TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY
*Julkaisu 794 • Publication 794*

Pasi Pertilä

# Acoustic Source Localization in a Room Environment and at Moderate Distances

Tampere 2009

Tampereen teknillinen yliopisto. Julkaisu 794
Tampere University of Technology. Publication 794

Pasi Pertilä

# Acoustic Source Localization in a Room Environment and at Moderate Distances

Thesis for the degree of Doctor of Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB222, at Tampere University of Technology, on the 30th of January 2009, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2009

ISBN 978-952-15-2106-5 (printed)
ISBN 978-952-15-2137-9 (PDF)
ISSN 1459-2045

# Abstract

THE pressure changes of an acoustic wavefront are sensed with a microphone that acts as a transducer, converting sound pressure into voltage. The voltage is then converted into digital form with an analog to digital (AD) -converter to provide a discrete time quantized digital signal. This thesis discusses methods to estimate the location of a sound source from the signals of multiple microphones.

Acoustic source localization (ASL) can be used to locate talkers, which is useful for speech communication systems such as teleconferencing and hearing aids. Active localization methods receive and send energy, whereas passive methods only receive energy. The discussed ASL methods are passive which makes them attractive for surveillance applications, such as localization of vehicles and monitoring of areas. This thesis focuses on ASL in a room environment and at moderate distances that are often present in outdoor applications. The frequency range of many commonly occurring sounds such as speech, vehicles, and jet aircraft is large. Time delay estimation (TDE) methods are suitable for estimating properties from such wideband signals. Since TDE methods have been extensively studied, the theory is attractive to apply in localization.

Time difference of arrival (TDOA) -based methods estimate the source location from measured TDOA values between microphones. These methods are computationally attractive but deteriorate rapidly when the TDOA estimates are no longer directly related to the source position. In a room environment such conditions could be faced when reverberation or noise starts to dominate TDOA estimation.

The combination of microphone pairwise TDE measurements is studied as a more robust localization solution. TDE measurements are combined into a spatial likelihood function (SLF) of source position. A sequential Bayesian method known as particle filtering (PF) is used to estimate the source position. The PF based localization accuracy increases when the variance of SLF decreases. Results from simulations and real-data show that multiplication (intersection operation) results in a SLF with smaller variance than the typically applied summation (union operation).

The above localization methods assume that the source is located in the near-field of the microphone array, i.e., the source emitted wavefront curvature is observable. In the far-field, the source wavefront is assumed planar and localization is considered by using spatially separated direction observations. The direction of arrival (DOA) of a source emitted wavefront impinging on a microphone array is traditionally estimated by steering the array to a direction that maximizes the steered response power. Such estimates can be deteriorated by noise and reverberation. Therefore, talker localization is considered using DOA discrimination.

The sound propagation delay from the source to the microphone array becomes significant at moderate distances. As a result, the directional observations from a moving sound source point behind the true source position. Omitting the propagation delay results in a biased location estimate of a moving or discontinuously emitting source. To solve this problem the propagation delay is proposed to be modeled in the estimation process. Motivated by the robustness of localization using the combination of TDE measurements, source localization by directly combining the TDE-based array steered responses is considered. This extends the near-field talker localization methods to far-field source localization. The presented propagation delay modeling is then proposed for the steered response localization. The improvement in localization accuracy by including the propagation delay is studied using a simulated moving sound source in the atmosphere.

The presented indoor localization methods have been evaluated in the Classification of Events, Activities and Relationships (CLEAR) 2006 and CLEAR'07 technology evaluations. In the evaluations, the performance of the proposed ASL methods was evaluated by a third party from several hours of annotated data. The data was gathered from meetings held in multiple smart rooms. According to the obtained results from CLEAR'07 development dataset (166 min) presented in this thesis, 92 % of speech activity in a meeting situation was located within 17 cm accuracy.

# Preface

This thesis was compiled during my work at Tampere University of Technology (TUT) in the Department of Signal Processing. My research on direction of arrival (DOA) -based sound source localization is summarized in the latter part of this thesis. This topic was introduced to me by my supervisor, Professor Ari Visa. During the years 2007 – 2008 I worked with the time delay estimation (TDE) -based source localization problem. This topic is discussed in the first part of the thesis. The financial support of Tampere Graduate School in Information Science and Engineering (TISE) is acknowledged. I wish also to thank the Nokia Foundation and the Industrial Research Fund at TUT (Tuula and Yrjö Neuvo fund).

I wish to acknowledge Tuomo Pirinen's activity in organizing the spatial audio research in the Department of Signal Processing before me. I wish to express my gratitude towards my colleagues in the Audio Research Group (ARG) for creating an inspiring environment for working. I thank Teemu Korhonen for his insightful approaches and mathematical visions – this is also evident in the number of papers we have co-authored. Thanks to Mikko Parviainen for contributing to the presented research and for being an active co-author in many of the included publications. Thanks to Anssi Klapuri for his advice, and thanks to Matti Ryynänen for helping with LaTeX formatting. Thanks to Jouni Paulus, Tuomas Virtanen, Marko Helén, Toni Mäkinen, Antti Löytynoja, Atte Virtanen, Sakari Tervo, Juuso Penttilä, Mikko Roininen, Elina Helander, Hanna Silén, Teemu Karjalainen, Konsta Koppinen, Toni Heittola, and Annamaria Mesaros.

I thank my parents Heikki and Liisa, and my brother Esa for supporting me throughout my studies. Last, but not least, I would like to thank Minna for her kind support.

