TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Mikko Parviainen

Self-localization in Ad Hoc Indoor Acoustic Networks

Julkaisu 1420 • Publication 1420

Tampere 2016

Tampereen teknillinen yliopisto. Julkaisu 1420
Tampere University of Technology. Publication 1420

Mikko Parviainen

Self-localization in Ad Hoc Indoor Acoustic Networks

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB109, at Tampere University of Technology, on the 7th of October 2016, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2016

Supervisor:

Ari Visa, Professor
Department of Signal Processing
Tampere University of Technology
Tampere, Finland

Instructor:
Pasi Pertilä, D. Sc. (Tech.)
Department of Signal Processing
Tampere University of Technology
Tampere, Finland

Pre-examiner:

Alessio Brutti, Ph. D.
Fondazione Bruno Kessler
Trento, Italy

Pre-examiner and opponent:

Martin Vermeer, Professor
Department of the Built Environment
School of Engineering
Aalto University

Opponent:
Laura Ruotsalainen, D. Sc. (Tech.)
Department of Navigation and Positioning
The Finnish Geospatial Research Institute (FGI)

ISBN 978-952-15-3819-3 (printed)
ISBN 978-952-15-3895-7 (PDF)
ISSN 1459-2045

Abstract

The increasing use of mobile technology in everyday life has aroused interest into developing new ways of utilizing the data collected by devices such as mobile phones and wearable devices. Acoustic sensors can be used to localize sound sources if the positions of spatially separate sensors are known or can be determined. However, the process of determining the 3D coordinates by manual measurements is tedious especially with increasing number of sensors. Therefore, the localization process has to be automated. Satellite based positioning is imprecise for many applications and requires line-of-sight to the sky. This thesis studies localization methods for wireless acoustic sensor networks and the process is called self-localization.

This thesis focuses on self-localization from sound, and therefore the term acoustic is used. Furthermore, the development of the methods aims at utilizing ad hoc sensor networks, which means that the sensors are not necessarily installed in the premises like meeting rooms and other purpose-built spaces, which often have dedicated audio hardware for spatial audio applications. Instead of relying on such spaces and equipment, mobile devices are used, which are combined to form sensor networks. For instance, a few mobile phones laid on a table can be used to create a sensor network built for an event and it is inherently dismantled once the event is over, which explains the use of the term ad hoc. Once positions of the devices are estimated, the network can be used for spatial applications such as sound source localization and audio enhancement via spatial filtering. The main purpose of this thesis is to present the methods for self-localization of such an ad hoc acoustic sensor network. Using off-the-shelf ad hoc devices to establish sensor networks enables implementation of many spatial algorithms basically in any environment.

Several acoustic self-localization methods have been introduced over the years. However, they often rely on specialized hardware and calibration signals. This thesis presents methods that are passive and utilize environmental sounds such as speech from which, by using time delay estimation, the spatial information of the sensor network can be determined. Many previous self-localization methods assume that audio captured by the sensors is synchronized. This assumption cannot be made in an ad hoc sensor network, since the different sensors are unaware of each other without specific signaling that is not available without special arrangement.

The methods developed in this thesis are evaluated with simulations and real data recordings. Scenarios in which the targets of positioning are stationary and in motion are studied. The real world recordings are made in closed spaces such as meeting rooms. The targets are approximately $1 - 5$ meters apart. The positioning accuracy is approximately five centimeters in a stationary scenario, and ten centimeters in a moving-target scenario on average. The most important result of this thesis is presenting the first self-localization method that uses environmental sounds and off-the-shelf unsynchronized devices, and allows the targets of self-localization to move.

Preface

The work in this thesis has been carried out in the Department of Signal of Signal Processing Tampere University of Technology during periods 2003 – 2006 and 2007 – 2016. First, I would like to thank my supervisor Prof. Ari Visa and instructor Dr. Tech. Pasi Pertilä for their support and guidance.

I thank the pre-examiners Prof. Martin Vermeer from Aalto University and Dr. Tech. Alessio Brutti from the Center for Information Technology of Fondazione Bruno Kessler for their valuable feedback for improving the thesis. I am thankful for the opponents of the public defense of my thesis Prof. Martin Vermeer and Dr. Tech Laura Ruotsalainen from The Finnish Geospatial Research Institute.

I am grateful to Dr. Tech. Heikki Huttunen, Dr. Tech. Anssi Klapuri, Dr. Tech. Jari Yli-Hietanen. Heikki's course Introduction to Signal Processing inspired me to direct my studies towards Signal Processing and apply for a job as a research assistant. Luckily, I was hired by Anssi and Jari in 1998. Anssi's great research ideas led to many collaboratory projects with Nokia that allow me to work with topics that resulted in Master's degree in 2003 for which I am also grateful for Dr. Tech. Tuomas Virtanen for his guidance while writing my thesis.

I want thank my colleagues Teemu Korhonen, Pasi Pertilä, Tuomo Pirinen, and Atte Virtanen, who accepted me to their research group in 2003. Working in such a great group of people with interesting research topics inspired me to pursue doctoral studies.

I am grateful for Prof. Ulla Ruotsalainen and Prof. Ari Visa who hired me to work as a teaching assistant in 2007 that allowed me to continue my doctoral studies and develop my teaching skills. I thank collectively all the students who have participated in my courses and provided feedback that motivated me continuously develop my teaching methods and material.

This thesis would have been impossible to finish without Dr. Tech Pasi Pertilä who hired me in 2013 to work as a researcher in a spatial signal processing project. His complete understanding of the problems encountered during the research allowed me to focus on most reasonable solutions.

I am also thankful for Nokia for funding and especially Mr. Matti Hämäläinen for his feedback and ideas during our collaboration.

Thanks to my parents Marja-Leena and Jorma for their support and their encouragement. Thanks to Mari and Heikki for their support and tolerance for me once in a while complaining about various things.

Thanks to personnel in the Department of Signal Processing especially secretary Ms. Virve Larmila for her help, advice, and guidance.

Thanks to my colleagues including Juha, Atte, Tuomo, Vesa, Katariina, Hanna, Elina, Joonas, Antti, Aleksandr, Toni, Annamaria, and Marko.

Contents

List of Included Publications

This thesis is a compound and consists of the following publications. The first publication is referred to as [P1], the second as [P2], and so forth.

P1 **M. Parviainen, P. Pertilä**, "Self-localization of Dynamic User-Worn Microphones From Observed Speech', in Applied Acoustics

P2 **M. Parviainen, P. Pertilä, and M. Hämäläinen**, "Self-localization of wireless acoustic sensors in meeting rooms," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*, pp. 152–156, May 2014.

P3 **M. Parviainen**, "Robust self-localization solution for meeting room environments," in *13th International Symposium on Consumer Electronics*, (Kyoto, Japan), 5 2009.

P4 **M. Parviainen, P. Pertilä, T. Korhonen, and A. Visa**, "A spatiotemporal approach for passive sound source localization — real-world experiments," in *International Workshop on Nonlinear Signal and Image Processing (NSIP2005)*, 2005.

P5 **M. Parviainen, T. Pirinen, and P. Pertilä**, "A speaker localization system for lecture room environment," in *3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, 2006.

Chapter 1

Introduction

Humans can infer a lot of information from sound. For instance, we can tell in which kind of environment we are, which kinds of sounds we hear, rough direction and location of a sound event, and what is the source or phenomenon that has caused the sound. It is appealing to research which properties of sound human hearing extracts to make such conclusions. Hearing is a very complicated system, but in recent decades several models have been presented that are able to mimic its capabilities [Bre90][Bla96]. Since sound can provide a vast amount of information, many computational systems are constantly being developed, and multiple systems based on sound already exist in everyday technology. This thesis focuses on inferring location from observed sounds such as speech.

Acoustic sensor networks are already common in specialized settings. For instance, lecture rooms and meeting venues often have microphones, speakers, and video cameras installed and they are utilized for various purposes. In a teleconference, the person speaking could have a name tag rendered over his body in the video stream. The location information of the person speaking can be used to steer and focus the camera, and adjust microphone gain towards her/him. Such a system is proposed in [LRGC01].

To enable this type of application, positions of sensory devices like microphones and microphone-camera units are needed. The positions could be determined manually during the installation of such a system. This is not a problem in case of a meeting room with fixed brackets mounted on the walls. In order to introduce this technology outside dedicated premises, mobile devices provide an appealing approach; instead of fixed equipment, could one utilize smartphones and other consumer-level lightweight devices for the same purposes as the dedicated hardware? Such devices are referred to as *ad hoc nodes* because they are not necessarily permanently on the premises. A node can be either a sensor, a source, or an entity consisting of a source and a sensor. For instance, a mobile phone contains at least one sound source and one microphone and therefore constitutes a node. In an acoustic positioning system installed in a meeting room, the nodes are the individual microphones used for positioning, and the participants of the meeting represent the sources[1]. A node can even consist of a sound source and multiple sensors, which in practice is a miniature microphone array with a speaker. Which node model is used, is a fundamental design parameter of a self-localization algorithm. That is, which capabilities can be expected from a node and can therefore be exploited in designing the self-localization method.

Figure 1.1 illustrates a dedicated microphone array and an ad hoc microphone array

1. A meeting room often contains other sources that are not necessarily of interest such as a video projector and air-conditioning.

which can be used for spatial audio applications. To obtain spatial information, the array geometry is needed. In dedicated microphone arrays, the geometry is a design parameter and therefore the information is available. However, dedicated devices are often expensive and non-ubiquitous. The geometry of an ad hoc array needs to be determined, but the equipment is generally available and inexpensive.

The determination of the geometry of an ad hoc microphone array can be done in some cases using e.g. a tape measure, but clearly such a step in deployment is impractical, not to mention unappealing or even impossible for a nontechnical user. Furthermore, measurements made by hand may be of insufficient accuracy, which leads to deteriorated spatial information. All in all, manual installation (positioning) is cumbersome. This is one of the major motivations for automatic positioning, or *self-localization*.

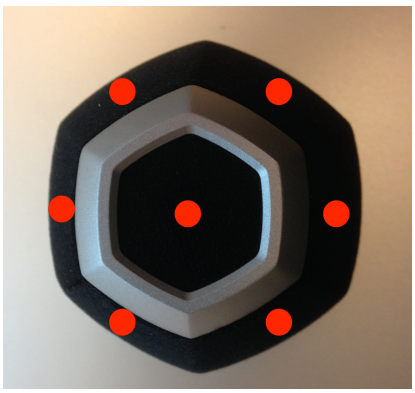
Self-localization for wireless sensor networks has been studied extensively. Such networks usually consist of nodes capable of exchanging information, such as received signal strength and time of arrival, via radio frequency signaling [PAK+05]. Using acoustic signals for self-localization sets a different scenario, e.g., sound propagation is often affected by objects present in a space including its building materials (the walls, floor, and ceiling).

In addition to business applications, the amount of technology in the home environment is increasing and therefore there is a need for self-localization to build new sound based applications. For instance, a home theater system could be tuned to provide the optimal acoustic experience based on determining the location of a person watching a movie. Speech based user interfaces benefit from signal enhancement that can be implemented using microphone arrays [Zel88].

Given the examples above, self-localization is enabling technology, i.e., self-localization facilitates the implementation of a variety of applications. There is a clear demand for self-localization algorithms especially when *smart rooms*[2] become more and more common in business and eventually in homes. Some functions of smart rooms are already in place. However, the utilization of data collected by the sensors is much wider in scope. A smart room mainly for research purposes is presented in [NCMH09] and it contains 85 microphones and 8 cameras that allow testing of various audio technologies including automatic speech recognition, speaker identification, speech activity detection, acoustic source localization, acoustic event detection, and speech synthesis. The video cameras enable the development of video technologies such as multi-camera localization and tracking, face detection and identification, body analysis and head pose estimation, gesture recognition, object detection and analysis, text detection, and global activity detection. Of course, the room of [NCMH09] contains expensive equipment preventing its widespread adoption as such. But testbeds are needed to research which kind of requirements apply for each part of the system. A simpler version than the smart room of [NCMH09] is presented in [BAR05], which researches the deployment of a sound source localization system in the home environment. The study [BAR05] explores the use of sound source localization to infer communication activity between people.

Besides smart room applications, self-localization makes some well-known tasks more feasible and easier to implement such as microphone array calibration[SSP02][SMSP05].

2. A smart room is equipped with sensors such as microphones, cameras and other sensory elements that can be used for various purposes, e.g., communication via video and audio.



(a) A dedicated microphone array.

(b) An ad hoc microphone array.

Figure 1.1. Information on the microphones' relative locations is needed to use such an array in speech enhancement. With a dedicated microphone array this information is easily obtained. Self-localization is needed with an ad hoc microphone array to acquire the distances between the nodes (e.g. mobile devices).

1.1 Self-localization Problem

Self-localization in this thesis refers to determining the relative positions of each node of a network. The term is sometimes used also to refer to fine-tuning of node position estimates. When discussing this type of self-localization, the term *self-calibration* is more appropriate as used by Rockah [RS87]. Sometimes also the opposite is true – self-calibration is used while self-localization is more appropriate term.

Self-localization typically utilizes internode measurements that relate to distances between nodes. Measurements can be for instance Time of Arrival (TOA), Time Difference of Arrival (TDOA), Angle of Arrival (AOA) and Received Signal Strength (RSS). The measurements are often made from calibration signals. Several methods that utilize ambient sounds have been presented including [PMH12][Thr05][MLH08]. These methods are often referred to as passive due to the fact that the self-localization system itself does not transmit or emit anything itself.

Since the measurements are often relative to an arbitrarily selected node, one is not able to derive physical locations of nodes. However, in many applications it is desirable to obtain the physical coordinates. Therefore, there has to be some mapping between the *virtual coordinates* derived from the internode measurements and the physical location in which the sensor network lies. This mapping can be implemented by measuring positions of a small subset of nodes manually. These special nodes are often referred to as anchor, or beacon, nodes. If there are enough beacon nodes, the self-localization problem turns into a source localization problem since measurements made by known nodes can be used to localize each unknown node provided that the unknown node is transmitting or emitting a signal that can be measured.

1.2 The Objectives and the Scope of the Research

The main objective of this thesis is to research and develop acoustic localization and self-localization methods. The aim is to develop techniques for conditions in which other methods cannot be used. Such conditions include closed spaces in which many localization technologies such as satellite positioning systems fail or are inaccurate. Passiveness of the developed localization methods is one of the

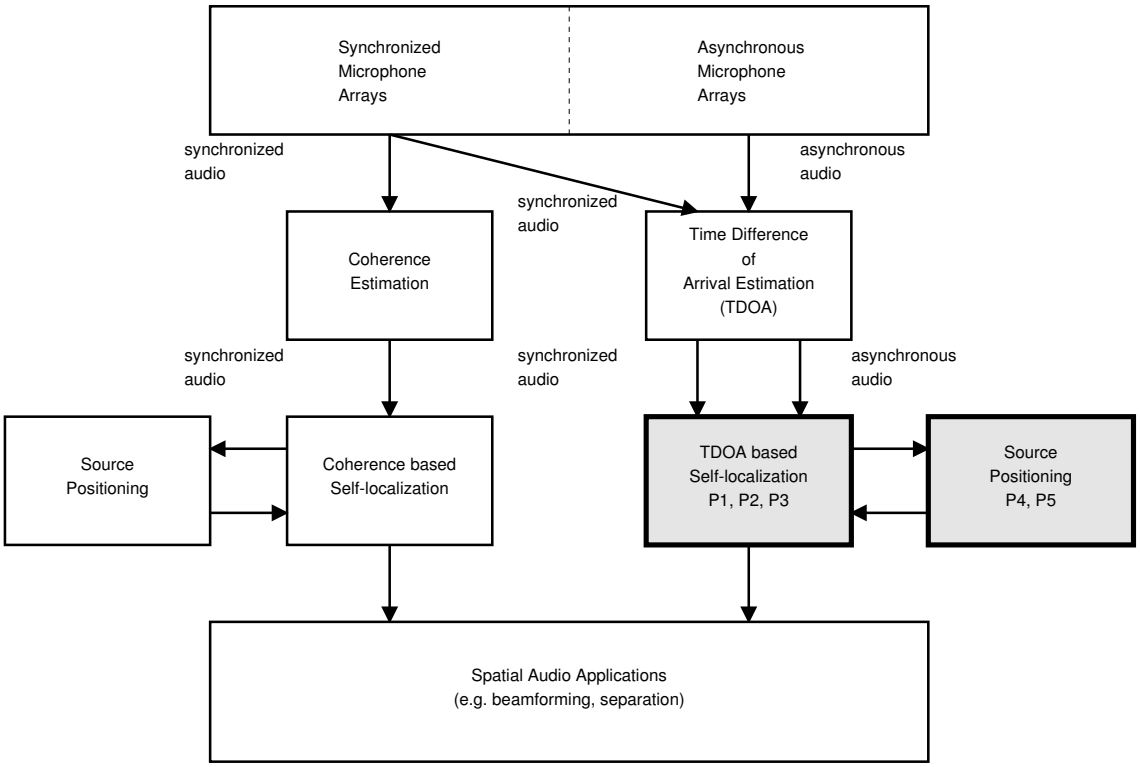


Figure 1.2. An overview of spatial audio processing for localization and self-localization. The work done in this thesis and its relation to other fields of spatial audio signal processing is highlighted in gray.

major advantages, which is important in surveillance applications. The developed self-localization methods aim at unobtrusiveness and ubiquitousness to provide techniques for adoption in consumer applications. Therefore, ad hoc wireless acoustic sensor networks using off-the-shelf equipment are specifically discussed in this thesis.

Figure 1.2 presents a typical processing path from sound signals collected by microphones to spatial applications. A microphone array is a formation of microphones that are placed in a known geometry. The number of microphones in an array varies depending on application. In general, the increase in number of microphones improves the accuracy and robustness. In an ad hoc microphone array, the distances between microphones are unknown. The combination of microphones can form an ad hoc array if they are close enough to measure the same acoustic events.

To obtain spatial information, the microphone signals are compared using techniques that are able to extract differences between the signals. Typically, a time delay is estimated using Time Difference of Arrival (TDOA) techniques and coherence estimation. Which technique is used depends on the expected sound field. This thesis assumes a point source model and therefore TDOA estimation is used to obtain spatial features of acoustic events. TDOA is used for the actual objective self-localization and source positioning. The results of self-localization can be used in many spatial audio applications such as beamforming and sound source separation. The work done in this thesis and its relation to other fields of spatial audio signal processing is highlighted in gray in Figure 1.2.

1.3 Author's Contributions

The author has carried out the majority of the research in each of the included publications [P1]-[P5] with the exception of the following contributions:

- The research proposal of [P1] came from M. Hämäläinen and the idea to use state based approach from P. Pertilä. The head of the research was P. Pertilä.
- The research proposal of [P2] came from M. Hämäläinen, and the work was conducted in guidance of P. Pertilä, who was the head of the research.
- The idea of [P3] is from the thesis author and the research was conducted in guidance of A. Visa, who was also the head of the research.
- The idea in [P4] to utilize time-of-flight information came from P. Pertilä. The head of the research was A. Visa.
- [P5] is collaborative research conducted with T. Korhonen, P. Pertilä, and P. Pirinen. The head of the research was A. Visa.

This thesis itself is a new publication that presents the role of the research in [P1]-[P5] in spatial signal processing, reviews similar and the author's methods, and presents applications for the developed methods.

The author's earliest work on self-localization [P3] used studied the use of regularization methods in a self-localization framework presented in [Fra06]. It was observed that in the given sensor/microphone scenario, regularization methods improve the robustness of the system.

This thesis provides a general framework for self-localization in quasi blind cases [P2]. The developed framework is able to operate with asynchronous data, estimates positions of microphones and sound sources, and labels sources based solely on TDOA measurements (i.e. without knowledge of microphone positions). It is assumed that sources and microphones are non-moving and therefore the approach is not entirely blind. However, the system can be used for many cases such as self-localization in a meeting scenario. The system [P2] achieves an accuracy of approximately 10 cm for real data measurements (meeting scenario).

A framework exploiting the node geometry is developed in [P1]. The system estimates node (microphone-source combination) positions from TDOA measured by the nodes themselves. The system is able to operate for non-moving nodes as well as for nodes in motion and achieves a 3D positioning accuracy of approximately 10 cm for real data in a four-node scenario, where two of the nodes are in motion simultaneously. The geometry approximates a wearable device scenario and therefore is an example that a reasonable assumption can make the self-localization problem easier to solve and the solution implementable.

The main strides of this thesis in the field of acoustic self-localization include utilizing ad hoc sensors and source, using environmental sound as source signal, allowing a dynamic scenario while the self-localization algorithm is run, and writing algorithms that enable use of consumer-level devices.

This thesis studies sound source localization using multiple arrays with known microphone geometries, which was the early work and motivation for work on self-localization. The investigated scenarios include large distance [P4] and room scale [P5] setups. The source localization using multiple arrays has been widely researched over the years. The main contribution of this thesis includes processing of data collected from large inter-array distance networks requiring acknowledging differences in propagation times from source to array, and approaches to combine information for source localization.

