sources sonores [Vermaak and Blake, 2001, Blandin et al., 2012] et de la détection d'activité vocale (*speech activity detection* ou SAD) [Ramírez et al., 2007, Zhang and Wu, 2013]. De plus, les sources sonores ne sont pas toujours actives, ce qui rend l'estimation encore plus difficile. Les silences et les transitions entre activité et silence augmentent l'incertitude sur la détection et la localisation de sources, et induisent des faux-positifs. Par conséquent, maîtriser l'incertitude est un défi-clé pour l'inférence d'informations précises à propos de la position des sources.

Le but principal de cette thèse est de s'attaquer à l'incertitude dans des environnements réverbérants, dans le contexte de sources sonores intermittentes, mobiles et multiples. Nous commençons par le problème de trouver la position d'une seule source intermittente et potentiellement mobile à l'aide d'un robot mobile dans un environnement réverbérant. Nous développons un modèle bayésien qui peut s'appliquer à n'importe quelle géométrie d'antenne de microphones pour estimer conjointement la position et l'activité d'une source. De plus, nous comparons le filtrage de Kalman étendu par mélange de gaussiennes (*extended mixture Kalman filter* ou MKF étendu) avec le filtrage particulaire en termes de performance de localisation et de temps de calcul. Nous étendons ensuite ce travail à plusieurs cibles intermittentes et mobiles. En implantant un MKF étendu avec association probabiliste des données (*Joint probabilistic data association filter* ou JPDAF), nous pouvons estimer conjointement les positions et les activités des deux sources dans le temps. Enfin, nous nous planifions le déplacement du robot pour réduire rapidement l'incertitude sur la position de la source sonore, mesurée par l'entropie de Shannon ou par l'écart-type de la croyance sur la position estimée. Nous adaptons alors

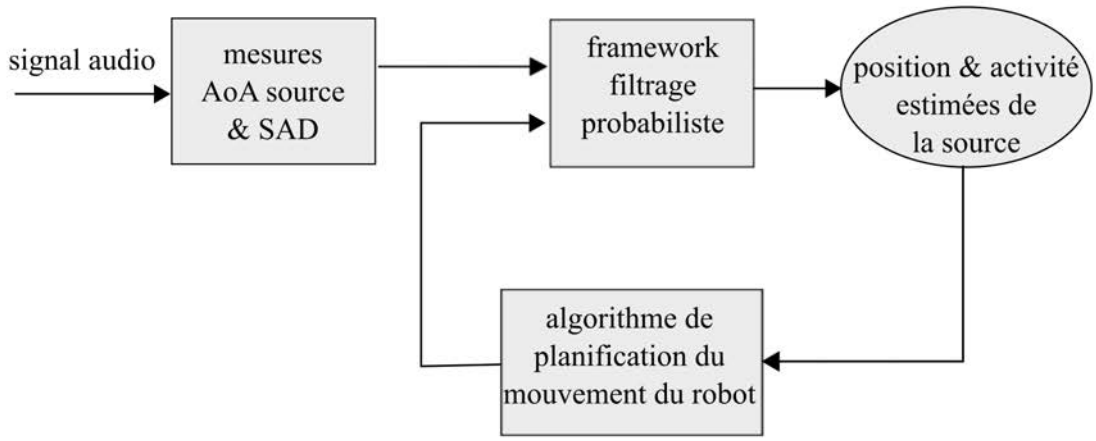

Figure A.1: Schéma général de localisation de sources sonores.

l'algorithme de recherche arborescente Monte Carlo (*Monte Carlo tree search* ou MCTS) pour trouver le mouvement optimal.

Le schéma général de localisation de sources sonores est présenté dans la Fig. A.1. Le signal audio est capturé par une antenne de microphones montée sur le robot. Ce signal peut être une source quelconque active pendant un temps suffisant, par exemple de la parole. Les mesures d'angle d'arrivée (*angle of arrival* ou AoA) et de SAD sont incertaines à cause de la réverbération et du bruit de fond. En fusionnant ces mesures avec le mouvement du robot dans un cadre de filtrage probabiliste, nous pouvons gérer l'incertitude des observations et estimer la position de la source et l'activité au cours du temps. La position estimée peut être améliorée en implantant un algorithme de planification du mouvement du robot. Les actions optimales sont sélectionnées et exécutées pour minimiser l'incertitude sur l'estimation.

A.2 État de l'art

Avec des microphones fixes, les méthodes classiques de localisation de source [DiBiase et al., 2001] estiment souvent simplement l'AoA. C'est le cas en champ lointain, lorsque la distance entre l'antenne et la source est plus grande que la taille de l'antenne. On peut ranger les méthodes existantes en trois classes principales [DiBiase et al., 2001]: celles exploitant les différences de temps d'arrivée entre paires de microphones [Knapp and Carter, 1976], celles basées sur la maximisation de la puissance du signal obtenu par formation de voies [Hahn and Tretter, 1973, Van Veen and Buckley, 1988, Johnson and Dudgeon, 1992], celles reposant sur l'analyse spectrale à haute résolution [Schmidt, 1986]. Parmi ces dernières se trouve la méthode appelée MUSIC-GSVD que nous utilisons par la suite. Des comparaisons expérimentales de ces algorithmes sont détaillées dans la littérature [DiBiase et al., 2001, Badali et al., 2009, Blandin et al., 2012].

Des robots équipés de microphones pour capter les signaux sonores ont été introduitsil y a assez longtemps [Kato et al., 1974, Kato et al., 1987]. Cependant ils ont été principalement utilisés pour reconnaître des commandes vocales simples. Dans les deux dernières décennies, l'audition a commencé à attirer plus d'attention en robotique et en traitement du signal [Hashimoto et al., 1997, Nakadai et al., 2000, Valin et al., 2007a, Nakadai et al., 2002, Okuno and Nakadai, 2015]. La plupart des travaux ont utilisé le robot d'une manière similaire à une antenne statique [Nakadai et al., 2000, Kim et al., 2008, Nakamura et al., 2012]. Quelques travaux se sont intéressés à

exploiter le mouvement du robot pour améliorer le traitement du signal, en particulier pour la localisation de sources [Valin et al., 2007b, Martinson and Schultz, 2009, Vincent et al., 2015]. Les robots fournissent une plateforme mobile pour relever l'AoA en continu avec plusieurs positions et orientations de l'antenne. En utilisant un algorithme de filtrage séquentiel pour intégrer le mouvement du robot avec les mesures d'AoA, il est possible de réduire l'incertitude sur l'AoA de la cible [Valin et al., 2007b, Portello et al., 2014, Vincent et al., 2015]. De plus, la distance à la source, indisponible pour une antenne fixe, peut être obtenue avec un robot.

Le suivi multi-cible a été appliqué pour suivre des sources sonores. [Gehrig and McDonough, 2006, Chakrabarty et al., 2014] ont utilisé l'algorithme JPDAF pour estimer les positions des sources. Un suivi multi-hypothèses (*multiple hypothesis tracking* ou MHT) a été implanté dans [Levy et al., 2011, Oualil et al., 2012], alors qu'une approche par ensemble fini aléatoire (*random finite set* ou RFS) est utilisée par [Evers et al., 2016b]. La croyance sur les positions des sources au cours du temps est souvent estimée par un MKF étendu [Chakrabarty et al., 2014, Oualil et al., 2012] ou un filtre à particules [Levy et al., 2011, Valin et al., 2007b, Pertilä and Hämäläinen, 2010]. Cependant la plupart des approches ci-dessus considèrent simplement le suivi multi-cibles avec une antenne fixe et les microphones sont souvent distribués dans tout l'espace [Gehrig and McDonough, 2006, Levy et al., 2011, Oualil et al., 2012]. De plus, l'estimation de l'activité de la source n'est pas incluse dans ces travaux.