Pasi Pertilä
Tampere, January 2009

# Contents

# List of Figures

# List of Tables

# List of Algorithms

# List of Terms, Symbols, and Mathematical Notations

## Terms and Acronyms

| Term or acronym | Explanation |
|---|---|
| AED | Adaptive Eigenvalue Decomposition |
| AMDF | Absolute Magnitude Difference Function |
| AMSF | Absolute Magnitude Sum Function |
| ASL | Acoustic Source Localization |
| BOL | Bearings Only Localization |
| CDF | Cumulative Distribution Function |
| CLEAR | CLassification of Events, Activities and Relationships evaluation and workshop |
| CRLB | Cramér-Rao Lower Bound |
| CSD | Cross Spectral Density |
| DFT | Discrete Fourier Transform |
| DOA | Direction Of Arrival |
| DOP | Dilution Of Precision |
| FIM | Fisher Information Matrix |
| FIR | Finite Impulse Response |
| GCC | Generalized Cross Correlation |
| GPS | Global Positioning System |
| IID | Independent and Identically Distributed |
| LASER | Light Amplification by Stimulated Emission of Radiation |
| LS | Least Squares |
| MAMDF | Modified Absolute Magnitude Difference Function |
| ML | Maximum Likelihood |
| MVDR | Minimum Variance Distortionless Response |
| PDF | Probability Density Function |
| PF | Particle Filter |

| Term or acronym | Explanation |
|---|---|
| PHAT | PHAse Transform |
| PSD | Power Spectral Density |
| RADAR | RAdio Detecting And Ranging |
| RMS | Root Mean Square |
| SAD | Speech Activity Detection |
| SIR | Sampling Importance Resampling algorithm |
| SLF | Spatial Likelihood Function (of source position) |
| SNR | Signal to Noise Ratio |
| SONAR | SOund NAvigation and Ranging |
| SRP-PHAT | Steered Response Power using PHAT |
| SSL | Sound Source Localization |
| TDE | Time Delay Estimation |
| TDOA | Time Difference Of Arrival |
| VAD | Voice Activity Detection |
| WLS | Weighted Least Squares |

# Mathematical Notations

## List of symbols

| Symbol | Explanation |
|---|---|
| $a$ | Scalar variable |
| $\mathbf{a}$ | A column vector of scalars, $\mathbf{a} = [a_1, a_2, \ldots, a_N]^{\mathrm{T}}$ |
| $\mathbf{1}$ | A column vector of values 1 |
| $\mathbf{I}$ | Identity matrix, $\mathbf{I} = \mathrm{diag}(\mathbf{1})$ |
| $\mathbf{W}$ | A matrix of scalars, $\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \ldots & w_{1N} \\ w_{21} & w_{22} & \ldots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{M1} & w_{M2} & \ldots & w_{MN} \end{bmatrix}$ |
| $\mathbb{R}^{\mathrm{N}}$ | A N-dimensional space of real numbers |
| $x(t)$ | Signal x value at time $t$ |
| $\omega$ | Angular frequency [rad/s] |
| $f$ | Frequency [Hz] |
| $f_s$ | Sampling frequency |
| $L$ | Length of processing frame [samples] |
| $T_w$ | Duration of processing frame of length $L$ [s] |
| $j$ | Scalar constant value of $\sqrt{-1}$ |
| $X(k)$ | DFT of frame $\mathbf{x}(t)$ (at discerete frequency index $k$) |
| $\mu_x$ | Mean value of variable $x$ |
| $\sigma_x^2$ | Variance of variable $x$ |
| $\Omega$ | A set of elements |
| $\lambda$ | Wavelength |

## List of operators

| Notation | Explanation |
|---|---|
| $\mathcal{U}(a,b)$ | Uniform distribution between a,b |
| $\mathcal{N}(\mu,\sigma^2)$ | Normal distribution with mean $\mu$ and variance $\sigma^2$ |
| $a^*$ | Complex conjugate of $a$ |
| $\lvert a\rvert$ | Absolute value of $a$ |
| $\lceil\cdot\rfloor$ | Rounding to nearest integer |
| $\hat{\mathbf{a}}$ | Estimate of $\mathbf{a}$ |
| $\lVert\mathbf{a}\rVert$ | Euclidean norm of vector $\mathbf{a}$ |
| $D(\mathbf{a},\mathbf{b})$ | Euclidean distance between $\mathbf{a}$ and $\mathbf{b}$, $\lVert\mathbf{a}-\mathbf{b}\rVert$ |
| $\mathbf{W}^{\mathrm{T}}$ | Matrix transpose |
| $\mathbf{W}^{-1}$ | Matrix inverse |
| $\mathrm{diag}(\mathbf{w})$ | A square matrix with non-diagonal values of 0 and diagonal values specified in vector $\mathbf{w}$. |
| $\mathrm{trace}(\mathbf{W})$ | sum of diagonal values of matrix $\mathbf{W}$ |
| $E[a]$ | Expected value of $a$ |
| $*$ | Convolution operator |
| $\otimes$,$\oplus$ | Binary operators |
| $p(a;\theta)$ | Probability of $a$ parameterized by $\theta$. |
| $P(a\lvert b)$ | The likelihood of $a$ conditioned on $b$ |
| $\mathrm{Proj}_{\mathbf{b}}\mathbf{a}$ | Projection of vector $\mathbf{a}$ onto vector $\mathbf{b}$ |
| $\lvert\Omega\rvert$ | Cardinality of set $\Omega$ |
| $f(x)=\mathcal{O}(g(x))$ | Function $g(x)$ is the asymptotic upper bound for the computational time of function $f(x)$. |

Chapter 1

# Introduction

Localization has been an important task in the history of mankind. In the beginning of modern navigation one could determine his/her position at sea by measuring the angles from the horizon of celestial objects at a known time. The angles were determined via measurements, e.g., using a sextant. The celestial object's angle above the horizon at a certain time determines a line of position (LOP) on a map. The crossing of LOPs is the location. Modern navigation and localization utilizes mainly electromagnetic signals. The applications of localization include radio detecting and ranging (RADAR) systems, global positioning system (GPS) navigation, and light amplification by stimulated emission of radiation (LASER) -based localization technology. Other means of localization include utilization of sound waves in, e.g., underwater applications such as the sound navigation and ranging (SONAR).

Localization methods can be divided between active and passive methods. Active methods send and receive energy whereas passive methods only receive energy. Active methods have the advantage of controlling the signal they emit which helps the reception process. Drawbacks of an active method include that the emitter position is revealed, more complex transducers are required, and the energy consumption is higher compared to passive systems. Passive methods are more suitable for surveillance purposes since no energy is intentionally emitted. This thesis focuses on passive acoustic source localization methods.