1.4 Outline of the thesis

The rest of the thesis is organized as follows. Chapter 2 presents background theory and supporting methods, formulates the general self-localization problem, and presents related and previous self-localization work. Chapter 3 presents the main work of this thesis. The most relevant related methods for acoustic self-localization as well as the methods developed in Publications [P1], [P2] and [P3] are reviewed. Chapter 4 presents sound source localization, which is one of the applications of the main work of this thesis. The principles of two acoustic source localization techniques Time Difference of Arrival (TDOA) based and Direction of Arrival (DOA) are presented. The DOA based localization methods presented in Publications [P4] and [P5] are reviewed. Chapter 5 concludes the results of the thesis.

Chapter 2

Background

This chapter presents the background knowledge of self-localization. First, the properties of sound, how sound source energy propagates from the source, and how the sound source can be measured are discussed. Then, the extraction of spatial information from the observed sound is presented followed by previous work on self-localization. Finally, supporting techniques and algorithms are discussed.

2.1 Sound

The concept of sound is easy to understand because it is a natural part of our everyday lives for most humans. However, to model and build systems that utilize this physical phenomenon, a more specific definition is needed. *A sound* is a disturbance that travels through an elastic material that cause changes in pressure by displacement of the particles in the material, which can be detected by an instrument or a person [Ber86].

One of the most significant properties of sound is its frequency content, that is, at which rate the particles of the medium are being made to vibrate by the sound source. Human beings are typically able to detect sound in a frequency range of $20 - 20000$ Hz, or vibrations per second.

Before reaching an observer or a sensor, the energy produced by a source propagates as waves away from the source. Thus, air acts as the conveyor of the sound source energy. Microphones measure the pressure change of the air and transform it to electric signals. The amplitude of the signal reflects the energy of acoustic waves at a given time. This chain of events is illustrated in Figure 2.1.

2.1.1 Acoustic Events

The definition of acoustic event may vary from one application to another. Acoustic event here is defined as an action that results in the production of sound. However,

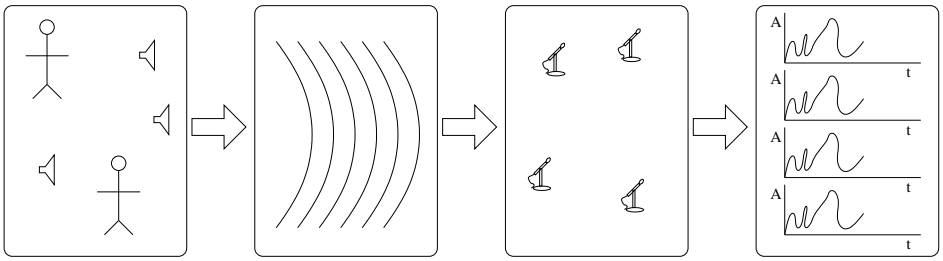


Figure 2.1. The chain of sound production to digital signals.

as distinct from any sound, an acoustic event has start and stop times. For instance, a door that is closed produces an acoustic event. Instead, a continuous humming sound produced by an air-conditioning device is often considered as background. Still, a temperature controlled air-conditioning device when activated would produce such a change in the acoustic environment, that it would be considered an acoustic event.

2.1.2 Sound Sources

Frequency, intensity, continuity and movement characterize sound sources. Having knowledge about the characteristics of a source is often necessary when designing acoustic systems.

Frequency: many natural signals that humans observe in their everyday life are considered as broadband (e.g. speech, wind, and other environmental sounds), i.e., a sound source emits at multiple frequencies. By contrast, a sound source emitting a sinusoidal tone at a single frequency or a source emitting at few closely spaced frequency components is considered a narrowband source. The frequency range at which a source emits is referred to as the bandwidth of the source.

Intensity: The energy produced by sound sources is transfered as waves. The intensity describes the rate at which the energy flows through the medium per unit area in a time unit (W/m^2). Strictly speaking, the intensity is a vector quantity defined as the product of particle velocity and the sound pressure. While designing algorithms and systems, the intensity of the target sound source has to be taken into account, e.g., a low intensity target source may limit the operating range of an acoustic system. Furthermore, a low intensity target may be masked by a higher intensity undesired source.

Movement: The knowledge on whether a sound source has a capability of moving and how fast is important in designing many acoustical systems. Loudspeakers installed in a room can be assumed to be stationary in many cases. Humans and other animals can be assumed to be able to move, but at a lower speed than cars for instance.

Continuity: Natural sound sources are often static or on/off type sources. For instance, a human speaker emits a sound for a while, pauses, and then continues to speak. Many artificial sounds may be impulse-like and brief (e.g. “beep”). From an acoustical system design viewpoint, continuity is important in the case of multiple sources. For instance, in human conversation of two speakers it is often the case that one speaker talks then pauses as the other responds. Sound sources that are active only for short time periods are hard to detect especially in the presence of multiple continuous sources and background noise. Knowledge of continuity besides the number of sources and source motion is an important feature from the data association viewpoint, that is, from which source the energy observed at a certain time originated.

2.1.3 Sound Propagation

Sound causes the particles of the medium to vibrate, resulting in sound observation in multiple locations. The propagation is actually a very complex process that is affected by the surroundings and environmental conditions (e.g., temperature and humidity). The modeling of propagation has been studied over the years and it is understood quite well [Ber86]. However, in real environments parameters affecting sound propagation vary all the time. On the other hand, if the amount of changes is small, the propagation of sound is mostly defined by static parameters such as walls, furniture, and other objects.

The relative distance between source and receiver is one of the major propagation modeling parameters. The distance defines whether a sound source is in *near-field* or in *far-field*. Far-field means that sound wave propagation can be modeled as a plane wave. For sound sources that are considered to be located in near-field with respect to the receiver, the wavefront model is spherical. The accuracy of the localization system is affected by whether the design assumes near-field conditions or far-field, and how well the assumption is met during the operation of the system. Usually, acoustical systems assume either near-field or far-field conditions.

In free field such as an anechoic chamber, sound wave propagation occurs along the direct path to the receiver. In indoors, multipath sound propagation occurs. This means that besides the direct path from the source to the receiver, the sound waves experience *reflections* from walls and other surfaces. Reflections can be either *specular* or *diffuse*. Specular reflections can be also referred to as mirror-like reflections. Waves coming from a single incoming direction continue to a single outgoing direction. The angle between propagation direction and surface normal of incoming and outgoing waves is the same. Diffuse reflections cause sound waves to reflect from a surface into multiple directions. *Diffraction* occurs when sound waves hit an object much smaller than their wavelength; the sound waves bend around such objects rather than reflect specularily from them [Ear03]. In this sense diffraction is similar to diffuse reflection since the outgoing propagation direction depends on several parameters.[Nir04]

The above mentioned phenomena affecting wave propagation are present in rooms and other enclosed spaces and they are defined by, e.g., room dimensions, wall materials, furniture, and other objects and their materials. The acoustic properties of a space is characterized by its *reverberation time*. It is defined as the time it takes for the sound pressure level to attenuate 60 dB after a continuous sound source is turned off [Ros90]. Reverberation time T_{60} is related to absorption coefficient γ, which defines how much sound energy is absorbed by a surface. γ for a certain surface is defined by incident angle, frequency, and material [Ber86]. Roughly speaking, a large space with surfaces that have low γ has a long T_{60}. A human observer would characterize this kind of space as echoic or reverberant [Ear03]. An anechoic chamber is an extreme example of a space with very low T_{60}. In anechoic chambers every surface is made of absorbing material. The floor is a floating net beneath which there is the same sound absorbing material used on the walls and on the ceiling. From a system design viewpoint one would desire a space with very low T_{60}. However, real-world spaces are often even designed to be reverberant (e.g., concert halls), since human beings perceive such spaces as livelier than rooms with little reverberation [Ros90]. Sound recording studios can often alter the level of reverberation using movable objects making the room “live” or “dead” [Ear03].

Sound attenuates as it propagates through the medium. Attenuation comprises of geometric spreading, atmospheric effects, and surface effects. In an anechoic chamber and outdoors (i.e., free-field conditions), sound attenuates from a point sound source approximately 6 dB each time the distance is doubled (inverse-square law) [Ear03]. Geometric spreading, i.e., the spreading of sound energy as a result of the expansion of the wavefronts is independent of frequency [Tru99].

Atmospheric effects include air absorption and wind/temperature gradients. Air absorption is caused by molecular relaxation and viscosity effects, the former of which is more important. The amount of absorption depends on temperature, humidity, and frequency. For instance, at 2 kHz the absorption is typically 0.25 dB / 100 m for 30 % relative humidity, and 20°C. At higher frequencies, the absorption is greater. For

instance, with 8 kHz sound, the absorption is 5 dB / 100 m for 10 % relative humidity and 20°C [Har66]. Since sound speed is dependent on temperature, local changes in temperature affect the speed if air temperature is cooler at higher altitudes. This results in bending of acoustic waves upwards (lower sound speed). Wind gradients cause sound waves to bend depending on wind direction. Temperature and wind gradients can cause greater sound level alteration than predicted by the geometric spreading model [Tru99].

In the outdoors, surface effects consist of ground absorption and attenuation due to trees and other obstacles. In urban environments, human built structures can be very effective sound attenuators when designed acoustically [Tru99]. On the other hand, many structures cause reflections and therefore cause echoic environments.

2.1.4 Sound Observation and Measurement Devices

After emission and propagation, the pressure changes of sound waves can be converted into electrical form using a transducer. Microphones can be used as transducers when frequencies of interest cover the typical human hearing range. There are several types of microphones designed for specific purposes. One design parameter is pickup pattern. Many pattern types have been introduced and the basic types include omnidirectional, cardioid, supercardioid, hypercardioid, subcardioid, and gradient [Ear03]. The omnidirectional pattern means that sound waves arriving from all directions have equal gain. There are many applications in which it is desirable that gain is different for certain directions. For measurement purposes, and mostly for applications discussed in this thesis, omnidirectional microphones are most useful, since amplifying sound from specific directions could leave target sound sources undetected in the worst case.

Over the years several microphone types have been developed [Micb]. Modern microphones operating at audible frequencies are typically condenser microphones. A microphone consists of a diaphragm that converts sound pressure changes into voltage changes. In a condenser microphone, the diaphragm is essentially a capacitor. The changes in air pressure lead to changes in capacitance, which in turn leads to changes in voltage [Ear03]. The voltage changes are converted into digital form by an analog-to-digital (AD) converter. From the viewpoint of mobile and embedded technology, the most interesting microphone type is a Micro Electro-Mechanical Systems (MEMS) based microphone [WBBZ06], which is usually a variant of a condenser microphone. MEMS microphones implement diaphragm, preamplifier, and an AD converter on a silicon chip making it very easy to install. These can be found, e.g., in mobile devices. Besides small size, reliability and good sound quality are benefits of MEMS microphones. For instance, a typical signal to noise ratio (SNR) is approximately 60 dB [MEM].

2.2 Time-delay Estimation

Time delay in this context refers to propagation time between two sensors. More specifically, sound is received first by the sensor which is closer to the source, and after a time delay, sound reaches the sensor farther away.

Time-delay estimation is the key to sound source localization and self-localization, since it can be directly measured from the sensors. Time-delay between two microphones can be used to determine the direction of arrival of a sound source and

multiple time delays measured between a set of microphones can be used to sound source localization. A common way to estimate the delay between two signals is to calculate cross-correlation [Wei] between the signals.

$$r_{x_1,x_2}(\tau) = \int_{-\infty}^{\infty} x_1(t)x_2(t+\tau)\,dt, \tag{2.1}$$

where τ is the time delay between x_1 and x_2. Because the signals are discretely sampled, cross-correlation can be only estimated. The time delay can be estimated by seeking a delay τ that maximizes the cross-correlation function:

$$\hat{\tau} = \underset{\tau}{\operatorname{argmax}}\, r_{x_1,x_2}(\tau) \tag{2.2}$$

The Fourier transform of cross-correlation function is defined as follows.

$$G_{x_1,x_2}(f) = \int_{-\infty}^{\infty} r_{x_1,x_2}(t)e^{-j2\pi ft}\,dt, \tag{2.3}$$

$G_{x_1,x_2}(f)$ denotes the *cross spectral density*.

2.2.1 Generalized Cross Correlation (GCC)

When working with real world signals, the cross-correlation function (2.1) is affected by e.g. background noise and reverberation, which causes spurious peaks in (2.1). The Generalized Cross Correlation introduces a weighting scheme to the correlation function. The weighting improves the robustness of TDE since it effectively filters the input signals in such a way that the maximum of GCC is more prominent than in (2.1). Generalized Cross Correlation is defined [KC76] as follows

$$r_P(\tau) = \int_{-\infty}^{\infty} \Phi_P(f)G_{x_1,x_2}(f)e^{j2\pi f\tau}\,df, \tag{2.4}$$

where the connection of cross spectral density and inverse Fourier transform is utilized. $\Phi_P(f)$ is a frequency weighting function enabling utilizing possible a priori knowledge about the input signals. Different kinds of weight functions can be used to shape the correlation function. The purpose of the different weighting schemes is to make the peak of the cross correlation function more prominent and thus provide a more robust estimate time delay estimate.

A popular weighting scheme is Phase Transform (PHAT)

$$\Phi_P(f) = \frac{1}{|G_{x_1,x_2}(f)|}. \tag{2.5}$$

PHAT weighting results in unity amplitude for all frequencies. Theoretically, PHAT should make the correlation function an impulse function, where the impulse occurs at the instant corresponding to the time delay between the signals. In general, the cross correlation function has several peaks resulting from echo and noise sources, and PHAT lowers the other peaks, making picking the peak corresponding to the direct path propagation easier [VDBBK+12]. Other weighting schemes such as the Roth Impulse Response, the Smoothed Coherence Transform (SCOT), the Eckart Filter, and the HT Processor are compared e.g. in [KC76] and in [Tas09].

2.3 Self-localization Methods – Related Work

The digitized signals measured spatially apart can be analyzed to derive information about the spatial properties of the sensor network. The goal of the self-localization algorithm is to localize the sensors. Some methods localize the sound sources after the sensors have been localized. The key to self-localization is to be able to detect differences from spatially separate observations, and use the differences to obtain the locations. The observed differences are by nature relative, which means that only relative positions can be estimated. To estimate physical positions, many methods also apply a priori knowledge to fix the relative location estimates to a physical coordinate system. The choice of a method depends on many factors including the application of the self-localized sensor, how accurately the sensor need to be localized, what kind of a priori knowledge there is if any, are the sensor and/or sources moving, what is the type and quality of the signals that sensors are measuring. Self-localization has been studied widely in wireless sensor networks and an overview and analysis can be found in [PAK+05]. In the following, self-localization and related technologies that have been developed and studied over the years are discussed.

2.3.1 RF and Ultrasound Based Self-localization Methods

Global Navigation Satellite System (GNSS) based node localization can be used to perform self-localization. However, this requires that every node in the network must contain a GNSS receiver. Furthermore, such a system is dependent on an auxiliary system and is only able to operate outdoors. For many applications the positioning accuracy of GNSS is far too coarse. Even with differential GPS the accuracy is of order tens of centimeters. For regular GPS the accuracy is of order several meters [HT08]. Other issues using GNSS include energy consumption and increase in cost.

A cricket location-support system [PCB00, MO07] is designed for indoors node localization. The cricket location support system design goals are privacy (listeners are passive), decentralized administration (owner of premises installs the beacons), network heterogeneity (the system is not coupled to a specific networking technology), and low cost per unit [PCB00]. The system uses *beacons* and nodes (static and mobile) that have *listeners* attached to them. Beacons' locations are known or obtained manually. Listeners use a combination of RF and ultrasound hardware and measure Time of Flight (TOF) to determine their ranges to beacons. Thus the system is suitable for dedicated spaces in which localizing of nodes is needed. Once in place, the cricket location-support system has shown to be an enabling technology (via its possibility to locate listeners) for various applications such as indoor active maps and device control [PCB00].

Simultaneous Localization and Tracking (SLAT) is a method developed for localizing sensor nodes which uses a moving target called mobile to localize sensor nodes. SLAT is related to Simultaneous Localization and Mapping (SLAM) ([DWB06]), a technique used in mobile robotics for tracking a target in a sensor network while simultaneously localizing the nodes of the network. SLAT determines distance to the nodes by the mobile emitting *events* measured by sensor nodes. Tracking techniques, e.g., Bayesian filtering are used to determine sensor node locations and mobile locations concurrently. The initial localization guesses can be obtained, e.g., from radio connectivity information. [TRB+06]

Ad hoc Positioning System [NN01] is a distributed, hop-by-hop positioning algorithm. It uses connectivity information to localize the nodes of the network. Connectivity

is a binary value of whether two nodes are within their communication range or not. The network is seen as a graph and the edges of the graph carry connectivity information. The network has special nodes called landmarks that are used to fix the graph to physical coordinates. The landmarks are equipped with GNSS functionality or other functionality that estimates the global coordinates of the landmark. The immediate neighbors can determine distance to landmarks nodes using signal strength measurement using a propagation method, the second hop neighbors are able to determine their distance to the landmark and so forth. Each node maintains information about landmark coordinates and hop count. A landmark calculates average hop size using other landmarks and transmits that information to the network.

Many techniques presented above assume the properties of RF wave propagation and therefore they do not account for the issues encountered with acoustic signals. The situation is very different with acoustic wave propagation when comparing, e.g., signal attenuation, propagation speed, and behavior when encountering physical objects.

A Self-Organizing Map (SOM) based self-localization method for a wireless sensor network is proposed in [GGM07]. The method however can be applied to other networks if connectivity information or more specifically, *a hop-count matrix* can be derived. Basically, an entry in the hop-count matrix is similar to the pairwise distance between two nodes. The method produces relative coordinates of nodes. A distinctive feature in the method as in many other connectivity information based methods is that it is able to operate with deficient input data, i.e., the existence of distance information between every node pair is not required. The main drawback of [GGM07] is convergence in case of random initialization of the algorithm.

In [ERB+14] a self-localization method of a moving receiver is presented. The method uses TDOA estimates from ultrasound, making it unobtrusive for humans. The drawback of the system is the requirement of an infrastructure of ultrasonic transmitters, their careful placement in a room, and a side channel for data association.

2.3.2 Self-localization Using Acoustic Signals

This class of methods uses acoustical signals to self-localize network nodes. Another emerging feature is that the methods presented here do not use auxiliary hardware or systems.

In [Fra06] an acoustic self-localization method is presented, which determines 3D sensor coordinates. The system estimates TDOA from several sources at known locations. Besides known source locations, it is assumed that the signals are synchronized. A system of non-linear equations similar to near-field sound source localization techniques (see e.g. [HBEM01][SA87b][MH95][WM97]) is solved for 3D sensor coordinates.

In [RKL03][RKL05] a method for self-localization of an ad hoc heterogeneous network is proposed. Heterogeneous network here means that nodes can be different types of devices such as laptops and tablets. It is assumed that each node has a microphone, a loudspeaker and wireless communication capabilities. The system estimates TDOA between microphones and TOF between microphones and loudspeakers, and finally the actual (unknown) and the estimated TDOAs are formulated as a nonlinear cost function. The nonlinear cost function is initialized with node location estimates obtained using Multidimensional scaling (MDS), which is described in Section 2.4.

In order to apply MDS, measurements between all node pairs i.e. microphone-loudspeaker, microphone-microphone and loudspeaker-loudspeaker are needed. Since the latter two cannot be measured, *clustering* is performed for microphones and loudspeakers that are close to each other (i.e. located in the same device). This allows the use of MDS.

Matrix factorization based methods have been proposed in [Thr05] and [PN08]. These methods use a matrix factorization technique to ease the nonlinear optimization problem. In [Thr05] a method called structure from sound (SFS) is proposed. It performs self-localization from environmental acoustic events, which are not generated by the network itself. The sensors measure TOF with respect to the first sensor, which is chosen to be in the origin. This relative measurement results in cancellation of emission times that are unknown in the formulation. Using SVD on the relative distance matrix (derived from relative TOFs) and assuming that acoustic events are in the far-field, results in a nonlinear optimization problem with only four unknown parameters (2D self-localization). This eases the optimization, since the original SFS problem is a function of the number of sound sources, sensors, and coordinate dimensions. In [LW08] Thrun's approach is extended to take measurement uncertainty into account. In [PN08], instead of the far-field assumption of [Thr05], spherical wave propagation is assumed. A rank-5 matrix factorization approach is proposed in [PN08] to ease the nonlinear optimization problem (in [Thr05] SVD is rank-3 factorization in 3D space). After factorization, constraints are applied to the estimated node locations to satisfy the formulation. This approach requires 10 microphones and 4 sources to solve the positions (or 10 sources and 4 microphones). Further refinements to matrix factorization based approaches to self-localization are presented in [CDBM12], [CDBBM12], and in [BKÅ15].

A method presented in [MLH08] is directed to microphone array self-localization or calibration in diffuse sound field. An extension to [MLH08] is presented in [HPF+09], which uses multiple microphone arrays and sound source localization to estimate relative rotation and translation of an array pair. Both methods are designed for relatively small intra-sensor distances of approximately 20 cm or smaller.

