Acoustic Echo Estimation using the model-based approach with Application to Spatial Map Construction in Robotics

Saqib, Usama

DOI (link to publication from Publisher):
10.54337/aau528831183

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Saqib, U. (2022). *Acoustic Echo Estimation using the model-based approach with Application to Spatial Map Construction in Robotics*. Aalborg Universitetsforlag. https://doi.org/10.54337/aau528831183

ACOUSTIC ECHO ESTIMATION USING THE MODEL-BASED APPROACH WITH APPLICATION TO SPATIAL MAP CONSTRUCTION IN ROBOTICS

BY
USAMA SAQIB

DISSERTATION SUBMITTED 2022

AALBORG UNIVERSITY
DENMARK

Acoustic Echo Estimation using the model-based approach with Application to Spatial Map Construction in Robotics

Ph.D. Dissertation

Usama Saqib

Aalborg University
Department of Architecture, Design and Medialogy
Fredrik Bajers Vej 7B
DK-9220 Aalborg

Dissertation submitted: December, 2022

PhD supervisor: Associate Prof. Jesper Rindom Jensen
Aalborg University

PhD committee: Associate Professor Claus B. Madsen (chairman)
Aalborg University

Dr. Akihiko Sugiyama
Yahoo! JAPAN Research

Professor Kazuhiro Nakadai
Tokyo Institute of Technology

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Architecture,
Design and Media Technology

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-991-2

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Stibo Complete, 2023

Curriculum Vitae

Usama Saqib

Usama received his B.Sc degree in Electrical Engineering from American University of Sharjah, U.A.E. (2010) and an M.Sc degree in Embedded Systems Engineering from University of Bedfordshire, U.K. (2015). In 2018, he joined Audio Analysis Lab at Aalborg University, Denmark, as a Ph.D. fellow to work on his research topic "Signal Processing for Robots and Drones." His research interest includes robotics, multi-channel acoustic localization, beamforming, embedded systems, and bio-mimicry. Prior to starting his Ph. D., Usama worked with several companies as an embedded systems engineer where he has gains experience in robotics, electronics design, and sensor technologies.

Abstract

Constructing an accurate map of an indoor environment is an active area of research within robotics. Camera- and laser-based technologies are commonly used to generate a spatial map of an environment. These maps are used to enable robots to move within an environment. However, these modalities have limitations as these technologies cannot detect transparent surfaces typically found in an office environment. Moreover, camera- and laser-based technologies suffer from a limited field of view, limiting the ability to generate a spatial map. These limitations can be addressed by utilizing sound. This inspiration comes from animals that utilize echolocation to make "sense" of their environment, i.e., get spatial information about the environment using sound. Animals such as bats, rats, and dolphins are among the few that have mastered the art of echolocation with accurate precision. The main research question that we attempt to answer in this thesis is that *if animals in nature can use echolocation for navigation, can we enable robots to utilize echolocation to navigate an environment?*

The central theme of this thesis is to propose the utilization of the concept of echolocation and combine it with advanced audio processing techniques that can complement existing robot perception technologies to estimate acoustic echoes. To this end, we propose a model-based approach to detecting and estimating acoustic echoes, i.e., we utilize a sound propagation model to formulate the problem of estimating a parameter of interest (POI), e.g., time of arrival (TOA) of acoustic echoes. Therefore, two methods to resolve the acoustic echo estimation problem are presented in this work: a non-linear least squares (NLS) estimator and an Expectation-Maximisation (EM) approach. Both methods estimate TOA/DOA directly from the observed signals. In this thesis, we first propose a single-channel TOA estimation technique using the NLS and EM methods. Later, these methods are extended into the multi-channel approach to estimate the direction of arrival (DOA) information of an acoustic reflector by estimating the time difference of arrival (TDOA) of acoustic echoes. Here, spatial filtering techniques, e.g., beamforming techniques, are utilized to localize the position of an acoustic reflector. Delay and sum beamformer (DSB), minimum variance distortionless response (MVDR) beamformer, and linear constraint minimum variance (LCMV) beamformer are used to estimate the DOAs of acoustic echoes. A proof of concept (POC) robotic platform was built to test the performance of the proposed methods. The estimators and beamforming techniques were implemented to acquire actual data and later used to localize acoustic landmarks for spatial map construction.

Finally, the detection problem was extended, and a new approach was proposed, where we utilize the ego-noise of a robotic platform to detect the presence of an acoustic landmark. This work paves the way for a novel sound-based collision avoidance system that gives 360^o of spatial awareness which could be utilized in robots and drones.

This thesis will begin by reviewing the overall architecture of the robotic platform and later narrow down the discussion to robotic perception. The problems associated with estimating acoustic reflectors, e.g., walls, using a traditional approach, e.g., cameras and lidar, are also discussed. Discussion related to utilizing echolocation for acoustic spatial map construction is also an important highlight of this thesis. Based on the discussion in Section 1, a research question is formulated and later the aims and objectives of this thesis are discussed. In Section 2, the acoustic echo model is presented based on which we derive the nonlinear least squares (NLS) and expectation-maximization (EM). Discussion related to estimating distance, directions of acoustic echoes, and representing this information to construct an acoustic map of an environment is found in Section 2 and Section 3. The contribution is found in Section 4 while the conclusion and future work are discussed in Section 5

Resumé

Dektektion og estimering af akustiske reflektorer såsom vægge og andre forhindringer i et fysisk miljø er et populært emne inden for robotik til eksempelvis robotnavigering. Traditionelt anvendes kamera- og laser-baseret teknologi til at detektere tilstedeværelsen af landemærker til generering af spatiale kort, der kan hjælpe robotterne med at navigere inden for et tredimensionelt rum. Disse teknologier har dog begrænsninger idet de f.eks. ikke kan detektere transparente overflader, der er typisk findes i eksempelvis kontormiljøer. Desuden lider kamera- og laser-baserede teknologier typisk af et begrænset synsfelt, hvilket også begrænser deres anvendelse i generering af spatiale kort. Begrænsningerne kan derimod adresseres ved brug af lyd. Dette er biologisk inspireret idet dyr anvender ekkolokalisering til at forstå det omkringliggende miljø ved at få spatial information omkring miljøet ved hjælp af lyd. Flagermus, rotter, og delfiner er nogle af de få dyrearter der har mestret ekkolokalisering med høj nøjagtighed. Det primære forskningspørgsmål som forsøges besvaret i denne afhandling er derfor: ”hvis dyr i naturen kan anvende echolocation til navigering, kan dette så overføres til robotter til navigering i et fysisk miljø?”

Det centrale tema i denne afhandlingen er derfor at foreslå brugen af ekkolokation og avanceret audioprocesseringsteknikker til estimering af akustiske ekkoer, hvilket kan komplimentære eksisterende robotperceptionsteknologier. Til dette formål foreslår vi en modelbaseret tilgang til detektering og estimering af akustiske ekkoer, dvs. vi anvender en model for lydudbredelse til at formulere estimeringsproblemer for de relevante parametre såsom ekkoers ankomsttid (TOA). To fremgangsmåder til estimering af akustiske ekkoer præsenteres derefter: en non-lineær least squares (NLS) estimator der estimere TOA'er direkte fra observerede signaler i frekvensdomænet og en expectation-maximization (EM) metode der estimerer TOA'er direkte fra observationerne i tidsdomænet. Først foreslåes en enkeltkanals TOA estimeringsteknikker, der anvender NLS og EM metoderne. Senere udvides disse metoder til flerkanals scenarier, hvilket muliggør estimering af ankomstretningen (DOA) af en akustisk reflektor ved at estimatere tidforskellen mellem ankomster (TDOA) af de akustiske ekkoer. Her anvendes spatiale filtreringsteknikker (beamforming) til at lokalisere positionen af en akustisk reflektor. Forsink-og-summér beamforming (DSB), minimumarians og forvrængningsfri beamforming (MVDR), og minimum varians med lineære betingelser beamforming (LCMV) anvendes alle til estimering af TOA'er og DOA'er af de akustiske ekkoer. Til test af ydeevnen af de foreslåede metoder, blev der bygget og anvendt en prototype robotplatform (POC). Estimatorene og beamformingsteknikkerne blev implementeret til opsamling af reel data, som senere blev anvendt til lokalisering af akustiske landemærker til konstruktion af spatiale kort. Endeligt blev detektionsproblemet udvidet og en ny metode foreslået, hvor egenstøjen fra en robotplatform blev anvendt til at detektere tilstedeværelsen af et akustisk landemærke. Dette arbejde vil potentielt give 360 graders spatial årvågenhed og bane vejen for nye lyd-baserede systemer til kollisionsforhindring, der kan anvendes i robotter og droner.

Dette speciale vil begynde med at gennemgå robotplatformens overordnede arkitektur og senere indsnævre diskussionen til robotperception. Problemerne forbundet med at estimere akustiske reflektorer, f.eks. vægge, ved brug af en traditionel tilgang, f.eks. kameraer og lidar, diskuteres også. Diskussion relateret til udnyttelse af ekkolokalisering til akustisk rumlig kortkonstruktion er også et vigtigt højdepunkt i denne afhandling. På baggrund af diskussionen i afsnit 1 formuleres et forskningsspørgsmål og senere diskuteres formålet med dette speciale. I afsnit 2 præsenteres den akustiske ekkomodel ud fra hvilken vi udleder de ikke-lineære mindste kvadraters (NLS) og forventningsmaksimering (EM). Diskussion relateret til estimering af afstand, retninger af akustiske ekkoer og repræsentation af denne information til at konstruere et akustisk kort over et miljø findes i afsnit 2 og afsnit 3. Bidraget findes i afsnit 4, mens konklusionen og det fremtidige arbejde diskuteres i afsnit 5.

Contents

II Papers 43

List of publications

The main body of this thesis consists of the following publications:

A U. Saqib and J. R. Jensen, "Sound-based Distance Estimation for Indoor Navigation in the Presence of Ego Noise, "*European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, 2019.

B J. R. Jensen, U. Saqib and S. Gannot, "An Em Method for Multichannel Toa and Doa Estimation of Acoustic Echoes," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2019.

C U. Saqib, S. Gannot, and J. R. Jensen, "Estimation of acoustic echoes using expectation-maximization methods," *EURASIP Journal on Audio, Speech, and Music Processing,* 2020.

D U. Saqib, and J. R. Jensen, "A model-based approach to acoustic reflector localization with a robotic platform," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* Las Vegas, NV, USA, 2020.

E U. Saqib, A. Deleforge, and J. R. Jensen, "Detecting Acoustic Reflectors using Robot's Ego-Noise", *Proc. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, June 2021.

F U. Saqib, M. G. Christensen, and J. R. Jensen, "Robust Acoustic Reflector Localization for Robots," *EURASIP Journal on Audio, Speech, and Music Processing (JASM)* 2022. (Submitted)

G U. Saqib, and J. R. Jensen, "A Framework for Spatial Map Generation using Acoustic Echoes for Robotic Platforms," *Elsevier's Robotics and Autonomous Systems* 2021.

Preface

This thesis was submitted to the Doctoral School of Engineering and Science at Aalborg University in partial fulfillment of the requirement of the doctoral degree in philosophy. This thesis is composed of two parts: The first part is an introduction to the research, while the second part contains the collection of papers that were published and submitted to respectable peer-reviewed conferences and journals. The research was carried out in the period between May 2018 to August 2021 at the Department of Architecture, Design, and Medialogy at Aalborg University, Denmark.

This thesis was not possible without the support of numerous people who came into my life and supported me in my journey to achieve a doctoral degree. I will always remember them in my memories and I am forever thankful for their love and support. First of all, I would like to express my gratitude towards my supervisor Jesper Rindom Jensen, who gave me an opportunity to work on this relatively new area of research and also gave me complete freedom to change the direction of my research. I always had good conversations with him regarding research and non-research-related topics. Secondly, I am grateful to my collaborators, Sharon Gannot from Bar-Ilan University, Ramat-Gan, Israel, for spurring my interest in optimization techniques such as expectation-maximization (EM) and helping me understand key acoustic localization concepts such as beamforming. Additionally, I would like to extend my gratitude to Antoine Deleforge who also invited me to Universite de Lorraine, CNRS, Inria, LORIA, France as a visiting researcher. Antoine was very supportive of my research ideas and offered unlimited help in my research.

Moreover, the Ph.D. process would have been very lonely if it was not for my friends in Aalborg. I really appreciate their presence during my stressful periods as they made every attempt to boost me morally and spiritually. Last but not the least, I would like to wholeheartedly send my thanks to my family, especially my mother, sister and late father whose only advice to me was to keep learning.

And to all the people I met over coffee breaks during my three years at Aalborg University. A huge thank you to you all.

Usama Saqib
Aalborg University, December 21, 2022

Part I

Introduction

Introduction

"Any sufficiently advanced technology is indistinguishable from magic"
- Arthur C. Clarke -

1 Background

To many of us, the term *robots* ignites excitement as it is an attempt by scientists and engineers to replicate human senses and movement. Robots have captured our imagination for centuries. The idea of having mechanical entities to help us ease our mundane lives has encouraged us to invest our monetary and intellectual resources to foster research and development on robotics. Initially, robots were used in academia and certain industries but their performance was slow and were often unreliable. With the advent of modern technology, we saw a rise in processing power, as well as smaller and cheaper electronics, that has enabled research and development of robotic systems [1]. Over time, robotics has spun into several sub-domains such as robot perception, robot navigation, and human-robot interaction (HRI) [2]. Nowadays, robots can detect and avoid obstacles, distinguish between a human and an object, understand semantic information from video and audio, navigate indoor/outdoor terrain, handle dangerous chemicals, manufacture cars as well as planes and explore extra-terrestrial environments autonomously. Additionally, teaching robots to learn human cultures [3] and social cues [4] is also a major area of research.

One of the most exciting areas of research within robotics is robot perception. This subdomain enables robots to sense the physical world which can aid robotic platforms to navigate an environment and possibly construct a spatial map of an environment. For example, indoor cleaning robots from *iRobot* and *Dyson* consist of cameras that enable a robot to localize its position in an environment [5]. Therefore, having a map of an environment is useful for robots as it can aid a robot to navigate an environment more efficiently. Great emphasis is given to replicating the senses of seeing and hearing. Thus, computer vision and robot audition techniques have dominated the robotic space in recent years. This is partly due to advancements in sensor technologies and processing power. However, as we shall see in this thesis, vision-based technologies, e.g., cameras and lasers, have limitations. For instance, a robot equipped with a camera-based system cannot detect obstacles and physical surfaces that are not in direct line-of-sight or are transparent. Furthermore, these systems are affected by changing light conditions and tend to be very expensive if wider coverage areas are required. Therefore, on their own vision-based technologies are not suitable in some environments for spatial map generation. Hence, other sensor modalities, such as the use of microphones, could improve the overall accuracy of environmental representation. This aspect of sensing the environment using other sensor modalities.

In this thesis, we will review various technologies that enable robots to explore their environment. More specifically, we will investigate different ways to utilize audio processing techniques that could be used by robots to estimate the geometry of an environment and generate a spatial map of an environment using the concept of echolocation. By investigating different technologies, this thesis would like to promote a novel and exciting area of research, *active sound localization*, which has received less attention within the scientific domain. Hence, in the following sections, we shall see how audio signal processing techniques could be combined with robot perception to overcome the limitations of current technologies, e.g., cameras and lasers. We will begin by investigating different technologies that are inspired by animals in nature before analyzing the robot's architecture with the aim to understand how a robot could incorporate audio signal processing to solve a particular problem, i.e., applying audio processing algorithms to enable a robot to construct a map of an indoor environment. Later, we will investigate different spatial filtering techniques, e.g., beamforming, to estimate the direction of a sound.

1.1 Bio-mimicry in robotics

The ultimate goal of robotics is to develop mechanical entities that look, react and think like humans [6]. To realize this dream, research in animals' senses is carried out to replicate the sensory organs. Researchers have been inspired by the five senses (sight, sound, touch, taste, and smell) for many years, as is evident from a large amount of scientific literature on the subject [6]. By studying how animals use their senses, researchers can come up with new techniques that can be applied to robotics. Some work on computer vision, and sound-source localization techniques are well investigated by studying animal vision and hearing but the remaining senses (touching, tasting, and smelling) are still an active area of research and beyond the scope of this thesis [6]. Interested readers are advised to read [7–15].

Research in understanding human sight has enabled researchers to develop algorithms that allow a computer system with a camera attached to detect, recognize and track the movement of obstacles or people. Computer vision is the study of enabling a computer system to understand and interpret visual information from images and videos. It started gaining traction in the 1950s [16]. Initially, it was not widely adopted by industry and academia because of the low processing power of the computer systems available at the time. As time went on, however, computer vision-based algorithms became popular as there was an exponential improvement in processing power and wider availability of cheaper electronics [1]. Computer vision algorithms have found refuge in many practical applications, e.g., object detection [17–19], and scene semantic recognition (SSR) [20], among others.

Techniques to locate the presence of sound sources within an environment were inspired by animal auditory systems. The field of robot audition focuses on using signal processing techniques to help robots understand and respond to sound in order to improve human-robot interactions. This subdomain of robotics involves developing algorithms that allow robots to locate and classify the source of a sound. [6]. For example, taking inspiration from nature, researchers in [21] mimics the auditory system of a parasitoid fly. This will be discussed in detail in Section 1.3. Bio-mimicry can also be seen in the use of active sound localization, which is based on the way that some animals use sound to explore their environment and gather spatial information. For example, some animals use echolocation to determine the distance to nearby objects by emitting sounds and listening to the echoes. This process is an example of bio-mimicry that has been applied in robotics to help robots navigate their surroundings [22–24] and is predominantly used by animals. e.g., bats, to aid in navigating in complete darkness [25]. Medical studies also suggest that humans could also train themselves to use echolocation to navigate and detect obstacles [26]. Echolocation-based systems, which are inspired by the way some animals use sound to navigate and gather information about their surroundings, have several advantages over vision-based systems. For example, they can detect glass surfaces and can be used in low-light or no-light conditions. The development of *Sound Navigation and Ranging (SONAR)* [27, 28], which uses sound waves to detect objects and measure distances, is also an example of bio-mimicry. SONAR is commonly used in naval vessels to locate schools of fish and to monitor the ocean floor, as well as to find sunken ships.

Based on the above discussion, using robots to acquire spatial information from the environment is an interesting and challenging area of research. In literature, this is known as *active sound localization*. In this thesis, we take inspiration from animals, e.g., bats, to aid a robot to navigate its environment in complete darkness while generating a map of the environment. To do so would require an in-depth analysis of the robot's architecture. In the following subsections, we shall examine the technological achievements by reviewing the building blocks that make up a robotic system. More specifically, we will investigate individual blocks within the robot's architecture to better understand the processes involved in enabling a robot to perceive its environment.

1.2 Robot Architecture

When designing robots, four main blocks are considered to enable a robot to perceive, process, and react to an environment. The four blocks are shown in Fig. 1: Sensor block, perception block, mapping block, and path planning and control block [29, 30]. These individual blocks are crucial in the development of a robotic system that can perform a task without human supervision. These include: moving within an environment autonomously or with a fixed trajectory, interacting with a human speaker or objects in a 3D space, avoiding colliding with obstacles, and executing complex tasks such as picking up an object. These blocks are discussed in detail in the following subsections which will help in identifying the limitations of the current

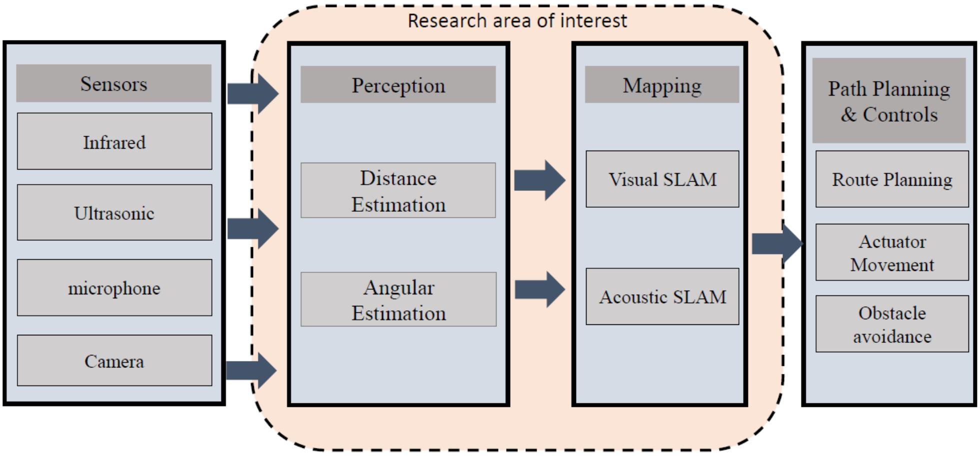


Fig. 1: Overview of robot architecture

technologies and define the objectives as well as the structure of this dissertation.

Sensor Block

The first block in a robotic architecture is called the Sensor block which enables a robotic platform to acquire raw data from the physical environment. Traditionally, two types of sensors are used on robotic platforms: Proprioceptive and exteroceptive sensors [31]. Proprioceptive sensors measure the internal state of the systems while exteroceptive sensors measure the parameters external to the robotic systems. Proprioceptive sensors include the accelerometer, gyroscope, potentiometer, motor encoders, Inertial Measurement Unit (IMU), and so on, whereas exteroceptive sensors include a camera, ultrasonic sensors, range sensors, pressure sensors, and so on.

Proprioceptive sensors are used to estimate the position of a robotic system by constantly monitoring the internal state of the robot. This is also called *odometry* or *dead reckoning*, whereas exteroceptive sensors use external parameters to correlate the position of a robotic platform [32]. However, proprioceptive sensors are sensitive to sensor drift, which results in errors. These errors multiply over time, which makes these types of sensors unsuitable for position estimation. Exteroceptive sensors, on the other hand, rely on external parameters to help a robotic platform localize its position in 3D space. Exteroceptive sensors are generally used to replace odometry data. The use of these external sensors to monitor the environment, i.e., to detect landmarks, could aid a robot in simultaneously locating the position of the platform and generating a spatial map of the environment. These sensors are commonly used in a framework called Simultaneous Localization and Mapping (SLAM) discussed in Section 1.2. The most famous exteroceptive sensors used for the SLAM framework are cameras and lasers, that aid a robotic platform to construct a spatial map of an environment. However, camera and laser-based technologies have their own limitations, as discussed later. The raw data obtained by these sensors are then processed by the perception block discussed next.

Perception block

In this block, the raw data obtained from the sensor block is processed in order to apply algorithms that enable a robot to comprehend, perceive, make decisions, and operate in a real-world setting. Some examples of robot perception are object detection [33], object recognition [34, 35], semantic environment classification [36, 37], activity detection [38], highway/road detection [38], voice and gesture detection [39], 3D environment representation [40], vehicle detection [41], pedestrian detection [42, 43], terrain classification [44] and environmental change detection [45].

Data can be obtained either from a single sensor or by using a fusion of different sensors. In the case of single sensors, only a single physical parameter is measured, e.g., distance, velocity, acceleration, etc. However, a single modality will not give distinct information about the environment. Hence, in scientific literature, the use of multiple sensor modalities is recommended. This makes sensing a challenging environment, e.g., fog [46], smoke [47], airborne dust [48], more resilient. For example, a lidar sensor bounces off a wall but passes through a transparent surface, whereas sound-based sensors, such as ultrasonic sensors, can detect transparent surfaces. Thus, these multiple sensors can complement each other [49]. Moreover, fusing data from multiple sensors enables a better representation of the environment [50].

Mapping Block

Robot mapping is an active area of research that has applications ranging from logistics and supply chains [51–54], military [55], search and rescue mission [56, 57], and autonomous driving, as it is considered to be the backbone of robot architecture to ensure effective navigation of a robotic platform within an environment. Here, a mapping consists of acquiring metric model information as well as a semantic representation of the environment, which is also known as scene/semantic representation in the scientific literature. Most of the time, the exteroceptive sensors used to acquire data from the environment are used by this block to ensure reasoning and inferences regarding the real world where the robot operates [58]. Therefore, different assumptions are made to map (represent) an environment. More specifically, robots are designed for two environments: indoor and outdoor. For an indoor environment, it is assumed that the ground is flat and regular, while this is not the case for the outdoor environment. Hence, the sensors used for these environments are also different, since these sensors acquire different environmental parameters to enable a robot to navigate.

Traditionally, robots were required to have prior knowledge of an environment to ensure accurate and efficient navigation [59]. However, prior knowledge of an environment is only possible if the environment that the robot is moving in is static. In reality, the environment, e.g., in a warehouse setting, is always changing. Hence, a robot cannot rely on prior knowledge of the environment for navigation as it is susceptible to obstacles, such as moving pedestrians.

To remedy this problem, a framework was developed by R.C. Smith and P. Cheeseman in their seminal work [60] where a framework for representing the relationship (position and orientation) of an object was proposed. More work on this framework was proposed by [61] to construct a map of an environment. These types of algorithms are called Simultaneous Localization and Mapping (SLAM). SLAM is a computational problem of generating a map of an environment while keeping track of the robotic platform as it moves within an environment. In the case of visual SLAM, the algorithm detects a visual landmark, e.g., chair, person, object, etc., using a camera- or a laser-based system against which the algorithm simultaneously estimates the position and orientation of a robot, while at the same time uses this information to build a topological or stochastic map of an environment.

One noticeable problem of using visual SLAM is that it would not work in the absence of light, i.e., in an environment with complete darkness. To resolve this problem, acoustic landmarks are of great importance. Compared to visual SLAM, acoustic SLAM [62, 63] treats sound sources as landmarks. This thesis extends this idea and proposes probing the environment which is then used to construct a stochastic or topological map of an indoor environment. Once a map of an environment is made by the robotic platform, then the robot finds the most efficient way to navigate an environment, i.e., the robot searches for a path that enables it to reach its target in as little time as possible. This is the job of the next block, Path Planning, and Control.

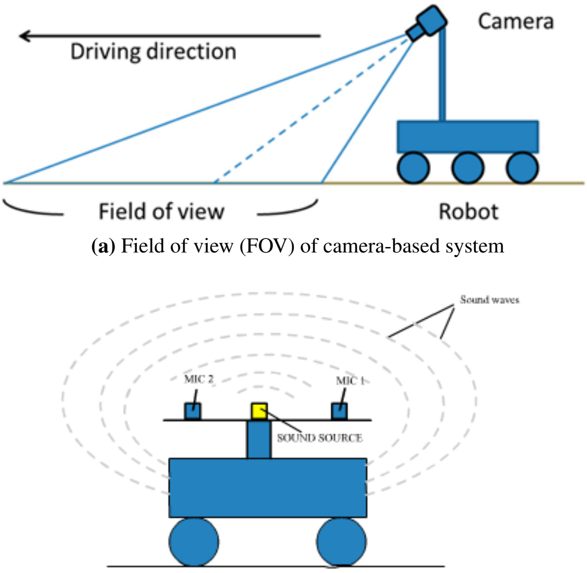
Path Planning and Control Block

Within robotics, the ability of a robotic platform to plan and navigate within an unknown or known environment while avoiding colliding with objects or obstacles autonomously is the job of this block. In the literature, this problem is called path planning, also known as, motion planning [64, 65]. This is the most important block that consolidates data from all the remaining blocks to ensure that a robot or a drone navigates an environment autonomously. There are numerous methods found in the literature that suggests ways to solve motion planning problems. According to [64], there are three approaches to address this problem: roadmaps, potential fields, and cell decomposition. But the most important and used methods to find the most optimum path are A* (A-Star) algorithm [66–69], probabilistic roadmaps [70–72] and optimization methods such as Particle Swarm or Genetic Algorithm [71, 73, 74].

However, in this thesis, we will not be exploring path planning and control algorithms, as they are beyond the scope of this research. Therefore, interested readers are advised to refer to [75–77].

Limitations of existing technologies

As discussed in the previous subsections, a robotic system comprises several blocks that play a crucial role in the development of robots intended to solve a particular problem. For instance, a robot intended for navigating an indoor environment will be equipped with visual sensors,



(a) Field of view (FOV) of camera-based system

(b) Field of view (FOV) of microphone-based system

Fig. 2: Comparison of a camera and microphone-based systems

e.g., a camera- and/or laser-based sensors to detect the presence of an obstacle, e.g., objects and humans. These signals are used by the perception block, where the algorithm processes the raw data to estimate the distance and position of the obstacles, while the mapping block will consist of SLAM algorithms to localize and track the position of the robotic platform itself in 3D space, i.e., for spatial map generation. The path planning and control block then uses the information from the other blocks to plan an effective route for the robot.

However, robotic platforms that utilize visual sensors, e.g., cameras and lasers, have limitations that can affect the performance and accuracy of spatial map generation. This is because these visual sensors are susceptible to low light intensity and offer a limited field of view, as depicted in Fig. 2. Moreover, visual sensors cannot detect transparent surfaces that are predominantly available in a typical office environment, leading to a construction of a spatial map with inaccuracies. One way to overcome these problems is to use multiple modalities, e.g., combining ultrasonic sensors with a camera or lidar to detect glass surfaces. But adding sensors to a robotic platform could potentially increase its overall cost. However, existing loudspeakers/microphones on the robot cannot generate and detect ultrasonic frequencies.

Another interesting area of research is to design algorithms in the audible frequency range because most off-the-shelf acoustic systems, e.g., loudspeakers and microphones, work in the audible frequency range. This does not increase the overall cost of the robotic platform since loudspeakers and microphones are already equipped for these robots, e.g., Softbank's NAO

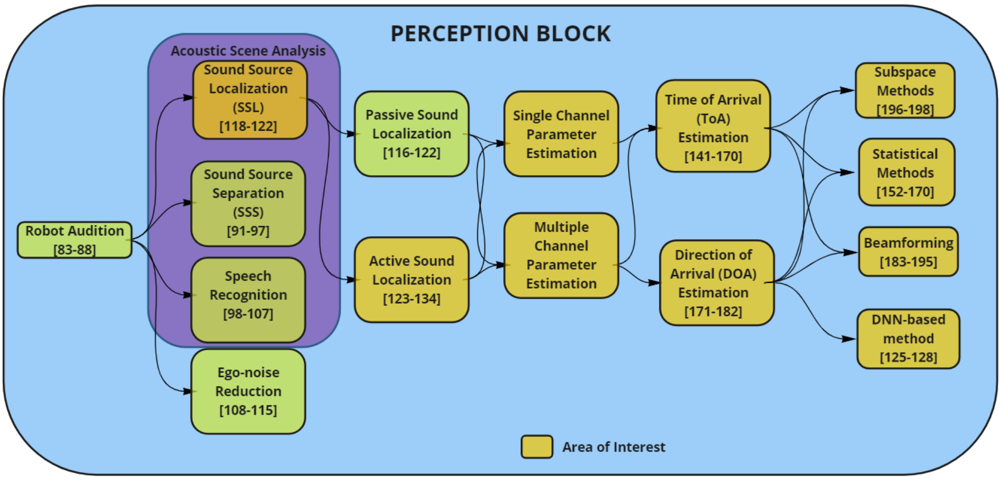

PERCEPTION BLOCK
Acoustic Scene Analysis
Sound Source Localization (SSL) [118-122]
Sound Source Separation (SSS) [91-97]
Speech Recognition [98-107]
Ego-noise Reduction [108-115]
Robot Audition [83-88]
Passive Sound Localization [116-122]
Active Sound Localization [123-134]
Single Channel Parameter Estimation
Multiple Channel Parameter Estimation
Time of Arrival (ToA) Estimation [141-170]
Direction of Arrival (DOA) Estimation [171-182]
Subspace Methods [196-198]
Statistical Methods [152-170]
Beamforming [183-195]
DNN-based method [125-128]
Area of Interest

Fig. 3: Overview of robot audition and the focus of this thesis

robot. Therefore, a new approach to detecting the presence of acoustic reflectors is required using off-the-shelf acoustic systems. Localization of an acoustic source is a known problem within the context of audio signal processing. Traditionally, this is done by estimating the time-of-arrival (TOA) information of an acoustic echo from the acoustic impulse response of an environment, which is used to estimate room geometry. In the following section, we will discuss audio perception techniques that are used in robotics to estimate the time of arrival of acoustic signals.

1.3 Audio Perception for Robots and Drones

Within the context of this dissertation, audio perception is defined as the ability of a robotic platform to detect, estimate and react to a presence of an acoustic source in an environment. Audio perception plays an important role within robotics to detect the presence of a human speaker, as in the case of human-robot interaction (HRI) [78–80], search and rescue robots [81, 82], obstacle detection and avoidance systems [83, 84], sound-based odometry and acoustic scene analysis [85–87]. In the following section, we will review the techniques that are implemented on robots for audio perception. First, we will begin by reviewing the artificial auditory system and also discuss the sub-domain within robotics called robot audition [88]. Later, we will review echolocation techniques that are being used in robotics to detect the presence of obstacles. The overall structure of this research is presented in Fig. 3.

Robot Audition

According to [88, 89], the research domain that aims to acquire audio data from the environment using microphones to localize, detect and recognize human speakers, as well as environmental noise on robots, is called *robot audition*. Robot audition is a relatively new area of research and was accepted as a research topic of its own only in the early 2000s [89, 90]. However, the work on robot audition is found earlier [91–93]. One of the objectives of this domain is to provide electrical ears to robots that enable them to understand commands given by human users. Robot audition encompasses many sub-domains of research. For instance, it establishes human-robot interaction (HRI) [78–80] by employing different algorithms found in sound source localization (SSL) techniques, sound source separation (SSS), speech recognition, and ego-noise reduction. Within the context of HRI, robots are expected to "understand" the sounds that are present within an environment, that is, robots should differentiate speech signals from non-speech signals and music. The research area that promotes this process is known as *acoustic scene analysis* (ASA). The main functions required for ASA are SSS techniques, SSL techniques, and speech recognition techniques [94, 95]. According to [90], two paradigms exist in robot audition, 1) microphone arrays are used to localize acoustic sources, which can be organized into different shapes, e.g., line, circular or spherical. 2) Binaural approach, which utilizes a pair of microphones to localize the acoustic source. However, from an engineering aspect, there are no limitations on how many microphones to use for localization but by utilizing the binaural approach, we have an opportunity to investigate human perception. In the following, we will investigate different sub-domains that constitute robot auditions.

Sound Source Separation

Through evolution, humans have developed an ability to classify and isolate audible sounds [96]. Isolating a sound of interest from the environment is called the *cocktail party effect* in literature. Sound source separation (SSS) is a sub-domain of robot audition that aims to replicate audible sound [97, 98]. SSS techniques are based on single and multiple microphones [99]. Some recent methods employ Gaussian complex models for estimating source parameters using an expectation-maximization (EM) algorithm [100, 101]. Several SSS techniques are proposed to implement this ability [102]. The approaches are categorized into four major groups:

- Blind source separation
- Steered adaptive beamforming
- Inverse filtering
- Binary masking

Speech Recognition

Speech recognition systems within robotics enable a robot to understand human speech which helps in HRI. The process involves recording speech signals and using a parametric approach or machine learning approach to estimate pitch, fundamental frequency, and timbre. These features such as statistical methods [103], spectral methods [104, 105], model-based methods [106, 107], and template methods [108–110] are used. Later, these features are used to identify human speech [111, 112]. Another important aspect of robot audition is to reduce the noise generated by the robotic system itself which is discussed next.

Ego-noise Reduction

One of the main sources of the issue within robotics is the presence of ego noise. Ego noise is caused by the moving part of the robotic platform, e.g., actuators. This becomes increasingly challenging because ego-noise cannot be modeled as a static point source. Hence, traditional statistical methods cannot be applied directly. The researchers in [113], proposed fusing motor data with a dictionary algorithm. The beamforming method is proposed in drone auditions to reduce ego-noise [114, 115]. Other methods found in the literature include, template matching [116] and data-driven methods for ego-noise reduction is also found in the literature [117, 118]. Traditionally, ego noise is considered a source of the problem but the work presented in [119, 120] shows that ego noise could be used constructively.

Sound Source Localization (SSL) and passive sound localization

There are two main categories of SSL techniques: passive sound localization and active sound localization [121, 122]. Passive sound localization involves detecting sounds that are already present in the environment and impinging them on the microphones. Feature extraction techniques are used to estimate the distance and the direction of the sound source. On the other hand, active sound localization probes the environment with a known signal. Feature extraction techniques are applied to acoustic echoes for distance and direction estimation. Echolocation employed by bats, active SONAR in submarines, and radars are some examples of active sound localization [123–126]. In literature, one way to find the direction of a sound source is to take the amplitude and phase difference of the target source from multiple microphones such that the directional differences could lead to locating the target sources [127]. SSL techniques are normally employed to extract two values:

- Direction-of-arrival (DOA) information
- Distance information which can be inferred from time-of-arrival (TOA) information of a sound source

There are several popular techniques used in robotics for estimating the direction of a sound source, including learning-based approaches like neural networks, beamforming-based approaches, subspace-based approaches, and tracking techniques like Kalman filters and particle filters. These techniques are commonly used in robot auditions to help robots understand and respond to sounds in their environment [89, 128–132]. When implementing these techniques on robots, multiple aspects are taken into account: the use of multiple microphones, robustness against noise and reverberation, and the choice of microphone array geometry. Passive source localization requires the presence of a sound source, e.g., a human speaker, as a prerequisite but this problem is addressed by active source localization.

Active Sound localization

As discussed in the previous subsection, the active sound localization technique requires that a sound signal is probed into an environment. The process involves processing the acoustic echoes for features that correspond to the distance and direction of an echo. In nature, Bats are known to use ultrasonic frequencies to probe the environment and use their elongated ears to detect the presence of obstacles while in the air. Moreover, bats are also known to distinguish prey from obstacles while in mid-flight. Researchers have studied and developed systems that utilize echolocation to locate objects [121, 133]. But in robotics, obstacle detection is done by exteroceptive sensors [134–138].

In this thesis, we propose utilizing active sound localization techniques to facilitate the task of spatial map generation [139]. Active source localization approach is used in [140] to classify outdoor terrains, e.g., grass, concrete, and sand, with 97% accuracy. This was achieved by employing a Support Vector Machine (SVM) classifier. Similarly, the work presented by [141], involves the use of a robotic platform that probes the outdoor environment to navigate and classify floras. In [141], the researchers use an ultrasonic emitter/receiver to probe a chirp signal to the environment. The acoustic echoes are then processed to determine the time-of-arrival (TOA) information of the acoustic echoes. The TOA of the acoustic echoes is used by a neural network for spatial map construction.

Moreover, ultrasonic transducers were used in [142]. Here, the authors proposed a biomimetic navigation model. The authors designed a sonar system mimicking the shape of a bat's pinnae. The process involves probing the environment with ultrasonic sound and using ultrasonic receivers to record acoustic echoes. These echoes serve as acoustic fingerprints which are used by the robotic systems to construct a spatial map of an environment. The authors took inspiration from RatSLAM, which is a navigation system based on a computational model of a rodent to show that spatial map construction of an environment is possible with sonar alone [143]. Similar work on the use of sonar (ultrasonic transducer) for spatial map construction on autonomous vehicles which can be found in [144]. In [145], The author proposes two methods for estimating the direction of a sound source in a robot audition system. The first method uses triangulation to calculate the position of the virtual sound source, while the second approach uses a Bayesian approach to estimate the acoustic echoes. Both methods are intended to help robots understand and respond to sounds in their environment. Estimating room geometry estimation based on cross-correlation and probabilistic mapping algorithms were implemented on a standard smartphone in [146] but the choice of frequency, pulse width, and duration were experimentally found in this work.

One similarity among most of these research studies is the use of transducers that work in the ultrasonic frequencies and some of these works only estimate first-order early reflections. However, estimating acoustic echoes in audible frequencies has not received much interest within the active sound localization domain because of its intrusive properties. Intrusiveness is not a problem if the application of interest is either a factory, a warehouse, or an underground tunnel. Therefore, in this thesis, we are interested in exploiting the use of audible frequencies to achieve active sound localization. Moreover, we are taking a model-based approach where we model early reflections and take reverberation, ego-noise, and background noise into account which has not been done in previous approaches.

1.4 Objective and Structure

The scientific literature is dominated by techniques that aid robotic platforms to localize an acoustic source passively, i.e., detecting a sound signal from the environment that impinges microphones. However, in the absence of a sound, e.g., in a quiet environment, these passive localization techniques fail to detect acoustic sources that can help a robot to perceive its environment. This problem could be resolved by incorporating active localization techniques such as echolocation. Therefore, the initial hypothesis that stems from the above discussion is:

> *"If the acoustical properties of sound could help animals interpret distance and angular estimation then it should be possible for us to combine the concept of echolocation (active sound localization) with a mathematical model of early reflections to develop a system that will enable a robot or a drone to represent an environment to aid in navigation while taking the background noise, reverberation, ego-noise and interference into account."*

Despite having great potential, the use of audio signal processing on robots and drones for spatial map generation has limited uses because it introduces new challenges and problems that are not addressed in the traditional audio signal processing domain. For example, the reduction of wind noise, ego noise, movement, and so on. Therefore, in this dissertation, we take inspiration from nature, e.g., bats, to facilitate a robot with active localization techniques so that we can estimate the TOA and DOA of an acoustic reflector and use these estimates to generate a spatial map. To realize this, we need to fulfill four objectives that could support enabling robots to utilize echolocation for spatial map generation. We propose using audio-based sensor technologies, e.g., microphones, to acquire raw data from the environment and propose audio processing techniques to extract acoustic features to enable a robot to construct a spatial map of an environment. The study has the following objectives:

- Formulate a mathematical model of acoustic echoes that incorporates the structure of probe signals, reverberation, interference, and other environmental parameters that affect the sound

- Investigate different parameter estimation techniques that could enable a robot to estimate the distance to acoustic reflectors

- Investigate different spatial filtering techniques to determine the DOA of the acoustic echoes

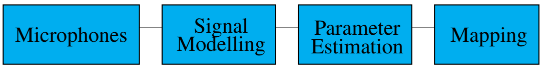


Fig. 4: Proposed acoustic map structure

- Incorporate TOA and DOA estimates to generate a map of an environment

Generating a spatial map of an environment is important for a robotic platform because it allows engineers and building maintenance officers to monitor the state of an environment, e.g., to monitor underground tunnels such as sewers [147]. Moreover, having an accurate map of an environment enables a robot to navigate an environment effectively and efficiently. A spatial map is traditionally generated using laser-based and vision-based technologies. But, as stated earlier, in a typical office environment that consists of glass partitions, these technologies fail to detect the glass surface, hence other modalities such as sound-based sensors are useful. Additionally, to reduce the cost of the robotic platform, we propose algorithms that utilize existing audio sensors on robots to detect acoustic reflectors and use them for spatial map construction. The proposed methods complement existing robotic modalities which makes them useful to generate an accurate map of an environment. The proposed architecture is shown in Fig. 4. As shown in the figure, the microphones on the robotic platform capture the acoustic echoes which are processed by NLS or EM approach for distance and direction of acoustic echoes estimation. This information is used for acoustic map construction. In Chapter 2, we formulate a mathematical model of the acoustic echoes. This model is used throughout this dissertation. Furthermore, we will also investigate acoustic impulse response (AIR) which is traditionally used for TOA estimation. In Chapter 3, we investigate different statistical and parameter estimators based on estimation theory to derive estimators to estimate acoustic parameters, e.g., time of arrival (TOA). Furthermore, we also investigate state-of-the-art spatial filtering techniques to estimate the DOAs of acoustic echoes. In Chapter 4, we introduce the contributions of our work and finally, in Section 5, we will conclude the thesis and discuss possible directions of future research.

2 Acoustic Echo Model

In acoustics, there are two physical regions of sound: the *near field* and the *far field* as seen in Fig. 5 and 6, respectively. *Near field* happens when you are closer to the emitting sound source. The sound receiving the observer is a curved shape. The *far field* happens when you are further away from the radiating sound source. The sound propagation is assumed to be a plane wave. In literature, far-field assumptions are generally used to simplify the problem. For this thesis, we also assumed a far-field sound propagation assumption. Hence, when a robot probes an

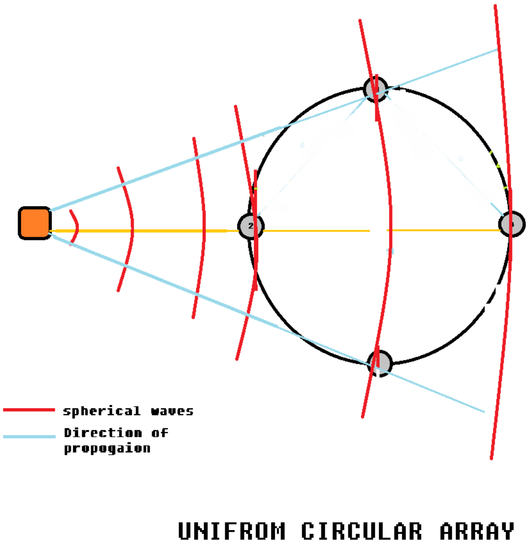


Fig. 5: Near-field assumption

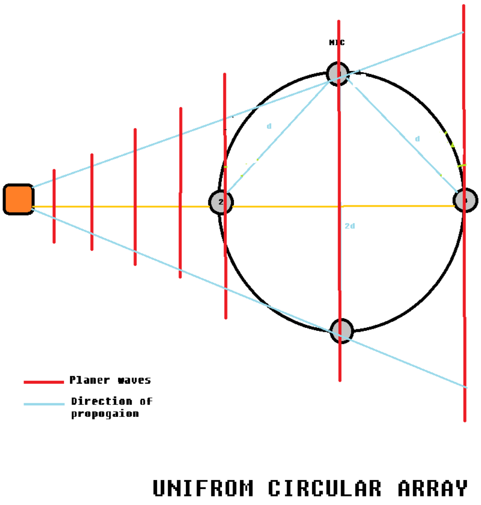


Fig. 6: Far-field assumption

environment, the sound recorded on the microphones is mathematically formulated as shown:

$$\begin{aligned} y(n) &= h(n) * s(n) + v(n), \qquad (1)\\ &= x(n) + v(n), \end{aligned}$$

where $y(n)$ is the observed signal recorded by the microphone while $h(n)$ and $s(n)$ denotes the acoustic impulse response (discussed in Section 2.1) and the probe signal, respectively. Furthermore, $x(n) = h(n) * s(n)$. The background noise $v(n)$ is represented as an additive white Gaussian noise (AWGN). The background noise assumption of AWGN is useful to formulate a mathematically feasible model so that close-form solutions to the estimators are found [148].

2.1 Acoustic impulse response (AIR)

The *acoustic impulse response* (AIR) or room impulse response (RIR) is the response of an environment to a sound signal, as measured by a microphone and a sound source, as shown in Fig. 7. It can be divided into three parts: the direct path and early reflections, followed by a stochastic long tail that represents reverberation or late reflections. The AIR is an important concept in robot audition, as it can provide valuable information about the acoustics of a space and help robots understand and respond to sounds in their environment. [149]. The shortest distance that a sound wave takes is called direct sound or direct-path component. When the emitted sound wave is reflected from an acoustic reflector, e.g., a wall. It is called early reflection which is a delayed version of the direct sound. The late reflections are sounds that get reflected by multiple walls before reaching a microphone. In this thesis, we are interested in estimating the distance of an acoustic reflector based on first-order early reflections. According to [150–152], first-order early reflection is useful to give information about the geometry of the environment. Based on these observations, we can decompose (1) as a sum of its direct path and early reflections. The transfer function between the emitter and the receiver can be formulated in terms of its gains and delay, the signal model can be rewritten as shown:

$$y(n) = \sum_{q=1}^{\infty} g_q s(n - \tau_q) + v(n), \qquad (2)$$

g_q is the gain or attenuation of the q-th reflection from the source to the microphone and τ_q is the TOA of the reflected signal while $*$ represents the convolution operator. In our definition of (2), the direct-path component corresponds to $q = 1$. Keeping the structure of AIR in mind, we can then rewrite (2) as the sum of the first R reflections as shown:

$$\begin{aligned} y(n) &= \sum_{q=1}^{R} g_q s(n - \tau_q) + d(n) + v(n), \qquad (3)\\ &= x(n) + v'(n), \qquad (4) \end{aligned}$$

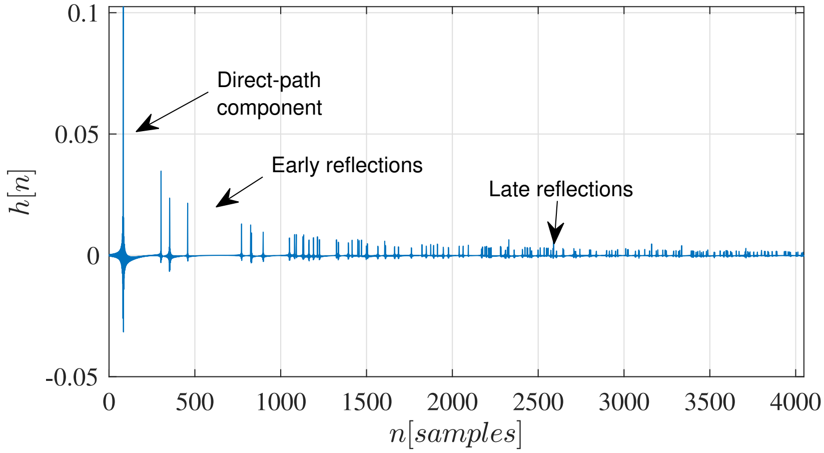


Fig. 7: Acoustic Impulse Response

where $d(n)$ is the stochastic and dense tail of the late reflections shown in Fig. 7. Hence, if N samples are taken then we can represent the model in (3) as shown:

$$\begin{aligned} \mathbf{y}(n) &= \sum_{q=1}^{R} g_q \mathbf{s}(n-\tau_q) + \mathbf{d}(n) + \mathbf{v}(n), && (5) \\ &= \mathbf{x}(n) + \mathbf{v}'(n), && (6) \end{aligned}$$

This model in (5) is used throughout this thesis to facilitate time of arrival (TOA) estimation. The task at hand is to estimate the TOA, τ, which can be used to infer the distance of an acoustic reflector to a robotic platform, In the following sections, we will review different techniques in the literature for TOA estimations.

2.2 Summary

We proposed a mathematical model of acoustic echo based on the structure of AIR. According to the structure of AIR, the direct-path component is the shortest distance a sound wave takes from a source to the receiver. The estimation of the direct-path component is useful in applications that require identifying and estimating the location of active sources, e.g., two human speakers within an environment. On the other hand, the early reflections refer to the distance a sound wave takes from the source to an acoustic reflector before reaching the receivers, while the late reflection is a stochastic long tail that refers to sound waves bouncing off multiple reflections before reaching the microphone. Moreover, early reflections are useful to infer the shape of the room. It is this property of the early reflections that are of interest to us within this thesis, hence a mathematical model based on the AIR structure is formulated using the gain and TOA of the acoustic echoes. In Chapter 3, we will investigate numerous statistical parametric methods for TOA/DOA estimation that are found in the literature. These two values are essential when constructing a spatial map of an environment.

3 Statistical Parametric Estimation

For a robot to generate a map of an environment, two values are required: it needs to know the distance of an acoustic reflector and the direction of an acoustic reflector. Combining these two values enable a robot to represent or map an environment. In this thesis, we have only investigated 2D mapping, and an extension to 3D representation of the environment is left for future work. In the following subsections, we will investigate different TOA and DOA estimation techniques. Later, we will discuss how these two parameters are used to represent acoustic maps of an environment.

3.1 Distance estimation

In literature, distance estimation of a wall is usually done from the TOA information of a sound signal [153–155]. The process involves estimating an AIR of an environment as shown in Fig. 7 and then using the peak-picking algorithm to get the time-of-arrival (TOA) information of acoustic echoes. Assuming far-field assumption and assuming that the distance of the speed of sound remains constant then distance estimation is straightforward. The distance can be estimated directly from the TOAs $d = \frac{c\tau}{2}$, where τ is the TOA of an acoustic echo and c is the speed of sound. However, within the context of robot audition, estimating the AIR and extracting the time of arrival (TOA) of each sound reflection can be a time-consuming process, especially if the robot is moving around and experiencing changes in its acoustic environment. This is because the AIR will need to be re-estimated each time the robot moves to a new location. As a result, it may be necessary to use more efficient methods for estimating the TOA of sound reflections in a robot audition system.

Along with TOAs, the second most commonly used feature for distance estimation is the time-difference-of-arrival (TDOA) between a pair of microphones. Ultrasonic sensors are commonly used in robotics to estimate the distance between a robot and an acoustic reflector, such as a wall or other nearby object. These sensors work by emitting high-frequency sound waves and measuring the time it takes for the echoes to return. This information can be used to help robots navigate their environment and avoid obstacles [6]. This requires attaching specialized transducers which could increase the overall cost of robotic platforms. Hence, estimating acoustic reflectors in the audible frequency range is an interesting and challenging problem to tackle. Moreover, the TDOA feature is also known as interaural time difference (ITP). This can be estimated by using zero-level-crossing (ZLC) [93] or onset time between each signal [156–158]. The frequency counterpart of ITP is interaural phase difference (IPD) and this is done by assuming a narrow-band signal.

An alternative approach to distance estimation in robot auditions is to use a data-driven approach. In this approach, features like inter-aural time difference (ITD), time of arrival (TOA), and time-difference of arrival (TDOA) are extracted from the signals observed at different distances and used to train a model for estimating distance. For instance, in [159], the researchers used an artificial neural network to localize a sound source. The input to the neural network in this approach is the coordinates of the microphone, and the features of the sound source are recorded at a different position within the environment. The neural network is trained using this input and output data to learn how to estimate the distance between the microphone and the sound source. In another example, researchers used a convolutional recurrent neural network (CRNN) to learn features related to distance [**?**, 160, 161]. The CRNN was trained by converting the recorded audio signals into a time-frequency representation, such as a Mel-spectrogram [162]. This allowed the network to learn how to estimate the distance to a sound source based on the characteristics of the audio signal. The use of CRNN and other machine-learning techniques can improve the accuracy and efficiency of distance estimation in robot audition systems.

Furthermore, as shown in [145, 163, 164], distance estimation could also be done by exploiting the robot's movement within an environment. This helps in constructing spatial maps of an environment as shown in [163], where a robot moves a predefined path to map spatial maps. A biomimetic sonar system based on bats' is proposed in [165] to navigate an environment using sonar sensing alone. The authors based their analysis on the number of conditional entropy [166] between the range data and the robot position.

Moreover, model-based approaches for distance estimation are found in the literature. In a model-based approach, a model of the observed signal is formulated which offers certain advantages. The advantage of the model-based approach is that it accommodates the background noise, reverberation, ego-noise, and the AIR of the environment which are then used for parameter estimation, e.g., TOAs, TDOAs, ITD, etc. The model-based approach provides a mechanism in the form of mathematical processes to incorporate underlying physical knowledge [167]. In this way, the model-based approach interprets results directly from observation. Model-based approaches are found abundantly in the literature and are used within different domains including the telecommunication domain. In [168], the researchers proposed a model based on Received Signal Strength Indication (RSSI) which enables distance estimation between sensor nodes in a Wireless Sensor Network (WSN). In audio signal processing, models of speech and noise signals are used to estimate the fundamental frequency and number of harmonics. The fundamental frequency is subsequently used to enhance the periodic signals [169]. Additionally, parameter estimation is also used in signal compression [**?**, 170–172], signal modification [173, 174], and so on.

The model of the sound source is used in the literature to estimate POIs and could be resolved using various data optimization techniques such as non-linear least squares (NLS) methods and expectation-maximization (EM) methods. NLS is a type of data optimization technique that fits multiple observations into a model which are non-linear. NLS estimators are found in many domains including in audio signal processing. It is used for SSL to estimate the TOA of the sound source within an environment. For example, in [175], the researchers jointly estimate the DOA as well as the pitch of the sound source in the presence of background noise. Moreover, in [176], the researchers estimated the harmonics of an acoustic signal. Furthermore, a TDOA-based estimator using the NLS method was also proposed in [177]. Another parameter estimation method within audio signal processing is the EM method. EM is an iterative method for performing maximum likelihood estimation in the presence of latent variables. For example, EM-based algorithms are used for pitch estimation as well as harmonic spectra estimation [178, 179]. Once distance estimation is done, the next step is to generate an acoustic map of an environment to estimate the direction from which the echo is originating. This is done by spatial filtering techniques, which are discussed next.

3.2 Estimation of Spatial filters for the direction of arrival (DOA) of acoustic echoes

While the TOA of an acoustic echo is used to infer the distance of an acoustic wall, the DOA of an acoustic echo is also required to make a spatial map of an environment. That is, the robotic platform needs to know where the acoustic echo is originating from. In acoustic signal processing, DOA estimation is done using multiple microphones. The signal model is formulated as follows:

$$\begin{aligned} y_m(n) &= h_m(n) * s(n) + v_m(n), && (7)\\ &= x_m(n) + v_m(n), && (8) \end{aligned}$$

where m represents the microphone number. Similarly, we can represent (4) for multi-channel scenario as:

$$y_m(n) = \sum_{q=1}^{R} g_{m,q} s(n - \tau_{1,q} - \eta_{m,q}) + v_{m,q}(n), \tag{9}$$

where η represents the TDOA of the q-th acoustic echoes between microphones and $\tau_{1,q}$ is the TOA of the q-th acoustic echoes from the reference microphone. The signal model in (9) is used throughout this thesis for DOA estimation, i.e., estimating the parameters η and τ, which will aid a robotic platform in generating an acoustic map of an environment. DOA estimation techniques are popularly used by the robotic community to enable robots to locate the position of acoustic sources. If the DOA of a signal is known then the observed signal can be preprocessed spatially to reduce noise from the signal. DOA estimation is important for other applications, e.g., autonomous vehicles [180], and automated camera steering [181, 182]. Many years of research have resulted in different methods for DOA estimation, e.g., neural networks, beamforming-based approaches, and subspace-based approaches. The TDOA between a pair of sensors or microphones is commonly used to estimate a sound source. Moreover, the most popular method of estimating TDOA is the cross-correlation technique [183] between the recorded signal and the probed signal. However, the dynamic range of the TDOA caused by distance variation is very small leading to non-linearity in close distances which could cause distance estimation errors [93]. Estimating TDOA using the cross-correlation method is also sensitive to reverberations as well as other noise sources [184]. Therefore, cross-correlation-based methods such as the GCC-PHAT-based TDOA method were also proposed in [185] for DOA estimation. GCC-PHAT consists of normalization that makes all frequencies have a magnitude of 1. This forces the correlated signal to have high peaks. Hence, it provides GCC-PHAT robustness against reverberation [186] and interfering sources [187]. GCC-PHAT is used in robotics to carry out acoustic map generation when a robot explores the environment [188]. Other applications of GCC-PHAT in SSL for service robots can be found in [185, 189–192].