In the era of electrical localization methods, why does one require acoustic localization? Typically the location of a source can be solved with several techniques, often even more accurately than with the use of sound. There are, however, situations where the use of sound for localization is natural. Consider the following video conference setup. A rotating camera is placed on the center of the meeting room table and the participants sit around the table. The remote end would like to see the video image of the active talker and hear his speech. How could the camera be steered to the direction of the active talker? All participants could have buttons which they press before speaking to turn

the pre-calibrated camera, a cameraman could manually turn the camera, or a microphone array could determine the speaker direction and steer the camera automatically. All these approaches would work in varying degree, but obviously the sound-based automatic camera steering is the most practical solution. Such systems have been widely developed and have been used for automatic camera management during lectures [Liu01]. However, more reverberation and noise tolerant solutions are called for. Microphones are becoming ubiquitous through the use of smart phones and laptops. They are relatively cheap and robust. Hence, acoustic localization methods hold a great potential for utilization.

Special rooms that are equipped with different sensors such as microphones, orientation sensors, and video cameras are referred as *Smart rooms*. Smart room data together with annotations are important resources for developing and evaluating automatic methods to sense human actions. For example, systems for locating people based on audio and video could be investigated separately or jointly if a smart room is equipped with microphones and video cameras. Public databases of such recordings are available [Gar07b]. Some localization methods presented in this thesis have also been evaluated in the "CLEAR technology evaluation" which uses a large database consisting of annotated smart room recordings [cle07, Mos07]. These recording rooms are located at the Society in Information Technologies at Athens Information Technology, Athens, Greece (AIT), the IBM T.J. Watson Research Center, Yorktown Heights, USA (IBM), the Centro per la ricerca scientica e tecnologica at the Instituto Trentino di Cultura[1], Trento, Italy (ITC-irst), the Interactive Systems Labs of the Universitat Karlsruhe, Germany (UKA), and the Universitat Politecnica de Catalunya, Barcelona, Spain (UPC).

## 1.1 List of Included Publications

This thesis is a compound thesis and is based on the following publications:

P1 **Pasi Pertilä, Teemu Korhonen, and Ari Visa**, Measurement Combination for Acoustic Source Localization in a Room Environment. *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2008, Article ID 278185, 14 pages, 2008.

P2 **Pasi Pertilä and Mikko Parviainen**, Robust Speaker Localization in Meeting Room Domain. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'07)*, vol. 4, pages 497 – 500, 2007.

P3 **Pasi Pertilä, Mikko Parviainen, Teemu Korhonen, and Ari Visa**, A Spatiotemporal Approach to Passive Sound Source Localization. In *Pro-*

[1]Fondazione Bruno Kessler

*ceedings of International Symposium on Communications and Information Technologies 2004 (ISCIT'04)*, pages 1150–1154, 2004.

P4 **Pasi Pertilä, Mikko Parviainen, Teemu Korhonen, and Ari Visa**, Moving Sound Source Localization in Large Areas. In *2005 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2005)*, pages 745–748, 2005.

P5 **Pasi Pertilä**, Array Steered Response Time-Alignment for Propagation Delay Compensation for Acoustic Localization. In *42nd Asilomar Conference on Signals, Systems, and Computers*. In press, 2008.

These publications are cited as [P1],[P2], etc.

### 1.1.1 List of Supplemental Publications

S1 **Teemu Korhonen and Pasi Pertilä,** TUT Acoustic Source Tracking System 2007. In R. Stiefelhagen, R. Bowers, and J. Fiscus, editors, *Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR 2007 and RT 2007. Revised Selected Papers*, volume 4625 of *Series: Lecture Notes in Computer Science*, pages 104-112. Springer, 2008.

S2 **Pasi Pertilä, Teemu Korhonen, Tuomo Pirinen, and Mikko Parviainen,** TUT Acoustic Source Tracking System 2006. In R. Stiefelhagen and J. Garofolo, editors, *Multimodal Technologies for Perception of Humans – First international Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006, Southampton, UK*, Lecture Notes in Computer Science 4122, pages 127–136. Springer, Southampton, UK, 2007.

S3 **Mikko Parviainen, Pasi Pertilä, Teemu Korhonen, and Ari Visa,** A Spatiotemporal Approach for Passive Source Localization — Real-World Experiments. In *Proceedings of International Workshop on Nonlinear Signal and Image Processing (NSIP 2005), Sapporo, Japan*, pages 468–473, 2005.

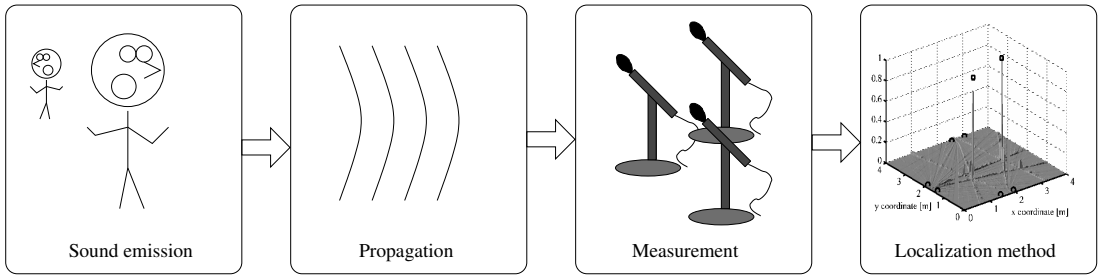



Figure 1.1: The process of sound localization can be divided into four stages: sound emission, propagation of the sound wave, reception of sound, and the actual localization algorithm.

## 1.2 Problem Description

The process of acoustic source localization (ASL) is illustrated in Fig. 1.1. The ASL problem is divided into four stages: sound emission, propagation, measurement, and localization. The first three stages represent the physical phenomena and the measurement taking place before the localization algorithm solves the source position. These stages are briefly discussed in the following subsections. This thesis focuses on the last stage and discusses signal processing methods to locate the sound source.