In [HF11] a self-localization method for an ad hoc network consisting of smartphones is proposed, which is similar to [RKL03, RKL05] in that microphones and loudspeakers are basically at the same position. Also the integrated compasses of the devices are used to obtain orientation of the device. The benefit is the reduction of free parameters because loudspeaker coordinates are solely dependent on microphone coordinates. The approach [HF11] takes into account the fact that in ad hoc networks the devices are asynchronous. To estimate the distances, TDOA are used. TDOA estimates are corrected with capture start times to take asynchronous infrastructure into account. Capture start times are estimated by using an arbitrary microphone as a reference and setting its capture time to zero and using a recursive method to estimate for the remaining microphones. Using the knowledge that certain microphone and loudspeaker pairs are near to each other, MDS is applied to obtain initial estimates. The MDS coordinates are then rotated and translated to satisfy coordinate restrictions that are set to fix a coordinate system. Finally, the coordinates are iteratively optimized.

[JSHU12] present a microphone array network self-calibration method that uses Direction of Arrival (DOA) measurements estimated from reverberant speech. The DOA measurements from all microphone arrays are collected to a cost function that has its minimum when the predicted DOA is close to the measured DOA. The measurements are collected from a speech source that moves around the space. The

drawback of this system is that the node has to be able to produce DOA information rather than raw audio.

[OKIS09] presents a self-localization method estimating TDOA from environmental speech. The method attacks the general self-localization problem using auxiliary functions that are optimized instead of the original cost function. The method is able to operate on asynchronously recorded signals. The solution of the optimization problem requires $(K-4)(L-4) \geq 9$, where K is the number of sources and L is the number of microphones. In practice the requirement is fulfilled by assuming a moving source and considering sound events at each time as a different source, which increases K easily. The method is able to estimate the positions of microphones, sources and temporal offsets. The drawback of this method is the large number of required iterations and possibility of getting stuck to a local minimum.

In [PMH12], a self-localization method designed for reverberant spaces is proposed. It is assumed that sound sources are surrounding the sensor network and as sound waves travel through a microphone pair, TDOA between them can be estimated. When a sound source is located at endfire directions (see Figure 3.1), the maximum TDOA value is observed corresponding to the distance between a microphone pair. The method is designed to take the reverberant environment into account by introducing a set of nonlinear techniques to detect TDOA outliers. Once pairwise microphone distances are estimated, MDS is used to convert distance information into relative Cartesian coordinates. The method does not require that sensors are synchronized as the unknown time offsets are cancelled out in the formulation.

2.3.3 Summary of Related Work

For many applications and as underlying technology self-localization is convenient and unobtrusive when using acoustic environmental signals. Self-localization using other measurements requires additional hardware and is therefore more inflexible than audio. However, there are many challenges conducting self-localization from environmental audio. These include complexity of the problem especially with increasing number of nodes, data association, reverberant environments, and asynchronous signals. Therefore, solving the self-localization problem benefits greatly if there is some knowledge about, e.g., node model, node motion, and sound field. Many proposed methods utilize a priori knowledge and are able to meet the required performance in the applications they are designed for. Completely blind cases are generally the hardest and possibly require several assisting subsystems such as Multiple Target Tracking (MTT) along with the core self-localization method.

Table 2.1 groups different types of self-localization systems that are relevant in the scope of this thesis, since they can use acoustic measurements[1]. The main categories are matrix factorization based methods, MDS based methods, Diffuse Sound Field based methods, and Blind methods.

Two kinds of sound field models are used: diffuse and point source. Diffuse sound is encountered in high-reverberant environments ($T_{60} > 400$ ms) such as cars while driven and office premises. Point source model, i.e., acoustic events from a single source, propagate along the direct path in low reverberant spaces. In reality, multipath propagation occurs, but the energy mainly travels via the direct path.

Especially in ad hoc sensor networks, the synchronization of audio streams measured at different locations is crucial. In practice this means connecting the input of each

1. However, many methods are themselves not necessarily limited to acoustic input.

Table 2.1. Taxonomy of acoustic self-localization methods.

	Matrix Factorization based methods	MDS based methods	Diffuse Sound Field methods	Blind methods
Sound field model	Point source	Point source	Diffuse	Point source
Synchronization	Estimated or side channel synchronization	Not needed	Side channel synchronization	Estimated
Solution	Closed-form+iterative	Iterative	Iterative	Iterative
Node model	microphones and sources are separated	Node consists of a microphone and a source	microphones and sources are separated	microphones and sources are separated
References	[BKÅ15],[BT04], [Thr05], [LW08], [PN08], [CDBM12]	[PMH12], [HF11], [P1]	[MLH08], [HPF$^+$09]	[OKIS09], [P2]

sensor to a single analog-to-digital converter, which can be inconvenient and at least requires additional hardware. Many proposed self-localization methods assume synchronization, or some methods use a special side channel (e.g. via RF). Yet again, additional protocol and/or hardware is required.

The choice of self-localization method is affected by the node model and the application. The methods can be iterative, closed-form, and a combination of iterative and closed-form solution. Often a closed-form solution is computed first after which iterative methods are used to force a set of constraints derived from a priori information on, e.g., the sound propagation model.

The matrix factorization based methods presented so far assume a point source model, time synchronization, or a side-channel synchronization. Often some assumption on whether a sound source is in the far-field or the near-field is used to simplify the problem.

MDS based methods assume that a node consists of a source and a sensor, which implies that they do not need accurate sample level synchronization. This results from cancellation of temporal offsets (see e.g. [PMH12]). However, rough alignment of signals collected from ad hoc networks is required in practice due to framewise processing of signals. The MDS algorithm used to estimate the positions from pairwise distance estimates is iterative.

Diffuse Sound Field methods are directed to high-reverberant environments. The node network consists of spatially separate sensors and sources.

Blind methods aim to estimate the positions of spatially separate sources, sensor, and temporal sensor offsets. The basic idea is to perform self-localization via optimization of the general self-localization problem. The input to the system is TDOA and data association information. The drawbacks of the blind methods stem from the nature of the optimization problem.

2.4 Multidimensional Scaling

Multidimensional scaling (MDS) is a technique that aims at representing the original data set $\boldsymbol{x}_i \in \mathbb{R}^{l\times 1}, i = 1, \ldots, n$ in a lower dimensional space while keeping the error between the newly represented data and the original data as low as possible (n denotes the number of samples). One of the obvious benefits of this process is visualization of the multidimensional data.

A similar technique that aims at dimensionality reduction is Principal Component Analysis (PCA). PCA retains the covariance matrix as precisely as possible [DHS01]. MDS aims to dimensionality reduction while retaining differences in the data. For

instance, if the data is clustered in some way, these clusters are preserved in the re-presentation obtained using MDS.

As input MDS has a *dissimilarity matrix* $D \in \mathbb{R}^{n \times n}$. That is, each element describes the difference between two data points. Therefore the dissimilarity matrix is by definition symmetric and its diagonal elements are zero. If the differences are metric D is also referred to as distance matrix.

MDS algorithms can be classified into different categories depending on interpretation of the dissimilarity matrix. *Classical* MDS assumes that the dissimilarities are distances and finds coordinates to explain them by minimizing a stress function called strain. *Metric* MDS generalizes the classical method, e.g., by introducing weights into the optimization, and relaxes the assumption on considering the dissimilarities as distances. *Non-metric* MDS further relaxes the assumption on dissimilarities, and non-metric models represent only the ordinal properties in the estimation, e.g., two entries $d_{12} = 5$ and $d_{34} = 4$ of the dissimilarity matrix are estimated in such a way that $\hat{d}_{12} > \hat{d}_{34}$, where $\hat{d}_{ij}$ denotes an element of the estimated dissimilarity matrix.

In general MDS estimates another data set $\boldsymbol{y}_i \in \mathbb{R}^{k \times 1}, i = 1, \ldots, n$, where k is the new dimension (typically $k < l$) and for which the estimated distance matrix $\hat{D} = (||y_i - y_j||)$ is similar to the distance matrix D.

It is important to note that also $\boldsymbol{z}_i = A\boldsymbol{y}_i + \boldsymbol{c}, i = 1, \ldots, n$ is a solution that fulfills the constraints set by MDS. The solution is thus non-unique and can be reflected (i.e. mirrored), rotated, or translated. A is referred to as rotation and reflection matrix, and $\boldsymbol{c}$ is a translation vector. The transformation that MDS is subject to is discussed in detailed in Section 2.5.

In summary, the fundamental idea of multidimensional scaling is to search a mapping that transforms the original data

$$X = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} \in \mathbb{R}^{n \times k} \text{ to the new data set } Y = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix} \in \mathbb{R}^{n \times l} .$$

In *metric* multidimensional scaling the entries in the dissimilarity matrix are representable, e.g., as distances in the Euclidean space. Then, the task is to search a configuration Y that minimizes the error

$$\sigma^2 = \sum_{i \neq j} (\hat{d}_{ij} - d_{ij})^2. \tag{2.6}$$

However, if the data is known to include errors, this implies that the actual distances are undiscoverable. The error function can be modified as follows

$$\sigma_r = \sqrt{\frac{\sum_{i \neq j} (\hat{d}_{ij} - d_{ij})^2}{\sum_{i \neq j} d_{ij}^2}}, \tag{2.7}$$

which is known as stress metric proposed by Kruskal [Kru64].

Kruskal's stress is also used in case the data is measured on an ordinal scale, that is, the entries in the dissimilarity matrix are measured by their relative order. This is known as non-metric multidimensional scaling[Wic03]. Besides (2.7), other error functions can be used to find the optimal configuration (see [BG97] for details). The optimization of (2.7), i.e., the estimation of the configuration, can be performed using the Shepard-Kruskal method [AWC+07].

Table 2.2. The properties and grouping of transformations by invariance.

Transformation Group	Transformation	Invariance
Rigid motion (isometry)	rotation reflection translation	distances
Similarity transformation	rotation reflection translation dilation (scaling)	ratio of distances

2.5 Coordinate Transformations and Graph Ambiguities

Methods that use the estimated distances between points (e.g. MDS) to determine their coordinates result in a representation that is non-unique. Rigid body transformations, like *rotation* and *reflection*, can be applied to the configuration without violating the conditions that led to the configuration estimate. Furthermore, *translation* can be applied without violations, since it means that all points are translated by exactly the same amount and in the same direction. Besides rigid motions (translation, rotation, and reflection), there is *dilation* (or scaling). It refers to the magnification or reduction of the whole configuration. Dilations, like rigid-body transformations, do not affect ratios of distances [BG97].

Dilations together with rigid motions are referred to as similarity transformations. The transformations preserve the shape, but not necessarily the size of the configuration. Table 2.2 (adapted from [BG97]) summarizes the transformation group, transformations and *invariance*. Invariance is a property of a configuration that is unchanged by the transformations listed in the respective column.

In Table 2.2 "ratio of distances" refers to the fact that if the distance of nodes i and j is doubled or trebled, also the distance between another node pair, say k and l, is doubled or trebled. Thus in MDS and similar methods even dilation has to be recovered if physical coordinates are of interest.

2.5.1 Flip and Flex Ambiguities

So called component-based methods localize the whole set of points as a shape or structure whereas single unit based methods localize each point sequentially.

Flip and flex ambiguities refer to alternative representations, or realizations, of the point configuration, i.e., the *component*, that still satisfy the same restrictions set by the distance measurements. An algorithm may converge to either of the representations, however, one of the two may differ drastically from the ground truth.

2.5.2 Flip ambiguity

Flip ambiguity occurs when there are points in the configuration that are collinear, that is, the points lie approximately on the same line. Then, a neighboring point can be flipped across the line without violating the distance constraints or changes in a cost function that is being optimized. [WLY+10]

Figure 2.2 illustrates the problem. The edge AD forms the "mirror" with respect to which vertex E is repositioned at E_f. The illustration is adapted from [MLRT04].

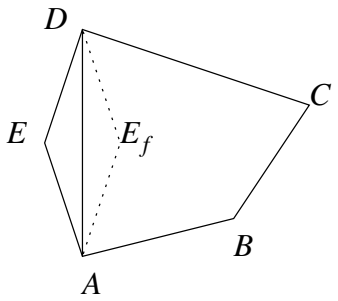


Figure 2.2. Flip ambiguity. The vertex E and its flipped representation E_f satisfy the distance constraints. However, one of them deteriorates the localization performance by possibly being far from the ground truth.

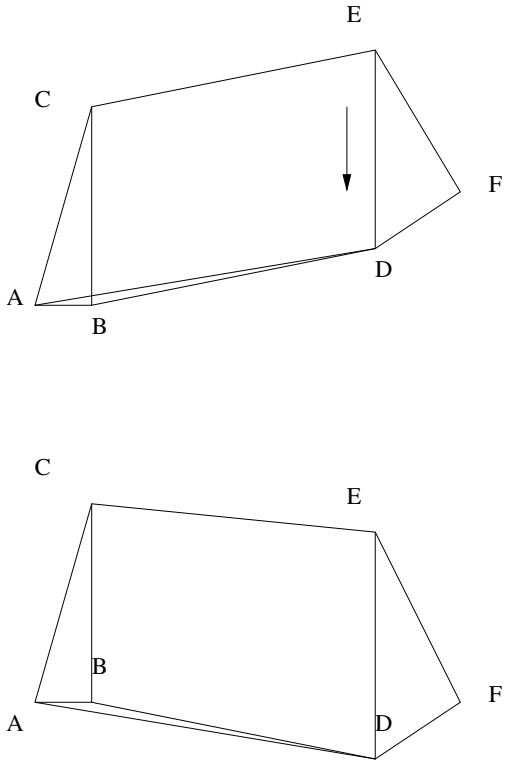


Figure 2.3. Discontinuous flex ambiguity. If edge AD is removed, the graph can flex in the direction of the arrow, taking on a different configuration after AD is reinserted while preserving all edge lengths, i.e., distance constraints.

The occurrence of the flex ambiguity has been researched in [KFM09]. Methods for detecting the existence of ambiguous realization have been developed to mitigate the problem. In [MLRT04], the authors propose the probabilistic method of *robust quadrilaterals*, where the point localization is seen as a two-dimensional graph realization problem and it is designed to localize the points using approximate edge lengths as input. Ultimately, the point locations are discovered up to a global rotation and translation.

2.5.3 Discontinuous Flex Ambiguity

This ambiguity occurs when the removal of one edge will allow part of the graph to be flexed to a different configuration and the removed edge reinserted with the same length. This type of deformation is distinct from continuous flex ambiguities which are present only in non-rigid graphs [MLRT04]. Figure 2.3 illustrates the

discontinuous flex ambiguity resulting in two different configurations with the same distance information. The illustration is adapted from [MLRT04].

2.6 The Kalman Filter

A brief introduction of the Kalman filter [Kal60] is presented, which is one of utilized techniques in this thesis. A comprehensive and an intuitive tutorial to the Kalman and General Kalman filters can be found e.g. in [WB95].

The Kalman filter is often used to adaptively learn, or model, a linear or a non-linear system. Ideally, the trained filter produces the same output as the real system with any defined input for the target system.

The Kalman filter models a system by measurement, state and update variables. The measurement is the value read from the system output. The state variables present the system internal condition and the update variables are used to update the state variables. In this sense, the Kalman filter is system with feedback that corrects the system towards the desired output. Figure 2.4 presents the cycle of the Kalman filter.

From the self-localization viewpoint, *system state*, or filter state, represents the locations or variable that reflects the locations of sensors and sources. The *measurement* is the quantity or a set of quantities that are being observed from the system. For instance, measurement can be sound intensity level or time difference of arrival. From the viewpoint of a filter it is irrelevant what is being measured. However, the quantity that is measured may have properties that lead to undesired operation of the Kalman filter if the assumptions and the parameters of the filter deviate too much from the actual phenomenon. For instance, the measurement may originate from other targets, be deteriorated by background noise and environmental conditions (e.g. reverberation in case of sound), and suffer from possible errors caused by measurement equipment.

Using the terminology of Figure 2.4, there are two kinds of equations related to the Kalman filter operation: time update equations and measurement update equations. The time update equations map the current state to obtain a priori estimates for the state variables in the next step. The measurement update equations connect the feedback (the new measurement) with the a priori estimate to get the posterior estimate of the state. Thus "Corrector" (see Figure 2.4) adjusts the estimates of "Predictor" into the correct direction by providing additional information. The state and covariance estimates should be closer to the true values since more information is added when updating measurements.

2.6.1 The Discrete Kalman Filter System Model

The Kalman filter models the state $\boldsymbol{x}_k \in \mathbb{R}^{n\times 1}$ of a system according to

$$\boldsymbol{x}_k = A\boldsymbol{x}_{k-1} + B\boldsymbol{u}_k + \boldsymbol{w}_{k-1}, \tag{2.8}$$

where $A \in \mathbb{R}^{n\times n}$ maps the previous state at $k-1$ of the system to the current state at k. $u_k \in \mathbb{R}^{l\times 1}$ is an input to the system. The matrix $B \in \mathbb{R}^{n\times l}$ maps the input to the current state. The input is omitted if there is no active element contributing to the state. $\boldsymbol{w}_{k-1}$ is the process noise since the previous step $k-1$.

The measurement $\boldsymbol{y}$ or the output of a system is modeled as

$$\boldsymbol{y}_k = H\boldsymbol{x}_k + \boldsymbol{v}_k, \tag{2.9}$$

where $H \in \mathbb{R}^{m \times n}$ maps the state to the system output. $\boldsymbol{v}_k$ is the measurement noise at the current step k. Also the matrices in (2.8) and (2.9) may change over time, which is then indicated with notation A_k, B_k, and H_k.

2.6.2 The Adjustment of the Kalman Filter

The question on how to adjust the filter to produce the desired output involves choosing the criterion. Intuitively, an error signal is used to achieve this. $\hat{\boldsymbol{x}}_k^-$ denotes the a priori state estimate at k. The posterior error estimate, i.e., the information provided by the measurement $\boldsymbol{y}_k$ used to update the a priori $\hat{\boldsymbol{x}}_k^-$ is denoted as $\hat{\boldsymbol{x}}_k$. [2] A priori error and a posteriori error are defined as follows.

$$\begin{aligned} \boldsymbol{e}_k^- &= \boldsymbol{x}_k^- - \hat{\boldsymbol{x}}_k^- \\ \boldsymbol{e}_k &= \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k \end{aligned} \tag{2.10}$$

The covariance matrices of the state errors is written as

$$\begin{aligned} P_k^- &= \mathsf{E}\{\boldsymbol{e}_\mathsf{k}^- \boldsymbol{e}_\mathsf{k}^{-\mathsf{T}}\} \\ P_k &= \mathsf{E}\{\boldsymbol{e}_\mathsf{k} \boldsymbol{e}_\mathsf{k}^\mathsf{T}\}. \end{aligned} \tag{2.11}$$

The posterior state estimate $\hat{\boldsymbol{x}}_k$ is updated using the prior information and the current measurement. The posterior estimate is written as

$$\hat{\boldsymbol{x}}_k = \hat{\boldsymbol{x}}_k^- + K_k(\boldsymbol{y}_k - H\hat{\boldsymbol{x}}_k^-), \tag{2.12}$$

where $K_k \in \mathbb{R}^{n \times m}$ is *Kalman gain* and $(\boldsymbol{y}_k - H\hat{\boldsymbol{x}}_k^-)$ is *innovation*. The latter is the difference between the actual and the predicted measurement. The Kalman gain is to be adjusted so that the posterior covariance is minimized. The minimization is performed by first substituting (2.12) into the definition of posterior error (2.10), substituting it to the definition of posterior covariance (2.11), performing the expectation operation, differentiating the trace of the result with respect to K_k, setting the result equal to zero, and then solving it for K_k [WB95]. The details of obtaining K_k can be found e.g. in [BH97].

The resulting gain is written as follows.

$$K_k = P_k^- H^T (H P_k^- H^T + R)^{-1} \tag{2.13}$$

R denotes the measurement error covariance.

The illustration in Figure 2.4 summarizes the operation of the discrete Kalman filter. Note that in the beginning the state of the filter has to be initialized. Choosing the initial values is non-trivial and any a priori knowledge should be utilized.

2.7 Matrix Decompositions

Matrix decompositions are useful, for instance, finding an inverse of a matrix. This results from the fact that often the constituent matrices are special matrices that are easier calculate analytically than the original matrix.

2. The minus in the superscript of a variable in the theory refers always to a priori estimate

Time Update ("Predict")	Measurement Update ("Correct")
$\hat{\boldsymbol{x}}_k^- = A\hat{\boldsymbol{x}}_{k-1} + B\boldsymbol{u}_k$	$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1}$
	$\hat{\boldsymbol{x}}_k = \hat{\boldsymbol{x}}_k^- + K_k(\boldsymbol{y}_k - H\hat{\boldsymbol{x}}_k)$
$P_k^- = \mathsf{E}\{\mathsf{e}_\mathsf{k}^- \mathsf{e}_\mathsf{k}^{-\mathsf{T}}\} = AP_{k-1}A^T + Q$	$P_k = (I - K_k H)P_k^-$

Figure 2.4. The Kalman filter update equations. The two phases keep repeating one after the other until some stopping criterion is met.