To detect multiple reflections, spatial filtering techniques such as beamforming are often applied [193–196]. Beamforming is a spatial filtering technique that captures signals in an array of sensors or microphones. These signals are then weighted such that the output points to the direction where the signals are coming from, i.e., the DOA of the source signal. The advantage of this technique is that it electronically steers the microphones, such that the microphones focus on the direction of the source signal. Beamforming techniques are useful to attenuate inferring sources. The simplest form of the beamformer is a delay-and-sum beamformer (DSB) [197]. It is the simplest beamformer to build and has applications ranging from gunshot detection [198] to microwave imaging [199]. DSB belongs to a class of beamformers known as filter-and-sum beamformers (FSB). DSB artificially shifts the incoming signals at each microphone to counter the time difference and later the signals are all added together to obtain an output. The DSB can be implemented either in the time domain or frequency domain. In the time domain, the DSB can be implemented by introducing different delays on each microphone to steer the microphone array, while in the frequency domain phase shifts are applied on each frequency bin. Furthermore, the advantage of DSB is that it is computationally less intensive but the disadvantage is that it does not take into account the statistical property of the background noise. If the weights of the beamformers are required to be updated automatically then adaptive beamformers are employed. The minimum variance distortionless response (MVDR) beamformer follows a different approach. The MVDR beamformer was first proposed by Capon [200]. An MVDR beamformer steers the direction of the beamformer such that it enhances the desired observed signal [201, 202]. The MVDR beamformer first tries to minimize the interfering sources and background noise while maximizing the total output power. Moreover, unity gain constraint is applied to the MVDR beamformer such that the signal coming from a target direction is undistorted while the signals coming from other directions are minimized. Some researchers have proposed modifications to the MVDR beamformer such that multiple linear constraints could be applied to the beamformer to attenuate multiple interfering sources. This type of beamformer is known as a linear constraint minimum variance (LCMV) beamformer. The LCMV beamformer was originally proposed by Frost in 1972 as an implementation of MVDR in time-domain [203–205]. Therefore, both MVDR and LCMV beamformers lay the foundation of spatial filters which allow the signals from the target directions to remain undistorted while the signals from other directions are minimized.

In this thesis, we use an adaptive beamforming technique to localize acoustic echoes. This can be done by steering the beamformer in a different direction with the hope that the acoustic echo from a particular direction will be detected similar to how radar-based systems work. The output power from each steered direction is then plotted and the direction with maximum output power corresponds to the direction of the acoustic echo [206]. Other spatial filtering methods are also found in the literature, for example, subspace-based methods such as MUSIC [207] are popularly used to estimate the DOAs of acoustic sources. Moreover, it has been shown in the literature that subspace-based methods provide better DOA estimations compared to beamforming-based methods but they require a lot of processing power [208–210]. The DOA estimates from beamforming along with the TOA estimates, are useful to generate an acoustic map of an environment.

3.3 Environmental representation using sound

The TOA and the DOA estimates are the two main values that are useful to construct a spatial map. Numerous approaches to mapping an environment exist in the literature but the most popular and influential approach to environmental representation is the occupancy grid [211]. Although a $2D$ representation of an environment is still used within robotics, works on $3D$ mapping can also be found in the literature [212]. The advantage of incorporating an extra dimension for environment representation is that robots that are built to navigate on outdoor terrain require depth and elevation knowledge to move efficiently.

The most popular framework to enable a robot to localize its position in $3D$ space is the use of simultaneous localization and mapping (SLAM) which ensures the geometric consistency of a map. Traditionally, the SLAM framework is used with cameras to detect the presence of a landmark and is also known in the literature as a visual SLAM. The algorithm associates the landmark with the position of a robot and then tracks the location of the landmark as the robot moves within the environment. Hence, the visual SLAM algorithm helps generate a spatial map of an environment. However, if microphones are used instead of cameras and laser-based technologies, then a framework for acoustic SLAM can be developed. Using sound localization such as TOA and DOA estimation discussed earlier and combining it with the SLAM framework presents new challenges. For example, audio landmarks are not active all the time, which makes it difficult for a robotic system to determine the position of a sound source. Moreover, acoustic SLAM is beneficial when there are sound sources available within an environment.

In recent years, some researchers have proposed algorithms that enable robots equipped with microphones to navigate an environment as well as to interact with their environment, e.g., for HRI. Acoustic SLAM was proposed by [62] to carry out the DOA estimation of an unknown environment using acoustic signals alone. To overcome the limitations of needing permanent sound sources to act as landmarks, the researchers in [62] based their method on using random finite sets (RFS). Additionally, the movement of a robotic platform has also been exploited to enable the construction of a spatial map of an environment [145]

One drawback of passively localizing sound sources for map generation is that this method will only work if sound sources are present in an environment. To address this problem, some researchers propose attaching robotic platforms with a sound source, e.g., a loudspeaker, such that a robot probes the environment in a similar process used by bats when navigating. This is directly inspired by animals in nature, e.g., bats, and is known as active sound localization. Therefore, SLAM-based algorithms with an active sound localization approach have been proposed in [145, 164]. Active acoustic SLAM-based algorithms are useful for developing $3D$ imaging sonar sensors for robotic platforms [213, 214].

3.4 Summary

In this chapter, we investigated different TOA estimation and spatial filtering techniques and audio signal processing tools that are used within robotics to construct a spatial map of an environment. As discussed earlier, the traditional approach to TOA estimation requires extracting TOA information directly from the estimated AIR using a standard pick-picking approach. Other popular methods for TOA estimation are cross-correlation and data-driven approaches. Moreover, spatial filtering techniques such as beamforming are used for DOA estimations. In the literature, DSB, MVDR, and LCMV beamformers are popularly used to estimate the DOA of the acoustic source. DSB belongs to a category of beamformers that are classified as fixed beamformers while MVDR and LCMV beamformers belong to a category known as adaptive beamformers. Adaptive beamformers take the statistics of the background noise into account while maximizing the output power of the beamformers. That is, MVDR uses criteria such that unity gain is maintained in the direction of the true acoustic source which minimizes the variance of other interfering sources. Similarly, additional constraints are used in LCMV beamformers which nullify multiple interfering sources. Additionally, parameter estimation is done by incorporating a model of the acoustic signal to estimate the POI. Two popular methods of parameter estimations are found in the literature: the NLS method and the EM-based method.

4 Contributions

This thesis addresses the need for algorithms that contribute to the construction of spatial maps using echolocation alone. This modality complements existing state-of-the-art techniques such as lidar and camera-based systems to detect glass/transparent surfaces which are typically found in office environments. The main body of this thesis, which is constituted by papers A-G, contributes to the design of new algorithms and techniques used to estimate acoustic echoes to construct a spatial map of an indoor environment. Papers A and B propose two estimators that are derived based on the signal model for POI estimation: a nonlinear least squares (NLS) estimator and an expectation-maximization (EM) method. Multiple EM-based methods are proposed in paper C and the robustness of the EM-based methods are evaluated under different noise conditions. In paper D, the NLS method was extended to incorporate and implement on a robotic platform and successfully exploit the movement of the robotic platform to generate a spatial map of an indoor environment. In paper E, we proposed a novel way to estimate the acoustic reflectors using the ego-noise of the robotic platforms, e.g., drones. In paper F two algorithms

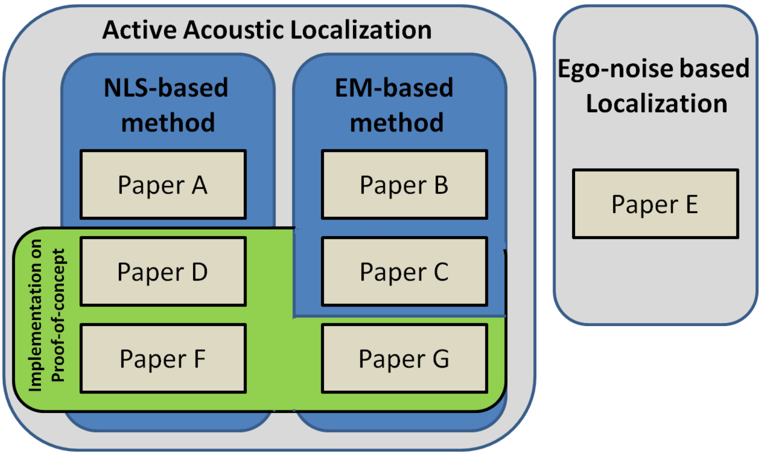

Active Acoustic Localization
NLS-based method
EM-based method
Paper A
Paper B
Implementation on Proof-of-concept
Paper D
Paper C
Paper F
Paper G
Ego-noise based Localization
Paper E

Fig. 8: Relationship of papers A-G. Papers A, D, and F are based on the NLS method while papers B, C and G are based on the EM method

based on NLS methods are proposed for estimating the TOA and the DOA of an acoustic reflector. Finally, in paper G, we proposed a robust EM estimator for estimating the nonlinearity of the audio systems, e.g., microphone and loudspeaker, for TOA and DOA estimation. All the papers in this thesis (except Paper E) follow a common procedure of probing the environment with a known sound. This is categorized as active acoustic localization while paper E is categorized separately as ego-noise-based localization. This is summarized in Fig. 8. The technical contribution of each paper are discussed below.

Paper A: This paper presents a frequency domain method based on the nonlinear least squares (NLS) method for estimating TOAs. In the literature, the TOAs are typically estimated from an estimated acoustic impulse response (AIR), but this is a computationally expensive process that is not well-suited for robotic platforms. In this paper, we propose estimating the TOA directly from an observed signal, which can be done more efficiently and effectively on robotic platforms. The estimator is based on the acoustic signal model which incorporates background noise, interfering sources, and ego-noise of the robotic platform. To estimate multiple TOAs which correspond to multiple acoustic reflectors, a cyclic approach, e.g., relaxation algorithm (RELAX) [215], is used. According to the simulation results, the proposed method could detect the location of an acoustic reflector up to a distance of 2 m.

Paper B: This paper presents a multichannel time domain method for estimating acoustic reflectors' location based on the expectation-maximization method. Instead of estimating TOAs from an estimated AIR, this paper proposes estimating the TOAs and the DOAs directly from the observed signal. The TOAs and the DOAs estimations are useful when inferring the distance and direction of an acoustic echo, respectively, to construct a spatial map. The simulated results show that the EM-based method can estimate the acoustic reflector's position up to a distance of 2 meters and offers robust estimation under low SNR compared to the peak-picking approach which is a method used to estimate the highest peaks from the estimated AIR.

Paper C: This journal paper extends the work in Paper B for spatial map construction by deriving multiple estimators for jointly estimating TOAs and DOAs directly from the observed signals. These different estimators are based on the expectation-maximization framework and are derived to be optimal under different conditions ranging from the simple white Gaussian noise scenario to scenarios with correlated and colored noise. Estimation of the covariance matrix directly from the observed signals and prewhitening of the observed signals before TOA/DOA estimation is also an important highlight of this paper. To make the evaluation more realistic, an analysis of the EM method in the presence of a faulty microphone was also discussed in this paper. The simulated results show that the proposed method could estimate the acoustic reflector's location under SNR of -10 dB with 60% accuracy.

Paper D: This paper presents an improvement to the work of Paper A by proposing a new algorithm that incorporates the movement of the robotic platform within the signal model. This method can enable a robot to construct a spatial map of an environment as it moves. This method was implemented and tested on a proof-of-concept robotic platform using a single microphone and loudspeaker. According to the evaluation, the NLS-based method could detect glass surfaces compared to the lidar sensor. The evaluation results also show that the proposed method could estimate the acoustic reflector's location up to a distance of 1.5 meters with an accuracy of 60%. The proposed method was found to be robust under low SNR conditions of 0 dB.

Paper E: This paper presents a method of estimating an acoustic reflector's location using only the ego-noise of the robotic platform, e.g., rotor noise. Instead of treating ego noise as a source of the problem, this work proposes using ego noise constructively. This is done by deriving the time difference of the echo (TDOE) estimator which is done by estimating the time difference of arrival (TDOA) between the direct sound source signal and its first echo in a given channel. Along with a TDOE estimator, this paper also proposes a probabilistic echo detector to distinguish echoes. This is achieved by deriving a classifier based on the generalized likelihood ratio test (GLRT). The simulated results show that the proposed estimator can detect an acoustic reflector's location up to a distance of 1 meters under a low signal to diffuse noise ratio of -10 dB and above.

Paper F: This journal paper presents two methods of estimating an acoustic reflector's location based on the nonlinear least square (NLS) method. The previous NLS-based methods in papers A and C are only used to infer distance and not the direction of acoustic echoes. This paper addresses the limitation of previous work and proposes a single-channel localization and mapping (ScLAM) method and a multi-channel localization and mapping (McLAM). The McLAM method jointly estimates the TOA and the DOA of an acoustic echo. Both methods are accompanied by an echo detector to distinguish whether an estimate belongs to an empty space or actually belongs to a wall. These two methods were implemented and tested on a proof-of-concept robotic platform. The evaluation results show that the McLAM method could detect the acoustic reflector's location under a signal-to-diffuse noise ratio (SDNR) of 10 dB with an accuracy of 80% of time. The McLAM method could also estimate the acoustic reflector's location up to a distance of 1.5 meters.

Paper G: This journal paper presents a method of improving the EM-based approach in papers B and C to robustly estimate the acoustic reflector's location. The previous work does not take into account the non-ideal response of acoustic systems, e.g., loudspeakers, for TOA/DOA estimation. The nonlinearity could hinder the TOA estimation of an acoustic echo. This paper proposes a method to estimate the transfer function or nonlinear response of an acoustic system using short filters. This method was evaluated and tested on two proof-of-concept setups. The experimental results show the proposed method could estimate multiple acoustic reflectors' locations up to a distance of 1.6 meters under low SNR of 0 dB.

5 Conclusion and direction of future research

This thesis investigates the perception technologies that are used in robotic platforms to acquire spatial information about the environment. The limitation of constructing a spatial map of an indoor environment with current technologies, e.g., camera and lidar, are addressed throughout this thesis. For instance, current sensing technologies are not suitable for detecting transparent and glass-like surfaces. Camera-based technologies are also susceptible to changing light conditions. This can hamper the accuracy of the spatial map of an environment. To address this limitation, active sound localization techniques based on non-linear least squares and expectation-maximization methods are proposed. The process involves probing the environment with a known signal and using the acoustic echoes to estimate the TOAs and DOAs of the signal. The proposed method could complement existing techniques for spatial map construction and helps in accurately representing an indoor environment.

As discussed in Section 4, the existing techniques of estimating the distance of a reflector, e.g., a wall, using acoustic signals are based on prior estimation of the AIR of an environment. The estimated AIR is then used with a peak-picking algorithm for estimating the TOAs of an acoustic echo. This is a computationally expensive process for the robotic platform because a new AIR will be required every time a robot moves to a new location. Hence, in this thesis, we propose two methods that can jointly estimate TOAs and DOAs directly from the observed signal. Jointly estimating TOAs and DOAs could enable a robotic platform to construct an acoustic spatial map of an environment. The current approach to estimating TOAs from AIR is not susceptible to changing background noise, interfering sources, the nonlinearity of acoustic systems, and the ego noise of the robotic platform. This makes the state-of-the-art approach unsuitable for distance and directional estimation under low SNR conditions.

To construct a spatial map of an environment, the contribution of this thesis starts with a single-channel approach to estimating the distance of an acoustic reflector using the NLS-based method (paper A). Since, estimating TOA was not sufficient for spatial map construction, hence a multi-channel EM-based method was proposed to estimate the TOA and the DOA of an acoustic echo using a uniform circular microphone array (paper B). The robustness of the multi-channel EM-based method was evaluated under different background noises, interfering sources, and the presence of a faulty microphone (paper C). Estimating nonlinearities of the acoustic system, e.g., the loudspeaker was addressed in paper F which can potentially improve the performance of the multi-channel EM method for joint TOA and DOA estimation. In the later stages of this research, the NLS-based methods were expanded to a multichannel approach for TOA/DOA estimation. Implementation on a robotic platform was done to exploit the movement of the robotic platform for spatial map generation. The performance of the NLS-based method in a practical setting is shown in papers D and F. The knowledge acquired from these papers gave way to another interesting approach for estimating acoustic reflectors' position. This was done by utilizing the ego-noise of the robotic platform (paper E). Compared to the other methods discusses so far, this approach does not require the use of loudspeakers to probe the environment.

The use of echolocation for spatial map generation has many benefits to society. For example, it could assist in the development of low-cost robotic platforms that can represent an environment. Such low-cost robots could be used in extreme environments, e.g., sewers, and underground tunnels. The use of the acoustic source location method could potentially be used to detect human survivors in search and rescue missions. Echolocation-based methods could also be used in autonomous vehicles to detect pedestrians. Furthermore, the method proposed in Paper E could potentially lead to the development of robotic platforms that can detect obstacles, e.g., walls, merely by utilizing a robot ego-noise

It is the opinion of the author that more research within active sound localization is required to enable robots to construct a spatial map of an environment using sound. The current algorithms are computationally expensive, however, in future iterations of this research, this could be optimized to enable robots to process audio data faster. The current research in this thesis was focused on beamforming techniques applied to robotic platforms, but subspace-based methods could be used in the future iteration of this research. Another direction of research could be to extend the approach of Paper E by investigating the structure of ego-noise of robotic platforms, e.g., drones. To extend the methods in paper E, the influence of airflow on the drone, and the varying speed of drones could also influence the TDOE estimation. Throughout this research, the acoustic echo propagation model employed was assumed to be far-field but if the robot is close to an acoustic reflector then it could be assumed to be near-field. This can change the derivation of the signal model. Moreover, the probe signal used is AWGN but experimentation with other probed signals that improve the accuracy and detection of the acoustic reflectors could also be investigated.

6 Appendix

The following equation from paper E (E. 12) is rewritten here since the derivation in the paper has a typo.

$$\frac{\delta J}{\delta \alpha} = -2\mathbf{y}^T \mathbf{D}_{\Delta\tau}\mathbf{x}_d + 2\mathbf{x}_d^T \mathbf{D}_{\Delta\tau}\mathbf{x}_d + 2\alpha \mathbf{x}_d^T \mathbf{D}_{\Delta\tau}^T \mathbf{D}_{\Delta\tau}\mathbf{x}_d = 0. \tag{10}$$

By observing that $\mathbf{D}_{\Delta\tau}^T \mathbf{D}_{\Delta\tau} = \mathbf{I}$, this becomes:

$$-2\left(\mathbf{y} - \mathbf{x}_d\right)^T \mathbf{D}_{\Delta\tau}\mathbf{x}_d + 2\alpha \|\mathbf{x}_d\|^2 = 0 \tag{11}$$

Hence,

$$\widehat{\alpha}(\Delta\tau) = \frac{\left(\mathbf{y} - \mathbf{x}_d\right)^T \mathbf{D}_{\Delta\tau}\mathbf{x}_d}{\|\mathbf{x}_d\|^2}. \tag{12}$$

References

[1] R. R. Schaller, "Moore's law: past, present and future," *IEEE spectrum*, vol. 34, no. 6, pp. 52–59, 1997.

[2] M. Rehm and E. André, "Catch me if you can: exploring lying agents in social settings," *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pp. 937–944, 2005.

[3] B. Endrass, E. André, and M. Rehm, "Towards culturally-aware virtual agent systems," *Handarticle of Research on Culturally-Aware Information Technology: Perspectives and Models*, pp. 412–430, 2011.

[4] M. Rehm, E. André, and M. Nischt, "Let's come together—social navigation behaviors of virtual and real humans," *International conference on intelligent technologies for interactive entertainment*, pp. 124–133, 2005.

[5] B. Tribelhorn and Z. Dodds, "Evaluating the roomba: A low-cost, ubiquitous platform for robotics research and education," *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pp. 1393–1399, 2007.

[6] U. Saqib and R. Kerstens, "Perceiving the world with sound," *IGI Global*, pp. 30–59, 9 2022.

[7] A. Loutfi and S. Coradeschi, "Smell, think and act: A cognitive robot discriminating odours," *Autonomous Robots*, vol. 20, no. 3, pp. 239–249, 2006.

[8] H. Miwa, T. Umetsu, A. Takanishi, and H. Takanohu, "Human-like robot head that has olfactory sensation and facial color expression," *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation*, vol. 1, pp. 459–464, 2001.

[9] G. Kowadlo and R. A. Russell, "To naively smell as no robot has smelt before," *IEEE Conference on Robotics, Automation and Mechatronics*, vol. 2, pp. 898–903, 2004.

[10] R. Rozas, J. Morales, and D. Vega, “Artificial smell detection for robotic navigation,” *Fifth International Conference on Advanced Robotics’ Robots in Unstructured Environments*, pp. 1730–1733, 1991.

[11] A. S. A. Yeon, R. Visvanathan, S. M. Mamduh, K. Kamarudin, L. Kamarudin, and A. Zakaria, “Implementation of behaviour based robot with sense of smell and sight,” *Procedia Computer Science*, vol. 76, pp. 119–125, 2015.

[12] G. Kowadlo and R. A. Russell, “Robot odor localization: a taxonomy and survey,” *The International Journal of Robotics Research*, vol. 27, no. 8, pp. 869–894, 2008.

[13] Z. Jie and H. Gunes, “Investigating taste-liking with a humanoid robot facilitator,” *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 131–136, 2020.

[14] B. Ciui, A. Martin, R. K. Mishra, T. Nakagawa, T. J. Dawkins, M. Lyu, C. Cristea, and R. S. J. Wang, “Chemical sensing at the robot fingertips: Toward automated taste discrimination in food samples,” *ACS sensors*, vol. 3, no. 11, pp. 2375–2384, 2018.

[15] H. Shimazu, K. Kobayashi, A. Hashimoto, and T. Kameoka, “Tasting robot with an optical tongue: Real time examining and advice giving on food and drink,” *Symposium on Human Interface and the Management of Information*, pp. 950–957, 2007.

[16] G. Bebis, D. Egbert, and M. Shah, “Review of computer vision education,” *IEEE Transactions on Education*, vol. 46, no. 1, pp. 2–21, 2003.

[17] A. Ahmed, A. Jalal, and A. A. Rafique, “Salient segmentation based object detection and recognition using hybrid genetic transform,” *2019 International Conference on Applied and Engineering Mathematics (ICAEM)*, pp. 203–208, 2019.

[18] P. Viola and M. Jones, “Robust real-time object detection,” *International journal of computer vision*, vol. 4, no. 34-47, p. 4, 2001.

[19] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, “Relation networks for object detection,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3588–3597, 2018.

[20] X. Song, S. Jiang, and L. Herranz, “Multi-scale multi-feature context modeling for scene recognition in the semantic manifold,” *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2721–2735, 2017.

[21] N. Ono, A. Saito, and S. Ando, “Design and experiments of bio-mimicry sound source localization sensor with gimbal-supported circular diaphragm,” *TRANSDUCERS ’03. 12th International Conference on Solid-State Sensors, Actuators and Microsystems. Digest of Technical Papers*, vol. 1, pp. 935–938 vol.1, 2003.

[22] J. Sohl-Dickstein, S. Teng, B. M. Gaub, C. C. Rodgers, C. Li, M. R. DeWeese, and N. S. Harper, “A device for human ultrasonic echolocation,” *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 6, pp. 1526–1534, 2015.

[23] G. Jones, “Echolocation,” *Current Biology*, vol. 15, no. 13, pp. R484–R488, 2005.

[24] J. A. Simmons and R. A. Stein, “Acoustic imaging in bat sonar: echolocation signals and the evolution of echolocation,” *Journal of comparative physiology*, vol. 135, no. 1, pp. 61–84, 1980.

[25] D. S. Edwards, R. Allen, T. Papadopoulos, D. Rowan, S. Y. Kim, and L. Wilmot-Brown, “Investigations of mammalian echolocation,” *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 7184–7187, 2009.

[26] L. J. Norman, C. Dodsworth, D. D. Foresteire, and L. Thaler, “Human click-based echolocation: Effects of blindness and age, and real-life implications in a 10-week training program,” *PloS one*, vol. 16, no. 6, p. e0252330, 2021.

[27] G. Bienvenu, “Signal sonar processing,” *2002 14th International Conference on Digital Signal Processing Proceedings. DSP 2002 (Cat. No.02TH8628)*, vol. 1, pp. 441–446 vol.1, 2002.

[28] M. Mansour, B. V. Smith, and D. R. Noaks, “Active sonar signal simulation,” *IEE Colloquium on Simulation Techniques Applied to Sonar*, pp. 2/1–2/4, 1988.

[29] S. Thrun, “Robotic mapping: A survey,” *Exploring artificial intelligence in the new millennium*, vol. 1, no. 1-35, p. 1, 2002.

[30] G. Lakemeyer and B. Nebel, “Exploring artificial intelligence in the new millennium,” *Morgan Kaufmann*, 2003.

[31] R. Siegwart, I. R. Nourbakhsh, and D. Scaramuzza, “Introduction to autonomous mobile robots,” *MIT press*, 2011.

[32] S. T. Pfister, “Algorithms for mobile robot localization and mapping, incorporating detailed noise modeling and multi-scale feature extraction,” *California Institute of Technology*, 2006.

[33] C. Rennie, R. Shome, K. E. Bekris, and A. F. A. F. De Souza, “A dataset for improved RGBD-based object detection and pose estimation for warehouse pick-and-place,” *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 1179–1185, 2016.

[34] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla, “Understanding real world indoor scenes with synthetic data,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4077–4085, 2016.

[35] M. Firman, “RGBD datasets: Past, present and future,” *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 19–31, 2016.

[36] N. Sünderhauf, F. Dayoub, S. McMahon, B. Talbot, R. Schulz, P. Corke, G. Wyeth, B. Upcroft, and M. Milford, “Place categorization and semantic mapping on a mobile robot,” pp. 5729–5736, 2016.

[37] M. Brucker, M. Durner, R. Ambruş, Z. C. Márton, A. Wendt, P. Jensfelt, K. O. Arras, and R. Triebel, “Semantic labeling of indoor environments from 3D RGB maps,” *Proc. IEEE Int. Conf. Robotics, Automation.*, pp. 1871–1878, 2018.

[38] D. R. Faria, M. Vieira, C. Premebida, and U. Nunes, “Probabilistic human daily activity recognition towards robot-assisted living,” *2015 24th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pp. 582–587, 2015.

[39] T. Fong, I. Nourbakhsh, and K. Dautenhahn, “A survey of socially interactive robots,” *Robotics and autonomous systems*, vol. 42, no. 3-4, pp. 143–166, 2003.

[40] J. Saarinen, H. Andreasson, and A. J. Lilienthal, “Independent markov chain occupancy grid maps for representation of dynamic environment,” *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3489–3495, 2012.

[41] A. Asvadi, L. Garrote, C. Premebida, P. Peixoto, and U. J. Nunes, “Multimodal vehicle detection: fusing 3D-LIDAR and color camera data,” *Pattern Recognition Letters*, vol. 115, pp. 20–29, 2018.

[42] C. Premebida and U. Nunes, “Fusing lidar, camera and semantic information: A context-based approach for pedestrian detection,” *The International Journal of Robotics Research*, vol. 32, no. 3, pp. 371–384, 2013.

[43] C. Dondrup, N. Bellotto, F. Jovan, and M. Hanheide, “Real-time multisensor people tracking for human-robot spatial interaction,” *Proc. IEEE Int. Conf. Robotics, Automation.*, 2015.

[44] R. Manduchi, A. Castano, A. Talukder, and L. Matthies, “Obstacle detection and terrain classification for autonomous off-road navigation,” *Autonomous robots*, vol. 18, no. 1, pp. 81–102, 2005.

[45] H. Andreasson, M. Magnusson, and A. Lilienthal, “Has somethong changed here? autonomous difference detection for security patrol robots,” *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3429–3435, 2007.

[46] C. Thorpe, O. Clatz, D. Duggins, J. Gowdy, R. MacLachlan, J. R. Miller, C. Mertz, M. Siegel, C. Wang, and T. Yata, “Dependable perception for robots,” 2001.

[47] A. Kelly, A. Stentz, O. Amidi, M. Bode, D. Bradley, A. Diaz-Calderon, M. Happold, H. Herman, R. Mandelbaum, and T. Pilarski, “Toward reliable off road autonomous vehicles operating in challenging environments,” *The International Journal of Robotics Research*, vol. 25, no. 5-6, pp. 449–483, 2006.

[48] T. Peynot, J. Underwood, and S. Scheding, “Towards reliable perception for unmanned ground vehicles in challenging conditions,” *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1170–1176, 2009.

[49] M. P. Gerardo-Castro, T. Peynot, F. Ramos, and R. Fitch, “Robust multiple-sensing-modality data fusion using gaussian process implicit surfaces,” *17th International Conference on Information Fusion (FUSION)*, pp. 1–8, 2014.

[50] H. B. Mitchell, “Data fusion: concepts and ideas,” *Springer Science & Business Media*, 2012.

[51] L. J. Mohan and J. Ignatious, “Navigation of mobile robot in a warehouse environment,” *2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR)*, pp. 1–5, 2018.

[52] P. Beinschob and C. Reinke, “Strategies for 3D data acquisition and mapping in large-scale modern warehouses,” *2013 IEEE 9th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 229–234, 2013.

[53] D. Hoang, T. Stoyanov, and A. J. Lilienthal, “Object-RPE: Dense 3D reconstruction and pose estimation with convolutional neural networks for warehouse robots,” *2019 European Conference on Mobile Robots (ECMR)*, pp. 1–6, 2019.

[54] P. Gupta, J. Jost, and B. Bordihn, “Multi-robot mapping for optically guided vehicles,” *2018 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pp. 1–8, 2018.

[55] W. Budiharto, V. Andreas, J. S. Suroso, A. A. S. Gunawan, and E. Irwansyah, “Development of tank-based military robot and object tracker,” *2019 4th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*, pp. 221–224, 2019.

[56] J. Zuo, J. Chen, Z. Li, Z. Li, Z. Liu, and Z. Han, “Research on maritime rescue uav based on beidou cnss and extended square search algorithm,” *2020 International Conference on Communications, Information System and Computer Engineering (CISCE)*, pp. 102–106, 2020.

[57] J. Suarez and R. Murphy, “A survey of animal foraging for directed, persistent search by rescue robotics,” *2011 IEEE International Symposium on Safety, Security, and Rescue Robotics*, pp. 314–320, 2011.

[58] C. Premebida, R. Ambrus, and Z. C. Marton, “Intelligent robotic perception systems,” *Applications of Mobile Robots*, 2018.

[59] S. Thrun, “An approach to learning mobile robot navigation,” *Robotics and Autonomous systems*, vol. 15, no. 4, pp. 301–319, 1995.

[60] R. C. Smith and P. Cheeseman, “On the representation and estimation of spatial uncertainty,” *The international journal of Robotics Research*, vol. 5, no. 4, pp. 56–68, 1986.

[61] R. Smith, M. Self, and P. Cheeseman, “Estimating uncertain spatial relationships in robotics,” *Autonomous robot vehicles*, pp. 167–193, 1990.

[62] C. Evers and P. A. Naylor, “Acoustic SLAM,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1484–1498, 2018.

[63] I. Dokmanić, L. Daudet, and M. Vetterli, “From acoustic room reconstruction to slam,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016.

[64] S. M. LaValle, “Planning algorithms,” *Cambridge university press*, 2006.

[65] R. Datouo, F. B. Motto, B. E. Zobo, A. Melingui, I. Bensekrane, and R. Merzouki, “Optimal motion planning for minimizing energy consumption of wheeled mobile robots,” *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 2179–2184, 2017.

[66] J. Yao, C. Lin, X. Xie, A. J. Wang, and C. Hung, “Path planning for virtual human motion using improved A* star algorithm,” *2010 Seventh international conference on information technology: new generations*, pp. 1154–1158, 2010.

[67] C. Wang, L. Wang, J. Qin, Z. Wu, L. Duan, Z. Li, M. Cao, X. Ou, X. Su, and W. Li, “Path planning of automated guided vehicles based on improved A-star algorithm,” *2015 IEEE International Conference on Information and Automation*, pp. 2071–2076, 2015.

[68] T. Nayl, M. Q. Mohammed, and S. Q. Muhamed, “Obstacles avoidance for an articulated robot using modified smooth path planning,” *2017 international conference on computer and applications (ICCA)*, pp. 185–189, 2017.

[69] X. Huang, Q. Jia, and G. Chen, “Collision-free path planning method with learning ability for space manipulator,” *2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 1790–1795, 2017.

[70] M. A. Baumann, D. C. Dupuis, S. Léonard, E. A. Croft, and J. J. Little, “Occlusion-free path planning with a probabilistic roadmap,” *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2151–2156, 2008.

[71] R. M. C. Santiago, A. L. D. Ocampo, A. T. Ubando, A. A. Bandala, and E. P. Dadios, “Path planning for mobile robots using genetic algorithm and probabilistic roadmap,” *2017IEEE 9th international conference on humanoid, nanotechnology, information technology, communication and control, environment and management (HNICEM)*, pp. 1–5, 2017.

[72] N. Kumar, Z. Vámossy, and Z. M. Szabó-Resch, "Robot path pursuit using probabilistic roadmap," *2016 IEEE 17th International Symposium on Computational Intelligence and Informatics (CINTI)*, pp. 000 139–000 144, 2016.

[73] M. Korkmaz and A. Durdu, "Comparison of optimal path planning algorithms," *2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET)*, pp. 255–258, 2018.

[74] Z. Elmi and M. Ö. Efe, "Multi-objective grasshopper optimization algorithm for robot path planning in static environments," *2018 IEEE International Conference on Industrial Technology (ICIT)*, pp. 244–249, 2018.

[75] B. Hernández and E. Giraldo, "A review of path planning and control for autonomous robots," *2018 IEEE 2nd Colombian Conference on Robotics and Automation (CCRA)*, pp. 1–6, 2018.

[76] L. C. Santos, F. N. Santos, E. J. Solteiro Pires, A. Valente, P. Costa, and S. Magalhães, "Path planning for ground robots in agriculture: a short review," *2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pp. 61–66, 2020.

[77] R. S. Pol and M. Murugan, "A review on indoor human aware autonomous mobile robot navigation through a dynamic environment survey of different path planning algorithm and methods," *2015 International Conference on Industrial Instrumentation and Control (ICIC)*, pp. 1339–1344, 2015.

[78] M. A. Goodrich, A. C. Schultz *et al.*, "Human–robot interaction: a survey," *Foundations and Trends® in Human–Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2008.

[79] S. Thrun, "Toward a framework for human-robot interaction," *Human–Computer Interaction*, vol. 19, no. 1-2, pp. 9–24, 2004.

[80] T. B. Sheridan, "Human–robot interaction: status and challenges," *Human factors*, vol. 58, no. 4, pp. 525–532, 2016.

[81] A. Davids, "Urban search and rescue robots: from tragedy to technology," *IEEE Intelligent systems*, vol. 17, no. 2, pp. 81–83, 2002.

[82] F. Matsuno and S. Tadokoro, "Rescue robots and systems in japan," *2004 IEEE International Conference on Robotics and Biomimetics*, pp. 12–20, 2004.

[83] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots." Springer, 1986, pp. 396–404.

[84] I. Ulrich and J. Borenstein, "VFH+: Reliable obstacle avoidance for fast mobile robots," *Proceedings. 1998 IEEE international conference on robotics and automation (Cat. No. 98CH36146)*, vol. 2, pp. 1572–1577, 1998.

[85] Y. Bando, H. Suhara, M. Tanaka, T. Kamegawa, K. Itoyama, K. Yoshii, F. Matsuno, and H. Okuno, "Sound-based online localization for an in-pipe snake robot," *2016 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pp. 207–213, 2016.

[86] S. Widodo, T. Shiigi, N. Hayashi, H. Kikuchi, K. Yanagida, Y. Nakatsuchi, Y. Ogawa, and N. Kondo, "Moving object localization using sound-based positioning system with doppler shift compensation," *Robotics*, vol. 2, no. 2, pp. 36–53, 2013.

[87] L. Marchegiani and P. Newman, "Learning to listen to your ego-(motion): Metric motion estimation from auditory signals," *Annual Conference Towards Autonomous Robotic Systems*, pp. 247–259, 2018.

[88] H. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2015, pp. 5610–5614, 08 2015.

[89] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," *AAAI/IAAI*, pp. 832–839, 2000.

[90] S. Argentieri, A. Portello, M. Bernard, P. Danes, and B. Gas, "Binaural systems in robotics." Springer, 2013, pp. 225–253.

[91] R. A. Brooks, C. Breazeal, M. Marjanović, B. Scassellati, and M. M. Williamson, "The cog project: Building a humanoid robot," *International workshop on computation for metaphors, analogy, and agents*, pp. 52–87, 1998.

[92] S. Hashimoto, S. Narita, H. Kasahara, A. Takanishi, S. Sugano, K. Shirai, T. Kobayashi, H. Takanobu, T. Kurata, K. Fujiwara, T. Matsuno, T. Kawasaki, and K. Hoashi, "Humanoid robot-development of an information assistant robot hadaly," *Proceedings 6th IEEE International Workshop on Robot and Human Communication. RO-MAN'97 SENDAI*, pp. 106–111, 1997.

[93] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi, and N. Sugie, "A model-based sound localization system and its application to robot navigation," *Robotics and autonomous systems*, vol. 27, no. 4, pp. 199–209, 1999.

[94] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, no. 4, pp. 297–336, 1994.

[95] Y. Sasaki, N. Hatao, K. Yoshii, and S. Kagami, "Nested igmm recognition and multiple hypothesis tracking of moving sound sources for mobile robot audition," *Proc. IEEE Int. Conf. Intell., Robot, Automation.*, pp. 3930–3936, 2013.

[96] M. R. Pimpale, S. Therese, and V. Shinde, "A survey on: Sound source separation methods," *International Journal*, vol. 3, no. 11, pp. 580–584, 2016.

[97] N. Q. K.Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[98] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2009.

[99] P. Comon and C. Jutten, "Handbook of blind source separation: Independent component analysis and applications," 2010.

[100] S. T. Roweis, "One microphone source separation," *NIPS*, vol. 13, no. 2000, 2000.

[101] S. Bensaid, A. Schutz, and D. T. M. Slock, "Single microphone blind audio source separation using em-kalman filter and short+ long term ar modeling," *International Conference on Latent Variable Analysis and Signal Separation*, pp. 106–113, 2010.

[102] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *International journal of neural systems*, vol. 25, no. 01, p. 1440003, 2015.

[103] S. Shabani and Y. Norouzi, "Speech recognition using principal components analysis and neural networks," *2016 IEEE 8th International Conference on Intelligent Systems (IS)*, pp. 90–95, 2016.

[104] S. S. Agrawal, N. Prakash, and A. Jain, “Transformation of emotion based on acoustic features of intonation patterns for hindi speech,” *African Journal of Mathematics and Computer Science Research*, vol. 3, no. 10, pp. 255–266, 2010.

[105] A. Madan and D. Gupta, “Speech feature extraction and classification: A comparative review,” *International Journal of computer applications*, vol. 90, no. 9, 2014.

[106] C. H. Lee, F. K. Soong, and B. H. Juang, “A segment model based approach to speech recognition,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 501–502, 1988.

[107] M. Russo, M. Stella, M. Sikora, and V. Pekić, “Robust cochlear-model-based speech recognition,” *Computers*, vol. 8, no. 1, p. 5, 2019.

[108] H. Seki, K. Yamamoto, and S. Nakagawa, “A deep neural network integrated with filterbank learning for speech recognition,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 5480–5484, 2017.

[109] L. Chai, J. Du, D. Y. Liu, Y. H. Tu, and C. H. Lee, “Acoustic modeling for multi-array conversational speech recognition in the chime-6 challenge,” *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 912–918, 2021.

[110] B. Jolad and R. Khanai, “An art of speech recognition: A review,” *2019 2nd International Conference on Signal Processing and Communication (ICSPC)*, pp. 31–35, 2019.

[111] A. S. Spanias and F. H. Wu, “Speech coding and speech recognition technologies: a review,” *1991., IEEE International Sympoisum on Circuits and Systems*, pp. 572–577 vol.1, 1991.

[112] T. Barman and N. Deb, “State of the art review of speech recognition using genetic algorithm,” *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, pp. 2944–2946, 2017.

[113] A. Schmidt, A. Deleforge, and W. Kellermann, “Ego-noise reduction using a motor data-guided multichannel dictionary,” *iros*, pp. 1281–1286, 2016.

[114] L. Wang and A. Cavallaro, “Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles,” *IEEE Sensors Journal*, vol. 17, no. 8, pp. 2447–2455, 2017.

[115] M. Guzik, K. Kowalczyk, S. Woźniak, M. Fraś, K. Juros, D. Kaczor, and P. Walas, “Acoustic source localization using drone-embedded microphone array,” *The Journal of the Acoustical Society of America*, vol. 146, no. 4, pp. 3058–3059, 2019.

[116] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. Imura, “Ego noise suppression of a robot using template subtraction,” *Proc. IEEE Int. Conf. Intell., Robot, Automation.*, pp. 199–204, 2009.

[117] A. Schmidt, H. W. L. W., and W. Kellermann, “A novel ego-noise suppression algorithm for acoustic signal enhancement in autonomous systems,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 6583–6587, 2018.

[118] A. Ito, T. Kanayama, M. Suzuki, and S. Makino, “Internal noise suppression for speech recognition by small robots,” *Ninth European Conference on Speech Communication and Technology*, 2005.

[119] A. Pico, G. Schillaci, V. V. Hafner, and B. Lara, “How do i sound like? forward models for robot ego-noise prediction,” *Joint IEEE Int. Conf. on Development and Learning and Epigenetic Robotics*, pp. 246–251, 2016.

[120] A. D. U. Saqib and J. R. Jensen, "Detecting acoustic reflectors using a robot's ego-noise," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 1–5, 2021.

[121] P. M. Schultheiss and K. Wagner, "Active and passive localization: Similarities and differences." Springer, 1989, pp. 215–232.

[122] W. Burgard, D. Fox, and S. Thrun, "Active mobile robot localization," *IJCAI*, pp. 1346–1352, 1997.

[123] H. Heli and H. R. Abutalebi, "Localization of multiple simultaneous sound sources in reverberant conditions using blind source separation methods," *2011 International Symposium on Artificial Intelligence and Signal Processing (AISP)*, pp. 1–5, 2011.

[124] T. Wang and Y. Choy, "An approach for sound sources localization and characterization using array of microphones," *2015 International Conference on Noise and Fluctuations (ICNF)*, pp. 1–4, 2015.

[125] Y. Sasaki, Y. Tamai, S. Kagami, and H. Mizoguchi, "2d sound source localization on a mobile robot with a concentric microphone array," *2005 IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, pp. 3528–3533 Vol. 4, 2005.

[126] P. Jeyasingh and M. Mohamed Ismail, "Real-time multi source speech enhancement based on sound source separation using microphone array," *2018 Conference on Emerging Devices and Smart Systems (ICEDSS)*, pp. 183–187, 2018.

[127] D. Morikawa, "Effect of interaural difference for localization of spatially segregated sound," *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 602–605, 2014.

[128] P. Ramezanpour and M. R. Mosavi, "DNN-based interference mitigation beamformer," *IET Radar, Sonar & Navigation*, vol. 14, no. 11, pp. 1788–1794, 2020.

[129] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "DNN-based speech mask estimation for eigenvector beamforming," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 66–70, 2017.

[130] S. E. Chazan, J. Goldberger, and S. Gannot, "DNN-based concurrent speakers detector and its application to speaker extraction with lcmv beamforming," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6712–6716, 2018.

[131] C. Deng, H. Song, Y. Zhang, Y. Sha, and X. Li, "DNN-based mask estimation integrating spectral and spatial features for robust beamforming," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 4647–4651, 2020.

[132] X. Zhang, "Deep ad-hoc beamforming," *Computer Speech & Language*, vol. 68, p. 101201, 2021.

[133] A. Quazi, "An overview on the time delay estimate in active and passive systems for target localization," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 527–533, 1981.

[134] X. Wu, M. Abrahantes, and M. Edgington, "MUSSE: A designed multi-ultrasonic-sensor system for echolocation on multiple robots," *2016 Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*, pp. 79–83, 2016.

[135] J. H. Lim and J. Leonard, "Mobile robot relocation from echolocation constraints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 9, pp. 1035–1041, 2000.

[136] M. Wang, H. Tamimi, and A. Zell, “Robot navigation using biosonar for natural landmark tracking,” *2005 International Symposium on Computational Intelligence in Robotics and Automation*, pp. 3–7, 2005.

[137] X. Wu, P. D’Orazio, M. Edgington, and M. Abrahantes, “Robotics echolocation test platform,” *2015 IEEE International Conference on Electro/Information Technology (EIT)*, pp. 558–562, 2015.

[138] T. Moreira, J. Lima, P. Costa, and M. Cunha, “Low-cost sonar based on the echolocation.” *ICINCO (1)*, pp. 818–825, 2019.

[139] U. Saqib, S. Gannot, and J. R. Jensen, “Estimation of acoustic echoes using expectation-maximization methods,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–15, 2020.

[140] N. Riopelle, P. Caspers, and D. Sofge, “Terrain classification for autonomous vehicles using bat-inspired echolocation,” *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6, 2018.

[141] I. Eliakim, Z. Cohen, G. Kosa, and Y. Yovel, “A fully autonomous terrestrial bat-like acoustic robot,” *PLoS computational biology*, vol. 14, no. 9, p. e1006406, 2018.

[142] J. Steckel and H. Peremans, “BatSLAM: Simultaneous localization and mapping using biomimetic sonar,” *PloS one*, vol. 8, no. 1, p. e54076, 2013.

[143] M. J. Milford, G. F. Wyeth, and D. Prasser, “RatSLAM: a hippocampal model for simultaneous localization and mapping,” *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA ’04. 2004*, vol. 1, pp. 403–408 Vol.1, 2004.

[144] M. Chen and W. Hu, “Research on batslam algorithm for uav based on audio perceptual hash closed-loop detection,” *International journal of pattern recognition and artificial intelligence*, vol. 33, no. 01, p. 1959002, 2019.

[145] M. Kreković, I. Dokmanić, and M. Vetterli, “Echoslam: Simultaneous localization and mapping with acoustic echoes,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 11–15, 2016.

[146] B. Zhou, M. Elbadry, R. Gao, and F. Ye, “Batmapper: Acoustic sensing based indoor floor plan construction using smartphones,” *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 42–55, 2017.

[147] R. Worley, Y. Yu, and S. Anderson, “Acoustic echo-localization for pipe inspection robots,” *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 160–165, 2020.

[148] S. M. Kay, “Fundamentals of statistical signal processing: estimation theory,” *Prentice Hall PTR*, 1993.

[149] G. Moschioni, “A new method for measurement of early sound reflections in theaters and halls,” *IMTC/2002. Proceedings of the 19th IEEE Instrumentation and Measurement Technology Conference*, vol. 1, pp. 425–430 vol.1, 2002.

[150] B. Gunel, “Room shape and size estimation using directional impulse response measurements,” *Proc. Forum Acusticum Sevilla*, 2002.

[151] F. Antonacci, A. Sarti, and S. Tubaro, “Geometric reconstruction of the environment from its response to multiple acoustic emissions,” *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2822–2825, 2010.

[152] J. Filos, E. A. P. Habets, and P. Naylor, “A two-step approach to blindly infer room geometries,” *Int. Workshop Acoustic Signal Enhancement*, 2010.

[153] Y. E. Baba, A. Walther, and E. A. P. Habets, “3D room geometry inference based on room impulse response stacks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 857–872, 2017.

[154] S. Tervo and T. Tossavainen, “3D room geometry estimation from measured impulse responses,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 513–516, 2012.

[155] A. H. Moore, M. Brookes, and P. A. Naylor, “Room geometry estimation from a single channel acoustic impulse response,” *Proc. European Signal Processing Conf.*, 2013.

[156] A. M. Flynn, R. A. Brooks, W. M. W. III, and D. S. Barrett, “Squirt: The prototypical mobile robot for autonomous graduate students,” Massachusetts Institute of Technology Cambridge Artificial Intelligence Lab, Tech. Rep., 1989.

[157] F. Wang, Y. Takeuchi, N. Ohnishi, and N. Sugie, “A mobile robot with active localization and discrimination of a sound source,” *Journal of the Robotics Society of Japan*, vol. 15, no. 2, pp. 223–229, 1997.

[158] K. Nagashima, T. Yoshiike, A. Konno, M. Inaba, and H. Inoue, “Attention-based interaction between human and the robot chiye,” *Proceedings 6th IEEE International Workshop on Robot and Human Communication. RO-MAN’97 SENDAI*, pp. 100–105, 1997.

[159] P. Kumarakulasingam and A. Agah, “Neural network-based single sensor sound localization using a mobile robot,” *Intelligent Automation & Soft Computing*, vol. 14, no. 1, pp. 89–103, 2008.

[160] M. Yiwere and E. J. Rhee, “Sound source distance estimation using deep learning: An image classification approach,” *Sensors*, vol. 20, no. 1, p. 172, 2020.

[161] D. Tang, B. Qin, and T. Liu, “Document modeling with gated recurrent neural network for sentiment classification,” *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1422–1432, 2015.

[162] S. S. Stevens, J. Volkmann, and E. B. Newman, “A scale for the measurement of the psychological magnitude pitch,” *The journal of the acoustical society of america*, vol. 8, no. 3, pp. 185–190, 1937.

[163] L. Nguyen, J. V. Miro, and X. Qiu, “Can a robot hear the shape and dimensions of a room?” *Proc. IEEE Int. Conf. Intell., Robot, Automation.*, pp. 5346–5351, 2019.

[164] J. Steckel and H. Peremans, “Biomimetic sonar for biomimetic SLAM,” *SENSORS*, pp. 1–4, 2012.

[165] F. Schillebeeckx, F. D. Mey, D. Vanderelst, and H. Peremans, “Biomimetic sonar: Binaural 3D localization using artificial bat pinnae,” *The International Journal of Robotics Research*, vol. 30, no. 8, pp. 975–987, 2011.

[166] C. E. Shannon, “A mathematical theory of communication,” *ACM SIGMOBILE mobile computing and communications review*, vol. 5, no. 1, pp. 3–55, 2001.

[167] J. V. Candy, “Model-based signal processing,” *John Wiley & Sons*, vol. 36, 2005.

[168] R. K. Mahapatra and N. S. V. Shet, “Experimental analysis of rssi-based distance estimation for wireless sensor networks,” *2016 IEEE Distributed Computing, VLSI, Electrical Circuits and Robotics*, pp. 211–215, 2016.

[169] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative l model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 731–740, 2001.

[170] M. G. Christensen, "Estimation and modeling problems in parametric audio coding," *Aalborg Universitetsforlag*, 2005.

[171] H. Purnhagen and N. Meine, "Hiln-the mpeg-4 parametric audio coding tools," *2000 IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 3, pp. 201–204, 2000.

[172] J. R. Jensen, "Enhancement of periodic signals: with application to speech signals," *Aalborg University*, 2012.

[173] T. Quatieri and R. L. McAulay, "Speech transformations based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 6, pp. 1449–1464, 1986.

[174] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE transactions on speech and audio processing*, vol. 5, no. 5, pp. 389–406, 1997.

[175] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint doa and pitch estimation," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 5, pp. 923–933, 2013.

[176] M. Ghogho, A. Swami, and A. K. Nandi, "Non-linear least squares estimation for harmonics in multiplicative and additive noise," *Signal Processing*, vol. 78, no. 1, pp. 43–60, 1999.

[177] T. Qiao, Y. Zhang, and H. Liu, "Nonlinear expectation maximization estimator for tdoa localization," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 637–640, 2014.

[178] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Signal Processing*, vol. 88, no. 4, pp. 972–983, 2008.

[179] R. Badeau, V. Emiya, and B. David, "Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 3073–3076, 2009.

[180] H. K. .Aghajan and T. Kailath, "Sensor array processing techniques for super resolution multi-line-fitting and straight edge detection," *IEEE Transactions on Image Processing*, vol. 2, no. 4, pp. 454–465, 1993.

[181] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 187–190, 1997.

[182] N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video object localization and tracking," *IEEE signal processing magazine*, vol. 18, no. 1, pp. 22–31, 2001.

[183] M. S. Hosseini, A. Rezaie, and Y. Zanjireh, "Time difference of arrival estimation of sound source using cross correlation and modified maximum likelihood weighting function," *Scientia Iranica*, vol. 24, no. 6, pp. 3268–3279, 2017.

[184] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, 2017.

[185] D. Bechler, M. S. Schlosser, and K. Kroschel, “System for robust 3D speaker tracking using microphone array measurements,” *Proc. IEEE Int. Conf. Intell., Robot, Automation.*, vol. 3, pp. 2117–2122, 2004.

[186] M. S. Brandstein and H. F. Silverman, “A robust method for speech signal time-delay estimation in reverberant rooms,” *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 375–378, 1997.

[187] B. Kwon, Y. Park, and Y. S. Park, “Analysis of the gcc-phat technique for multiple sources,” *ICCAS 2010*, pp. 2070–2073, 2010.

[188] E. Martinson and A. Schultz, “Robotic discovery of the auditory scene,” *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pp. 435–440, 2007.

[189] V. M. Trifa, A. Koene, J. Morén, and G. Cheng, “Real-time acoustic source localization in noisy environments for human-robot multimodal interaction,” *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 393–398, 2007.

[190] E. Ben-Reuven and Y. Singer, “Discriminative binaural sound localization,” *Proceedings of the 15th International Conference on Neural Information Processing Systems*, pp. 1253–1260, 2002.

[191] R. Stiefelhagen, H. K. Ekenel, C. Fugen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel, “Enabling multimodal human–robot interaction for the karlsruhe humanoid robot,” *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 840–851, 2007.

[192] G. I. Parisi, J. Bauer, E. Strahl, and S. Wermter, “A multi-modal approach for assistive humanoid robots.” *Proc. IEEE Int. Conf. Intell., Robot, Automation.*, pp. 10–15, 2015.

[193] B. D. V. Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[194] J. Li and P. Stoica, “Robust adaptive beamforming,” *Wiley Online Library*, 2006.

[195] J. C. Chen, Y. Kung, and R. E. Hudson, “Source localization and beamforming,” *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 30–39, 2002.

[196] A. C. Luchies and B. C. Byram, “Deep neural networks for ultrasound beamforming,” *IEEE transactions on medical imaging*, vol. 37, no. 9, pp. 2010–2021, 2018.

[197] V. Perrot, M. Polichetti, F. Varray, and D. Garcia, “So you think you can das? a viewpoint on delay-and-sum beamforming,” *Ultrasonics*, vol. 111, p. 106309, 2021.

[198] A. L. L. Ramos, S. Holm, S. Gudvangen, and R. Otterlei, “Delay-and-sum beamforming for direction of arrival estimation applied to gunshot acoustics,” *Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense X*, vol. 8019, p. 80190U, 2011.

[199] T. Reimer, M. Solis-Nepote, and S. Pistorius, “The application of an iterative structure to the delay-and-sum and the delay-multiply-and-sum beamformers in breast microwave imaging,” *Diagnostics*, vol. 10, no. 6, p. 411, 2020.

[200] J. Capon, “High-resolution frequency-wavenumber spectrum analysis,” *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[201] D. A. Pados and G. N. Karystinos, “An iterative algorithm for the computation of the MVDR filter,” *IEEE Transactions On signal processing*, vol. 49, no. 2, pp. 290–300, 2001.

[202] J. Benesty, J. Chen, and Y. Huang, “A generalized MVDR spectrum,” *IEEE Signal Processing Letters*, vol. 12, no. 12, pp. 827–830, 2005.

[203] O. L. Frost, “An algorithm for linearly constrained adaptive array processing,” *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[204] M. Souden, J. Benesty, and S. Affes, “A study of the LCMV and MVDR noise reduction filters,” *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4925–4935, 2010.

[205] E. A. P. Habets, J. Benesty, S. Gannot, P. A. Naylor, and I. Cohen, “On the application of the lcmv beamformer to speech enhancement,” *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, pp. 141–144, 2009.

[206] W. Kellermann, “A self-steering digital microphone array,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 3581–3584 vol.5, 1991.

[207] A. Tanaka and H. Imai, “Music-based doa estimation by oblique projection along the signal subspace,” *2014 IEEE Workshop on Statistical Signal Processing (SSP)*, pp. 300–303, 2014.

[208] A. L. Swindlehurst and T. Kailath, “A performance analysis of subspace-based methods in the presence of model errors. i. the music algorithm,” *IEEE Transactions on signal processing*, vol. 40, no. 7, pp. 1758–1774, 1992.

[209] M. Hawkes, A. Nehorai, and P. Stoica, “Performance breakdown of subspace-based methods: Prediction and cure,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 6, pp. 4005–4008, 2001.

[210] W. Zuo, J. Xin, N. Zheng, and A. Sano, “Subspace-based localization of far-field and near-field signals without eigendecomposition,” *IEEE Transactions on Signal Processing*, vol. 66, no. 17, pp. 4461–4476, 2018.

[211] H. Moravec and A. Elfes, “High resolution maps from wide angle sonar,” *Proceedings. 1985 IEEE international conference on robotics and automation*, vol. 2, pp. 116–121, 1985.

[212] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, “Octomap: An efficient probabilistic 3D mapping framework based on octrees,” *Autonomous robots*, vol. 34, no. 3, pp. 189–206, 2013.

[213] R. Kerstens, D. Laurijssen, and J. Steckel, “eRTIS: A fully embedded real time 3D imaging sonar sensor for robotic applications,” *Proc. IEEE Int. Conf. Robotics, Automation.*, pp. 1438–1443, 2019.

[214] J. Steckel, A. Boen, and H. Peremans, “Broadband 3D sonar system using a sparse array for indoor navigation,” *IEEE Transactions on Robotics*, vol. 29, no. 1, pp. 161–171, 2012.

[215] J. Li and P. Stoica, “Efficient mixed-spectrum estimation with applications to target feature extraction,” *IEEE transactions on signal processing*, vol. 44, no. 2, pp. 281–295, 1996.

Part II

Papers

Paper A

Sound-based distance estimation for indoor navigation in the presence of ego-noise

Usama Saqib and Jesper Rindom Jensen

The paper has been published in the
27th European Signal Processing Conference (EUSIPCO)

Abstract

An off-the-shelf drone for indoor operation would come with a variety of different sensors that are used concurrently to avoid collision with, e.g., walls, but these sensors are typically unidirectional and offers limited spatial awareness. In this paper, we propose a model-based technique for distance estimation using sound and its reflections. More specifically, the technique is estimating Time-of-Arrivals (TOAs) of the reflected sound that could infer knowledge about room geometry and help in the design of sound-based collision avoidance. Our proposed solution is thus based on probing a known sound into an environment and then estimating the TOAs of reflected sounds recorded by a single microphone. The simulated results show that our approach to estimating TOAs for reflector position estimation works up to a distance of at least 2 meters even with significant additive noise, e.g., drone ego noise.

1 Introduction

One of the key issues when it comes to indoor operation of Unmanned Aerial Vehicles (UAVs), also known as drones, is the estimation of the physical boundaries' (e.g., walls) position in order to avoid collision. A common approach to estimating such positions is to use active sensors such as ultrasonic or infrared. Alternatively, camera-based technology combined with advanced computer vision techniques such as Simultaneous Localization And Mapping (SLAM) can be used for landmark or wall position estimation [A.1]. These techniques, however, have certain limitations. For instance, computer vision based techniques are susceptible to changing lightening conditions and does not work well under low-light conditions. Also, SLAM-based algorithms tends to have difficulty tracking a plain, white surface or landmarks making it harder for SLAM algorithm to estimate a wall position [A.2]. Moreover, such sensors have a limited field-of-view, so multiple sensors are required to cover all directions around the drone to avoid collisions with walls or other acoustic reflectors, e.g., glass windows. However, localization of a reflector position can be achieved using sound, by estimating the Time-of-Arrivals (TOAs) of acoustic reflections. This is a known estimation problem within the area of acoustic signal processing, which can potentially be implemented on moving robotic platforms or drones. TOAs estimation can thus be important in, e.g., robot and drone (UAV) applications, where it can facilitate acoustic SLAM (ASLAM) [A.3] and room geometry estimation (RGE) [A.4]. Moreover, if knowledge of TOAs is obtained, then distance estimation to acoustic reflectors is a straight-forward process given that the speed of sound is known.