When discussing solutions to a problem, it is useful to classify the type of problem. According to [Tar] the prediction of results from measurements requires 1) a model of the system under investigations and 2) a physical theory linking the parameters of the model to the parameters being measured. A *forward problem* is to predict the measurement parameters from the model parameters, which is often straightforward. An *inverse problem* uses measurement parameters to infer model parameters. An example of a forward problem would state: *output the received signal at the given microphone location by using the known source position and source signal.* Assuming a free-field scenario this would be achieved by simply delaying the source signal the amount of sound propagation delay between source and microphone position and attenuating the signal relative to the propagation path length. The inverse problem would state: *solve the source location by using the measured microphone signals at known locations.* The example inverse problem is much more difficult to answer than the forward problem.

Hadamard's definition of a *well-posed* problem is

1. A solution exists
2. The solution is unique
3. The solution depends continuously on the data

A problem that violates one or more of these rules is termed *ill-posed*. During

this thesis it will become evident that sound source localization is an ill-posed inverse problem in most of the realistic scenarios.

### 1.2.1 Sound Source

Sound source localization system is often designed for a specific application which leads to assumptions about the source. For example, in the case of locating a talker for video conferencing some assumptions about the movement of humans can be applied. In addition, the speech signal has special characteristics originating from the speech production system that differentiate it from other signals. Signal characteristics such as bandwidth and center frequency can also guide the selection of a suitable localization scheme. A coarse characterization of the source signal between a narrowband and wideband signal is typically made. Many commonly occurring audio signals are wideband, e.g., human speech and jet aircraft represent typical wideband signals, while, e.g., some bird calls could be considered as narrowband, consisting of a few individual frequencies. The source can also be directive, possibly as a function of frequency, such as a human talker [Dun39]. However, the presented methods do not exploit directionality. It is also noted that detecting a source or enumerating sources is a separate problem from localization although they are somewhat related. These problems are not discussed here.

### 1.2.2 Sound Propagation

Sound is mechanical vibration of particles, and is propagated as a longitudinal wave. It therefore requires a medium (here, air) to exist. Accurately modeling sound propagation from an unknown source position to the sensor is not trivial, and the physical properties of sound propagation are therefore briefly reviewed.

A rough division between near-field and far-field sources can be made based on the geometry of the problem setting. Far-field methods assume that the source emitted wavefront is a plane wave at the receiving microphone array. In the near-field situation, the received wavefront is curved. In a way, the far-field assumption is an approximation of the near-field situation.

This work discusses sound source localization for indoor and outdoor applications. In both scenarios the received waveform is disturbed by background noise and multipath propagation effects. In indoors the multipath effects are caused by sound *reflections* from room surfaces and objects larger than the wavelength. Sound bends around objects that are smaller than the wavelength. This phenomenon is called *diffraction*. For example, a 2000 Hz signal having the wavelength of 17 cm would reflect from an office wall but not from a coffee mug. Reflections can be *specular* (mirror-like) or *diffuse* where sound is reflected into directions not assumed by the specular reflection. Diffuse reflections cause *scattering* of the wave, i.e., difference between the ideal wave behavior and the actual

wave behavior.

Enclosures can be characterized by their acoustical properties. A typical measure is the amount of *reverberation* expressed as the time sound pressure takes to attenuate 60 dB after switching off a continuous source [Ros90]. The reverberation time is noted as $T_{60}$ (s). Reverberation is related to the surface absorption coefficient $\alpha_i$ which determines how much sound is reflected from the surface and how much is absorbed. The absorption coefficient is a function of incident angle, frequency, and material properties [Ber86]. In practical calculations the coefficient may be thought to be averaged over random incidence angle. The *reverberation time* is related to the absorption coefficient through Sabine's equation [Ros90], which can be used to approximate $T_{60}$ when room volume $V$, reflection surface area $S_i$, and respective absorption coefficients $\alpha_i$ are known:

$$T_{60} = 0.163\frac{V}{A}, \tag{1.1}$$

where $A$ is total absorption surface area obtained by $A = \sum_i \alpha_i S_i$. Different desirable reverberation times for various activities exist. The optimum reverberation time is a compromise between clarity (requires short reverberation time), liveliness (requires long reverberation time), and sound intensity (requires a high reverberation level) [Ros90]. An auditorium designed for speech should have a lower reverberation time than an auditorium designed for music.

In a free-field environment the sound pressure level drops approximately 6 dB for doubling the distance from a point-like sound source. This is also known as geometric spreading attenuation $A_s$ (dB). Such conditions are sometimes assumed to exist in an anechoic chamber or outdoors. Atmospheric attenuation $A_a$ and excess attenuation $A_e$ also contribute to sound attenuation. Atmospheric absorption increases at high frequencies and is detailed and empirically quantified in [IE93]. For example, in 20 °C temperature and at 70 % humidity a 250 Hz tone will experience an attenuation of 1.1 dB/km whereas an 8 000 Hz tone will experience a large 76.6 dB/km attenuation. The excess attenuation term $A_e$ is used to group other attenuation contributions. Also ground causes attenuation due to interference.

When a wave is incident at an oblique angle to a boundary of two mediums the passing sound is *refracted*. Refraction changes the heading of the wave into the direction of lower sound velocity medium. For example, the wind speed normally grows when altitude increases [Cro97] and the sound is therefore refracted downwards in the direction of wind and upwards in headwind. Similarly, if the air cools upwards the sound will bend upwards since the speed of sound decreases as a function of temperature. Such a situation can occur on a sunny day when the sun warms the ground. When the temperature increases upwards an inversion exists and the sound is refracted downwards. For a tutorial on sound propagation in outdoors refer to [Emb96].

### 1.2.3 Measurement

A transducer is used to convert sound pressure changes into corresponding voltage changes. The voltage changes are converted into digital form with an analog-to-digital (AD) converter. The sound signal is captured with multiple spatially separated microphones. These installations are often referred as *microphone arrays* or arrays in short. In the scope of this thesis the microphone locations are assumed to be known, and the microphone radiation pattern is assumed omnidirectional. The choice of microphone positioning can favor or even hinder the use of different localization methods. In the case of *ad hoc sensor networks*, one does not get to choose the geometry.