2.7.1 Singular Value Decomposition (SVD)

Any matrix A can be presented as $A = UDV^H$ [GVL96], where matrix U represents the *column space* and matrix V the *row space* of A, respectively. The column space is defined as follows. Let B be an $m \times n$ matrix and $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n$ its n columns and they are called column vectors. A linear combination of these is $\beta_1\boldsymbol{c}_1 + \beta_2\boldsymbol{c}_2 +, \ldots, +\beta_n\boldsymbol{c}_n$, where $\beta_i, i = 1, \ldots, n$ are scalars. A set of all such combinations of $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n$ is called the column space of B [Gro95]. The row space is defined similarly, but the column vectors are replaced with row vectors $\boldsymbol{r}_i, i = 1, \ldots, m$. D is a matrix with non-negative real numbers on the diagonal and they are the singular values of A. H denotes the conjugate-transpose operation.

2.7.2 Eigenvalue Decomposition (EVD)

Let B be an $N \times N$ matrix with linearly independent eigenvectors $\boldsymbol{q}_i, i = 1, \ldots, N$. Then B can be factorized as $B = Q\Lambda Q^{-1}$, where Q is an $N \times N$ matrix and its ith column is the eigenvector $\boldsymbol{q}_i$ of B. Λ is the diagonal matrix and its ith diagonal element λ_i is the ith eigenvalue value of B.

Matrix Q is by definition orthogonal, since its columns are orthogonal. Thus, it is always invertible and $Q^T = Q^{-1}$. As a consequence, the inverse of the original matrix is written as follows: $B^{-1} = Q\Lambda^{-1}Q^{-1} = Q\Lambda^{-1}Q^T$, where $\Lambda^{-1} = \text{diag}\lceil\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_N}\rfloor$.

Note especially that B has to be a square matrix, which is one of the essential differences between this method and SVD.

2.7.3 Principal Component Analysis (PCA)

One of the main uses of PCA is in dimensionality reduction in such a way that the reconstructed lower dimensional data contains as much as possible of original information [JW92]. Principal components can be thought of as axes of an imaginary ellipsoid surrounding a data cloud. Mathematically PCA is an orthogonal linear transformation that maps data into a new coordinate system, where the greatest variation of data shows on the first axis (i.e. the first principal component), and the second largest variation on the second axis, and so on. All the axes are orthogonal to each other.

The procedure to perform PCA via eigenvalue decomposition is as follows. The covariance matrix C of the data points is estimated, i.e., $C = XX^T$, where $X = \big[x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}\,\big]^T$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then, eigenvalue decomposition is done on the covariance matrix resulting in $C = MDM^T$ (see Section 2.7.2). The principal components are the columns of M. Using the principal components, the data X is projected on the principal axes:

$$\tilde{X} = XM, \tag{2.14}$$

where $M = [m_{1c}, \ldots, m_{nc}]^T, c = 1, \ldots, n'$. n' is the new dimensionality typically chosen to be $n' < n$. $\tilde{X}$ is the reconstructed data in the new dimension n'. The contribution of each principal component in the reconstruction decreases as c increases. A detailed presentation of PCA can be found in [Jol02].

2.7.4 Rank-N Matrix Factorization

Let us define a matrix $A \in \mathbb{R}^{m \times p}$ with rank $r \leq \min\{m, p\}$. Any matrix can be presented with its SVD: $A = UDV^H$, where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{p \times p}$. Now if

$$D = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix},$$

where Σ is a diagonal matrix with strictly positive singular values, i.e. [Bjfrm[o]–5]:

$$\Sigma = \text{diag}\lceil \sigma_1, \ldots, \sigma_r \rfloor,$$

then the full-rank factorization and thus rank-r approximation for SVD of the original A is

$$U_r \Sigma V_r^H = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r] \Sigma [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r]^H,$$

where U_r and V_r are submatrices of U and V keeping the first r columns, respectively.

2.8 Ill-posed Problems and Their Solution

The problem (3.13) is of form

$$A\boldsymbol{x} = \boldsymbol{m}, \tag{2.15}$$

where $A \in \mathbb{R}^{m \times n}$, $\boldsymbol{x} \in \mathbb{R}^{n \times 1}$ and $\boldsymbol{m} \in \mathbb{R}^{m \times 1}$.

In contrast to an ill-posed problem, a well-posed problem is defined as follows. (H1) the solution $\boldsymbol{x}$ exists, (H2) the solution is unique, and (H3) the solution depends continuously on the data. H1, H2 and H3 are Hadamard's conditions [Jac04]. The failure of H1 and H2 is noticeable often in problem formulation, whereas the failure of H3 is more hazardous, since the problem may be solved but the solution can be erroneous.

2.8.1 Truncated SVD

The regularization is based on performing the matrix inverse via SVD (see Section 2.7.1).

Matrix A can only be classically inverted if $m = n$ and all its singular values are strictly positive. Otherwise the inverse is written as $A^{-1} = VD^{-1}U^T$, where

$$D^{-1} = \text{diag}\lceil \tfrac{1}{d_1}, \tfrac{1}{d_2}, \cdots, \tfrac{1}{d_j}, \cdots, \tfrac{1}{d_{\min\{m,n\}}} \rfloor. \tag{2.16}$$

$\text{diag}\lceil \cdot \rfloor$ denotes a diagonal matrix.

Let r be the biggest index for which $d_r > 0$: $r = \max\{r | 1 \leq j \leq n, d_r > 0\}$. The pseudoinverse of A is written as $A^+ = VD^+U^T$, where

$D^+ = \text{diag}\lceil \frac{1}{d_1}, \frac{1}{d_2}, \cdots, \frac{1}{d_j}, \cdots, \frac{1}{d_r}, 0, \cdots, 0 \rfloor$ and the solution of problem (2.15) is

$$\boldsymbol{x} = A^+ \boldsymbol{m}. \tag{2.17}$$

If d_j is very small, H3 may fail. Therefore another pseudoinverse is defined as $D_\alpha^+ = \text{diag}\lceil \frac{1}{d_1}, \frac{1}{d_2}, \cdots, \frac{1}{d_j}, \cdots, \frac{1}{d_{p(\alpha)}}, 0, \cdots, 0 \rfloor$, where $p(\alpha) = \max\{j | d_j \geq \alpha\}$ and the approximate solution of problem (2.15) is

$$\boldsymbol{x}_\alpha = A_\alpha^+ \boldsymbol{m}. \tag{2.18}$$

The tolerance α can be used to prevent H3 from failing.

2.8.2 Tikhonov Regularization

Let us define

$$\boldsymbol{x}_\delta = \underset{\boldsymbol{x}}{\text{argmin}} \{||A\boldsymbol{x} - \boldsymbol{m}||^2 + \delta ||\boldsymbol{x}||^2\}, \tag{2.19}$$

where $\delta > 0$ is the regularization parameter. δ is used to balance between two requirements for the solution $\boldsymbol{x}_\delta$: (A) the error should be as small as possible (term $||A\boldsymbol{x} - \boldsymbol{m}||^2$), and (B) $||\boldsymbol{x}||$ should be as small as possible.

The solution is obtained via SVD of matrix A. The regularized inverse of (2.16) is

$$D_\delta = \text{diag}\lceil \frac{d_1}{d_1^2 + \delta}, \ldots, \frac{d_{\min\{m,n\}}}{d_{\min\{m,n\}}^2 + \delta} \rfloor \tag{2.20}$$

and the solution is

$$\boldsymbol{x}_\delta = V D_\delta U^T \boldsymbol{m}. \tag{2.21}$$

When $\delta \to 0$, the solution is close to the minimum norm solution, which may deviate from the true value significantly since the magnitude of the solution is not limited in any way. When $\delta \to \infty$, the solution is not fitted to model the data at all since it forces the solution towards zero. [Jac04]

Chapter 3

Acoustic Self-localization

Acoustic self-localization aims at determining absolute or relative positions of nodes. Here, a node can be a microphone, a sound source, or an entity consisting of a microphone and source. The primary interest in this work is to localize nodes that contain microphones. Once positions are determined, the nodes with microphones form an ad hoc microphone array. The positioning of nodes that contain only a sound source may be useful in some cases, since it can be used to assist solving the positions of nodes of interest (as in [PPH14]).

This chapter is organized as follows. Sections 3.1-3.3 present the signal model and the acoustic self-localization problem formulation. Section 3.4 presents a summary and the main results of the author's early work on acoustic self-localization [P3]. Section 3.5 and Section 3.6 present two self-localization methods that provide the basis to the author's work [P2], which, in turn, is summarized in Section 3.7. A competitive method that uses environmental sound for self-localization is presented in 3.8. Finally, Section 3.9 presents the author's latest work [P1] on acoustic self-localization.

3.1 Signal Model

Let $\mathbf{m}_i \in \mathbb{R}^3$ be the ith node position and $i \in 1, \ldots, N$. In an anechoic room the signal $m_i(t)$ can be modeled as a delayed source signal $s_k(t)$ as

$$m_i(t) = s_k(t - \Delta_{i,k}) + n_i(t), \tag{3.1}$$

where t is time, $k \in [1, \ldots, K]$ denotes active node index, $n_i(t)$ is noise component, and $\Delta_{i,k}$ is TOA from active node k to the ith node

$$\Delta_{i,k} = c^{-1}\|\mathbf{s}_k - \mathbf{m}_i\| + \delta_i, \tag{3.2}$$

where δ_i is unknown time offset, c is speed of sound, and $\mathbf{s}_k \in \mathbb{R}^3$ are source positions.

3.2 Spatial information: TDOA and TOA

Clearly, TOAs are characterized by node positions. However, TOA cannot be measured directly in the passive self-localization problem, but the differences between each node pair TOA, that is TDOA, can be estimated. TDOA between microphone pair $\{i, j\}$ for source k is

$$\Delta_{i,j,k} \triangleq \Delta_{i,k} - \Delta_{j,k} = c^{-1}(\|\mathbf{s}_k - \mathbf{m}_i\| - \|\mathbf{s}_k - \mathbf{m}_j\|) + \delta_{ij}, \tag{3.3}$$

where pairwise time offset is $\delta_{ij} \triangleq \delta_i - \delta_j$. The time offsets result from the fact that, in an ad hoc network, nodes have their own time axis. It is acknowledged that ad hoc homogeneous devices in general have drift in their analog-to-digital converter. The drift is assumed to be small compared to the TDOA and therefore it is omitted in the formulation above.

3.3 Problem Formulation

The self-localization problem in acoustic sensor networks can be stated as follows. Estimate the coordinates of all nodes in the network that are either capable of receiving or emitting sound or both. More specifically, the coordinates of each microphone and sound speaker are estimated. The problem can be formulated as follows.

$$\langle S^*, M^*, \boldsymbol{\delta} \rangle = \underset{S,M,\boldsymbol{\delta}}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left\{ c^{-1} \|\boldsymbol{s}_k - \boldsymbol{m}_i\| + \delta_i - (c^{-1} \|\boldsymbol{s}_k - \boldsymbol{m}_j\| + \delta_j) - \Delta_{i,j,k} \right\}^2, \tag{3.4}$$

where $S = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_K]$, $M = [\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_N]$, $\boldsymbol{\delta} = [\delta_1, \delta_2, \ldots, \delta_N]^T$. $\boldsymbol{\delta}$ are temporal offsets of the microphones.

The key to the self-localization problem is Δ, which is utilized in many self-localization methods discussed below. Δ contains range information between node pairs of the network. Solving the general self-localization problem of (3.4) involves solving the temporal offsets between the receiving nodes. Many self-localization methods ignore the offsets $\boldsymbol{\delta}$ by assuming that the devices are synchronized implying $\delta_i = \delta_j, \forall i = 1, \ldots, N$. Also other assumptions are often set to ease solving (3.4) such as using special anchor nodes the coordinates of which are known.

Using acoustic measurements, Δ can be measured e.g. using TDOA estimation. Since the measurements in the general self-localization problem are relative (e.g. time differences between all receiving node pairs), the configuration of nodes can be translated, rotated, flipped across any coordinate axis and still satisfy (3.4). The solution can be fixed to a chosen coordinate system with appropriate transformation matrices [Fol96]. See Section 2.5 for coordinate transformations.

3.4 Robust Self-localization Solution for Meeting Room Environments [P3]

[P3] is a self-localization method intended for localization of multiple microphones. This enables deploying a room with microphones at unknown locations for source localization and other spatial audio applications. [P3] uses multiple calibration sources at known locations to localize the microphones and the number of needed calibration sources depends on the number of microphones to be localized. The system uses the framework of [Fra06] and introduces robustness in presence of timing errors via regularization of possibly ill-posed problem [Han97].

For K sound sources at known locations, [P3] simplifies the problem setting of (3.4). Furthermore, all microphones are synchronized and therefore the pairwise offsets

between all microphone pairs are zero. The self-localization problem (3.4) is simplified to the form

$$M^* = \underset{M}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left\{ c^{-1} \left\| \boldsymbol{s}_k - \boldsymbol{m}_i \right\| - c^{-1}(\left\| \boldsymbol{s}_k - \boldsymbol{m}_j \right\|) - \Delta_{i,j,k} \right\}^2. \quad (3.5)$$

3.4.1 Spatial Information

In [P3] the TDOA between a reference sensor denoted 1 and sensor i is written

$$\Delta_{i1k} = \Delta_{ik} - \Delta_{1k}, \quad (3.6)$$

where Δ_{1k} denotes the time at which the reference sensor node detects the sound. Δ_{ik} denotes the detection time of ith sensor. $i = 2, \ldots, N$ and N is the number of sensors in the network.

3.4.2 The Least Squares Estimator for Microphone Positions

The distance d between the emitting source at $\boldsymbol{s} = [s^x, s^y, s^z]^T$ and the reference node at $\boldsymbol{m}_1 = [m_1^x, m_1^y, m_1^z]^T$ can be written

$$d = ||\boldsymbol{m}_1 - \boldsymbol{s}||, \quad (3.7)$$

where $d = c \cdot \Delta_{1k}$ and c is the speed of sound.

Similar equations to (3.7) can be defined for sensors $\boldsymbol{m}_i, i = 2, \ldots, N$ as follows

$$\begin{aligned} d^2 &= ||\boldsymbol{m}_1 - \boldsymbol{s}||^2 \\ (d + c\Delta_{21k})^2 &= ||\boldsymbol{m}_2 - \boldsymbol{s}||^2 \\ (d + c\Delta_{31k})^2 &= ||\boldsymbol{m}_3 - \boldsymbol{s}||^2 \\ (d + c\Delta_{41k})^2 &= ||\boldsymbol{m}_4 - \boldsymbol{s}||^2 \\ &\vdots, \end{aligned} \quad (3.8)$$

where Δ_{i1k} denotes detection time difference between $\boldsymbol{m}_1$ and $\boldsymbol{m}_i$. The right side of (3.8) is the squared Euclidean distance between a node and a source.

Each equation in (3.8) can be solved independently [Fra06]. For an unknown node position, the problem is formulated as follows.

$$\begin{aligned} (d_i + c\Delta_{i11})^2 &= ||\boldsymbol{m}_i - \boldsymbol{s}_1||^2 \\ (d_i + c\Delta_{i12})^2 &= ||\boldsymbol{m}_i - \boldsymbol{s}_2||^2 \\ (d_i + c\Delta_{i13})^2 &= ||\boldsymbol{m}_i - \boldsymbol{s}_3||^2 \\ &\vdots, \end{aligned} \quad (3.9)$$

where $d_i = ||\boldsymbol{m}_i - \boldsymbol{s}||$. The location of the reference source $\boldsymbol{s}$ defines the origin for the solution.

Expanding and rearranging of (3.9) results in

$$\begin{aligned} c^2\Delta_{i11}^2 + 2c\Delta_{i11}d_i + d_i^2 - \boldsymbol{s}_1^T\boldsymbol{s}_1 &= \boldsymbol{m}_i^T\boldsymbol{m}_i - 2\boldsymbol{m}_i^T\boldsymbol{s}_1 \\ c^2\Delta_{i12}^2 + 2c\Delta_{i12}d_i + d_i^2 - \boldsymbol{s}_2^T\boldsymbol{s}_2 &= \boldsymbol{m}_i^T\boldsymbol{m}_i - 2\boldsymbol{m}_i^T\boldsymbol{s}_2 \\ &\vdots \end{aligned} \quad (3.10)$$

Rewriting the first row of (3.10) yields

$$c^2\Delta_{i11}^2 + 2c\Delta_{i11}d_i + (\boldsymbol{m}_i - \boldsymbol{s})^T(\boldsymbol{m}_i - \boldsymbol{s}) - \boldsymbol{s}_1^T\boldsymbol{s}_1 = \boldsymbol{m}_i^T\boldsymbol{m}_i - 2\boldsymbol{m}_i^T\boldsymbol{s}_1 \tag{3.11}$$

.

Rearranging equations in (3.10) and using (3.11) yields the matrix representation:

$$\begin{bmatrix} 2(\boldsymbol{s}_1^T - \boldsymbol{s}^T) & 2c\Delta_{i11} \\ \vdots & \vdots \end{bmatrix} \begin{bmatrix} \boldsymbol{m}_i \\ d_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{s}_1^T\boldsymbol{s}_1 - c^2\Delta_{i11}^2 - \boldsymbol{s}^T\boldsymbol{s} \\ \vdots \end{bmatrix}, \tag{3.12}$$

or, in symbolic form,

$$M_i \begin{bmatrix} \boldsymbol{m}_i \\ d_i \end{bmatrix} = \boldsymbol{b}_i. \tag{3.13}$$

The solution of (3.13) is $\boldsymbol{m}_i^{d_i} = M_i^{-1}\boldsymbol{b}_i$, where M_i^{-1} denotes the inverse matrix and $\boldsymbol{m}_i^{d_i} = \begin{bmatrix}\boldsymbol{m}_i^T & d_i\end{bmatrix}^T$, $\boldsymbol{m}_i^{d_i} \in \mathbb{R}^{4\times 1}$.

In practice, the inverse matrix in general cannot be solved and therefore the pseudoinverse of M_i is calculated:

$$\boldsymbol{m}_{i,\mathrm{LS}}^{d_i} = \underbrace{(M_i^T M_i)^{-1} M_i^T}_{M_i^+}\boldsymbol{b}_i, \tag{3.14}$$

where M_i^+ denotes the pseudoinverse. $\boldsymbol{m}_{i,\mathrm{LS}}^{d_i}$ is the least squares solution of (3.13). It is assumed that M_i is non-singular. However, especially with random sensor deployments the non-singularity can not be guaranteed and the problem can be *ill-posed*.

[P3] introduces the use of Truncated Singular Value Decomposition (TSVD) and Tikhonov Regularization to handle possible ill-posedness of (3.14). TSVD and Tikhonov Regularization are presented in Section 2.8.1 and 2.8.2, respectively.

3.4.3 The Performance of [P3]

The focus of the evaluation is on comparing the accuracy of node positioning between pseudoinverse, truncated SVD, and Tikhonov regularization in cases where (2.15) is prone to ill-posedness. Such cases occur when there is error in time information and error in calibration source positions (i.e., the expected spatial distances between estimated and true positions) combined with few measurements. Therefore, the performance is evaluated first using five calibration sources (the minimum for the existence of a solution is four), and with varying timing and calibration source positioning errors. The performance is evaluated using 50000 random simulated deployments of the sensor network.

When there are no timing or positioning errors, the pseudoinverse is the most accurate in positioning the unknown nodes. But when normally distributed noise with 10^{-2} s standard deviation is added to timing information, the pseudoinverse estimator achieves an error of 1080 cm, whereas truncated SVD and Tikhonov regularization achieve errors of 220 cm and 131 cm, respectively.

A 10 cm calibration source positioning error, with no timing error, results in a 305 cm positioning error of the unknown nodes for pseudoinverse estimator. Truncated SVD and Tikhonov regularization achieve 73 cm and 70 cm errors, respectively.

3.4.4 Conclusions on Robust Self-localization Solution for Meeting Room Environments

[P3] focused on calibration of unknown microphones with inaccuracies in calibration source positions and TDOA errors. Based on the experiments, it can be stated that there is a clear benefit of using a regularization method over the baseline pseudoinverse estimator in case of possible ill-posedness of the problem. Work in [P3] is highly focused on the microphone position calibration and therefore many simplifications in the self-localization problem (3.4) are made including assumed knowledge of source positions and the assumption of synchronized audio streams.

3.5 Affine Structure From Sound

Structure from sound (SFS) means localizing simultaneously a set of sound sources and a set of sensors whose locations and emission times are unknown. Furthermore, SFS relies on environmental sounds that are not generated by the sensor network. Thrun presents in [Thr05] a self-localization method for the SFS scenario.

In [Thr05], the SFS is problem is eased by assuming clock synchronization between sensors. This is generally not true without accurate infrastructure dedicated to providing synchronization. Acoustic events are assumed to occur sufficiently sparse in time so that there is no data association problem. Also reverberation is assumed to be absent, which in practice corresponds to valid distance estimates all the time.