La planification de mouvement pour l'audition robotique a commencé à émerger [Martinson and Schultz, 2009, Vincent et al., 2015, Schymura et al., 2017, Bustamante et al., 2017]. Une stratégie simple, consistant à suivre un circuit fixe couvrant une zone de manœuvre potentielle, permet d'améliorer la précision de la localisation [Martinson and Schultz, 2009]. Cette approche est sous-optimale parce qu'elle prend souvent beaucoup de temps pour trouver la zone potentielle, De plus, si la source est mobile, la zone détectée est susceptible de changer. Des stratégies de mouvement plus sophistiquées basée sur des critères de théorie de l'information, par exemple l'entropie de Shannon, ont été proposées [Sommerlade and Reid, 2008, Bustamante et al., 2017, Kumon et al., 2010]. L'idée générale est de conduire le robot dans la direction qui mène à l'incertitude minimale sur la position de la source. Une approche utilisant l'exploration de Monte Carlo pour échantillonner et sélectionner la prochaine action a été proposée [Schymura et al., 2017]. Dans cette approche gloutonne, le but final est prédéfini. Une méthode de planification à long terme a été introduite en utilisant la programmation dynamique pour trouver la trajectoire optimale pour localiser une source statique [Vincent et al., 2015]. Cette méthode approche la somme des entropies sur un horizon fini en faisait l'hypothèse que l'entropie à chaque pose future ne dépend pas du trajet utilisé pour atteindre cette pose.

A.3 Localisation d'une source en environnement réverbérant

Nous considérons dans un premier temps le cas d'une source unique, intermittente et mobile. La plupart des méthodes de la littérature considèrent que la source est toujours active, ou que l'information de SAD est parfaite [Portello et al., 2012, Portello et al., 2014]. À l'opposé, la méthode que nous proposons prend aussi en compte l'activité de la source dans le vecteur

d'état. Ainsi nous définissons le vecteur d'état de la manière suivante :

$$\begin{bmatrix} X \\ a \end{bmatrix} = \begin{bmatrix} X_\mathrm{r} \\ X_\mathrm{s} \\ a \end{bmatrix} = \begin{bmatrix} x_\mathrm{r} \\ y_\mathrm{r} \\ \theta_\mathrm{r} \\ x_\mathrm{s} \\ y_\mathrm{s} \\ \theta_\mathrm{s} \\ v_\mathrm{s} \\ w_\mathrm{s} \\ a \end{bmatrix}, \tag{A.1}$$

avec X_r la pose du robot, c'est-à-dire sa position absolue $[x_\mathrm{r}, y_\mathrm{r}]$ et son orientation θ_r par rapport à l'axe x ; X_s la partie continue de l'état de la source sonore, c'est-à-dire sa position absolue $[x_\mathrm{s}, y_\mathrm{s}]$, son orientation θ_s et ses vitesses linéaire et angulaire $[v_\mathrm{s}, w_\mathrm{s}]$; a l'activité de la source représentée par une variable binaire pour laquelle $a = 1$ indique que la source est active, sinon $a = 0$. Ainsi nous avons un vecteur d'état qui contient la variable continue X et la variable discrète a. Dans ce travail, nous supposons que la pose du robot X_r est connue et que nous avons simplement besoin d'estimer la variable continue X_s et l'activité a de la source.

Nous supposons que le vecteur d'observation Z_k à un instant k est constitué d'une mesure d'AoA Z_k^l obtenue par une technique de localisation et d'une mesure d'activité Z_k^a obtenue par une technique de SAD.

La vraisemblance du vecteur d'état par rapport à cette observation peut s'écrire

$$P(Z_k|X_k, a_k) = \begin{cases} P_\mathrm{sn}(Z_k^\mathrm{l}|X_k)P(Z_k^\mathrm{a}|a_k) & \text{pour } a_k = 1 \\ P_\mathrm{n}(Z_k^\mathrm{l})P(Z_k^\mathrm{a}|a_k) & \text{pour } a_k = 0, \end{cases} \tag{A.2}$$

avec P_sn et P_n les distributions respectives de l'AoA mesuré quand la source est active ou inactive. Dans ce dernier cas, il est supposé que le signal est constitué d'un bruit diffus et qu'ainsi P_n ne dépend pas de X_k.

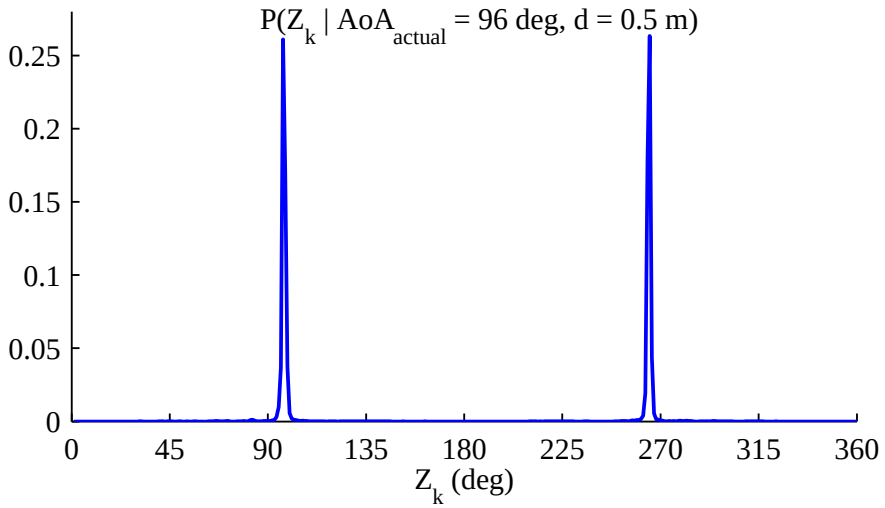

Figure A.2: Distribution de l'AoA mesuré quand la source réelle est à 96° et 0,5 m de l'antenne.

Un exemple de P_sn est montré dans la Fig. A.2 pour une antenne linéaire et la technique de localisation MUSIC-GSVD. La densité de probabilité se concentre autour de l'AoA réel et

son symétrique par rapport à l'axe de l'antenne : ce phénomène est appelé « ambigüité avant-arrière ». La densité de probabilité pour d'autres AoA est non nulle mais bien plus petite. Ainsi on peut approcher le modèle d'observation par un mélange de deux gaussiennes :

$$P_{\text{sn}}(Z_k^{\text{l}}|X_k) = \sum_{j=1}^{2} \frac{1}{2} \mathcal{N}(Z_k^{\text{l}}; \hat{Z}_k^{\text{l},j}, R_k^j). \tag{A.3}$$

Pour une antenne non-linéaire, le modèle d'observation pourrait être représenté par une seule gaussienne.

La distribution a posteriori du vecteur d'état peut être calculée récursivement en alternant ces deux étapes :

- **prédiction** : calcul de $P(X_k, a_k|Z_{1:k-1})$ à partir de l'estimation précédente $P(X_{k-1}, a_{k-1}|Z_{1:k-1})$ et du modèle de transition :

$$\begin{aligned} P(X_k, a_k|Z_{1:k-1}) &= \sum_{a_{k-1}} \int P(X_k, a_k|X_{k-1}, a_{k-1}) P(X_{k-1}, a_{k-1}|Z_{1:k-1}) dX_{k-1} \\ &= \sum_{a_{k-1}} \int P(a_k|a_{k-1}) P(X_k|X_{k-1}) P(X_{k-1}, a_{k-1}|Z_{1:k-1}) dX_{k-1}. \end{aligned} \tag{A.4}$$

- **observation** : calcul de $P(X_k, a_k|Z_{1:k})$ à partir du résultat de la prédiction et d'une nouvelle mesure Z_k :

$$P(X_k, a_k|Z_{1:k}) = \eta P(Z_k|X_k, a_k) P(X_k, a_k|Z_{1:k-1}), \tag{A.5}$$

avec η une constante de normalisation.

Comme le vecteur d'état inclut à la fois des variables continues et discrètes et que le modèle d'observation est un mélange de gaussiennes, nous proposons un MKF étendu et un filtre à particules. À l'instant k, la distribution sur l'état (X_k, a_k) est représentée soit par un mélange de gaussiennes soit par un ensemble d'échantillons de l'espace d'état (X^i, a^i) (particules) avec des poids correspondants w^i, $i = 1, \ldots, N$.