In acoustic signal processing, the sound recorded by a microphone consists of a direct path component, first-order early reflections and later reflectins. This acoustic signal propagation from a loudspeaker to a microphone in a room is described by the room impulse response (RIR). The RIR contains information about the TOAs of acoustic reflections, which can be extracted. In the following, we review recent examples of methods utilizing this approach. For instance, in [A.5], a cell phone is used to probe the walls at different locations of the room with a chirp signal. The sound signals are reflected by the wall which are then correlated against the source signal to find TOAs which in turn helps determining the distances of the reflectors. The distance estimation was done by successfully extracting TOAs from a RIR. This knowledge helps the authors generate a map of the environment. Similarly, in [A.6], a single collocated microphone and loudspeaker arrangement was placed on a moving robotic platform to estimate distance between the robot and the reflecting surface from TOAs obtained from RIR. The authors in [A.6] proposes two estimators to calculate distance from TOAs; one involving multilateration techniques that uses the measured TOA values to construct a tangent line of the circle that indicates the position of the wall while the other approach is a Bayesian approach that gives a general solution to the RGE. Common for these state-of-the-art methods is that they require information about the TOA's of the early reflections. Typically, it is assumed that these estimates can be simply obtained through peak picking on an estimated RIR [A.7]- [A.8]. This approach is problematic in practice, however, because the individual peaks corresponding to the true TOA's can be small due to dispersion, diffusion, etc., and additive noise (e.g., drone ego noise) can introduce spurious peaks in the estimated RIR [A.9]. Moreover, the accuracy of the TOA estimates will be limited by the sampling rate [A.10], unless heuristic interpolation methods are used.

Since a moving drone is always accompanied by ego noise due to the motion of the rotors, we therefore propose and alternative approach to TOA estimation. This is a model-based approach for estimating TOA's based on a model for the early reflections. This enable us to derive a statistically optimal estimator for obtaining TOA estimates directly from observed microphone recordings instead of the traditional peak picking on an estimated RIR. This is inspired by the work in [A.11] on DOA estimation in reverberant environments. When it is desired to estimated multiple TOA's, e.g., to estimate the distance to multiple reflectors, our proposed estimator becomes computational complex due to its multidimensional nature. To tackle this, we propose an iterative estimation procedure based on the RELAX procedure [A.12].

The remaining part of this paper is organized as follows: Section 2 formulates the signal model and the problem. Section 3 describe the proposed TOA estimator based on the model, Section 4 describe an iterative procedure for handling multiple reflections, while Section 5 evaluates the performance and robustness of the proposed solution. Furthermore, Section 6 contains our conclusions and future work.

2 Signal Model and Problem Formulation

Consider the setup where a single loudspeaker is situated at $r_s \triangleq [x_s, y_s, z_s]$ that emits a known signal $s(n)$ which is recorded by a microphone placed at location $r_m \triangleq [x_m, y_m, z_m]$. The microphone and sound source are assumed to be collocated and placed inside a room. The observed signal recorded by microphone $y(n)$ is then modeled as follows:

$$y(n) = s(n) * h(n) + v(n) = x(n) + v(n) \tag{A.1}$$

where $h(n)$ is the impulse response of the room measured from r_s to r_m, $x(n) = s(n) * h(n)$ is the sound source signal including reverberation, $v(n)$ is additive background noise, e.g., ego noise, and $*$ represents the convolution operator. If we decompose (A.1) as a sum of its direct-path component and its first few reflections, then the observed signal model can be written as:

$$y(n) = \sum_{q=1}^{R} g_q s(n - \tau_q) + v'(n) \tag{A.2}$$

where g_q is the attenuation of the q^{th} order sound reflection from the source to the microphone, and $v'(n)$ is a combined noise term constituted by the late reverberation (i.e., the $q > R$ components) and the additive background noise. This can be further decomposed as

$$y(n) = x_D(n) + x_R(n) + v'(n), \tag{A.3}$$

where $x_D(n) = g_1 s(n - \tau_1)$ is the direct path component, and $x_R(n) = \sum_{q=2}^{R} g_q s(n - \tau_q)$ is the early reflection components. The signal decomposition, can also be expressed using simple first order FIR filters, h_q, for $q = 1, \ldots, R$, as

$$y(n) = \sum_{q=1}^{R} h_q * s(n) + v'(n), \tag{A.4}$$

The transfer function of these filters are given by

$$H_q(z) = g_q z^{-\tau_q}, \tag{A.5}$$

for $q = 1, \ldots, R$. In many applications, the microphone and the sound source will be placed in fixed positions. In such cases the transfer function of h_1 can be either measured offline or computed analytically using the geometry, i.e., by computing g_1 and τ_1. In such cases, we can thus work with a modified signal model:

$$\overline{y}(n) = \sum_{q=2}^{R} h_q * s(n) + v'(n), \tag{A.6}$$

where $\overline{y}(n) = y(n) - x_D(n)$, and only the gains and delays of the early reflections are unknown. The estimation problem at hand, is thus to estimate these unknown quantities, τ_q and g_q for $q = 2, \ldots, R$, which are key components in acoustic SLAM and room geometry estimation methods.

3 Non-Linear Least Square (NLS) Estimator

If we take N samples of the observed signals $\mathbf{y}(n) = \begin{bmatrix} y(n) & y(n+1) & \cdots & y(n+N-1) \end{bmatrix}^T$ and assume that we know $s(n)$ we can formulate a nonlinear least squares (NLS) estimator, which is the maximum likelihood estimator when the noise is white Gaussian. Mathematically, this can be formulated as

$$\{\widehat{\mathbf{g}}, \widehat{\boldsymbol{\tau}}\} = \underset{\mathbf{g},\boldsymbol{\tau}}{\arg\min} \, \|\overline{\mathbf{y}}(n) - \mathbf{x}(n)\|^2 \tag{A.7}$$

$$= \underset{\mathbf{g},\boldsymbol{\tau}}{\arg\min} \left\| \overline{\mathbf{y}}(n) - \sum_{q=2}^{R} h_q * \mathbf{s}(n) \right\|^2 , \tag{A.8}$$

where

$$\widehat{\boldsymbol{\tau}} = \begin{bmatrix} \widehat{\tau}_2 & \widehat{\tau}_3 & \cdots & \widehat{\tau} \end{bmatrix}^T , \tag{A.9}$$

$$\widehat{g} = \begin{bmatrix} \widehat{g}_2 & \widehat{g}_3 & \cdots & \widehat{g}_R \end{bmatrix}^T . \tag{A.10}$$

and $\overline{\mathbf{y}}(n)$, $\mathbf{x}_R(n)$ and $\mathbf{s}(n)$ are defined similarly to $\mathbf{y}(n)$. Moreover, the notation $a * \mathbf{b}$ denotes the convolution of each entry in the vector $\mathbf{b}$ with the scalar a, while $\widehat{c}$ denotes an estimate of the parameter c. Using Parseval's theorem, we can transfer (D.4) to the frequency domain, which yields

$$\{\widehat{\mathbf{g}}, \widehat{\boldsymbol{\tau}}\} = \underset{\mathbf{g},\boldsymbol{\tau}}{\arg\min} \, \|\overline{\mathbf{Y}} - \mathbf{X}\|^2 \tag{A.11}$$

$$= \underset{\mathbf{g},\boldsymbol{\tau}}{\arg\min} \left\| \overline{\mathbf{Y}} - \sum_{q=2}^{R} \mathbf{H}_q \odot \mathbf{S} \right\|^2 , \tag{A.12}$$

where $\overline{\mathbf{Y}}$ and $\mathbf{X}$ are the length K DFT vectors of $\overline{\mathbf{y}}(n)$ and $\mathbf{x}(n)$, respectively. Moreover, $\mathbf{H}_q = g_q \mathbf{Z}(\tau_q)$ and

$$\mathbf{Z}(\tau) = \begin{bmatrix} 1 & e^{-j\tau 2\pi \frac{1}{K}} & \cdots & e^{-j\tau 2\pi \frac{K-1}{K}} \end{bmatrix}^T . \tag{A.13}$$

That is, when the noise is white Gaussian, the maximum likelihood estimator can also be written as

$$\{\widehat{\mathbf{g}}, \widehat{\boldsymbol{\tau}}\} = \underset{\mathbf{g},\boldsymbol{\tau}}{\arg\min} \left\| \overline{\mathbf{Y}} - \sum_{q=2}^{R} g_q \mathbf{Z}(\tau_q) \odot \mathbf{S}) \right\|^2 \tag{A.14}$$

$$= \underset{\mathbf{g},\boldsymbol{\tau}}{\arg\min} \, J(\mathbf{g}, \boldsymbol{\tau}) \tag{A.15}$$

4 RELAX non-linear least square (RNLS) estimator

The estimator in $(A.14)$ can be shown to be statistically optimal when estimating $\mathbf{g}$ and $\boldsymbol{\tau}$ in the presence of additive white Gaussian noise. However, it is computationally expensive when estimating multiple TOA's as it will require a multi-dimensional search for different values of τ and $\mathbf{g}$, limiting its use in real-time, practical applications. Therefore, a RELAX procedure, originally proposed by [A.12] and later used in [A.11], will be adopted to iteratively calculate the value of $\boldsymbol{\tau}$ and $\mathbf{g}$.

In order to implement the RELAX method, we will introduce a modified observed signal:

$$\mathbf{Y}_r = \overline{\mathbf{Y}} - \sum_{q=2, q \neq r}^{R} g_q \mathbf{Z}(\tau_q) \odot \mathbf{S} \tag{A.16}$$

where $\mathbf{Y}_r$ is a modified observation vector containing only the r'th early reflection and additive noise.With this we can then estimate the r'th gain and TOA as

$$\{\widehat{g}_r, \widehat{\tau}_r\} = \arg\min_{g,\tau} \|\mathbf{Y}_r - g_r \mathbf{Z}(\tau_r) \odot \mathbf{S})\|^2 \tag{A.17}$$

We can then solve for the linear gain parameter g_r by taking the derivative of the cost function and setting it equal to zero, yielding

$$\widehat{g}_r = \frac{\mathbf{Y}_r^H \overline{\mathbf{Z}}(\tau_r) + \overline{\mathbf{Z}}^H(\tau_r)\mathbf{Y}_r}{2\overline{\mathbf{Z}}^H(\tau_r)\overline{\mathbf{Z}}(\tau_r)} \tag{A.18}$$

where $\overline{\mathbf{Z}}(\tau_r) = \mathbf{Z}(\tau_r) \odot \mathbf{S}$. This can be inserted back into estimator in (A.17) to obtain the τ_r as

$$\widehat{\tau}_r = \arg\min_{\tau} \left\| \mathbf{Y}_r - \frac{\mathbf{Y}_r^H \overline{\mathbf{Z}}(\tau) + \overline{\mathbf{Z}}^H(\tau)\mathbf{Y}_r}{2\overline{\mathbf{Z}}^H(\tau)\overline{\mathbf{Z}}(\tau)} \overline{\mathbf{Z}}(\tau) \right\|^2 \tag{A.19}$$

$$\widehat{\tau}_r = \arg\max_{\tau} \mathbb{R}\{\mathbf{Y}_r^H \overline{\mathbf{Z}}(\tau)\} \tag{A.20}$$

That is, by solving the optimization problem in (A.20), we can calculate $\widehat{\tau}_r$ and its corresponding $\widehat{g}_r$ of the r'th reflection. This leads to the iterative RELAX-based procedure:

- Step 1: Assume that $R = 2$, i.e., that we have one first-order reflection of the sound. Estimate g_2 and τ_2 using (A.18) and (A.19) from $\mathbf{Y}_2 = \overline{\mathbf{Y}}$.

- Step 2: Assume $R = 3$. Estimate g_3 and τ_3 using (A.18) and (A.19) from $\mathbf{Y}_3$ computed with the current estimates of τ_2 and g_2. Then re-estimate g_2 and τ_2 from $\mathbf{Y}_2$ computed using the newly estimated values of g_3 and τ_3. Continue Step 2 until it converges (e.g., $\|J^i - J^{i+1}\|^2 < \epsilon$ where i is the iteration index and ϵ is a threshold value.

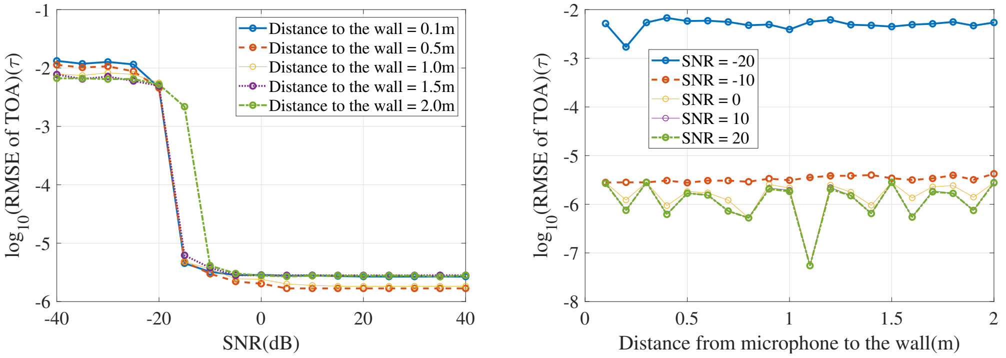
- Step 3: Assume $R = 4$. Estimate g_4 and τ_4 using (A.18) and (A.19) from $\mathbf{Y}_4$ computed with the current paramater estimates of the other reflections. Then re-estimate g_2 and τ_2 from $\mathbf{Y}_3$ computed using newly estimated reflection parameters. Then re-estimate g_3 and τ_3 from $\mathbf{Y}_3$ computed using the newly estimated reflection parameters. Continue until convergence.

- Remaining Steps: Continue until R is equal to the desired number of early reflections.

5 Experimental results and Evaluation

In this section, we will evaluate our proposed solution in a simulated room environment obtained with the Multichannel Room Acoustic Simulator (MCRoomSim) [A.13]. The performance was measured in terms of root mean squared error (RSME) with respect to the distance from the microphone and loudspeaker arrangement to the acoustic reflector, but also with respect to the noise level. Two experiments were conducted; one involving a random noise signal that is transmitted by the loudspeaker for different drone positions while the background noise is white Gaussian; and the other involved using more realistic drone ego noise (e.g., rotor noise) as the background noise. The drone sound was obtained from the DREGON dataset [A.14].

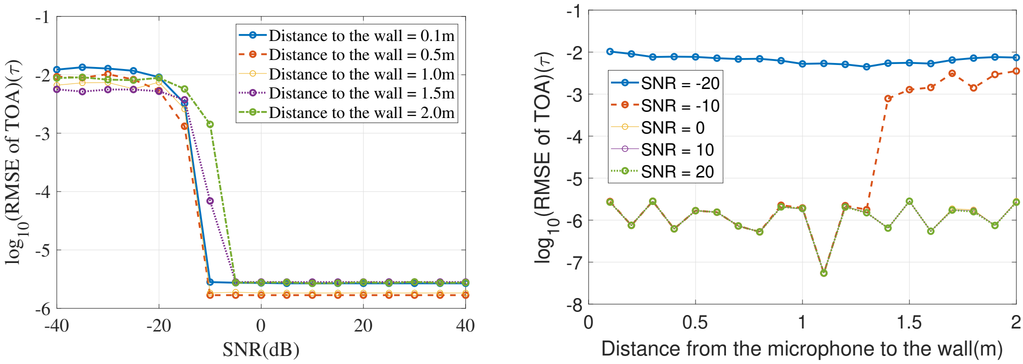
A room with a dimension of $10 \times 10 \times 6$ m was considered. To test the validity of our proposed solution, we use a collocated microphone-loudspeaker arrangement where the loudspeaker generate a known sound signal and a microphone is placed at a fixed distance of 0.1m directly underneath the loudspeaker. The microphone-loudspeaker arrangement was placed parallel to the x-axis of the room and was located at a position $r_s = [0.1, 5, 3]$ m while the microphone position is $r_m = [0.1, 5, 2.9]$ m. The position of the source and the microphone arrangement in relation to the wall is then varied from 0.1 m to 2 m in 0.2 m steps. Moreover, the sampling frequency was set to 44.1 kHz and the signal length was set to 2000 samples. As discussed in the previous section, we generate a known sound signal. For this particular experiment, we use a random noise signal as our sound source constituted by 2000 samples drawn from a Gaussian distribution. Furthermore, the speed of sound was fixed at 343 m/s. Then, additive white Gaussian noise was introduced at varying SNR levels ranging from -40 dB to 40 dB in 5dB steps. Similarly, the two evaluations (i.e., versus distance and SNR) was carried out with realistic drone ego noise as well. The ϵ value was set to 1×10^{-5}, which we found through experiments to be suitable for accurate estimation of the gains and TOAs with the RELAX procedure. Finally, 50 Monte Carlo simulation were conducted for each of the settings and the average results for each setting are shown.



(a) RMSE vs SNR

(b) RMSE vs distance

Fig. A.1: Performance metrics of proposed method using a Gaussian noise as the background noise. RMSE of TOA were measured against varying (a) SNR and (b) distance of collocated microphone-loudspeaker from one of the wall



(a) RMSE vs SNR

(b) RMSE vs distance

Fig. A.2: Performance metrics of propose method using a drone sound as the background noise for a large room. RMSE of TOA were measured against varying (a) SNR and (b) distance of collocated microphone-loudspeaker from one of the wall

5.1 Algorithm testing with additive white Gaussian noise as the sensor noise

In the first experiment, we tested the performance of our proposed method with white Gaussian background noise. As seen in Fig.A.1(a), the proposed method give low estimation errors for SNRs above -15 dB for distances between 0.1 m and 1.0 m, whereas for the higher distances, this is the case for SNRs above -10 dB. Moreover, as seen in A.1(b), the proposed method could estimate reflector's distance up to 2m when the background noise level is above -20 dB. Furthermore, the algorithm was tested on a standard desktop computer using MATLAB as the simulation environment running on Microsoft Windows 10 operating system with a an Intel Core i7 CPU with 3.40 GHz processing speed and 16 GB of Random Access Memory (RAM). The average time for the algorithm for estimating first-order early reflection is around 1.71 seconds which we believe would be suitable for any drone application. The average computation time could be further reduced when estimating the distances over time and reducing the grid size τ in (A.19). This is possible if we estimate distances at time instances zero and then at time instance one, the algorithm could use previous estimates of distance to search for TOAs using a reduced grid size.

5.2 Algorithm testing with drone noise as a background noise

In this experiment, we tested the performance of the proposed method in the presence of drone ego noise as the background noise. As seen in A.2(b), the performance is comparable to A.1(b). Moreover, it show the TOAs estimator starts to break down at -10 dB when increasing the distance above 1 m. These observations are expected, because the local SNR decreases as the distance of the proposed microphone- loudspeaker setup is increased against the wall. Moreover, similar behaviour will be expected across the remaining SNR values if we evaluate the estimator beyond $2m$.

5.3 Detecting multiple peaks using RELAX procedure

In a real-world situation, drones could be placed in near multiple acoustic reflectors, in which case we want to estimate multiple TOAs. This can be done with the RELAX procedure, we can estimate all the reflections associated with the reflecting surfaces. This was evaluated with a room of dimensions $6 \times 6 \times 2.4$ m that was simulated in MCRoomSim and the collocated loudspeaker-microphone pair was placed at a location of $r_s = [0.1, 1, 3]$m and $r_m = [0.1, 1, 2.9]$m, respectively. As seen in Fig. A.3, multiple reflections are recorded by the microphone, each associated with a wall inside a room. The estimated TOAs are close to strongest of the true TOAs of the walls.

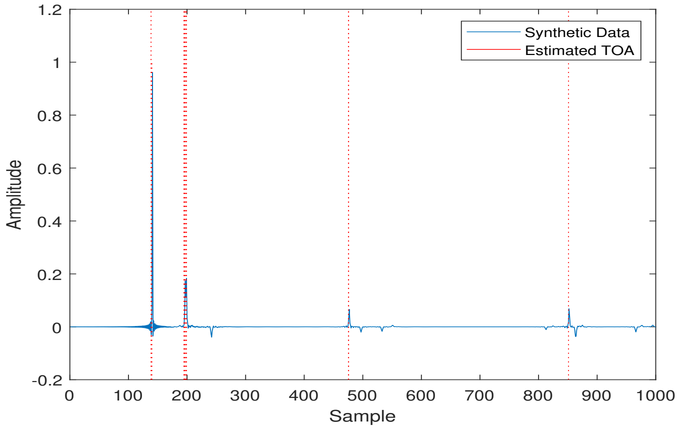


Fig. A.3: Detection of multiple reflections using the proposed iterative procedure.

6 Discussion and Future work

In this paper, we proposed an active approach to estimate TOAs using a collocated loudspeaker-microphone arrangement. Our iterative and model-based approach to TOA estimation could, e.g, be implemented on a UAV as part of a collision-avoidance system. The proposed method, is based on a model of early reflections leading to a statistically optimal NLS estimator. To handle the computationally complex problem of estimating multiple TOAs of multiple reflectors in this way, also proposed and iterative implementation of the estimator. In the experiments, we evaluated the method in different noisy scenarios, showing that our proposed method is robust and accurate up to at least a distance of 2 m with negative SNRs, both with additive white Gaussian noise and more realistic ego noise from the rotors of a drone. This indicate that the propose probing approach would not be too intrusive, as the TOAs can be estimated even when the ego noise is louder than the probing sound. In the future iteration of this research, we will test the performance of our proposed method on an actual UAV. Moreover, we aim at extending the proposed method to use an array of microphones so we can estimate both the distance and the direction of the early reflections.

References

[A.1] M. A. Al-Ammar, S. Alhadhrami, A. Al-Salman, A. Alarifi, H. S. Al-Khalifa, A. Alnafessah, and M. Alsaleh, "Comparative survey of indoor positioning technologies, techniques, and algorithms," *2014 International Conference on Cyberworlds*, pp. 245–252, 2014.

[A.2] E. Eade and T. Drummond, “Edge landmarks in monocular slam,” *Image and Vision Computing*, vol. 27, no. 5, pp. 588–596, 2009, the 17th British Machine Vision Conference (BMVC 2006). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885608000978

[A.3] I. Dokmanić, L. Daudet, and M. Vetterli, “From acoustic room reconstruction to slam,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 6345–6349, 2016.

[A.4] Y. E. Baba, A. Walther, and E. A. P. Habets, “3d room geometry inference based on room impulse response stacks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 857–872, 2018.

[A.5] T. Wang, F. Peng, and B. Chen, “First order echo based room shape recovery using a single mobile device,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 21–25, 2016.

[A.6] M. Kreković, I. Dokmanić, and M. Vetterli, “Echoslam: Simultaneous localization and mapping with acoustic echoes,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 11–15, 2016.

[A.7] S. Tervo, J. Pätynen, and T. Lokki, “Acoustic reflection localization from room impulse responses,” *Acta Acustica United With Acustica*, vol. 98, pp. 418–440, 2012.

[A.8] C. Falsi, D. Dardari, L. Mucchi, and M. Z. Win, “Time of arrival estimation for uwb localizers in realistic environments,” *Proc. European Signal Processing Conf.*, vol. 2006, pp. 1–13, 2006.

[A.9] I. J. Kelly and F. M. Boland, “Detecting arrivals in room impulse responses with dynamic time warping,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 7, pp. 1139–1147, 2014.

[A.10] J. R. Jensen, J. K. Nielsen, M. G. Christensen, and S. H. Jensen, “On frequency domain models for tdoa estimation,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 11–15, 2015.

[A.11] J. R. Jensen, J. K. Nielsen, R. Heusdens, and M. G. Christensen, “Doa estimation of audio sources in reverberant environments,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 176–180, 2016.

[A.12] J. Li and P. Stoica, “Efficient mixed-spectrum estimation with applications to target feature extraction,” *IEEE transactions on signal processing*, vol. 44, no. 2, pp. 281–295, 1996.

[A.13] A. Wabnitz, N. Epain, C. Jin, and A. V. Schaik, “Room acoustics simulation for multichannel microphone arrays,” *Proceedings of the International Symposium on Room Acoustics*, pp. 1–6, 2010.

[A.14] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, “DREGON: Dataset and methods for UAV-embedded sound source localization,” pp. 1–8, 2018.

Paper B

An EM method for multichannel TOA and DOA estimation of acoustic echoes

Jesper Rindom Jensen, Usama Saqib and Sharon Gannot

The paper has been published in the
2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)

The layout has been revised.

Abstract

The time-of-arrivals (TOAs) of acoustic echoes is a prerequisite in, e.g., room geometry estimation and localization of acoustic reflectors, which can be an enabling technology for autonomous robots and drones. However, solving these problems alone using TOAs introduces the difficult problem of echolabeling. Moreover, it is typically suggested to estimate the TOAs by estimating the room impulse response, and finding the peaks of it, but this approach is vulnerable against noise (e.g., ego noise). We therefore propose an expectation-maximization (EM) method for estimating both the TOAs and direction-of-arrivals (DOAs) of acoustic echoes using a loudspeaker and a uniform circular array (UCA). Our results show that this approach is more robust against noise compared to the traditional peak finding approach. Moreover, they show that the TOA and DOA information can be combined to estimate wall positions directly without considering echolabeling.

1 Introduction

Robot and drone audition are topics that have emerged during the past decade [B.1, B.2, B.3]. In addition to more established applications of audio, such as for human-robot interaction [B.4], audio has proven to be useful for estimation of acoustic source locations, robot/drone position, acoustic reflector positions and room geometries, which potentially can be an enabling technology for, e.g., indoor operation of robots and drones. The existing approaches for solving these estimation problems can be broadly classified as either passive or active approaches. In the passive approach, localization is conducted using external sources in the environment, such as human speech. This approach was considered previously for acoustic simultaneous localization and mapping (aSLAM) [B.5, B.6, B.7], which enables estimation of a robot's position in relation to a number of external acoustic sources. An advantage of the passive approach is that it is non-intrusive in the sense that it use already present sound in the environment, but it is unreliable if there are long periods of sound inactivity, and current approaches do not consider the estimation of acoustic reflectors.

The active approach, which is considered in this paper, uses one or more loudspeakers to probe the environment and one or more microphones to record the propagated probe sound. This enables the estimation of the TOAs of both the direct and reflected sounds. This can further increase the localization accuracy compared to the passive approach, and facilitate the locations of acoustic reflectors. One example of an active approach was proposed in [B.8], where the authors consider the problem of estimating both the room geometry and a robot's position within the room using a collocated microphone and loudspeaker setup. The approach utilizes TOA estimates of the first order reflections, which are assumed known. To resolve the ambiguity of how each TOA is mapped to a reflector position, they consider multiple observations over time and assume the robot is moving. Based on this, they propose two algorithms, one based on basic trigonometry and the other based on Bayesian filtering. Another approach was considered in [B.9], where the TOAs corresponding to first-order echoes are used for estimating the shape of arbitrary convex room shapes. Commonly for these and many other active approaches, is that they do not consider the TOA estimation problem although it is a difficult one due to, e.g., spurious estimates [B.10], and for methods relying on first- and second-order echoes only, it introduces the subsequent problem of echolabeling [B.11]. Moreover, if only one microphone and one loudspeaker is used, the mapping of first-order TOA estimates to reflector positions is ambiguous and requires either more transducers or the exploitation of movement.

To address some of these issues with the current active approaches, we consider a setup with a loudspeaker located inside a uniform circular microphone array. Based on this setup, we propose an expectation-maximization (EM) based method, which can estimate both the TOA and direction-of-arrival (DOA) of the sound echoes. In addition to yielding more accurate TOAs due to the use of multiple microphone recordings, the DOA estimation reduces the ambiguity of the estimated echoes, since the estimates corresponding to the first-order echoes directly reveal the reflector position. The estimation is carried out in the time-domain and directly from the recorded signals, and not from, e.g., estimated room impulse responses. Joint TOA and DOA estimation has been considered previously in multiuser and multipath communication systems [B.12, B.13, B.14], but to the best of our knowledge it has not been considered in active approaches for acoustic reflector localization.

2 Problem Formulation

Consider a setup where M microphones are recording the sound from a loudspeaker including its reflections from the physical objects and the boundaries of the acoustic enclosure. In its most general form, we can thus model the signal received by microphone m as

$$y_m(n) = h_m * s(n) + v_m(n) = x_m(n) + v_m(n) \tag{B.1}$$

where, $x_m(n) = h_m * s(n)$, h_m is the acoustic impulse response from the loudspeaker to microphone m, $s(n)$ is the audio signal played back by the loudspeaker, and $v_m(n)$ is an additive background noise (e.g., ego-noise from a robot or drone platform). In this paper, the audio signal $s(n)$ is assumed to be a known signal, which is then used to probe the acoustic environment and facilitate TOA and DOA estimation of the individual early reflections. To this end, we rewrite $y_m(n)$ as

$$y_m(n) = \sum_{r=1}^{R} g_{m,r} s(n - \tau_{1,r} - \eta_{m,r}) + w_m(n), \tag{B.2}$$

where R is the number of early reflections including the direct-path sound component, and $g_{m,r}$ is the attenuation of the rth sound component from the loudspeaker to microphone m. Moreover, $\eta_{m,r} = \tau_{m,r} - \tau_{1,r}$ is the TDOA of the rth component measured between microphone #1 and microphone $\#m$, $\tau_{m,r}$ is the TOA of the rth component on microphone m, and $w_m(n)$

is a noise term comprising both the additive noise component $v_m(n)$, and late reverberation, i.e., the late arrivals $r > R$. We note that microphone $\#1$ was arbitrarily chosen as the reference microphone, but the reference could be any of the microphones or even a virtual location like the array center.

If the geometry of the microphones and the loudspeaker are known, the model can be further specified. In this paper, we consider a setup where the loudspeaker is placed in the center of a uniform circular array (UCA) with M microphones. This enable us to write the TDOAs as

$$\eta_{m,r} = d \sin \psi_r [\cos(\theta_1 - \phi_r) - \cos(\theta_m - \phi_r)] \frac{f_s}{c}, \tag{B.3}$$

where d is the radius of the UCA, ψ_k and ϕ_k are the inclination and azimuth angles of the rth reflection, respectively, and θ_m is the angle of the mth microphone on the circle forming the UCA. Furthermore, f_s is the sampling frequency and c denotes the speed of sound. If we collect N time samples from each microphone and assume stationarity across those samples, we can vectorize our data and extend our signal model as:

$$\mathbf{y}_m(n) = \sum_{r=1}^{R} g_{m,r} \mathbf{s}(n - \tau_{1,r} - \eta_{m,r}) + \mathbf{w}_m(n) \tag{B.4}$$

with $\mathbf{y}_m(n)$, $\mathbf{s}(n)$, and $\mathbf{w}_m(n)$ being vectors comprising N time samples of $y_m(n)$, $s(n)$ and $w(n)$, respectively, e.g., $\mathbf{y}_m(n) = \begin{bmatrix} y_m(n) & y_m(n+1) & \cdots & y_m(n+N+1) \end{bmatrix}^T$.

The task at hand is then to estimate the unknown TOAs and DOAs of the R early reflections from N time samples from each of the M microphones.

3 Expectation-Maximization based TOA and DOA Estimation

We proceed to propose a method for solving the estimation problem in Sec. 2 as a maximum likelihood criterion solved by the application of the expectation-maximization (EM) algorithm. Before that, we briefly present an EM-based method for estimating the TOAs of the reflections when having only one loudspeaker and one microphone, which we refer to as a single-channel TOA estimation. This method was proposed in [B.15], and serves as our reference method.

3.1 Single-Channel TOA Estimation

In the following, we omit the microphone index since only a single microphone is considered. If we assume that the additive noise term is white Gaussian, the maximum likelihood estimator for the unknown TOAs is given by [B.16]

$$\{\hat{\boldsymbol{\tau}}, \hat{\mathbf{g}}\} = \min_{\boldsymbol{\tau}, \mathbf{g}} \left\| \mathbf{y}(n) - \sum_{r=1}^{R} g_r \mathbf{s}(n - \tau_r) \right\|^2, \tag{B.5}$$

where $\boldsymbol{\tau} = \begin{bmatrix}\tau_1 & \cdots & \tau_R\end{bmatrix}^T$, and $\mathbf{g} = \begin{bmatrix}g_1 & \cdots & g_R\end{bmatrix}^T$. While this estimator is statistically efficient under the white Gaussian noise assumption, it is non-convex with respect to the unknown TOAs and thus requires an exhaustive, computationally demanding and multidimensional search over numerous candidate TOAs.

Alternatively, an EM approach for superimposed signals [B.15] can be adopted. The basic idea behind this approach is to define the complete data as observations of all the individual signals. For the problem at hand, these observations are given by

$$\mathbf{x}_r(n) = g_r\mathbf{s}(n-\tau_r) + \mathbf{w}_r(n), \tag{B.6}$$

for $r = 1,\ldots,R$, where $\mathbf{w}_r(n)$ is obtained by an arbitrary decomposition of the total noise $\mathbf{w}(n)$ into the R components such that

$$\sum_{r=1}^{R}\mathbf{w}_r(n) = \mathbf{w}(n) \quad \wedge \quad \mathbf{y}(n) = \sum_{r=1}^{R}\mathbf{x}_r(n). \tag{B.7}$$

As suggested in [B.15], we let the individual noise terms be independent, zero-mean, white Gaussian and distributed as $\mathcal{N}(0, \beta_r\mathbf{C})$, and $\mathbf{C}$ is the covariance matrix of $\mathbf{w}(n)$. Moreover, the β_r's are arbitrary, non-negative and real-valued scalars satisfying $\sum_{r=1}^{R}\beta_r = 1$. With these assumptions, it can be shown that the EM algorithms assumes the following form

E-step: For $r = 1,\ldots,R$, compute

$$\widehat{\mathbf{x}}_r^{(i)}(n) = \widehat{g}_r^{(i)}\mathbf{s}\left(n-\widehat{\tau}_r^{(i)}\right) + \beta_r\left[\mathbf{y}(n) - \sum_{k=1}^{R}\widehat{g}_k^{(i)}\mathbf{s}\left(n-\widehat{\tau}_k^{(i)}\right)\right]. \tag{B.8}$$

M-step: For $r = 1,\ldots,R$, compute

$$\{\widehat{g}_r, \widehat{\tau}_r\}^{(i+1)} = \underset{g,\tau}{\operatorname{argmin}}\, \|\widehat{\mathbf{x}}_r^{(i)}(n) - g\mathbf{s}(n-\tau)\|^2, \tag{B.9}$$

where $^{(i)}$ denotes the iteration index. It can be shown that the M-step can be simplified if the analysis window is long compared to the length of the known signal used for the TOA estimation, in which case the estimator in (B.9) can be decomposed as

$$\widehat{\tau}_r = \underset{\tau}{\operatorname{argmax}}\, \widehat{\mathbf{x}}_r^T(n)\mathbf{s}(n-\tau), \tag{B.10}$$

$$\widehat{g}_r = \frac{\widehat{\mathbf{x}}_r^T(n)\mathbf{s}(n-\widehat{\tau}_r)}{\|\mathbf{s}(n)\|^2}. \tag{B.11}$$

This reveals an interesting interpretation of the EM-based estimator. First, the individual observations are processed by matched filters using the known source signal to find the unknown TOAs. Then, based on the estimated TOAs, closed-form estimates of the unknown gains are found by a least squares fit between the known input signal and the estimated contribution to the rth component.

3.2 Proposed TOA and DOA Estimation Method

Based on this single-channel approach, we now propose an EM algorithm for estimating both the TOAs and the DOAs of individual reflections, that are generated and observed using the setup described in Section 2. For this method, the complete data is considered as the observations of all the individual reflections from all microphones, where each of these observations, for $r = 1, \dots, R$ and $m = 1, \dots, M$ are given by

$$\mathbf{x}_{m,r} = g_{m,r}\mathbf{s}(n - \tau_{1,r} + \eta_{m,r}) + \mathbf{w}_{m,r}(n). \tag{B.12}$$

As for the single-channel case, the signals $\mathbf{w}_{m,r}(n)$ represents an arbitrary decomposition of the noise into R components. Here, the decomposition is applied for each microphone such that, for $m = 1, \dots, M$,

$$\sum_{r=1}^{R} \mathbf{w}_{m,r} = \mathbf{w}_m(n). \tag{B.13}$$

Moreover, we assume that the additive noise is uncorrelated between frames and sensors and that the variance of the noise is the same on each channel. Then, based on the EM algorithm for superimposed signals, the E- and M-steps can be stated:

E-step: For $r = 1, \dots, R$ and $m = 1, \dots, M$, compute

$$\begin{aligned}\widehat{\mathbf{x}}_{m,r}^{(i)}(n) &= \widehat{g}_{m,r}^{(i)}\mathbf{s}(n - \widehat{\tau}_{m,r}^{(i)}) \\ &\quad + \beta_r \left[\mathbf{y}_m(n) - \sum_{k=1}^{R} \widehat{g}_{m,k}^{(i)}\mathbf{s}(n - \widehat{\tau}_{m,k}^{(i)})\right],\end{aligned} \tag{B.14}$$

where $\widehat{\tau}_{m,r} = \widehat{\tau}_{1,r} + \widehat{\eta}_{m,r}$.

M-step: For $r = 1, \dots, R$, compute

$$\begin{aligned}\{\widehat{\tau}_{1,r}, \widehat{\boldsymbol{\eta}}_r, \widehat{\mathbf{g}}_r\}^{(i+1)} &= \underset{\tau_1,\boldsymbol{\eta},\mathbf{g}}{\operatorname{argmin}} \|\widehat{\mathbf{x}}_r^{(i)}(n) - \mathbf{D}(\boldsymbol{\eta}, \mathbf{g})\mathbf{s}(n - \tau_{1,r})\|^2 \\ &= \underset{\tau_1,\boldsymbol{\eta},\mathbf{g}}{\operatorname{argmin}} J(\tau_{1,r}, \boldsymbol{\eta}_r, \mathbf{g}_r)\end{aligned} \tag{B.15}$$

where

$$\begin{aligned}\widehat{\mathbf{x}}_r(n) &= \begin{bmatrix}\widehat{\mathbf{x}}_{1,r}^T & \cdots & \widehat{\mathbf{x}}_{M,r}^T\end{bmatrix}^T, \\ \mathbf{g}_r &= \begin{bmatrix}g_{1,r} & \cdots & g_{M,r}\end{bmatrix}^T, \ \boldsymbol{\eta}_r = \begin{bmatrix}\eta_{2,r} & \cdots & \eta_{M,r}\end{bmatrix}^T, \\ \mathbf{D}(\boldsymbol{\eta}_r, \mathbf{g}_r) &= \begin{bmatrix}g_{1,r}\mathbf{I}_N & g_{2,r}\mathbf{D}_{\eta_{2,r}}^T & \cdots & g_{M,r}\mathbf{D}_{\eta_{M,r}}^T\end{bmatrix}^T,\end{aligned}$$

and $\mathbf{D}_\eta$ is a cyclic shift matrix which delays a signal by $-\eta$ samples. It turns out that the cost function, $J(\tau_{1,r}, \boldsymbol{\eta}_r, \mathbf{g}_r)$, can be written as

$$J(\tau_{1,r}, \boldsymbol{\eta}_r, \mathbf{g}_r) = \|\widehat{\mathbf{x}}_r(n)\|^2 + \|\mathbf{g}_r\|^2\|\mathbf{s}(n - \tau_{1,k})\|^2 - 2\widehat{\mathbf{x}}_r^T(n)\mathbf{D}(\boldsymbol{\eta}_r, \mathbf{g}_r)\mathbf{s}(n - \tau_{1,r}), \tag{B.16}$$

The first term does not depend on any parameter of interest. If we assume that the analysis window is long compared to the length and the delay of the source signal $s(n)$, we also have that the second term does not depend on either $\tau_{1,k}$ or $\boldsymbol{\eta}_r$. That is, to estimate the TOA's and TDOA's, we only need to consider a simpler estimation problem:

$$\begin{aligned}\{\widehat{\tau}_{1,r}, \widehat{\boldsymbol{\eta}}_r\} &= \underset{\tau,\boldsymbol{\eta}}{\operatorname{argmax}}\, \mathbf{x}_r^T(n)\mathbf{D}(\boldsymbol{\eta}_r, \mathbf{g}_r)\mathbf{s}(n-\tau_{1,r}) \\ &= \underset{\tau,\boldsymbol{\eta}}{\operatorname{argmax}} \left(\sum_{m=1}^{M} g_{m,r}\mathbf{x}_{m,r}^T(n)\mathbf{D}_{\eta_{m,r}}\right)\mathbf{s}(n-\tau_{1,r}).\end{aligned} \tag{B.17}$$

The unknown gains can be replaced by their estimates obtained from minimizing the cost function with respect to these, yielding

$$\widehat{g}_{m,r} = \frac{\widehat{\mathbf{x}}^T\mathbf{D}_{\widehat{\eta}_{m,r}}\mathbf{s}(n-\widehat{\tau}_{1,r})}{\|\mathbf{s}(n)\|^2}. \tag{B.18}$$

If the reflections are assumed to be in the far-field of the array, we can simplify the estimators even further, since the gains will be independent of the microphone and will only depend on the reflection index r, namely $g_{m,r} = g_r$. If this is the case, the TOA and TDOA estimators become

$$\{\widehat{\tau}_{1,r}, \widehat{\boldsymbol{\eta}}_r\} \approx \underset{\tau,\boldsymbol{\eta}}{\operatorname{argmax}} \left(\sum_{m=1}^{M} \mathbf{x}_{m,r}^T(n)\mathbf{D}_{\eta_{m,r}}\right)\mathbf{s}(n-\tau_{1,r}), \tag{B.19}$$

and, accordingly, the gain estimator can be reformulated as

$$\widehat{g}_r = \frac{\sum_{m=1}^{M}\widehat{\mathbf{x}}_{m,r}^T\mathbf{D}_{\widehat{\eta}_{m,r}}\mathbf{s}(n-\widehat{\tau}_{1,r})}{M\|\mathbf{s}(n)\|^2}. \tag{B.20}$$

For the considered setup with a loudspeaker centered inside a UCA, we can use the model in (B.3) to further simplify the estimation problem, i.e., by searching over DOAs rather than TDOAs. That is, the TOA and TDOA estimator in (B.19) of the M-step is replaced by

$$\{\widehat{\tau}_{1,r}, \widehat{\phi}_r, \widehat{\psi}_r\} \approx \underset{\tau,\phi,\psi}{\operatorname{argmax}} \left(\sum_{m=1}^{M} \mathbf{x}_{m,r}^T(n)\mathbf{D}_{\eta_{m,r}}\right)\mathbf{s}(n-\tau_{1,r}), \tag{B.21}$$

where $\eta_{m,r}$ is computed using (B.3). This can reduce the dimensionality and thus the complexity of the estimation problem, since we then only need to estimate the DOAs of the individual reflections rather than all the TDOAs between the reference microphone and all the other microphones for each reflection. Moreover, the estimators in (B.19) and (B.21) have very interesting interpretations, namely that the EM-based estimators corresponds to estimating the TOAs and the DOAs by maximizing the output power of a matched filter at the output of a delay-and-sum beamformer. Further reductions in the computational complexity may be achieved by employing the space alternating generalized expectation (SAGE) algorithm rather than the EM algorithm [B.17], or by employing a recursive EM procedure [B.18] if the TOAs and DOAs need to be tracked over time.

4 Experimental Results

In our experimental study, we investigate two issues: the benefit of using multiple microphones for TOA estimation, and the application of the proposed TOA and DOA estimation method for acoustic reflector localization. In both experiments, the methods are tested using signals that are spatially synthesized using a room impulse response generator [B.19] with the following setup: the room dimensions were $8\times6\times5$ m, the reverberation time (T_{60}) was set to 0.6 s, and the sound speed was 343 m/s. Moreover, the loudspeaker was placed at the location $(1, 1.3, 2.5)$ m, and the UCA had $M = 3$ microphones centered around this position with at a radius of $d = 0.1$ m. A white Gaussian noise burst of $1,500$ samples was used as the known signal, $s(n)$, at a sampling frequency of $f_s = 22,050$ Hz. Since the UCA and loudspeaker configuration is fixed, we assumed that the direct path sound components can be estimated offline and subtracted these from the recorded signals before estimating the parameters of the reflections. The background noise was constituted by two parts: diffuse spherical noise and thermal sensor noise. The diffuse spherical noise was generated using the method described in [B.20] using noise from the rotors of a drone running at 70 RPS, which is available from the DREGON database [B.3]. The thermal sensor noise was spatially and spectrally white noise. These noises were then added to the microphone recordings to obtain certain signal-to-diffuse-noise and signal-to-noise ratios (SDNR and SNR). Both the SDNR and the SNR were the same across all microphones. The EM algorithm was setup to estimate $R = 3$ early reflections using 30 EM iterations, and the β_r's were all set to $1/R$. To initialize the method, the gain estimates, $\widehat{g}_r$ were sampled from a uniform distribution over the interval $[0; 1]$, the TOAs, $\widehat{\tau}_{1,r}$, were sampled from a uniform discrete distribution over the time indices corresponding to the analysis window, and the DOAs, $\widehat{\phi}_r$, were sampled from a uniform distribution over the interval $[0°; 360°]$. After emitting and recording the known source signal, an analysis window of each recording was considered starting from τ_{min} samples to τ_{max} samples after the source signal was emitted. For the first experiment, the interval was chosen such that the first-order reflections between distances of 0.5 m and 2 m from the array center were captured. With this setup, we then carried out an experiment where we evaluated the accuracy of the TOA estimates obtained with the proposed EM method (EM-UCA) for joint TOA and DOA estimation. Since the ψ's is ambiguous with the chosen array structure, we only estimated the ϕ's. The accuracy, was compared with that of the TOA estimates obtained with the single-channel EM method in Sec. 3.1 (EM-SC) using the observations from the reference microphone only, and with the commonly suggested approach of first estimating the RIR and then estimate the TOAs using peak picking on the estimated RIR (RIR-PP). The RIR was estimated by using dual channel analysis [B.21], i.e., by computing $\widehat{H}_1(f) = Y_1(f)/S(f)$ and then taking the inverse DFT to get $\widehat{h}_1 = \mathcal{F}^{-1}\{\widehat{H}_1(f)\}$. The accuracy was defined as the percentage of TOA estimates that were within ±1 sample of one of the TOAs of the first- and second-order reflections. This was measured for different SDNRs while the SNR was fixed to 40 dB, and for each SDNR it was measured over 100 Monte-Carlo simulations. Eventually, this led to the results depicted in Fig. B.1. These results show that the EM methods for TOA estimation clearly outperforms the RIR-PP approach. To achieve

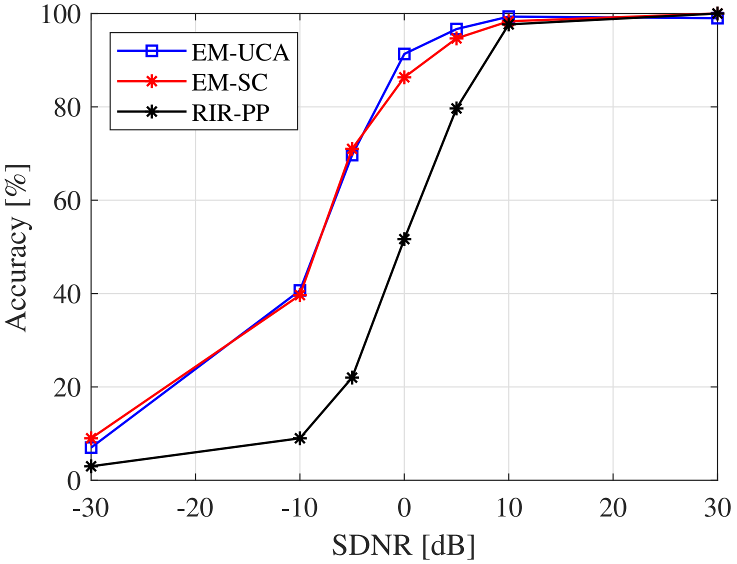


Fig. B.1: TOA estimation accuracy of the UCA and single-channel EM methods versus the SDNR.

similar accuracy with RIR-PP as with the EM methods, the SDNR needs to be almost 10 dB larger. Moreover, the results show that the proposed EM-UCA approach slightly outperforms the EM-SC approach in the SDNR region from 0 to 10 dB, and otherwise they show similar performance. It is important to note that EM-UCA achieves this while estimating one additional unknown parameter (i.e., the DOA) compared to EM-SC. Estimating the DOA is more difficult with the TOA-only based approaches, since it requires echolabeling to associate the TOAs estimated at different time instances, across multiple microphones, or both. Furthermore, it is expected that the performance of EM-UCA may be further improved by, e.g., not assuming far field.

In the second experiment, we consider an application example of the proposed method (EM UCA), where it was applied to acoustic reflector localization at different positions inside a room of dimensions $6 \times 4 \times 3$ m. This could be used on a robot or drone platform equipped with a UCA and loudspeaker setup to map the surroundings by estimating the distances and angles to physical objects including walls. For this experiment, the SNR was 40 dB and the SDNR was 5 dB. The microphone and loudspeaker setup was similar to the previous experiment, and was assumed to follow the path indicated in Fig. B.2 at a height of 1.5 m. We simulated this by estimating two reflector positions (i.e., $R = 2$) for 80 equispaced grid points on the depicted path. Aside from this, the simulation setup was identical to that in the previous experiment. The results from the experiment is depicted in Fig. B.2. As it can be seen, the proposed method is clearly able to provide accurate estimates of the acoustic reflector positions in most cases. The few erroneous outliers primarily happens when the UCA is only near one wall, e.g., at

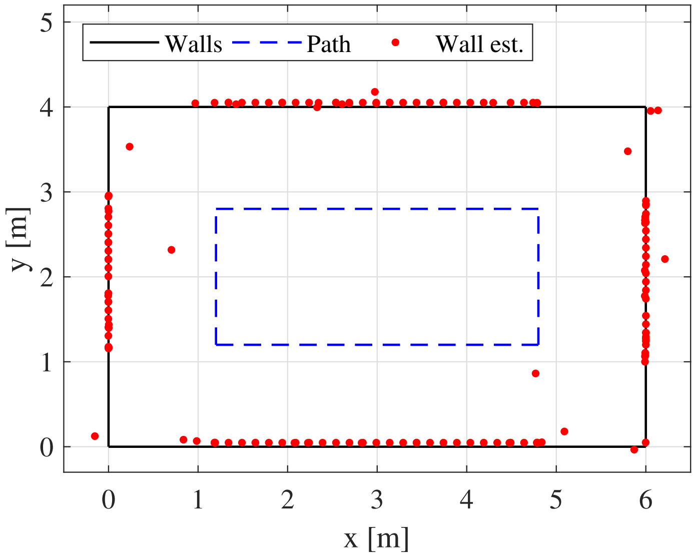


Fig. B.2: Examples of reflector positions estimates obtained using the TOA and DOA estimates from the proposed EM method.

$(3, 1.2, 1.5)$ m, in which case $R = 2$ is an inappropriate choice. While it is out of the scope of this paper, it is expected that the amount of errors can be reduced further by either choosing R adaptively or smoothing the estimates.

5 Conclusion

We considered estimation of the TOAs and DOAs of acoustic reflections with an active approach, assuming a hardware setup with a loudspeaker in the center of a UCA. Using this, we proposed an EM based method for estimating these parameters that can be used for example in estimating the position of walls and other physical objects, e.g., to enable autonomous indoor robots and drones. Existing methods for estimating acoustic reflector positions typically only use TOA information, and assume these can be obtained through peak finding on estimated RIRs. However, this approach is not robust to noise as opposed to the proposed approach as shown in our experiments. In the considered setup, the peak picking approach requires the SDNR to be 10 dB higher to yield the same results as the EM methods. In addition, our proposed method includes DOA estimation, so we can directly and accurately estimate the positions of first-order acoustic reflectors by combining the TOA and DOA estimates as shown in the experiments. This is difficult in existing TOA-only based approaches, where echolabeling is required to associate echoes over time, across microphones, or both.

References

[B.1] C. Rascon and I. Meza, “Localization of sound sources in robotics: A review,” *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889016304742

[B.2] H. W. Löllmann, A. Moore, P. A. Naylor, B. Rafaely, R. Horaud, A. Mazel, and W. Kellermann, “Microphone array signal processing for robot audition,” *Hands-free Speech Comm. and Microphone Arrays*, pp. 51–55, Mar 2017.

[B.3] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, “DREGON: Dataset and methods for UAV-embedded sound source localization,” *IEEE/RJS Int. Conf. Intelligent Robots and Systems*, pp. 5735–5742, Oct 2018.

[B.4] F. Badeig, Q. Pelorson, S. Arias, V. Drouard, I. D. Gebru, X. Li, G. Evangelidis, and R. Horaud, “A distributed architecture for interacting with NAO,” *Int. Conf. Multimodal Interaction*, Nov. 2015.

[B.5] J.-S. Hu, C.-Y. Chan, C.-K. Wang, M.-T. Lee, and C.-Y. Kuo, “Simultaneous localization of a mobile robot and multiple sound sources using a microphone array,” *Adv. Robotics*, vol. 25, no. 1–2, pp. 135–152, 2011. [Online]. Available: https://doi.org/10.1163/016918610X538525

[B.6] S. Ogiso, T. Kawagishi, K. Mizutani, N. Wakatsuki, and K. Zempo, “Self-localization method for mobile robot using acoustic beacons,” *ROBOMECH J.*, vol. 2, no. 1, p. 12, Sep 2015. [Online]. Available: https://doi.org/10.1186/s40648-015-0034-y

[B.7] C. Evers and P. A. Naylor, “Acoustic SLAM,” *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 26, pp. 1484–1498, 2018. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2018.2828321

[B.8] M. Kreković, I. Dokmanić, and M. Vetterli, “EchoSLAM: Simultaneous localization and mapping with acoustic echoes,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 11–15, Mar 2016.

[B.9] T. Wang, F. Peng, and B. Chen, “First order echo based room shape recovery using a single mobile device,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 21–25, Mar 2016.

[B.10] I. J. Kelly and F. M. Boland, “Detecting arrivals in room impulse responses with dynamic time warping,” *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 22, no. 7, pp. 1139–1147, Jul 2014.

[B.11] M. D. Plumbley, “Hearing the shape of a room,” *Proc. Natl. Acad. Sci. U S A*, vol. 110, no. 30, pp. 12 162–12 163, 2013.

[B.12] L. B. Nelson and H. V. Poor, “Iterative multiuser receivers for CDMA channels: an EM-based approach,” *IEEE Trans. Commun.*, vol. 44, no. 12, pp. 1700–1710, Dec 1996.

[B.13] M. C. Vanderveen, C. B. Papadias, and A. Paulraj, “Joint angle and delay estimation (JADE) for multipath signals arriving at an antenna array,” *IEEE Commun. Lett.*, vol. 1, no. 1, pp. 12–14, Jan 1997.

[B.14] J. Verhaevert, E. V. Lil, and A. V. de Capelle, “Direction of arrival (DOA) parameter estimation with the SAGE algorithm,” *Signal Processing*, vol. 84, no. 3, pp. 619–629, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168403003402

[B.15] M. Feder and E. Weinstein, “Parameter estimation of superimposed signals using the EM algorithm,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 4, pp. 477–489, Apr. 1988.

[B.16] J. E. Ehrenberg, T. E. Ewart, and R. D. Morris, “Signal processing techniques for resolving individual pulses in multipath signal,” *J. Acoust. Soc. Am.*, vol. 63, no. 6, pp. 1861–1865, 1978.

[B.17] J. A. Fessler and A. O. Hero, “Space-alternating generalized expectation-maximization algorithm,” *IEEE Trans. Signal Process.*, vol. 42, no. 10, pp. 2664–2677, Oct 1994.

[B.18] O. Schwartz and S. Gannot, “Speaker tracking using recursive EM algorithms,” *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 22, no. 2, pp. 392–402, Feb 2014.

[B.19] E. A. P. Habets, “Room impulse response generator,” Technische Universiteit Eindhoven, Tech. Rep., 2010, ver. 2.0.20100920. [Online]. Available: https://github.com/ehabets/RIR-Generator

[B.20] E. A. P. Habets, I. Cohen, and S. Gannot, “Generating nonstationary multisensor signals under a spatial coherence constraint,” *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, 2008. [Online]. Available: https://doi.org/10.1121/1.2987429

[B.21] H. Herlufsen, “Dual channel FFT analysis (part I),” *Brüel & Kjær Technical Review*, no. 1984-1, 1984.

Paper C

Estimation of acoustic echoes using expectation-maximization methods

Usama Saqib, Jesper Rindom Jensen and Sharon Gannot

The paper has been published in the
EURASIP Journal on Audio, Speech, and Music Processing (JASM)

The layout has been revised.

Abstract

Estimation problems like room geometry estimation and localization of acoustic reflectors are of great interest and importance in robot and drone audition. Several methods for tackling these problems exist, but most of them rely on information about times-of-arrival (TOAs) of the acoustic echoes. These need to be estimated in practice, which is a difficult problem in itself, especially in robot applications which are characterized by high ego-noise. Moreover, even if TOAs are successfully extracted, the difficult problem of echolabeling needs to be solved. In this paper, we propose multiple expectation-maximization (EM) methods, for jointly estimating the TOAs and directions-of-arrival (DOA) of the echoes, with a uniform circular array (UCA) and a loudspeaker in its center for probing the environment. The different methods are derived to be optimal under different noise conditions. The experimental results show that the proposed methods outperform existing methods in terms of estimation accuracy in noisy conditions. For example, it can provide accurate estimates at SNR of 10 dB lower compared to TOA extraction from room impulse responses, which is often used. Furthermore, the results confirm that the proposed methods can account for scenarios with colored noise or faulty microphones. Finally, we show the applicability of the proposed methods in mapping of an indoor environment.

1 Introduction

During the past decade, there has been an increased research interest in robot and drone audition [C.1, C.2, C.3]. Hearing capabilities enable robots to, understand and interact with humans [C.4]. Moreover, it has also been proven useful for sensing the physical environment. For example, it can be used for estimating the locations of acoustic sources, the position of a robot or drone, the positions of acoustic reflectors, and for inferring room geometry [C.5, C.6, C.7]. Potentially, this can enable autonomous indoor operation of robots and drones.

Some different approaches for tackling the above estimation problems have already been considered. In a broad sense, these can be classified as being either passive or active. The passive approach relies on using external sound sources in the environment to conduct the localization. Examples of such sources could be human speech, noise from machinery, or ego-noise from other robots or drones. This approach was, e.g., used for solving the acoustic simultaneous localization and mapping (aSLAM) problem [C.8, C.9, C.10]. With aSLAM, it is possible to estimate the robot location relative to a number of passive acoustic sources in its vicinity. One obvious advantage of such passive approaches, is that they are non-intrusive since only already existing sounds are used in the estimation. This comes at a price, however, since many acoustic sources, such as human speech, contains periods of inactivity, which can lead to unreliable estimates. This is particularly true with moving objects such as robots and drones. Moreover, to facilitate autonomous indoor operation, it is of great importance to also estimate the location of acoustic reflectors, e.g., walls, which is difficult with the passive approach, where only relative timing information is available.