The structure from sound problem in [Thr05] is formulated as an optimization problem and the locations of microphones, sound sources, and emission times of acoustic events are unknown. The optimization task is cumbersome since the cost function has several local minima to which common optimization methods may get stuck. To avoid this, it assumed that the incident angle of acoustic events is the same for all the sensors. This implies that acoustic events occur relatively far away from the sensors.

The input data consists of detection times for each sensor-source pair. That is, the data matrix is defined as follows.

$$D = \begin{bmatrix} d_{1,1} & \dots & d_{1,M} \\ \vdots & \ddots & \vdots \\ d_{N,1} & \dots & d_{N,M} \end{bmatrix}, \tag{3.15}$$

where $d_{i,j}$ denotes the detection time between ith sensor and jth source. $d_{i,j}$ is defined as follows.

$$d_{ij} = t_j + c^{-1} \left\{ \left| \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} a_j \\ b_j \end{pmatrix} \right| \right\}, \tag{3.16}$$

where t_j is the emission time of source j and c is speed of sound. x_i and y_i denote the location of the sensor in a 2D Euclidean coordinate system. a_j and b_j denote the location of the sound source in the same coordinate system as that of the sensors. The formulation here is for the 2D case but can be extended to 3D.

Redefining the data matrix using the first sensor $i = 1$ as a reference and setting $x_1 = y_1 = 0$, the distance between sensor-source pairs can be written as

$$\Delta_{i,j} = t_j + c^{-1} \left| \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} a_j \\ b_j \end{pmatrix} \right| - t_j - c^{-1} \left| \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} a_j \\ b_j \end{pmatrix} \right| \tag{3.17}$$

$$= c^{-1} \left\{ \left| \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} a_j \\ b_j \end{pmatrix} \right| - \left| \begin{pmatrix} a_j \\ b_j \end{pmatrix} \right| \right\} \tag{3.18}$$

This results in cancellation of the emission times t_j, which eases the optimization problem.

The input data matrix becomes

$$\Delta = \begin{bmatrix} d_{2,1} - d_{1,1} & \dots & d_{2,M} - d_{1,M} \\ \vdots & \ddots & \vdots \\ d_{N,1} - d_{1,1} & \dots & d_{N,M} - d_{1,M} \end{bmatrix}. \tag{3.19}$$

Note that the first row is omitted since it is all zeros. Thus the data matrix size is $(N-1) \times M$.

3.5.1 A Least-squares Problem Definition of SFS

The optimization problem discussed above is as follows.

$$\langle A^*, X^* \rangle = \underset{X,A}{\operatorname{argmin}} \sum_{j=2}^{N} \sum_{i=1}^{M} \left\{ \left| \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} a_j \\ b_j \end{pmatrix} \right| - \left| \begin{pmatrix} a_j \\ b_j \end{pmatrix} \right| - \Delta_{i,j} \right\}^2, \tag{3.20}$$

where A denotes the acoustic event location matrix and X is the sensor location matrix and they are written

$$X = \begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_N & y_N \end{bmatrix} A = \begin{bmatrix} a_1 & b_1 \\ \vdots & \vdots \\ a_N & b_N \end{bmatrix}.$$

The global minimum of (3.20) is the solution to the SFS problem.

The emission times were canceled out in the above formulation. To estimate t_j, (3.16) is utilized and having source locations and microphone locations solved using (3.20), the emission times can be solved as follows.

$$T^* = \frac{1}{N} \sum_{i=1}^{N} d_{ij} - t_j + c^{-1} \left| \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} a_j \\ b_j \end{pmatrix} \right|, \tag{3.21}$$

where $T = [t_1, \dots, t_N]^T$.

The solution of the SFS problem (3.20) is not unique, but can be rotated, mirrored, and translated. A more serious issue with (3.20) is that minimization algorithms tend to get stuck in local minima. In [Thr05] this problem is mitigated by reducing the number of variables in the optimization task. This is obtained by writing the distance between a sensor and a source as function of sound event incidence angle. Also, it assumed that sources are far-field, which yields that the incidence angles of an acoustic event are the same for all the sensors. Using SVD and the above assumptions the original SFS problem is reduced to optimization of only four variables (2D case) of a constraint matrix.

3.5.2 Performance of Affine Structure From Sound Self-localization

The method [Thr05] is tested with simulations and with real data measurements. The simulated data consists of 1000 random configurations, in which sensors were sampled uniformly within an interval of 1×1 m and sound sources were placed at varying ranges from 2 m to 10 m. The author highlights the proposed algorithm by

comparing it to gradient descent optimization of the problem (3.20). Based on the curves presented in [Thr05], the positioning error of the gradient descent method increases with the number of sources and microphones whereas the error of the proposed method remains approximately same. As the distance between the sources and the microphones is increased, the error decreases towards zero as expected with the far-field assumption. The error of the baseline gradient descent method is lower than the proposed method when the microphones and sources are near and thus the far-field assumption is violated. The real data experiments are made with Crossbow sensor motes similar to [MICa], which are placed arbitrarily around the space and with mutual distances ranging from 14 cm to 125 cm. As to the simulation results, the real data results are presented using performance curves and based on them it can be stated that the performance is similar to that of the simulated data.

3.5.3 Conclusions on Affine Structure From Sound Self-localization

The system [Thr05] estimates microphone positions, source positions and the emission times of acoustic events. The system is a matrix factorization based self-localization method (see Table 2.1). SVD factorization is used and its use is made possible by assuming that sound sources are so far away from the microphones that the incident angle of acoustic events is approximately the same. This enables simplification of the original problem (3.20) and the resulting simplified problem involves four parameters in 2D space. The microphone positioning results obtained with simulations and real data measurements indicate that the proposed method is far better than the baseline gradient descent method especially with increasing microphone count and source-microphone distances. The author of [Thr05] acknowledges that there is a need to solve the data association issue and evaluation in 3D space. The temporal offset estimation is not needed with [MICa] type hardware since radio frequencies can be used as a side channel for synchronization.

3.6 Passive Self-localization of Microphones Using Ambient Sounds

In [PMH12] a self-localization method is presented that estimates the distance between each microphone pair in the network from TDOA of acoustic events that occur on the axis passing through the two microphones and the location of the acoustic event. Such occurrences of acoustic events are referred to as endfire directions. In Figure 3.1 such a case is illustrated. The method [PHM13] is presented briefly below using the same notation as in the original article.

Pairwise Distance Estimation

The center point on the axis connecting microphone pair i and j located at $\boldsymbol{m_i}$ and $\boldsymbol{m_j}$, respectively, is $\boldsymbol{r} = \frac{1}{2}(\boldsymbol{m_i} + \boldsymbol{m_j})$. Assuming that sound events occur in the far field, the propagation of sound can be modeled as plane waves. The direction of propagation is represented by vector $\boldsymbol{k} \in \mathbb{R}^{3\times1}$ with length $||c^{-1}||$, where c is the sound speed. The time of arrival of a wave front at microphone i is

$$\tau_i = \langle \boldsymbol{m_i} - \boldsymbol{r}, \boldsymbol{k} \rangle + \Delta_i \tag{3.22}$$

[Zio94]. $\langle \cdot, \cdot \rangle$ is the dot product and Δ_i the microphone time offset to a reference time (i.e., to a reference microphone). If all sensors i are synchronized, then $\Delta_i = 0 \forall i$.

The TDOA between microphones i and j is

$$\tau_{ij} = \tau_i - \tau_j = \langle \boldsymbol{m_i} - \boldsymbol{m_j}, \boldsymbol{k} \rangle + \Delta_{ij}, \tag{3.23}$$

where $\Delta_{ij} = \Delta_i - \Delta_j$. The wave front propagation vector $\boldsymbol{k}$ in an endfire case (3.1) can be written as follows.

$$\boldsymbol{k} = \beta c^{-1} \frac{\boldsymbol{m}_j - \boldsymbol{m}_i}{||\boldsymbol{m}_j - \boldsymbol{m}_i||}, \tag{3.24}$$

where $\beta \in \{-1, 1\}$. In Figure 3.1 $\boldsymbol{k}_{+1}$ represents the case when $\beta = 1$ and $\boldsymbol{k}_{-1}$ when $\beta = -1$.

Now TDOA can be written as a function of β using (3.24) and (3.23).

$$\tau_{ij}(\beta) = \beta c^{-1} ||\boldsymbol{m}_j - \boldsymbol{m}_i|| + \Delta_{ij}, \tag{3.25}$$

If $\Delta_{ij} = 0$, the TDOA magnitude (either $\beta = -1$ or $\beta = 1$) is the propagation time between microphones i and j when an acoustic event occurs at an endfire direction.

Let us denote $\tau_{ij}(-1) = \tau_{ij}^{\min}$ and $\tau_{ij}(1) = \tau_{ij}^{\max}$. Now the distance between the microphones i and j is

$$d_{ij} = \frac{c}{2}(\tau_{ij}^{\max} - \tau_{ij}^{\min}) \tag{3.26}$$

This can be shown by assigning (3.25) to (3.26). Note that the pairwise time offsets Δ_{ij} are cancelled out in the distance estimate d_{ij}. To estimate d_{ij} the maximum and minimum TDOA values can be measured when sources are at endfire directions, not between the microphones i and j.

Clearly, at least one acoustic event has to be detected followed by successful TDOA estimation to obtain the distance between a microphone pair i, j. The TDOA is estimated using GCC (see 2.2.1 [KC76]). Next, a set of nonlinear operations is performed for the correlation function $r_{ij}(t)$.

Time Delay Estimation from GCC function

The time delay in samples is

$$\hat{\tau}_{ij} = \underset{t}{\operatorname{argmax}}\, r_{ij}(t). \tag{3.27}$$

The estimation of the GCC is presented in Section 2.2.

Sequential TDOA gating

TDOA estimates $\hat{\tau}_{ij}$ are processed using *sequential gating*. The idea of this is that that acoustic events are produced by natural sounds occurring in the environment. Thus, TDOA estimates change from one time frame to the next smoothly. Otherwise TDOA is resulting from background noise, some other source, or from reverberation. The following filtering scheme is used.

$$\bar{\boldsymbol{\tau}}_{ij} = \{\hat{\tau}_{ij} | \lambda_G > |\hat{\tau}_{ij}(t) - \hat{\tau}_{ij}(t-n)|, \forall t\}, \tag{3.28}$$

where n defines the amount of sequential frames observed. λ_G is a threshold in samples that TDOA estimates n samples apart cannot exceed. In [PMH12], the first or second order filter is used, i.e. $n \in [1, 2]$.

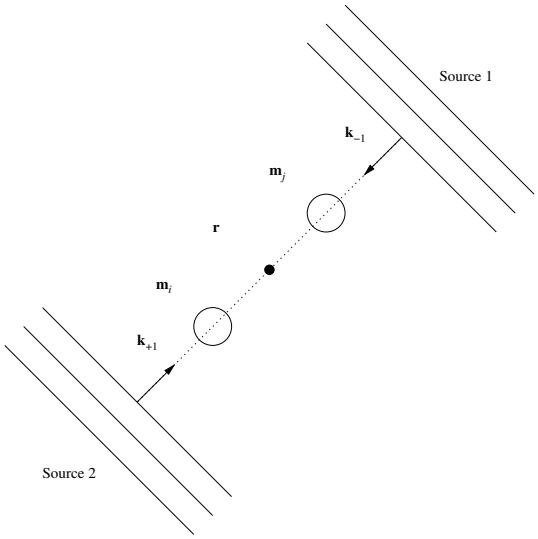


Figure 3.1. Two sound sources are located at endfire positions. The measured TDOA can be used to estimate the distance between microphone pair i and j.

TDOA Histogram Filtering

Next, a histogram of filtered $\bar{\boldsymbol{\tau}}_{ij}$ is formed. Let us denote the number of entries in the kth bin as n_{ij}^k. That is, the kth bin contains those TDOA values $\bar{\boldsymbol{\tau}}_{ij}$ that are closest to the value of k, where $k \in [-K, K]$ and K is the upper limit of the histogram. Histogram thresholding is performed to preserve only bins with the entries exceeding a certain threshold as follows.

$$\widetilde{\boldsymbol{\tau}}_{ij} = \{\bar{\tau}_{ij}^k | n_{ij}^k > \alpha \cdot \max(n_{ij}^{-K}, \ldots, n_{ij}^K), \forall k\}, \tag{3.29}$$

where $\alpha \in [0, 1]$ is the threshold. Setting $\alpha = 0$ results in preserving all TDOA values whereas values close to unity preserve only histogram bins with most occurrences.

Finally, the maximum and the minimum TDOA estimators are

$$\hat{\tau}_{ij}^{\max} = \max(\widetilde{\boldsymbol{\tau}}_{ij}), \hat{\tau}_{ij}^{\min} = \min(\widetilde{\boldsymbol{\tau}}_{ij}). \tag{3.30}$$

Once all pairwise distances are estimated, MDS (see Section 2.4 for MDS) is used to map the pairwise distance matrix into the relative microphone positions in 3D Cartesian coordinates.

3.6.1 Performance of [PMH12]

The system [PMH12] is tested with simulations and real data and the goal is to estimate microphone positions from environmental audio. In the simulated data scenario two sources and six linear microphone arrays are placed in a reverberant space. The image method [AB79] is used to generate the data corresponding to a $2.4 \times 5.9 \times 2.8$ m space with reverberation time T_{60} from 0.0 s to 2.0 s. The speech sources are at 1.1 m distance from the array. Different amounts of noise are added resulting in signal-to-noise ratios (SNR) in the range from 0 dB to 30 dB. Two major observations can be made from this simulation: the positioning error increases significantly when the SNR is lower than 15 dB, and the method seems to be robust against reverberation when SNR is above 15 dB.

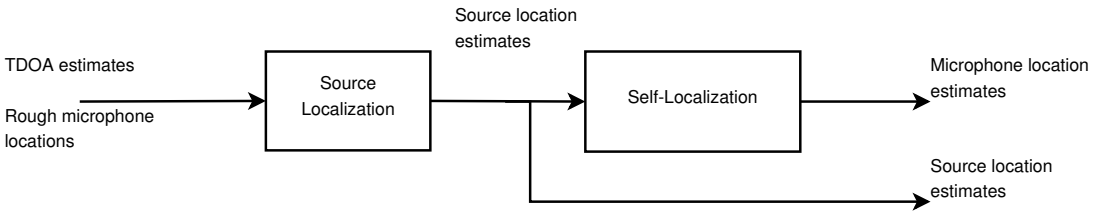


Figure 3.2. Self-localization via source localization. Rough microphone positions are used as initial guesses. The microphone array is used for source localization. Finally, source locations and rough microphone positions along with the TDOA measurements are used to improve microphone position estimates via optimization.

Another simulation scenario is tested in which the sources are not at endfire directions, but on a three-meter-radius circle around the array. The sources are rotated and each configuration is evaluated. The number of sources used ranges from two to eleven and the amount of reverberation time T_{60} from 0 to 1.6 s. The error decreases approximately logarithmically as a function of number of sources when $T_{60} \leq 0.4$ s. The error is high with few sources even in low reverberation. When $T_{60} \geq 0.8$ s, the error is high until the number of sources increases to seven and then slowly decreases with the increasing number of sources. The increase of the number of sources results in more TDOA estimates from endfire directions.

The real data measurements were made in a 6 m $\times$ 4 m office space with a varying ceiling height of 2.9 m to 3.5 m and with $T_{60} = 440$ ms. Ten Nokia N900 devices were placed on a table to capture the audio of three persons talking in turns. The speakers move behind each device (corresponding to endfire directions) and spoke sentences sequentially until speech has been emitted behind every handset. The device positioning error is in the beginning of recording approximately 30 cm and decreases as more data is received to 6.9 cm averaged over devices.

3.6.2 Conclusions on Passive Self-localization of Microphones Using Ambient Sounds

The method [PMH12] is categorized as a node based self-localization (See Table 2.1) and therefore the method does not require synchronous microphone signals. The pairwise distances between microphones are estimated from speech and therefore the method is unobtrusive and provides an appealing underlying technique for many higher-level applications that benefit from microphone locations. The method obtains less than 10 cm positioning accuracy in relatively high-reverberant spaces if the SNR is sufficiently high. The amount of data from the endfire directions is crucial and therefore longer recording period lowers the positioning error.

3.7 Self-localization via Source Localization [P2]

Given initial estimates of relative microphone position and TDOA measurements, the self-localization problem (3.20) can be approached by solving source positions. The idea of this approach is illustrated in Figure 3.2. The source localization can be conducted using trilateration [MH95]. TDOA input along with rough microphone position estimates can be used as initial points of an optimization problem (3.4). Such a system is presented in [P2] and it is designed for an ad hoc meeting room

scenario. The goal is to estimate relative positions of each device containing a single microphone. Microphones and sound sources are assumed non-moving.

The result of the minimization of (3.4) is subject to arbitrary rotation, reflection, translation, and arbitrary common time offset. This implies $N_u = 3(N + K) + N - 7$ number of unknown variables (3D space), which cannot exceed the $N_m = S(N - 1)$ independent measurements [OKIS09]. The degrees of freedom (DOF) is defined here as $DOF = N_m - N_u$, where $DOF \geq 0$ is forced by fixing a coordinate system $\boldsymbol{m}_1 = [0, 0, 0]^T$ (translation), $\boldsymbol{m}_2 = [m_{2x}, 0, 0]^T$ and $\boldsymbol{m}_3 = [m_{3x}, m_{3y}, 0]^T$. Furthermore, the temporal offset of the first microphone is set to zero i.e. $\delta_1 = 0$.

3.7.1 Temporal Offset Estimation

[P2] uses asynchronous signals and therefore temporal offsets δ_i of (3.4) are needed in contrast to [PMH11], where pairwise offsets are canceled out due to co-location of microphones and sources. The offset estimation can be made using the method presented in [PHM13]. The pairwise offsets δ_{ij} are obtained as follows:

$$\delta_{ij} = \frac{1}{2}(\tau_{ij}^{\max} + \tau_{ij}^{\min}), \tag{3.31}$$

where $\tau_{ij}^{\min}$ and $\tau_{ij}^{\max}$ are the minimum and the maximum TDOA values for microphone pair (i, j). The observation of the minimum and the maximum requires that there are measurements from sources at endfire positions as in Figure 3.1. Individual offsets relative to an arbitrary reference node can be recovered using the Maximum Likelihood (ML) estimator presented in [PHM13]. In practice, microphone 1 is set as reference with zero offset i.e. $\delta_1 = 0$, the remaining $N - 1$ individual microphone offsets $\delta_2, \ldots, \delta_N$ are recoverable.

3.7.2 Source Localization

Using rough microphone positions obtained using the procedure of Section 3.6 and estimated temporal offsets, source localization can be performed using the time-aligned signals. Each source can be localized via optimization of (3.4) for each source separately. The optimization problem (3.4) is easier when good initial guesses for microphone positions and temporal offsets are available. The number of unknowns is three since $\boldsymbol{s}_k = [s_{kx}, s_{ky}, s_{kz}]^T$ and only one source is assumed active at a time. Furthermore, closed-form source localization methods (see Section 4.5) can be used, but they contain linearizations of non-linear equations that may result in too large inaccuracies in source position estimates.

3.7.3 Data Association

The data association provides information from which source a certain TDOA measurement is originated. The meeting scenario is assumed in [P2]; the attendants are seated at a table with a mobile device in front of them.

The idea of data association is that TDOA is different for sources at different positions. As different persons are speaking, changes in TDOA measurements imply the change of the source. Due to the large number of microphones ($N = 10$), there are $N(N - 1)/2 = 45$ TDOA measurements per analysis frame. Therefore detection of speaker change is difficult. Using dimensionality reduction techniques such as PCA, the detection of changes can be done provided that sources are non-moving.

3.7.4 Iterative Optimization

The refined microphone positions are obtained by initializing (3.4) with the rough microphone positions obtained using the approach in Section 3.6, all estimated source positions, and ML temporal offset estimates. Furthermore, the number of variables is reduced by eliminating arbitrary rotation, reflection, translation, and arbitrary common time offset. Optimization methods such as Levenberg-Marquardt [Lev44] [Mar63] are used.

3.7.5 Performance of [P2]

The system is evaluated with real data recordings made in two meeting rooms and the scenarios imitate a meeting of four participants. The target is to self-localize mobile devices (Nokia N900) that were placed on the table in front of each person. The content of each recording is non-overlapping speech; each participant utters a sentence sequentially. The recorded signals are asynchronous, and the positions of devices and persons are unknown. This kind of setup corresponds to an ad hoc deployment of an acoustic sensors network. The ground truth coordinates are obtained using a tape measure.

The performance of [P2] is measured using the root mean square (RMS) error in device position and it achieves an RMS error from 7.2 cm to 14.7 centimeters averaged over 10 devices. Four sources were present in each recording.