La Fig. A.3 montre les premières secondes du suivi d'une cible mobile intermittente à l'aide du MKF étendu. À l'instant $t = 0\,\text{s}$, le mélange est initialisé avec plusieurs composantes distribuées régulièrement dans la pièce de manière à approcher un a priori uniforme. Après 1 s, la moitié des hypothèses sur la position de la source sont distribuées dans la direction entre le robot et la source et l'autre moitié sont distribuées symétriquement par rapport à l'axe de l'antenne. Cette incertitude symétrique est due à l'ambigüité avant-arrière illustrée dans la Fig. 3.1. Ces hypothèses symétriques deviennent plus petites et disparaissent après 3 s grâce au mouvement du robot. Plus précisément, le mouvement de la source pour les composantes symétriques est plus important et moins cohérent et par conséquent moins probable que le mouvement des composantes correctes. Pour une antenne non-linéaire, la phase transitoire avec ces deux directions serait plus courte voire inexistante.

Nous avons conduit un certain nombre d'expériences pour montrer la robustesse de l'approche par MKF étendu aux observations erronées de l'AoA et de la SAD. Lorsqu'un faux positif arrive pour l'activité, c'est-à-dire que la source est inactive mais la SAD la détecte comme active, notre méthode peut gérer cette situation alors qu'un MKF étendu sans modélisation de l'activité en souffre. Lorsqu'une observation erronée de l'angle arrive, elle a des conséquences importantes sur le filtre sans modèle d'activité, même si la probabilité d'une telle observation est faible. À

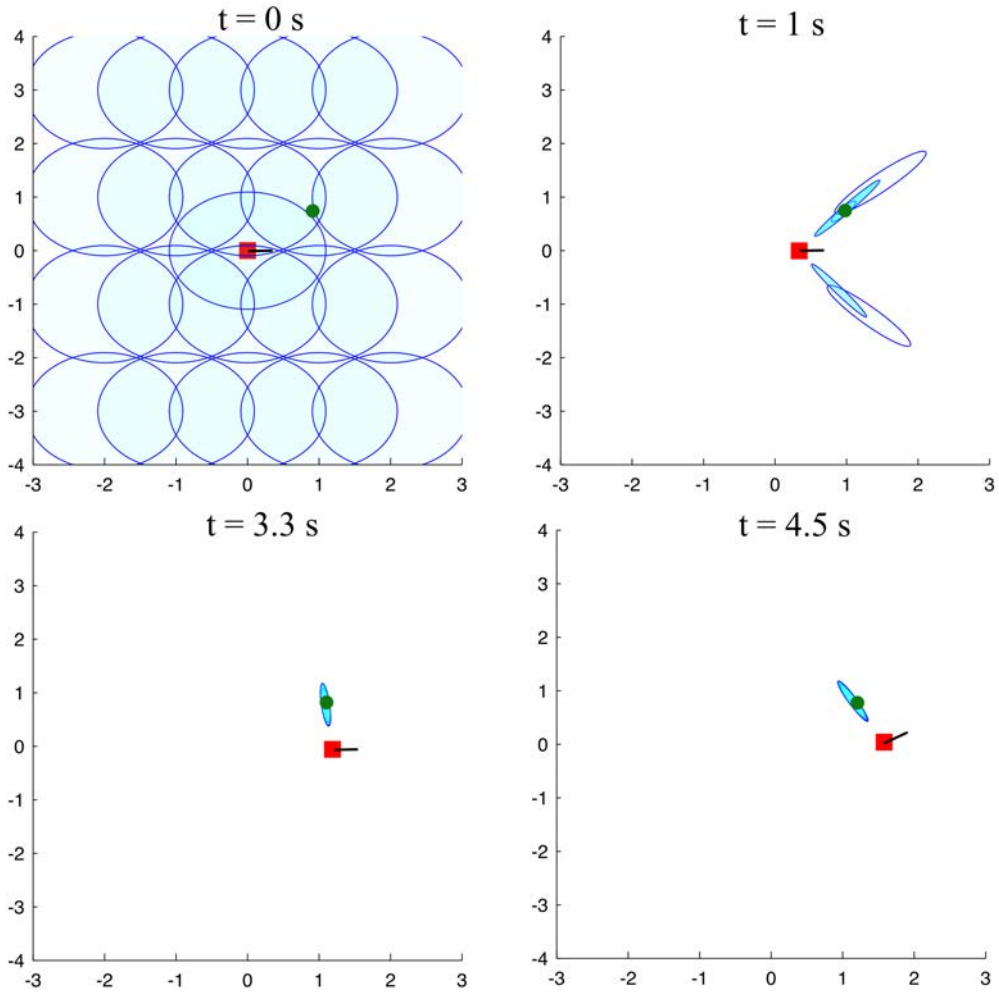

Figure A.3: Visualisation de notre MKF étendu sur un exemple. La position du robot est montrée par un carré rouge et la position réelle de la source par un disque vert. Les ellipses bleues représentent les régions de confiance à 95% des différentes composantes gaussiennes du mélange sur l'estimation de la position de la source, avec une transparence indiquant le poids de la composante.

l'opposé, l'erreur d'estimation de notre filtre avec activité ne change pas beaucoup. La raison en est que lorsqu'une observation erronée de l'angle arrive, le poids de l'hypothèse correspondant à une source inactive augmente de sorte que la distribution est moins affectée. De manière générale, les expériences et les résultats statistiques montrent que la méthode que nous proposons surpasse un filtre qui ne considère pas explicitement l'activité dans son modèle.

Nous comparons aussi la performance du MKF étendu avec un filtre à particules, en termes d'erreur d'estimation en fonction du temps, de nombre de composantes ou particules utilisées et de temps de calcul. L'erreur moyenne d'estimation à l'instant final est légèrement plus faible pour le filtre à particules que pour le MKF étendu, mais pas de manière significative. Pour le MKF étendu, 50 composantes suffisent pour l'estimation de la position de la source et le temps de calcul moyen pour chaque itération est de 1,7 s. Pour obtenir une performance similaire, le filtre à particule a besoin d'environ 600 particules et le temps de calcul moyen par itération est alors de 5,8 s.

A.4 Localisation de plusieurs sources

Dans un deuxième temps, nous considérons le problème du suivi de deux sources sonores mobiles et intermittentes à l'aide d'un robot mobile dans un environnement bruité et réverbérant. Contrairement à d'autres méthodes existantes de suivi multi-source [Gehrig and McDonough, 2006, Chakrabarty et al., 2014, Evers et al., 2016b], nous estimons conjointement la position et l'activité des deux sources au cours du temps. Nous définissons donc le vecteur d'état pour chaque source avec l'activité comme dans la partie A.3.

Dans le modèle d'observation, nous supposons que la SAD fournit une valeur d'activité unique, indiquant soit l'absence totale d'activité soit l'activité d'une source au moins. Nous supposons qu'il y a deux mesures d'angle d'arrivée. Chaque mesure d'angle j peut être générée par au plus une source i. Avec une antenne linéaire, la distribution de chaque mesure $Z^{\mathrm{l},j}$ sachant l'état X_k^i de la source i associée est bimodale lorsque la source est active et identique à celle de la section A.3. Lorsque la source n'est pas active, le modèle d'observation est uniforme. Nous supposons en outre que le premier AoA détecté correspond à une des sources actives et que le deuxième AoA détecté peut correspondre à l'autre source active ou à une fausse alarme. Nous faisons l'hypothèse que les mesures d'angle pour chaque source sont conditionnellement indépendantes les unes des autres et des fausses alarmes.

Dans le contexte de plusieurs sources sonores, il faut aborder le problème de l'association entre une mesure d'AoA et ka source qui peut en être l'origine. Ainsi, nous définissons un événement d'association β qui est un ensemble de paires constituées d'une mesure j et d'une cible i. Nous considérons aussi le cas d'une fausse alarme (false alarm (FA)) : la mesure $z_{l,k}^j$ ne correspond à aucune des sources suivies. La distribution a posteriori est mise à jour à partir de la probabilité d'association :

$$\begin{aligned} P(X_k, a_k|Z_{1:k}) =& \eta \sum_{\beta} P(Z_k|\beta, X_k, a_k) P(X_k, a_k|Z_{1:k-1}) P(\beta) \\ =& \eta \sum_{\beta} \prod_{i=1}^{n} P(X_k^i, a_k^i|Z_{1:k-1}) \prod_{j=1}^{m} P(Z_k^j|\beta, X_k, a_k) P(\beta), \end{aligned} \tag{A.6}$$

avec η une constante de normalisation.