The alternative, which we consider in this paper, is the active approach. In this approach, one or more loudspeakers are used to probe the environment using a known signal. Subsequently, a number of microphones are used to record the sound after it has propagated through the environment. Compared to the passive approach, this facilitate the estimation of the times-of-arrival (TOAs) of both the direct and reflected sound components. With this information, the localization accuracy can be increased significantly compared to the passive approach, and the task of acoustic reflector localization becomes less complex. In the following, we briefly outline some of the most recent and relevant work on active approaches. Some authors have considered the problem of estimating both room geometry and a robot's position with a setup consisting of a collocated microphone and speaker pair [C.11]. To achieve this, they utilize TOA estimates of the first order reflections. The TOAs are assumed known or estimated beforehand. To tackle the estimation problem with the considered single-channel setup (i.e., one microphone and one loudspeaker), they consider multiple observations from different time instances and locations, i.e., movement is assumed. Based on this, they then proposed two different methods: a method based on basic trigonometry, and another one based on Bayesian filtering. A similar approach also based on a priori RIR/TOA knowledge was considered using a multichannel setup in the context of robotics in [C.12]. Other authors considered an approach where the TOAs of the first order echoes are utilized for estimating the arbitrary convex room shapes [C.13]. As briefly mentioned, these as well as other active approaches, do not consider the TOA estimation problem, which is an equally important and difficult problem in itself due to, e.g., spurious estimates [C.14]. Moreover, methods relying on first- and second-order reflections only suffer from the inevitable problem of echolabeling [C.15]. In addition to this, many methods are based on only one microphone and one loudspeaker, but this lead to ambiguity in the mapping of the TOA estimates of the first-order reflections unless more transducers are included or movement is exploited.

These issues will be addressed in this paper, where we consider a setup consisting of a microphone array which is collocated with a single loudspeaker. More specifically, we consider a uniform circular array that could be placed on the perimeter of, e.g., a drone or robot platform, with a loudspeaker located in its center. With this setup in mind, we propose a number of expectation-maximization (EM) methods for estimating both the TOAs and directions-of-arrival (DOA) of a number of the acoustic reflections. This has the benefit of not only yielding more accurate TOAs compared to a single-channel approach, but also of reducing the ambiguity of the estimated reflections since the DOA is estimated simultaneously. In fact, this means that the estimates directly reveal the locations of mirror sources, which greatly simplifies the task of localizing the acoustic reflector positions. The proposed methods are derived in the time-domain, and, thus, estimates the parameters of interest directly from the recorded signals, i.e., not from estimated room impulse responses as in numerous state-of-the-art methods. While joint TOA and DOA estimation is a new topic in the context of robot and drone audition, it has been considered previously in multiuser and multipath communication systems [C.16, C.17, C.18]. However, it has not yet been considered for acoustic reflector localization to the best of our knowledge. The paper builds on the results reported in our earlier paper [C.19], and

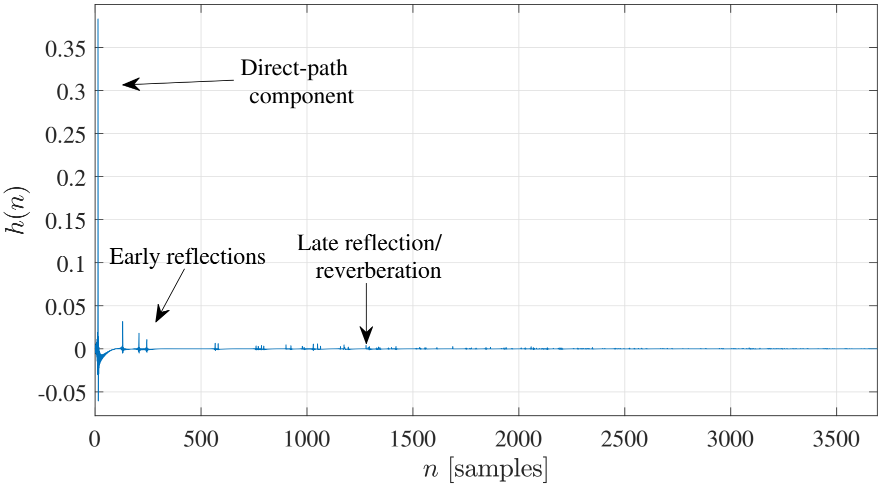


Fig. C.1: An example of a synthetic room impulse response illustrating its different parts, i.e., the direct-path component, early reflections and late reflections/ reveberation

extends on this work in several ways. First, we relax our previous noise assumptions and derive the optimal estimators for these more realistic scenarios. The first scenario deals with spatially independent white Gaussian noise with different noise variances across the microphones, e.g., to simulate low quality or faulty microphones. The second scenario considered deals with spatio-temporarily correlated noise, which we tackle using prewhitening. Here, we include different approaches for the prewhitening. Moreover, we have included a beamformer interpretation of one of the proposed multichannel estimators, which provides an intuitive understanding of the EM-based method. In addition to this, we included further experimental work to show case the merits of the different proposed estimators and how they compare with traditional methods.

The rest of the paper is organized as follows. In Section 2, we propose the signal model for the considered setup along with a problem formulation. Then, in Section 3, we briefly revisit the single-channel EM method for TOA estimation, which serves as our reference method. Inspired by this, we then proceed with the derivation of the different TOA and DOA estimators in Section 4. Finally, the paper closes with the experimental results and conclusions in Sections 5 and 6, respectively.

2 Problem Formulation

We now proceed to lay the foundation for the derivation of EM-based methods for estimating the TOA and TDOA of the acoustic echoes. This is done by formulating the relevant temporal and spatial signal models.

2.1 Time-domain model

Consider a setup with a single loudspeaker and M microphones that are assumed to be collocated on some hardware platform, e.g., a mobile robot or a drone. The loudspeaker is used to probe the environment with a known sound while the microphones are used to record the sound emitted by the loudspeaker including its acoustic reflections from physical objects and boundaries, e.g., walls. Both the microphones and loudspeakers are assumed to be omnidirectional and ideal. While this assumption might not hold in practice, we do not consider the handling of non-ideal characteristics in this paper. As suggested in other work [C.5], this might be partly addressed by estimating and introducing another filter accounting for the hardware characteristics, which may also be included in the methods proposed later. Moreover, the non-ideal characteristic of the hardware, i.e., loudspeakers could be modelled as shown in [C.5] but this is not included when formulating the following estimator.

We can then formulate a general model for the signal recorded by microphone m, for $m = 1, \ldots, M$, as

$$y_m(n) = h_m * s(n) + v_m(n) = x_m(n) + v_m(n), \tag{C.1}$$

where, $x_m(n) = h_m * s(n)$, h_m is the acoustic impulse response as measured from the loudspeaker to the mth microphone, $s(n)$ is a known signal being played back by the loudspeaker. Finally, $v_m(n)$ is an additive noise term, which is supposed to model ego-noise from a robot/drone platform, interfering sound sources (e.g., human speakers), thermal sensor noise, etc. That is, the signal $s(n)$ is used to probe the environment to, eventually, facilitate the estimation of the parameters of the acoustic echoes, such as their TOA and TDOA. Thus, we proceed by rewriting the observation model as a sum of the individual reflections[1] in noise, i.e.,

$$y_m(n) = \sum_{r=1}^{\infty} g_{m,r} s(n - \tau_{\text{ref},r} - \eta_{m,r}) + v_m(n), \tag{C.2}$$

with $g_{m,r}$ being the attenuation of the rth reflection from the loudspeaker to the mth microphone, e.g., due to the inverse square law for sound propagation and sound absorption in the acoustic reflectors. Furthermore, $\eta_{m,r} = \tau_{m,r} - \tau_{\text{ref,r}}$ is the TDOA of the rth component measured between a reference point and microphone m, while $\tau_{m,r}$ and $\tau_{\text{ref},r}$ are the TOAs of the rth component on microphone m and the reference point, respectively.

Acoustic impulse responses often exhibit a certain structure, which can be characterized by two parts: the early part, which is sparse in time and contains the direct-path and early reflections, and the late part, which is a more stochastic, dense, and characterized by decaying

[1] In our definition, the direct-path component is one of the reflections, i.e., the 0th order reflection corresponding to $r = 1$.

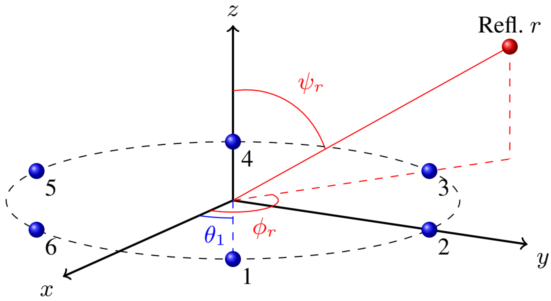


Fig. C.2: Example of a uniform circular array with six microphones.

tail of late reflections. This suggests that we can split the model as [C.20]

$$y_m(n) = \sum_{r=1}^{R} g_{m,r} s(n - \tau_{\text{ref},r} - \eta_{m,r}) + d_m(n) + v_m(n), \tag{C.3}$$

where R is the number of early reflections, and $d_m(n)$ is the late reverberation. A common assumption is that the late reverberation can be modeled as a spatially homogeneous and isotropic sound field with time-varying power but known coherence function [C.21]. If we collect N samples from each microphone and assume stationarity within the corresponding time frame, the vector model for our observations becomes:

$$\mathbf{y}_m(n) = \sum_{r=1}^{R} g_{m,r} \mathbf{s}(n - \tau_{\text{ref},r} - \eta_{m,r}) + \mathbf{d}_m(n) + \mathbf{v}_m(n), \tag{C.4}$$

with $\mathbf{y}_m(n)$, $\mathbf{s}(n)$, $\mathbf{d}(n)$, and $\mathbf{v}_m(n)$ being vectors comprising N time samples of $y_m(n)$, $s(n)$, $d_m(n)$, and $v_m(n)$, respectively, e.g.,

$$\mathbf{y}_m(n) = \begin{bmatrix} y_m(n) & \cdots & y_m(n+N-1) \end{bmatrix}^T,$$

This leaves us with the problem of estimating R unknown TOAs and MR TDOAs from the observations $\mathbf{y}_m(n)$, for $m = 1, \dots, M$. However, if we know the geometry of the loudspeaker and microphone array configuration, we can significantly reduces the dimensionality of this problem by further parametrizing the TDOAs in terms of the directions-of-arrival (DOAs).

2.2 Array model

While the array model can in principle be chosen arbitrarily, we choose to exemplify the TDOA modeling with a setup where the loudspeaker is placed in the center of a uniform circular array

(UCA). Such a setup could be placed on, e.g., a robot or drone platform to enable the estimation of the angle of and distance to acoustic reflectors, e.g., to facilitate autonomous and sound-based navigation.

If we assume the reference point to be the center of the UCA, it can be shown that the TDOA's, for a setup like this, can be modeled as

$$\eta_{m,r} = d \sin \psi_r \cos(\theta_m - \phi_r) \frac{f_s}{c} \tag{C.5}$$

where d is the radius of the UCA, ψ_r and ϕ_r are the inclination and azimuth angles of the rth reflection, respectively, and θ_m is the angle of the mth microphone on the circle forming the UCA. These definitions are illustrated in the UCA example in Figure C.2. In addition to this, f_s is the sampling frequency, and c is the speed of sound.

The TDOA model in (C.5) can then be combined with the observation model in (C.4). By doing this, the estimation problem at hand is then simplified to the estimation of $2R$ angles, i.e., ψ_r and ϕ_r, for $r = 1, \dots, R$, rather than MR TDOA's. It should be noted here that the considered UCA configuration introduces ambiguities, e.g., an acoustic reflection impinging from an elevation of $0°$ will result in the same TDOAs as an acoustic reflection mirrored around the UCA plane, i.e., at an elevation angle of $180°$. However, this ambiguity can easily be accounted for by applying the proposed methods on array structures with microphones in all three dimensions, e.g., spherical microphone arrays [C.22].

3 Single-Channel Estimation

Before presenting the proposed TOA and TDOA estimators, we briefly revisit an EM-based method for single-channel TOA estimation, i.e., that is with a setup consisting of one loudspeaker and one microphone. The original version of this method was proposed in [C.23] under a white Gaussian noise assumption and serves as a reference for the proposed methods.

3.1 White Gaussian noise

In the following, we leave out the microphone index, i.e., subscript m, since only a single microphone is considered. If we assume that the additive noise, i.e., both the late reverberation and the background noise is independent and identically distributed white Gaussian and zero-mean. Later, as part of the proposed multichannel methods, this assumption is substituted with a more realistic one, where the late reverberation is modelled as being spatio-temporarily correlated. The signal model in (C.4) then reduces to

$$\mathbf{y}(n) = \sum_{r=1}^{R} g_r \mathbf{s}(n - \tau_r) + \mathbf{v}(n), \tag{C.6}$$

where $\mathbf{v}(n)$ is distributed as $\mathcal{N}(\mathbf{0}, \mathbf{C})$, with $\mathbf{0}$ being a vector of zeroes, $\mathbf{C} = \mathrm{E}[\mathbf{v}(n)\mathbf{v}^T(n)] = \sigma_v^2 \mathbf{I}_N$ is the $N \times N$ covariance matrix of $\mathbf{v}(n)$, σ_v^2 is its variance, $\mathbf{I}_N$ denotes the $N \times N$ identity matrix, and $\mathrm{E}[\cdot]$ is the mathematical expectation operator. The maximum likelihood (ML) estimator of the unknown parameters, i.e., the gains and the TOAs, is well known to be the nonlinear least squares (NLS) criterion in this case, i.e.,

$$\{\widehat{\boldsymbol{\tau}}, \widehat{\mathbf{g}}\} = \underset{\boldsymbol{\tau}, \mathbf{g}}{\operatorname{argmin}} \left\| \mathbf{y}(n) - \sum_{r=1}^{R} g_r \mathbf{s}(n - \tau_r) \right\|^2 , \tag{C.7}$$

where

$$\boldsymbol{\tau} = \begin{bmatrix} \tau_1 & \cdots & \tau_R \end{bmatrix}^T ,$$
$$\mathbf{g} = \begin{bmatrix} g_1 & \cdots & g_R \end{bmatrix}^T .$$

While this estimator is statistically efficient, it also requires computationally costly search since the cost function is high-dimensional and non-convex with respect to the TOAs.

A computationally more efficient way of implementing this estimator could be to adopt the expectation-maximization (EM) approach for superimposed signals proposed in [C.23]. The concept behind this approach is to define the complete data as the observation of all individual signals, i.e., each of the individual early reflections in our case. According to the previously stated signal model in (C.4), the individual observations can be modeled as

$$\mathbf{x}_r(n) = g_r \mathbf{s}(n - \tau_r) + \mathbf{v}_r(n), \tag{C.8}$$

for $r = 1, \ldots, R$, where $\mathbf{v}_r(n)$ is obtained by arbitrarily decomposing the combined noise term, $\mathbf{v}(n)$, into R different components adhering to

$$\sum_{r=1}^{R} \mathbf{v}_r(n) = \mathbf{v}(n). \tag{C.9}$$

Moreover, the observed signal can be written as the sum of individual observations such as:

$$\mathbf{y}(n) = \sum_{r=1}^{R} \mathbf{x}_r(n). \tag{C.10}$$

Following [C.23], we let the individual noise terms be independent, zero-mean, white Gaussian, and distributed as $\mathcal{N}(\mathbf{0}, \beta_r \mathbf{C})$. Furthermore, the scaling factors, β_r are non-negative, real-valued scalars that satisfy

$$\sum_{r=1}^{R} \beta_r = 1. \tag{C.11}$$

Under these assumptions, it can be shown that the EM algorithm for estimating the gains and the time-of-arrivals is given by [C.23]

E-step: for $r = 1, \dots, R$, compute

$$\widehat{\mathbf{x}}_r^{(i)}(n) = \widehat{g}_r^{(i)} \mathbf{s}\left(n - \widehat{\tau}_r^{(i)}\right) + \beta_r \left[\mathbf{y}(n) - \sum_{k=1}^{R} \widehat{g}_k^{(i)} \mathbf{s}\left(n - \widehat{\tau}_k^{(i)}\right)\right]. \tag{C.12}$$

M-step:

$$\{\widehat{g}_r, \widehat{\tau}_r\}^{(i+1)} = \underset{g,\tau}{\operatorname{argmin}} \, \|\widehat{\mathbf{x}}_r^{(i)}(n) - g\mathbf{s}(n-\tau)\|^2, \tag{C.13}$$

where $^{(i)}$ is denoting the iteration index. If the length, N, of the analysis window is long compared to the length of the known signal, $s(n)$, the M-step can be simplified as

$$\widehat{\tau}_r = \underset{\tau}{\operatorname{argmax}} \, \widehat{\mathbf{x}}_r^T(n)\mathbf{s}(n-\tau), \tag{C.14}$$

$$\widehat{g}_r = \frac{\widehat{\mathbf{x}}_r^T(n)\mathbf{s}(n-\widehat{\tau}_r)}{\|\mathbf{s}(n)\|^2}. \tag{C.15}$$

We see that the estimation problem has been greatly simplified with this signals decomposition, since we now have $2R$ one-dimensional estimators rather than a $2R$-dimensional estimator as in (C.7). From this simplified version of the M-step, we can make some interesting interpretations. First in (C.14), the individual observations are applied with a matched filter based on the known source signal. The TOA is estimated as the one maximizing the output power of the matched filter. Secondly, the estimated TOA's are used to obtain closed-form estimated of the gains in (C.15), which is based on a least squares fit between the known source signal and the estimated contribution of the r'th component.

4 Multichannel Estimation

We now proceed to consider the multichannel case, where we have one loudspeaker and multiple microphones. First, we consider a white Gaussian noise scenario similar to Section 3.1 where the noise is independent across the microphones, after which we turn to the more realistic scenarios with correlated noise.

4.1 Spatially independent white Gaussian noise

If we first assume that the noise is temporally white Gaussian and independent and the late reverberation is negligible, the signal model in (C.4) reduces to

$$\mathbf{y}_m(n) = \sum_{r=1}^{R} g_{m,r}\mathbf{s}(n - \tau_{\text{ref},r} - \eta_{m,r}) + \mathbf{v}_m(n), \tag{C.16}$$

for $m = 1, \ldots, M$. Subsequently, we can aggregate the observations from all microphones in one model as

$$\begin{aligned} \mathbf{y}(n) &= \sum_{r=1}^{R} \mathbf{H}(\boldsymbol{\eta}_r, \mathbf{g}_r)\mathbf{s}(n - \tau_{\text{ref},r}) + \mathbf{v}(n) \\ &= \begin{bmatrix} \mathbf{y}_1^T & \mathbf{y}_2^T & \cdots & \mathbf{y}_M^T \end{bmatrix}^T, \end{aligned} \tag{C.17}$$

where $\mathbf{v}(n)$ is the stacked noise terms from each microphone defined similarly to $\mathbf{y}(n)$, and

$$\begin{aligned} \boldsymbol{\eta}_r &= \begin{bmatrix} \eta_{1,r} & \eta_{2,r} & \cdots & \eta_{M,r} \end{bmatrix}^T, \\ \mathbf{g}_r &= \begin{bmatrix} g_{1,r} & g_{2,r} & \cdots & g_{M,r} \end{bmatrix}^T. \end{aligned}$$

In addition to this, we note that, under the assumptions of spatial independent white Gaussian noise, the covariance matrix, $\mathbf{C}$ of the stacked noise, $\mathbf{v}(n)$ is diagonal and given by

$$\mathbf{C} = \text{diag}\left(\sigma_{v_1}^2 \mathbf{I}_N, \sigma_{v_2}^2 \mathbf{I}_N, \ldots, \sigma_{v_M}^2 \mathbf{I}_N\right), \tag{C.18}$$

where diag$(\cdot)$ is the operator constructing a diagonal matrix from the input of scalars(/matrices) and $\mathbf{C}$ is the $MN \times MN$ covariance matrix. Furthermore,

$$\mathbf{H}(\boldsymbol{\eta}_r, \mathbf{g}_r) = \begin{bmatrix} g_{1,r}\mathbf{D}_{\eta_{1,r}}^T & \cdots & g_{M,r}\mathbf{D}_{\eta_{M,r}}^T \end{bmatrix}^T, \tag{C.19}$$

and $\mathbf{D}_\eta$ is a circular shift matrix which delays a signal by $-\eta$ samples.

With these definitions, the ML estimator for the problem at hand becomes

$$\{\widehat{\mathbf{g}}, \widehat{\boldsymbol{\tau}}, \widehat{\boldsymbol{\eta}}\} = \underset{\mathbf{g}, \boldsymbol{\tau}, \boldsymbol{\eta}}{\text{argmin}}\, J(\mathbf{g}, \boldsymbol{\tau}, \boldsymbol{\eta}), \tag{C.20}$$

where

$$J(\mathbf{g}, \boldsymbol{\tau}, \boldsymbol{\eta}) = \left\| \mathbf{y}(n) - \sum_{r=1}^{R} \mathbf{H}(\boldsymbol{\eta}_r, \mathbf{g}_r)\mathbf{s}(n - \tau_{\text{ref},r}) \right\|_{\mathbf{C}^{-1}}^2 \tag{C.21}$$

such that $\|\mathbf{x}\|_{\mathbf{W}}^2 = \mathbf{x}^T\mathbf{W}\mathbf{x}$, where $\mathbf{W}$ denotes the weighted 2-norm of $\mathbf{x}$. Moreover, $\mathbf{g}, \boldsymbol{\tau}$ and $\boldsymbol{\eta}$ are the parameter vectors containing all unknown gains, TOAs and TDOAs, respectively. In the single-channel case, the ML estimator ends up being high-dimensional and non-convex, resulting in a practically infeasible computational complexity if implemented directly. Therefore, we propose to adopt the EM framework also for the multichannel scenario.

Like in the single-channel approach, we consider the complete data to be all the individual observations of the reflections, but in this case from all the M microphones. Each of the observations can thus, for $r = 1, \ldots, R$, be modeled as

$$\mathbf{x}_r = \mathbf{H}(\boldsymbol{\eta}_r, \mathbf{g}_r)\mathbf{s}(n - \tau_{\text{ref},r}) + \mathbf{v}_r(n). \tag{C.22}$$

The decomposition is assumed to satisfy the conditions in (C.9)–(C.11). Then, it can be shown that the EM-algorithm for the multichannel estimation problem is given by

E-step: for $r = 1, \dots, R$, compute

$$\begin{aligned}\widehat{\mathbf{x}}_r^{(i)}(n) &= \mathbf{H}\left(\widehat{\boldsymbol{\eta}}_r^{(i)}, \widehat{\mathbf{g}}_r^{(i)}\right)\mathbf{s}\left(n - \widehat{\tau}_{\text{ref},r}^{(i)}\right) \\ &+ \beta_r\left[\mathbf{y}(n) - \sum_{k=1}^{R}\mathbf{H}\left(\widehat{\boldsymbol{\eta}}_k^{(i)}, \widehat{\mathbf{g}}_k^{(i)}\right)\mathbf{s}\left(n - \widehat{\tau}_{\text{ref},k}^{(i)}\right)\right].\end{aligned} \tag{C.23}$$

M-step: for $r = 1, \dots, R$,

$$\{\widehat{\mathbf{g}}_r, \widehat{\tau}_r, \widehat{\boldsymbol{\eta}}_r\}^{(i+1)} = \underset{\mathbf{g},\tau,\boldsymbol{\eta}}{\operatorname{argmin}}\, J_r(\mathbf{g}, \tau, \boldsymbol{\eta}), \tag{C.24}$$

with $J_r(\mathbf{g}, \tau, \boldsymbol{\eta})$ being a weighted least squares estimator defined as

$$J_r(\mathbf{g}, \tau, \boldsymbol{\eta}) = \left\|\widehat{\mathbf{x}}_r^{(i)}(n) - \mathbf{H}(\boldsymbol{\eta}, \mathbf{g})\mathbf{s}(n - \tau)\right\|_{\mathbf{C}^{-1}}^2. \tag{C.25}$$

If we explicitly write the cost function, we get

$$\begin{aligned}J_r(\mathbf{g}, \tau, \boldsymbol{\eta}) = \sum_{m=1}^{M}\frac{\|\widehat{\mathbf{x}}_{m,r}(n)\|^2}{\sigma_{v_m}^2} \\ + \|\mathbf{s}(n-\tau)\|^2\sum_{m=1}^{M}\frac{g_{m,r}^2}{\sigma_{v_m}^2} \\ - 2\sum_{m=1}^{M}\frac{g_{m,r}\widehat{\mathbf{x}}_{m,r}^T(n)\mathbf{D}_{\eta_m}}{\sigma_{v_m}^2}\mathbf{s}(n-\tau),\end{aligned} \tag{C.26}$$

This can be used to simplify the M-step by making a few observations. Clearly, the first term in this expression does not depend on any parameter of interest. Moreover, if we assume that the analysis window is long compared to the length of the known source signal, $s(n)$, we observe that the second term does not depend on either the TOAs or the TDOAs. That is, to estimate these time parameters, we only need to consider the maximization of the last term, i.e.,

$$\{\widehat{\tau}_{\text{ref},r}, \widehat{\boldsymbol{\eta}}_r\} = \underset{\tau,\boldsymbol{\eta}}{\operatorname{argmax}}\sum_{m=1}^{M}\frac{g_{m,r}\widehat{\mathbf{x}}_{m,r}^T(n)\mathbf{D}_{\eta_m}}{\sigma_{v_m}^2}\mathbf{s}(n-\tau), \tag{C.27}$$

The gains, $g_{m,r}$, and the noise statistics, $\sigma_{v_m}^2$, are unknown in practice. However, if the noise is assumed (quasi-)stationary, its variance can be estimated from microphone recordings acquired before emitting the known source signal, $s(n)$. By taking the partial derivative of (C.26) with respect to $g_{m,r}$, we obtain the following closed-form estimate for $g_{m,r}$

$$\widehat{g}_{m,r} = \frac{\widehat{\mathbf{x}}_{m,r}^T(n)\mathbf{D}_{\widehat{\eta}_m}\mathbf{s}(n-\widehat{\tau}_{\text{ref},r})}{\|\mathbf{s}(n)\|^2}, \tag{C.28}$$

If the reflections are assumed to be in the far-field of the array, we can further simplify the estimators. In this case, the gains of reflection r will be the same across all microphones for $r = 1, \dots, R$. That is, we can instead estimate the TOAs and TDOAs as

$$\{\widehat{\tau}_{\text{ref},r}, \widehat{\boldsymbol{\eta}}_r\} \approx \underset{\tau,\boldsymbol{\eta}}{\operatorname{argmax}} \left(\sum_{m=1}^{M} \frac{\widehat{\mathbf{x}}_{m,r}^T(n)\mathbf{D}_{\eta_m}}{\sigma_{v_m}^2}\right) \times \mathbf{s}(n-\tau). \tag{C.29}$$

Subsequently, the gain estimator can then be reformulated as

$$\widehat{g}_r = \left(\sum_{m=1}^{M} \frac{1}{\sigma_{v_m}^2}\right)^{-1} \sum_{m=1}^{M} \frac{\widehat{\mathbf{x}}_{m,r}^T\mathbf{D}_{\widehat{\eta}_m}}{\sigma_{v_m}^2} \frac{\mathbf{s}(n-\widehat{\tau}_{\text{ref},r})}{\|\mathbf{s}(n)\|^2}, \tag{C.30}$$

If the geometry of the loudspeaker and microphone configuration is known, we further reduce the dimensionality of the estimation problem. This is achieved by parameterizing the TDOA's, $\eta_{m,r}$, for $r = 1, \dots, R$ and $m = 1, \dots, M$ using the array model, e.g., the one for a UCA configuration formulated in (C.5). Then, the TOA and TDOA estimator in the M-step can be written as

$$\{\widehat{\tau}_{\text{ref},r}, \widehat{\phi}_r, \widehat{\psi}_r\} \approx \underset{\tau,\phi,\psi}{\operatorname{argmax}} \left(\sum_{m=1}^{M} \frac{\widehat{\mathbf{x}}_{m,r}^T(n)\mathbf{D}_{\eta_m}}{\sigma_{v_m}^2}\right) \times \mathbf{s}(n-\tau), \tag{C.31}$$

where η_m is replaced by the expression in (C.5). In this way, we only need to estimate two angles for each reflection, whereas the estimator in, e.g., (4.1) requires the estimation of M TDOAs (or $M-1$ if one of the microphone positions is used as the reference point). That is, the computational benefits of using the array model increases as we increase the number of microphones. It can be shown that the resulting estimators in the M-step has an interesting interpretation as minimum variance distortionless response (MVDR) beamforming followed by a matched filter as we show in the following subsection.

4.2 Beamformer interpretation

Intuitively, if we were able to observe the reflections individually in noise and the noise is differently distributed across the microphones. Then, it would be natural to apply an MVDR beamformer to these to optimally account for the noise when estimating the TOA's and TDOA's. Let us consider the scenario where we have a filtering matrix, $\mathbf{W}$, which we use to process the individually observed reflections in (C.22):

$$\mathbf{z}(n) = \mathbf{W}^T \mathbf{x}_r(n). \tag{C.32}$$

Then, we define the residual noise power after this filtering as the normalized sum of the residual noise variances over the different time indices included in $\mathbf{z}(n)$, i.e., $n, n+1, \ldots, n+N-1$. Mathematically, this is equivalent to

$$\begin{aligned}\sigma_{v,f}^2 &= \mathrm{E}\left[\frac{1}{N}\mathrm{Tr}\left\{\mathbf{W}^T \mathbf{v}_r(n)\mathbf{v}_r^T(n)\mathbf{W}\right\}\right] \\ &= \frac{\beta_r}{N}\mathrm{Tr}\left\{\mathbf{W}^T\mathbf{C}\mathbf{W}\right\},\end{aligned} \tag{C.33}$$

where Tr$\{\cdot\}$ is the trace operator. Obviously, by inspection of the individual observation model in (C.22), we can see that the following expression needs to be satisfied for the filter to be distortionless with respect to the known source signal:

$$\mathbf{W}^T \mathbf{H}(\boldsymbol{\eta}_r, \mathbf{g}_r) = \mathbf{I}_N. \tag{C.34}$$

That is, omitting the arguments of the steering matrix $\mathbf{H}(\boldsymbol{\eta}_r, \mathbf{g}_r)$ for brevity, the problem of finding the MVDR solution for $\mathbf{W}$ can be formulated as

$$\min_{\mathbf{W}} \mathrm{Tr}\left\{\mathbf{W}^T\mathbf{C}\mathbf{W}\right\} \quad \text{s.t.} \quad \mathbf{W}^T\mathbf{H} = \mathbf{I}_N. \tag{C.35}$$

It can be shown that the solution to the quadratic optimization problem with linear constraints is given by

$$\mathbf{W}_{\mathrm{M}} = \mathbf{C}^{-1}\mathbf{H}\left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}. \tag{C.36}$$

If we then apply the MVDR filtering matrix to the estimated observation of the rth reflection in noise, careful inspection reveals that

$$\mathbf{x}_r^T(n)\mathbf{W}_{\mathrm{M}} = \frac{\displaystyle\sum_{m=1}^{M} \frac{g_m \mathbf{x}_{m,r}^T(n)\mathbf{D}_{\eta_m}}{\sigma_{v_m}^2}}{\displaystyle\sum_{m=1}^{M} \frac{g_m^2}{\sigma_{v_m}^2}}. \tag{C.37}$$

The denominator is clearly independent of either the TOA or the TDOAs of the rth reflection, so if the objective is to estimate these, we only need to consider the numerator. Interestingly, the numerator resembles the first part of the cost function in (C.27). This reveals the following interpretation of the M-step. First, the individual observations of the reflections are filtered by an MVDR filter, and the resulting output is then processed by a matched filter with the transmitted signal. The TOA and TDOAs that maximizes the output power of this operation are then the estimates for the rth reflection. This is in line with the findings in [C.24, C.25, C.26], where it was shown that the output of an MVDR/LCMV beamformer provide the sufficient statistics for estimating individual signals.

4.3 Spatio-temporarily correlated noise

We now consider the scenario, where the noise is spatio-temporarily correlated, a scenario practically encountered. For example, the late reverberation is often modeled as spatially homogeneous and isotropic sound field [C.20], resulting in a degree of spatial coherence which is dependent on the distance between the measurement points. Moreover, there might be interfering, quasi-periodic noise sources in the the recording environment, like human talkers, ego-noise from a drone/robot, etc. For such scenarios, we can rewrite the model in (C.4) as

$$\mathbf{y}(n) = \sum_{r=1}^{R} \mathbf{H}(\boldsymbol{\eta}_r, \mathbf{g}_r)\mathbf{s}(n - \tau_{\text{ref},r}) + \mathbf{d}(n), \tag{C.38}$$

where

$$\mathbf{d}(n) = \begin{bmatrix} \mathbf{d}_1^T(n) & \mathbf{d}_2^T(n) & \cdots & \mathbf{d}_M^T(n) \end{bmatrix}^T. \tag{C.39}$$

To deal with scenarios like this, we can preprocess the observed signals, such that the white Gaussian noise assumptions of the EM method is satisfied.

One way to achieve this is to use spatio-temporal decorrelation technique. Let us consider the correlated noise terms of the model in (C.4), i.e., $\mathbf{d}_m(n)$, for $m = 1, \ldots, M$. First, we define the spatio-temporal correlation matrix as

$$\mathbf{C}_d = \text{E}\left[\mathbf{d}(n)\mathbf{d}^T(n)\right]. \tag{C.40}$$

If we assume that this matrix is Hermitian and positive definite, the Cholesky factorization of it is given by

$$\mathbf{C}_d = \mathbf{L}\mathbf{L}^T, \tag{C.41}$$

where $\mathbf{L}$ is a lower triangular matrix with real and positive diagonal entries. That is, to whiten the noise term before estimating the unknown parameters, we can left-multiply the observation in (C.38) with $\mathbf{L}^{-1}$ [C.27]. The prewhitened observations are thus given by

$$\begin{aligned} \bar{\mathbf{y}}(n) &= \mathbf{L}^{-1}\mathbf{y}(n) \qquad \text{(C.42)} \\ &= \mathbf{L}^{-1}\sum_{r=1}^{R} \mathbf{H}(\boldsymbol{\eta}_r, \mathbf{g}_r)\mathbf{s}(n - \tau_{\text{ref},r}) + \bar{\mathbf{d}}(n), \end{aligned}$$

where $\overline{\mathbf{d}}(n) = \mathbf{L}^{-1}\mathbf{d}(n)$. Based on this and [C.23], we end up with the following EM method for estimating the acoustic reflection parameters when the noise is correlated in time and space:

E-step: for $r = 1, \ldots, R$, compute

$$\begin{aligned}\widehat{\mathbf{x}}_r^{(i)}(n) = \mathbf{H}\left(\widehat{\boldsymbol{\eta}}_r^{(i)}, \widehat{\mathbf{g}}_r^{(i)}\right)\mathbf{s}\left(n - \widehat{\tau}_{\text{ref},r}^{(i)}\right) \\ + \beta_r\left[\mathbf{y}(n) - \sum_{k=1}^{R}\mathbf{H}\left(\widehat{\boldsymbol{\eta}}_k^{(i)}, \widehat{\mathbf{g}}_k^{(i)}\right)\mathbf{s}\left(n - \widehat{\tau}_{\text{ref},k}^{(i)}\right)\right].\end{aligned} \tag{C.43}$$

M-step: for $r = 1, \ldots, R$,

$$\{\widehat{\mathbf{g}}_r, \widehat{\tau}_r, \widehat{\boldsymbol{\eta}}_r\}^{(i+1)} = \underset{\mathbf{g},\tau,\boldsymbol{\eta}}{\operatorname{argmin}}\, \overline{J}_r(\mathbf{g}, \tau, \boldsymbol{\eta}). \tag{C.44}$$

where

$$\overline{J}_r(\mathbf{g}, \tau, \boldsymbol{\eta}) = \left\|\mathbf{L}^{-1}\left(\widehat{\mathbf{x}}_r^{(i)}(n) - \mathbf{H}(\boldsymbol{\eta}, \mathbf{g})\mathbf{s}(n-\tau)\right)\right\|^2, \tag{C.45}$$

Eventually, we can explicitly write the cost function for the M-step as

$$\begin{aligned}\overline{J}_r(\mathbf{g}, \tau, \boldsymbol{\eta}) = \mathbf{x}_r^T(n)\mathbf{C}_d^{-1}\mathbf{x}_r(n) \\ + \mathbf{s}^T(n-\tau)\mathbf{H}^T(\boldsymbol{\eta}, \mathbf{g})\mathbf{C}_d^{-1}\mathbf{H}(\boldsymbol{\eta}, \mathbf{g})\mathbf{s}(n-\tau) \\ - 2\mathbf{x}_r^T(n)\mathbf{C}_d^{-1}\mathbf{H}(\boldsymbol{\eta}, \mathbf{g})\mathbf{s}(n-\tau),\end{aligned} \tag{C.46}$$

Compared with the cost function in (C.26), the minimization of (C.46) is more challenging. For example, the second term in (C.46) will generally depend on the DOA/TDOAs. That is, if we assume the reflections to be in the far-field of the array, we can adopt an iterative estimation scheme, where we first estimate the TOA and TDOAs, then update the TDOAs, and, finally, estimate the gains, i.e., for $r = 1, \ldots, R$:

Step 1: Obtain estimates of the TOA and TDOAs as

$$\{\widehat{\tau}_r, \widehat{\boldsymbol{\eta}}_r\} = \underset{\tau,\boldsymbol{\eta}}{\operatorname{argmax}}\, \mathbf{x}_r^T(n)\mathbf{C}_d^{-1}\overline{\mathbf{H}}(\boldsymbol{\eta}, \mathbf{g})\mathbf{s}(n-\tau), \tag{C.47}$$

where

$$\overline{\mathbf{H}}(\boldsymbol{\eta}) = \begin{bmatrix}\mathbf{D}_{\eta_1}^T & \cdots & \mathbf{D}_{\eta_M}^T\end{bmatrix}^T.$$

Step 2: Update the TDOA estimates as

$$\widehat{\boldsymbol{\eta}}_r = \arg\min_{\boldsymbol{\eta}} \overline{J}_{2,r}(g_r, \boldsymbol{\eta}) + \overline{J}_{3,r}(g_r, \boldsymbol{\eta}), \tag{C.48}$$

where

$$\overline{J}_{2,r}(g_r,\boldsymbol{\eta}) = g_r^2\mathbf{s}\left(n-\widehat{\tau}_r\right)\overline{\mathbf{H}}^T(\boldsymbol{\eta})\mathbf{C}_d^{-1}\overline{\mathbf{H}}(\boldsymbol{\eta}) \times \mathbf{s}(n-\widehat{\tau}_r) \tag{C.49}$$

$$\overline{J}_{3,r}(g_r,\boldsymbol{\eta}) = -2g_r\mathbf{x}_r^T(n)\mathbf{C}_d^{-1}\overline{\mathbf{H}}(\boldsymbol{\eta})\mathbf{s}(n-\widehat{\tau}_r). \tag{C.50}$$

Step 3: Estimate the unknown gain as

$$\widehat{g}_r = \frac{\mathbf{x}_r^T(n)\mathbf{C}_d^{-1}\overline{\mathbf{H}}(\widehat{\boldsymbol{\eta}}_r)\mathbf{s}(n-\widehat{\tau}_r)}{\mathbf{s}^T(n-\widehat{\tau}_r)\overline{\mathbf{H}}^T(\widehat{\boldsymbol{\eta}}_r)\mathbf{C}_d^{-1}\overline{\mathbf{H}}(\widehat{\boldsymbol{\eta}}_r)\mathbf{s}(n-\widehat{\tau}_r)}. \tag{C.51}$$

with the TOA and TDOA estimates from (C.47) and (C.48), respectively. If needed, these steps can then be repeated until convergence. It is also possible to simplify the M-step further by using particular signals as the known signal, $s(n)$. By close inspection of the second term of the cost function in (C.48), we get

$$\overline{J}_{2,r}(g_r,\boldsymbol{\eta}) = g_r^2\sum_{i=1}^{M}\sum_{j=1}^{M}c_{i,j} \times \mathbf{s}^T\left(n-\tau-\eta_i\right)\mathbf{s}\left(n-\tau-\eta_j\right), \tag{C.52}$$

where $c_{i,j}$ denotes the (i,j)th element of $\mathbf{C}_d^{-1}$. This reveals that, if the known probe signal is an uncorrelated noise sequence, it is reasonable to assume that this term is independent of both the TOA and the TDOAs, meaning that we can skip the update step in (C.48).

4.4 Kronecker decomposition

Another challenge with the prewhitening based estimator is the inversion of the noise covariance matrix, $\mathbf{C}_d$, which has a high dimension of $NM \times NM$. However, if we assume that the covariance matrix is separable, we can approximate it with two smaller matrices [C.28], i.e.,

$$\mathbf{C}_d \approx \mathbf{C}_\text{s} \otimes \mathbf{C}_\text{t}. \tag{C.53}$$

where $\mathbf{C}_\text{s}$ and $\mathbf{C}_\text{t}$ represents the spatial and temporal correlation matrices of dimensions $M \times M$ and $N \times N$, respectively, and $\otimes$ denotes the Kronecker product operator. Since $(\mathbf{C}_\text{s} \otimes \mathbf{C}_\text{t})^{-1} = \mathbf{C}_\text{s}^{-1} \otimes \mathbf{C}_\text{t}^{-1}$, we now only need to invert these smaller matrices, which is both numerically and computationally preferable. Moreover, we can now conduct the prewhitening using the Cholesky factorization of these smaller matrices due to the mixed-product property, yielding

$$\mathbf{C}_\text{s} \otimes \mathbf{C}_\text{t} = \mathbf{L}_\text{s}\mathbf{L}_\text{s}^T \otimes \mathbf{L}_\text{t}\mathbf{L}_\text{t}^T = (\mathbf{L}_\text{s} \otimes \mathbf{L}_\text{t})(\mathbf{L}_\text{s}^T \otimes \mathbf{L}_\text{t}^T). \tag{C.54}$$

In other words, by assuming separability, we can approximate $\mathbf{L}$ in (C.41) by $\mathbf{L}_\text{s} \otimes \mathbf{L}_\text{t}$. Eventually, it can be shown that, for uncorrelated probe signals, the Kronecker product decomposition allows us to rewrite the first step of the M-step in (C.44) as

Step 1:

$$
\begin{aligned}
\{\widehat{\tau}_r, \widehat{\boldsymbol{\eta}}_r\} &= \underset{\tau,\boldsymbol{\eta}}{\operatorname{argmax}}\, \mathbf{x}_r^T(n) \left(\mathbf{C}_\mathrm{s}^{-1} \otimes \mathbf{C}_\mathrm{t}^{-1}\right) \overline{\mathbf{H}}(\boldsymbol{\eta}, \mathbf{g})\mathbf{s}(n-\tau), \\
&= \underset{\tau,\boldsymbol{\eta}}{\operatorname{argmax}}\, \mathrm{tr}\left(\mathbf{X}_r^T(n)\mathbf{C}_\mathrm{t}^{-1}\mathbf{S}_{\tau,\boldsymbol{\eta}}(n)\mathbf{C}_\mathrm{s}^{-1}\right) && \text{(C.55)} \\
&= \underset{\tau,\boldsymbol{\eta}}{\operatorname{argmax}} \sum_{m=1}^{M} \widetilde{\mathbf{x}}_{m,r}^T(n)\widetilde{\mathbf{s}}(n-\tau-\eta_m) && \text{(C.56)}
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{X}_r(n) &= \begin{bmatrix}\mathbf{x}_{1,r}(n) & \cdots & \mathbf{x}_{M,r}(n)\end{bmatrix}, && \text{(C.57)} \\
\mathbf{S}_{\tau,\boldsymbol{\eta}}(n) &= \begin{bmatrix}\mathbf{D}_{\eta_1}\mathbf{s}(n-\tau) & \cdots & \mathbf{D}_{\eta_M}\mathbf{s}(n-\tau)\end{bmatrix}, \\
&= \begin{bmatrix}\mathbf{s}(n-\tau-\eta_1) & \cdots & \mathbf{s}(n-\tau-\eta_M)\end{bmatrix}, && \text{(C.58)}
\end{aligned}
$$

and the vectors $\widetilde{\mathbf{x}}_{m,r}(n)$ and $\widetilde{\mathbf{s}}(n-\tau-\eta_m)$ are the prewhitened observation and probe signals for microphone m, respectively, defined as the m'th columns of the following matrices:

$$
\begin{aligned}
\widetilde{\mathbf{X}}_r(n) &= \mathbf{L}_\mathrm{t}^{-1}\mathbf{X}_r(n)\mathbf{L}_\mathrm{s}^{-T} && \text{(C.59)} \\
\widetilde{\mathbf{S}}_{\tau,\boldsymbol{\eta}}(n) &= \mathbf{L}_\mathrm{t}^{-1}\mathbf{S}_{\tau,\boldsymbol{\eta}}(n)\mathbf{L}_\mathrm{s}^{-T}. && \text{(C.60)}
\end{aligned}
$$

These expressions can be interpreted in the following way. The left hand multiplication with $\mathbf{L}_t^{-1}$ corresponds to temporal prewhitening of all the microphone signals, whereas the right hand multiplication with $\mathbf{L}_s^{-T}$ corresponds to spatial prewhitening of all time snapshots.

Step 2: With the Kronecker decomposition, the second term of the cost function in (C.49) becomes

$$
\overline{J}_{2,r}(g_r, \boldsymbol{\eta}) = g_r^2 \mathrm{tr}(\widetilde{\mathbf{S}}_{\tau,\eta}^T(n)\widetilde{\mathbf{S}}_{\tau,\eta}(n)). \tag{C.61}
$$

This does not depend on the TOAs and TDOAs, so the Kronecker decompositions allow us to skip the intermediate step of updating the TDOAs as in (C.48). We can therefore directly proceed to conducting the closed form estimate of the gains as

$$
\widehat{g}_r = \frac{\sum_{m=1}^{M} \widetilde{\mathbf{x}}_{m,r}^T(n)\widetilde{\mathbf{s}}(n-\tau-\eta_m)}{M\|\widetilde{\mathbf{s}}(n)\|^2}. \tag{C.62}
$$

Even after all the presented simplifications and assumptions, the computational complexity of the proposed methods might still be considered relatively high due to their iterative and multidimensional nature. However, although not considered in this paper, we expect that further reductions in the computational complexity can be obtained by employing, e.g., the space alternating generalized expectation (SAGE) algorithm rather than the EM algorithm [C.29], or through a recursive EM procedure as suggested in [C.30], where the number of iterations per time instance can be reduced by instead tracking the parameters of interest over time.

4.5 Temporal prewhitening with filter

One issue with this prewhitening approach still is that the number samples in time might be relatively high in practice. The consequence of this is that, even with the Kronecker decomposition of the noise correlation matrix, the inversion of $\mathbf{L}_\mathrm{t}$ might be intractable in practice since its dimensions equal the number of time samples. An alternative approach could be to use a lower order filter for the prewhitening instead [C.31]. If we assume that the noise follows an autoregressive model, we can approximate it as:

$$d(n) \approx \sum_{p=1}^{P} a_p d(n-p). \tag{C.63}$$

Given the noise correlation matrix, $\mathbf{C}_\mathrm{t}$, we can obtain the AR coefficients of the noise using the Levinson-Durbin recursion. The prewhitening filter is then formed using the AR coefficients as the coefficients of a P'th order FIR filter, $h_\mathrm{pw}(p) = a_p$. Subsequently, the prewhitened signals are obtained as

$$\widetilde{x}_{m,r}(n) = \sum_{p=0}^{P} h_\mathrm{pw}(p) x_{m,r}(n-p), \tag{C.64}$$

$$\widetilde{s}(n) = \sum_{p=0}^{P} h_\mathrm{pw}(p) s(n-p), \tag{C.65}$$

where $h_\mathrm{pw}(0) = 1$.

4.6 Covariance estimation

In the previous subsections, we have considered the covariance matrices as known quantities. However, we need to estimate these from the observed data in practice. If no particular structure is assumed for the covariance matrix, a common approach is to use the following estimator [C.32]

$$\widehat{\mathbf{C}}_d = \frac{1}{N-K+1} \sum_{n=0}^{N-K} \mathbf{d}(n)\mathbf{d}(n)^T, \tag{C.66}$$

where

$$\mathbf{d}(n) = \begin{bmatrix} \mathbf{d}_1(n) & \cdots & \mathbf{d}_M(n) \end{bmatrix}^T, \tag{C.67}$$

$$\mathbf{d}_m(n) = \begin{bmatrix} d_m(n) & \cdots & d_m(n+K-1) \end{bmatrix}^T. \tag{C.68}$$

As evident from, e.g., (C.47), the estimated covariance needs to be invertible. This requires that

$$K \leq \frac{N+1}{M+1}. \tag{C.69}$$

Algorithm 1: Flip-flop algorithm [C.33].

```
**Result:** Estimates of temporal and spatial covariance matrices, $\widehat{\mathbf{C}}_{\text{t}}$ and $\widehat{\mathbf{C}}_{\text{s}}$.
$\mathbf{D}(n) = \begin{bmatrix} \mathbf{d}_1(n) & \cdots & \mathbf{d}_M(n) \end{bmatrix}$;
$\widehat{\mathbf{C}}_{\text{s}} = \mathbf{I}$;
$\widehat{\mathbf{C}}_{\text{t}} = \frac{1}{M(N-K+1)} \sum_{n=0}^{N-K} \mathbf{D}(n)\widehat{\mathbf{C}}_{\text{s}}^{-1}\mathbf{D}^T(n)$;
**repeat**
    $\widehat{\mathbf{C}}_{\text{s}} = \frac{1}{K(N-K+1)} \sum_{n=0}^{N-K} \mathbf{D}^T(n)\widehat{\mathbf{C}}_{\text{t}}^{-1}\mathbf{D}(n)$;
    $\widehat{\mathbf{C}}_{\text{t}} = \frac{1}{M(N-K+1)} \sum_{n=0}^{N-K} \mathbf{D}(n)\widehat{\mathbf{C}}_{\text{s}}^{-1}\mathbf{D}^T(n)$;
**until** *convergence*;
```

where K is the number of snapshots, N is the number of samples of the signal and M is the number of microphones. Consequently, we can only use relatively short temporal subvectors, $\mathbf{d}_m(n)$ in the estimation of the covariance matrix when the number of microphones is increased.

If it is assumed that the multichannel noise samples in $\mathbf{d}(n)$ follows a multichannel matrix normal distribution, the maximum likelihood (ML) estimator for the noise covariance matrix can be derived [C.33]. Unfortunately, the resulting estimator is not closed-form, but it can be implemented using the iterative flip-flop algorithm in Algorithm 1. In some cases, e.g., if one of the covariance matrices are close to being rank deficient, this iterative procedure can be problematic, since their inverses are required. Different approaches for dealing with this and the computational complexity of the iterative procedure have been considered [C.32, C.34]. Alternatively, a non-iterative estimator can be used such as [C.32]

$$\widehat{\mathbf{C}}_{\text{s}} = \frac{1}{(N-K+1)\text{tr}\left(\mathbf{C}_{\text{t}}\right)} \sum_{n=0}^{N-K} \mathbf{D}^T(n)\mathbf{D}(n), \tag{C.70}$$

$$\widehat{\mathbf{C}}_{\text{t}} = \frac{1}{(N-K+1)\text{tr}\left(\widehat{\mathbf{C}}_{\text{s}}\right)} \sum_{n=0}^{N-K} \mathbf{D}(n)\mathbf{D}^T(n), \tag{C.71}$$

where

$$\mathbf{D}(n) = \begin{bmatrix} \mathbf{d}_1(n) & \mathbf{d}_2(n) & \cdots & \mathbf{d}_M(n) \end{bmatrix}. \tag{C.72}$$

As indicated in (C.70), the trace of the temporal covariance is assumed to be known. This might not be the case in practice, however, in most situations we can simply replace it by an arbitrary value, since its main purpose is to resolve the ambiguity

$$\mathbf{C}_d = \mathbf{C}_{\mathrm{s}} \otimes \mathbf{C}_{\mathrm{t}} = \left(\frac{1}{\alpha}\mathbf{C}_s\right) \otimes (\alpha \mathbf{C}_{\mathrm{t}}). \tag{C.73}$$

4.7 Non-stationary noise

While the stationarity assumption may not hold in practice, there are a number of ways to address this problem. For example, we may reduce the length, N, of the probe signal and the analysis window, which would naturally increase the validity of the assumption. Alternatively, we may decouple the prewhitening and estimation parts, as suggested in Section 4.5. In this way, We may first prewhiten our signal using a filter, and then apply the proposed estimators with a white Gaussian noise assumption on the prewhitened signals. This approach can be exploited to take the non-stationarity of the noise into account by updating the prewhitening filters over time, according to the changing AR coefficients of the noise. Estimating non-stationary noise parameters, however, is more difficult, since the statistics need to be tracked during the presence of the desired signal, i.e., the probe signal and its reflections in our case. This problem has been well-investigated in other audio signal processing problems, such as speech enhancement [C.35, C.36, C.37, C.38].

5 Results and Discussion

In this section, we investigate the performance of the different variants of the proposed EM method. More specifically, we consider the variant assuming spatially independent white Gaussian in Section 4.1 resulting in noise variance weighting (EM-UCA-NW), and its special case where the noise variance is assumed equal (EM-UCA) [C.19]. Moreover, we consider the setup with correlated noise proposed in Section 4.3 resulting in the prewhitening-based approach (EM-UCA-PW). The experiments were carried out using signals that were generated using the room impulse response generator [C.39]. The dimensions of the simulated room were set to $8 \times 6 \times 5$ m, the reverberation time (T_{60}) was set to 0.6 s while the speed of sound is fixed at 343 m/s. The loudspeaker was positioned at the center of an UCA at $(1 \times 1.5 \times 2.5)$ m while the UCA has $M = 4$ microphones with a radius of $d = 0.2$ m. Although, any type of known broadband signal could be used to probe the environment, such as a chirp signal or maximum length sequences (MLS) [C.40], we decided to use a white Gaussian noise sequence as the known sound source, $s(n)$, consisting of $1,500$ samples from a Gaussian distribution. This sequence was subsequently zero-padded to get a total signal length of $20,000$ samples. The objective of the zero-padding was to get a longer analysis window to ensure that the first few reflections are present in the observation. Moreover, as discussed in Section 4.3, the reason for using a WGN sequence is that the EM estimator can be simplified if the probe signal is an uncorrelated signal. In addition to this, using such a broadband sequence minimizes the effects of spatial aliasing [C.41]. The sampling frequency f_s was set to $22,050$ Hz. We assumed that

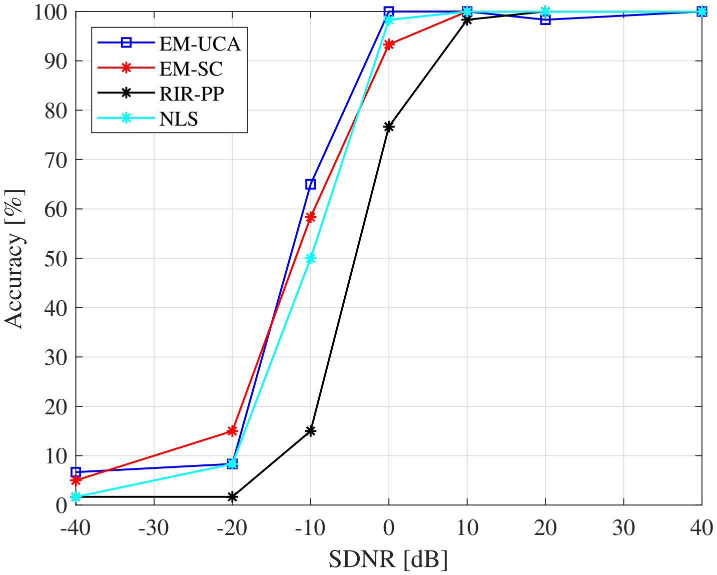


Fig. C.3: Comparison of the proposed EM-UCA method with state-of-the-art methods in terms of TOA estimation accuracy.

the direct-component is subtracted from the observed signal given that we know the arrangement of the loudspeaker and the microphones. Knowing the array geometry, enables either: offline measurement of the impulse response of the direct-path component offline; or analytical computation of the impulse response of the direct-path component based on the geometry. The background noise comprises of two components: one being diffuse spherical noise and the other being thermal sensor noise. The diffuse spherical noise was generated using the method described in [C.42] using the rotor noise of a drone from the DREGON database [C.3]. The drone audio file used to generate the diffuse spherical noise corresponds to rotors running at 70 revolutions per second (RPS). The thermal sensor noise was simulated as spatially independent white Gaussian noise. Both these noises were added to the observed signal before estimating the parameters. The evaluation was then conducted for different signal-to-diffuse noise ratios (SDNRs) and signal-to-sensor noise ratios (SSNRs). In the following subsections, we evaluate the performance of our propose method in various conditions.

5.1 Comparison of with state-of-the-art

The aim of the first experiment was to compare the proposed method with existing state-of-the-art methods. The EM algorithm was set to estimate $R = 3$ reflections with 40 iterations and β was set to $\frac{1}{R}$. The main application for this manuscript is acoustic reflector mapping for robot audition. For this application, the mapping should be possible in unknown, complex environments, and we therefore do not rely on trivial room geometry models as opposed to many of the traditional methods for room geometry estimation [C.11, C.12, C.13]. Therefore, we chose to use a small number of reflections in the estimation (i.e., $R = 3$), to mainly estimate the TOA's/DOA's of first-order reflections impinging from nearby acoustic reflectors. These can be directly mapped to acoustic reflector positions based on the estimated time and angle of arrival. While this will not facilitate the localization of all acoustic reflectors at any given time instance, we can carry out such estimation over time and space, to generate a map of an arbitrary room geometry (see Section 5.4). An alternative to choosing a fixed reflection order, would be to combine the proposed method with order estimation methods [C.43, C.44]. To initialize the method, the gain estimates, $\widehat{g}_{m,r}$, were sampled from a uniform distribution over the interval $[0; 1]$, the TOAs, $\widehat{\tau}_{1,r}$, were sampled from a uniform discrete distribution over the time indices corresponding to the analysis window, and the DOAs, $\widehat{\phi}_r$, were sampled from a uniform distribution over the interval $[0°; 360°]$. After emitting and recording the known source signal, an analysis window of each recording was considered starting from τ_{min} samples to τ_{max} samples after the source signal was emitted. In this experiment, the analysis window was set such that the search is made between 0.5 m to 2 m. This was done to primarily capture the first order reflections. The lower bound was chosen because we can only search for reflectors that are outside the geometry of the array, which, in our experiments, had a radius of 0.2 m. After 2 m, the performance of the proposed method degrades because the energy of the reflected signals decrease quadratically over distance, which motivated the choice of the upper limit.

The proposed EM method (EM-UCA) was compared to the single-channel EM method (EM-SC) in [C.23] in terms of TOA accuracy, applied to the first microphone. Moreover, these were compared with a common approach to extracting TOAs from estimated RIR through peak-picking (RIR-PP). Finally, the performance was also compared with our previous work [C.45] termed the non-linear least squares estimator (NLS). The results for the TOA estimation are shown in Fig. C.3, where the accuracy was defined as the percentage of TOA estimates that were within ± 2 % tolerance of one of the true parameters of the first-order reflections computed using the image-source method. This was measured for different SDNRs while the SSNR was fixed to 10 dB, and for each SDNR the accuracy was measured over 100 Monte-Carlo simulations. As seen in Figure C.3, the proposed method clearly outperforms the existing method by providing higher accuracy at lower SDNRs.