3.7.6 Conclusions on Self-localization via Source Localization

The system presented in [P2] is a blind self-localization method (see Table 2.1). However, it assumes that the number of sources is known and that the sources and the microphones are non-moving. The assumption on knowing the source count can be relaxed by temporal analysis of the used data association scheme, but the assumption on the non-movement is strict. A dynamic scenario, i.e. sources are moving while self-localizing, requires more advance data association using Multiple Target Tracking (MTT) techniques [Pul05].

Another blind self-localization method is presented in [OKIS09], which assumes moving sources and considers that each TDOA estimate is originated from a different source, and therefore the number of sources inherently increases with signal length. The performance of [OKIS09] in terms of comparable RMS values is not reported, but visual comparison made from the illustrations in [OKIS09] indicates similar accuracy to [P2].

3.8 Coherence Based Self-localization

Coherence based self-localization is designed for environments where diffuse sound field can occur. Diffuse sound field consist of large number of wave fronts travelling in all directions with equal probability [PK15]. These environments are highly reverberant ($T_{60} > 400$ms) or for some other reason have sounds arriving from spatially different directions. The fundamental idea of coherence based self-localization methods is to estimate intra-microphone distances by measured noise coherence with its theoretical model [MLH08]. Relative coordinates are obtained from pairwise distance estimates using MDS. Thus, the key difference to other self-localization methods of Table 2.1 is in distance estimation. The main advantages of coherence

based self-localization include not having to rely on initialization, calibration signals, and there is no need for data association.

3.8.1 Coherence of Diffuse Sound Field

The complex coherence function is [CKN73]

$$\Gamma_{ij}(f) = \frac{\phi_{ij}(f)}{\sqrt{\phi_{ii}(f)\phi_{jj}(f)}}, \tag{3.32}$$

where ϕ_{ij} it the cross spectral density at frequency f between signals i and j. $\phi_{ii}(f)$ and $\phi_{ii}(f)$ are auto spectra of the signals i and j, respectively.

It can be shown [CWB+55] that in diffuse noise sound field, the coherence between two locations with distance d_{ij} is

$$\Gamma_{ij}^{\text{diffuse}}(f) = \text{sinc}\left(\frac{2\pi f d_{ij}}{c}\right), \tag{3.33}$$

where $\text{sinc}\,(a) \triangleq \frac{\sin a}{a}$ and c is speed of sound.

3.8.2 Estimation of Pairwise Distances in Diffuse Noise Fields

Assuming fixed speed of sound, (3.32) states that the coherence between two signals at frequency f depends only on the distance d_{ij} for the nodes i and j. Therefore, coherence based self-localization can be formulated as an optimization problem in which the measured coherence is compared to the theoretical coherence. The distance estimator d_{ij} is obtained as follows [MLH08]

$$\hat{d}_{ij} = \underset{d_{ij}}{\text{argmin}} \sum_{f=0}^{f_s/2} \left| \Re\{\Gamma_{ij}\} - \text{sinc}\left(\frac{2\pi f d_{ij}}{c}\right)\right|^2, \tag{3.34}$$

where f_s is the sampling frequency. $\Re\{\cdot\}$ is the real part operator. Traditional optimization algorithms such as Levenberg-Marquardt [Lev44][Mar63] can be used to obtain the pairwise distance estimates for each node pair i, j. Once distance estimates $\hat{d}_{ij}$ are obtained for all node pairs, MDS is applied to obtain relative coordinates.

3.8.3 Temporal Filtering of Distance Estimates

Inaccuracies between the diffuse noise field model (3.33) and the measurements result in deviations in the distances estimates $\hat{d}_{ij}$ and the ground truth. Therefore, tracking or clutter detection scheme has to be introduced to prevent outliers from deteriorating the distances estimates. One way to address the outliers is to calculate the current distance estimate over multiple frames. In [MLH08] input frames are classified to two classes A and B. The frames that belong to class A fit well with the diffuse noise model (3.33) whereas class B frames do not. In case of non-moving sensors, the distance estimates of class A representatives are relatively constant, whereas sequential distance estimates of class B members vary significantly. An unsupervised learning scheme can be used to perform the classification and in [MLH08] k-means clustering is used to cluster an $L \times 2$ matrix. Each line of the matrix contains a distance estimate on its first column and the residual error of (3.34) on its second column.

An alternative scheme for outlier mitigation is to average coherence function over multiple frames. In [TAGB14] average over 100 sequential frames is taken and then the average is compared to the theoretical using (3.34).

3.8.4 Performance of Coherence Based Self-localization

The coherence based self-localization system in [MLH08] is evaluated with real data collected using a 5-microphone linear antenna with 5 cm intra-microphone distance and using a 8-microphone circular array with 10 cm radius. The recording environments are rooms with the reverberation times $T_{60} = 400$ ms and $T_{60} = 700$ ms and the system achieves distance error of 0.43 cm and 1.50 cm on average, respectively. After distance estimation, MDS is applied to obtain relative microphone coordinates. After MDS, rotation and translation are estimated to allow comparison to ground truth coordinates. The Euclidean error in microphone coordinates is from 0.48 cm to 1.72 cm on average and again the lower error is obtained in the $T_{60} = 700$ ms environment. The difference in performance between environments may result from diffuse noise model being more accurate in rooms with higher reverberation time.

In [HPF+09] a coherence based self-localization method is presented. Rather than determining positions of individual microphones directly, microphones form multiple sensors arrays and utilize source positioning in the process. The difference to [MLH08] is that [HPF+09] is designed for scenarios where multiple arrays are several meters apart from each other. In [HPF+09] real data is collected in a conference room with $T_{60} \approx 500$ ms with two 8-microphone arrays and with four 4-microphone T-shaped arrays. Furthermore, a linear array is used to test the diffuse noise model (3.33). The error between the model and the ground truth ranges from 1 mm for 15 cm intra-microphone distance to 12.9 cm for 87.5 cm intra-microphone distance. Intra-microphone distance of approximately 60 cm can be regarded as a sort of threshold for the method to be operational, since the error is 3.7 cm. The distance estimation error for T-shaped array with 14.1 cm inter-microphone is approximately 1 cm. For the circular array with 10 cm radius, the error ranges from 1 cm to 4 cm. A moving source is used to estimate the inter-array distance. The distance estimate errors range from approximately 10 cm to 27 cm depending on input signal type. The lowest error is obtained using white noise and the highest with speech.

3.8.5 Conclusions on Coherence Based Self-localization

The distinctive feature of the coherence based self-localization is the assumption of diffuse sound field. Even if the conditions are approximating diffuse sound field, the methods achieve less than 10 cm errors in estimating microphone positions. The drawback of coherence based self-localization methods is that diffuse sound field model (3.33) is accurate at relatively small inter-microphone distances (e.g. less than 1 m for $T_{60} = 500$ ms environments [HPF+09]). A major advantage over TDOA based self-localization methods is avoiding the problems related to optimization such as (3.20) and (3.4), and there is no need for data association.

3.9 Self-localization of Moving Nodes [P1]

Many self-localization methods assume that the targets of self-localization are at the same position during, which is in general a reasonable expectation. Some methods

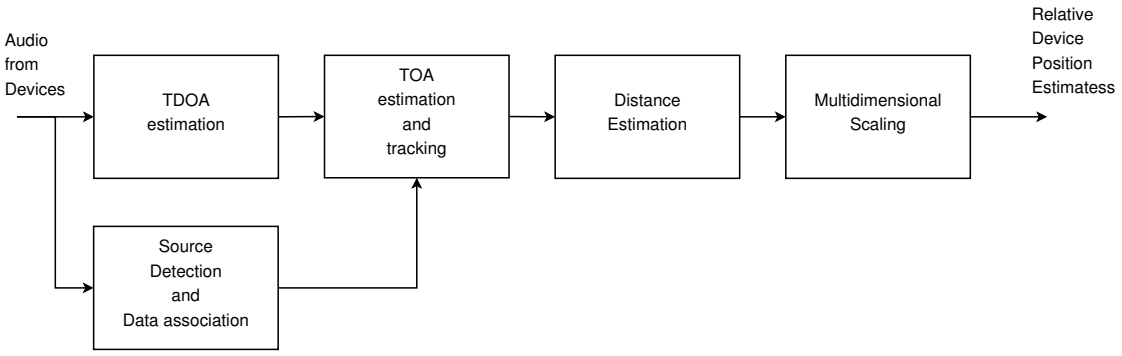


Figure 3.3. An overview of acoustic self-localization system for dynamic scenario.

such as [OKIS09] expect a source to move, but the microphones are stationary. With increasing wearable and embedded sensor technology, the assumption on stationarity may have to be relaxed to enable continuous self-localization in dynamic scenario. A method capable of self-localization of moving unsynchronized nodes is presented in [P1], which estimates the node positions from speech emitted by the nodes. The system takes advantage of the concept of a wearable device and therefore nodes of the network contain a sound source and microphone. The system diagram is presented in Figure 3.3. The node model is the same as in other node based self-localization methods such as the one presented in Section 3.6.

3.9.1 Spatial information and Tracking

The spatial information is extracted using TDOA estimation. Since the system is designed for nodes in motion while not emitting, a tracking scheme is introduced. In addition to TDOA, node positions are characterized by TOA. In a network of N nodes, there are $N(N-1)/2$ TDOA estimates per analysis frame and in terms of TOA N estimates. Therefore, tracking of TOA is more appealing than tracking of TDOA.

TDOA can be formulated as the matrix product of an observation matrix H and TOA vector Δ [PHM13]:

$$\boldsymbol{\tau} = H\Delta \tag{3.35}$$

where $\boldsymbol{\tau}$ is the TDOA vector, $\boldsymbol{\Delta} = [\Delta_1, \ldots, \Delta_N]^T$ is the TOA vector, and the observation matrix $H = [\boldsymbol{e}_1 - \boldsymbol{e}_2, \boldsymbol{e}_1 - \boldsymbol{e}_3, \ldots, \boldsymbol{e}_1 - \boldsymbol{e}_N, \boldsymbol{e}_2 - \boldsymbol{e}_3, \ldots, \boldsymbol{e}_2 - \boldsymbol{e}_N, \ldots, \boldsymbol{e}_{N-1} - \boldsymbol{e}_N]^T$. $\boldsymbol{e}_i = [\delta_{1i}, \ldots, \delta_{Ni}]^T$, where δ_{ij} are Kronecker's delta function. [P1] uses (3.35) as the measurement model in the tracking.

3.9.2 Data Association and Tracking

Data Association

Since there are multiple nodes that are tracked, a data association scheme is needed, that is, assigning each measurement to a correct source. In [P1], the data association and tracking method exploits the geometry of the node network: each node contains a microphone and a source. Therefore, it is highly likely that the loudest speech energy that exceeds background noise energy level can be detected in the nearest microphone where the source is active at current time. The data association assumes that there is only one source active at a time and the content of data contains mainly

signal originated from mostly speech source. Therefore, Sound Pressure Level (SPL) based threshold for source detection and identification is used. More sophisticated data association [Pul05] and speech detection methods [TR06] could be applied if SPL thresholding is too error prone.

Tracking

In order to maintain the TOA state of the nodes, [P1] uses the Kalman Filter (KF) [Kal60] presented in Section 2.6. The use of Kalman filtering in TOA tracking is inspired by [PT13]. Each node has its own Kalman filter and the correct filter is selected using data association information. The Wiener motion model is used as in [PT13] [SVL07]. $\boldsymbol{q} \sim \mathcal{N}(\mathbf{0}, Q)$ and $\boldsymbol{r} \sim \mathcal{N}(\mathbf{0}, R)$ are system noise and measurement noise with variances Q and R, respectively, and $\boldsymbol{y}_t$ are TDOA measurements in time t.

For instance, for the case $N = 4$, the state of the KF is written

$$\boldsymbol{x} = [\Delta_2, \Delta_3, \Delta_4, \dot{\Delta}_2, \Delta_3, \dot{\Delta}_4]^T, \tag{3.36}$$

where $\dot{\Delta}$ denotes velocity. Δ_1 and $\dot{\Delta}_1$ are set to zero and omitted due to insufficient degrees of freedom.

The state transition matrix is written

$$A = \begin{bmatrix} 1 & 0 & 0 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 & \Delta t & 0 \\ 0 & 0 & 1 & 0 & 0 & \Delta t \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}_{6\times 6} \tag{3.37}$$

and the observation matrix (3.35) is

$$H_0 = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \end{bmatrix}_{6\times 6} \tag{3.38}$$

Pairwise Distance and Relative Coordinate Estimation using MDS: The TOA estimates are converted to TDOA

The TOA estimates are converted to TDOA:

$$\tau_{ij} = \Delta_i - \Delta_j. \tag{3.39}$$

Using the a priori knowledge of sound speed $c \approx 344\ \frac{m}{s}$ in indoors, TDOA information can be transformed into pairwise distance matrix D, which is written as

$$D = \begin{bmatrix} 0 & d_{12} & \dots & d_{1N} \\ 0 & 0 & \dots & d_{2N} \\ \vdots & \vdots & \ddots & d_{N-1,N} \\ 0 & 0 & 0 & 0 \end{bmatrix}_{N\times N}, \tag{3.40}$$

where $d_{ij} = c \cdot \tau_{ij}$.

The distance matrix is converted to relative coordinates using MDS [BG97, Chapter 7.9] (see Section 2.4).

3.9.3 Performance of [P1]

The self-localization system of moving nodes is evaluated using real data measurements in a 4.5 m $\times$ 3.9 m $\times$ 2.6 m room, which can be characterized as low reverberant with $T_{60} \approx 260$ ms. The target is to self-localize nodes continuously while they are moving around the room from the speech emitted by the nodes. Nokia N900s are attached to user's chest to simulate a wearable device of an ad hoc network. The ground truth coordinates used for evaluation are obtained using an image based node positioning system. The structure with reference points is illustrated in Figure 3.5.

The measurement scenario consists of a static phase and dynamic phase. During the static phase a user behind each node, while being stationary, speaks a sentence sequentially. The static phase is used to obtain initial estimates for node positions using [PMH11] (see Section 3.6). During the dynamic phase two nodes are moving and emitting speech sequentially. All nodes are continuously being self-localized.

The performance of [P1] is measured using RMS error of node positions. The self-localization RMS error during non-moving phase is below 10 cm. During the dynamic phase, RMS error varies between 25 cm and 27.5 cm averaged over four nodes. The mean RMS over the whole recording (including non-moving and dynamic phases) is less than 9.2 cm on average. An example of system output and the video based reference method is depicted in Figure 3.4.

3.9.4 Conclusions on Self-localization of Moving Nodes

The self-localization system presented in [P1] is the first self-localization method using environmental sounds that allows the targets of self-localization to move. The system can be classified as a node based self-localization system (see Table 2.1) and it has its origins in [PMH12]. Therefore, the fundamental assumption is that a node consists of a microphone and a sound source, which in turn removes the requirement on knowing the source count. The accuracy of the system in positioning is appealing and therefore [P1] introduces a viable framework for an unobtrusive system for many applications such as indoor positioning.

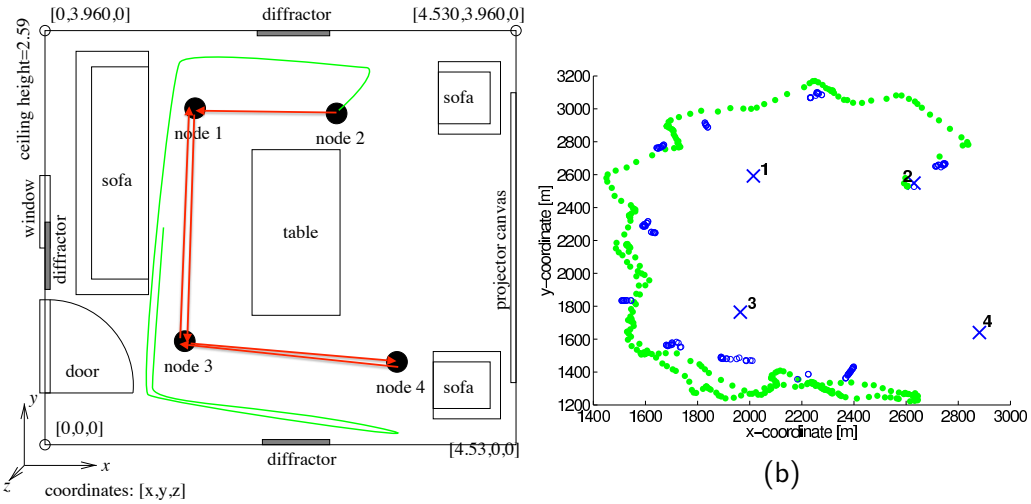


Figure 3.4. Panel (a) presents the planned route (red arrows) and track (green path) for node 2, which is node 1→node 3→node 4→node 3→node 1. Panel (b) presents the reference coordinates (green dots) and the estimated coordinates (blue circles) from audio data. The position estimates are shown when the node is emitting. The crosses × denote the initial position of the nodes which are labeled '1', '2', '3', and '4'.

Figure 3.5. The frame with calibration marks. The calibration points are highlighted by the red arrows and are visible to four cameras that used to obtain the reference 3D coordinates (see [P1] for details). The two points highlighted by two green arrows outside the calibration pole are also used to increase the accuracy of the transformation.

3.10 Summary

This chapter presented the work conducted in the field of acoustic self-localization during this dissertation and the most related methods.

[P3] introduced the use of Tikhonov Regularization and Truncated Singular Value Composisition (TSVD) in acoustic self-localization to increase the robustness against measurement errors. The results of [P3] show that in presence of the errors in time information and calibration source location, it is beneficial to use either Tikhonov Regularization or TSVD to increase robustness in the used microphone/source setup. [P3] assumes known calibration source positions and synchronized audio streams, which are not often available especially in case of ad hoc networks.

[P2] introduced a self-localization method especially for ad hoc device scenario that uses source localization as an intermediate step to localize the microphones. It performs self-localization using the rough microphone positions and temporal offsets obtained using the method presented in [PMH12] and in [PHM13], respectively (both methods are presented briefly). Then, it estimates the source positions using the rough microphone positions as a microphone array and aligns the microphone signals using the estimated temporal offsets. Finally, the rough microphone positions, the estimated source positions, and the temporal offsets are used as initial guesses for the general self-localization problem (3.4), which is then optimized to obtain the finer microphone position estimates. The system of [P2] assumes only knowledge about the number of sources and that sources and microphones are non-moving. A source enumeration subsystem can be easily added to [P2] to estimate the number of sources. Otherwise, the system very flexible in terms of source and microphone configurations.

A distinct self-localization method [MLH08] is presented, since it is suitable for ad hoc device scenario. Like [P2], [MLH08] uses environmental audio for self-localization of the microphones. [MLH08] assumes diffuse sound field and estimates intra-microphone distances by comparing estimated coherence to its theoretical model. One of the advantages of [MLH08] is that does not need the knowledge about the number of sources and a data association method.

[P1] introduced a self-localization method that is intended for wearable and embedded device scenario. It is assumed that the target of self-localization, called a node, contains a microphone and a source. The system self-localizes the nodes even if they are moving. [P1] estimates the node positions from their pairwise distance estimates, which, in turn, are obtained from the time difference of arrival estimates. Moving node scenario is taken into account by tracking of the spatial information. At the time of writing this thesis, [P1] is the first system that enables continuous self-localization of the moving node setup.

Chapter 4

Application: Sound Source Localization

The sound source localization is one of the applications that can benefit from self-localization and is required by traditional beamforming and related array signal processing methods. This chapter presents two Direction of Arrival (DOA) based approaches [P4] and [P5] to source localization. Furthermore, Time Difference of Arrival (TDOA) based methods are presented for completeness.

4.1 DOA Based Sound Source Localization

DOA based source localization methods use multiple microphone arrays, which estimate direction of arrival and combine their measurements resulting in source location estimate. Such systems for two different scenarios are presented in [P4] and in [P5]. Other DOA based localization systems include [Dom87],[KLM01], and [HN03].

4.2 DOA Based Localization Problem

The fundamental idea is to solve the position where the lines between each microphone array and the target cross. The direction of the line is obtained as result of DOA estimation in each array. For simplicity, let us consider a single source scenario. Let $\mathbf{p}_i, i = 1, \ldots, N$ denote the positions of N spatially separate microphone arrays. The microphone array positions are presented in the 3D Cartesian coordinates, i.e. $\mathbf{p}_i = [p_i^x, p_i^y, p_i^z]^T$. The microphone array-to-source position arrays are 3D vectors $\mathbf{r_i} = [r_i^x, r_i^y, r_i^z]^T$, and the source position is $\mathbf{r}_s = [r_s^x, r_s^y, r_s^z]^T$. The 3D DOA vectors $\mathbf{k_i}, i = 1, \ldots, N$ can be estimated using beamforming techniques [MGW07][VVB88], TDOA based closed-form estimators [YHKA96][YHKA99], Time-Delay Estimation (TDE) based methods [JGN02], and using acoustic vector sensor arrays [HN03]. If an array produces azimuth and elevation angle estimates, they can be transformed into 3D DOA vector $\mathbf{k_i}$ using the relations between spherical coordinates and the Cartesian coordinates. Furthermore, $||\mathbf{k_i}|| = 1, \forall i$ and the microphone array orientations to a common origin is known. Figure 4.1 presents a two-sensor-station localization network and illustrates the process.