Nous proposons un MKF étendu avec une version de l'algorithme JPDAF pour localiser deux sources mobiles et intermittentes. Nous construisons un modèle d'observation en simulant deux sources avec du bruit de fond et de la réverbération et nous apprenons la probabilité d'observer correctement chaque AoA. Avec deux sources actives, deux observations et en considérant les FA, nous pouvons avoir 4 associations différentes. L'équation A.6 peut alors se réécrire

$$\begin{aligned} P(X_k, a_k|Z_{1:k}) =& \eta \sum_{\beta} \prod_{i=1}^{2} P(X_k^i, a_k^i|Z_{1:k-1}) \prod_{j=1}^{2} P(Z_k^j|\beta, X_k, a_k) P(\beta) \\ =& \eta P(\beta_1) P(X_k^1, a_k^1|Z_{1:k-1}) P_{\mathrm{sn}}(Z_k^{l,1}|X_k^1) P(X_k^2, a_k^2|Z_{1:k-1}) P_{\mathrm{n}}(Z_k^{l,2}) P(Z_k^a|a_k) \\ &+ \eta P(\beta_2) P(X_k^1, a_k^1|Z_{1:k-1}) P_{\mathrm{n}}(Z_k^{l,2}) P(X_k^2, a_k^2|Z_{1:k-1}) P_{\mathrm{sn}}(Z_k^{l,1}|X_k^2) P(Z_k^a|a_k) \\ &+ \eta P(\beta_3) P(X_k^1, a_k^1|Z_{1:k-1}) P_{\mathrm{sn}}(Z_k^{l,1}|X_k^1) P(X_k^2, a_k^2|Z_{1:k-1}) P_{\mathrm{sn}}(Z_k^{l,2}|X_k^2) P(Z_k^a|a_k) \\ &+ \eta P(\beta_4) P(X_k^1, a_k^1|Z_{1:k-1}) P_{\mathrm{sn}}(Z_k^{l,2}|X_k^1) P(X_k^2, a_k^2|Z_{1:k-1}) P_{\mathrm{sn}}(Z_k^{l,1}|X_k^2) P(Z_k^a|a_k). \end{aligned} \tag{A.7}$$

Nous menons une évaluation expérimentale et montrons la capacité du modèle proposé à gérer l'incertitude et les fausses alarmes pour la localisation de deux sources sonores mobiles et intermittentes dans un environnement bruité et réverbérant.

A.5 Planification de mouvement pour l'audition

Dans un dernier temps, nous considérons la planification optimale de mouvement pour l'audition robotique. Il s'agit de trouver une séquence de mouvements du robot qui minimise l'erreur d'estimation sur la position d'une source mobile et intermittente. En pratique la position réelle de la source est inconnue. Nous quantifions donc l'incertitude sur l'estimation de la position de la source par deux critères : l'entropie de Shannon et l'écart-type de la croyance estimée. Les deux critères sont cumulés au cours le temps à l'aide d'un facteur d'actualisation. En minimisant l'incertitude, nous minimisons indirectement l'erreur d'estimation.

Supposons que le robot a déjà fait des observations jusqu'à un certain instant k. Toute la connaissance sur la position et l'activité de la source, ainsi que la pose du robot à l'instant k est représentée pas la distribution $P(X_k, a_k|Z_{1:k})$. Maintenant considérons le déplacement du robot vers une nouvelle pose à l'instant $k+1$. Pour trouver la mouvement optimal à l'instant $k+1$, il faut calculer l'entropie ou l'écart-type sachant la séquence de mouvement $u_{k+1:k+T}$ pour toutes les séquences possibles de mouvements jusqu'à un horizon $k+T$ fixé. Ces grandeurs ne peuvent pas être calculées de manière déterministes puisque les observations futures $Z_{k+1:k+T}$ ne sont pas connues, mais leurs espérances le sont. Le mouvement optimal est sélectionné de manière à minimiser ces fonctions de coût.

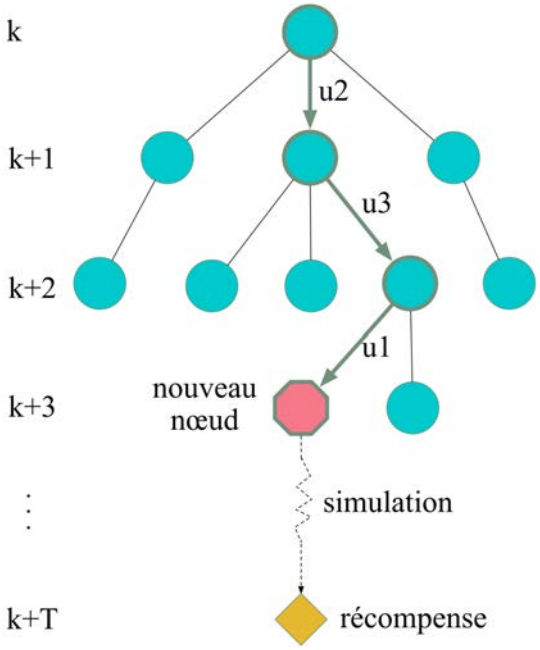

Figure A.4: Une itération de l'algorithme MCTS.

En pratique, considérer toutes les séquences possibles de mouvements $u_{k+1:k+T}$ est insurmontable. Nous proposons d'adapter l'algorithme MCTS pour résoudre ce problème. MCTS [Chaslot et al., 2008, Browne et al., 2012] est un algorithme en profondeur d'abord guidé par les résultats de simulations stochastiques pour trouver une action optimale à partir d'un état racine. Il est beaucoup employé dans les jeux et a révolutionné le jeu de Go [Silver et al., 2016]. MCTS est en train de remplacer des algorithmes de recherche traditionnels comme méthode par défaut dans les domaines difficiles. Dans notre cas, nous définissons la récompense comme fonction de l'entropie ou de l'écart-type de la distribution sur l'état.

La Fig. A.4 montre une itération de l'algorithme MCTS à un instant k. Chaque niveau

de l'arbre correspond à un instant futur. Chaque nœud n contient l'information sur : la pose du robot, la distribution $b(n)$ sur la position de la source, les actions pas encore essayées dans un ensemble fini prédéterminé, la récompense accumulée $\bar{Q}(n)$ (voir ci-dessous) et le compteur de visites $N(n)$. Le nœud racine n_0 représente la pose du robot à l'instant k et contient la distribution estimée $P(X_k, a_k|Z_{1:k})$. Les liens entre un nœud et ses enfants sont les différentes actions possibles.

À partir de la racine, un arbre est construit itérativement en sélectionnant un nœud, en ajoutant un enfant correspondant à une des actions non essayées de ce nœud, et en suivant ensuite une trajectoire aléatoire du robot depuis cet enfant jusqu'à l'instant $k+T$. L'opposé de l'entropie ou l'écart-type correspondant aux distributions calculées le long de cette trajectoire est propagé en remontant l'arbre pour mettre à jour la récompense accumulée et le compteur de visites. Les itérations de construction de l'arbre s'arrêtent à l'épuisement du temps de calcul imparti. La pose optimale à l'instant $k+1$ est alors choisie à partir de la récompense moyenne évaluée pour cette pose. Le robot applique alors l'action correspondante pour se déplacer jusqu'à cette pose, prend une nouvelle mesure Z_{k+1}, construit un nouvel arbre pour décider de la prochaine pose, etc.