Furthermore, the computation time of the RIR-PP and the proposed method, EM-UCA, were measured. This test was performed in MATLAB using the built-in function *timeit* on a standard desktop computer running a Microsoft Windows 10 operating system with an Intel Core i7 CPU with 3.40 GHz processing speed and 16 GB of RAM. A Monte Carlo simulation with 100 trials was performed on each method and an average time was calculated. The measured computation times of the RIR-PP and the EM-UCA were 0.0063 s and 25.74 s, respectively, for $R = 1$ and an SDNR of 40 dB. This shows that the improved estimation accuracy with the proposed method comes at the cost of a higher computational complexity. It is important to stress, however, that in

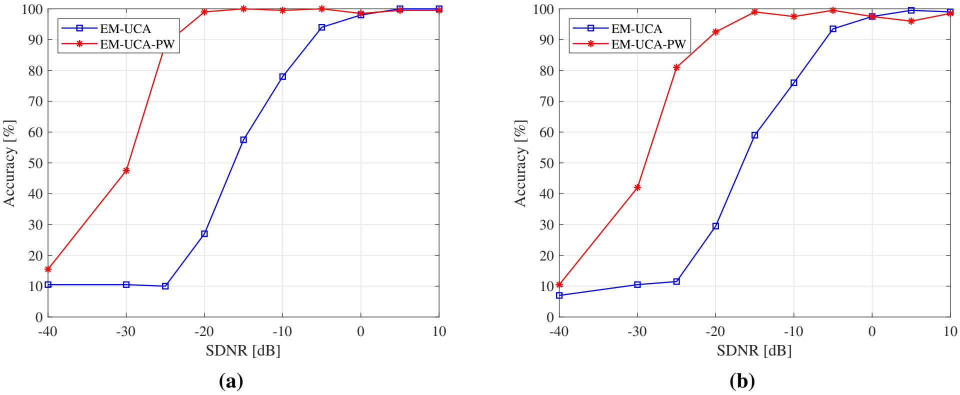
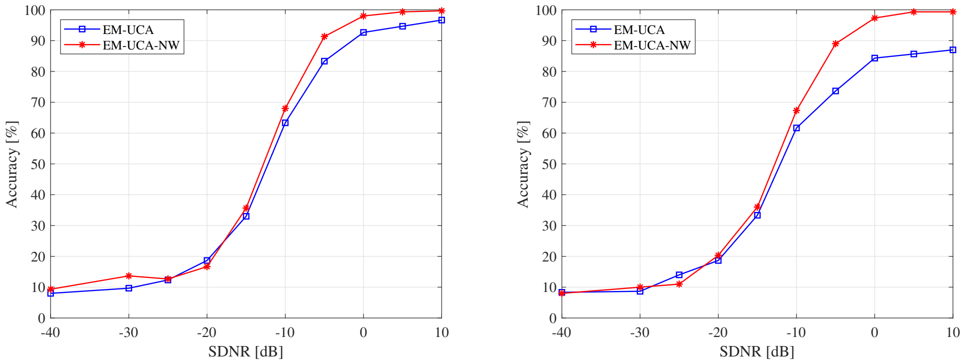


Fig. C.4: a) TOA estimation accuracy of the proposed EM method with and without prewhitening. b) DOA estimation accuracy of the proposed EM method with and without prewhitening.

applications such as acoustic reflector localization with a drone, it is common to have negative SNR conditions [C.46], where the RIR-PP method may fail to provide accurate estimates as opposed to the proposed method (see, e.g., Figure C.3). Moreover, the computational cost could be reduced further by, e.g., employing the recursive EM approach [C.30, C.47]. If the TOA/DOA estimation is carried out continuously over time and space, the EM algorithm may be initialized using previous estimates, which may significantly reduce the number of iterations needed for convergence. Another potential computational saving may be obtained by deriving the proposed methods in the frequency domain.

5.2 Evaluation for different diffuse noise conditions

In the second experiment, we evaluated the effect of the proposed prewhitening approach under different diffuse noise conditions. To test the performance of the EM algorithm under such realistic scenarios, we test our estimator for different SDNRs in the interval $[-40; 10]$ dB while setting the SSNR to 40 dB. Here, we are comparing the EM algorithm with and without the prewhitening in terms of both TOA and DOA estimation accuracy as seen in Fig. C.4(a) and Fig. C.4(b), respectively. The diffuse rotor noise is indeed correlated with strong periodic components, but the results show that the proposed prewhitening approach can successfully account for this, and can retain a high estimation accuracy at SDNRs levels 20 dB lower than those needed for the EM-UCA approach.



(a)

(b)

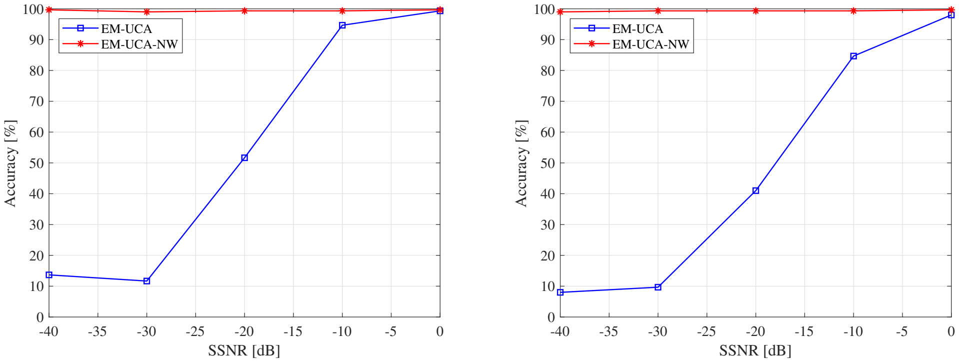
Fig. C.5: a) TOA estimation accuracy of the proposed EM method with and without noise variance weighting when 1 microphone has a lower SSNR of -10 dB while the remaining microphones has a SSNR of 40 dB. b) DOA estimation accuracy of the proposed EM method with and without noise variance weighting when 1 microphone has a lower SSNR of -10 dB while the remaining microphones has a SSNR of 40 dB.

5.3 Evaluation for faulty/noisy microphone conditions

In this experiment, we consider a scenario where one microphone is excessively noisy compared to the other microphones. An example of this could be a robot platform, where one microphone is placed closer to an ego-noise source such as a fan, leading to TOA and DOA estimation errors. To simulate this effect, we set thermal noise of a single microphone to an SSNR level of -10 dB, while the thermal noise of the remaining microphones are set to an SSNR level of 40 dB. As seen in Fig. C.5(a) and C.5(b), the performance of the EM algorithm with noise variance weighting is less affected by the high thermal sensor noise in terms of both TOA and DOA estimation accuracy. Moreover, we conducted an experiment without diffuse noise, where the SSNR level of the faulty microphone was changed from -40 dB to 0 dB. These results are shown in Figures C.6(a) and C.6(b), and show that the estimation accuracy is already degrading from 0 dB SSNR and downwards when using the EM-UCA approach, whereas the proposed EM-UCA-NW approach retains a high accuracy.

5.4 Application example of the proposed method

We consider an application example where the localization of the acoustic reflectors is done using the proposed EM method with and without prewhitening. More specifically, we have used filter-based prewhitening approach as discussed in Sub-section 4.5. This experiment thus shows how the proposed method can be used to map an environment using a moving robot platform. The room parameters were kept the same as the earlier experiment. Furthermore, the SDNR was



(a) **(b)**

Fig. C.6: a) TOA estimation accuracy of the proposed EM method with and without noise variance weighting for different SSNR levels for one of the microphones. b) DOA estimation accuracy of the proposed EM method with and without noise variance weighting for different SSNR levels for one of the microphones.

set to -10 dB corresponding to a strong ego-noise. The loudspeaker-microphone arrangement was similar to the previous experiments, and follows the a predefined path as shown in Fig. C.7 indicated by the blue dashed line. As depicted in the figure, the EM algorithm with prewhitening performs better at estimating acoustic reflector using the estimated TOAs and DOAs, compared to EM algorithm without prewhitening.

6 Conclusion

In this paper, we consider the problem of estimating the time- and direction-of-arrivals of acoustic echoes using a loudspeaker emitting a known source signal and multiple microphones. Among other examples, this is an important problem in robot and drone audition, where these parameters can reveal the positions of nearby acoustic reflectors, and thus facilitate mapping and navigation of a physical environment. Some methods exist for solving the problems of acoustic reflector localization and room geometry estimation, however, most of these rely on a priori information, e.g., of the TOAs or DOAs of the acoustic echoes. However, estimating these is a difficult problem on its own, which is dealt with by the methods proposed herein. Moreover, even when the TOAs are estimated for some of the traditional approaches, the difficult problem of echolabeling needs to be solved, since the order of the corresponding reflection is generally unknown. We therefore propose different methods for estimating, not only the TOAs, but also the DOAs of acoustic echoes. By estimating the DOAs also, it is possible to resolve some of the ambiguity introduced by knowing only the TOAs. The proposed method is based on the expectation-maximization framework, and are derived to be optimal under different conditions

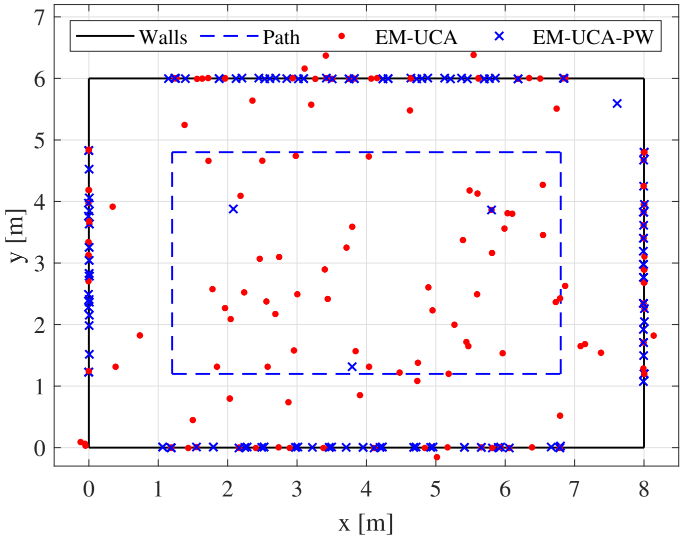


Fig. C.7: Example of reflector localization for different array positions using the proposed EM method with and without prewhitening based on TOA and DOA estimates at an SDNR of -10 dB.

ranging from the simple white Gaussian noise scenario to scenarios with correlated and colored noise. In the experiments, we show that proposed methods are able to estimate the TOAs and DOAs with higher accuracy and noise robustness compared to existing methods. Moreover, we show that some of the proposed variants can account for colored noise and scenarios where a microphone is faulty or more noisy than the other microphones of the array. Finally, we conducted a more applied experiment, where it is illustrated how a room can be mapped from the estimated parameters, which is relevant to, e.g., autonomous robot and drone applications. While the proposed method has a higher computation time than traditional methods, this can be reduced significantly by adopting the recursive EM scheme and deriving the proposed methods in the frequency domain.

Abbreviations

TOA: Time-of-arrival; EM: Expectation-maximization; UCA: Uniform circular array; SNR: Signal-to-noise ratio; DOA: Direction-of-arrival; aSLAM: Acoustic simultaneous localization and mapping; RIR: Room impulse response; TDOA: Time difference-of-arrival; ML: Maximum likelihood; MVDR: Minimum variance distortionless response; LCMV: Linearly constrained minimum variance; SAGE: Space alternating generalized expectation; FIR: Finite impulse response; AR: Autoregressive; EM-UCA: proposed method without prewhitening or noise weighting; EM-UCA-NW: proposed method with only noise weighting; EM-UCA-PW: proposed method with only prewhitening; T_{60}: Reverberation time (60 dB); RPM: revolutions per minute; DREGON: Database of drone audio recordings; SDNR: Signal-to-diffuse-noise ratio; SSNR: Signal-to-sensor-noise ratio; EM-SC: Single channel EM method; RIR-PP: RIR-based method with peak picking; NLS: nonlinear least squares;

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

Authors' contributions

JRJ and SG designed the idea for the manuscript. JRJ and US conducted the experiments. All the authors contributed to the writing of this work. Moreover, all author(s) read and approved the final manuscript.

Acknowledgements

Not applicable.

References

[C.1] Rascon, C., Meza, I.: Localization of sound sources in robotics: A review. Robotics and Autonomous Systems **96**, 184–210 (2017)

[C.2] Löllmann, H.W., Moore, A., Naylor, P.A., Rafaely, B., Horaud, R., Mazel, A., Kellermann, W.: Microphone array signal processing for robot audition. Hands-free Speech Comm. and Microphone Arrays, 51–55 (2017)

[C.3] Strauss, M., Mordel, P., Miguet, V., Deleforge, A.: DREGON: Dataset and methods for UAV-embedded sound source localization. IEEE/RJS Int. Conf. Intelligent Robots and Systems, 5735–5742 (2018)

[C.4] Badeig, F., Pelorson, Q., Arias, S., Drouard, V., Gebru, I.D., Li, X., Evangelidis, G., Horaud, R.: A distributed architecture for interacting with NAO. Int. Conf. Multimodal Interaction (2015)

[C.5] Antonacci, F., Filos, J., Thomas, M.R.P., Habets, E.A.P., Sarti, A., Naylor, P.A., Tubaro, S.: Inference of room geometry from acoustic impulse responses. IEEE Trans. Audio, Speech, and Language Process. **20**(10), 2683–2695 (2012)

[C.6] Coutino, M., Møller, M.B., Nielsen, J.K., Heusdens, R.: Greedy alternative for room geometry estimation from acoustic echoes: A subspace-based method. Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 366–370 (2017)

[C.7] Saqib, U., Gannot, S., Jensen, J.R.: Estimation of acoustic echoes using expectation-maximization methods. EURASIP Journal on Audio, Speech and Music **2020** (2020). Accepted for publication

[C.8] Hu, J.-S., Chan, C.-Y., Wang, C.-K., Lee, M.-T., Kuo, C.-Y.: Simultaneous localization of a mobile robot and multiple sound sources using a microphone array. Adv. Robotics **25**(1–2), 135–152 (2011)

[C.9] Ogiso, S., Kawagishi, T., Mizutani, K., Wakatsuki, N., Zempo, K.: Self-localization method for mobile robot using acoustic beacons. ROBOMECH J. **2**(1), 12 (2015)

[C.10] Evers, C., Naylor, P.A.: Acoustic SLAM. IEEE/ACM Trans. Audio, Speech, and Language Process. **26**, 1484–1498 (2018)

[C.11] Kreković, M., Dokmanić, I., Vetterli, M.: EchoSLAM: Simultaneous localization and mapping with acoustic echoes. Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 11–15 (2016)

[C.12] Nguyen, L., Miro, J.V., Qiu, X.: Can a robot hear the shape and dimensions of a room? IEEE/RSJ Int. Conf. Intell. Robots and Syst., 5346–5351 (2019)

[C.13] Wang, T., Peng, F., Chen, B.: First order echo based room shape recovery using a single mobile device. Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 21–25 (2016)

[C.14] Kelly, I.J., Boland, F.M.: Detecting arrivals in room impulse responses with dynamic time warping. IEEE/ACM Trans. Audio, Speech, and Language Process. **22**(7), 1139–1147 (2014)

[C.15] Plumbley, M.D.: Hearing the shape of a room. Proc. Natl. Acad. Sci. U S A **110**(30), 12162–12163 (2013)

[C.16] Nelson, L.B., Poor, H.V.: Iterative multiuser receivers for CDMA channels: an EM-based approach. IEEE Trans. Commun. **44**(12), 1700–1710 (1996)

[C.17] Vanderveen, M.C., Papadias, C.B., Paulraj, A.: Joint angle and delay estimation (JADE) for multipath signals arriving at an antenna array. IEEE Commun. Lett. **1**(1), 12–14 (1997)

[C.18] Verhaevert, J., Lil, E.V., de Capelle, A.V.: Direction of arrival (DOA) parameter estimation with the SAGE algorithm. Signal Processing **84**(3), 619–629 (2004)

[C.19] Jensen, J.R., Saqib, U., Gannot, S.: An EM method for multichannel TOA and DOA estimation of acoustic echoes. Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust. (2019)

[C.20] Braun, S., Kuklasiński, A., Schwartz, O., Thiergart, O., Habets, E.A.P., Gannot, S., Doclo, S., Jensen, J.: Evaluation and comparison of late reverberation power spectral density estimators. IEEE/ACM Trans. Audio, Speech, and Language Process. **26**(6), 1056–1071 (2018)

[C.21] Cron, B.F., Sherman, C.H.: Spatial-correlation functions for various noise models. J. Acoust. Soc. Am. **34**(11), 1732–1736 (1962)

[C.22] Sun, H., Abhayapala, T.D., Samarasinghe, P.N.: Active noise control over 3D space with multiple circular arrays. Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust., 135–139 (2019)

[C.23] Feder, M., Weinstein, E.: Parameter estimation of superimposed signals using the EM algorithm. IEEE Trans. Acoust., Speech, Signal Process. **36**(4), 477–489 (1988)

[C.24] Schwartz, O., Gannot, S., Habets, E.A.P.: Multispeaker LCMV beamformer and postfilter for source separation and noise reduction. IEEE/ACM Trans. Audio, Speech, and Language Process. **25**(5), 940–951 (2017)

[C.25] Balan, R., Rosca, J.: Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase. Proc. IEEE Workshop Sensor Array and Multichannel Signal Process., 209–213 (2002)

[C.26] Scharf, L.L.: Statistical Signal Processing: Detection, Estimation, and Time Series Analysis. Addison-Wesley Publishing Company, Michigan (1991)

[C.27] Hansen, P.C., Jensen, S.H.: Prewhitening for rank-deficient noise in subspace methods for noise reduction. IEEE Trans. Signal Process. **53**(10), 3718–3726 (2005)

[C.28] Reinsel, G.: Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. J. Am. Statist. Assoc. **77**(377), 190–195 (1982)

[C.29] Fessler, J.A., Hero, A.O.: Space-alternating generalized expectation-maximization algorithm. IEEE Trans. Signal Process. **42**(10), 2664–2677 (1994)

[C.30] Schwartz, O., Gannot, S.: Speaker tracking using recursive EM algorithms. IEEE/ACM Trans. Audio, Speech, and Language Process. **22**(2), 392–402 (2014)

[C.31] Nørholm, S.M., Jensen, J.R., Christensen, M.G.: Instantaneous fundamental frequency estimation with optimal segmentation for nonstationary voiced speech. IEEE/ACM Trans. Audio, Speech, and Language Process. **24**(12), 2354–2367 (2016)

[C.32] Castaneda, M.H., Nossek, J.A.: Estimation of rank deficient covariance matrices with Kronecker structure. Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 394–398 (2014)

[C.33] Dutilleul, P.: The MLE algorithm for the matrix normal distribution. J. Statist. Comput. Simul. **64**(2), 105–123 (1999)

[C.34] Werner, K., Jansson, M., Stoica, P.: On estimation of covariance matrices with Kronecker product structure. IEEE Trans. Signal Process. **56**(2), 478–491 (2008)

[C.35] Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**(5), 504–512 (2001)

[C.36] Gerkmann, T., Hendriks, R.C.: Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio, Speech, and Language Process. **20**(4), 1383–1393 (2012)

[C.37] Hendriks, R.C., Heusdens, R., Jensen, J.: MMSE based noise PSD tracking with low complexity. Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 4266–4269 (2010)

[C.38] Nielsen, J.K., Kavalekalam, M.S., Christensen, M.G., Boldt, J.: Model-based noise PSD estimation from speech in non-stationary noise. Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 5424–5428 (2018)

[C.39] Habets, E.A.P.: Room impulse response generator. Technical report, Technische Universiteit Eindhoven (2010). Ver. 2.0.20100920. `https://github.com/ehabets/RIR-Generator`

[C.40] Florencio, D., Zhang, Z.: Maximum a posteriori estimation of room impulse responses. Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 728–732 (2015)

[C.41] Dmochowski, J., Benesty, J., Affes, S.: On spatial aliasing in microphone arrays. IEEE Trans. Signal Process. **57**(4), 1383–1395 (2009)

[C.42] Habets, E.A.P., Cohen, I., Gannot, S.: Generating nonstationary multisensor signals under a spatial coherence constraint. J. Acoust. Soc. Am. **124**(5), 2911–2917 (2008)

[C.43] Han, K., Nehorai, A.: Improved source number detection and direction estimation with nested arrays and ULAs using jackknifing. IEEE Trans. Signal Process. **61**(23), 6118–6128 (2013)

[C.44] Stoica, P., Selen, Y.: Model-order selection: a review of information criterion rules. IEEE Signal Process. Mag. **21**(4), 36–47 (2004)

[C.45] Saqib, U., Jensen, J.R.: Sound-based distance estimation for indoor navigation in the presence of ego noise. Proc. European Signal Processing Conf., 1–5 (2019)

[C.46] Deleforge, A., Di Carlo, D., Strauss, M., Serizel, R., Marcenaro, L.: Audio-based search and rescue with a drone: Highlights from the IEEE signal processing cup 2019 student competition [SP competitions]. IEEE Signal Process. Mag. **36**(5), 138–144 (2019)

[C.47] Weisberg, K., Gannot, S., Schwartz, O.: An online multiple-speaker DOA tracking using the CappÉ-Moulines recursive expectation-maximization algorithm. Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 656–660 (2019)

Paper D

A model-based approach to acoustic reflector localization using robotic platform

Usama Saqib and Jesper Rindom Jensen

The paper has been published in the
International Conference on Intelligent Robots and Systems (IROS)

The layout has been revised.

Abstract

Constructing a spatial map of an indoor environment, e.g., a typical office environment with glass surfaces, is a difficult and challenging task. Current state-of-the-art, e.g., camera- and laser-based approaches are unsuitable for detecting transparent surfaces. Hence, the spatial map generated with these approaches are often inaccurate. In this paper, a method that utilizes echolocation with sound in the audible frequency range is proposed to robustly localize the position of an acoustic reflector, e.g., walls, glass surfaces etc., which could be used to construct a spatial map of an indoor environment as the robot moves. The proposed method estimate the acoustic reflector's position, using only a single microphone and a loudspeaker that are present on many socially assistive robot platforms such as the NAO robot. The experimental results show that the proposed method could robustly detect an acoustic reflector up to a distance of 1.5 *m in more than* 60% *of the trials and works efficiently even under low SNRs. To test the proposed method, a proof-of-concept robotic platform was build to construct a spatial map of an indoor environment.*

1 Introduction

Constructing a spatial map of a dynamic environment using a robotic platform is useful, for example, in navigation, analysis and monitoring environments such as buildings, tunnels, etc., for maintenance purposes [D.1]. Simultaneous Localization And Mapping (SLAM) is a popular framework among the robotic community to generate a spatial map of an indoor environment as well as to localize and orient the pose of the robot [D.2, D.3, D.4]. SLAM is often used in conjunction with external sensors such as cameras, lasers, etc., to receive environmental data. However, camera-based systems are susceptible to changing light conditions which affects the accuracy of the system. Moreover, camera has limited field of view (FOV) which makes it unsuitable to detect objects around the corner [D.5]. Similarly, Light Detection and Ranging (LiDAR) is a laser-based range sensing technology that is often used with SLAM to accurately generate a spatial map of an environment [D.6]. However, both LiDAR and camera systems are unsuitable for detecting transparent surfaces that are typically found in an office environment [D.7].

These issues could be resolved by employing sound. Sound has been widely used in robotics to detect acoustic sources [D.8] but in this paper we consider echolocation [D.9, D.10]. The advantage of incorporating echolocation on a robotic platform is that it enable robots to navigate under low light conditions or even under complete darkness [D.11, D.12]. Moreover, microphones can be omnidirectional as opposed to common cameras which make their FOV larger. Additionally, microphones can be used to detect audible sources not within direct line of sight. Therefore, constructing a spatial map of an environment using echolocation can be desirable in such difficult scenarios. In recent years, some works on combining echolocation with camera-based systems have been conducted to generate a spatial map of a room. For instance, in [D.7], the authors proposed using laser and ultrasonic sensor to detect glass surfaces that aids a robot in navigating a room. However, most robotic platforms especially those used for Human-Robot Interaction (HRI), only includes loudspeakers and microphones in the audible frequency range such as Softbank's NAO robots. In this paper, we consider echolocation with audible sound for mapping an environment with platforms like these, which would not require additional sensors (e.g., ultrasonic, LiDAR).

Localization of acoustic reflectors is a known problem in acoustic signal processing, which can be achieved by estimating the Time-of-Arrivals (TOAs) of reflected sounds. The echoes recorded by a microphone has a certain structure and is distinctively described in two parts: the direct-path plus early reflections and late reflections which are often described as a stochastic and dense tail. This is described by the room impulse response (RIR), which contains important information about the TOAs of the echoes, eventually enabling estimation of the acoustic reflectors' position. Based on this knowledge, several methods have been proposed to infer the shape of a room from the RIR. For instance, in [D.13, D.14], a collocated loudspeaker and microphone arrangement was used to detect first-order echoes from RIRs, which are then utilized to construct a spatial map of a room. Another attempt to construct a spatial map of a room using a mobile robot is proposed in [D.15], where the environment is probed to estimate RIRs as the robot moves in a predefined path. Based on this setup, the authors then proposed two room geometry estimation methods, one using simple trigonometry, while the other uses Bayesian filtering. Moreover, echolocation was proposed in [D.16] for robotic platforms. The authors proposed a multichannel approach to room geometry and robot position estimation, but their approach is also based on *a priori* knowledge of RIRs. Relying on RIRs is problematic in at least two ways. Firstly, it is a difficult estimation problem in itself to obtain the RIRs, and, secondly, it is non-trivial to extract TOAs from the RIR estimates. Commonly, TOAs are extracted by employing peak picking from estimated RIRs [D.17]. This is also seen in [D.11], which is an attempt to generate a spatial map of an outdoor environment using echolocation. The authors in [D.11] probes the environment and extract TOAs from the echoes as the platform moves to a new location in an environment. However, this approach to the TOA estimation from RIRs is problematic because the individual peaks relating to the true TOAs can be small and smeared as a result of dispersion, diffusion, etc. [D.18]. Our previous work in [D.19] and [D.20] addresses these issues by estimating TOAs directly from the recorded probe signal and its reflections. This is facilitated through modelling of the echoes from which a statistically optimal estimator is derived.

In this paper, a method is proposed based on the TOA estimation method in [D.19], which was only evaluated in a simulated environment that does not take into account the important practical aspects, e.g., robot movement, non-ideal hardware and propagation models, etc. In the proposed method, the TOA estimates are used in the construction of a spatial map of an indoor environment by exploiting the robot's movement. The method is tested on a proof-of-concept robotic platform that was build at Aalborg University. While we here focus on using audible sound for the mapping, the proposed method could in theory be used with sound in any frequency range as long as the source signal is known.

The remaining part of this paper is organised as follows: Section 2 introduces the signal model and problem statement. In Section 3, the TOA estimation method is described followed by the proposed mapping method in Section 3.2. Furthermore, Section 6 describes the hardware and software setup while Section 5 details the results obtained with the platform followed by a discussion and a conclusion, in Sections 5 and Section 6, respectively.

2 Signal Model and Problem Formulation

Consider a single microphone that is attached on top of a single loudspeaker. Both are placed on top of a robotic platform and is controlled by a single computer. The loudspeaker emits a known signal $s(n)$ which is recorded by a microphone. The observed signal $w(n)$ recorded by the microphone is then modelled as follows:

$$w_k(n) = (s * h_k)(n) + v(n) = x_k(n) + v(n), \tag{D.1}$$

where $h_k(n)$ is the acoustic impulse response of the room measured from the loudspeaker at robot position, r_k, to the microphone position, and $v(n)$ is the additive background noise plus ego-noise. The different positions of the robot, r_k, for $k = 1, \dots, K$, are assumed known in this work, and can in practice be estimated using, e.g., odometry or accerelometers. The background noise is assumed to be white Gaussian noise, $*$ represents the convolution operator, and $x_k(n) = (s * h_k)(n)$. By decomposing (D.1) as a sum of its direct-path component and its reflections, the signal model in (D.1) can be written as:

$$w_k(n) = \sum_{q=1}^{\infty} g_{q,k} s(n - \tau_{q,k}) + v(n) \tag{D.2}$$

where $g_{q,k}$ is the attenuation of the q^{th} reflections from the loudspeaker at the k'th robot position to the microphone, and $\tau_{q,k}$ is the TOA of the reflected sound[1].

The acoustic impulse response has a certain structure and is distinctively described in two parts: the direct-path plus early reflections and late reflections often described as a stochastic and dense tail. This means that we could rewrite the equation in (D.2) as:

$$w_k(n) = \sum_{q=1}^{R} g_{q,k} s(n - \tau_{q,k}) + v'(n), \tag{D.3}$$

where R is the number of early reflections including the direct-path component and $v'(n)$ is the collective noise term comprised of all the late reverberation (i.e., the $q > R$ components) and the additive background noise. The problem at hand is then to estimate the position of the acoustic reflectors from estimates of the TOAs, $\tau_{q,k}$, at the different robot positions, r_k.

[1] In our definition in (D.2), the direct-path component corresponds to $q = 1$

3 TOA Estimation and Mapping

We start this section by showing how the TOAs mentioned in (D.3) can be estimated based on the method in [D.19]. Then, based on these estimates, we propose the echolocation approach to mapping of the acoustic reflectors.

3.1 Nonlinear least squares TOA estimation

If N samples of the reflected signals $\mathbf{w}_k(n) = \begin{bmatrix} w_k(n) & w_k(n+1) & \cdots & w_k(n+N-1) \end{bmatrix}^T$ is taken while assuming that $s(n)$ is known and the robot position is assumed fixed within these N samples, then a nonlinear least squares (NLS) estimator can be formulated which is statistically optimal under white Gaussian noise conditions. This is expressed as follows:

$$\{\widehat{\mathbf{g}}_k, \widehat{\boldsymbol{\tau}}_k\} = \underset{\mathbf{g},\boldsymbol{\tau}}{\arg\min} \left\| \mathbf{w}_k(n) - \sum_{q=1}^{R} g_q s(n - \tau_q) \right\|^2, \tag{D.4}$$

where

$$\boldsymbol{\tau} = \begin{bmatrix} \tau_1 & \tau_2 & \cdots & \tau_q \end{bmatrix}^T, \tag{D.5}$$

$$\mathbf{g} = \begin{bmatrix} g_1 & g_2 & \cdots & g_q \end{bmatrix}^T, \tag{D.6}$$

and $\widehat{\boldsymbol{\tau}}_k$ and $\widehat{\mathbf{g}}_k$ are defined similarly.

Using Parseval's theorem, (D.4) can be transformed into the frequency domain. This helps in reducing the computational cost of the estimator and could facilitate using only selected frequencies for the estimation. This is expressed below:

$$\{\widehat{\mathbf{g}}_k, \widehat{\boldsymbol{\tau}}_k\} = \underset{\mathbf{g},\boldsymbol{\tau}}{\arg\min} \left\| \mathbf{W}_k - \sum_{q=1}^{R} g_q \mathbf{Z}(\tau_q) \odot \mathbf{S} \right\|^2, \tag{D.7}$$

$$\mathbf{Z}(\tau) = \begin{bmatrix} 1 & e^{-j\tau 2\pi \frac{1}{K}} & \cdots & e^{-j\tau 2\pi \frac{K-1}{K}} \end{bmatrix}^T, \tag{D.8}$$

where the upper case vectors, e.g., $\mathbf{W}_k$ denotes the frequency domain version of their time-domain counterparts, e.g., $\mathbf{w}_k(n)$. The matrix, $\mathbf{Z}(\tau)$, delays the source signal $\mathbf{S}$ by τ samples while $\odot$ is the element-wise product operator. In order to estimate $\widehat{\mathbf{g}}_k$ and $\widehat{\tau}_k$ parameters of the multiple reflections, R, various cyclic methods could be used like the RELAX method proposed in [D.21] that iteratively estimates the values of $\tau_{q,k}$ and $g_{q,k}$. Solving (D.7) for g_q by taking the derivative of the cost function yields:

$$\widehat{g}_q = \frac{\mathbf{W}_k^H \overline{\mathbf{Z}}(\tau_q) + \overline{\mathbf{Z}}^H(\tau_q)\mathbf{W}_k}{2\overline{\mathbf{Z}}^H(\tau_q)\overline{\mathbf{Z}}(\tau_q)} \tag{D.9}$$

Algorithm 2: Proposed method for spatial mapping.

```
Input: Trajectory $\mathcal{R} = \{(r_{x_1}, r_{y_1}), \dots, (r_{x_K}, r_{y_K})\}$;
Output: Reflector position estimates $\mathcal{P} = \{(p_{x_1}, p_{y_1}), \dots, (p_{x_K}, p_{y_K})\}$;
Initialization: $\mathcal{P} = \{\}$;
for $k = 1, \dots, r_K$ do
    Acquire direction of robot movement: $\theta_{r,k}$;
    Acquire direction of loudspeaker: $\theta_{l,k}$;
    Probe the environment with $\mathrm{s}(n)$ ;
    Record echoes: $\mathbf{w}_k$;
    Transform signals to frequency domain $\mathrm{s}(n), \mathbf{w}_k(n) \xrightarrow{\mathrm{FFT}} \mathbf{S}, \mathbf{W}_k$;
    Estimate TOA, $\widehat{\tau}_k$, using (D.12);
    Estimate acoustic reflector position $p_k$ using (D.14) and add it to $\mathcal{P}$;
end
```

where $\overline{\mathbf{Z}}(\tau_q) = \mathbf{Z}(\tau_q) \odot \mathbf{S}$ is the frequency domain probe signal delayed by τ_q samples. By inserting this back into (D.7), we get

$$\widehat{\boldsymbol{\tau}}_k = \arg\min_{\boldsymbol{\tau}} \left\| \mathbf{W}_k - \sum_{q=1}^{R} \frac{\mathbf{W}_k^H \overline{\mathbf{Z}}(\tau_q) + \overline{\mathbf{Z}}^H(\tau_q)\mathbf{W}_k}{2\overline{\mathbf{Z}}^H(\tau_q)\overline{\mathbf{Z}}(\tau_q)} \overline{\mathbf{Z}}(\tau_q) \right\|^2 . \tag{D.10}$$

In the special case, where we assume $R = 1$, e.g., if we are interested in estimating only the nearest acoustic reflector position, we get that

$$\widehat{\tau}_k = \arg\max_{\tau} \mathrm{I\!R}\{\mathbf{W}_k^H \overline{\mathbf{Z}}(\tau)\} \tag{D.11}$$

where the operator $\mathrm{I\!R}$ represents taking the real part of the signal. As seen, our derivation leads to cross-correlations.

3.2 TOA-based acoustic reflector mapping

The NLS estimator described earlier estimates τ_k for every robot position, r_k. By taking multiple observation at different time instance and position, i.e. taking the robot's movement into account, the NLS estimator can be used to generate a spatial map of an environment. Consider the platform moving in a predefined trajectory $\mathcal{R} = \{r_1, \dots, r_K\}$ with $r_k = (r_{x_k}, r_{y_k})$, such that the platform moves from r_k to r_{k+1} etc. Here, we implicitly considered mapping in two dimensions, but if additional microphones or loudspeakers are included the principle could be extended to three dimensions. For every position, r_k, the platform will probe the environment with $\mathrm{s}(n)$ and record the observed signal $\mathbf{w}_k(n)$. The probed signal and the observed signal are then converted into the frequency domain before passing it to the NLS estimator. In practice, the analysis window for the TOA could be restricted to a search interval from $\tau_{\min}$ up to $\tau_{\max}$ samples. This leads to

$$\widehat{\tau}_k = \underset{\tau \epsilon [\tau_{\min};\tau_{\max}]}{\arg\max} \; \mathbb{R}\{\mathbf{W}_k^H \overline{\mathbf{Z}}(\tau)\} \tag{D.12}$$

To estimate the position of the acoustic reflector from the estimated TOA, $\widehat{\tau}_k$, we assume the loudspeaker to be directional and place the reflector position at a distance corresponding to the estimated TOA in the direction of the loudspeaker. The direction in which the robot platform is moving, $\theta_{r,k}$, at position r_k, is related to the direction that the loudspeaker is facing, $\theta_{l,k}$, by a fixed offset angle, $\Delta\theta$, i.e.,

$$\theta_{l,k} = \theta_{r,k} + \Delta\theta. \tag{D.13}$$

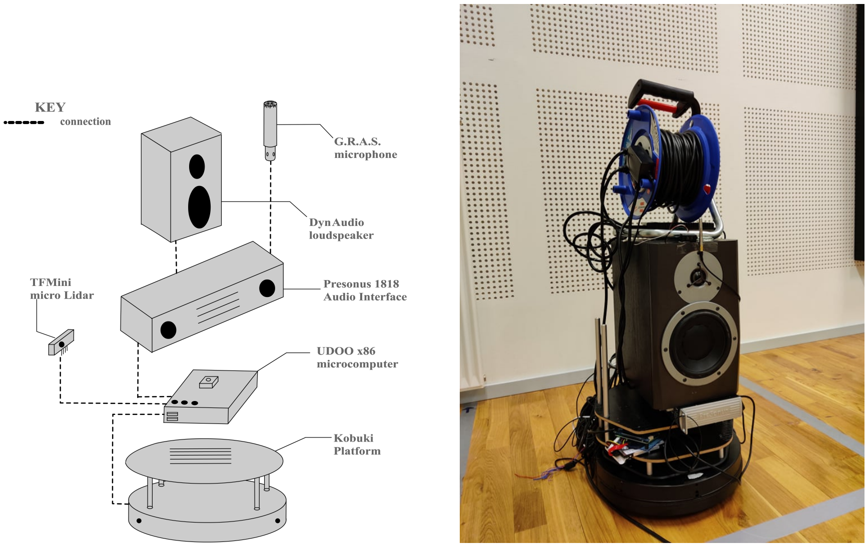
Based on the above information, the coordinates of the position of the acoustic reflector is then estimated as follows:

$$\begin{aligned} p_{x_k} &= r_{x_k} + c\frac{\widehat{\tau}_k}{2}\cos\theta_{l,k} \\ p_{y_k} &= r_{y_k} + c\frac{\widehat{\tau}_k}{2}\sin\theta_{l,k} \end{aligned} \tag{D.14}$$

where c is the speed of the sound. The procedure is then to estimate the acoustic reflector positions for each of the known robot positions, r_k, along its trajectory. The estimated acoustic reflector positions are then concatenated in the set $\mathcal{P} = \{p_1, \ldots, p_K\}$ with $p_k = (p_{x_k}, p_{y_k})$ for $k = 1, \ldots, K$. The resulting method for mapping the environment in two dimensions based on TOA estimates are outlined in Algorithm 2. However, the fixed offset $\Delta\theta$ could be avoided when employing multiple microphones but this is left for future iteration of this project.

4 Robotic Platform Overview

In this section, the hardware and software of the robotic platform is discussed. The proposed method discussed in Section 3 was implemented on embedded platform with a microcomputer running `Windows 10`. The microcomputer was developed by `UDOO`. The `UDOO x86` is a single board development platform. On the platform, `MATLAB` was used to implement the proposed method in Algorithm 4. Moreover, for multichannel audio data acquisition, `Playrec` [D.22] was used to emit and record the sounds. A `Kobuki TMR-K01-W1` platform was used as the base unit of the robot. It is a wheeled platform with on-board odometry sensor that allows for precise control and movement. The `Kobuki` platform has a built-in microcontroller that was programmed with a predefined trajectory. The loudspeaker and microphone arrangement is placed on top of the platform, which was connected to a `Presonus 1818VSL` audio interface. The Presonus interface was then subsequently connected to the `UDOO x86` microcomputer. The sampling frequency of the audio interface was set to $48,000$ Hz. Moreover, a pre-calibrated laser range sensor, `TFMini micro Lidar`, was also attached to the an `Arduino Uno` on the UDOO platform, which was used as the ground truth in our experiments. This helps in



(a) (b)

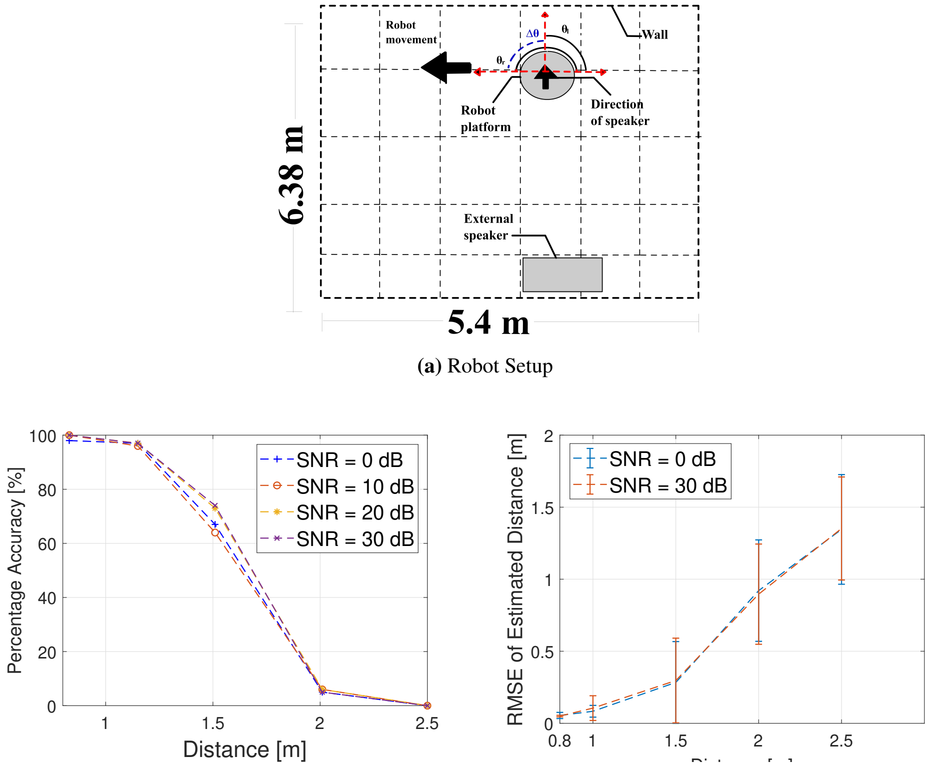
Fig. D.1: a) A proof-of-concept robotic platform b) An overview of the hardware required to design the platform used in this research

evaluating the performance of the proposed method at varying distances under different Signal-to-Noise Ratios (SNRs). The recorded data was processed by the `UDOO x86` microcomputer in real-time as the robot was moving along its trajectory. The system diagram is shown in Fig. D.1(a) and the final assembly is shown in Fig. D.1(b).

5 Experimental Setup and Results

Two experiments were performed to evaluate the performance of the proposed method. Both were tested on the proof-of-concept robotic platform discussed in the previous section. The first experiment evaluate the performance of the used NLS estimator under different SNRs and distances while placing it against one reflector as shown in Fig. D.2(a). The second experiment tested the system in a real scenario of generating a spatial map as the robot moves in a predefined trajectory. Furthermore, in the second experiment, two indoor environments were conducted in the *Sound Lab* and an office area with a glass partition, respectively. Both environments are located in the CREATE building at Aalborg University, Denmark.

For our experiments, we assumed the speed of sound to be 343 m/s and considered an analysis window starting from τ_{min} samples to τ_{max} samples corresponding to distances from 0.8 m to 2 m. The interval was selected such that the first-order reflections between distances of 0.8 m



(a) Robot Setup

(b) Performance of NLS estimator

(c) RMSE of proposed method

Fig. D.2: The performance of the NLS estimator was tested under different SNR level at varying distances

and 2 m from the microphone were captured, without capturing the direct-path component. Moreover, R was set to 1 so that only one reflection from the first-order early reflection was considered for the estimation. For both experiments, the source signal, $s(n)$, was selected as a broadband signal of length $1,500$ samples drawn from a Gaussian burst with zero padding to form a signal with length of $N = 20,000$ samples. Furthermore, a LiDAR was as placed adjacent to the microphone to measure the distance to the acoustic reflector as shown in Fig. D.1. This distance serves as the ground truth for our experiments.

5.1 Evaluation for different SNRs and distances

This experiment was performed inside the *Sound Lab* at Aalborg University, Denmark. The *Sound Lab* has dimensions of $6.38 \times 5.4 \times 4.05$ m and has sound absorbing materials embedded into the wall. The test was performed to evaluate the performance of the TOA estimation under different SNRs and distances. To simulate low SNRs, a separate loudspeaker was used as the

Table D.1: Evaluation of the proposed method against ground truth and SNRs

	SNR = 0 dB			**SNR = 10 dB**			**SNR = 20 dB**			**SNR = 30 dB**		
LiDAR [m]	μ **[m]**	σ **[m]**	**RMS error [m]**	μ **[m]**	σ **[m]**	**RMS error [m]**	μ **[m]**	σ **[m]**	**RMS error [m]**	μ **[m]**	σ **[m]**	**RMS error [m]**
0.83	0.8796	0.0214	0.0540	0.8974	0.0055	0.0497	0.8827	0.0060	0.0530	0.8827	0.0060	0.0530
1.15	1.0774	0.0408	0.0832	1.0888	0.0859	0.1051	1.0847	0.0403	0.0766	1.0977	0.0871	0.1013
1.51	1.4865	0.2837	0.2833	1.4613	0.2940	0.2966	1.4939	0.2718	0.2709	1.4523	0.2652	0.2701
2.01	1.1584	0.3521	0.9208	1.1834	0.3480	0.8962	1.1559	0.3428	0.9197	1.1763	0.3481	0.9028
2.50	1.2088	0.3806	1.3456	1.1956	0.3577	1.3521	1.1932	0.3562	1.3540	1.1863	0.3512	1.3594

interfering source playing an audio clip called *Cocktail Party*[2]. The loudspeaker was placed at the corner of the lab at a distance of 6.4 m away from the robot as shown in Fig. D.2(a). The SNR is defined as the ratio between the variance of the recorded probed signal $\mathbf{x}(n)$ against the variance of the background noise $\mathbf{v}(n)$, i.e.,

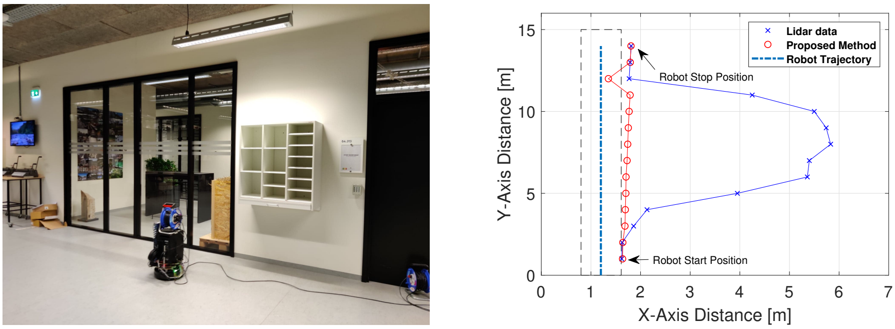
$$\text{SNR} = \frac{\sigma_x^2}{\sigma_v^2}, \tag{D.15}$$

where $\sigma_x^2 = E[\|x(n)\|^2]$ and $\sigma_v^2 = E[\|v(n)\|^2]$. Both the background noise and probe signal were recorded for 1 second. Then, 4 SNR values $[30, 20, 10, 0]$ dB were selected for this experiment. To test the performance of the proposed method in estimating the distance of the wall, the platform was placed at the following distances from one of the walls $[0.8, 1.0, 1.5, 2.0, 2.5]$ m. The results from this experiment are depicted in Fig. D.2 and Table D.1.

5.2 Mapping of an indoor environment

The second experiment was conducted to evaluate the spatial mapping method in Algorithm 4 on the robotic hardware platform. The objective here is to successfully generate a spatial map of the environment, while detecting transparent surfaces, e.g., glass partitions. Two trajectories were predefined for both environment. To construct the spatial map of the office area with the glass surface in Fig. D.3(a), the robot was predefined to move in straight line along the glass surface, while a rectangular trajectory was selected for the experiment in the *Sound Lab* shown in Fig. D.4(a). In both environments, the robot follows the predefined trajectories and stops every 1 m before coming to a momentary stop for 2 seconds. During these 2 seconds, the platform then probes the environment using a known signal, $s(n)$ and uses the proposed method to determine the location of the acoustic reflector before moving to a new location and repeating the echolocation process for each of the positions, r_k, for $k = 1, \dots, K$ as outlined in Algorithm 2. The results from the mapping experiment are shown in Fig. D.3 and D.4.

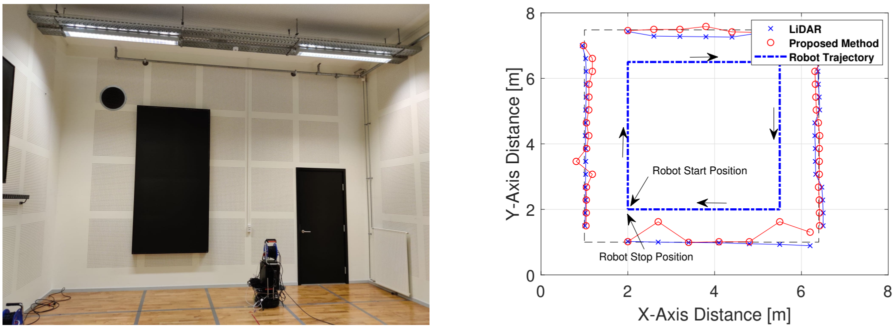
[2] Cocktail Party audio clip can be found on YouTube



(a) Layout of office with glass surfaces.

(b) 2D map of an office with glass surfaces.

Fig. D.3: Detecting of glass surface at Aalborg University.



(a) Layout of the Sound Lab.

(b) 2D map of the Sound Lab.

Fig. D.4: Generating a spatial map of the Sound Lab.

6 Discussion

The data collected from the first experiment is summarized in Fig. D.2 and Table D.1. These results show that the proposed method can accurately measure the acoustic reflector positions even when testing with sound absorbing materials and under highly noisy conditions. In general, the material of the environment is expected to be more reflective then in the considered experimental setting, which should only make the reflector position estimation easier, since the reflections will be stronger. To measure the performance of the TOA estimator, we considered its accuracy defined as the percentage of the TOA estimates that are within $\pm\epsilon$ of the ground truth (LiDAR) data, where ϵ was chosen as 10 %. The results in Fig. D.2(b) shows that the proposed method has above 60% accuracy up to a distance of 1.5 m even under low SNRs. For each SNR and distance configuration, we conducted 100 experiments. In each of these, the probe signal and the background noise were sampled randomly.

Furthermore, as seen in the Table D.1, as the distance between the platform and robot increases, the standard deviation, σ, and Root Mean Square Error (RMSE), also seen in Fig. D.2(c), increases. Additionally, the mean, μ, is close to the ground truth value for distances up to 1.5 m and for all SNRs. In conducting the second experiment, we only considered a single high SNR level of approximately 30 dB. As seen in both Fig. D.3(b) and Fig. D.4(b), a spatial map of both environment was obtained based on the sound recordings at the different robot position along the trajectory. Moreover, in Fig. D.3(b), the algorithm accurately constructed reflector position estimates even in the presence of transparent glass surfaces as opposed to the LiDAR data. Careful examination of Fig. D.3(b) and Fig. D.4(b) would reveal that the data points are not aligned to the layout of the wall. This is due to the drift in the robotic platform. One way to overcome this drift is to monitor the motor state and compensate the movement of the platform. However, this aspect is beyond the scope of this paper and will be tackled in the future iteration of this research.

7 Conclusion

The contribution of this paper is to propose a new mapping algorithm that could benefit many robotic applications by localizing acoustic reflectors from recorded microphone data. Our proposed method make use of a single loudspeaker-microphone arrangement which are commonly found in many robotic platforms used for Human Robot Interaction (HRI). Two experiments were conducted: one to test the performance of the proposed method, and another to construct a spatial map of two indoor environments. According to the data, the proposed method can robustly detect acoustic reflector at distances up to 1.5 m with above 60 % accuracy. It is also seen from the results that at higher distances, the standard deviation and the RMSE also increases which reduces the overall performance of the algorithm. Furthermore, in the second experiment, spatial maps of two environments were estimated using the proposed method on a robotic platform following a predefined trajectory. Fig. D.3 and D.4 shows that the method accurately detects even sound absorbing and transparent surfaces when compared with traditional sensing technologies. In the future iteration of this research, the proposed method will be extended to a multichannel approach, which should increase the accuracy further and enable estimation of multiple reflectors. Moreover, movement will be taken into account, in which case the reflector localizaing can be conducted while the robot is moving.

References

[D.1] D. Hahnel, R. Triebel, W. Burgard, and S. Thrun, "Map building with mobile robots in dynamic environments," *IEEE International Conference on Robotics and Automation*, vol. 2, pp. 1557–1563, 2003.

[D.2] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *IEEE robotics and automation magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[D.3] G. P. Huang, A. I. Mourikis, and S. I. Roumeliotis, "A quadratic-complexity observability-constrained unscented kalman filter for slam," *IEEE Transactions on Robotics*, vol. 29, no. 5, pp. 1226–1243, 2013.

[D.4] G. Deng, J. Li, W. Li, and H. Wang, "SLAM: Depth image information for mapping and inertial navigation system for localization," *Asia-Pacific Conference on Intelligent Robot Systems*, pp. 187–191, 2016.

[D.5] C. Hui and M. Shiwei, "Visual SLAM based on EKF filtering algorithm from omnidirectional camera," *IEEE 11th International Conference on Electronic Measurement and Instruments*, vol. 2, pp. 660–663, 2013.

[D.6] S. Chan, P. Wu, and L. Fu, "Robust 2d indoor localization through laser SLAM and visual SLAM fusion," *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1263–1268, 2018.

[D.7] H. Wei, X. Li, Y. Shi, B. You, and Y. Xu, "Multi-sensor fusion glass detection for robot navigation and mapping," *WRC Symposium on Advanced Robotics and Automation*, pp. 184–188, 2018.

[D.8] L. Marchegiani and P. Newman, "Learning to listen to your ego-(motion): Metric motion estimation from auditory signals," *Towards Autonomous Robotic Systems*, pp. 247–259, 2018.

[D.9] R. Kuc, "Echolocation with bat buzz emissions: Model and biomimetic sonar for elevation estimation," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 561–568, 2012.

[D.10] S. Kim and H. Kim, "Simple and complex obstacle detection using an overlapped ultrasonic sensor ring," *12th International Conference on Control, Automation and Systems*, pp. 2148–2152, 2012.

[D.11] I. Eliakim, Z. Cohen, G. Kosa, and Y. Yovel, "A fully autonomous terrestrial bat-like acoustic robot," *PLoS computational biology*, vol. 14, no. 9, 2018.

[D.12] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, 2017.

[D.13] F. Peng, T. Wang, and B. Chen, "Room shape reconstruction with a single mobile acoustic sensor," *IEEE Global Conference on Signal and Information Processing*, pp. 1116–1120, 2015.

[D.14] T. Wang, F. Peng, and B. Chen, "First order echo based room shape recovery using a single mobile device," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 21–25, 2016.

[D.15] M. Kreković, I. Dokmanić, and M. Vetterli, "Echoslam: Simultaneous localization and mapping with acoustic echoes," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 11–15, 2016.

[D.16] L. Nguyen, J. V. Miro, and X. Qiu, "Can a robot hear the shape and dimensions of a room?," *International Conference on Robots and Systems*, 2019.

[D.17] S. Tervo, J. Pätynen, and T. Lokki, "Acoustic reflection localization from room impulse responses," *ACTA Acustica united with Acustica*, vol. 98, no. 3, pp. 418–440, 2012.

[D.18] I. J. Kelly and F. M. Boland, "Detecting arrivals in room impulse responses with dynamic time warping," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 7, pp. 1139–1147, 2014.

[D.19] U. Saqib and J. R. Jensen, "Sound-based distance estimation for indoor navigation in the presence of ego noise," *27th European Signal Processing Conference*, pp. 1–5, 2019.

[D.20] J. R. Jensen, U. Saqib, and S. Gannot, "An EM method for multichannel toa and doa estimation of acoustic echoes," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 120–124, 2019.

[D.21] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with applications to target feature extraction," *IEEE transactions on signal processing*, vol. 44, no. 2, pp. 281–295, 1996.

[D.22] R. Humphrey, "Playrec: Multi-channel matlab audio," *URL http://www.playrec.co.uk*, 2007.

Paper E

Detecting Acoustic Reflectors using a Robot's Ego-Noise

Usama Saqib, Antoine Delaforge and Jesper Rindom Jensen

The paper has been published in the
International Conference on Acoustic, Speech and Signal Processing (ICASSP)

The layout has been revised.

Abstract

In this paper, we propose a method to estimate the proximity of an acoustic reflector, e.g., a wall, using ego-noise, i.e., the noise produced by the moving parts of a listening robot. This is achieved by estimating the times of arrival of acoustic echoes reflected from the surface. Simulated experiments show that the proposed non-intrusive approach is capable of accurately estimating the distance of a reflector up to 1 meter and outperforms a previously proposed intrusive approach under loud ego-noise conditions. The proposed method is helped by a probabilistic echo detector that estimates whether or not an acoustic reflector is within a short range of the robotic platform. This preliminary investigation paves the way towards a new kind of collision avoidance system that would purely rely on audio sensors rather than conventional proximity sensors.

1 Introduction

Within the context of robot audition, ego-noise is defined as the noise generated by the moving parts of a robotic platform, e.g., the rotors of a drone [E.1]. Ego-noise is a source of problems in many robotic applications, as it corrupts audio recordings captured by microphones, as available in many Human-Robot Interaction (HRI) systems [E.2, E.3]. For this reason, ego-noise reduction is an active area of research that plays an important role in many autonomous systems, and has enabled applications such as speech recognition for HRI [E.4] or acoustic scene analysis [E.5].

The structure of ego-noise has been investigated by several authors in the past. For instance, [E.6] investigates the spectral content of a multimotor aerial vehicle (MAV) and shows that the ego-noise is a combination of harmonic and broadband components. According to the authors, the noise spectra vary dynamically with the motor speed. Furthermore, because of the rigid mounting of the microphones with respect to the motors, the acoustic mixing can be assumed stationary. In [E.5], the authors exploit both the spectral and the spatial characteristics of ego-noise to train a dictionary that is used for ego-noise reduction.

While robotic platforms are almost always accompanied by ego-noise, only a few studies have attempted to use it constructively in the literature. For instance, in [E.7, E.8], the authors emphasize that ego-noise carries useful information about the motor system's movements and the characteristics of the environments. More specifically, in [E.7], the authors propose a forward model to predict the dynamics of the motor system of a wheeled robot. Two experiments are set to test the predictive capabilities of the model. The first experiment uses ego-noise predictions to classify velocity profiles from the auditory signals acquired by the robot, while the second experiment shows that auditory predictions can be used to detect changes in the environment, e.g., changes in the inclination of the surface where the robot is moving. Furthermore, in [E.9], the authors investigate the possibility of estimating a robot's motion from its ego-noise, i.e., "audio-based odometry". According to the authors, audio-based odometry presents advantages over laser- and visual-based odometry because it is not affected by changing light conditions.