### 1.2.4 Localization Algorithm

After converting the pressure changes into digital form, several ways to obtain information about the spatial properties of the sound source exist. Assumptions about signal propagation, background noise, source signal type, and source directivity must be made. All these assumptions together are used as the justification for the selected localization method.

#### Energy-Based Methods

For example, let us assume that we are interested in locating a sound source in an environment where background noise is negligible. The source is assumed isotropic and the sound pressure attenuation is assumed inverse proportional to the distance. The received signal energy is measured using two microphones. Substituting the ratio of measured energies to be the ratio of two (squared) distances (from the microphones to the source) defines a set of points. The set is a circle on which the source must reside (circle of Apollonius). Using three microphones gives two ratios and therefore two circles, which intersect at two points. The source location is now either one of them, assuming they are two separate points. Adding a fourth microphone resolves the location ambiguity. If the amount of background noise is suddenly increased a bias is introduced in the energy measurements. The final location estimate is therefore biased in low signal-to-noise (SNR) conditions, without further improvements to the method. Knowing the conditions of the final application space is therefore essential for choosing and developing suitable localization method. In [She05] an energy-based maximum likelihood approach is described.

#### Beamforming

*Beamforming* is a popular method for source localization [Tre02, Mor07, Yan03, Joh93]. A basic sum-and-delay beamformer steers the received array signals into the desired direction by applying a microphone placement specific steering delay

to each array signal. The resulting signals are then summed to acquire the *directional response* of the array. Traditionally, the direction that maximizes the response power represents the dominant source direction. To avoid spatial aliasing, the sensors should be spaced less than half a wavelength apart $d \leq \lambda/2$. The maximum frequency detected without spatial aliasing is then $f_{\max} = c/(2d)$. Statistical beamforming utilizes the characteristics of the received signal to form an optimal beamformer. Such methods include the optimal beamformer, also known as Capon or minimum variance distortionless response (MVDR) beamformer. This subclass of beamformers utilizes the frequency dependent covariance matrix estimate of the received signals [Tre02]. In practice the covariance matrix is not available and must be estimated from ensemble averages. However, if the signal is not stationary between adjacent frames, such as in the case of speech, this estimation can be problematic [DiB01b]. Also, the estimation introduces errors in to the process if the assumptions about noise and signal characteristics do not hold [Mor07]. The assumption about the signal of interest being narrowband causes additional computational load in the case of wideband audio signals, since the covariance matrix has to be calculated for all frequency bands to which the processing bandwidth is divided. The wideband approaches include the subband decomposition scheme, where the signal is divided into several subbands that are shifted to the baseband. Each baseband signal is then processed separately and combined after the direction estimation [Yan03].

In [Moh08] direction estimation of multiple sources is considered. The method assumes that the sources are sparse, i.e., locate in different time-frequency regions. A coherence test is provided to detect such low-rank time-frequency bins which are then used to estimate the narrowband spectrum of each bin (using MUSIC algorithm). The directional spectrum is summed over time and frequency to obtain DOA estimates. The clustering the low-rank covariance matrices and estimation of the narrowband spectra of each cluster is also proposed. The method is tested in reverberation time $T_{60}$ value 250 ms and SNR range 15–30 dB for a hearing-aid application using a small array, and for moving vehicles and gunfire.

Although spectral estimation techniques for direction finding are not considered in this thesis, they provide a well studied alternative to time delay based methods.

### Time Delay Estimation -Based Localization

Time delay estimation (TDE) methods [Che06, Kna76, Has81, Car87, Bra99, Ros74, Che05a, Ros74, Jac93, Ben00, Doc03, Ree81, You84, You86, Hah06, Bra99, Yeg05, Ray05, Lai99] are suitable for wideband signal processing. It is assumed that a coherent wavefront passes two microphones at time instants depending on the microphone locations and the shape and direction of the arriving wavefront. The propagation delay between the microphones can be estimated based on the temporal similarity of the microphone signals – ideally the signals

differ only temporally. The theoretical behavior of the TDE methods has been extensively studied in literature [Car81, Wei83b, Sad06, Cho81, Ash05, Koz04, Cha96, Gus03, Zha08, DB03] and a brief description is given in Section 2.7.

TDE-based near-field localization methods can be divided into two classes. The two-step TDE-based approach utilizes microphone pairwise time difference of arrival (TDOA) estimates. The location is solved in a closed-form [Sto06, Zhe07, Gil08, Smi87a, Smi87b, Fri87, Hua00, Hua01, Cha94, Ho04, So03], iteratively [Bra95, Sva97, Ray05, Sil05], or in a sequential Bayesian framework [Kle06, Gan06, Ver01, Vog07]. The TDOA-based localization problem is non-linear in respect to the unknown source position which has resulted in multiple solution schemes for the problem.

The two-step methods are not robust towards corrupted TDOA values that are present in noisy and reverberant environments. The one-step TDE-based ASL methods utilize directly the TDE measurements to infer source position [Aar03, Omo94, DiB01a, Bra01, Che01, Val07, Leh04, Cir08, Ber91, Do07a, Do07b, Dmo07, Zot04, Gar07a, War03, Leh03, Leh06],[P1],[S1] and are generally more robust towards noise and reverberation. The microphone pairwise TDE measurements are combined to obtain a *spatial response*. Similarly to the directional response methods the traditional approach is to maximize the spatial response to locate the source.

In the far-field scenario the sound wavefront is planar instead of spherical. The localization can be performed by combining several wavefront direction of arrival (DOA) measurements from spatially separated arrays [Tor84, Blu00, Haw03, Kap01, Dom87, Guo08],[P2],[P3],[P4],[S2],[P5] or from a single array [Kar05]. The DOA estimate is traditionally obtained by parameterizing the steered response by the direction that maximizes the response power. A more robust way, which is similar to TDE-based likelihood localization, is to combine directly the array steered responses [P5],[Ali07] to build the spatial likelihood function of source position.