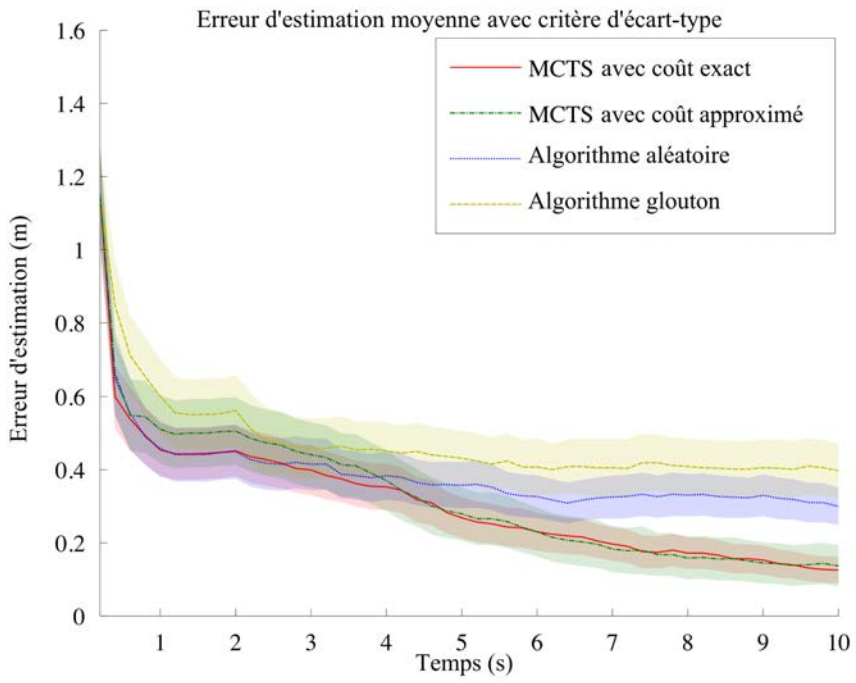

Figure A.5: Erreur d'estimation moyenne et intervalle de confiance à 95% en fonction du temps pour les quatre algorithmes avec l'écart-type comme critère calculés à l'aide de 200 expériences.

Nous conduisons un grand nombre d'expériences pour obtenir des résultats significatifs. Les expériences montrent que la méthode proposée surpasse les autres méthodes de planification de mouvement du robot à long terme. L'erreur d'estimation moyenne avec le critère d'écart-type est présentée pour tous les algorithmes dans la Fig. 5.12. À long terme, l'erreur moyenne des deux méthodes MCTS est plus faible comparée à une méthode gloutonne ou un mouvement aléatoire. De plus, l'erreur de toutes les méthodes ne change pas durant les intervalles de silence de la source si nous utilisons le MKF pour l'estimation. Nous évaluons aussi les performances

de notre approche MCTS avec différents facteurs d'actualisation. Les résultats de MCTS avec l'écart-type montrent l'importance de ce facteur sur le long terme. L'analyse de tous les résultats montre une erreur d'estimation plus cohérente en utilisant l'écart-type comme critère au lieu de l'entropie.

A.6 Conclusion et perspectives

Dans cette thèse, nous avons proposé des contributions sur la localisation d'une source sonore intermittente et mobile à l'aide de modèles de filtrage bayésien estimant conjointement la position et l'activité de la source. Nous en avons montré les avantages dans un environnement réverbérant dans lequel de fausses observations de l'angle d'arrivée et de l'activité peuvent arriver. Nous avons étendu ces modèles à plusieurs sources intermittentes et mobiles. Les expériences avec un MKF étendu avec JPDAF montrent la capacité de suivre deux sources. De plus, nous avons proposé une autre contribution sur une technique de planification à long terme en adaptant l'algorithme MCTS pour trouver le mouvement optimal du robot minimisant l'incertitude de l'estimation. De nombreuses expériences ont permis de montrer que cet algorithme présente une meilleure performance que d'autres algorithmes, par exemple glouton ou aléatoire.

Plusieurs directions peuvent être explorées pour étendre ces modèles. Dans un travail futur, nous pourrions nous intéresser à l'incertitude sur la position du robot et résoudre le problème de localisation et cartographie simultanées (*simultaneous localization and mapping* ou SLAM) en exploitant l'audition. Une autre direction est d'améliorer l'efficacité, en terme de temps de calcul, de notre algorithme de planification en fournissant des informations préalables aux étapes de simulation et sélection. Cette technique de planification peut aussi être étendue au contexte multi-source. Une des autres perspectives serait de suivre des sources sonores mobiles en fusionnant des informations audio et vidéo. Le MKF pourrait être utilisé pour la fusion de ces informations pour améliorer la performance de localisation et gérer les situations où les deux sources sont silencieuses ou hors du champ de vision. Tous les algorithmes proposés dans cette thèse peuvent aussi être implantés sur un robot réel pour être testés dans un environnement réel. Pour ce faire, il faudrait régler le problème du bruit propre généré par le robot (ses capteurs) et son mouvement.

Bibliography

[Agin, 1979] Agin, G. J. (1979). Real time control of a robot with a mobile camera. Technical Report 179, AI Center, SRI International.

[Alam et al., 2014] Alam, J., Kenny, P., Ouellet, P., Stafylakis, T., and Dumouchel, P. (2014). Supervised/unsupervised voice activity detectors for text-dependent speaker recognition on the RSR2015 corpus. In *Proc. Odyssey*.

[Amanatiadis et al., 2013] Amanatiadis, A. A., Chatzichristofis, S. A., Charalampous, K., Doitsidis, L., Kosmatopoulos, E. B., Tsalides, P., Gasteratos, A., and Roumeliotis, S. I. (2013). A multi-objective exploration strategy for mobile robots under operational constraints. *IEEE Access*, 1:691–702.

[Araya-López et al., 2010] Araya-López, M., Buffet, O., Thomas, V., and Charpillet, F. (2010). A POMDP extension with belief-dependent rewards. In *Advances in Neural Information Processing Systems 23 (NIPS-10)*.

[Arulampalam et al., 2002] Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188.

[Asano et al., 2008] Asano, F., Kimura, M., Shibuya, D., and Kamitani, Y. (2008). Localization and extraction of brain activity using generalized eigenvalue decomposition. In *Proc. ICASSP*, pages 565–568.

[Auer et al., 2002] Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.

[Badali et al., 2009] Badali, A., Valin, J.-M., Michaud, F., and Aarabi, P. (2009). Evaluating real-time audio localization algorithms for artificial audition in robotics. In *Proc. IROS*, pages 2033–2038.

[Bar-Shalom, 1990] Bar-Shalom, Y. (1990). *Multitarget-multisensor tracking: advanced applications*. Artech House.

[Berglund and Sitte, 2005] Berglund, E. and Sitte, J. (2005). Sound source localisation through active audition. In *Proc. IROS*, pages 509–514.

[Blackman and Popoli, 1999] Blackman, S. and Popoli, R. (1999). *Design and analysis of modern tracking systems*. Artech House.

[Blackman, 2004] Blackman, S. S. (2004). Multiple hypothesis tracking for multiple target tracking. *IEEE Aerospace and Electronic Systems Magazine*, 19(1):5–18.

[Blandin et al., 2012] Blandin, C., Ozerov, A., and Vincent, E. (2012). Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. *Signal Processing*, 92(8):1950–1960.

[Boor et al., 1999] Boor, V., Overmars, M. H., and van der Stappen, A. F. (1999). The Gaussian sampling strategy for probabilistic roadmap planners. In *Proc. ICRA*, volume 2, pages 1018–1023.

[Brandstein and Silverman, 1997] Brandstein, M. S. and Silverman, H. F. (1997). A robust method for speech signal time-delay estimation in reverberant rooms. In *Proc. ICASSP*, volume 1, pages 375–378.

[Browne et al., 2012] Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., et al. (2012). A survey of Monte Carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43.

[Brutti and Nesta, 2013] Brutti, A. and Nesta, F. (2013). Tracking of multidimensional TDOA for multiple sources with distributed microphone pairs. *Computer Speech and Language*, 27(3):660–682.

[Bustamante and Danès, 2017] Bustamante, G. and Danès, P. (2017). Multi-step-ahead information-based feedback control for active binaural localization. In *Proc. IROS*.

[Bustamante et al., 2016] Bustamante, G., Danès, P., Forgue, T., and Podlubne, A. (2016). Towards information-based feedback control for binaural active localization. In *Proc. ICASSP*, pages 6325–6329.