In this paper, we follow this notion of using the ego-noise constructively rather than treating it as an interference, by proposing an estimator for acoustic reflectors based on the *time difference of echo* (TDOE). The TDOE was introduced in [E.10] as the time difference of arrival between the direct sound source signal and its first echo in a given channel. When the source is near the receiver, the distance from the source-receiver system to the nearest acoustic reflector is half the TDOE multiplied by the speed of sound. Hence, TDOE estimation is identified to distance estimation in this paper. To estimate the TDOE, we will exploit the comb-filtering effect that emerges from the direct-path component of the sound source mixing with its delayed version, due to the presence of the acoustic reflector [E.11]. Recently, a number of methods have been proposed to use early acoustic echoes constructively for audio signal processing applications, e.g., in sound source localization [E.10], sound source separation [E.12], robust speaker verification [E.13] or room geometry reconstruction [E.14, E.15, E.16]. While the latter generalizes our study, it uses clean multi-channel room impulse responses (RIR) rather than a single noisy ego-noise signal. Accurately measuring RIRs is a time-consuming and costly process that is not suitable for robotic applications.

Conventionally, proximity sensors based on ultrasounds, lasers, or infrared lights are used to detect and localize rigid surfaces in an environment. The authors of [E.17] notably used laser sensors and a Kinect to estimate the positions of acoustic reflectors in a room to inform a sound source localization method. Here, we postulate that the acoustic structure of the ego-noise naturally produced by a robot carries enough information that may help in detecting and localizing acoustic reflectors solely based on audio recordings. In our previous works [E.18, E.19, E.20], we proposed an active/intrusive approach where a loudspeaker emitting a known broadband signal was attached to a drone in order to probe the environment using times of arrival. In contrast, in this work, we propose removing the loudspeaker from the setup and develop a method solely utilizing the drone's ego-noise to detect an acoustic reflector with a single microphone. Throughout this study, we assume that the direct-path component of the ego-noise within the microphone signal is known. A number of techniques could be envisioned to estimate it, e.g., dictionary learning or model-based methods calibrated using prior measurements in an anechoic chamber, e.g., [E.5, E.18, E.19, E.20], or using close-range microphones placed next to the ego-noise sources as references. This aspect is beyond the scope of this paper and is left for future iterations of this research. Here, we focus on the specific question of whether an uncontrolled ego-noise signal, as opposed to a controlled emitted broadband signal, is sufficient to probe an environment for nearby reflectors.

First, we develop a statistically optimal TDOE estimator to solve this problem in the least-square sense. Then, we introduce a probabilistic echo-detector that helps our estimator distinguishing an acoustic reflector from empty space. Simulated experiments show that the proposed non-intrusive method is capable of accurately estimating distances of 1 meters or less and outperforms our previously proposed intrusive approach under loud ego-noise conditions. To the best of our knowledge, this is the first time that ego-noise echoes stemming from acoustic reflectors are used in a constructive way in the context of robot audition, and in particular in drone audition.

The remainder of this paper is organized as follows. Section 2 formulates the signal model and the problem. Section 3 describes the proposed TDOE estimator based on our model. Section 4 evaluates the performance and robustness of the proposed solution. Finally, Section 5 concludes and provides directions for future work.

2 Problem formulation

Consider a setup with a single microphone that records both the ego-noise generated by the rotors of a drone, $x[n]$, and a background noise from the environment. The signal model is then:

$$y[n] = (h * s)[n] + v[n] = x[n] + v[n], \tag{E.1}$$

where h is the impulse response from the ego-noise source to the microphone. The source signal $s[n]$ is generated by the rotors of the drone while $v[n]$ is the white Gaussian background noise. The signal $x[n]$ is the ego-noise of the drone that we will use to facilitate TDOE estimation. We now proceed to decompose the observed ego-noise signal, $y[n]$, as the sum of individual reflections from the source signal:

$$y[n] = \sum_{q=1}^{\infty} g_q s[n - \tau_q] + v[n], \tag{E.2}$$

where $q = 1$ indexes the direct path component and $q > 1$ the acoustic reflections, τ_q represents the time of arrival of the q-th direct or reflected source signal, while g_q is the corresponding gain or attenuation due to the inverse square law of sound propagation and to the sound-absorbent material at the acoustic reflector, assuming frequency-independence in this work. Acoustic impulse responses have a certain structure, which can be classified into two parts: an early part and a late part. The early part is sparse in time and contains the direct-path as well as the early reflections while the late part is characterized by a stochastic, dense and decaying tail of late reflections [E.21]. This suggests to divide the signal model as follows:

$$y[n] = \sum_{q=1}^{R} g_q s[n - \tau_q] + v'[n], \tag{E.3}$$

where R is the number of considered early reflections and $v'[n]$ is composed of late reflections and background noise. Furthermore, we can rewrite (E.3) in a compact expression by separating $x[n]$ into the direct-path and early reflection components as:

$$y[n] = x_d[n] + x_r[n] + v'[n], \tag{E.4}$$

where $x_d[n] = g_1 s[n-\tau_1]$ is the direct path component, and $x_r[n] = \sum_{q=2}^{R} g_q s[n-\tau_q]$ contains all the early reflections. While the direct path component $x_d[n]$ is always present, the reflection component $x_r[n]$ vanishes from the microphone signal if the robotic platform is not near any acoustic reflector. In drone audition applications, the microphones are often set in a fixed location with respect to the rotors. This fact could be used to estimate the direct path component using additional close-range microphones placed next to the rotors. Alternatively, a direct path estimation method calibrated in anechoic conditions could be derived. This is out of the scope of this paper, and we assume here that the direct path component $x_d[n]$ is known. Moreover, we are only interested in detecting one acoustic reflector, e.g., the closest one for obstacle avoidance, so we set $R = 2$ to estimate the first reflection. If we vectorize (E.4) and express it in terms of the gains and delays, we can approximate the signal model as shown:

$$\begin{aligned} \mathbf{y}[n] &\approx g_d \mathbf{D}_{\tau_d} s[n] + g_r \mathbf{D}_{\tau_r} s[n] + \mathbf{v}'[n], \qquad \text{(E.5)} \\ \mathbf{y}[n] &= \begin{bmatrix} y[n] & y[n+1] & \cdots & y[n+N-1] \end{bmatrix}^T, \end{aligned}$$

where $\mathbf{D}_\tau$ is a cyclic shift register that delays the unknown rotor signal $s[n]$ by τ samples and g is the gain of the signal. Note that $\mathbf{D}_\tau$ is an identity matrix whose columns are cyclically shifted to the right by τ, which approximates a true delay operator. Similarly, we can decompose (E.5) into vectorized direct-path and reflection components, $\mathbf{x}_d[n]$ and $\mathbf{x}_r[n]$. Since $\mathbf{x}_r[n]$ is a delayed version of the direct-path component, (E.4) can be expressed as shown:

$$\begin{aligned} \mathbf{y}[n] &= \mathbf{x}_d[n] + \frac{g_r}{g_d} \mathbf{D}_{\Delta\tau} \mathbf{x}_d[n] + \mathbf{v}'[n], \qquad \text{(E.6)} \\ &= (\mathbf{I} + \alpha \mathbf{D}_{\Delta\tau})\, \mathbf{x}_d[n] + \mathbf{v}'[n], \qquad \text{(E.7)} \end{aligned}$$

where $\Delta\tau$ is the TDOE of the observed signal, such that $\Delta\tau = \tau_r - \tau_d$ and $\alpha = \frac{g_r}{g_d}$, while $\mathbf{I}$ is the identity matrix. The task at hand is then to estimate $\Delta\widehat{\tau}$ and $\widehat{\alpha}$ in order to detect the presence of an acoustic reflector and possibly infer its distance to the nearest acoustic reflector.

3 TDOE estimation based on Least-Squares Fit

Let $\mathbf{y} \in \mathbb{R}^N$ contain N consecutive samples of the observed signal at a given time. Assume that the corresponding direct-path component $\mathbf{x}_d$ is known. Then, we can estimate $\Delta\tau$ and α from (E.7) in the least-squares sense by solving the following problem:

$$\begin{aligned} \{\Delta\widehat{\tau}, \widehat{\alpha}\} &= \underset{\Delta\tau, \alpha}{\arg\min} \, \|\mathbf{y} - (\mathbf{I} + \alpha \mathbf{D}_{\Delta\tau})\, \mathbf{x}_d\|^2 \qquad \text{(E.8)} \\ &= \underset{\Delta\tau, \alpha}{\arg\min} \, J(\Delta\tau, \alpha). \qquad \text{(E.9)} \end{aligned}$$

Note that, assuming the background noise is white and Gaussian, the resulting estimators will also be maximum likelihood estimators. The cost function in (E.9) can be rewritten as follows:

$$\begin{aligned} J(\Delta\tau, \alpha) &= \|\mathbf{y} - (\mathbf{I} + \alpha\mathbf{D}_{\Delta\tau})\,\mathbf{x}_d\|^2 \\ &= \|\mathbf{y}\|^2 - 2\mathbf{y}^T\,(\mathbf{I} + \alpha\mathbf{D}_{\Delta\tau})\,\mathbf{x}_d \\ &\quad + \mathbf{x}_d^T\,(\mathbf{I} + \alpha\mathbf{D}_{\Delta\tau})^T\,(\mathbf{I} + \alpha\mathbf{D}_{\Delta\tau})\,\mathbf{x}_d. \end{aligned} \tag{E.10}$$

By zeroing the derivative of (E.10) with respect to α we get:

$$\begin{aligned} \frac{\delta J}{\delta\alpha} &= -\mathbf{y}^T\mathbf{D}_{\Delta\tau}\mathbf{x}_d - \mathbf{x}_d^T\mathbf{D}_{\Delta\tau}^T\mathbf{y} \\ &\quad + \mathbf{x}_d^T\mathbf{D}_{\Delta\tau}^T\mathbf{x}_d + \mathbf{x}_d^T\mathbf{D}_{\Delta\tau}\mathbf{x}_d + 2\alpha\mathbf{x}_d^T\mathbf{D}_{\Delta\tau}^T\mathbf{D}_{\Delta\tau}\mathbf{x}_d = 0. \end{aligned} \tag{E.11}$$

By observing that $\mathbf{D}_{\Delta\tau}^T\mathbf{D}_{\Delta\tau} = \mathbf{I}$, this becomes:

$$-\left(\mathbf{y} - \mathbf{x}_d\right)^T\mathbf{D}_{\Delta\tau}\mathbf{x}_d + \mathbf{x}_d^T\mathbf{D}_{\Delta\tau}^T\left(\mathbf{y} - \mathbf{x}_d\right) + 2\alpha\|\mathbf{x}_d\|^2 = 0 \tag{E.12}$$

Hence,

$$\widehat{\alpha}(\Delta\tau) = \frac{\left(\mathbf{y} - \mathbf{x}_d\right)^T\mathbf{D}_{\Delta\tau}\mathbf{x}_d}{\|\mathbf{x}_d\|^2}. \tag{E.13}$$

We see here that the estimated gain ratio α has an interesting interpretation. It can be viewed as a cross-correlation between the known direct path $\mathbf{x}_d$ and the observed signal without direct path $\mathbf{y} - \mathbf{x}_d$. Now, by inserting back (E.13) into (E.8), removing constant terms and simplifying, we obtain:

$$\Delta\widehat{\tau} = \arg\max_{\Delta\tau}\ \widehat{\alpha}(\Delta\tau)^2. \tag{E.14}$$

This expression can be maximized over a finite, predefined set of candidate delays $\Delta\tau$ to obtain the desired least-square estimate $\Delta\widehat{\tau}$.

4 Echo detector

Solving (E.8) will always give a TDOE estimate no matter where the drone is positioned in 3D space. However, we require a mechanism to distinguish whether this estimate belongs to an acoustic reflector or is an artifact stemming from the background noise. This detection is thus vital if using the TDOE estimator for, e.g., collision avoidance as it will help remove spurious estimates. Therefore, we resolve this problem by introducing an echo detector. The decision about whether an observation contains an acoustic reflection can be formulated as a detection problem [E.22]. Let us consider the following two hypotheses:

$$\mathcal{H}_0 : \mathbf{y}[n] = \mathbf{x}_d[n] + \mathbf{v}[n], \tag{E.15}$$

$$\mathcal{H}_1 : \mathbf{y}[n] = \mathbf{x}_r[n] + \mathbf{x}_d[n] + \mathbf{v}[n], \tag{E.16}$$

where $\mathcal{H}_0$ is the null hypothesis and refers to a situation when the observation only includes the direct-path component $\mathbf{x}_d[n]$ and white Gaussian background noise $\mathbf{v}(n)$, with variance σ^2 and mean 0, i.e., $\mathcal{N}(0, \sigma^2)$. In contrast, $\mathcal{H}_1$ refers to the situation when a reflected signal $\mathbf{x}_r[n]$ from an acoustic reflector is observed, in addition to $\mathbf{v}[n]$ and $\mathbf{x}_d[n]$. The observation interval is $n \in [0, N-1]$ and the generalized likelihood ratio test (GLRT) is given as:

$$\mathcal{L}(n) = \frac{p(\mathbf{y}; \mathbf{x}_r[n], \mathcal{H}_1)}{p(\mathbf{y}; \mathcal{H}_0)} > \gamma. \tag{E.17}$$

In other words, in order to detect if an observation n belongs to $\mathcal{H}_1$, its GLRT should be greater than γ. The probability density functions (PDFs) for the two hypotheses are given as shown:

$$\begin{aligned} p(\mathbf{y}; \mathbf{x}_r[n], \mathcal{H}_1) &= \frac{1}{(2\pi\sigma_v^2)^{N/2}} \exp\left(\frac{-\|\mathbf{y}[n] - \mathbf{x}_r[n] - \mathbf{x}_d[n]\|^2}{2\sigma_v^2}\right), \\ p(\mathbf{y}; \mathcal{H}_0) &= \frac{1}{(2\pi\sigma_v^2)^{N/2}} \exp\left(\frac{-\|\mathbf{y}[n] - \mathbf{x}_d[n]\|^2}{2\sigma_v^2}\right), \end{aligned} \tag{E.18}$$

where σ_v^2 is the variance of the background noise, $v[n]$.

Note that in order to compute $\mathcal{L}(n)$, an estimate $\widehat{\mathbf{x}}_r[n]$ of the unknown reflected component $\mathbf{x}_r[n]$ is needed. One way of obtaining such estimate would be to use the estimates of section 3, i.e., $\widehat{\mathbf{x}}_r[n] = \widehat{\alpha}\mathbf{D}_{\Delta\widehat{\tau}}\mathbf{x}_d[n]$. However, this approach would strongly rely on the hypothesis that only one reflection exists in the observation. Instead, we propose to use a more straight-forward estimate for $\mathbf{x}_r[n]$ which is agnostic to the reflection model. Under the hypothesis $\mathcal{H}_1$, directly maximizing the observed-data likelihood with respect to $\mathbf{x}_r$ by zeroing the derivative of the logarithm of (E.18) yields:

$$\frac{d \ln p(\mathbf{y}; \mathcal{H}_1)}{d\mathbf{x}_r} = -\left(\mathbf{x}_r[n] - \mathbf{y}[n] + \mathbf{x}_d[n]\right) = 0, \tag{E.19}$$

that is, the reflected signal is found by subtracting the direct-path component $\mathbf{x}_d[n]$ from the observation $\mathbf{y}[n]$, as shown:

$$\widehat{\mathbf{x}}_r[n] = \mathbf{y}[n] - \mathbf{x}_d[n]. \tag{E.20}$$

By inserting (E.20) into (E.18) and this back into (E.17) we get:

$$\begin{aligned} \ln \mathcal{L}(\mathbf{x}) &= \frac{\ln p(\mathbf{y}; \mathbf{x}_r[n], \mathcal{H}_1)}{\ln p(\mathbf{y}; \mathcal{H}_0)} && \text{(E.21)} \\ &= (\mathbf{y}[n] - \mathbf{x}_d[n])^T(\mathbf{y}[n] - \mathbf{x}_d[n]) > 2\sigma_v^2 \ln \gamma. && \text{(E.22)} \end{aligned}$$

Hence, the criterion to detect an acoustic reflector is:

$$T(\mathbf{y}) = \|\mathbf{y}[n] - \mathbf{x}_d[n]\|^2 \underset{H_0}{\overset{H_1}{\gtrless}} 2\sigma_v^2 \ln \gamma. \tag{E.23}$$

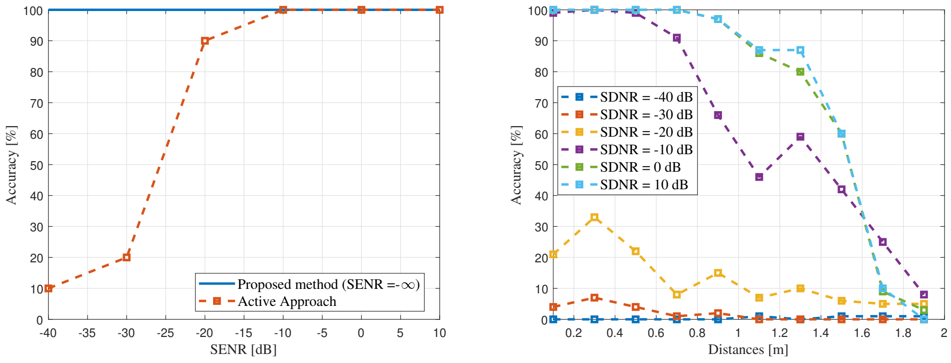
In other words, for a reflector to be detected, the power of the reflected signal should be greater than a threshold that depends on the variance of the background noise σ_v^2. Note that this criterion will change under different noise conditions.

5 Experiments

Two experiments were conducted within a simulated room of dimensions $8 \times 6 \times 5$ m. For these experiments, we simulated the drone ego-noise as a point source. The distance between the source and the microphone is 0.2 m. A signal generator [E.23] was used to generate the response of a moving sound source and a receiver. The signal generator convolves the sound source, i.e., the rotor noise, with a time varying RIR. The RIR is generated using the image-source method, first proposed in [E.24]. The reverberation time was set to $T_{60} = 0.4$ s, the FFT length was set to $2,048$ samples while the speed of sound was set to 343 m/s. For the ego-noise sound source, we used the signal `allMotors_70.wav` from the DREGON dataset [E.1], which was recorded from a drone whose four rotors were set to a fixed speed of 70 rotations per second. A diffuse cylindrical background noise was generated from this signal using the method described in [E.25]. The background noise has two parts, the first part is the ego-noise of the drone which contributes to late reverberation [E.26] while the latter part is the white Gaussian noise set at 40 dB. The signal was then down-sampled from 44.1 kHz to $5,512.5$ Hz in order to decrease the computational cost of simulating the moving source and the receiver.

5.1 Comparison and evaluation of the proposed estimator

In the first experiment, we compare the proposed method against our previously published approach [E.18]. The latter also estimates the distance of a nearby acoustic reflector based on its TDOE, but it utilizes an embedded loudspeaker to probe the environment with a known white-noise signal $s[n]$, i.e., an *intrusive* approach. Hence, in that case, the ego-noise is a disturbance in the acoustic reflector estimation. The purpose of this experiment is to test the limits of the intrusive approach under varying signal-to-ego-noise ratios (SENR). The SENR is computed in decibel (dB) as the variance of the probe signal, σ^2_{probe}, divided by the variance of the ego-noise σ^2_{ego}. The evaluation metric used is the accuracy, defined as the percentage of TDOEs that are within $\pm 10\%$ of the true TDOE, calculated from the actual distance of the robotic platform to the acoustic reflector. In this experiment, the distance to the acoustic reflector was fixed to 0.5 m. As seen in Fig. E.1(a), the intrusive approach gradually fails for SENR values below -10 dB. For comparison, the figure also shows the accuracy obtained with the proposed approach, from the same distance, without using any probe signal (SENR=-∞) and under an observed-signal to diffuse-background-noise ratio (SDNR) of 40 dB. The SDNR is calculated as the variance of the observed signal including the direct-path and reflections, σ_x^2, divided by the variance of the diffuse background environment noise σ_v^2. As can be seen, the accuracy of the proposed approach is 100% in that case. We then evaluate the performance of the proposed approach against varying distances and under different SDNR values. In this experiment, the robotic



(a) **(b)**

Fig. E.1: a)Comparison between the proposed TDOE estimator in (E.14) and the intrusive method in [E.18] against varying SENRs values, when the robotic platform is 0.5 m from an acoustic reflector (SDNR = 40dB) b) Evaluation of the proposed method in (E.14) against varying distances and SDNR values when the platform is near one acoustic reflectors

platform moves at a distance of $[0.1 : 0.2 : 2]$ m from the acoustic reflector, while the SDNR of the environment changes for every simulation within an interval of $[-40 : 10 : 10]$ dB. We conducted 100 Monte-Carlo trials for each (distance, SDNR) combination to obtain the results in Fig. E.1(b). As can be seen, the proposed TDOE estimator (E.14) provides high accuracy and offers robustness from 0 dB and above but starts to fail under low SDNR values. Moreover, the proposed method can robustly estimate the distance of an acoustic reflector up to around 1 m.

5.2 Application Example

In Fig.E.2, we simulate a scenario where we move the drone from one acoustic reflector to another, i.e., from $r_{s_x} = 0.1$ m to $r_{s_x} = 7.9$ m, in order to test the TDOE estimator in (E.14) and the echo-detector in (E.23) on a larger set of distances. The value of $\ln(\gamma)$ was empirically set to $2,500$. If the echo-detector is close to an acoustic reflector, then the detector assigns a value of 1. Otherwise, a value of 0 is assigned to indicate empty space. The experiment was conducted under an SDNR of 10 dB. In Fig. E.2(a), we see that the gain estimate using (E.13) becomes higher as the drone gets closer to an acoustic reflector. Furthermore, as seen from Fig. E.2(b), as the drone approaches an acoustic reflector, we are able to correctly estimate the TDOE up to a distance of around 1 m. The red line on Fig. E.2(b) indicates the true distance and the corresponding TDOE (ground truth) to the acoustic reflector of the drone. However, at distances larger than 1 m, the estimator fails to estimate the TDOE. This is illustrated by the fluctuations in the center of the figure, followed by a linear decrease of TDOE estimates as

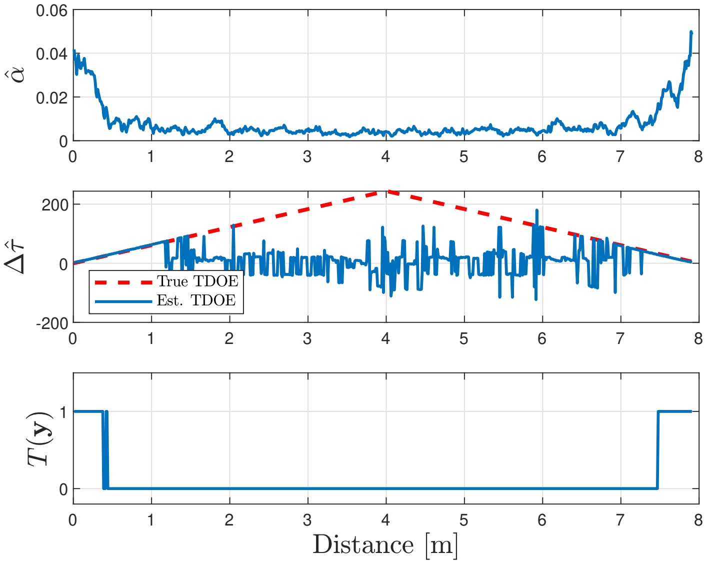


Fig. E.2: A moving microphone-rotor setup was tested within a simulated environment to represent a moving drone platform from one acoustic surface to another. The performance of the a) gain estimator b) TDOE estimator and c) GLRT detector is shown in the figure.

the drone approaches the other wall. Finally, results from the echo-detector using the threshold value $T(\mathbf{y})$ in (E.23) are shown in Fig. E.2(c). From the figure, it is seen that the power of the reflected sound is higher than the threshold for reflectors that are closer than 0.5 m, allowing the method to detect them.

6 Conclusion and Future work

In this paper, we proposed a TDOE estimator and an echo detector to estimate the proximity of an acoustic reflector. These make use of the natural ego-noise of a robotic platform equipped with a microphone. The proposed method could lead to the development of new sound-based collision avoidance systems for, e.g., drones. With such a system, the platform would not need to utilize proximity sensors, e.g., infrared lights or ultrasounds, to prevent collisions into walls. According to preliminary simulated experiments, the proposed method is able to estimate a distance of up to 1 m and can distinguish an acoustic reflector from empty space based on the energy of the reflected signal which is compared to a predefined threshold. In future iterations of this project, we aim to investigate the estimation of the direct-path component, $\mathbf{x}_d[n]$, which was assumed to be known throughout this paper. This is a very challenging problem on its own, which requires further investigation. It could be addressed using the known microphone placement together with close-range microphones placed next to the ego-noise sources, or using direct-path ego-noise models trained and calibrated in anechoic conditions.

References

[E.1] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "DREGON: Dataset and methods for uav-embedded sound source localization," *Proc. IEEE Int. Conf. Intell., Robot, Automation.*, pp. 1–8, 2018.

[E.2] K. Nakadai, T. Lourens, H.G. Okuno, and H. Kitano, "Active audition for humanoid," *American Asso. on Artificial Intell.*, pp. 832–839, 2000.

[E.3] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 5610–5614, 2015.

[E.4] N. Wake, M. Fukumoto, H. Takahashi, and K. Ikeuchi, "Enhancing listening capability of humanoid robot by reduction of stationary ego-noise," *IEEE Transactions on Electrical and Electronic Engineering*, vol. 14, no. 12, pp. 1815–1822, 2019.

[E.5] A. Schmidt, H. W. Löllmann, and W. Kellermann, "A novel ego-noise suppression algorithm for acoustic signal enhancement in autonomous systems," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 6583–6587, 2018.

[E.6] L. Wang and A. Cavallaro, "Ear in the sky: Ego-noise reduction for auditory micro aerial vehicles," *IEEE Int. Conf. on Adv. Video and Signal Based Surveillance*, pp. 152–158, 2016.

[E.7] A. Pico, G. Schillaci, V. V. Hafner, and B. Lara, "How do i sound like? forward models for robot ego-noise prediction," *Joint IEEE Int. Conf. on Development and Learning and Epigenetic Robotics*, pp. 246–251, 2016.

[E.8] A. Pico, G. Schillaci, V. V. Hafner, and B. Lara, "On robots imitating movements through motor noise prediction," *Joint IEEE Int. Conf. on Development and Learning and Epigenetic Robotics*, pp. 318–323, 2017.

[E.9] L. Marchegiani and P. Newman, "Learning to listen to your ego-(motion): Metric motion estimation from auditory signals," *Towards Autonomous Robotic Systems*, pp. 247–259, 2018.

[E.10] D. D. Carlo, A. Deleforge, and N. Bertin, "MIRAGE: 2D source localization using microphone pair augmentation with echoes," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 775–779, 2019.

[E.11] M.G. Christensen, *Introduction to Audio Processing*, Springer International Publishing, 2019.

[E.12] R. Scheibler, D. D. Carlo, A. Deleforge, and I. Dokmanic, “Separake: Source separation with a little help from echoes,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 6897–6901, 2018.

[E.13] Khamis A Al-Karawi and Duraid Y Mohammed, “Early reflection detection using autocorrelation to improve robustness of speaker verification in reverberant conditions,” *Int. J. of Speech Technology*, vol. 22, no. 4, pp. 1077–1084, 2019.

[E.14] Fabio Antonacci, Jason Filos, Mark RP Thomas, Emanuël AP Habets, Augusto Sarti, Patrick A Naylor, and Stefano Tubaro, “Inference of room geometry from acoustic impulse responses,” *J. Audio, Speech, Language Process.*, vol. 20, no. 10, pp. 2683–2695, 2012.

[E.15] I. Dokmanić, R. Parhizkar, A. Walther, Y. Lu, and M. Vetterli, “Acoustic echoes reveal room shape,” *Proc. of the National Academy of Sciences*, vol. 110, no. 30, pp. 12186–12191, 2013.

[E.16] M. Crocco, A. Trucco, V. Murino, and A. D. Bue, “Towards fully uncalibrated room reconstruction with sound,” *Proc. European Signal Processing Conf.*, pp. 910–914, 2014.

[E.17] I. An, M. Son, D. Manocha, and S. Yoon, “Reflection-aware sound source localization,” *Proc. IEEE Int. Conf. Robotics, Automation.*, pp. 66–73, 2018.

[E.18] U. Saqib and J. R. Jensen, “Sound-based distance estimation for indoor navigation in the presence of ego noise,” *Proc. European Signal Processing Conf.*, 2019.

[E.19] J. R. Jensen, U. Saqib, and S. Gannot, “An EM method for multichannel TOA and DOA estimation of acoustic echoes,” *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, pp. 120–124, Oct. 2019.

[E.20] U. Saqib, S. Gannot, and J.R. Jensen, “Estimation of acoustic echoes using expectation-maximization methods,” *EURASIP J. on Audio, Speech, and Music Process.*, vol. 2020, no. 1, pp. 1–15, 2020.

[E.21] H. Kuttruff, *Room acoustics*, Crc Press, 2016.

[E.22] S. M. Kay, *Fundamentals of statistical signal processing*, vol. 2, Prentice Hall PTR, 1993.

[E.23] E. A. P. Habets, “Signal generator,” 2017.

[E.24] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[E.25] E. A. P. Habets, I. Cohen, and S. Gannot, “Generating nonstationary multisensor signals under a spatial coherence constraint,” *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, 2008.

[E.26] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, “Evaluation and comparison of late reverberation power spectral density estimators,” *J. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1056–1071, 2018.

Paper F

Robust Acoustic Reflector Localization for Robots

Usama Saqib, Mads Græsbøll Christensen and Jesper Rindom Jensen

This paper is submitted to the
EURASIP Journal on Audio, Speech, and Music Processing (JASM)

The layout has been revised.

Abstract

In robotics, echolocation has been used to detect acoustic reflectors, e.g., walls, as it aids robotic platform to navigate in darkness and also help detect transparent surfaces. However, the transfer function or response of an acoustic system, e.g., loudspeakers/emitters, contributes to non-ideal behaviour within the acoustic systems that can contributes to a phase lag due to propagation delay. This non-ideal response can hinder the performance of a time-of-arrival (TOA) estimator intended for acoustic reflector localization especially when estimation of multiple reflections are require. In this paper, we therefore propose a robust Expectation-Maximization (EM) algorithm that takes into account the response of acoustic systems to enhance the TOA estimation accuracy when estimating multiple reflections when the robot is placed at a corner of a room. A non-ideal transfer function is built with two parameters, which are estimated recursively within the estimator. To test the proposed method, a hardware proof-of-concept setup was built with two different designs. The experimental results show that the proposed method could detect an acoustic reflector up to a distance of 1.6 *m with* 60% *accuracy under the signal-to-noise ratio (SNR) of* 0 *dB.*

1 Introduction

Within the context of robot audition, the use of echolocation for acoustic reflector localization and estimation has been proposed by various researchers in the past [F.1, F.2, F.3]. Within this domain, researchers are utilizing acoustic signal processing techniques and propose combining echolocation with state-of-the-art technologies, e.g., laser- and camera-based technologies to aid a robot to construct a spatial map of an indoor environment. This can be accomplished by a collocated microphone-loudspeaker combination. One major disadvantage of camera and laser-based technologies is that it cannot work in complete darkness and it cannot detect transparent surfaces that are typically found in an office environment, This makes accurate construction of spatial map of an environment a difficult process.

The process involved in the aforementioned echolocation techniques is to probe the environment with a known sound, so that the reflected signal acquired by a microphone can be processed to estimate the time of arrival (TOA) of the acoustic echo that aid a robot to estimate the distance between the acoustic reflector. Traditionally, TOA information is extracted from room impulse response (RIR) estimates, Fig. F.1 which is normally done using a peak-picking approach [F.2, F.3, F.4, F.5, F.6]. This model is broadly divided into two distinct parts: the direct-path including early reflections and late reflections which are comprised by a stochastic dense tail [F.7]. The direct-path component is the shortest distance a sound can take ,i.e., it provides information about the distance between the transmitter and receiver while early reflections helps in inferring the distance of the closest acoustic reflector [F.8, F.3, F.2]. While, TOA estimation enables a robot to determine the distance of an acoustic reflector, direction-of-arrival (DOA) of an acoustic source is required to determine the location of an acoustic source. This is

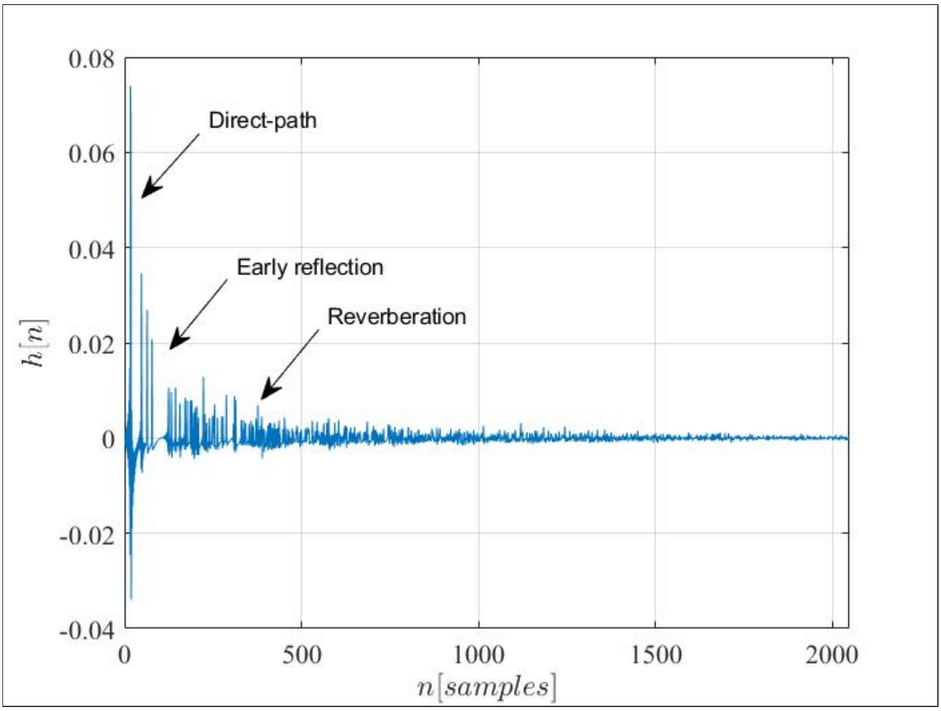


Fig. F.1: Transfer function of the room between source and microphone, RIR. The direct-path contains the highest energy followed by the early reflection and reverberation which is represented by a dense tail

done by incorporating multiple receivers attached to a robot [F.9, F.10, F.11].

Recent advancement in machine learning techniques has also enabled robotic platform to incorporate echolocation for terrain classification and detecting echoes from noisy data. For example, in [F.12], the author proposed training using advanced signal filtering and machine learning techniques which could be used to accurately classify terrain types for a small mobile robot. One potential for such method is to help robot navigation, i.e., detecting road from other surfaces. Moreover, echolocation is used to map a spatial map of an indoor environment. For example, in [F.13], the authors propose training a neural network to predict depth maps and gray-scale images from sound alone. The work presented in [F.13] was later improved by [F.14] by improving the neural network and reducing the computation time to run the model. The contribution of the paper was a full 360^o $3D$ depth reconstruction with 4 microphones and a lidar-based SLAM for training a model. One notable difference between model-based approach and data-driven approaches is the availability of large data-sets required to train a neural network. Comparatively, model-based approach finds the feature of interest directly from the signal model.

While, ultrasonic sensors are popular within robotics to detect obstacles, these require specialized hardware to transmit/receive acoustic echoes and could potentially increase the overall cost of a robotic platform. However, most robots intended for human-robot interaction (HRI) consist of a collocated microphone-loudspeaker setup, e.g, Softbank's NAO robot. Therefore, we propose an algorithm that utilize the existing loudspeaker-microphone setup to estimate the distance of an acoustic reflector [F.15, F.16]. However, the loudspeaker's and microphone's response was not taken into account when deriving a TOA estimator. The response of the acoustic systems are required to enable a robot to effectively detect and estimate the location of additional echoes that originates from other walls. In practice, loudspeakers/microphones are non-ideal. One reason is that the transfer function of the acoustical systems, e.g., loudspeakers/microphones, contributes to a phase lag due to propagation delay [F.17]. This can lead to model-mismatch and have a detrimental impact on the performance of the TOA estimation method proposed in [F.8, F.18] and may hinder estimation of multiple acoustic echoes.

Traditionally, estimating the transfer function of the loudspeaker is usually done using a loudspeaker enclose microphone (LEM) setup which involves placing the setup within an anechoic environment. However, in [F.19], the researchers proposed a method to measure the transfer-function of the loudspeaker within an echoic environment. This is done by utilizing two loudspeakers, one of them calibrated and its transfer-function already estimated within an anechoic chamber. The loudspeaker is placed in a fixed location within the environment. The process involves transmitting a white noise signal through the calibrated loudspeaker to measure its impulse response (IR) and later replacing the loudspeaker with the uncalibrated loudspeaker and repeating the IR measurement. The transfer-function of the uncalibrated loudspeaker is estimated using least-squares. Furthermore, TOA estimation can also be influenced by the materials that acoustic reflectors are composed off, e.g., concrete, glass, and cardboards. This is because, some materials absorb certain sound frequencies that could lead to non-ideal characteristics of the observed signals [F.20]. The aforementioned method requires access to anechoic chamber which is a time consuming process, hence, there is a need to estimate the response of the acoustic system directly from the model.

In this paper, we therefore extend the model-based method originally proposed in [F.21] and later used in our previous work [F.8] to accommodate the non-ideal transfer function of an acoustic system, i.e., the loudspeaker, the microphone and the reflecting materials. We take a model-based approach to TOA estimation where the model of the early reflections is used to derive a statistically optimal estimator. More specifically, we include an unknown filter to model the uncertainties of the acoustic system which may alleviate the need to estimate loudspeaker IR measurement suggested in [F.19]. Moreover, to test the proposed method, a proof-of-concept setup is built to conduct experiments using real data.[1]

The remaining part of this paper is organized as follows: Section 2 introduces the problem formulation, Section 3 proposed the TOA estimation method based on EM. Finally, the experimental results followed by discussion and conclusion can be found in Section 4, 5 and 6, respectively.

[1] The dataset and code for this work can be found here: https://doi.org/10.5281/zenodo.5082224

2 Problem formulation

Consider the scenario where a loudspeaker is emitting a known probe signal, which is then propagating an acoustic environment, and recorded by a microphone. This can be mathematically modeled as

$$\begin{aligned} y(n) &= h(n) * s(n) + w(n) \\ &= x(n) + w(n), \end{aligned} \tag{F.1}$$

where $h(n)$ is the acoustic impulse response from the loudspeaker to the microphone, $s(n)$ is the known probe signal, and $w(n)$ is additive background noise while $x(n) = h(n) * s(n)$. The acoustic impulse response can be further modelled by decomposing the reverberation into early and late reverberation components. The early reflections are modelled as time-delayed and filtered versions of the known probe signal, where the filter represents the responses of the loudspeaker, microphone, and acoustic reflectors. Mathematically, we formulate this as

$$y(n) = \sum_{r=1}^{R} g_r * s(n - \tau_r) + v(n), \tag{F.2}$$

where R is the number of early reflections, g_r is the filter pertaining to the rth reflection, τ_r is the delay of the r'th reflection, and $v(n)$ is a noise term embracing both the additive background noise and the late reflections. In the special case where g_r is a Dirac function for all $r = 1, \ldots, R$, we get the ideal model used in [F.8], which does not account for the non-ideal hardware responses that are inevitable in real scenarios. We then assume stationarity and that we have N observations following this model, i.e.,

$$\mathbf{y}(n) = \sum_{r=1}^{R} \mathbf{G}_r \mathbf{s}(n - \tau_r) + \mathbf{v}(n), \tag{F.3}$$

$$= \sum_{r=1}^{R} \mathbf{S}(n - \tau_r)\mathbf{g}_r + \mathbf{v}(n), \tag{F.4}$$

$$\mathbf{G} = \left[[\mathbf{D}_0\mathbf{g}_r]^T, [\mathbf{D}_1\mathbf{g}_r]^T, \cdots, [\mathbf{D}_{M-N}\mathbf{g}_r]^T\right]^T \tag{F.5}$$

$$\mathbf{g}_r = [g_{0,r}, g_{1,r}, \cdots, g_{M-1,r}]^T. \tag{F.6}$$

$$\mathbf{S}(n-\tau) = \begin{bmatrix} s(n-\tau+M-1) & \cdots & s(n-\tau) \\ s(n-\tau+M) & \cdots & s(n-\tau+1) \\ \vdots & & \vdots \\ s(n-\tau+N-1) & \cdots & s(n-\tau+N-M) \end{bmatrix}, \tag{F.7}$$

$$\mathbf{s}(n-\tau) = [s(n-\tau), s(n-\tau+1), \cdots, s(n-\tau+N-1)]^T, \tag{F.8}$$

Here, $\mathbf{D}$ is a cyclic shift register that delays filter gain $\mathbf{g}_r$. The matrix $\mathbf{G}_r$ has a dimension of $(N-M+1)\times N$ while $\mathbf{S}_r$ has a dimension of $(N-M+1)\times M$, where N is the length of the signal while M is the filter length. The filter $\mathbf{g}_r$ is a $1 \times M$ vector of the r-th reflection. If we assume that the noise term is white Gaussian noise, the maximum likelihood estimator for the unknown filters, $\mathbf{g}_r$, and delays, τ_r, for $r = 1, \ldots, R$, is given by

$$\{\widehat{\boldsymbol{\tau}}, \widehat{\mathbf{g}}\} = \arg \min_{\boldsymbol{\tau}_r, \mathbf{g}_r \forall r \in [1;R]} \left\| \mathbf{y}(n) - \sum_{r=1}^{R} \mathbf{S}(n-\tau_r)\mathbf{g}_r \right\|^2 . \tag{F.9}$$

Compared to [F.21], we do not assume that the gain or filter $\mathbf{g}_r$ is the set to 1. Hence, the problem at hand is to estimate the delay τ_r and the filter parameters $\mathbf{g}_r$. Moreover, in this paper, we are interested to estimate these parameter to localize the position of an acoustic reflector using echolocation which was not addressed in [F.21]. Furthermore, resolving (F.9) to estimate τ_r and $\mathbf{g}_r$ clearly, leaves us with a computationally complex and multidimensional task. However, as we shall see next, this can be solved by incorporating iterative procedures such as expectation-maximization (EM).

3 Robust EM-based acoustic reflector localization

The EM algorithm developed in [F.22] is a general method intended to solve Maximum-likelihood (ML) estimation problem given incomplete data [F.21]. It is intended to alleviate the complexity of parameter estimation. The EM algorithm requires that the complete data be specified. Here, we may define our complete data as all the observations of the individual reflections, each defined as

$$\mathbf{x}_r(n) = \mathbf{S}(n-\tau_r)\mathbf{g}_r + \mathbf{v}_r(n), \tag{F.10}$$

for, $r = 1, \ldots, R$, where $\mathbf{v}_r(n)$ are individual noise terms obtained by arbitrarily decomposing the noise term $\mathbf{v}(n)$ into R components, such that

$$\sum_{r=1}^{R} \mathbf{v}_r(n) = \mathbf{v}(n). \tag{F.11}$$

Moreover, we can write the observed signal as the sum of the individual observed reflections, i.e.,

$$\mathbf{y}(n) = \sum_{r=1}^{R} \mathbf{x}_r(n). \tag{F.12}$$

We let the individual noise terms be independent, zero-mean, white Gaussian, and distributed as $\mathcal{N}(\mathbf{0}, \beta_r \mathbf{C})$, where $\mathbf{0}$ is a vector of zeros and $\mathbf{C} = \mathrm{E}[\mathbf{v}(n)\mathbf{v}^T(n)] = \sigma_v^2 \mathbf{I}_N$ is an $N \times N$ matrix of $\mathbf{v}(n)$, σ_v^2 is the variance. E[.] is the mathematical expectation. Moreover, the scaling factors, β_r, are non-negative, real-valued scalars, that satisfy the following:

$$\sum_{r=1}^{R} \beta_r = 1. \tag{F.13}$$

Here, the β_r must satisfy the condition above but it is arbitrary free variable and could be used to control the rate of convergence. The choice of β could be resort to more investigation as noted by [F.21] but here we choose the $\beta = 1/\mathbf{R}$. The EM algorithm for the problem at hand is given by

E-step:

$$\widehat{\mathbf{x}}_r^{(i)}(n) = \mathbf{S}(n - \widehat{\tau}_r^{(i)})\widehat{\mathbf{g}}_r^{(i)} + \beta_r \left[\mathbf{y} - \sum_{r=1}^{R} \mathbf{S}(n - \widehat{\tau}_r^{(i)})\widehat{\mathbf{g}}_r^{(i)}. \right] \tag{F.14}$$

M-step:

$$\{\widehat{\mathbf{g}}_r, \widehat{\tau}_r\}^{(i+1)} = \arg\min_{\mathbf{g},\tau} \left\| \mathbf{x}_r^{(i)}(n) - \mathbf{S}(n-\tau)\mathbf{g} \right\|^2, \tag{F.15}$$

where $^{(i)}$ denotes the iteration index. The M-step can be simplified, since the estimator is linear in with respect to the unknown filter coefficients. Moreover, under white Gaussian conditions, the estimator in (F.15) becomes a maximum likelihood estimator. We can thus solve for these first, which yields

$$\widehat{\mathbf{g}}_r^{(i+1)} = \left[\mathbf{S}^T(n-\tau)\mathbf{S}(n-\tau)\right]^{-1} \mathbf{S}^T(n-\tau)\mathbf{x}_r^{(i)}(n), \tag{F.16}$$

If we insert this back into (F.15), we get

$$\widehat{\tau}_r^{(i+1)} = \arg\max_{\tau} \mathbf{x}_r^{(i)} \mathbf{S}(n-\tau) \left[\mathbf{S}^T(n-\tau)\mathbf{S}(n-\tau)\right]^{-1} \mathbf{S}^T(n-\tau)\mathbf{x}_r^{(i)}(n),$$

A potential problem with these estimators is that the filter estimates $\widehat{\mathbf{g}}_r$ are unconstrained, which may lead to unreasonably large filter coefficients, since the reflections may partly cancel each other out. One way of addressing such problems is by introducing a constraint on the white noise gain of the filter:

$$\{\widehat{\mathbf{g}}_r, \widehat{\tau}_r\}^{(i+1)} = \arg\min_{\mathbf{g},\tau} \left\| \mathbf{x}_r^{(i)}(n) - \mathbf{S}(n-\tau)\mathbf{g} \right\|^2 \quad \text{s.t.} \quad \|\mathbf{g}\| < \epsilon. \tag{F.17}$$

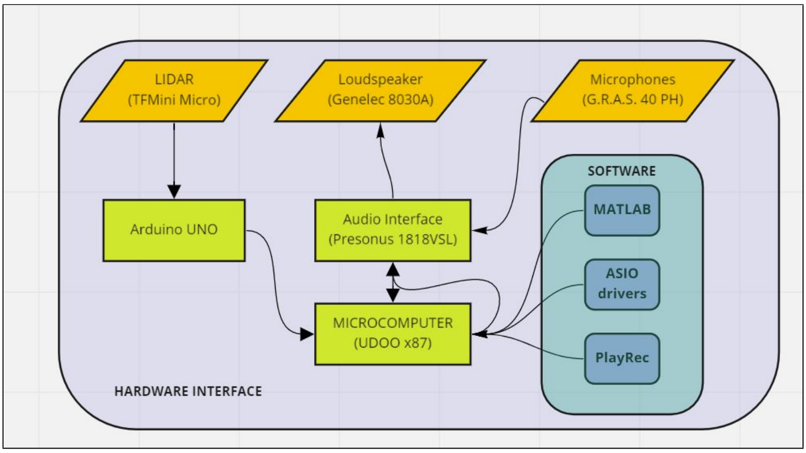


Fig. F.2: An overview of the hardware required to design the platform used in this research

This can be solved using the method of Lagrange multipliers, i.e., to solve for the constrained filter, we write

$$\begin{aligned}\{\widehat{\mathbf{g}}_r, \widehat{\tau}_r\} &= \arg\min_{\mathbf{g},\tau} -2\mathbf{x}_r^T(n)\mathbf{S}(n-\tau)\mathbf{g} + \mathbf{g}^T\mathbf{S}^T(n-\tau)\mathbf{S}(n-\tau)\mathbf{g} + \lambda(\mathbf{g}^T\mathbf{g} - \epsilon)\\ &= \arg\min_{\mathbf{g},\tau} J(\mathbf{g},\tau) \qquad \text{(F.18)}\end{aligned}$$

By taking the partial derivative with respect to the filter, we get

$$\frac{\partial J}{\partial \mathbf{g}_r} = -\mathbf{S}^T(n-\tau)\mathbf{x}_r(n) + \mathbf{S}^T(n-\tau)\mathbf{S}(n-\tau)\mathbf{g}_r + \lambda\mathbf{g}_r = 0. \qquad \text{(F.19)}$$

That is, the filter estimate becomes

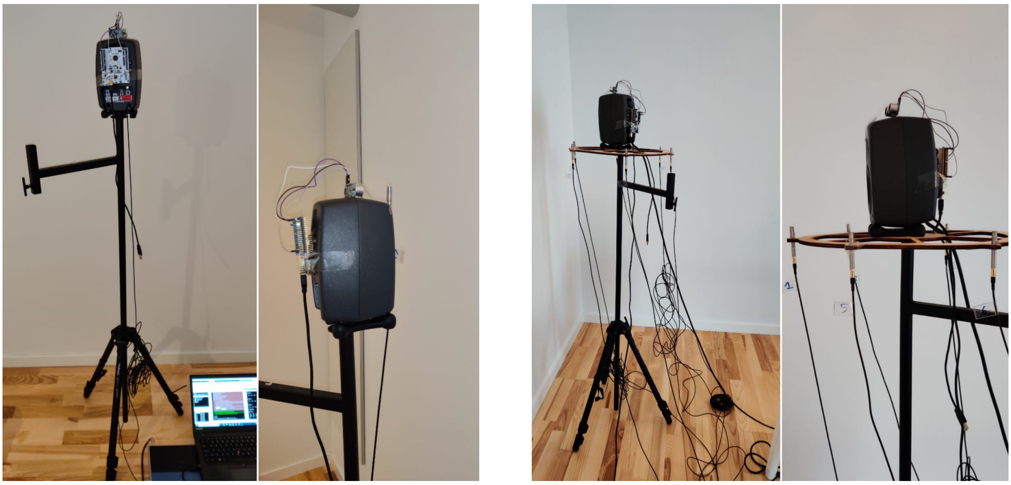
$$\widehat{\mathbf{g}}_r = \left[\mathbf{S}^T(n-\tau)\mathbf{S}(n-\tau) + \lambda\mathbf{I}\right]^{-1}\mathbf{S}^T(n-\tau)\mathbf{x}_r(n). \qquad \text{(F.20)}$$

where λ is the tuning parameter that is empirically set while the $\mathbf{I}$ is the identity matrix. The estimated τ_r of an acoustic reflector could be converted into distance estimate if we assume that the speed of sound is known for the given environment and that we are interested in estimating only the first-order early reflection. This simple conversion can be done as follows:

$$d = c \times \tau, \qquad \text{(F.21)}$$

where c is the speed of sound and d is the distance of an acoustic reflector with respect to a source.

However, by taking the acoustic response within the model, we can estimate multiple reflections originating from two acoustic reflector, i.e., first-order and second-order reflection. By combining the proposed method with echolabeling, we can estimate the position of multiple acoustic echoes.



(a) Hardware setup for experiments with single channel microphone-loudspeaker

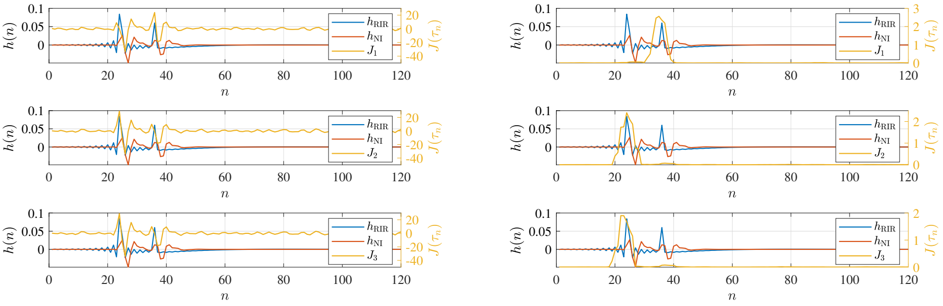
(b) Hardware setup for experiments with multi channel microphones organized in a uniform circular array with a loudspeaker placed at the center of the array

Fig. F.3: Proof of concept used to obtained acoustic data

4 Experimental results

In this section, we investigate two issues, the performance of the proposed method under different conditions, and the benefit of estimating multiple acoustic echoes. In the first experiment, the proposed method was tested using signals that are synthesized using the room impulse response generator [F.23] with the following setup. The synthetic room has a dimension of $6.38 \times 5.4 \times 4.05$ m. The analysis window considered were set to τ_{min} and τ_{max} corresponding to a distance of 0.5 m to 3 m similar to the work performed in [F.15]. This analysis window also helps in estimating the first-order early reflection and prevents direct-path component from being estimated. Moreover, the probe signal $s(n)$ is a broadband signal of length $2,000$ samples drawn from a Gaussian burst with zero padding to form a signal of length $20,000$ samples.

The experimental platform used to evaluate the performance of the proposed method. The overall system architecture is shown in Fig. F.2. Two design variations are proposed to test the proposed method for acoustic reflector's position and distance estimation. One variation consists of a loudspeaker (Genelec 8030A) with a microphone (G.R.A.S 40 PH) attached on top of the loudspeaker. The distance between the acoustic center of a loudspeaker and the center of a microphone is 0.15 m. This is shown in Fig. F.3(a). The second variation consist of a 6 microphones arranged in a uniform circular array (UCA) of radius 0.2 m with a loudspeaker placed at the center of the UCA. This is shown in Fig. F.3(b). The loudspeaker-microphone was



(a) Cost functions of the M-step for $M = 1$ using the EMI method in [F.21]

(b) Cost functions of the M-step for $M = 5$ and $\lambda = 100$ using the proposed method (EMR).

Fig. F.4: Estimating multiple acoustic echoes using simulated data

placed 1.5 m above the floor inside Aalborg University's `Sound Lab` that has a dimension of $6.38 \times 5.4 \times 4.05$ m^3. Furthermore, both the loudspeaker and microphones are connected to an audio interface (Presonus 1818VSL). A Lidar sensor (TFMini Micro) is used to measure the distance between the wall and the platform and is used as a ground truth for further analysis. The audio interface is subsequently connected to a laptop via a USB port. To ensure low latency from hardware, ASIO drivers[2] is installed from the internet. Moreover, MATLAB is used as a data acquisition software tool to record and save the observed signals and for statistical analysis on the proposed method. Furthermore, for multichannel data acquisition PlayRec [F.24] is used to transmit and record sound simultaneously. The sampling frequency is set to $48,000$ Hz while the speed of sound is assumed as 343 m/s

4.1 Proof-of-concept

4.2 Simulated and real results

In the first experiment, the non-ideal characteristic of acoustic systems is modelled by filtering the room impulse response, h_{RIR} using a bandpass filter with impulse response, h_{BP}, to obtain our non-ideal impulse response, h_{NI}, i.e.,

$$h_{\text{NI}} = h_{\text{RIR}} * h_{\text{BP}}. \tag{F.22}$$

The band pass filter was a second order Butterworth filter with cutoff frequencies, $\boldsymbol{\omega} = [0.2\pi, 0.6\pi]$. The non-ideal room impulse response was then applied to a known probe signal, $s(n)$, to generate the observation used for the experiment. Here, the search interval for the delays, or TOAs,

[2] https://www.asio4all.org/

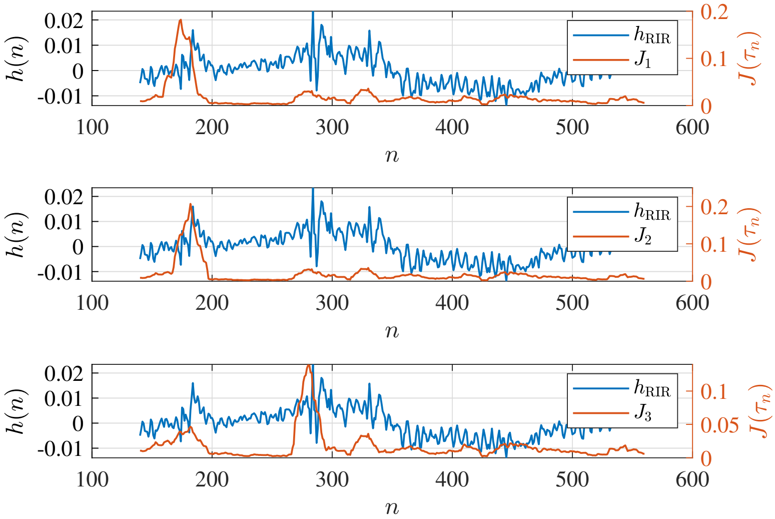
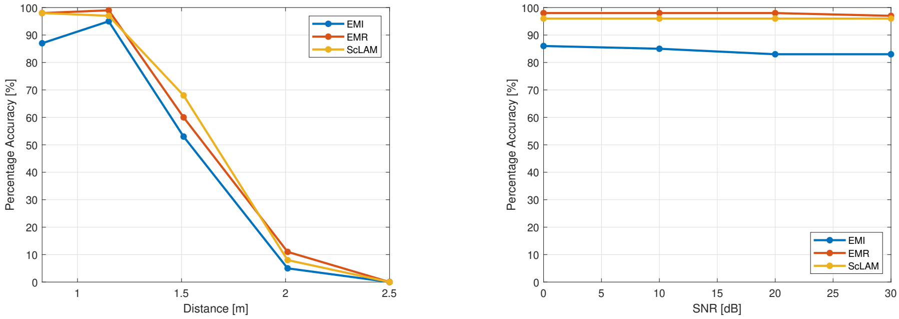


Fig. F.5: Estimating multiple acoustic echoes using real data obtained from hardware platform in Fig. F.3(a)

was chosen as $\tau \in [1, 80]$ samples, and therefore we set N to $2,080$. The number of reflections was set to $R = 3$ because this number give us better estimates of 2 acoustic reflectors, the number of EM iterations was set to 100, and $\beta_r = 1/R$. Furthermore, the direct-path component was removed from the observed signal using RIR generator. Using this setup, we ran the Ideal-EM (EMI) method with a filter length $M = 1$ as proposed in [F.21], and the presented Robust-EM method (EMR) with filter length $M = 5$ and $\lambda = 100$. The resulting cost functions are depicted in Figures F.4(a) and Fig.F.4(b), respectively. From the results, we can first see how the ideal impulse responses are affected by the bandpass filter applied to it, which smears out the peaks. When applying the EMI method, we therefore also do not see two clearly defined peaks around the time-of-arrivals of the two components. If we instead use the EMR method, we can model the effects of the bandpass filter, which results in two broader, but clearly defined peaks at the TOA.

Furthermore, we repeat the simulated experiment in practical setting using the hardware platform in Fig. F.3(a). The platform was placed at a corner of a room with a distance to the walls, 1 m and 0.65m, respectively. The collocated microphone-loudspeaker setup probes the environment with a known sound and the received echoes are recorded by the microphone. The observed signal was later used to estimate RIR of the environment using dual-channel method [F.25]. This is done by computing $\widehat{H}(f) = Y(f)/S(f)$ and then taking the inverse DFT to get $\widehat{h} = \mathcal{F}^{-1}\{\widehat{H}(f)\}$. The EMR's filter length was set to $M = 15$, $\lambda = 500$ and $R = 3$.