In large spaces the problem is complicated by the limited sound propagation speed [Blu00, Kap01, Dom87, Guo08],[P3],[P4],[P5]. A directional estimate of a moving sound source therefore points behind the true source position. Including a simple propagation delay model is discussed for two cases:

1. Array output is the wavefront DOA estimate [Blu00, Kap01, Dom87, Guo08],[P3],[P4],[S3].

2. Array output is the directional steered response [P5].

## 1.3 Overview of Thesis

Chapter 2 discusses the free-field and room impulse response signal models. The source localization geometry is illustrated in the room environment. Time delay

estimation (TDE) theory is then reviewed and time difference of arrival (TDOA) estimation methods are discussed along with signal processing concepts required by the sound source localization task. A practical measurement room environment and a simulated room environment are described. In the simulated room environment, the reverberation time and noise conditions are varied. The TDOA performance bounds are then briefly introduced.

Chapter 3 first presents the problem of locating a sound source from TDOA measurements in the near-field. The Maximum Likelihood (ML) method is then introduced and the Cramér-Rao lower bound (CRLB) is given. The dilution of precision is also introduced. Sequential localization methods using TDOA measurements are then discussed briefly. The spatial response constructed from TDE measurements is then discussed for localization purposes. The widely known steered response power using phase transform (SRP-PHAT) is one such method. Source localization with steered response methods is discussed by first considering the direct maximization of the response. This is followed by the sequential Bayesian approach with the numerically effective method known as Particle Filtering (PF). It is shown that the localization performance of SRP-PHAT using PF is improved by changing the way the TDE measurements are combined [P1]. This is also verified with the CLEAR'07 database. The localization performance of a ML TDOA method is compared to the TDE measurement based method in the simulated room environment.

Chapter 4 discusses direction of arrival (DOA) -based localization methods. First, the closed-form localization is discussed followed by a more robust localization method based on DOA discrimination [P2]. The problem of limited speed of sound and delayed observations is then described. The propagation delay from source to array is then proposed to be included into the DOA-based localization model [P3],[P4]. The chapter proceeds by discussing TDE-based array directional responses, i.e., DOA estimation using TDE measurements. The combination of array directional responses is then presented for localization using the principles discussed in the near-field case in Chapter 3. The propagation delay is then proposed to be included in this model [P5]. Chapter 5 concludes the discussion and presents future work ideas. Errata of included publications is given in Chapter 6.

## 1.4 Author's Contributions

The author has written and contributed the majority of the research in each of the included publications P1–P5. This section lists the main contributions of this thesis.

Section 3.1 is a literature review and groups existing TDOA based closed-form localization methods.

Section 2.6.3 is based on [P1] and presents a novel way of combining the TDE likelihood measurements for near-field localization. The main result is that

combining the TDE likelihoods with an intersection operation results in a source likelihood distribution with smaller variance compared to the union of the TDE likelihoods. This has been numerically verified and tested with real data in [P1] using particle filters. The method has been further tested with 3.3 hours of data (CLEAR 2007 evaluation) in different smart rooms [S1].

Section 4.3 is based on [P2] and presents a robust DOA-based localization scheme. The scheme has been tested with 3.2 hours of real data in different smart rooms (CLEAR 2006 evaluation) and the results are published in [S2].

Section 4.4 is based on [P3] and [P4] and presents a novel method of using the propagation delay between the source and observer in DOA-based localization. Real data results are presented in [S3].

The near-field TDE likelihood localization is extended to far-field localization in Section 4.5. Section 4.6 is based on [P5] and applies the propagation delay model presented in Section 4.4 for TDE likelihood based far-field localization presented in Section 4.5.

## 1.5 Related Work

The ASL problem has been extensively studied in recent years and several theses on the topic have been written. In [Bru07] the indoor localization methods are reviewed and main contributions are in the use of global coherence field (GCF) in localization and in determining the speaker's head orientation. In [Gar07a] the use of multiple microphones inside a smart room for perceiving humans is discussed with the focus on beamforming approach. The thesis also contributes on speaker head orientation and microphone array speech enhancement and recognition. In [Leh04] a general framework for acoustic source localization using sequential Monte Carlo methods (particle filters) is presented.

Chapter 2

# Time Delay Estimation

The capability of estimating the time difference of arrival (TDOA) of a source emitted wavefront between two microphones is essential for many localization methods. Therefore, it is important to describe the TDOA estimation problem and to review its solutions before describing TDOA-based localization methods.

This chapter's outline is the following. Section 2.1 discusses signal model for the free-field case. Section 2.2 then reviews the impulse response model, suitable for reverberant enclosures. The utilized measurement room is described in Section 2.3. A simulated environment is described in Section 2.4. Section 2.5 defines and illustrates the TDOA estimation problem, and Section 2.6 reviews TDOA estimation methods. In Section 2.7 the correlation-based TDOA estimation performance bounds are discussed for the free-field and reverberant environments. Finally, Section 2.8 summarizes the discussion.

## 2.1 Signal Model

The sound propagation in air can be modeled with the (linear) wave equation [Joh93]. The one-dimensional equation of motion, or acoustical wave equation relates the second derivative of pressure with the $x$ coordinate to the second derivative of pressure with time $t$ and square of the speed of sound $c$

$$\frac{\partial^2 p}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = 0. \tag{2.1}$$

In this work, an approximation of propagation is adopted where the speed of the wavefront is parametrized by the temperature. The speed of sound waves in gases is given by [Ros90]

$$c = \sqrt{\frac{\gamma R T}{m}}, \tag{2.2}$$

where $T$ is absolute gas temperature [K], and $m$ is the molecular weight of gas [kg/mol]. For air $m$ equals $2.88 \cdot 10^{-2}$ kg/mol, $R$=8.31 [J/mol K] and the adiabatic constant for air is $\gamma$=1.4. The speed of sound in air is then reduced to

$$c = 20.1\sqrt{T}. \tag{2.3}$$

For a room of 19 °C temperature $c$ is 343 m/s, which will be used as the default value hereafter.