[Bustamante et al., 2017] Bustamante, G., Danès, P., Forgue, T., Podlubne, A., and Manhès, J. (2017). An information based feedback control for audio-motor binaural localization. *Autonomous Robots*.

[Chakrabarty et al., 2014] Chakrabarty, S., Kowalczyk, K., Taseska, M., and Habets, E. A. (2014). Extended Kalman filter with probabilistic data association for multiple non-concurrent speaker localization in reverberant environments. In *Proc. ICASSP*, pages 7445–7449.

[Champagne et al., 1996] Champagne, B., Bédard, S., and Stéphenne, A. (1996). Performance of time-delay estimation in the presence of room reverberation. *IEEE Transactions on Speech and Audio Processing*, 4(2):148–152.

[Chaslot et al., 2008] Chaslot, G., Bakkes, S., Szita, I., and Spronck, P. (2008). Monte-Carlo tree search: A new framework for game AI. In *Proc. AIIDE*.

[Chaumette and Hutchinson, 2006] Chaumette, F. and Hutchinson, S. (2006). Visual servo control. I. Basic approaches. *IEEE Robotics & Automation Magazine*, 13(4):82–90.

[Chengalvarayan, 1999] Chengalvarayan, R. (1999). Robust energy normalization using speech/nonspeech discriminator for German connected digit recognition. In *Proc. Eurospeech*.

[Colas et al., 2013] Colas, F., Mahesh, S., Pomerleau, F., Liu, M., and Siegwart, R. (2013). 3D path planning and execution for search and rescue ground robots. In *Proc. IROS*, pages 722–727.

[Cox, 1993] Cox, I. J. (1993). A review of statistical data association techniques for motion correspondence. *International Journal of Computer Vision*, 10(1):53–66.

[DeJong, 2012] DeJong, B. P. (2012). Auditory occupancy grids with a mobile robot. *Journal of Automation, Mobile Robotics & Intelligent Systems*, 6(3).

[DiBiase, 2000] DiBiase, J. H. (2000). *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. PhD thesis, Brown University.

[DiBiase et al., 2001] DiBiase, J. H., Silverman, H. F., and Brandstein, M. S. (2001). Robust localisation in reverberant rooms. In *Microphone Arrays: Signal Processing Techniques and Applications*. Springer.

[Dijkstra, 1959] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271.

[Dmochowski et al., 2007] Dmochowski, J. P., Benesty, J., and Affes, S. (2007). A generalized steered response power method for computationally viable source localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2510–2526.

[Doucet et al., 2001] Doucet, A., De Freitas, N., and Gordon, N. (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer.

[Doucet et al., 2000] Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208.

[Duan et al., 2016] Duan, J., Ren, L., Li, L., and Liu, D. (2016). Moving objects detection in evidential occupancy grids using laser radar. In *Proc. IHMSC*, volume 02, pages 73–76.

[Espiau et al., 1992] Espiau, B., Chaumette, F., and Rives, P. (1992). A new approach to visual servoing in robotics. *IEEE Transactions on Robotics and Automation*, 8(3):313–326.

[Evers et al., 2016a] Evers, C., Moore, A. H., and Naylor, P. A. (2016a). Acoustic simultaneous localization and mapping (a-SLAM) of a moving microphone array and its surrounding speakers. In *Proc. ICASSP*, pages 6–10.

[Evers et al., 2016b] Evers, C., Moore, A. H., and Naylor, P. A. (2016b). Localization of moving microphone arrays from moving sound sources for robot audition. In *Proc. EUSIPCO*, pages 1008–1012.

[Fallon and Godsill, 2012] Fallon, M. F. and Godsill, S. J. (2012). Acoustic source localization and tracking of a time-varying number of speakers. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1409–1415.

[Flanagan et al., 1985] Flanagan, J., Johnston, J., Zahn, R., and Elko, G. (1985). Computer-steered microphone arrays for sound transduction in large rooms. *The Journal of the Acoustical Society of America*, 78(5):1508–1518.

[Gebru et al., 2017] Gebru, I. D., Evers, C., Naylor, P. A., and Horaud, R. (2017). Audio-visual tracking by density approximation in a sequential Bayesian filtering framework. In *Proc. HSCMA*, pages 71–75.

[Gehrig and McDonough, 2006] Gehrig, T. and McDonough, J. (2006). Tracking multiple speakers with probabilistic data association filters. In *International Evaluation Workshop on Classification of Events, Activities and Relationships*, pages 137–150.

[Germain et al., 2013] Germain, F. G., Sun, D. L., and Mysore, G. J. (2013). Speaker and noise independent voice activity detection. In *Proc. Interspeech*.

[Giagkiozis and Fleming, 2014] Giagkiozis, I. and Fleming, P. J. (2014). Pareto front estimation for decision making. *Evolutionary Computation*, 22(4):651–678.

[Gonzalez-Banos and Latombe, 2002] Gonzalez-Banos, H. H. and Latombe, J.-C. (2002). Navigation strategies for exploring indoor environments. *The International Journal of Robotics Research*, 21(10-11):829–848.

[Gordon et al., 1993] Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113.

[Goutsias et al., 2012] Goutsias, J., Mahler, R. P., and Nguyen, H. T. (2012). *Random sets: theory and applications*, volume 97. Springer Science & Business Media.

[Hahn and Tretter, 1973] Hahn, W. and Tretter, S. (1973). Optimum processing for delay-vector estimation in passive signal arrays. *IEEE Transactions on Information Theory*, 19(5):608–614.

[Hart et al., 1968] Hart, P. E., Nilsson, N. J., and Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107.

[Hashimoto et al., 1997] Hashimoto, S., Narita, S., Kasahara, H., Takanishi, A., Sugano, S., Shirai, K., Kobayashi, T., Takanobu, H., Kurata, T., Fujiwara, K., Matsuno, T., Kawasaki, T., and Hoashi, K. (1997). Humanoid robot-development of an information assistant robot hadaly. In *Proc. RO-MAN*, pages 106–111.

[Hoeffding, 1963] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

[Hsiao et al., 2007] Hsiao, K., Kaelbling, L. P., and Lozano-Perez, T. (2007). Grasping POMDPs. In *Proc. ICRA*, pages 4685–4692.

[Huber et al., 2008] Huber, M. F., Bailey, T., Durrant-Whyte, H., and Hanebeck, U. D. (2008). On entropy approximation for Gaussian mixture random vectors. In *Proc. MFI*, pages 181–188.

[Hutchinson et al., 1996] Hutchinson, S., Hager, G. D., and Corke, P. I. (1996). A tutorial on visual servo control. *IEEE Transactions on Robotics and Automation*, 12(5):651–670.

[Ince et al., 2010] Ince, G., Nakadai, K., Rodemann, T., Tsujino, H., and Imura, J.-I. (2010). Robust ego noise suppression of a robot. *Trends in Applied Intelligent Systems*, pages 62–71.

[Jazwinski, 1970] Jazwinski, A. H. (1970). *Stochastic processes and filtering theory*. Academic Press.

[Johnson and Dudgeon, 1992] Johnson, D. H. and Dudgeon, D. E. (1992). *Array signal processing: concepts and techniques*. Simon & Schuster.

[Julier and Uhlmann, 2004] Julier, S. J. and Uhlmann, J. K. (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422.

[Kaelbling et al., 1998] Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134.

[Karray and Martin, 2003] Karray, L. and Martin, A. (2003). Towards improving speech detection robustness for speech recognition in adverse conditions. *Speech Communication*, 40(3):261–276.

[Katehakis and Veinott Jr, 1987] Katehakis, M. N. and Veinott Jr, A. F. (1987). The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research*, 12(2):262–268.

[Kato et al., 1974] Kato, I., Ohteru, S., Kobayashi, H., Shirai, K., and Uchiyama, A. (1974). Information-power machine with senses and limbs. In *On Theory and Practice of Robots and Manipulators*, pages 11–24. Springer.

[Kato et al., 1987] Kato, I., Ohteru, S., Shirai, K., Matsushima, T., Narita, S., Sugano, S., Kobayashi, T., and Fujisawa, E. (1987). The robot musician 'WABOT-2' (WAseda roBOT-2). *Robotics*, 3(2):143–155.