(a) Comparison of the proposed method Robust EM with $M = 5$ and $\lambda = 100$ against Ideal EM $M = 1$ for acoustic reflector estimation at varying distances

(b) Comparison of the proposed method Robust EM $M = 5$ and $\lambda = 100$ against Ideal EM with $M = 1$ for acoustic reflector estimation against different background noise.

Fig. F.6: Comparison of the EMI, EMR and ScLAM methods under different distances and background noise

As seen in Fig. F.5, the EMR method successfully estimates all the peaks corresponding to individual acoustic reflector. In this experiment, both M and λ are set empirically. However, in the future iteration of this work, we can adaptively select these parameters.

4.3 Impact of distances and background noises

In this experiment, we evaluate the performance of the proposed TOA estimator and compared it against varying distances. The setup was placed at a distance of $[0.8, 1.0, 1.5, 2.0, 2.5]$ m and 100 acoustic echoes were recorded at each interval. The data was collected using the single channel setup shown in Fig. F.3(a). The Accuracy is defined as the percentage of TOA that are within $\pm 10\%$ of the ground truth value obtained from lidar. The proposed method (EMR) is compared with previous method (EMI) proposed by [F.21], and single-channel localization and mapping (ScLAM) [F.16]. These results are shown in Fig. F.6(a). The data obtained from this experiment is also summarized in Table F.1.

Additionally, a comparison of the proposed method against different background noise was also performed. To simulate different noise levels, a separate loudspeaker was placed at a distance of 6.4 m away from the setup within the lab. This separate loudspeaker was used to simulate low signal-to-noise ratio (SNR). The separate loudspeaker is playing an audio clip from YouTube called cocktail party[3]. The SNR is defined as the variance of the observed signal,

[3] https://youtu.be/IKB3Qiglyro

Table F.1: Comparison of EMI against the other TOA estimation methods under different distances and background noise

	EMI SNR = 30dB			EMI SNR = 0dB		
Lidar Data [m]	μ **[m]**	σ **[m]**	**RMSE [m]**	μ **[m]**	σ **[m]**	**RMSE [m]**
0.83	0.8886	0.0403	0.0710	0.8856	0.0436	0.0704
1.15	1.1306	0.1274	0.1282	1.1151	0.1108	0.1156
1.51	1.4185	0.2522	0.2671	1.4288	0.2739	0.2844
2.01	1.2356	0.2772	0.8221	1.2348	0.2689	0.8201
	EMR $M = 5$ $\lambda = 100$ **SNR = 30dB**			**EMR** $M = 5$ $\lambda = 100$ **SNR = 0dB**		
Lidar Data/m	μ **[m]**	σ **[m]**	**RMSE [m]**	μ **[m]**	σ **[m]**	**RMSE [m]**
0.83	0.8734	0.0105	0.0447	0.8703	0.0233	0.0464
1.15	1.0772	0.0252	0.0769	1.0705	0.0246	0.0831
1.51	1.4370	0.2585	0.2674	1.4541	0.2549	0.2597
2.01	1.2379	0.3434	0.8443	1.2837	0.3531	0.8067
	ScLAM = 30dB			**ScLAM = 0dB**		
Lidar Data [m]	μ **[m]**	σ **[m]**	**RMSE [m]**	μ **[m]**	σ **[m]**	**RMSE [m]**
0.83	0.8826	0.0059	0.0709	0.8796	0.0214	0.0704
1.15	1.0977	0.0871	0.1281	1.0776	0.0395	0.1156
1.51	1.4789	0.2301	0.2670	1.5312	0.2245	0.2843
2.01	1.2658	0.3276	0.8221	1.2648	0.3197	0.8200

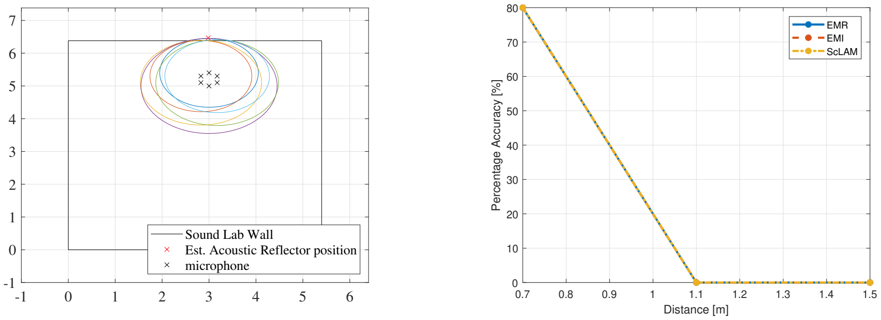
$\mathbf{x}(n)$, against the variance of the background noise, $\mathbf{v}(n)$.

$$\text{SNR} = \frac{\sigma_x^2}{\sigma_v^2}, \tag{F.23}$$

where $\sigma_x^2 = E[\|\mathbf{x}(n)\|^2]$ and $\sigma_v^2 = E[\|\mathbf{v}(n)\|^2]$. Both the observed signal and the background noise is recorded for 1 sec. The background noise was recorded before the system probed the environment with a known signal. Based on this configuration, 4 SNRs were selected by adjusting the loudness of the separate speaker, $[0, 10, 20, 30]$ dB. Furthermore, 100 audio recordings were obtained at each SNRs to evaluate the proposed method (EMR). The evaluation results are shown in Fig. F.6(b).

4.4 Evaluation of Robust EM using multilateration technique

In this experiment, we test the performance of the proposed method using multilateration technique. In this way, we can estimate the DOA of the acoustic echoes which can aid robotic platforms to locate the source of the acoustic echoes. The idea here is that the proposed method will estimate TOAs from each of the microphone-loudspeaker combinations, which will then be used with a multilateration technique. Multilateration is a localization techniques popularly used in telecommunication to estimate the direction and distance of a transmitter/source



(a) EMR and multilateration technique to localize an acoustic echo situated at a distance of 0.7 m. The convergence of the individual circles indicate the location of the acoustic reflectors

(b) Performance of proposed and existing methods against varying distances

Fig. F.7: a) Evaluation of the proposed method with multilateration to detect a single acoustic reflector b) Evaluation of the proposed and existing method against distance using the setup in Fig. F.3(b)

Table F.2: Performance of the proposed method using multilateration technique evaluated over distances

	EMI SNR = 30			**EMR SNR = 30**			**ScLAM SNR = 30**		
Lidar Data [m]	μ[m]	σ[m]	RMSE [m]	μ[m]	σ[m]	RMSE [m]	μ[m]	σ[m]	RMSE [m]
0.7	0.6240	0.1442	0.1617	0.6154	0.15383	0.16176	0.65628	0.072963	0.08443
1.1	0.8428	0.0689	0.2660	0.77155	0.058971	0.26605	0.77155	0.058971	0.3336
1.5	1.1686	0.3247	0.4617	3.1354	0.18567	1.9132	1.6851	1.5701e-15	0.18509

[F.26, F.27, F.28]. Moreover, multilateration was also used to estimate robot's position in 3D space as proposed in [F.29]. Within the context of this paper, multilateration is used to estimated the location of the acoustic reflector. Multilateration techniques rely on the TOAs knowledge of the acoustic reflections and also assume that the locations of the sensor nodes are known with respect to the same coordinate system. To locate an acoustic reflector, we need to set a reference with respect to a coordinate system. This information could be known from robot's motor encorder or Inertial Measurement Unit (IMU) but this aspect of robot navigation is beyond the scope of this paper. More specifically, let us assume that we have M microphones and the source is placed on the same xy-plane. Using (3), we can estimate the TOA and (F.21), the range value vector, $\mathbf{d}$. If the microphones are located on the xy-plane or 2D plane, at positions, $[\mathbf{x}_p, \mathbf{y}_p] = [(x_1, y_1), (x_2, y_2), \ldots, (x_P, y_P)]$, where P is the number of microphones, then based on the range data $\mathbf{d}_p$ a circle can be draw from each microphones. The point of intersection of these individual circles would yield the location of the acoustic reflector as seen in Fig. F.7(a). The true acoustic reflector position (x, y) is at the intersection of all the circles and satisfies the following equations:

$$(x - x_p)^2 = d_p^2, \quad p = 1, \cdots, P. \tag{F.24}$$

In the presence of noise, the estimations of $[\mathbf{d}]$, the circles will not intersect at a single point. Therefore, a least-square fit can used to obtain the acoustic reflector location estimate [F.30], i.e.,

$$\mathbf{r}_s = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}, \tag{F.25}$$

where

$$\mathbf{A} = \begin{bmatrix} 2(x_1 - x_P) & 2(y_1 - y_P) \\ \vdots & \\ 2(x_{P-1} - x_P) & 2(y_{P-1} - y_P) \end{bmatrix} \tag{F.26}$$

$$\mathbf{b} = \begin{bmatrix} x_1^2 - x_P^2 + y_1^2 - y_P^2 + d_P^2 - d_1^2 \\ \vdots \\ x_{P-1}^2 - x_P^2 + y_{P-1}^2 - y_P^2 + d_P^2 - d_{P-1}^2 \end{bmatrix} \tag{F.27}$$

The setup used for this experiment is shown in Fig.F.3(b). Here, the setup was fixed at distances $[0.7, 1.1, 1.5]$ m against an acoustic reflector. Furthermore, 50 recordings were made at each distance which was later evaluated. The results are depicted in Fig.F.7 and listed in Table F.2.

5 Discussion and limitations

Two platform designs were proposed to test the algorithm: A collocated microphone-loudspeaker as seen in Fig. F.3(a) and a uniform circular microphone array with a loudspeaker positioned at the center of the array as seen in Fig. F.3(b). The results obtained from the first experiment reveals that the proposed method can be used to estimate multiple acoustic reflections as EMR can account for acoustic system's response which can hinder the estimation accuracy of multiple acoustic reflections. As seen in Fig. F.4, EMR estimates multiple peaks that corresponds to an acoustic reflectors while EMI estimates a single acoustic reflector. Estimating multiple acoustic reflectors is beneficial for spatial map construction in an indoor environment.

In the second experiment, the performance of EMR and EMI is evaluated using the proof-of-concept described in Section 4.1. The results in Fig. F.6(a) reveals that EMR provides significant improvements in estimating the acoustic reflector as it can account for acoustic system's response that affects the performance of TOA estimator while Fig. F.6(b) shows that the proposed method is 10% better than the EMI method over all SNR values which is on par with the ScLAM techniques. According to the results, the proposed method can estimate an acoustic reflector up to a distance of 1.5 m with 60% under low SNR of 0 dB. According to Table.F.1, both the standard deviation σ and Root Mean Square Error (RMSE) of the EMI and EMR increases when the distance between the acoustic reflector and the platform increases while the mean value μ is close to the ground truth for distance up to 1.5 and for all SNRs.

In the last experiment, we combined the proposed method with multilateration technique so that the direction as well as the location of the acoustic reflector is determined by a robotic system as it navigates an indoor environment. Here, we test EMI, EMR and ScLAM under an SNR of 30 dB and place the multi-channel setup at varying distances. According to the results obtained in Fig. F.7(b), All methods can estimate an acoustic reflector up to a distance of 0.7 m with 80% accuracy. This reduction in accuracy could be due to the loudspeaker blocking the acoustic echoes from reaching one of the microphone placed behind the loudspeaker that could affect the TOA estimation. This could results in estimating spurious estimates that can reduce the performance of the multilateration technique when locating an acoustic source. Furthermore, according to Table F.2, the σ and RMSE values of the proposed method increases as the platform's distance with respect to the wall is also increase while μ value is close to 0.7 m at a SNR of 30. Similar performance is seen in the remaining methods. However, for multilateration technique to work, the robotic platform requires the knowledge of its Cartesian position of the environment, i.e., the position of the loudspeaker and microphones should be known. One way to acquire this information is by utilizing sensors used for tracking the odometry and orientation of a robot, e.g., Inertial Measurement Unit. However, in this paper, we assume that the location of the loudspeaker and microphones will be known.

6 Conclusion and future work

The contribution of this paper is to propose a Robust Expectation-Maximization technique for acoustic reflector localization, intended for robotic platform using echolocation. The proposed method builds on existing work proposed by [F.21], i.e., their work assumed that the gain or filter parameters are assumed to be the same which in practice is not a valid assumption as this can hinder the acoustic reflector estimation process. Hence, in this paper, we introduced this uncertainty within the signal formulation. Three experiments were performed in simulated and practical environment. To test the performance of the proposed method, two proof-of-concept platforms are used: One consist of a collocated microphone-loudspeaker arrangement while the other consist of a uniform circular microphone array with a loudspeaker placed at the center of an array. From our experimental results, we deduce that our proposed method can estimate an acoustic reflector up to a distance of 1.5 m with 60% accuracy and can be combined with multilateration technique to locate the direction of an acoustic reflector. Our proposed method can be beneficial to robotic platform as it can complement existing laser- and camera-based technologies for generating a spatial map of an indoor environment as done in our previous works. Our proposed echolocation method can aid a robotic platform to detect and estimate transparent surfaces and can also estimate multiple acoustic echoes when a robot moves to a corner of a room.

In the future iteration of this work, we aim to implement the proposed method on existing robotic platform, e.g., Softbank's NAO robot and also improve the algorithm and combining it with echolabeling techniques as proposed in [F.31] so that multiple acoustic echoes are estimated and categorized to represent a indoor environment. Moreover, these method could also be used in a wireless acoustic sensor networks (WASN) to detect acoustic sources [F.32, F.16].

References

[F.1] J. Steckel and H. Peremans, “BatSLAM: Simultaneous localization and mapping using biomimetic sonar,” *PLOS ONE*, vol. 8, no. 1, pp. 1–11, 01 2013.

[F.2] R. Kuc, “Echolocation with bat buzz emissions: Model and biomimetic sonar for elevation estimation,” vol. 131, no. 1, pp. 561–568, 2012.

[F.3] M. Kreković, I. Dokmanić, and M. Vetterli, “EchoSLAM: Simultaneous localization and mapping with acoustic echoes,” pp. 11–15, 2016.

[F.4] S. Tervo, J. Pätynen, and T. Lokki, “Acoustic reflection localization from room impulse responses,” *ACTA Acustica united with Acustica*, vol. 98, no. 3, pp. 418–440, 2012.

[F.5] G. Defrance, L. Daudet, and J.-D. Polack, “Detecting arrivals within room impulse responses using matching pursuit,” *Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx-08), Espoo, Finland*, vol. 10, pp. 307–316, 2008.

[F.6] G. Defrance, L. Daudet, and J. Polack, “Using matching pursuit for estimating mixing time within room impulse responses,” *ACTA Acustica united with Acustica*, vol. 95, no. 6, pp. 1071–1081, 2009.

[F.7] G. Moschioni, “A new method for measurement of early sound reflections in theaters and halls,” *IMTC/2002. Proceedings of the 19th IEEE Instrumentation and Measurement Technology Conference (IEEE Cat. No.00CH37276)*, vol. 1, pp. 425–430 vol.1, 2002.

[F.8] U. Saqib, S. Gannot, and J. Jensen, “Estimation of acoustic echoes using expectation-maximization methods,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–15, 2020.

[F.9] Y. Geng and J. Jung, “Sound-source localization system for robotics and industrial automatic control systems based on neural network,” *2008 International Conference on Smart Manufacturing Application*, pp. 311–315, 2008.

[F.10] S. Dey, S. Boppu, and M. S. Manikandan, “Design of a real-time automatic source monitoring framework based on sound source localization,” *2019 Seventh International Conference on Digital Information Processing and Communications (ICDIPC)*, pp. 35–40, 2019.

[F.11] H. Zhu and H. Wan, “Single sound source localization using convolutional neural networks trained with spiral source,” *2020 5th International Conference on Automation, Control and Robotics Engineering (CACRE)*, pp. 720–724, 2020.

[F.12] N. Riopelle, P. Caspers, and D. Sofge, “Terrain classification for autonomous vehicles using bat-inspired echolocation,” *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6, 2018.

[F.13] J. H. Christensen, S. Hornauer, and S. X. Yu, “Batvision: Learning to see 3d spatial layout with two ears,” *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1581–1587, 2020.

[F.14] E. Tracy and N. Kottege, “Catchatter: Acoustic perception for mobile robots,” *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7209–7216, 2021.

[F.15] U. Saqib and J. Jensen, “A model-based approach to acoustic reflector localization using robotic platform,” pp. 1–8, 2018.

[F.16] U. Saqib and J. R. Jensen, “A framework for spatial map generation using acoustic echoes for robotic platforms,” *Robotics and Autonomous Systems*, vol. 150, p. 104009, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889021002633

[F.17] D. W. Gunness, “Loudspeaker transfer function averaging and interpolation,” *journal of the audio engineering society*, november 2001.

[F.18] U. Saqib and J. R. Jensen, “Sound-based distance estimation for indoor navigation in the presence of ego noise,” 2019.

[F.19] P. Ahgren and P. Stoica, “A simple method for estimating the impulse responses of loudspeakers,” *IEEE Transactions on Consumer Electronics*, vol. 49, no. 4, pp. 889–893, 2003.

[F.20] Z. Sü and M. Çalışkan, “Acoustical design and noise control in metro stations: Case studies of the ankara metro system,” *Building Acoustics*, vol. 14, no. 3, pp. 203–221, 2007.

[F.21] M. Feder and E. Weinstein, “Parameter estimation of superimposed signals using the em algorithm,” *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 4, pp. 477–489, 1988.

[F.22] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[F.23] E. A. P. Habets, I. Cohen, and S. Gannot, “Generating nonstationary multisensor signals under a spatial coherence constraint,” vol. 124, no. 5, pp. 2911–2917, 2008.

[F.24] R. Humphrey, “Playrec: Multi-channel matlab audio,” *URL http://www.playrec.co.uk*, 2007.

[F.25] H. Herlufsen, “Dual channel FFT analysis (part I),” *Brüel & Kjær Technical Review*, no. 1984-1, 1984.

[F.26] J. Yang, H. Lee, and K. Moessner, “Multilateration localization based on singular value decomposition for 3d indoor positioning,” *Int. Conf. Indoor Positioning and Indoor Navigation*, pp. 1–8, 2016.

[F.27] J. Wan, N. Yu, R. Feng, Y. Wu, and C. Su, “Localization refinement for wireless sensor networks,” *Computer Communications*, vol. 32, no. 13, pp. 1515–1524, 2009.

[F.28] Y. Zhou, Jun Li, and L. Lamont, “Multilateration localization in the presence of anchor location uncertainties,” *IEEE Global Communications Conference (GLOBECOM)*, pp. 309–314, 2012.

[F.29] A. Yazici, U. Yayan, and H. Yücel, “An ultrasonic based indoor positioning system,” *Int. Symposium on Innovations in Intell. Sys. and Applications*, pp. 585–589, 2011.

[F.30] C. Chen and K. Yao, “Source and node localization in sensor networks,” T. E. Tuncer and B. Friedlander, Eds. Boston: Academic Press, 2009, pp. 343–383.

[F.31] I. Dokmanic, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, “Acoustic echoes reveal room shape,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 30, pp. 12 186–12 191, 2013. [Online]. Available: https://www.pnas.org/content/110/30/12186

[F.32] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, “A survey of sound source localization methods in wireless acoustic sensor networks,” *Wireless Communications and Mobile Computing*, vol. 2017, 2017.

Paper G

A Framework for Spatial Map Generation using Acoustic Echoes using for Robotic Platforms

Usama Saqib and Jesper Rindom Jensen

The paper has been published in the
Elsevier's Robotics and Autonomous Systems

The layout has been revised.

Abstract

In this work, we present a framework for constructing a spatial map of an indoor environment using the concept of echolocation. More specifically, we propose a non-linear least squares (NLS) estimator which is combined with a spatial filtering technique, e.g., beamforming, to estimate both the time-of-arrival (TOA) and direction-of-arrival (DOA) of the acoustic echoes. The proposed framework is complemented with an echo detector to classify a spurious estimate and an acoustic reflector, i.e., a wall. Based on these estimators, we propose two algorithms that complement existing range sensors and aid robotic platforms in acoustic reflector localization and mapping: single-channel localization and mapping (ScLAM) and a multi-channel localization and mapping (McLAM). Compared to commonly used sensors, such as lidar, cameras and ultrasonic sensors, our proposed model-based approach can detect transparent surfaces that are typically found in an office environment and could work in audible frequency ranges. A proof-of-concept robotic platform was built to test our algorithms. According to our evaluation, both qualitative and quantitative experiments reveal that the proposed methods can detect an acoustic reflector up to a distance of 1.5 *m at a signal-to-diffuse-noise ratio (SDNR) of* 0 *dB in a simulated environment and* 10 *dB in a real environment with an accuracy of* 80 %.

1 INTRODUCTION

Robotic platforms, e.g., drones and unmanned ground vehicles (UGVs), have become an essential part of our society. We use them for tasks that are often monotonous and too dangerous for human workers to handle. With the advancement of perception technology, i.e., the ability of a robot to perceive its environment, robots are now able to perform complicated tasks that make them suitable for work in different sectors, such as agriculture [G.1], construction [G.2], supply chain and logistics [G.3], hospitals [G.4], etc. Within a warehouse setting, robots are often programmed to follow a predefined trajectory within an environment to transport goods. Over time, robots were equipped with proximity sensors, like lidars, cameras, ultrasonic sensors, etc., for navigation, which then led to the robots being able to plan their own paths without human intervention, making these robots more autonomous. According to the IEEE Standard for Robot Map Data Representation for Navigation [G.5], one way to effectively navigate an indoor environment is to construct a spatial map of said environment, which is normally done using a very popular framework called Simultaneous Localization and Mapping (SLAM) [G.6, G.7, G.8]. One of the advantages of constructing a spatial map of a surrounding environment is that it could also aid engineers and building planners doing asset maintenance and survey related work. Additionally, SLAM-based robots aid rescue workers and surveyors in constructing spatial maps of unknown environments, like sewers [G.9, G.10], underground tunnels, etc. Traditionally, lidar and camera-based technologies are used to provide input data to SLAM algorithms in constructing spatial maps of different environments [G.11].

However, lidar and camera-based technologies are susceptible to changing light conditions which makes them unsuitable for detecting acoustic reflectors such as walls, glass partitions and wooden surfaces [G.12], hence also making them unsuitable to accurately generate a spatial map of a typical indoor environment [G.13]. Furthermore, lidar and camera-based technologies have a limited field of view (FOV) and, thus, offers limited coverage when localizing targets around the corner of the room [G.14]. These issues can be resolved by employing sound [G.15]. By probing their environment, sound is used by animals (e.g., bats, dolphins, and rats) in nature for orienting themselves within an environment and hunting prey [G.16]. This process is known as echolocation. An advantage of using echolocation for spatial map generation is that it can enable a robotic platform to navigate an environment under poor lighting conditions. Furthermore, compared to camera and lidar-based technologies, microphones are typically cheaper and may offer omni-directionality. In the past, the concept of echolocation was studied by several researchers to build active SONAR (sound navigation and ranging) technologies for naval submarines to detect incoming ships and hostile submarines [G.17].

The use of SONAR in air-borne applications is a challenging and complicated task but an attempt to study this was proposed in [G.18]. The authors utilizes two ultrasonic transmitters/receivers to effectively localize multiple targets up to a distance of 8 m and classification of the targets was done using template matching. Moreover, the authors in [G.19, G.20, G.21] also proposed several techniques that utilizes sound to make a distance estimate of an acoustic reflector. However, these studies assume that the time-of-arrival (TOA) information of the acoustic echoes are known prior to estimation. TOA measurement of an acoustic echo is usually extracted from the estimated room impulse response (RIR) using a standard peak-picking approach, but this has proven to be a non-trivial and time consuming process [G.22]. In acoustic signal processing, the RIR is the transfer function between the source and the microphone. It has a distinctive characteristic, i.e., it contains two main components: a direct-path plus early reflections and a stochastic long tail representing late reflections that contributes to the reverberation [G.23]. The direct-path component corresponds to the shortest distance that a sound travels to reach a receiver while the early reflections correspond to the sound bouncing off an acoustic reflector before reaching the receiver as shown in Fig. G.1. Within the context of robotic platforms, the individual RIRs must be estimated as the robot moves within an environment for TOA estimation of the early reflections as it corresponds to information about the geometry of the room [G.24]. The estimation of TOAs and direction of arrival (DOA) from an observed signal is relatively a new area of research within the context of robotics and has been addressed previously in multipath communication systems [G.25, G.26, G.27, G.28]. However, to the best of our knowledge, it is not addressed to estimate acoustic reflectors using the model-based approach. Furthermore, in order to construct a spatial map of an indoor environment, the DOA of the acoustic echo is also required. This helps a robot estimate the orientation of the acoustic reflectors, i.e., walls. Several techniques of DOA estimation exist in the literature [G.29, G.30, G.31, G.32, G.33].

The notion of utilizing echolocation for spatial map generation on robots has been addressed by several researchers in the past. For instance, [G.34] proposed an algorithm called BatSLAM which utilizes a transmitter/receiver within ultrasonic range to generate a spatial map of an

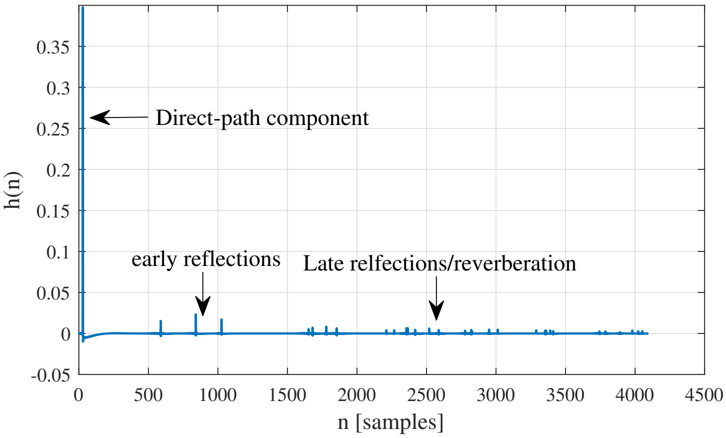


Fig. G.1: An example illustrating synthetic Room Impulse Response (RIR)

indoor environment. Furthermore, in [G.16], the authors built a robotic platform that navigates an outdoor environment in order to construct a spatial map as well as classify flora using an artificial neural network. However, these approaches utilizes specialized sensors that operate in a specific frequency range of sound, i.e., the ultrasonic range. On the other hand, most robotic platforms that exist in the market are intended for human-robot interaction (HRI), such as the NAO robot which is equipped with standard microphones and loudspeakers that operate in audible frequency range. This is addressed by the authors of [G.35, G.36] who propose techniques that utilizes standard smartphones to construct spatial maps of environments using cross-correlation techniques. However, in [G.35], the author uses empirically determined pulse modulation, duration and frequency of the probe signals. The algorithm proposed in [G.35] also requires multiple walks of the same space to generate the contour. Additionally, it requires multiple training data for its probabilistic model, as pointed out by [G.36].

This current study builds on our previous work [G.25, G.12] and extends it in many ways. First, we propose a model-based approach to acoustic distance estimation, where we provide a general model of the early reflections and take ego-noise, interfering sources and background noise within the signal model to estimate TOAs directly from the observed signals instead of relying on RIR knowledge. This way we can probe the environment with a known sound (audible or ultrasonic) and do not have to empirically design probe signals. We intentionally use audible frequency because the existing research on audible frequency sound for distance estimation is comparatively less in scientific literature. Secondly, we resolve our estimate using a least-squares approach such that we can estimate the R order of acoustic reflections. Thirdly, we propose utilizing spatial filtering techniques such as beamforming, to estimate the DOA of the acoustic reflector. In our earlier work [G.25], the presence of the direct-path component was

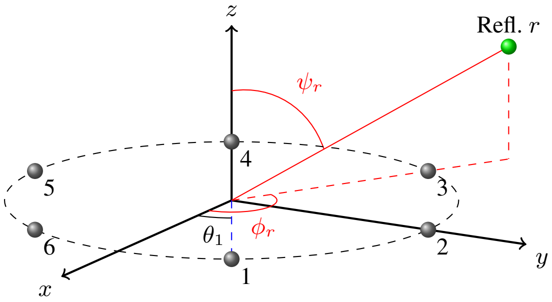


Fig. G.2: Example of a uniform circular array with six microphones.

assumed to be removed during pre-processing to ensure accurate TOA and DOA estimation but, in practice, direct-path elimination is difficult and its presence could have a detrimental effect when estimating TOA and DOA. Therefore, in this work we use a uniform circular array (UCA) and beamforming techniques to reduce the influence of the direct-path component. Moreover, we evaluate different types of beamformers to see which variation works best for our estimation. Finally, we propose a novel acoustic echo detector within our framework using detection theory [G.37]. The detector is a binary classifier, where the statistics of the background noise is used to optimally define a threshold value against which the spurious estimate is differentiated from those corresponding to actual acoustic reflections.

The remainder of the paper is structured as follows: Section 2 describes the signal model and problem formulation, Section 3 describes the non-linear least squares (NLS) estimator, Section 4 describes the first multi-channel localization and mapping (McLAM) algorithm while Section 5 describes single-channel localization and mapping (ScLAM) algorithm. Additionally, the robotic architecture and components descriptions are detailed in Section 6 before proceeding to Section 7 which describes the simulated results of the proposed methods. In Section 8, we test the performance of the proposed method on a robotic platform and then we discuss our findings and conclude the paper in Section 9 and 10, respectively. Moreover, in Sections 10, we propose different ways on how this research could be extended. The dataset used for simulation and evaluation can be found online [1].

2 Signal Model and Problem Statement

2.1 Time-domain model

Consider an array with M microphones recording a sound emitted from a loudspeaker, including its acoustic reflections from walls, etc. The microphones and loudspeaker are collocated, and the loudspeaker is assumed to be a point source. We can then formulate a general model for the

[1] http://homes.create.aau.dk//ussa/journal/index.php

recorded signal at microphone m, for $m = 1, \ldots, M$ at the kth robot position, as

$$\begin{aligned} y_{m,k}(n) &= (h_{m,k} * s)(n) + v_{m,k}(n), \\ &= x_{m,k}(n) + v_{m,k}(n) \end{aligned} \tag{G.1}$$

where $h_{m,k}(n)$ is the acoustic impulse response of the room measured from the loudspeaker to the microphone m at robot position, w_k, for $k = 1 \ldots K$. Moreover, $v_{m,k}(n)$ is the additive background noise, including interfering sources plus the ego-noise of the robot at position, w_k. The operator $*$ represents the convolution operator, and $x_{m,k}(n) = (h_{m,k} * s)(n)$. In what follows, the background noise is assumed to be white Gaussian noise, but prewhitening techniques could be employed in cases where such assumptions are not met, as seen in other studies [G.38, G.39]. By decomposing (G.1) as a sum of its direct-path component and its reflections and expressing the transfer function between the loudspeaker and a microphone in terms of its gain and delay, the signal model in (G.1) can be written as:

$$\begin{aligned} y_{m,k}(n) = &\sum_{r=1}^{\infty} g_{m,r,k} s(n - \tau_{ref,r,k} - \eta_{m,r,k}) \\ &+ v_{m,k}(n) \end{aligned} \tag{G.2}$$

where $g_{m,r,k}$ is the attenuation of the rth reflection from the loudspeaker to the microphone m at position, w_k, while $\tau_{ref,r,k}$ is the TOA of the reflected sound received at the reference point of the UCA at robot position, w_k, while $\eta_{m,r,k}$ is the time-difference-of-arrival (TDOA) between the reference point and the microphone m. In our definition in (G.2), the direct-path component corresponds to $r = 1$. The acoustic impulse response has a certain structure and is distinctively described in two parts: the direct-path plus early reflections and late reflections often described as a stochastic and dense tail. This means that we could rewrite (G.1) as the sum of the first R reflections to facilitate TOA and DOA estimation as shown

$$\begin{aligned} y_{m,k}(n) = &\sum_{r=1}^{R} g_{m,r,k} s(n - \tau_{ref,r,k} - \eta_{m,r,k}) \\ &+ d_{m,k}(n) + v_{m,k}(n), \qquad \text{(G.3)} \\ = &\, x_{m,k}(n) + v'_{m,k}(n) \qquad \text{(G.4)} \end{aligned}$$

where $d_{m,k}(n)$ is the stochastic and dense tail of the late reflections. Often, we can combine the late reflections, $d_{m,k}(n)$, with the background noise as shown in (G.4) [G.40]. If we collect N time samples from each microphone and assume stationarity across those samples within the corresponding time frame, we can vectorize our data and extend our signal model as shown:

$$\begin{aligned}\mathbf{y}_{m,k}(n) &= \sum_{r=1}^{R} g_{m,r,k}\mathbf{s}(n - \tau_{ref,r,k} - \eta_{m,r,k}) \\ &+ \mathbf{d}_{m,k}(n) + \mathbf{v}_{m,k}(n), \\ &= \mathbf{x}_{m,k}(n) + \mathbf{v}'_{m,k}(n) && \text{(G.5)} \\ &= \begin{bmatrix} y_{m,k}(0) & y_{m,k}(1) & \cdots & y_{m,k}(N-1) \end{bmatrix}^T, && \text{(G.6)}\end{aligned}$$

where the time-stacked probe signal, $\mathbf{s}(n)$, early reflections, $\mathbf{x}_{m,k}(n)$, and noise, $\mathbf{v}'_{m,k}(n)$, are defined similarly to $\mathbf{y}_{m,k}(n)$. While not considered in this paper, the stationarity assumption may lifted by extending the model to include the Doppler shift of the probe signal in the presence of, e.g., robot movement [G.41].

Hence, the signal formulation above yields an interesting problem to solve, namely, how to estimate $\tau_{ref,r,k}$ and $\eta_{m,r,k}$ of an acoustic reflector that will aid in simultaneously localizing and mapping an indoor environment. However, this requires us to estimate R unknown TOA and MR TDOAs from the observations $\mathbf{y}_{m,k}(n)$, at position, w_k. If we assume a known array configuration, however, we can reduce the dimensionality of this problem by incorporating the geometry of the loudspeaker and the microphone array.

2.2 Array Model

The array model can be chosen to be of any geometry but in this paper, we use a uniform circular array (UCA) with a loudspeaker placed at the center of the array. Although any reference point could be chosen to solve the TOA and DOA problems, we assume the center of the UCA to be the reference point. Assuming that the reflectors are in the far-field of the array and given the geometry of the microphones and the loudspeaker where the center of the microphone array is chosen as the reference point, we can then write the TDOAs of the acoustic echoes as follows:

$$\eta_{m,r}(\boldsymbol{\zeta}_r) = d \sin\psi_r \cos(\phi_r - \beta_m)\frac{f_s}{c}, \tag{G.7}$$

where $\boldsymbol{\zeta}_r = [\psi_r \ \phi_r]^T$, and ψ_r and ϕ_r are the elevation and azimuth angles, respectively, while d is the radius of the UCA. Furthermore, $\beta_m = \frac{2\pi i}{M} + \alpha$ is the angular position of the mth element on the UCA circle counted in an anti-clockwise manner from the x-axis and α is the offset angle. Moreover, f_s is the sampling frequency and c is the speed of sound. The TDOA model in (G.7) can be combined with the observation model in (G.5) to simplify the dimension of the estimation problem from MR to $2R$. The problem of interest is thus to estimate the unknown orientation parameters, i.e., ψ_r and ϕ_r, and the distance-related parameter, τ_r, based on the posed array (G.7) and observation (G.5) models. Additionally, a classification of the estimates as either belonging to an actual acoustic reflector or an empty space is needed to generate a spatial map of the acoustic reflectors. Finally, the parameter estimates of the acoustic echoes need to be mapped into the acoustic reflector positions based on the robots movement and orientation.

3 Non-Linear Least Squares (NLS) Estimator

We can resolve the problem of estimating the unknown parameters in (G.5), i.e., $\tau_{ref,r,k}$ and $\eta_{m,r,k}$, by using an NLS estimator, which is statistically optimal under the assumed white Gaussian noise conditions. Mathematically, this can be formulated as

$$\{\widehat{\mathbf{g}}_k, \widehat{\boldsymbol{\tau}}_k, \widehat{\boldsymbol{\zeta}}_k\} = \underset{\mathbf{g},\boldsymbol{\tau},\boldsymbol{\zeta},}{\arg\min} \sum_{m=1}^{M} \left\| \mathbf{y}_{m,k}(n) - \sum_{r=1}^{R} g_{m,r,k}\mathbf{s}(n - \tau_{ref,r,k} - \eta_{m,r,k}(\boldsymbol{\zeta}_r)) \right\|_2^2, \quad \text{(G.8)}$$

where

$$\boldsymbol{\tau} = \begin{bmatrix} \tau_1 & \tau_2 & \cdots & \tau_R \end{bmatrix}^T, \quad \text{(G.9)}$$

$$\mathbf{g} = \begin{bmatrix} \mathbf{g}_1^T & \mathbf{g}_2^T & \cdots & \mathbf{g}_R^T \end{bmatrix}^T, \quad \text{(G.10)}$$

$$\mathbf{g}_r = \begin{bmatrix} g_{1,r} & g_{2,r} & \cdots & g_{M,r} \end{bmatrix}^T, \quad \text{(G.11)}$$

$$\boldsymbol{\zeta} = \begin{bmatrix} \boldsymbol{\zeta}_1^T & \boldsymbol{\zeta}_2^T & \cdots & \boldsymbol{\zeta}_R^T \end{bmatrix}^T, \quad \text{(G.12)}$$

with $\widehat{x}$ denoting an estimate of x, and x_k denoting a parameter x related to the kth robot position. The displacement k of the robot can be estimated using an accelerometer or can be pre-programmed within the robot so that the robot follows a predefined trajectory. We can also solve (G.8) by converting it into a frequency domain because 1) it will reduce the computational load when estimating the desired parameters [G.42], and 2) by working in the frequency domain, we will have the flexibility to work in specific frequency ranges or account for frequency dependency. For instance, if we want to work in the ultrasonic range, we can select and utilize only the frequency bins corresponding to these high frequencies. This may also help us design probe signals that are non-intrusive to human hearing, but this is left for future iterations of this research. Using Parseval's theorem and omitting the frequency dependency in the notation, we can transfer (G.8) to the frequency domain, which yields the following:

$$\{\widehat{\mathbf{g}}_k, \widehat{\boldsymbol{\tau}}_k, \widehat{\boldsymbol{\zeta}}_k\} = \underset{\mathbf{g},\boldsymbol{\tau},\boldsymbol{\zeta}}{\arg\min}\, J(\mathbf{g}, \boldsymbol{\tau}, \boldsymbol{\zeta}), \quad \text{(G.13)}$$

where

$$J(\mathbf{g},\boldsymbol{\tau},\boldsymbol{\zeta}) = \sum_{m=1}^{M}\left\|\mathbf{Y}_{m,k} - \sum_{r=1}^{R} g_{m,r}\mathbf{Z}(\tau_r,\boldsymbol{\zeta}_r)\odot\mathbf{S}\right\|^2, \tag{G.14}$$

$$\mathbf{Y}_{m,k} = \begin{bmatrix} Y_{m,k}(0) & \cdots & Y_{m,k}(F-1)\end{bmatrix}^T, \tag{G.15}$$

$$\mathbf{Z}(\tau,\boldsymbol{\zeta}) = \begin{bmatrix} 1 & e^{-j(\tau+\eta(\boldsymbol{\zeta}))2\pi\frac{1}{F}} & \dots & e^{-j(\tau+\eta(\boldsymbol{\zeta}))2\pi\frac{F-1}{F}}\end{bmatrix}^T \tag{G.16}$$

with F denoting the number of frequency bins, $Y_{m,k}(f)$ denoting the DFT of $y_{m,k}(n)$ in frequency bin f. Moreover, $\mathbf{S}$ is the DFT vector of $\mathrm{s}(n)$ defined similarly to $\mathbf{Y}_{m,k}$. This estimation problem is multidimensional and, thus, computationally expensive in practice. To minimize the computational complexity, the multidimensional estimator could instead by implemented using various cyclic methods like the RELAX method proposed in [G.43] and later used in [G.44] to iteratively estimate the values of $\widehat{\boldsymbol{\tau}}_k$ and $\widehat{\mathbf{g}}_k$. In the special case where we are only concerned with estimating one acoustic reflection, and assuming that the direct-path component has been removed via preprocessing, we can set $R = 1$. Additionally, if we assume that the gain of each microphone is the same, then we can solve (G.13) for the gain $\widehat{g}_k$ by taking the derivative of the cost function, yielding:

$$\begin{aligned}\frac{\partial J(g_k,\tau_k,\zeta_k)}{\partial g_k} &= \frac{\partial}{\partial g_k}(\mathbf{Y}^H\mathbf{Y} - g_k\mathbf{Y}^H\overline{\mathbf{Z}}(\tau_k,\zeta_k) - g_k\overline{\mathbf{Z}}^H(\tau_k,\zeta_k)\mathbf{Y} + g_k^2\overline{\mathbf{Z}}^H(\tau_k,\zeta_k)\overline{\mathbf{Z}}(\tau_k,\zeta_k))\\ &= -\mathbf{Y}^H\overline{\mathbf{Z}}(\tau_k,\zeta_k) - \overline{\mathbf{Z}}^H(\tau_k,\zeta_k)\mathbf{Y} + 2g_k\overline{\mathbf{Z}}^H(\tau_k,\zeta_k)\overline{\mathbf{Z}}(\tau_k,\zeta_k) = 0, \end{aligned} \tag{G.17}$$

where $\overline{\mathbf{Z}}(\tau_k,\zeta_k) = \mathbf{Z}(\tau_k,\zeta_k)\odot\mathbf{S}$ is the frequency domain probe signal delayed by τ_k samples at angle ζ_k. Solving for the linear gain parameter g_k gives:

$$\widehat{g}_k = \frac{\mathbf{Y}_k^H\overline{\mathbf{Z}}(\tau_k,\zeta_k) + \overline{\mathbf{Z}}^H(\tau_k,\zeta_k)\mathbf{Y}_k}{2\overline{\mathbf{Z}}^H(\tau_k,\zeta_k)\overline{\mathbf{Z}}(\tau_k,\zeta_k)}. \tag{G.18}$$

By inserting this back into (G.13), we get

$$\widehat{\tau}_k = \arg\min_{\tau}\left\|\mathbf{Y}_k - \frac{\mathbf{Y}_k^H\overline{\mathbf{Z}}(\tau,\zeta_k) + \overline{\mathbf{Z}}^H(\tau,\zeta_k)\mathbf{Y}_k}{2\overline{\mathbf{Z}}^H(\tau,\zeta_k)\overline{\mathbf{Z}}(\tau,\zeta_k)}\overline{\mathbf{Z}}(\tau)\right\|^2 \tag{G.19}$$

$$= \arg\max_{\tau}\mathbb{R}\{\mathbf{Y}_k^H\overline{\mathbf{Z}}(\tau,\zeta_k)\} \tag{G.20}$$

where the operator $\mathbb{R}$ represents taking the real part of the signal. The expression in (G.20) estimates TOA for a single reflector at position, w_k. That is, for the special case with one acoustic echo, the NLS estimator in (G.20) can be interpreted as a cross-correlation based technique, which is widely used within robotics for source localization [G.45].

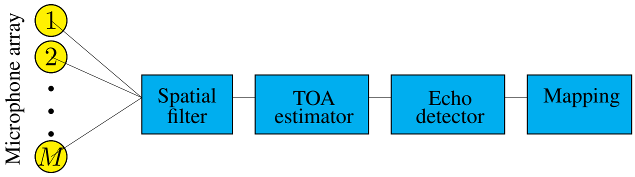


Fig. G.3: Proposed McLAM Architecture

Therefore, the problem at hand is to estimate the TOA, τ and DOA, ζ of an acoustic reflector that enables construction of acoustic reflectors given an arrat geometry. Based on the problem formulation and methods in Sections 3, we propose two algorithms that could aid different robotic platform for spatial map construction: a multichannel localization and mapping (McLAM) algorithm and a single-channel localization and mapping algorithm (ScLAM).

4 Multi-channel Localization and Mapping (McLAM)

When constructing a spatial map of an environment using sound, a robotic platform requires both DOA and TOA information of the acoustic echoes while distinguishing estimates belonging to an acoustic reflector from spurious estimates. Furthermore, this should be carried out under the presence of a strong direct-path component originating from the sound source, which detrimentally influences the estimation of the acoustic parameters. To address these problems, we propose mounting a microphone array on a robotic platform so that both the DOA and TOA of the acoustic echoes could be estimated, while suppressing the direct-path component. The McLAM architecture has four important components as shown in Fig. G.3. First, we introduce a spatial filter, i.e., a beamformer, to determine the DOAs of the acoustic echoes impinging from the reflectors, e.g., walls. Second, we feed the filtered observation into an NLS estimator to find the TOAs of the acoustic echoes. Then, we introduce a binary classifier to distinguish between spurious and real estimates, to exclude spurious estimates in the subsequent mapping of the acoustic reflectors, which constitutes the final block.

4.1 Spatial filter block

The DOA information of an acoustic echo can, for example, be determined using the traditional spatial filtering techniques, e.g., beamforming [G.31], as considered in this paper. Later, a TOA estimate technique is applied, so that acoustic echoes corresponding to the distance of acoustic reflectors are estimated. Apart from DOA estimation, the other advantage of using spatial filtering before TOA estimation is that it can suppress the direct-path component that can affect the parameter estimation. Beamforming is based on the spatial weighting of the signals recorded by a microphone array such that the output signal is the weighted summation of all the signals to extract the signal impinging from a particular DOA [G.46]. In this way, we first employ a beamformer for estimating the angle of an acoustic echo using the steered response power approach. Subsequently, the echo is extracted by applying a beamformer steered towards the estimated angle to produce the output signal for the later TOA estimation using an NLS estimator (G.20) [G.44]. Therefore, we seek to estimate the signal of interest (SOI), while minimizing the influence of the direct-path component of the probe sound and other noise sources, e.g., from the rotors of a drone. With this aim, we consider the use of an adaptive beamformer [G.31]. In addition, this idea also builds on the statistical foundation of the EM method [G.25], which indicates that this is the optimal way of solving the problem of localizing acoustic reflectors in an indoor environment.

Due to the broadband nature of the signals involved, we implement the beamformer in the frequency domain. Therefore, the observations in (G.1) were first converted into frequency domain as shown:

$$\begin{aligned}\mathbf{Y}_k &= \mathbf{X}_k + \mathbf{V}'_k \\ &= \mathbf{d}(\boldsymbol{\zeta}_{r,k})S_{r,k} + \mathbf{U}_k \\ &= \begin{bmatrix} Y_{1,k}(\omega) & Y_{2,k}(\omega) & \cdots & Y_{M,k}(\omega)\end{bmatrix}^T,\end{aligned} \tag{G.21}$$

where $\mathbf{X}_k$ and $\mathbf{V}'_k$ is defined similarly to $\mathbf{Y}_k$. Moreover, $\mathbf{U}_k$ contains the remaining $R-1$ early reflections as well as the late reverberation and background noise, and $S_{r,k}$ is the complex amplitude of the rth reflection at frequency ω. Assuming a UCA with the center of the array chosen as the reference point, the steering vector can be written as follows:

$$[\mathbf{d}(\boldsymbol{\zeta}_k)]_m = e^{-j\frac{\omega}{c}d\sin(\psi_k)\cos(\phi-\beta_m)}. \tag{G.22}$$

Here, $\boldsymbol{\zeta}_k$ is the look direction of the beamformer. The objective of the beamformer is then to recover the desired signal $S_{r,k}$ given the observation $\mathbf{Y}_k$, i.e.,

$$\overline{Y}_{\boldsymbol{\zeta}_k} = \mathbf{w}^H\mathbf{Y}_k, \tag{G.23}$$

where $\mathbf{w} \in \mathbb{C}^M$ and $\overline{Y}_{\boldsymbol{\zeta}_k}$ is the recovered signal from the observed signal from direction $\boldsymbol{\zeta}_k$ at position w_k, which should be an estimate of S_k. Here, several beamforming filters could be used, while, in this paper, we consider three types of beamformers which, facilitate a trade-off between computational efficiency, estimation accuracy, and direct-path component suppression. These are 1) the minimum power distortionless response (MPDR) beamformer, 2) the delay-and-sum (DSB) beamformer, and 3) the linearly constrained minimum variance (LCMV) beamformer [G.47]. The MPDR beamformer is derived by minimizing the power of the of the output of the beamformer $\overline{Y}_{\boldsymbol{\zeta}_k}$ subject to a distortionless constraint, i.e.,

$$\begin{aligned}\mathbf{w}_{\text{MPDR}} = &\arg\min \mathbf{w}^H\mathbf{R}_{Y_k}\mathbf{w} \\ &\text{subject to } \mathbf{w}^H\mathbf{d}(\boldsymbol{\zeta}_k) = 1.\end{aligned} \tag{G.24}$$

The solution to this is then well known to be given by [G.44]

$$\mathbf{w}_{\text{MPDR}} = \frac{\mathbf{R}_{Y_k}^{-1}\mathbf{d}(\boldsymbol{\zeta}_k)}{\mathbf{d}^H(\boldsymbol{\zeta}_k)\mathbf{R}_{Y_k}^{-1}\mathbf{d}(\boldsymbol{\zeta}_k)}, \tag{G.25}$$

where $\mathbf{R}_{Y_k} = E[\mathbf{Y}_k\mathbf{Y}_k^H]$ is the $M \times M$ covariance matrix of the observed signal, $E[\cdot]$ is the mathematical expectation operator, and $\mathbf{w}_{\text{MPDR}}$ is the complex weight vector corresponding to the MPDR beamformer. If the observed signal is assumed to be white Gaussian noise, e.g., $\mathbf{R}_{Y_k} = \mathbf{I}_M$, where $\mathbf{I}_M$ is the $M \times M$ identity matrix, the MPDR design resembles the DSB, i.e.,

$$\mathbf{w}_{\text{DSB}} = \frac{\mathbf{d}(\boldsymbol{\zeta}_k)}{M}. \tag{G.26}$$

Similarly, the LCMV beamformer is derived by extending the MPDR beamformer with additional constraints such that the optimization problem is solved as shown:

$$\mathbf{w}_{\text{LCMV}} = \arg\min \mathbf{w}^H\mathbf{R}_{Y_k}\mathbf{w} \quad \text{subject to } \mathbf{w}^H\mathbf{D} = \mathbf{f}^T. \tag{G.27}$$

Here, $\mathbf{D}$ is a matrix containing all the steering vector for the C different constraints in $\mathbf{f} \in \mathbb{R}^L$. In this paper, we choose $\mathbf{f}$ as

$$\mathbf{f} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T \tag{G.28}$$

By utilizing this, we can reject the interference of the direct-path component by introducing a null in the direction of the loudspeaker, i.e., the center of the UCA. The solution to the LCMV beamforming problem is

$$\mathbf{w}_{\text{LCMV}} = \mathbf{R}_{Y_k}^{-1}\mathbf{D}[\mathbf{D}^H\mathbf{R}_{Y_k}^{-1}\mathbf{D}]^{-1}\mathbf{f}. \tag{G.29}$$

4.2 TOA estimator block

The output of these beamformers is subsequently fed to the NLS estimator for TOA estimation in (G.20). This estimator is statistically optimal when estimating τ and g for a single reflection while the background noise is white Gaussian. By preprocessing the observation with the adaptive beamformer, this assumption is better met since we can reduce the impact of directional and colored noise [G.48]. The resulting NLS estimator is then given by

$$\{\widehat{g}_k, \widehat{\tau}_k, \widehat{\boldsymbol{\zeta}}_k\} = \underset{g,\tau,\boldsymbol{\zeta}}{\arg\min} \left\|\overline{Y}_{\boldsymbol{\zeta}} - g\mathbf{Z}(\tau) \odot \mathbf{S})\right\|^2, \tag{G.30}$$

where $\overline{Y}_{\boldsymbol{\zeta}}$ is the output of the beamformer (G.23) extracted from direction $\boldsymbol{\zeta}$ at a position w_k at frequency ω, while $\odot$ is the element-wise multiplication operator. By solving for the linear parameters in (G.30) by expanding and then taking the derivative of the expression with respect to the gain parameter g_k as shown in (G.17), (G.18) and (G.19), we get the concentrated estimator for the TOA and DOAs:

$$\{\widehat{\tau}_k, \boldsymbol{\zeta}_k\} = \arg\max_{\tau,\boldsymbol{\zeta}} \mathbb{R} \left\{ \overline{Y}_{\boldsymbol{\zeta}}^H \overline{\mathbf{Z}}(\tau) \right\} \tag{G.31}$$

where $\overline{\mathbf{Z}}(\tau) = \mathbf{Z}(\tau) \odot \mathbf{S}$ and the operator $\mathbb{R}$ represents taking the real part of the signal.

4.3 Echo detector block

If the robotic platform is expected to move autonomously based on echolocation, a significant problem will be to detect whether the observed signal received by the microphones represent an acoustic reflector, or if it only contains noise, e.g., the ego-noise of the robotic platform. This is because the TOA estimator in (G.31) provides estimates even when no acoustic reflector is present, which may lead to spurious localization estimates. To prevent false estimation when no acoustic reflector is present, several approaches could be applied including machine learning approaches [G.49], and deep learning [G.50] to categorize acoustic reflectors. Another approach could be to include a Generalized Likelihood Ratio Test (GLRT) detector [G.37] within our framework to distinguish whether the observed signal contains an acoustic reflection or not. Compared to the data-driven machine learning approaches, the GLRT is based on a priori model assumptions, and does thus not require training data. In this paper, we therefore employ the GLRT detection approach as discussed in the following.

If we assume the acoustic reflection to be in the far-field of the array, the decision about whether an observation contains an acoustic reflection can be formulated as a detection problem [G.37]:

$$\mathcal{H}_0 : \mathbf{y}_{m,k}(n) = \mathbf{v}'_{m,k}(n) \tag{G.32}$$

$$\mathcal{H}_1 : \mathbf{y}_{m,k}(n) = g_k \mathbf{s}(n - \tau_{m,k}) + \mathbf{v}'_{m,k}(n), \tag{G.33}$$

for $m = 1, \ldots, M$, where $\mathcal{H}_0$ is the null hypothesis referring to a situation when the observation only includes white Gaussian background noise and late reverberation, $v'_{m,k}(n)$, with variance σ^2, while $\mathcal{H}_1$ refers to the situation when the observation includes a reflected version of the known probe signal $s(n)$ in noise. Here, we assume that the direct-path component is absent, i.e., suppressed via preprocessing. The GLRT is then given by

$$\frac{p(\mathbf{y}_k; \widehat{g}_k, \mathcal{H}_1)}{p(\mathbf{y}_k; \mathcal{H}_0)} > \gamma, \tag{G.34}$$

$$\mathbf{y}_k = \begin{bmatrix} \mathbf{y}_{1,k}^T(0) & \cdots & \mathbf{y}_{M,k}^T(0) \end{bmatrix}^T, \tag{G.35}$$

It can then be shown that, in order to detect whether or not the observation belongs to $\mathcal{H}_1$, we

Algorithm 3: Proposed method McLAM.

```
Input  : Trajectory $\mathcal{W} = \{(w_{x_1}, w_{y_1}), \ldots, (w_{x_K}, w_{y_K})\}$;
Output: Reflector position estimates $\mathcal{P} = \{(p_{x_1}, p_{y_1}), \ldots, (p_{x_K}, p_{y_K})\}$;
Initialization:
    $\mathcal{P} = \{\}, \mathbf{DOA} = \{\}, \mathbf{TOA}, = \{\}, \mathbf{\Phi} = [0°; 360°]$;
for $k = 1, \ldots, K$ do
    Probe the environment with $\mathbf{s}(n)$;
    Record echoes in $\mathbf{y}_k$;
    Transform signals to frequency domain $\mathbf{s}(n), \mathbf{y}_k(n) \xrightarrow{\text{FFT}} \mathbf{S}, \mathbf{Y}_k$;
    for $\phi \in \mathbf{\Phi}$ do
        Compute $\mathbf{w}$, e.g., using (G.24);
        $\overline{Y}_{\phi,k}(\omega) = \mathbf{w}^H \mathbf{Y}_k$;
    $\{\widehat{\tau}_k, \widehat{\phi}_k\} = \arg\max_{\tau,\phi} \mathbb{R}\left\{\overline{Y}^H_{\phi,k} \overline{\mathbf{Z}}(\tau_k)\right\}$;
    $\widehat{\phi}_k \xrightarrow{\text{update}} \mathbf{DOA}$;
    $\widehat{\tau}_k \xrightarrow{\text{update}} \mathbf{TOA}$;
    Apply the echo detector in (G.36);
    if $\mathbf{y}_k^H(n)\mathbf{H}(\boldsymbol{\tau}_k)\mathbf{s}(n) > \widehat{g}_k \frac{\epsilon}{2} + \frac{\sigma^2 \ln \gamma}{2\widehat{g}_k}$ then
        Compute $p_k$ using (G.40);
        $p_k \xrightarrow{\text{update}} \mathcal{P}$;
```

can use a threshold that depends on the power of the attenuated probe signal, the noise variance, and γ. If the power, $T(\mathbf{y}_k)$, of a matched filtering between the probe signal and the observed signal at the reference microphone exceeds this threshold, we decide $\mathcal{H}_1$, i.e., if

$$T(\mathbf{y}_k) = \mathbf{y}_k^H(n)\mathbf{H}(\boldsymbol{\tau}_k)\mathbf{s}(n) > \widehat{g}_k \frac{\epsilon}{2} + \frac{\sigma^2 \ln \gamma}{2\widehat{g}_k} \tag{G.36}$$

with

$$\mathbf{H}(\boldsymbol{\tau}_k) = \begin{bmatrix} \mathbf{D}^T_{\tau_{1,k}} & \cdots & \mathbf{D}^T_{\tau_{M,k}} \end{bmatrix}^T, \tag{G.37}$$

$$\epsilon = M\|\mathbf{s}(n)\|^2, \tag{G.38}$$

$$\widehat{g}_k = \frac{2\mathbf{y}_k^H(n)\mathbf{H}(\boldsymbol{\tau}_k)\mathbf{s}(n)}{M\|\mathbf{s}(n)\|^2}, \tag{G.39}$$

where $\mathbf{D}_\tau$ is a cyclic shift register that delays a signal by τ samples.

4.4 Mapping block

In this block, the DOA and TOA estimates are used alongside the robot's position within an environment to localize the position of an acoustic reflector. The aspect of the robot's navigation and path planning is beyond the scope of this paper, however, by utilizing common on-board sensors, e.g., Inertial Measurement Units (IMUs), of the robotic platform, we can estimate the robot's position. By combining this information with the estimates of the acoustic echoes obtained using for example, the methods considered in this paper, a spatial map of the environment can be generated for the robotic platform. The resulting spatial map may then enable the robotic platform to plan its path and move autonomously within the environment.