4.3 DOA Based Source Localization for Long Inter-array Distances [P4]

In a microphone array network based localization the reception time of acoustic events is usually different for each microphone array. The difference has to be taken

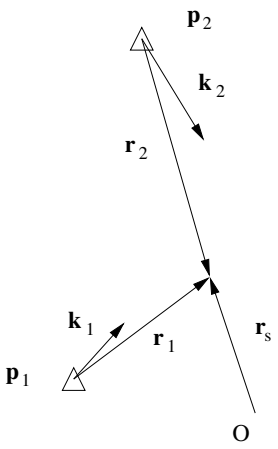


Figure 4.1. A Two-array DOA Based Localization. The arrays are at locations $\mathbf{p}_1$ and $\mathbf{p}_2$, the source position is denoted as $\mathbf{r}_s$, which measure the signal and compute the DOA vectors $\mathbf{k_1}$ and $\mathbf{k_2}$. The array-to-source vectors $\mathbf{r_1}$ and $\mathbf{r_2}$ point from each sensor station to the source location $\mathbf{r}_s$. O denotes the origin of the coordinate system.

into account when combining the information of spatially separate microphone arrays especially when the distance between the arrays is large. [P4] estimates the propagation times to each array, and combines appropriate DOA estimates for source localization.

Other such systems are presented in [BA00], where the first iteration produces an initial location estimate, which is used to estimate propagation delays, then DOA is re-estimated, and finally the location is re-estimated. In [Dom87] correction of location estimates is done directly. The approach in [P4] is originally presented in [PPKV04].

4.3.1 Propagation Time

The propagation time is calculated using

$$\Delta t_i = \frac{||\mathbf{r_i}||}{c} = \frac{||\mathbf{r}_s - \mathbf{p}_i||}{c}, \tag{4.1}$$

where $\mathbf{r_i}$, $i = 1, \ldots, N$ are the candidate array-to-source vectors, and c is the speed of sound. The Δt_i are used to find the DOA estimates in each array's history.

4.3.2 Localization Algorithm

For sensor station i, the estimated $\mathbf{k_i}$ are compared to a vector pointing from each array towards a candidate source location. The comparison outputs a value, which indicates deviation between the estimated direction of arrival and the hypothetical DOA. In Figure 4.1, DOA vectors $\mathbf{k_1}$ and $\mathbf{k_2}$ slightly deviate from the ground truth DOA, which is parallel to vectors $\mathbf{r_1}$ and $\mathbf{r_2}$.

The likelihood of a source for each hypothetical point in space is obtained via angle deviation between a measured DOA vector $\mathbf{k_i}$ and a array-to-source vector $\mathbf{r_i}$. The dot product between two vectors is suitable for measuring the angle deviation and the likelihood field is obtained as

$$l(\mathbf{r}) = \sum_{i=1}^{N} \mathbf{r_i}^{\circ} \cdot \mathbf{k_i}(t + \Delta t_i) = \sum_{i=1}^{N} e_i, \tag{4.2}$$

where $\mathbf{r_i}^\circ$ is unit vector of $\mathbf{r_i}$, $\mathbf{k}_i(t+\Delta t_i)$ is the DOA at time $t+\Delta t_i$, and e_i measures the angle between the two vectors and it varies in the interval of $[-1, 1]$. (4.2) has its maximum when DOA from each array indicates the same position. Therefore, the source location estimate $\hat{\mathbf{r}}_s$ at time t is

$$\hat{\mathbf{r}}_s = \underset{\mathbf{r}}{\operatorname{argmax}}\, l(\mathbf{r}). \tag{4.3}$$

The exhaustive search is used to obtain $\hat{\mathbf{r}}_s$ i.e. all potential source locations are evaluated by using (4.2).

4.3.3 Performance of [P4]

The system is evaluated using real data, which is collected using two four-microphone arrays. The distance between the arrays is 75.2 m and the arrays are approximately on the same height on open ground. A loudspeaker was used as sound source and it was located at two different positions. Pink noise was used as a stimulus to obtain robust DOA information and the DOA is calculated using a system similar to [YHKA96].

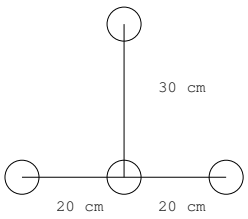
[P4] operates on 350 m $\times$ 350 m area, which is divided into a grid with cell size 4 m $\times$ 4 m. The likelihood of source location is calculated at each cell. DOA is fed unfiltered to the system to test the robustness. The absolute accuracy of [P4] is not evaluated, but visual evaluation of likelihood fields indicates robust performance.

4.3.4 Conclusions on DOA Based Source Localization for Long Inter-array Distances

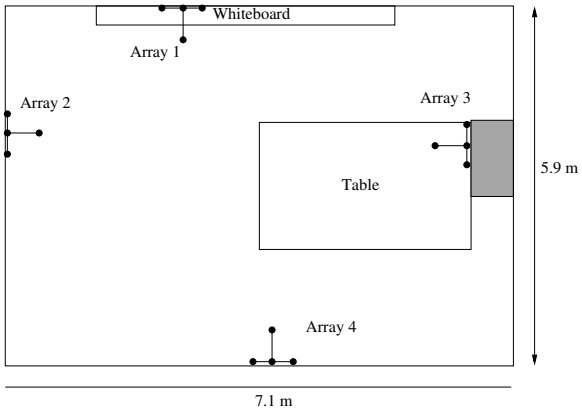
[P4] presents a DOA based localization system that takes into account the relatively slow speed of sound, which causes the observation time differences at spatially separate microphone arrays. The phenomenon is emphasized as the distance between the source and the array is tens of meters. [P4] used a grid based search method that can be an issue in some applications, especially with dense grid and large coverage area. The performance showed itself to be robust even with noisy DOA.

4.4 A Speaker Localization System for Lecture Room Environments [P5]

A DOA based speaker localization system designed for indoors such as meeting rooms is presented in [P5] and is evaluated in the Rich Transcription Evaluation Series [NIS]. The system uses four-microphone arrays mounted on the walls of a meeting room, which is illustrated in Figure 4.2b. The scenario is different from [P4] since the distances between the arrays are short. Closed spaces result in reverberation that pose challenges in DOA estimation. Therefore some method for DOA filtering or validation is needed. In [P5] median filtering is used. More advanced techniques such as confidence metric [Pir05], DOA exclusion [PP07], and direct TDOA filtering based on inspection of the GCC function [BK02] may be needed for low signal-to-noise ratio conditions. More robust, but computationally more intensive systems based on particle filtering and direct use of the GCC function are presented in [PKPP07][PP07].



(a) The microphone geometry of [P5]. The stations are mounted at a 2.3 m height, one on each wall of the meeting room.



(b) The Rich Transcription Evaluation Series [NIS] meeting room. The room is equipped with four microphone arrays with the geometry of Figure 4.2a.

4.4.1 Localization Method

The approach in [P5] is similar to other DOA based localization methods [KLM01] [HN03][PPKV04]. A weighted least squares solution [HN03] is used. The method is presented briefly using the same notation as in Section 4.3.

$$\hat{\mathbf{r}}_s = \underset{\mathbf{r}}{\operatorname{argmin}} \sum_{i}^{N} ||\mathbf{p}_i + \mathbf{k_i}^T(\mathbf{r} - \mathbf{k_i}) - \mathbf{r}||^2 \cdot w_i \tag{4.4}$$

where w_i is a weight corresponding to the confidence of each DOA estimate $\mathbf{k_i}$. It is shown in [HN03] that (4.4) has a closed-form solution:

$$\hat{\mathbf{r}}_s = \left[\left(\sum_{i=1}^{N} w_i \right) I - \hat{U} W \hat{U}^T \right]^{-1} A\boldsymbol{w}, \tag{4.5}$$

where $\boldsymbol{w} = [w_1, \ldots, w_N]^T$, $W = \mathrm{diag}\lceil w_1, \ldots, w_N \rfloor$, $\hat{U} = [\mathbf{k_1}, \ldots, \mathbf{k_N}]$,

$$A = \left[\left(I - \mathbf{k_1}\mathbf{k_1}^T \right) \mathbf{p}_1, \ldots, \left(I - \mathbf{k_N}\mathbf{k_N}^T \right) \mathbf{p}_N \right],$$

and I is an $N \times N$ identity matrix.

4.4.2 Performance of [P5]

The real data measurements are conducted in a room with dimensions of 5.9 m × 7.1 m × 3.0 m. The room is characterized as an ordinary room used for meetings and studying. Including the measurement equipment, there is a table and a few seats [Sti04]. The microphone arrays are installed on the walls and the geometry is presented in Figure 4.2a.

The system is evaluated with data collected from 13 meeting sessions and the RMS error of the speaker position estimates varies approximately from 40 cm to 140 cm averaged over the whole recording, and the average error over all recordings is approximately 80 cm.

4.4.3 Conclusions on [P5]

The presented DOA based speaker localization system is designed a for meeting room scenario. However, the system can be used in other scenarios where the intra-array distances are relatively small. Otherwise, [P5] should integrate an approach such as [P4] to acknowledge source-to-array propagation times. The accuracy of [P4] in general is acceptable for many applications. The performance is most degraded by the quality of the DOA. For instance, [P5] lacks a Speech Activity Detection (SAD) subsystem. Air-conditioning and other noise sources may be the dominant sound source from time to time, which results in large errors. Including SAD, tracking schemes such as Kalman and particle filtering are likely to enhance the performance, but increase the computational burden. Therefore, it depends on the demands of the applications whether additional DOA improvement techniques should be implemented.

4.5 TDOA Based Source Localization

There are several techniques to localize a sound source from TDOA estimates. Iterative and closed-form estimates have been developed and a summary can be found e.g. in [Per09]. Closed-form techniques are attractive due to computational efficiency and they can be used as initialization for more accurate methods.

Relying on TDOA, the success and accuracy of source localization, as well as the general self-localization problem, depends directly on the quality of the TDOA data, which is affected by a reverberant environment, the signal-to-noise ratio, and the chosen time-delay estimation method. Some time-delay estimation methods take into account the target signal type and the acoustic environment (see Section 2.2 and [KC76]). Next, two closed-form TDOA based localization methods are presented. More comprehensive reviews can be found in [SL06] and in [Per09].

4.6 Sound Source Localization Using Range Differences

The range difference between microphone $k \in 1, \ldots, M$ and reference microphone $\boldsymbol{m}_0$ is

$$\Delta t_{k0} \cdot c = d_{k0} = ||\boldsymbol{r} - \boldsymbol{m}_k|| - ||\boldsymbol{r} - \boldsymbol{m}_0||, \tag{4.6}$$

where Δt_{k0} is the TDOA estimate for microphone pair $(k, 0)$, c is speed of sound, d_{k0} is the distance between microphone 0 and k, $\boldsymbol{r} = [r_x, r_y, r_z]^T$ is the source position, and $\boldsymbol{m}_k = [m_x, m_y, m_z]^T$ denotes the microphone position in Cartesian coordinates.

Rearranging and squaring (4.6), it follows that

$$||\boldsymbol{r} - \boldsymbol{m}_k||^2 = (d_k + ||\boldsymbol{r} - \boldsymbol{m}_0||)^2 \tag{4.7}$$

The expansion of the above yields

$$d_k||\boldsymbol{r} - \boldsymbol{m}_0|| + (\boldsymbol{m}_k^T - \boldsymbol{m}_0^T)\boldsymbol{r} = b_k, \tag{4.8}$$

where

$$b_k = \frac{1}{2}(||\boldsymbol{m}_k^T||^2 - ||\boldsymbol{m}_0^T||^2 - d_k^2)$$

The above is a generalization of the presentation in [SL06], where the reference microphone is at the origin, i.e., $\boldsymbol{m}_0 = [0, 0, 0]^T$. The generalized formulation is also derived in [Fri87].

4.7 Unconstrained Least Squares Method (UC)

Let us define

$$\boldsymbol{y}(\boldsymbol{r}) = \begin{bmatrix} R_0 \\ \boldsymbol{r} \end{bmatrix}, \Theta = \begin{bmatrix} d_{10} & \boldsymbol{m}_1^T \\ d_{20} & \boldsymbol{m}_2^T \\ \vdots & \vdots \\ d_{M0} & \boldsymbol{m}_M^T \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_M \end{bmatrix}, \tag{4.9}$$

where R_0 is the distance to the reference microphone. The least squares criterion is

$$J_{UC} = ||\Theta \boldsymbol{y}(r) - \boldsymbol{b}||^2 \tag{4.10}$$

and its LS solution is

$$\widetilde{\boldsymbol{y}} = (\Theta^T \Theta)^{-1} \Theta^T \boldsymbol{b} \tag{4.11}$$

[SA87a, SL06]. The source location estimate is the lower part of $\widetilde{y}$:

$$\widetilde{\boldsymbol{r}} = \begin{bmatrix} \boldsymbol{0} & I \end{bmatrix} \widetilde{\boldsymbol{y}}, \tag{4.12}$$

where $\boldsymbol{0}$ is a vector of zeros, I is an identity matrix of size $\mathbb{R}^{d \times d}$, and d is the dimension of the coordinate space. The first element of $\boldsymbol{y}$ should be R_0. However, in (4.10) the dependence is ignored (hence the term unconstrained). Thus, the first element of estimator $\widetilde{\boldsymbol{y}}$ does not have an interpretation of range.

4.7.1 Performance of Unconstrained Least Squares Method

In [SA87a] the spherical interpolation (SI) method is introduced, which is developed based on the (4.12) estimator. In [SL06], it is stated that the SI method is identical to the Unconstrained Least Squares Method and therefore the performance reported for SI is same as that of (4.12). The performance of SI is presented in two simulated setups where the source is 454.5 and 1467 distance units away (the authors of [SA87a] specify no absolute metric) from the microphone array. The measurement errors are simulated by adding white noise to distance differences. In the scenario where source is at the distance of 454.5, the root mean square error between the ground truth and the estimated coordinates is approximately 2 distance units and when source is at the distance of 1467, the error is 28.6 distance units. The standard deviation of noise was 0.1 distance units.

4.8 Extended Unconstrained Least Squares Method (UCExt)

An extension to UC method is to use more than one reference microphone, which results in more TDOA measurements. In the following formulation two reference microphones are used. The formulation is based on [GS08, PMH12].

As in the case of UC, the LS solution for UCExt can be written as

$$\widetilde{\boldsymbol{y}}_{UCExt} = (\Theta_{UCExt}^T \Theta_{UCExt})^{-1} \Theta_{UCExt}^T \boldsymbol{w}. \tag{4.13}$$

In this case the measurement matrix Θ_{UCExt} is

$$\Theta_{UCExt} = \begin{bmatrix} d_{10} & 0 & (\boldsymbol{m}_1 - \boldsymbol{m}_0)^T \\ d_{20} & 0 & (\boldsymbol{m}_2 - \boldsymbol{m}_0)^T \\ \vdots & \vdots & \vdots \\ d_{M0} & 0 & (\boldsymbol{m}_M - \boldsymbol{m}_0)^T \\ 0 & d_{1N} & (\boldsymbol{m}_1 - \boldsymbol{m}_N)^T \\ 0 & d_{2N} & (\boldsymbol{m}_2 - \boldsymbol{m}_N)^T \\ \vdots & \vdots & \vdots \\ 0 & d_{MN} & (\boldsymbol{m}_M - \boldsymbol{m}_N)^T \end{bmatrix}, \tag{4.14}$$

where d_{ij} are range differences between two microphones:

$$d_{ij} = ||\boldsymbol{r} - \boldsymbol{m}_i|| - ||\boldsymbol{r} - \boldsymbol{m}_j||$$

$\boldsymbol{w}$ is defined as follows

$$w = [w_{10}, w_{20}, \ldots, w_{M0}, w_{1N}, w_{2N}, \ldots, w_{MN}]^T, \tag{4.15}$$

where $w_{ij} = \frac{1}{2}(||\boldsymbol{m}_i||^2 - ||\boldsymbol{m}_j||^2 - d_{ij}^2)$.

The estimator $\widetilde{\boldsymbol{y}}_{UCExt}$ is written as follows.

$$\widetilde{\boldsymbol{y}}_{UCExt}(\boldsymbol{r}) = \begin{bmatrix} R_0 \\ R_N \\ \boldsymbol{r} \end{bmatrix} \tag{4.16}$$

The final source location estimate is

$$\widetilde{\boldsymbol{r}} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & I \end{bmatrix} \widetilde{\boldsymbol{y}}_{UCExt}(\boldsymbol{r}). \tag{4.17}$$

4.8.1 The Performance of Extended Unconstrained Least Squares Method (UCExt)

The performance of [GS08] is similar to that of the Steered Response Power (SRP) [DSB01] based source localization system. The UCExt method achieves similar performance to SRP with the PHAT weighting method. The drop of computational cost compared to SRP-PHAT ranges from 15 to 60 times depending on the number of reference microphones. Increasing the number of reference microphones lowers the positioning error.

4.9 Conclusions on Closed-Form TDOA Based Localization

The localization methods that use TDOA directly assume a spherical wave propagation model and the localization is a non-linear optimization problem. Therefore, closed-form solutions are desirable in many applications of this localization technique. The closed-form methods [SA87a][GS08] contain linearizations of the non-linear problem, which can lead to inaccuracies in positioning. A hybrid method that uses the closed-form solution as initialization is presented in [PK05] and therefore the search can be much faster while retaining the accuracy of the localization via optimization.

Furthermore, using closed-form solution as initialization prevents the algorithm from converging to a wrong local optimum.

Constrained closed-form methods impose additional requirements for the solution. However, solving the constraint parameters often requires iterative methods as e.g. in [HBEM01]. A closed-form approximate least squares (LS) method is presented in [SL06] that utilizes the dependence between the elements of vector $\boldsymbol{y}(\boldsymbol{r})$ of (4.9). The least squares criterion is rewritten using (4.10) and (4.11) and the result is linearized in the vicinity of the unconstrained solution $\widetilde{\boldsymbol{y}}$ using a Taylor series expansion. Clearly, the accuracy of the approximate LS method is dependent on estimation errors in the unconstrained solution (4.11). Several constrained closed-form TDOA based localization methods can be found e.g. in [BCA+14].

Chapter 5

Conclusions

The work presented in this thesis focuses on acoustic self-localization and source localization, which belong to the realm of spatial signal processing and are used in many applications such as speaker localization, speech enhancement, surveillance, and entertainment (e.g., gaming). Acoustic methods for localization and self-localization are especially useful in indoors and closed spaces where other technologies cannot be used.

Self-localization has been researched over the years and recently its application has shifted from microphone array calibration to ad hoc sensor networks as acoustic sensors in consumer products are continuously proliferating. Sensors of devices such as mobile phones can be utilized as a single unit to extract information from an acoustic scene using, e.g., sound source recognition techniques. Accessing the spatial properties of an acoustic scene requires the combination of the sensors into networks, to which well-known array processing algorithms can be applied. Many self-localization systems rely on specific calibration signals that can be distracting and impractical. Therefore using sound sources that are a natural part of an environment makes self-localization unobtrusive and furthermore relieves one from the use of specific signals, hardware, and protocols. The speech signal and its utilization as a stimulus in self-localization is one of the focal points in this thesis. The methods of the work can be utilized, e.g., in sound localization and in speech enhancement, which require knowledge of microphone positions, and therefore ad hoc microphone arrays benefit from self-localization as the underlying technology. Besides recording raw audio, surveillance applications can benefit from spatial information provided by technologies presented in this thesis.

The self-localization methods developed so far introduce certain assumptions and one of the focus areas and the work in progress in self-localization is getting rid of the assumptions and advancing towards more blind cases, where operation happens on minimum a priori knowledge. However, in many applications certain assumptions are reasonable and are useful to exploit in the design of the self-localization method. On the other hand, developing generic self-localization methods opens new application areas and helps utilization of data collected by heterogeneous sensor networks.

The development of the methods towards ad hoc sensors and consumer-level equipment have been important factors during this thesis. The motivation of development of a self-localization method in most recent works [P1][P2] has arisen from an application (e.g, utilization of mobile devices in a meeting) and therefore some assumption on the scenario are reasonable and exploitable. The assumptions on the self-localization scenarios made in these methods include knowledge of the source count. However, a source enumeration subsystem can be integrated into the self-localization systems

using, e.g., Time Difference of Arrival (TDOA) as input. Well-known techniques such as Multiple Target Tracking (MTT) can be applied in more complicated scenarios at the expense of computational cost and online operation.

This thesis studied sound source localization using multiple arrays with known microphone geometries[P4][P5], which was the early work and motivation for work on self-localization. The sound source localization methods that use direction measurements from the target are quite mature and are already in use in many applications. One challenge is the requirement of having a framework, hardware, and protocols to collect and communicate the direction estimates to a common processing unit for localization. As with the self-localization from TDOA, Direction of Arrival (DOA) based source localization requires mitigation for outliers, which can result from poor SNR, reverberation, and lower quality hardware. Furthermore, sources, other than those of interest, may need to be acknowledged in some applications. Therefore, DOA based source localization benefits from MTT techniques, but the increase in computational cost and processing delay has to be considered in each application.