Two solutions of the wave equation (2.1) are considered in the free-field scenario:

- The far-field sound source is modeled as the solution to the plane wave equation and the signal is written in point $\mathbf{m} = [m_x, m_y, m_z]^{\mathrm{T}}$ at time $t$ as [Joh93, Ch.2]

  $$x(\mathbf{m}, t) = A\exp(j\omega(t - \mathbf{k}^{\mathrm{T}}\mathbf{m})), \tag{2.4}$$

  where $j$ is the imaginary unit, $A$ is amplitude, $\omega$ is angular frequency, and $\mathbf{k} = [k_x, k_y, k_z]^{\mathrm{T}}$ is the propagation vector or *slowness vector* pointing towards wave travel direction with magnitude equal to reciprocal of $c$, i.e., $\|\mathbf{k}\| = c^{-1}$.

- The near-field sound source signal is modeled after the solution to the spherical wave equation (in spherical coordinates) and is written [Joh93, Ch.2]

  $$x(r, t) = \frac{A}{r}\exp(j\omega(t - r/c)), \tag{2.5}$$

  where $r$ is the range between sensor and source.

The linearity of the wave equation implies that any sum of solutions is also a solution. This fact can be used in (2.4) and (2.5) to encompass wider bandwidth signal models by integrating the equations over the desired frequency range.

## 2.2 The Impulse Response Model

When a sound wave reaches a surface, both transmitted and reflected waves are formed. The absorption coefficient of the surface determines the amount of absorbed sound energy. Specular sound reflection can be modeled with the concept of an *image source* [All79]. From a geometrical perspective the reflected sound originates from a mirror image of the source (mirrored from the wall surface), see Fig. 2.1 for illustration. The distance from the mirrored source to the receiver determines the propagation time. The received signal is therefore a sum of delayed and decayed source signals[1]. The reverberation process can be modeled with a linear impulse response. The impulse response includes the direct

[1]The propagation process is more complicated, e.g., when the surface has small shapes and irregularities of the size of wavelength the wave will be scattered into many directions. The process is referred as diffuse reflection [Cro97].

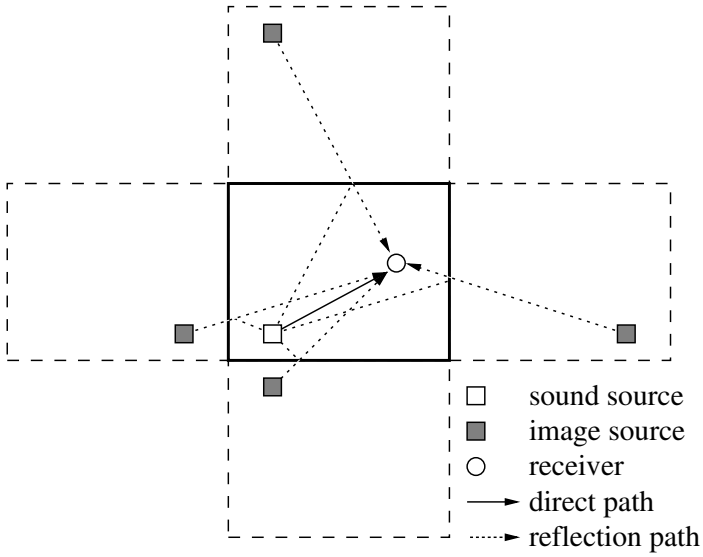



Figure 2.1: The concept of image source is illustrated in a rectangular room using first order reflections and the direct path. The reflections from room walls seem to originate from mirrored sound sources called image sources. The mirroring can be continued to encompass higher order reflections by adding more virtual rooms.

path signal and the reflected signals, along with the measurement equipment responses. In the case of multiple sources the received signal $x_k(t)$ is written as a sum of source signals $s_i(t)$ convolved with the corresponding linear impulse response $\mathbf{a}_{i,k}(t)$ between source $i$ and microphone $k$ at time $t$ [Bra01, Ch.8]:

$$x_k(t) = \sum_{i=1}^{N} \mathbf{a}_{i,k}(t) * s_i(t) + w_k(t), \tag{2.6}$$

where $*$ is the linear convolution operator, and $N$ is the number of sources. The noise term $w_k(t)$ is assumed independent and identically distributed (IID). Note that the room impulse response $\mathbf{a}_{i,k}(t)$ can be time varying.

The distance between the source and the sensor determines the propagation time of the direct path

$$\tau_{i,k} = c^{-1} D(\mathbf{r}_i, \mathbf{m}_k) = c^{-1} \parallel \mathbf{r}_i - \mathbf{m}_k \parallel, \tag{2.7}$$

where $\mathbf{r}_i$ is the position of $i$th source, $\mathbf{m}_k$ is the position of $k$th microphone $k = 1, \ldots, M$, $\parallel \cdot \parallel$ represent the Euclidean norm, and $D(\cdot, \cdot)$ is the Euclidean distance between two points.

In a simplified case, only the direct propagation path exists and the isotropic source radiates in a lossless medium. The simplified signal model is written