To estimate the position of the acoustic reflector from the estimated TOA, $\widehat{\tau}_k$, we assume that the sound propagates in plane waves (i.e., the source is in the far-field of the array). If we assume the speed of sound to be fixed then the distance of the acoustic reflector with respect to the robotic platform is estimated as $\delta_k = \frac{c \cdot \tau_k}{2}$. Additionally, the direction of the acoustic reflector at position w_k is determined from the DOA estimates ψ and ϕ. In a $2D$ scenario, where the reflections and the hardware are located in the same plane, we can utilize the far-field assumption and the choice of our reference point to conduct the mapping as:

$$\begin{aligned} p_{x_k} &= w_{x_k} + \delta_k \cos \phi_k \\ p_{y_k} &= w_{y_k} + \delta_k \sin \phi_k \end{aligned} \tag{G.40}$$

After this, the procedure is to estimate the acoustic reflector positions for each of the known robot positions, w_k, along its trajectory. The estimated acoustic reflector positions are then concatenated in the set $\mathcal{P} = \{p_1, \ldots, p_K\}$ with $p_k = (p_{x_k}, p_{y_k})$ for $k = 1, \ldots, K$. The spatial filtering, the TOA estimator, the echo detector and the mapping block are then combined to form the basis of our proposed McLAM method. The algorithm describing the proposed McLAM method is outlined in Algorithm 3. However, in some applications, only one microphone and loudspeaker pair may be available for the mapping. In the following section, we therefore consider, how the hardware directivity properties may be exploited to localize the acoustic reflectors.

5 Single Channel Localization and Mapping (ScLAM)

In some applications, robotic platforms, such as those intended for HRI, may consist of only a single loudspeaker and microphone. In such a scenario, it is therefore necessary to reduce the McLAM algorithm to a single-channel localization and mapping (ScLAM) algorithm. The ScLAM algorithm was proposed and evaluated in our previous published work [G.12]. However, using such a single-channel approach has certain limitations. For instance, it cannot generally be used to estimate the DOA of the acoustic echoes because of the lack of spatial information. Some possible ways of combating this are to exploit the movement of the robot [G.19], or, as considered in this paper, to exploit the directionality of the employed hardware [G.12].

```
Algorithm 4: Proposed method ScLAM.
input   : Trajectory $\mathcal{W} = \{(w_{x_1}, w_{y_1}), \dots, (w_{x_K}, w_{y_K})\}$, Initialization $\mathcal{P} = \{\}$,
          $\mathbf{TOA}, = \{\}$;
output  : Reflector position estimates $\mathcal{P} = \{(p_{x_1}, p_{y_1}), \dots, (p_{x_K}, p_{y_K})\}$;
for $k = 1, \dots, w_k$ do
    Acquire direction of robot movement: $\theta_{r,k}$;
    Acquire direction of loudspeaker: $\theta_{l,k}$;
    Probe the environment with $\mathbf{s}(n)$ ;
    Record echo: $\mathbf{y}_k$;
    Transform signals to frequency domain $\mathbf{s}(n), \mathbf{y}_k(n) \xrightarrow{\text{FFT}} \mathbf{S}, \mathbf{Y}_k$;
    $\widehat{\tau}_k = \arg\max_{\tau_k} \mathbb{R}\{\mathbf{Y}_k^H \overline{\mathbf{Z}}(\tau)\}$;
    $\{\widehat{\tau}_k\} \xrightarrow{\text{update}} \mathbf{TOAs}$;
    Apply the echo detector in; (G.36);
    if $\mathbf{y}_k^H(n)\mathbf{H}(\boldsymbol{\tau}_k)\mathbf{s}(n) > \widehat{g}_k \frac{\epsilon}{2} + \frac{\sigma^2 \ln \gamma}{2\widehat{g}_k}$ then
        $\tau_k \xrightarrow{\text{remove}} \mathbf{TOAs}$;
        $p_k$ using (G.43) $\xrightarrow{\text{update}} \mathcal{P}$;
```

As with the McLAM, the loudspeaker probes the room with a known sound, $s(n)$, which is recorded by a microphone as the robot moves via positions w_k, for $k = 1, \dots, K$. The NLS estimator described in (G.20) estimates τ_k for every robot position, w_k. Consider the platform moving in a predefined trajectory $\mathcal{W} = \{w_1, \dots, w_K\}$ with $w_k = (w_{x_k}, w_{y_k})$, such that the platform moves from w_k to w_{k+1} etc. Therefore, for every position, w_k, the platform will probe the environment with $\mathbf{s}(n)$ and record the observed signal $\mathbf{y}_k(n)$. The probed and observed signals are then converted into the frequency domain before passing them to the NLS estimator. In practice, the analysis window for the TOA could be restricted to a search interval from $\tau_{\min}$ up to $\tau_{\max}$ samples. This leads to

$$\widehat{\tau}_k = \underset{\tau \epsilon [\tau_{\min}; \tau_{\max}]}{\arg\max} \; \mathbb{R}\{\mathbf{Y}_k^H \overline{\mathbf{Z}}(\tau)\} \tag{G.41}$$

In ScLAM, the position of the acoustic reflector is then inferred from the estimated TOA, $\widehat{\tau}_k$, by exploiting the typical directionality of a loudspeaker. More specifically, we assume the acoustic reflector to be located at the distance corresponding to the estimated τ_k in the direction of the loudspeaker. Additionally, the direction in which the robot platform is moving, $\theta_{\text{rob},k}$, at position w_k, is related to the direction that the loudspeaker is facing, θ_{l_k}, by a fixed offset angle, $\Delta\theta$, i.e.,

$$\theta_{l_k} = \theta_{\text{rob},k} + \Delta\theta. \tag{G.42}$$

Based on the above information, the coordinates of the position of the acoustic reflector are then

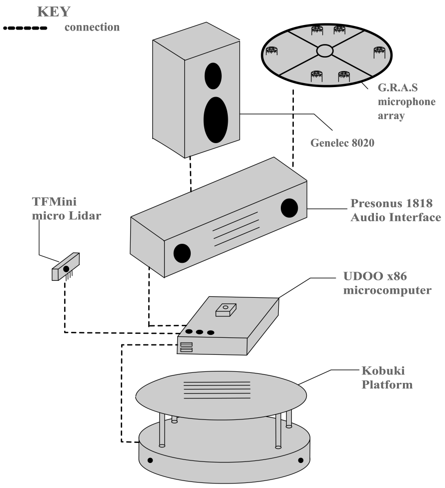


Fig. G.4: An overview of components required to built a multichannel robotic platform used for this research

estimated as follows:

$$\begin{aligned} p_{x_k} &= w_{x_k} + \delta_k \cos\theta_{l_k}, \\ p_{y_k} &= w_{y_k} + \delta_k \sin\theta_{l_k}. \end{aligned} \tag{G.43}$$

The resulting ScLAM algorithm is then proposed in Algorithm 4, which can be used to construct a spatial map of a 2D environment with a single-channel loudspeaker/microphone setup.

6 Robotic Platform Overview

The proposed methods discussed in Section 4 and Section 5 were implemented on an embedded platform running a Windows 10 Operating System. The microcomputer used for the proof-of-concept robotic platform is an UDOO x86, which is a single board development platform. On the platform, we used MATLAB to implement the proposed McLAM and ScLAM methods in Algorithm 3 and 4, respectively. Moreover, for multichannel audio data acquisition, the Playrec [G.51] was used to probe and record the acoustic signals. The base of the robot used for moving the microphone and loudspeaker array, as shown in Fig. G.4, is a Kobuki (TMR-K01-W1), which is a wheeled platform with on-board sensors such as an accelerometer, an odometer, etc., for precise control and movement. The Kobuki platform has a built-in microcontroller (Arduino) that can be programmed with a predefined trajectory to conduct experiments. The microphone and loudspeaker array is connected to a Presonus (1818VSL) audio interface, which was subsequently connected to the UDOO x86 microcomputer. The sampling frequency of the

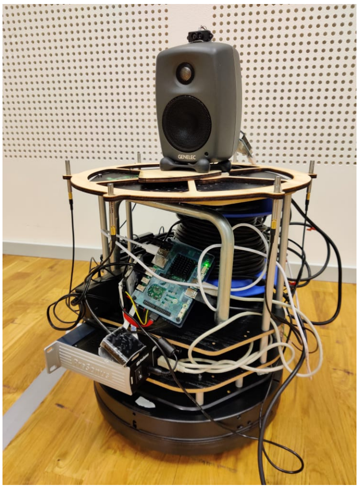


Fig. G.5: The multi-channel proof of concept robotic platform.

audio interface was set to $48,000$ Hz. Furthermore, a pre-calibrated laser range sensor (TFMini micro Lidar), was also attached to an external microcontroller (Arduino Uno) which was then connected to the UDOO microcomputer. This was done to receive a ground truth distance value for the experiments. The laser range finder helped in evaluating the performance of the proposed method at varying distances under different noise conditions. The recorded data was processed by the UDOO x86 microcomputer in real-time as the robot was moving along its trajectory. The final assembly is shown in Fig. G.5 where the microphone and loudspeaker array is attached on top of the Kobuki base. The microphones are organized as a UCA with a radius of 0.2 m.

7 Simulated Results

In this section, we evaluate the performance of the proposed method presented in the earlier sections. We evaluate the performance of the ScLAM and McLAM using simulation data and later, implement the methods on a proof-of-concept hardware platform that was built to test the proposed method in a lab setting. In the first simulated experiments the performance of the considered TOA/DOA estimators in terms of their accuracy were evaluated and compared against existing methods under different background noise levels. Additionally, the presence of the direct-path component on the effect of TOA/DOA estimation was also evaluated. Similarly, in the second experiment, we evaluated the TOA/DOA accuracy of the proposed methods against varying distances from the acoustic reflector. The simulated experiments were conducted using the room impulse response generator [G.52]. The dimension of the simulated room was set to $8 \times 6 \times 5$ m., the reverberation time (T_{60}) was set to 0.6 s, while the speed of sound was fixed at 343 m/s. The loudspeaker was positioned at the center of an UCA with a radius of 0.2 m and $M = 6$ microphones. A white Gaussian noise sequence was used as the known probe signal, $s(n)$, consisting of $1,500$ samples from a Gaussian distribution. Using such a broadband signal minimizes the effect of spatial aliasing [G.53] and was also used in [G.25] to simplify the EM estimator. However, any type of known broadband signal could be used to probe the environment, such as a chirp signal or a maximum length sequence (MLS) [G.54]. Additionally, we have zero-padded the probe signal to get a total length of $20,000$ samples in order to we get a longer analysis window which will ensure that all of the reflections are captured in the observed signal. The sampling frequency f_s was set to $48,000$ Hz. The background noise for the evaluation was composed of three components: a cylindrical diffuse noise $\mathbf{e}_{m,k}$, the sensor noise, $\mathbf{f}_{m,k}$, and an interfering source, $\mathbf{i}_{m,k}$, e.g., external and directional noise source. The diffuse cylindrical noise was generated using the method in [G.55] with the rotor noise of a drone from the DREGON database [G.56]. The audio file used to generate the cylindrical noise has a rotor speed of 70 revolutions per second (RPS). The thermal sensor noise was simulated as a white Gaussian noise while the interfering source is modelled as a point source. These noises were then added to the observed probe signal before estimating the parameters of interest from the observations, which can be mathematically written as:

$$\mathbf{y}_{m,k}(n) = \mathbf{x}_{m,k}(n) + \mathbf{v}'_{m,k}(n), \tag{G.44}$$

$$= \mathbf{x}_{m,k}(n) + \mathbf{e}_{m,k}(n) + \mathbf{f}_{m,k}(n) + \mathbf{i}_{m,k}(n). \tag{G.45}$$

The noise was added to achieve certain signal-to-diffuse noise ratios (SDNR's), signal-to-sensor noise ratios (SSNR's), and signal-to-inteference noise ratios (SINR's). These are defined, for the microphones $m = 1, \dots, M$, as

$$\mathrm{SDNR}_m = \frac{\sigma^2_{x_m}}{\sigma^2_{e_m}}, \tag{G.46}$$

$$\mathrm{SSNR}_m = \frac{\sigma^2_{x_m}}{\sigma^2_{f_m}}, \tag{G.47}$$

$$\mathrm{SINR}_m = \frac{\sigma^2_{x_m}}{\sigma^2_{i_m}}, \tag{G.48}$$

where σ_y^2 denotes the variance, $\sigma_y^2 = E[y^2(n)]$ of a zero-mean signal $y(n)$. In the following experiments, we then compared our proposed method with existing TOA/DOA methods found in the literature. This included the multi-channel expectation-maximization method (EM-UCA) method proposed in [G.25] and the common approach to extracting TOAs from the estimated RIR using dual-channel method [G.57] through the peak-picking approach (RIR-PP). This is done by computing $\widehat{H}(f) = Y(f)/S(f)$ and then taking the inverse DFT to get $\widehat{h} = \mathcal{F}^{-1}\{\widehat{H}(f)\}$. These methods were compared with different variations of the proposed beamforming and NLS-based approach, utilizing DS (DS-NLS), MPDR (MPDR-NLS), and LCMV (LCMV-NLS) beamforming, respectively. Moreover, the ScLAM algorithm [G.12] was also used to make the comparison.

Although the proposed methods can be extended and applied to $3D$ scenarios, we focus on the construction of $2D$ maps in our experiments and therefore set $\psi = 0$. The generalization to $3D$ is left for future research. In contrast to earlier works in [G.44, G.25], the direct-path component is accounted for and thus included within the simulations. Within the experiments, we assume that the robotic platform is closer to one acoustic reflector. Therefore, we choose $R = 1$ to estimate the TOA and the DOA of the nearby acoustic reflector. In order to estimate multiple reflections $R > 1$, we can adopt several iterative methods, such as, RELAX and EM method [G.43, G.58] but this method will be left for future work.

7.1 Implementation of the proposed DOA estimator

To implement the beamformers, we used the overlap-add technique [G.31]. The output of the microphone was divided into overlapping frames with a frame width of 960 samples (20 ms with a sampling rate of 48 kHz) with a window overlap of 50 %. Later, each frame is multiplied with a Hanning window. These frames are then transformed using a short-time Fourier transform (STFT). For each frequency bin, a beamformer was designed and applied to the received signals $\mathbf{Y}_k$. Furthermore, for each sub-band, the observed signal covariance matrix, needed in forming the MPDR and LCMV beamformers, is estimated as

$$\mathbf{R}_{Y_k} = \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{Y}_k\mathbf{Y}_k^H. \tag{G.49}$$

Moreover, to make the beamformers robust against miscalibration and reverberation, as example, we regularized the covariance matrix of the observed signal as in [G.59]

$$\overline{\mathbf{R}}_{Y_k} = (1-\beta)\mathbf{R}_{Y_k} + \beta\frac{\text{Tr}\{\mathbf{R}_{Y_k}\}\mathbf{I}}{M} \tag{G.50}$$

where β is the regularization parameter, Tr$(\cdot)$ is the trace of a matrix, and $\mathbf{I}_M$ is the $M \times M$ identity matrix. When evaluating the performance of our estimator, a value of $\beta = 0.1$ was selected for the MPDR beamformer. The noise covariance matrix, $\mathbf{R}_{Y_k}$, in (G.49) is then replaced by the regularized noise variance matrix (G.50), $\overline{\mathbf{R}}_{Y_k}$. For the LCMV beamformer, we added an additional regularization using γ to mitigate poor matrix conditioning for certain constraint and frequency combinations. This was done as $\mathbf{w}_{\text{LCMV}} = \mathbf{R}_{Y_k}^{-1}\mathbf{D}[\mathbf{A}(\gamma)]^{-1}\mathbf{f}$, where

$$\mathbf{A}(\gamma) = (1-\gamma)\mathbf{D}^H\mathbf{R}_{Y_k}^{-1}\mathbf{D} + \gamma\frac{\text{Tr}\{\mathbf{D}^H\mathbf{R}_{Y_k}^{-1}\mathbf{D}\}\mathbf{I}}{M}. \tag{G.51}$$

Values of $\gamma = 0.1$ and $\gamma = 1$ were selected empirically and used in the simulations. To initiate the method, we probed the environment with a known sound. The observed signals recorded by the microphone array were first processed to determine the DOA of the acoustic echoes. To estimate the DOA and the TOA of the acoustic echoes, a uniform grid of DOAs over the interval $[0°; 360°]$ and a uniform grid of TOAs corresponding to a distance interval from 0.5 m up to 3 m

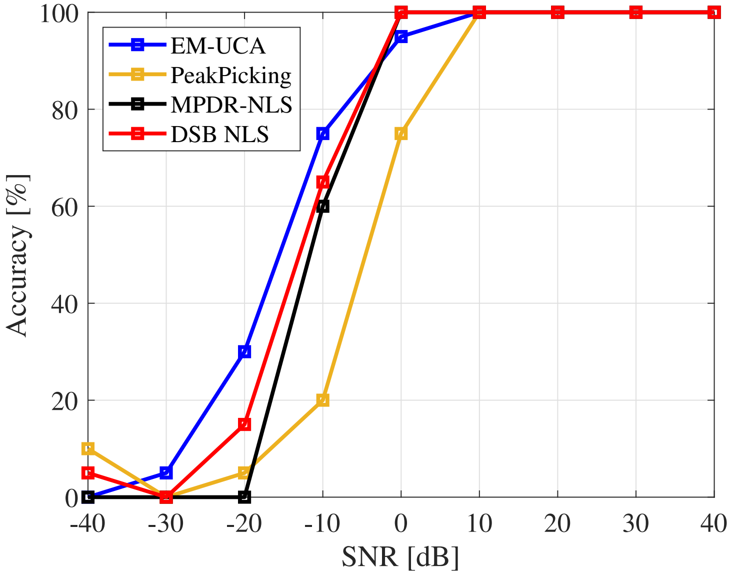
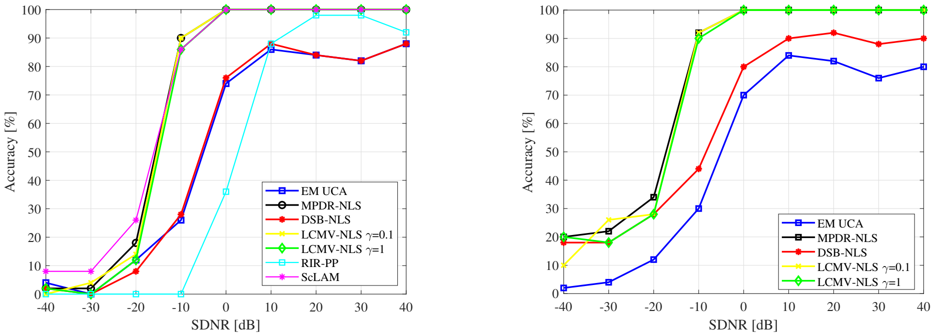


Fig. G.6: TOA evaluation in the absence of direct-path component

were considered. The estimators were then evaluated over these grids of candidate DOAs and TOAs. The reason for selecting 0.5 m as the lower bound was done to search for acoustic echoes that were outside the UCA which has a radius of 0.2 m, and so that the direct-path component was not included within the estimation window. Moreover, the upper bound of 3 m was selected because the performance of the proposed method degrades after 3 m according to [G.25].

7.2 Comparison of the proposed methods

In our first experiment, we compared our proposed method with the existing TOA/DOA methods. We compared the proposed methods against different SDNRs with and without the presence of the direct-path component while placing the setup at a distance of 1 m close to an acoustic reflector for Fig. G.7 and placing the other setup at the corner of the room with a distance of 1 m from one wall and 1.5 m from another wall for Fig. G.6. The performance of the proposed methods is shown in Fig. G.6 and Fig. G.7. The accuracy is defined as a percentage of the estimated TOAs that are within ± 10 $\%$ of the true TOA/DOA parameter of the first order acoustic echo computed using the image-source method [G.60]. This was measured for different SDNRs while the SSNR was fixed to 40 dB and the interfering source was absent in this experiment. For each SDNR value, the accuracy was measured over 50 Monte-Carlo simulations.In the absence of a direct-path component, all of the methods provided an estimate at -10



(a) TOA evaluation of the proposed method and state-of-the-art

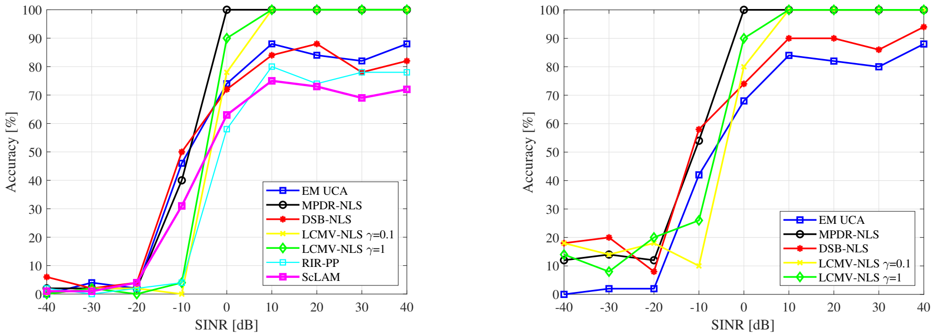
(b) DOA evaluation of the proposed method and state-of-the-art

Fig. G.7: Comparison of the proposed method against state-of-the-art

dB with around 60% to 70% accuracy while RIR-PP is only 20% accuracy. But as seen in Fig. G.7, the proposed methods, MPDR-NLS and LCMV-NLS, outperforms the existing TOA/DOA methods, EM-UCA and RIR-PP, in the presence of a direct-path component, for SDNR levels greater than -10 dB. The DSB-NLS method offers similar performance to EM-UCA both in terms of TOA and DOA estimation for most SDNRs as seen in Fig. G.7(b).

7.3 Evaluation of the proposed method in the presence of a point source interference

In this experiment, we investigated a scenario where the robot is placed within an environment in the presence of an external interfering source, e.g, a human-speaker, machinery, a radio, etc. In such a scenario, the proposed method will be affected from the external elements present in the environment. Therefore, the objective of this experiment is to evaluate both the TOA and the DOA performance of the proposed method against different SINR values. The interfering source was modelled as a point source for this experiment. More specifically, within this experiment, the robotic platform was placed close to an acoustic reflector at a position, $[1, 3, 2.5]$ m within an environment of dimension $8 \times 6 \times 5$ m. Furthermore, the external interfering point source was positioned at a location $[2, 1, 2.5]$ m such that the acoustic reflector was at a fixed angle of $180°$ while the point source was placed at an angle of $300°$ with respect to the robotic platform. The performance is shown in Fig. G.8. The SINR level selected for this experiment is within the interval $[-40; 40]$ dB while the SDNR and SSNR were both set to 40 dB. Moreover, some additional consideration was taken into account when modelling the interfering point source. For instance, if a human talker is considered as a point source, then it is natural for the human to move within the environment. To model this, the position of the point source was randomize in both the x-axis and y-axis. The interval selected to model the point source movement for



(a) TOA evaluation

(b) DOA evaluation

Fig. G.8: Evaluation of proposed McLAM method and state-of-the-art against different SINR

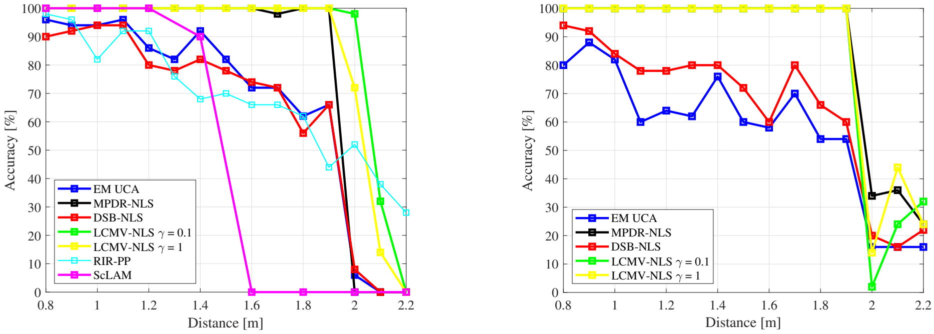
both x-axis and y-axis are $[1; 3]$ m and $[1; 2]$ m, respectively. As seen in Fig. G.8, the TOA of the MPDR-NLS and LCMV-NLS offers more robustness at a low SINR compared to EM-UCA, RIR-PP and DSB-NLS. A similar performance was seen in the DOA estimation. The accuracy is defined similarly to the previous method with a tolerance of $\pm 10\%$ of true TOA and DOA. Furthermore, the ScLAM method provided TOA estimates with around 70% at higher SINR values.

7.4 Evaluation of proposed methods against distance

In this experiment, we considered a scenario where the robotic platform was placed closer to an acoustic reflector and its distance with respect to the acoustic reflector was changed after every 50 iterations. With this setup, the performance of the proposed method and existing methods over distance interval $[0.8; 2.2]$ m was investigated. Here, the SDNR and SSNR values were set to 40 dB while the interfering source was absent. As seen in Fig. G.9, the MPDR-NLS, and the LCMV-NLS variants outperformed other methods in terms of TOA estimation and accurately estimate the DOA of the acoustic reflector as it can detect an acoustic reflector up to a distance of around 2 m. However, a distance of $1, 5$ m was estimated using ScLAM. This is because at larger distance the acoustic echoes loses its energy quadratically due to inverse square law.

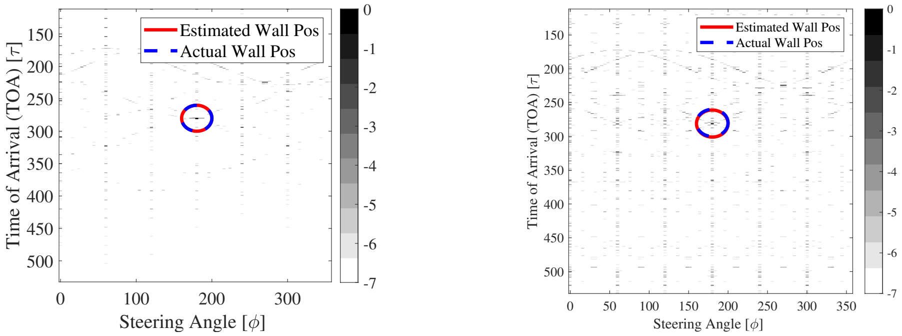
7.5 Visualizing acoustic echoes

Microphone array imaging has been around for quite some time and is used in aviation [G.61] for structural analysis as well as to study low frequencies [G.62]. Similarly, our proposed method could also be used to generate an acoustic image of acoustic echoes which could aid researchers in analyzing the direction and distance of acoustic reflectors or be used as input data for the development of deep learning based methods. To generate an acoustic image using



(a) TOA evaluation **(b)** DOA evaluation

Fig. G.9: Evaluation of proposed McLAM method and state-of-the-art against different distances



(a) MPDR-NLS costfunction **(b)** DSB-NLS costfunction

Fig. G.10: Acoustic image of MPDR-NLS and DSB-NLS beamformer

the methods proposed in this paper, we considered an estimation of the reflector in $2D$ only, i.e., we only estimated ϕ. For each beamformer we considered a grid of candidate steering angles with a resolution of 4° in the interval $[0°; 360°]$. The output of the beamformer was then passed to the NLS estimator in (G.20), which then estimated τ from candidate grid of delays in $[\tau_{min}; \tau_{max}]$. The resulting $2D$ cost functions are shown in Fig. G.10(a) and (b), respectively for one of these experiments. Both plots in Fig. G.10(a) and (b) were generated at an SINR of 40 dB with the observed signal including the direct-path component. As seen, the cost function of the MPDR beamformer Fig.G.10(a) shows a peak at times and angles corresponding to the TOAs and the DOAs at which the beamformer received the acoustic echo, these regions are marked by a red circle. In comparison, the DSB cost function in Fig.G.10(b) was very noisy, despite being evaluated under the high SINR of 40 dB, which made it difficult to extract the true

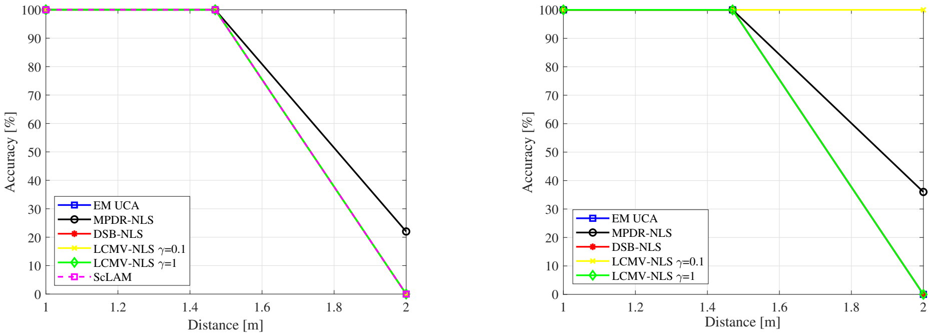
TOAs and DOAs. This was partly caused by the presence of the direct-path component which could not be sufficiently suppressed by the DSB.

7.6 Computational cost

The computational cost of the proposed methods were measured using MATLAB's built-in function *timeit*. These were tested on a standard desktop computer running a Microsoft Windows 10 operating system with an Intel Core i7 CPU with a 3.40 GHz processing speed and 16 GB of RAM. A Monte Carlo Simulation of 50 trials was conducted and an average time was calculated. The measured computational time of EM-UCA, RIR-PP, LCMV-NLS and MPDR-NLS are 63.25 s, 0.024, 59.75, and 60.65 s, respectively, for $R = 1$ for SINR = 40 dB. The proposed algorithms are computationally expensive when implemented within a robotic platform compared to lidar technologies. This is partly due to the choice of hardware used to process the acoustic signals. UDOO is a low end microcomputer with limited memory and processing power available. Replacing UDOO with a faster processor could enable faster processing. Another way to optimize the McLAM algorithm is to probe the environment and use echo detector first to determine whether the robot is closer to an acoustic reflector before proceeding with the proposed DOA/TOA estimates. This accelerate processing and prevents the robot from estimating the parameters when not in the presence of an acoustic reflector. Moreover, tracking, e.g., in the form of gradient searches, may be employed instead of performing a full grid search for every new robot position.

8 Experiments using Proof-of-Concept Robotic Platform

In this section, we evaluated the performance of the proposed algorithm (McLAM) using a robotic platform under different SINRs and distances. The objective of these experiments was to compare our simulated data with real data to test the performance of the proposed method in real scenarios. Two sets of experiments were conducted using the proof-of-concept robotic platform described earlier in Section 7. The first set of experiments were performed under different SINR and distances while the second set of experiments were performed as qualitative test to show the mapping ability of the robotic platform while comparing the MPDR-NLS algorithm against the lidar data (ground truth). The data is also summarized in Table G.1 and Table G.2. In both environments, the proof-of-concept robotic platform stops momentarily for 3 seconds before proceeding to the next location. During these 3 seconds, the robot probes the environment with a known signal, $s(n)$, and then use the recorded signal to determine the location of an acoustic reflector.



(a) TOA evaluation

(b) DOA evaluation

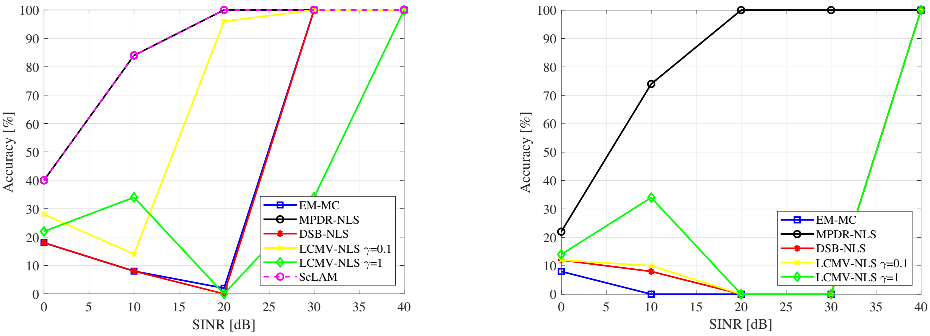
Fig. G.11: Evaluation of proposed McLAM method and state-of-the-art against varying Distances

8.1 Evaluation of the proposed method against different SINRs and distances

Similar to the experiments performed in Section 7.2, the proof-of-concept robotic platform was placed against an acoustic reflector within Aalborg University's Sound Lab. In the first part of the experiment, the robotic platform was placed at varying distances while the SINR value was set fixed to 40 dB. The platform was placed at an interval of $[1, 1.5, 2]$ m. At each distances, the robotic platform probed the environment with a known sound and 50 samples were collected at each distances. The TOA/DOA obtained from the robotic platform are shown in Fig. G.11. As seen in the figures, the proposed McLAM algorithm gives an accurate TOA estimate up to a distance of 1.5 m for all combinations of spatial filters. The accuracy is defined as the number of estimates that are $\pm 10\%$ of the true TOAs obtained from the lidar data: DOA accuracy is defined similarly.

The next experiment was performed to test the proposed method against different SINR values of the environment. The SINR value of the environment was changed by using a separate loudspeaker playing an audio file from YouTube called Cocktail party[2]. The loudspeaker was placed 6.3 m away from the robotic platform while the robotic platform was fixed at a distance of 1 m away from the acoustic reflector. The SINR of the environment was estimated by dividing the variance of the probed signal, σ_x^2, with the variance of the background noise, σ_v^2. The background noise $\mathbf{v}(n)$ was recorded by the robot before probing the environment. By tuning the volume of the loudspeaker, we then selected 5 SINR values, $[0, 10, 20, 30, 40]$ dB. The results for this experiment are shown in Fig. G.12. Here, we see that the proposed MPDR-NLS is robust under low SINR value of 10 dB for both TOA and DOA estimation with 80% accuracy. The changes seen in these experiments are discussed in Section 9.

[2] https://youtu.be/IKB3Qiglyro



(a) TOA estimation

(b) DOA estimation

Fig. G.12: Evaluation of the proposed method using proof of concept robotic platform against different SDNR

LIDAR = 1 m	SINR = 0 dB		SINR = 10 dB		SINR = 20 dB		SINR = 30 dB		SINR = 40 dB	
Methods	μ **[m]**	**RMS error [m]**	μ **[m]**	**RMS error [m]**	μ **[m]**	**RMS error [m]**	μ **[m]**	**RMS error [m]**	μ **[m]**	**RMS error [m]**
MPDR-NLS	1.0558	0.2843	0.9797	0.0992	0.9718	0.0281	0.9890	0.0118	0.9861	0.0138
DSB-NLS	1.0231	0.2802	0.8647	0.8896	0.1103	0.1174	0.9075	0.0924	0.9861	0.0138
EM-UCA	1.0229	0.2803	0.8647	0.1485	0.8908	0.1094	0.9075	0.0924	1.0040	0.0039
LCMV-NLS $\gamma = 0.1$	0.9899	0.2603	0.8758	0.1647	1.0387	0.0556	1.0647	0.0647	0.9861	0.0138
LCMV-NLS $\gamma = 1$	1.0325	0.2819	0.8813	0.1409	0.7996	0.2134	0.8084	0.2042	1.0254	0.0254
ScLAM	1.0774	0.0832	1.0888	0.0859	1.0847	0.0766	1.0977	0.2042	1.067	0.200

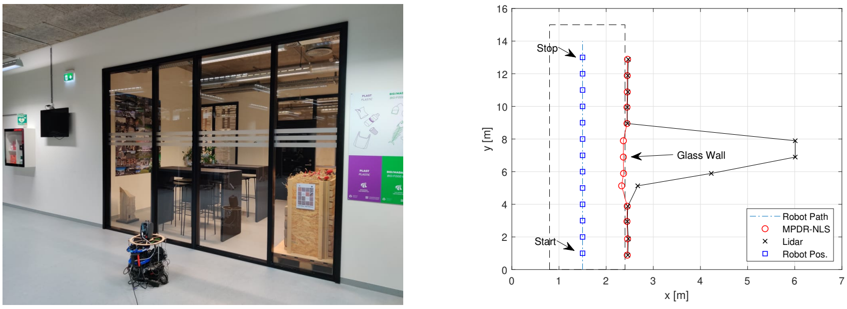
Table G.1: Evaluation of the proposed McLAM and ScLAM against ground truth and SDNRs

8.2 Application examples

Two qualitative experiments were performed to test the performance of the proposed method (MPDR-NLS) in constructing a spatial map of an indoor environments. Two environments were selected to perform this task: 1) a typical office environment with a glass partition and 2) Aalborg University's Sound Lab. These experiments are similar to the one performed in our earlier work with ScLAM [G.12]. In the first experiment, the McLAM algorithm was used to move within an office environment in a predefined trajectory (straight line). The objective of this experiment was to compare the proposed method against lidar, e.g., in detecting a glass surface. The robot moved a distance of 0.5 m and stopped momentarily to probe the environment with a known sound before moving to a new location. The robot repeated this process for $k = 1, \dots K$,

SINR = 40 dB	**MPDR-NLS**		**DSB-NLS**		**EM-MC**		**LCMV-NLS gamma = 0,1**		**LCMV-NLS gamma = 1**		**ScLAM**	
LiDAR	**mean [m]**	**RMSE [m]**	**mean [m]**	**RMSE [m]**	**mean [m]**	**RMSE [m]**	**mean [m]**	**RMSE [m]**	**mean [m]**	**RMSE [m]**	**mean [m]**	**RMSE [m]**
1,01	0,9861	0,0138	0,9861	0,0138	1,004	0,0039	0,9861	0,0138	1,0254	0,00254	1,0977	0,1013
1,47	1,4327	0,0387	1,5899	0,1199	1,4542	0,0158	1,4327	0,0372	1,5899	0,1199	1,4523	0,2701
2,0	1,4480	0,6142	0,7610	1,239	0,7610	1,2390	1,3321	0,6716	1,2734	0,8174	1,1863	1,3594

Table G.2: Evaluation of the proposed McLAM and ScLAM against ground truth and distances



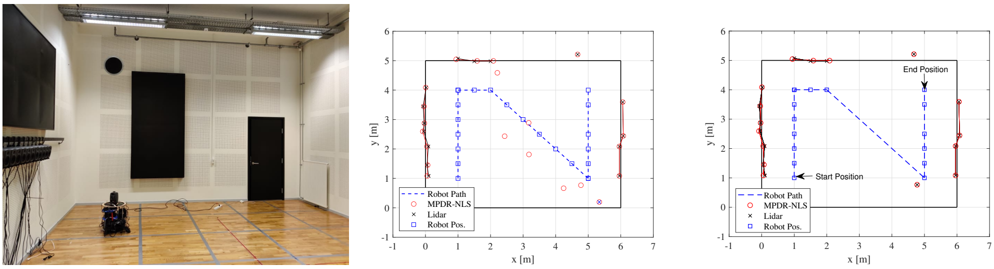
(a) Layout of office with glass surfaces.

(b) 2D map of an office with glass surfaces.

Fig. G.13: Detecting of glass surface at Aalborg University.

positions. The results are shown in Fig. G.13. As seen from the experiment, the proposed method is capable of detecting a glass surface compared to the commonly used lidar sensor. This shows that the proposed method is suitable for constructing a spatial map of a typical office environment.

In the second experiment, Algorithm 3 was used within Aalborg University's Sound Lab, which has a dimension of $5.4 \times 6.38 \times 4.05$ m^3, to construct a spatial map. The objective of this experiment was to move the robot in a more elaborate path within a $3D$ space such that the robot encounters acoustic reflectors as well as empty space along its trajectory. This was done to construct a spatial map of an enclosed environment and also to test the echo detector method presented in Section 4.3. To accomplish this task, the room was divided in to a grid of 20 square boxes, each box has a size of 1 m^2. This was done to ensure that the robot moved along its predefined trajectory and robot's location with respect to the acoustic reflector was always known. Autonomous navigation is also possible, but this would require additional on-board sensors, e.g., using IMU, odometer, gyroscope, etc., to estimate the robot's current position which can then be combined with our estimates to generate a spatial map. As the robot moved within the square grids and followed a predefined trajectory as shown in Fig. G.14(b), the robot probed the environment with a known sound. The recorded sound is spatially filtered using MPDR beamformer which was later fed to a NLS estimator for TOA estimation. Later, the estimated data is passed to a echo-detector, to determine whether it belongs to an acoustic reflector or is an spurious estimate. Finally, the estimated data are combined with the trajectory of the robotic platform to localize acoustic reflectors. As seen in Fig. G.14(b), if the robot moves without the echo detector then it will estimate spurious estimates even when the robot is away from any reflecting surfaces. However, these spurious estimates are removed when echo detector is applied as seen in Fig. G.14(c).



(a) Layout of the Sound Lab.

(b) 2D map of the Sound Lab without GLRT.

(c) 2D map of the Sound Lab with GLRT.

Fig. G.14: Generating a spatial map of the Sound Lab.

9 Discussion

In the experimental section, the performance of the different methods was evaluated in both simulated and practical environments. In the first simulated experiment, Fig. G.6 shows that in the absence of the direct-path component, all method provide good TOA accuracy under different SDNR values. But in the presence of the direct-path component as shown in Fig. G.7 and Fig. G.9, the MPDR-NLS, the LCMV-NLS $\gamma = 0.1$ and the LCMV-NLS $\gamma = 1$ methods detect an acoustic reflector up to a distance of around 2 m under the SDNR of -10 dB while the ScLAM method also provides good accuracy under low SDNR value but it can estimate up to a distance of around $1, 5$ m. Similarly, in a practical scenario, the McLAM methods detect an acoustic reflector up to a distance of around 1.5 m as seen in Fig. G.11. This is similar to the TOA estimation published in our ScLAM paper [G.12]. However, in Fig. G.12, only the MPDR-NLS and ScLAM are seen to provide good accuracy at low SINR for TOA/DOA estimation while LCMV-NLS $\gamma = 0.1$ is the second-best choice for TOA estimation compared to its other variant LCMV-NLS $\gamma = 1$ which performs less then EM-UCA and DSB-NLS when evaluating under different SINR. From these experiments, we can deduce that the MPDR-NLS estimator provides better performance compared to other methods when estimating TOAs and DOAs. The results from practical experiments are also detailed in Table. G.2. The RMSE of all beamformer variants are robust when the distance of the acoustic reflector is less than 1.5 m with respect to the robot. At higher distances, the RMSE increases while the RMSE decreases with higher SDNR values. One noticeable difference that can be seen between simulated and practical evaluation is the low accuracy in practical experiments. There could be various reasons for a lower accuracy in real scenarios. For instance, in our proposed method, we do not take sensor calibration or sensor drift into account.

Additionally, the simulation showed that the presence of a point interfering source does not limit the proposed methods' robustness as seen in Fig. G.8, but it can limit the performance of ScLAM algorithm with 70% accuracy. This show that an MPDR-NLS and both implementations of the LCMV-NLS method are effective in localizing the acoustic reflector of the environment when compared to other beamformer variants in the presence of an interfering source. In the qualitative experiments, we exploited robot's movement to construct a spatial map of an environment. Here, the current technologies were compared, e.g., lidar, with the proposed McLAM algorithm. As seen in Fig. G.14, our proposed method successfully constructed a spatial map of an environment. However, one obvious limitation of the proposed method is that lidar provides more accurate distance measurements over longer distances. This is because sound intensity decreases quadratically over distance due to the inverse squares law. One major advantage of our proposed method, on the other hand, is that it can be used to detect transparent surfaces as seen in Fig. G.13 that are typically found in an office environment, hence our proposed method could complement existing technologies for spatial map generation. Additionally, we also tested our echo-detector in the qualitative experiment. As seen in Fig. G.14(b), without the echo-detector enabled, spurious estimates were seen when the robotic platform is at an empty space. However, as seen in Fig. G.14(c), with the echo-detector enabled, the spurious estimates were removed.

Moreover, both variants of LCMV beamformers behave differently using real data and offers lower performance compared to the simulated results. This could be due to a mismatch of the microphones/loudspeaker positions in the array that leads, in which case the null constraint of the LCMV beamformers are not aligned with the direct-path component. Moreover, the LCMV beamformer implicitly assumes the loudspeaker to be a point source, which will not hold for larger loudspeakers located close to the array in practice. In our future work, we plan to incorporate these inaccuracies within our models and methods to improve their robustness.

As stated in Section 7.6, the proposed method with different variants of beamformers takes around 60 s (per trials) of computational load to estimate the location of an acoustic reflector. A total of around $3,000$ s is required to compute a single experiment. One reason for the long computation time is the matrix inversion operation of the covariance matrix, $\mathbf{R}_{Y_k}^{-1}$, which takes significant time. This can be resolved in practice by exploiting the structure of the covariance matrix as described in [G.63]. Moreover, instead of conducting a full grid search with the proposed estimators for every robot position, a gradient search may be conducted based on previous acoustic reflector estimates.

10 Conclusion

In this paper, we proposed a non-traditional method of constructing a spatial map of an indoor environment using the concept of echolocation. We proposed a model-based approach to distance estimation. In our work, we provided a general model of early reflection and took into account the environmental parameters such as background noise, ego-noise and interfering sources which enabled us to estimate TOAs and DOAs directly from the source signals instead of RIR which is assumed to be known in literature. Instead of working in the ultrasonic range, we proposed working in audible frequency range because most ordinary loudspeaker/microphones work in audible frequency range. Compared to our earlier work, we proposed utilizing adaptive beamforming techniques on a UCA such that it can suppress the direct-path component which can have a detrimental impact on the estimation of TOAs and DOAs. Another contribution of this paper is that it proposed a novel echo detector which was used to distinguish an acoustic reflector estimate from an empty space by using the statistic of the background noise into account. We proposed two algorithms in this paper, McLAM and ScLAM. ScLAM was proposed in our previous work [G.12] which was part of a more general algorithm called McLAM.

Hence, we propose a framework which incorporate four blocks to enable TOA and DOA estimation and these are 1) Spatial filtering block 2) TOA estimation Block 3) Echo Detector and 4) Mapping block. One obvious advantage of this framework is that each module in Fig. G.3 could be separately improved over time in order to increase the performance of the acoustic echo localization. Our simulation results have shown that by using an array geometry with adaptive beamformer, we can robustly improve the estimation of TOAs even in the presence of a direct-path component as shown in Fig. G.6 and interfering sources as shown in FigG.8. As seen from our experimentation, both of the proposed algorithms, McLAM and ScLAM, can detect acoustic reflectors up to a distance of 1.5 m at an SINR of 40 dB and robustly estimate TOAs at an SINR of 10 dB with 80% accuracy in realistic scenarios. The knowledge of TOAs and DOAs could help a robot map an environment to facilitate its autonomous planning and movement. The qualitative experiments demonstrated that compared to the commonly used lidar technology, the proposed method can detect transparent surfaces as seen in Fig. G.13 and it can also construct a spatial map of an indoor environment as seen in Fig. G.14. The later experiment also revealed that the proposed McLAM algorithm could provide a similar performance to lidar sensor if the directionality of the loudspeaker is known.

In a future iteration of this research, we aim to include a method to estimate the loudspeaker's directivity and transfer function within the signal model for TOA and DOA estimation because this is assumed to be known in this work. This will enable our algorithms to work more efficiently and help us understand and develop a sophisticated sound propagation model that could more accurate construction of spatial maps in an indoor environment. We also intend to combine our algorithm with echo-labeling techniques such as [G.64], to classify acoustic echoes that could enable to assign echoes to the corresponding acoustic reflectors. A different approach to acoustic reflector localization could be to modify the proposed algorithm such that it takes the ego-noise of the robotic platform to detect and estimate acoustic reflectors. This way the robotic platform, e.g., drones, does not require probing the environment. Additionally, we aim to reduce the computation load of the proposed method to make it run faster on resource constrained embedded devices. Currently, our proof-of-concept robotic platform moves in a predefined trajectory, where the robot only estimate $R = 1$ acoustic reflector. That is, we already have prior information of the environment to enable spatial mapping but in future, we aim to remove this constraint and enable the robot to move autonomously by employing path-planning algorithms which will enable the robot to estimate multiple acoustic reflectors as it moves. This is possible if the proposed method estimates $R > 1$ acoustic reflectors. By estimating multiple reflectors, objects within the robot's range such as furniture and boxes may also be located and positioned within the map.

References

[G.1] F. Rovira-Más, V. Saiz-Rubio, and A. Cuenca-Cuenca, "Augmented perception for agricultural robots navigation," *IEEE Sensors Journal*, pp. 1–1, 2020.

[G.2] N. Melenbrink, J. Werfel, and A. Menges, "On-site autonomous construction robots: Towards unsupervised building," *Automation in Construction*, vol. 119, p. 103312, 2020.

[G.3] J. Lima, V. Oliveira, T. Brito, J. Gonçalves, V. H. Pinto, P. Costa, and C. Torrico, "An industry 4.0 approach for the robot@factory lite competition," *IEEE Int. Conf. on Autonomous Robot Systems and Competitions*, pp. 239–244, 2020.

[G.4] X. Huang, Q. Cao, and X. Zhu, "Mixed path planning for multi-robots in structured hospital environment," *The Journal of Engineering*, vol. 2019, no. 14, pp. 512–516, 2019.

[G.5] "IEEE standard for robot map data representation for navigation," *IEEE Standard for Robot Map Data Representation for Navigation*, pp. 1–54, 2015.

[G.6] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part 1," *IEEE robotics and automation magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[G.7] A. I. M. G. P. Huang and S. I. Roumeliotis, "A quadratic-complexity observability-constrained unscented kalman filter for SLAM," *IEEE Transactions on Robotics*, vol. 29, no. 5, pp. 1226–1243, 2013.

[G.8] W. L. G. Deng, J. Li and H. Wang, "SLAM: Depth image information for mapping and inertial navigation system for localization," *Asia-Pacific Conference on Intelligent Robot Systems*, pp. 187–191, 2016.

[G.9] T. R. Wanasinghe, R. G. Gosine, O. De Silva, G. K. I. Mann, L. A. James, and P. Warrian, "Unmanned aerial systems for the oil and gas industry: Overview, applications, and challenges," *IEEE Access*, vol. 8, pp. 166 980–166 997, 2020.

[G.10] R. Worley, Y. Yu, and S. Anderson, "Acoustic echo-localization for pipe inspection robots," *IEEE Int. Conf. on Multisensor Fusion and Integration for Intell. Syst.*, pp. 160–165, 2020.

[G.11] D. Hahnel, R. Triebel, W. Burgard, and S. Thrun, "Map building with mobile robots in dynamic environments," *Proc. IEEE Int. Conf. Robotics, Automation.*, vol. 2, pp. 1557–1563, 2003.

[G.12] U. Saqib and J. R. Jensen, "A model-based approach to acoustic reflector localization using robotic platform," *Proc. IEEE Int. Conf. Intell., Robot, Automation.*, pp. 1–8, 2018.

[G.13] H. Wei, X. Li, Y. Shi, B. You, and Y. Xu, “Multi-sensor fusion glass detection for robot navigation and mapping,” *WRC Symposium on Advanced Robotics and Automation*, pp. 184–188, 2018.

[G.14] C. Hui and M. Shiwei, “Visual SLAM based on EKF filtering algorithm from omnidirectional camera,” *IEEE 11th International Conference on Electronic Measurement and Instruments*, vol. 2, pp. 660–663, 2013.

[G.15] H. G. Okuno and K. Nakadai, “Robot audition: Its rise and perspectives,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 5610–5614, 2015.

[G.16] I. Eliakim, Z. Cohen, G. Kosa, and Y. Yovel, “A fully autonomous terrestrial bat-like acoustic robot,” *PLOS Computational Biology*, vol. 14, no. 9, 2018.

[G.17] T. G. Muir and D. L. Bradley, “Underwater acoustics: A brief historical overview through world war 2,” *Acoustics today*, vol. 12, no. 3, 2016.

[G.18] L. Kleeman and R. Kuc, “Mobile robot sonar for target localization and classification,” *The International Journal of Robotics Research*, vol. 14, no. 4, pp. 295–318, 1995.

[G.19] M. Kreković, I. Dokmanić, and M. Vetterli, “EchoSLAM: Simultaneous localization and mapping with acoustic echoes,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 11–15, 2016.

[G.20] J. V. M. L. Nguyen and X. Qiu, “Can a robot hear the shape and dimensions of a room?” *Proc. IEEE Int. Conf. Intell., Robot, Automation.*, 2019.

[G.21] M. Boutin and G. Kemper, “A drone can hear the shape of a room,” *SIAM Journal on Applied Algebra and Geometry*, vol. 4, no. 1, pp. 123–140, 2020.

[G.22] I. J. Kelly and F. M. Boland, “Detecting arrivals in room impulse responses with dynamic time warping,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1139–1147, 2014.

[G.23] G. Moschioni, “A new method for measurement of early sound reflections in theaters and halls,” *IMTC/2002. Proceedings of the 19th IEEE Instrumentation and Measurement Technology Conference*, vol. 1, pp. 425–430 vol.1, 2002.

[G.24] Y. E. Baba, A. Walther, and E. A. P. Habets, “3d room geometry inference based on room impulse response stacks,” *J. Audio, Speech, Language Process.*, vol. 26, no. 5, pp. 857–872, 2018.

[G.25] U. Saqib, S. Gannot, and J. R. Jensen, “Estimation of acoustic echoes using expectation-maximization methods,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–15, 2020.

[G.26] L. B. Nelson and H. V. Poor, “Iterative multiuser receivers for CDMA channels: an EM-based approach,” vol. 44, no. 12, pp. 1700–1710, Dec 1996.

[G.27] M. C. Vanderveen, C. B. Papadias, and A. Paulraj, “Joint angle and delay estimation (JADE) for multipath signals arriving at an antenna array,” vol. 1, no. 1, pp. 12–14, Jan 1997.

[G.28] J. Verhaevert, E. V. Lil, and A. V. de Capelle, “Direction of arrival (DOA) parameter estimation with the SAGE algorithm,” *Signal Processing*, vol. 84, no. 3, pp. 619–629, 2004.

[G.29] D. D. Carlo, A. Deleforge, and N. Bertin, “Mirage: 2d source localization using microphone pair augmentation with echoes,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 775–779, 2019.

[G.30] J. Valin, F. Michaud, B. Hadjou, and J. Rouat, “Localization of simultaneous moving sound sources for mobile robot using a frequency- domain steered beamformer approach,” *icra*, vol. 1, pp. 103–1038 Vol.1, 2004.

[G.31] C. Pan, J. Chen, and J. Benesty, “Performance study of the MVDR beamformer as a function of the source incidence angle,” *J. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 67–79, 2014.

[G.32] H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, “Joint DOA and TDOA estimation for 3d localization of reflective surfaces using eigenbeam MVDR and spherical microphone arrays,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 113–116, 2011.

[G.33] R. C. Maher and E. Hoerr, “Audio forensic gunshot analysis and multilateration,” *Audio Engineering Society Convention 145*, 2018.

[G.34] J. Steckel and H. Peremans, “BatSLAM: Simultaneous localization and mapping using biomimetic sonar,” *PLOS ONE*, vol. 8, no. 1, pp. 1–11, 01 2013.

[G.35] B. Zhou, M. Elbadry, R. Gao, and F. Ye, “Batmapper: Acoustic sensing based indoor floor plan construction using smartphones,” *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 42–55, 2017.

[G.36] S. Pradhan, G. Baig, W. Mao, L. Qiu, G. Chen, and B. Yang, “Smartphone-based acoustic indoor space mapping,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 2, pp. 1–26, 2018.

[G.37] S. M. Kay, *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993, vol. 2.

[G.38] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, “Multi-pitch estimation,” vol. 88, no. 4, p. 972–983, Apr. 2008.

[G.39] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, “A study on how pre-whitening influences fundamental frequency estimation,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 6495–6499, 2019.

[G.40] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, “Evaluation and comparison of late reverberation power spectral density estimators,” *J. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1056–1071, 2018.

[G.41] S. Widodo, T. Shiig, H. Kikuchi, K. Yanagida, Y. Nakatsuchi, and N. Kondo, “Moving object localization using sound-based positioning system with Doppler shift compensation,” *Robotics*, vol. 2, no. 2, pp. 36–53, 2013.

[G.42] Y. Avargel and I. Cohen, “System identification in the short-time fourier transform domain with crossband filtering,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[G.43] J. Li and P. Stoica, “Efficient mixed-spectrum estimation with applications to target feature extraction,” *IEEE transactions on signal processing*, vol. 44, no. 2, pp. 281–295, 1996.

[G.44] U. Saqib and J. R. Jensen, “Sound-based distance estimation for indoor navigation in the presence of ego noise,” *Proc. European Signal Processing Conf.*, 2019.

[G.45] S. Hirata, M. K. Kurosawa, and T. Katagiri, “Real-time ultrasonic distance measurements for autonomous mobile robots using cross correlation by single-bit signal processing,” *Proc. IEEE Int. Conf. Robotics, Automation.*, pp. 3601–3606, 2009.

[G.46] E. A. P. Habets, I. Cohen, and S. Gannot, “Generating nonstationary multisensor signals under a spatial coherence constraint,” *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, 2008. [Online]. Available: https://doi.org/10.1121/1.2987429

[G.47] O. L. Frost, “An algorithm for linearly constrained adaptive array processing,” *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[G.48] O. Schwartz, S. Gannot, and E. A. P. Habets, “Multispeaker LCMV beamformer and postfilter for source separation and noise reduction,” *J. Audio, Speech, Language Process.*, vol. 25, no. 5, pp. 940–951, 2017.

[G.49] L. Liang, F. Kong, C. Martin, T. Pham, Q. Wang, J. Duncan, and W. Sun, “Machine learning–based 3-d geometry reconstruction and modeling of aortic valve deformation using 3-d computed tomography images,” *International journal for numerical methods in biomedical engineering*, vol. 33, no. 5, p. e2827, 2017.

[G.50] W. Yu and W. B. Kleijn, “Room geometry estimation from room impulse responses using convolutional neural networks,” *arXiv e-prints*, p. arXiv:1904.00869, Apr. 2019.

[G.51] R. Humphrey, “Playrec: Multi-channel matlab audio,” *URL http://www.playrec.co.uk*, 2007.

[G.52] E. A. P. Habets, “Room impulse response generator,” Technische Universiteit Eindhoven, Tech. Rep., 2010, ver. 2.0.20100920. [Online]. Available: https://github.com/ehabets/RIR-Generator

[G.53] J. Dmochowski, J. Benesty, and S. Affes, “On spatial aliasing in microphone arrays,” vol. 57, no. 4, pp. 1383–1395, 2009.

[G.54] D. Florencio and Z. Zhang, “Maximum a posteriori estimation of room impulse responses,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 728–732, 2015.

[G.55] E. A. P. Habets, I. Cohen, and S. Gannot, “Generating nonstationary multisensor signals under a spatial coherence constraint,” *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, 2008.

[G.56] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, “DREGON: Dataset and methods for UAV-embedded sound source localization,” *Proc. IEEE Int. Conf. Intell., Robot, Automation.*, pp. 5735–5742, Oct 2018.

[G.57] H. Herlufsen, “Dual channel FFT analysis (part I),” *Brüel & Kjær Technical Review*, no. 1984-1, 1984.

[G.58] M. Feder and E. Weinstein, “Parameter estimation of superimposed signals using the em algorithm,” *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 4, pp. 477–489, 1988.

[G.59] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise reduction in speech processing*. Springer Science & Business Media, 2009, vol. 2.

[G.60] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[G.61] M. Legg and S. Bradley, “Automatic 3d scanning surface generation for microphone array acoustic imaging,” *Applied acoustics*, vol. 76, pp. 230–237, 2014.

[G.62] E. G. Williams, J. D. Maynard, and E. Skudrzyk, “Sound source reconstructions using a microphone array,” *The Journal of the Acoustical Society of America*, vol. 68, no. 1, pp. 340–344, 1980.

[G.63] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, “Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient,” *Signal Processing*, vol. 135, pp. 188–197, June 2017.

[G.64] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, “Euclidean distance matrices: Essential theory, algorithms, and applications,” *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 12–30, 2015.

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-991-2

AALBORG UNIVERSITY PRESS