[Khatib, 1986] Khatib, O. (1986). Real-time obstacle avoidance for manipulators and mobile robots. *The International Journal of Robotics Research*, 5(1):90–98.

[Kim et al., 2008] Kim, U.-H., Kim, J., Kim, D., Kim, H., and You, B.-J. (2008). Speaker localization using the TDOA-based feature matrix for a humanoid robot. In *Proc. RO-MAN*, pages 610–615.

[Kitagawa, 1996] Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25.

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

[Kocsis et al., 2006] Kocsis, L., Szepesvári, C., and Willemson, J. (2006). Improved Monte-Carlo search. Technical Report 1, University of Tartu, Estonia.

[Kumon et al., 2010] Kumon, M., Fukushima, K., Kunimatsu, S., and Ishitobi, M. (2010). Motion planning based on simultaneous perturbation stochastic approximation for mobile auditory robots. In *Proc. IROS*, pages 431–436.

[Latombe, 1991] Latombe, J.-C. (1991). *Robot Motion Planning*. Kluwer Academic Publishers.

[LaValle, 2006] LaValle, S. M. (2006). *Planning algorithms*. Cambridge University Press.

[Lee et al., 2010] Lee, Y., Wada, T. S., and Juang, B.-H. (2010). Multiple acoustic source localization based on multiple hypotheses testing using particle approach. In *Proc. ICASSP*, pages 2722–2725.

[Lehmann and Williamson, 2006] Lehmann, E. A. and Williamson, R. C. (2006). Particle filter design using importance sampling for acoustic source localization and tracking in reverberant environment. *EURASIP Journal on Advances in Signal Processing*, 2006:017021.

[Levy et al., 2011] Levy, A., Gannot, S., and Habets, E. A. (2011). Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1540–1555.

[Liu and Chen, 1998] Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044.

[Lu and Cooke, 2011] Lu, Y.-C. and Cooke, M. (2011). Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners. *Speech Communication*, 53(5):622–642.

[Magassouba, 2016] Magassouba, A. (2016). *Aural servo: towards an alternative approach to sound localization for robot motion control*. PhD thesis, Université Rennes 1.

[Mahler, 2007] Mahler, R. P. (2007). *Statistical multisource-multitarget information fusion*. Artech House.

[Marković et al., 2013] Marković, I., Portello, A., Danès, P., Petrović, I., and Argentieri, S. (2013). Active speaker localization with circular likelihoods and bootstrap filtering. In *Proc. IROS*, pages 2914–2920.

[Martinson and Schultz, 2006] Martinson, E. and Schultz, A. (2006). Auditory evidence grids. In *Proc. IROS*, pages 1139–1144.

[Martinson and Schultz, 2009] Martinson, E. and Schultz, A. (2009). Discovery of sound sources by an autonomous mobile robot. *Autonomous Robots*, 27:221–237.

[Marzinzik and Kollmeier, 2002] Marzinzik, M. and Kollmeier, B. (2002). Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Transactions on Speech and Audio Processing*, 10(2):109–118.

[Meyer and Filliat, 2003] Meyer, J.-A. and Filliat, D. (2003). Map-based navigation in mobile robots: II. A review of map-learning and path-planning strategies. *Cognitive Systems Research*, 4(4):283–317.

[Nakadai et al., 2000] Nakadai, K., Lourens, T., Okuno, H. G., and Kitano, H. (2000). Active audition for humanoid. In *Proc. AAAI*, pages 832–839.

[Nakadai et al., 2002] Nakadai, K., Okuno, H. G., and Kitano, H. (2002). Real-time sound source localization and separation for robot audition. In *Proc. Interspeech*, pages 193–196.

[Nakadai et al., 2003] Nakadai, K., Okuno, H. G., and Kitano, H. (2003). Robot recognizes three simultaneous speech by active audition. In *Proc. ICRA*, pages 398–405.

[Nakadai et al., 2010] Nakadai, K., Takahashi, T., Okuno, H. G., Nakajima, H., Hasegawa, Y., and Tsujino, H. (2010). Design and implementation of robot audition system 'HARK' — open source software for listening to three simultaneous speakers. *Advanced Robotics*, 24(5–6):739–761.

[Nakamura et al., 2009] Nakamura, K., Nakadai, K., Asano, F., Hasegawa, Y., and Tsujino, H. (2009). Intelligent sound source localization for dynamic environments. In *Proc. IROS*, pages 664–669.

[Nakamura et al., 2012] Nakamura, K., Nakadai, K., and Ince, G. (2012). Real-time super-resolution sound source localization for robots. In *Proc. IROS*, pages 694–699.

[Nguyen et al., 2016] Nguyen, Q. V., Colas, F., Vincent, E., and Charpillet, F. (2016). Localizing an intermittent and moving sound source using a mobile robot. In *Proc. IROS*.

[Nguyen et al., 2017] Nguyen, Q. V., Colas, F., Vincent, E., and Charpillet, F. (2017). Long-term robot motion planning for active sound source localization with Monte Carlo tree search. In *Proc. HSCMA*.

[Nilsson, 1969] Nilsson, N. J. (1969). A mobile automaton: An application of artificial intelligence techniques. Technical report, DTIC Document.

[Okuno and Nakadai, 2015] Okuno, H. and Nakadai, K. (2015). Robot audition: Its rise and perspectives. In *Proc. ICASSP*, pages 5610–5614.

[Omologo and Svaizer, 1996] Omologo, M. and Svaizer, P. (1996). Acoustic source location in noisy and reverberant environment using CSP analysis. In *Proc. ICASSP*, volume 2, pages 921–924.

[Oualil et al., 2012] Oualil, Y., Faubel, F., and Klakow, D. (2012). A multiple hypothesis Gaussian mixture filter for acoustic source localization and tracking. In *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*, pages 1–4.

[Panchea et al., 2017] Panchea, A., Chapoutot, A., and Filliat, D. (2017). Extended reliable robust motion planners.

[Pertilä and Hämäläinen, 2010] Pertilä, P. and Hämäläinen, M. S. (2010). A track before detect approach for sequential Bayesian tracking of multiple speech sources. In *Proc. ICASSP*, pages 4974–4977.

[Popoviciu, 1935] Popoviciu, T. (1935). Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica (Cluj)*.

[Portello et al., 2014] Portello, A., Bustamante, G., Danès, P., Piat, J., and Manhès, J. (2014). Active localization of an intermittent sound source from a moving binaural sensor. In *Proc. Forum Acusticum*.

[Portello et al., 2011] Portello, A., Danès, P., and Argentieri, S. (2011). Acoustic models and Kalman filtering strategies for active binaural sound localization. In *Proc. IROS*, pages 137–142.

[Portello et al., 2012] Portello, A., Danès, P., and Argentieri, S. (2012). Active binaural localization of intermittent moving sources in the presence of false measurements. In *Proc. IROS*, pages 3294–3299.

[Ramírez et al., 2007] Ramírez, J., Górriz, J. M., and Segura, J. C. (2007). Voice activity detection. Fundamentals and speech recognition system robustness. In *Robust Speech Recognition and Understanding*.

[Ramirez et al., 2003] Ramirez, J., Segura, J. C., Benitez, C., Torre, A. d. l., and Rubio, A. J. (2003). A new adaptive long-term spectral estimation voice activity detector. In *Proc. Eurospeech*.

[Rayleigh, 1907] Rayleigh, L. (1907). On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232.

[Reid, 1979] Reid, D. (1979). An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6):843–854.

[Schmidt et al., 2016] Schmidt, A., Deleforge, A., and Kellermann, W. (2016). Ego-noise reduction using a motor data-guided multichannel dictionary. In *Proc. IROS*, pages 1281–1286.

[Schmidt, 1986] Schmidt, R. (1986). Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280.

[Schymura et al., 2017] Schymura, C., Grajales, J. D. R., and Kolossa, D. (2017). Monte Carlo exploration for active binaural localization. In *Proc. ICASSP*, pages 491–495.