References

[AB79] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[AWC+07] Sameer Agarwal, Josh Wills, Lawrence Cayton, Gert Lanckriet, David J Kriegman, and Serge Belongie. Generalized non-metric multidimensional scaling. In *International Conference on Artificial Intelligence and Statistics*, pages 11–18, 2007.

[BA00] Reinhard Blumrich and Jürgen Altmann. Medium-range localisation of aircraft via triangulation. *Applied Acoustics*, 61(1):65–82, 2000.

[BAR05] Xuehai Bian, GregoryD. Abowd, and JamesM. Rehg. Using sound source localization in a home environment. In Hans-W. Gellersen, Roy Want, and Albrecht Schmidt, editors, *Pervasive Computing*, volume 3468 of *Lecture Notes in Computer Science*, pages 19–36. Springer Berlin Heidelberg, 2005.

[BCA+14] P Bestagini, M Compagnoni, F Antonacci, A Sarti, and S Tubaro. TDOA-based acoustic source localization in the space–range reference frame. *Multidimensional Systems and Signal Processing*, 25(2):337–359, 2014.

[Ber86] Leo Beranek. *Acoustics*. Acoustical Society of America, 1986.

[BG97] I. Borg and P. Groenen. *Modern Multidimensional Scaling*. Springer-Verlag New York, Inc., 1997.

[BH97] Robert Grover Brown and Patrick Y.C. Hwang. *Introduction to Random Signals and Applied Kalman Filtering*, volume 3. John Wiley & Sons, Inc., 1997.

[Bjfrm[o]–5] Åke Björck. Linear least squares problems. In *Numerical Methods in Matrix Computations*, volume 59 of *Texts in Applied Mathematics*, pages 211–430. Springer International Publishing, 2015.

[BK02] Dirk Bechler and Kristian Kroschel. Confidence scoring of time difference of arrival estimation for speaker localization with microphone arrays. In *13. Konferenz Elektronische Sprachsignalverarbeitung ESSV*, 2002.

[BKÅ15] Simon Burgess, Yubin Kuang, and Kalle Åström. TOA sensor network self-calibration for receiver and transmitter spaces with difference in dimension. *Signal Processing*, 107:33–42, 2015.

[Bla96] J. Blauert. *Spatial hearing: the psychophysics of human sound localization*. The MIT Press, revised edition, 1996.

[Bre90] A. S. Bregman. *Auditory Scene Analysis*. The MIT Press, 1990.

[BT04] R. Biswas and S. Thrun. A passive approach to sensor network localization. In *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 2, pages 1544–1549 vol.2, Sept 2004.

[CDBBM12] M. Crocco, A. Del Bue, M. Bustreo, and V. Murino. A closed form solution to the microphone position self-calibration problem. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 2597–2600, March 2012.

[CDBM12] M. Crocco, A. Del Bue, and V. Murino. A bilinear approach to the position self-calibration of multiple sensors. *Signal Processing, IEEE Transactions on*, 60(2):660–673, 2012.

[CKN73] G.Clifford Carter, C. Knapp, and Albert H. Nuttall. Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing. *Audio and Electroacoustics, IEEE Transactions on*, 21(4):337–344, Aug 1973.

[CWB+55] Richard K. Cook, R. V. Waterhouse, R. D. Berendt, Seymour Edelman, and M. C. Thompson. Measurement of correlation coefficients in reverberant sound fields. *The Journal of the Acoustical Society of America*, 27(6):1072–1077, 1955.

[DHS01] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2001.

[Dom87] F. M. Dommermuth. A simple procedure for tracking fast maneuvering aircraft using spatially distributed acoustic sensors. *The Journal of the Acoustical Society of America*, 82(4):1418–1424, 1987.

[DSB01] Joseph H DiBiase, Harvey F Silverman, and Michael S Brandstein. Robust localization in reverberant rooms. In *Microphone Arrays*, pages 157–180. Springer, 2001.

[DWB06] H. Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *Robotics Automation Magazine, IEEE*, 13(2):99–110, 2006.

[Ear03] J. Eargle. *Handbook of Recording Engineering*. Springer, 4th edition, 2003.

[ERB+14] Alexander Ens, Leonhard M Reindl, Joan Bordoy, Johannes Wendeberg, and Christian Schindelhauer. Unsynchronized ultrasound system for tdoa localization. *Indoor Positioning and Indoor Navigation (IPIN)*, 2014.

[Fol96] J.D. Foley. *Computer Graphics: Principles and Practice*. Addison-Wesley systems programming series. Addison-Wesley, 1996.

[Fra06] K. D. Frampton. Acoustic self-localization in a distributed sensor network. *IEEE Sensors Journal*, 6(1), 2006.

[Fri87] Benjamin Friedlander. A passive localization algorithm and its accuracy analysis. *Oceanic Engineering, IEEE Journal of*, 12(1):234–245, 1987.

[GGM07] G. Giorgetti, S. K. S. Gupta, and G. Manes. Wireless localization using self-organizing maps. In *IPSN'07*, Cambridge, Massachusetts, USA, 2007.

[Gro95] S.I. Grossman. *Multivariable Calculus, Linear Algebra, and Differential Equations*. Saunders College Publishing, 1995.

[GS08] M.D. Gillette and H.F. Silverman. A linear closed-form algorithm for source localization from time-differences of arrival. *Signal Processing Letters, IEEE*, 15:1–4, 2008.

[GVL96] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 1996.

[Han97] P.C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems*. SIAM, 1997.

[Har66] Cyril Harris. Absorption of sound in air versus humidity and temperature. *Journal of the Acoustical Society of America*, 1966.

[HBEM01] Yiteng Huang, Jacob Benesty, Gary W Elko, and Russell M Mersereati. Real-time passive source localization: A practical linear-correction least-squares approach. *Speech and Audio Processing, IEEE Transactions on*, 9(8):943–956, 2001.

[HF11] M.H. Hennecke and G.A. Fink. Towards acoustic self-localization of ad hoc smartphone arrays. In *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on*, pages 127–132. IEEE, 2011.

[HN03] M. Hawkes and A. Nehorai. Wideband source localization using a distributed acoustic vector. *IEEE Transactions on Signal Processing*, 51(6), June 2003.

[HPF+09] M. Hennecke, T. Plotz, G.A. Fink, J. Schmalenstroer, and R. Hab-Umbach. A hierarchical approach to unsupervised shape calibration of microphone array networks. In *Statistical Signal Processing, 2009. SSP '09. IEEE/SP 15th Workshop on*, pages 257–260, Aug 2009.

[HT08] Yuichi S. Hayakawa and Hiro'omi Tsumara. Accuracy assessment of a post-processing differential GPS device:a case study in kaman and haçıtuğrul, central turkey. Technical report, Japanese Institute of Anatolian Archaeology, 2008.

[Jac04] M. Jacobsen. *Modular Regularization Algorithms*. PhD thesis, Technical University of Denmark, 2004.

[JGN02] Anders Johansson, Nedelko Grbic, and Sven Nordholm. Speaker localisation using the far-field SRP-PHAT in conference telephony. In *IEEE International Symposium on Intelligent Sig. Proc. and Comm. Systems*, 2002.

[Jol02] Ian Jolliffe. *Principal Component Analysis*. Wiley Online Library, 2002.

[JSHU12] Florian Jacob, Joerg Schmalenstroeer, and Reinhold Haeb-Umbach. Microphone array position self-calibration from reverberant speech input. In *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*, pages 1–4. VDE, 2012.

[JW92] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1992.

[Kal60] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45, 1960.

[KC76] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(4):320 – 327, aug 1976.

[KFM09] Anushiya A Kannan, Baris Fidan, and Guoqiang Mao. Derivation of flip ambiguity probabilities to facilitate robust sensor network localization. In *Wireless Communications and Networking Conference, 2009. WCNC 2009. IEEE*, pages 1–6. IEEE, 2009.

[KLM01] L.M. Kaplan, Qiang Le, and N. Molnar. Maximum likelihood methods for bearings-only target localization. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, volume 5, pages 3001–3004 vol.5, 2001.

[Kru64] Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[Lev44] K. Levenberg. A method for the solution of certain problems in least squares. *Quart. Applied Math.*, 2:164–168, 1944.

[LRGC01] Qiong Liu, Yong Rui, Anoop Gupta, and J J Cadiz. Automating Camera Management for Lecture Room Environments, 2001.

[LW08] Chi-Hao Lin and Chieh-Chih Wang. Probabilistic structure from sound and probabilistic sound source localization. In *Advanced robotics and Its Social Impacts, 2008. ARSO 2008. IEEE Workshop on*, pages 1–6, aug. 2008.

[Mar63] Donald W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11(2):431–441, 1963.

[MEM] Tutorial for MEMS microphones. Retrieved Sep 17, 2016. `www.st.com/resource/en/application_note/dm00103199.pdf`.

[MGW07] Mark L. Moran, Roy J. Greenfield, and D. Keith Wilson. Acoustic array tracking performance under moderately complex environmental conditions. *Applied Acoustics*, 68(10):1241 – 1262, 2007.

[MH95] W Murphy and Willy Hereman. Determination of a position in three dimensions using trilateration and approximate distances. *Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, Colorado, MCS-95*, 7:19, 1995.

[MICa] MICA2 wireleess measurement system. `http://www.eol.ucar.edu/isf/facilities/isa/internal/CrossBow/DataSheets/mica2.pdf`. Retrieved Mar 27, 2015.

[Micb] *Science of Sound Recording, Chapter 5: Microphones*. `https://ccrma.stanford.edu/courses/192a/SSR/Microphones.pdf`.

[MLH08] Iain McCowan, Mike Lincoln, and Ivan Himawan. Microphone array shape calibration in diffuse noise fields. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(3):666–670, 2008.

[MLRT04] David Moore, John Leonard, Daniela Rus, and Seth Teller. Robust distributed network localization with noisy range measurements. In *Proceedings of the 2nd international conference on Embedded networked sensor systems*, SenSys '04, pages 50–61, New York, NY, USA, 2004. ACM.

[MO07] Rainer Mautz and Washington Yotto Ochieng. Indoor positioning using wireless distances between motes. *Proceedings of TimeNav*, 7:1530–1541, 2007.

[NCMH09] Joachim Neumann, Josep Casas, Dušan Macho, and Javier Hidalgo. Integration of audiovisual sensors and technologies in a smart room. *Personal and Ubiquitous Computing*, 13:15–23, 2009. 10.1007/s00779-007-0172-1.

[Nir04] Heli Nironen. Diffuse Reflections in Room Acoustics Modelling. Master's thesis, Helsinki University of Technology, 2004.

[NIS] *Rich Transcription Spring 2005 Evaluation*. `http://www.itl.nist.gov/iad/mig//tests/rt/2005-spring/index.html`.

[NN01] D. Niculescu and B. Nath. Ad hoc positioning system (aps). In *Global Telecommunications Conference, 2001. GLOBECOM '01. IEEE*, volume 5, pages 2926–2931 vol.5, 2001.

[OKIS09] N. Ono, H. Kohno, N. Ito, and S. Sagayama. Blind alignment of

asynchronously recorded signals for distributed microphone array. In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09. IEEE Workshop on*, pages 161–164, 2009.

[PAK+05] N. Patwari, J.N. Ash, S. Kyperountas, A.O. Hero, R.L. Moses, and N.S. Correal. Locating the nodes: cooperative localization in wireless sensor networks. *Signal Processing Magazine, IEEE*, 22(4):54–69, 2005.

[PCB00] Nissanka B. Priyantha, Anit Chakraborty, and Hari Balakrishnan. The cricket location-support system. In *Proceedings of the 6th annual international conference on Mobile computing and networking*, MobiCom '00, pages 32–43, New York, NY, USA, 2000. ACM.

[Per09] Pasi Pertilä. *Acoustic Source Localization in a Room Environment and at Moderate Distances*. Doctoral thesis, Tampere University of Technology, Jan 2009.

[PHM13] P. Pertilä, M.S. Hämäläinen, and M. Mieskolainen. Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(11):2393–2402, 2013.

[Pir05] T.W. Pirinen. Normalized confidence factors for robust direction of arrival estimation. In *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on*, pages 1429–1432 Vol. 2, May 2005.

[PK05] J Michael Peterson and Chris Kyriakakis. Hybrid algorithm for robust, real-time source localization in reverberant environments. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 4, pages iv–1053. IEEE, 2005.

[PK15] V. Pulkki and M. Karjalainen. *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics*. Wiley, 2015.

[PKPP07] Pasi Pertilä, Teemu Korhonen, Tuomo Pirinen, and Mikko Parviainen. Tut acoustic source tracking system 2006. In Rainer Stiefelhagen and John Garofolo, editors, *Multimodal Technologies for Perception of Humans*, volume 4122 of *Lecture Notes in Computer Science*, pages 127–136. Springer Berlin Heidelberg, 2007.

[PMH11] P. Pertilä, M. Mieskolainen, and M.S. Hämäläinen. Closed-form self-localization of asynchronous microphone arrays. In *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on*, pages 139–144, 2011.

[PMH12] P. Pertilä, M. Mieskolainen, and M.S. Hämäläinen. Passive self-localization of microphones using ambient sounds. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 1314–1318, 2012.

[PN08] M. Pollefeys and D. Nister. Direct computation of sound and microphone locations from time-difference-of-arrival data. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2445 – 2448, 2008.

[PP07] P. Pertilä and M. Parviainen. Robust speaker localization in meeting room domain. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–497–IV–500, April 2007.

[PPH14] M. Parviainen, P. Pertilä, and M.S. Hämäläinen. Self-localization of wireless acoustic sensors in meeting rooms. In *Hands-free Speech*

Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on, pages 152–156, May 2014.

[PPKV04] P. Pertilä, M. Parviainen, T. Korhonen, and A. Visa. A spatiotemporal approach to passive sound source localization. In *Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium on*, volume 2, pages 1150–1154 vol.2, Oct 2004.

[PT13] P. Pertilä and A. Tinakari. Time-of-arrival estimation for blind beamforming. In *Digital Signal Processing (DSP), 2013 18th International Conference on*, pages 1–6, July 2013.

[Pul05] G.W. Pulford. Taxonomy of multiple target tracking methods. *Radar, Sonar and Navigation, IEE Proceedings*, 152(5):291–304, October 2005.

[RKL03] V. Raykar, I. Kozintsev, and R. Lienhart. Self localization of acoustic sensors and actuators on distributed platforms. In *International Workshop on Multimedia Technologies in E-Learning and Collaboration*, 2003.

[RKL05] V.C. Raykar, I.V. Kozintsev, and R. Lienhart. Position calibration of microphones and loudspeakers in distributed computing platforms. *Speech and Audio Processing, IEEE Transactions on*, 13(1):70 – 83, jan. 2005.

[Ros90] Thomas D. Rossing. *The Science of Sound*. Addison-Wesley, Reading, UK, 1990.

[RS87] Y. Rockah and P.M. Schultheiss. Array shape calibration using sources in unknown locations–part i: Far-field sources. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 35(3):286–299, 1987.

[SA87a] J.O. Smith and J.S. Abel. Closed-form least-squares source location estimation from range-difference measurements. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 35(12):1661–1669, 1987.

[SA87b] Julius O. Smith and J.S. Abel. The spherical interpolation method of source localization. *Oceanic Engineering, IEEE Journal of*, 12(1):246–252, Jan 1987.

[SL06] Petre Stoica and Jian Li. Lecture notes-source localization from range-difference measurements. *Signal Processing Magazine, IEEE*, 23(6):63–66, 2006.

[SMSP05] Joshua M. Sachar, Student Member, Harvey F. Silverman, and William R. Patterson. Microphone position and gain calibration for a large-aperture microphone array. *IEEE Transactions on Speech and Audio Processing*, pages 42–52, 2005.

[SSP02] J.M Sachar, H.F. Silverman, and W.R. Patterson. Position calibration of large-aperture microphone arrays. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages II–1797 – II–1800, 2002.

[Sti04] R. Stiefelhagen. CHIL evaluation data – overview of sensor setup and recordings. Technical report, Computers in the Human Interaction Loop (CHIL) Consortium, 2004.

[SVL07] Simo Särkkä, Aki Vehtari, and Jouko Lampinen. Rao-blackwellized particle filter for multiple target tracking. *Information Fusion*, 8(1):2 – 15, 2007. Special Issue on the Seventh International Conference on Information Fusion-Part {II} Seventh International Conference on Information Fusion.

[TAGB14] Mohammad J. Taghizadeh, Afsaneh Asaei, Philip N. Garner, and Hervé

Bourlard. Ad-hoc microphone array calibration from partial distance measurements. In *Proceeding of 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, May 2014.

[Tas09] I.J. Tashev. *Sound Capture and Processing: Practical Approaches*. Wiley, 2009.

[Thr05] Sebastian Thrun. Affine structure from sound. In *In NIPS*, pages 1353–1360. MIT Press, 2005.

[TR06] S.E. Tranter and D.A. Reynolds. An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1557–1565, Sept 2006.

[TRB+06] C. Taylor, A. Rahimi, J. Bachrach, H. Shrobe, and A. Grue. Simultaneous localization, calibration, and tracking in an ad hoc sensor network. In *Information Processing in Sensor Networks, 2006. IPSN 2006. The Fifth International Conference on*, pages 27–33, 2006.

[Tru99] Barry Truax, editor. *Handbook for Acoustic Ecology*. Cambridge Street Publishing, second edition, 1999.

[VDBBK+12] Bert Van Den Broeck, Alexander Bertrand, Peter Karsmakers, Bart Vanrumste, Marc Moonen, et al. Time-domain generalized cross correlation phase transform sound source localization for small microphone arrays. In *Education and Research Conference (EDERC), 2012 5th European DSP*, pages 76–80. IEEE, 2012.

[VVB88] Barry D Van Veen and Kevin M Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE assp magazine*, 5(2):4–24, 1988.

[WB95] Greg Welch and Gary Bishop. An Introduction to the Kalman Filter, 1995.

[WBBZ06] JW Weigold, TJ Brosnihan, J Bergeron, and X Zhang. A MEMS condenser microphone for consumer applications. In *Micro Electro Mechanical Systems, 2006. MEMS 2006 Istanbul. 19th IEEE International Conference on*, pages 86–89. IEEE, 2006.

[Wei] E. W. Weinstein. Cross-correlation. MathWorld – A Wolfram Web Resource. Retrieved April 8, 2015.

[Wic03] Florian Wickelmaier. An introduction to MDS. Technical report, Sound Quality Research Unit, University of Aalborg, Denmark, 2003. `https://homepage.uni-tuebingen.de/florian.wickelmaier/pubs/Wickelmaier2003SQRU.pdf`.

[WLY+10] Xiaoping Wang, Yunhao Liu, Zheng Yang, Junliang Liu, and Jun Luo. Etoc: Obtaining robustness in component-based localization. In *Proceedings of the 18th annual IEEE International Conference on Network Protocols, ICNP 2010, Kyoto, Japan, 5-8 October, 2010*, pages 62–71. IEEE Computer Society, 2010.

[WM97] M. Walworth and A. Mahajan. 3d position sensing using the difference in the time-of-flights from a wave source to various receivers. In *Advanced Robotics, 1997. ICAR '97. Proceedings., 8th International Conference on*, pages 611–616, Jul 1997.

[YHKA96] Jari Yli-Hietanen, Kari Kalliojärvi, and Jaakko Astola. Robust time-delay based angle of arrival estimation. In *Proceedings of the 1996 IEEE Nordic Signal Processing Symposium (NORSIG 96)*, pages 219–222. Citeseer, 1996.

[YHKA99] Jari Yli-Hietanen, Konsta Koppinen, and Jaakko Astola. Time-delay

selection for robust angle of arrival estimation. In *SIP*, pages 81–83. Citeseer, 1999.

[Zel88] R. Zelinski. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pages 2578–2581 vol.5, Apr 1988.

[Zio94] Lawrence Ziomek. *Fundamentals of acoustic field theory and space-time signal processing*. CRC press, 1994.

Publications

Publication 1

M. Parviainen and P. Pertilä, "Self-localization of dynamic user-worn microphones from observed speech," in *Applied Acoustics*, vol. 117, pp. 76–85, 2017.

Publication 2

M. Parviainen, P. Pertilä, and M. Hämäläinen, "Self-localization of wireless acoustic sensors in meeting rooms," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*, pp. 152–156, May 2014.

Publication 3

M. Parviainen, "Robust self-localization solution for meeting room environments," in *13th International Symposium on Consumer Electronics*, (Kyoto, Japan), 5 2009.

Publication 4

M. Parviainen, P. Pertilä, T. Korhonen, and A. Visa, "A spatiotemporal approach for passive sound source localization — real-world experiments," in *International Workshop on Nonlinear Signal and Image Processing (NSIP2005)*, 2005.

Publication 5

M. Parviainen, T. Pirinen, and P. Pertilä, "A speaker localization system for lecture room environment," in *3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, 2006.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-3819-3
ISSN 1459-2045