[Siegwart et al., 2011] Siegwart, R., Nourbakhsh, I. R., and Scaramuzza, D. (2011). *Introduction to autonomous mobile robots*. MIT Press.

[Silver et al., 2016] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

[Siméon et al., 2000] Siméon, T., Laumond, J.-P., and Nissoux, C. (2000). Visibility-based probabilistic roadmaps for motion planning. *Advanced Robotics*, 14(6):477–493.

[Sohn et al., 1999] Sohn, J., Kim, N. S., and Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3.

[Sommerlade and Reid, 2008] Sommerlade, E. and Reid, I. (2008). Information theoretic active scene exploration. In *Proc. CVPR*.

[Song et al., 2011] Song, K., Liu, Q., and Wang, Q. (2011). Olfaction and hearing based mobile robot navigation for odor/sound source search. *Sensors*, 11:2129–2154.

[Su et al., 2015a] Su, D., Vidal-Calleja, T., and Miro, J. V. (2015a). Simultaneous asynchronous microphone array calibration and sound source localisation. In *Proc. IROS*, pages 5561–5567.

[Su et al., 2015b] Su, D., Vidal-Calleja, T., and Miro, J. V. (2015b). Split conditional independent mapping for sound source localisation with inverse-depth parametrisation. In *Proc. IROS*, pages 2000–2006.

[Sun et al., 2005] Sun, Z., Hsu, D., Jiang, T., Kurniawati, H., and Reif, J. H. (2005). Narrow passage sampling for probabilistic roadmap planning. *IEEE Transactions on Robotics*, 21(6):1105–1115.

[Svaizer et al., 1997] Svaizer, P., Matassoni, M., and Omologo, M. (1997). Acoustic source location in a three-dimensional space using crosspower spectrum phase. In *Proc. ICASSP*, volume 1, pages 231–234.

[Tanyer and Ozer, 2000] Tanyer, S. G. and Ozer, H. (2000). Voice activity detection in nonstationary noise. *IEEE Transactions on Speech and Audio processing*, 8(4):478–482.

[Tesch et al., 2013] Tesch, M., Schneider, J., and Choset, H. (2013). Expensive multiobjective optimization for robotics. In *Proc. ICRA*, pages 973–980.

[Thrun, 2003] Thrun, S. (2003). Learning occupancy grid maps with forward sensor models. *Autonomous Robots*, 15(2):111–127.

[Torras, 1992] Torras, C. (1992). *Robot Motion Planning: A Survey*. Springer Netherlands, Dordrecht.

[Valin et al., 2007a] Valin, J., Yamamoto, S., Rouat, J., Michaud, F., Nakadai, K., and Okuno, H. (2007a). Robust recognition of simultaneous speech by a mobile robot. *IEEE Transactions on Robotics*, 23(4):742–752.

[Valin et al., 2007b] Valin, J.-M., Michaud, F., and Rouat, J. (2007b). Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems*, 55(3):216–228.

[Van Veen and Buckley, 1988] Van Veen, B. D. and Buckley, K. M. (1988). Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24.

[Vermaak and Blake, 2001] Vermaak, J. and Blake, A. (2001). Nonlinear filtering for speaker tracking in noisy and reverberant environments. In *Proc. ICASSP*, volume 5, pages 3021–3024.

[Vincent et al., 2015] Vincent, E., Sini, A., and Charpillet, F. (2015). Audio source localization by optimal control of a mobile robot. In *Proc. ICASSP*, pages 5630–5634.

[Vo and Ma, 2006] Vo, B. N. and Ma, W. K. (2006). The Gaussian mixture probability hypothesis density filter. *IEEE Transactions on Signal Processing*, 54(11):4091–4104.

[Wang and Kaveh, 1985] Wang, H. and Kaveh, M. (1985). Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(4):823–831.

[Ward et al., 2003] Ward, D. B., Lehmann, E. A., and Williamson, R. C. (2003). Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Transactions on Speech and Audio Processing*, 11(6):826–836.

[Wightman and Kistler, 1999] Wightman, F. L. and Kistler, D. J. (1999). Resolution of front–back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America*, 105(5):2841–2853.

[Wilcoxon, 1945] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

[Woo et al., 2000] Woo, K.-H., Yang, T.-Y., Park, K.-J., and Lee, C. (2000). Robust voice activity detection algorithm for estimating noise spectrum. *Electronics Letters*, 36(2):180–181.

[Yamauchi, 1997] Yamauchi, B. (1997). A frontier-based approach for autonomous exploration. In *Proc. CIRA*, pages 146–151.

[Zhang and Wu, 2013] Zhang, X.-L. and Wu, J. (2013). Deep belief networks based voice activity detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(4):697–710.

[Zotkin et al., 2002] Zotkin, D. N., Duraiswami, R., and Davis, L. S. (2002). Joint audio-visual tracking using particle filters. *EURASIP Journal on Applied Signal Processing*, 2002(1):1154–1164.

Résumé

L'audition est une modalité utile pour aider un robot à explorer et comprendre son environnement sonore. Dans cette thèse, nous nous intéressons à la tâche de localiser une ou plusieurs sources sonores mobiles et intermittentes à l'aide d'un robot mobile équipé d'une antenne de microphones en exploitant la mobilité du robot pour améliorer la localisation. Nous proposons d'abord un modèle bayésien pour localiser une seule source mobile intermittente. Ce modèle estime conjointement la position et l'activité de la source au cours du temps et s'applique à tout type d'antenne. Grâce au mouvement du robot, il peut estimer la distance de la source et résoudre l'ambiguïté avant-arrière qui apparaît dans le cas des antennes linéaires. Nous proposons deux implémentations de ce modèle, l'une à l'aide d'un filtre de Kalman étendu basé sur des mélanges de gaussiennes et l'autre à l'aide d'un filtre à particules, que nous comparons en termes de performance et de temps de calcul. Nous étendons ensuite notre modèle à plusieurs sources intermittentes et mobiles. En combinant notre filtre avec un *joint probability data association filter* (JPDAF), nous pouvons estimer conjointement les positions et activités de deux sources sonores dans un environnement réverbérant. Enfin nous faisons une contribution à la planification de mouvement pour réduire l'incertitude sur la localisation d'une source sonore. Nous définissons une fonction de coût avec l'alternative entre deux critères: l'entropie de Shannon ou l'écart-type sur l'estimation de la position. Ces deux critères sont intégrés dans le temps avec un facteur d'actualisation. Nous adaptons alors l'algorithme de *Monte-Carlo tree search* (MCTS) pour trouver, efficacement, le mouvement du robot qui minimise notre fonction de coût. Nos expériences montrent que notre méthode surpasse, sur le long terme, d'autres méthodes de planification pour l'audition robotique.

Mots-clés: audition robotique, localisation de sources sonores, planification de mouvement.

Abstract

Robot audition provides hearing capability for robots and helps them explore and understand their sound environment. In this thesis, we focus on the task of sound source localization for a single or multiple, intermittent, possibly moving sources using a mobile robot and exploiting robot motion to improve the source localization. We propose a Bayesian filtering framework to localize the position of a single, intermittent, possibly moving sound source. This framework jointly estimates the source location and its activity over time and is applicable to any microphone array geometry. Thanks to the movement of the robot, it can estimate the distance to the source and solve the front-back ambiguity which appears in the case of a linear microphone array. We propose two implementations of this framework based on an extended mixture Kalman filter (MKF) and on a particle filter, that we compare in terms of performance and computation time. We then extend our model to the context of multiple, intermittent, possibly moving sources. By implementing an extended MKF with joint probabilistic data association filter (JPDAF), we can jointly estimate the locations of two sources and their activities over time. Lastly, we make a contribution on long-term robot motion planning to optimally reduce the uncertainty in the source location. We define a cost function with two alternative criteria: the Shannon entropy or the standard deviation of the estimated belief. These entropies or standard deviations are integrated over time with a discount factor. We adapt the Monte Carlo tree search (MCTS) method for efficiently finding the optimal robot motion that will minimize the above cost function. Experiments show that the proposed method outperforms other robot motion planning methods for robot audition in the long run.

Keywords: robot audition, source localization, robot motion planning, Bayesian filtering.