$$x_k(t) = s(t - \tau_{i,k}) + w_k(t). \tag{2.8}$$

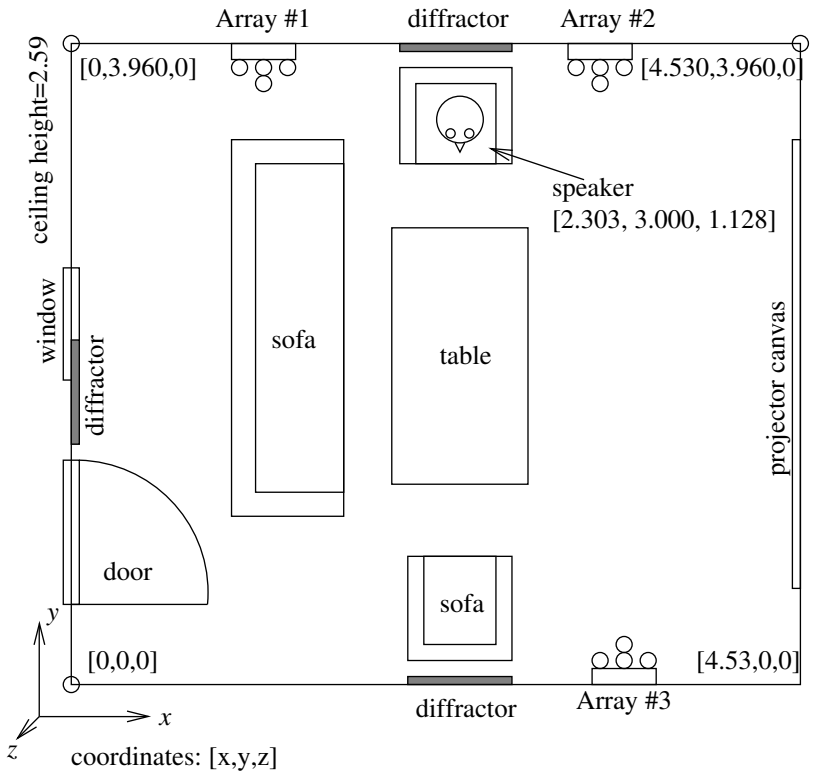



Figure 2.2: A floor plan of the recording room. The room is a meeting room and is equipped with microphones grouped into three arrays at locations specified in Table 2.1. The room is also equipped with furniture and other small objects.

For a short time period the signal's statistical properties are assumed unchanged. For example, the properties of human speech signal can be considered unchanged for short a period of 10 – 20 ms [Che05b, p.864]. Therefore, in several practical signal processing applications the signal is processed in frames of data during which the assumptions about signal properties hold. A frame of data from microphone $k$ is noted as

$$\mathbf{x}_k(t) = [x_k(Lt), x_k(Lt+1), \ldots, x_k(Lt+L-2), x_k(Lt+L-1)]^{\mathrm{T}}, \tag{2.9}$$

where $L$ is the frame length in samples, $t$ is now frame index[2], and $x(n)$ indicates value of signal $x$ at discrete time index $n$. The data from all $M$ microphones at time index $t$ is noted

$$\mathbf{X}(t) = [\mathbf{x}_1(t), \mathbf{x}_2(t), \ldots, \mathbf{x}_M(t)]. \tag{2.10}$$

The frame length in seconds is $T_w = L/f_s$, where $f_s$ is the sampling frequency. Typical values of $f_s$ range from 8000 to 48000 Hz and the frame length is commonly selected between 10 ms and hundreds of milliseconds.

[2]Note that $t$ depends on whether indexing a signal or a frame of signals.

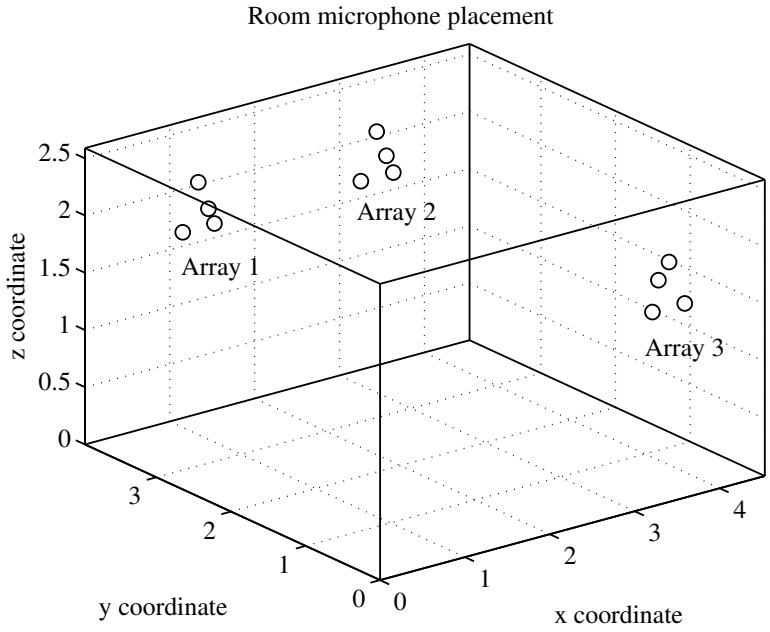



Figure 2.3: The microphone locations on the recording room walls are illustrated ("∘"). The floor plan is given in Fig. 2.2 and microphone coordinates are given in Table 2.1.

## 2.3 Practical Measurement Environment

A description of a meeting room equipped with microphones is displayed in Figs. 2.2 and 2.3. Recordings from this environment are considered in this work and used as illustrative examples. Microphones are grouped into arrays of four microphones and their coordinates are given in Table 2.1. The microphone locations are also visualized in Fig. 2.3. An example of an impulse response from the given room is displayed in Fig. 2.4. The example is calculated with the method proposed by Farina [Far00] from a logarithmic sine-sweep of duration 20 seconds on frequency band 100–20000 Hz. The direct path has the strongest peak and

Table 2.1: Microphone coordinates are given (mm) and their corresponding array 1–3 is shown. The coordinate system is the same used in Figs. 2.2 and 2.3. For a 3D visualization of the geometry, see Fig. 2.3.

| | Array 1 | | | | Array 2 | | | | Array 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **mic** | x | y | z | **mic** | x | y | z | **mic** | x | y | z |
| **1** | 1029 | 3816 | 1690 | **5** | 3127 | 3816 | 1715 | **9** | 3714 | 141 | 1630 |
| **2** | 1405 | 3818 | 1690 | **6** | 3507 | 3813 | 1715 | **10** | 3335 | 144 | 1630 |
| **3** | 1215 | 3819 | 2088 | **7** | 3312 | 3814 | 2112 | **11** | 3527 | 140 | 2030 |
| **4** | 1215 | 3684 | 1898 | **8** | 3312 | 3684 | 1940 | **12** | 3517 | 270 | 1835 |

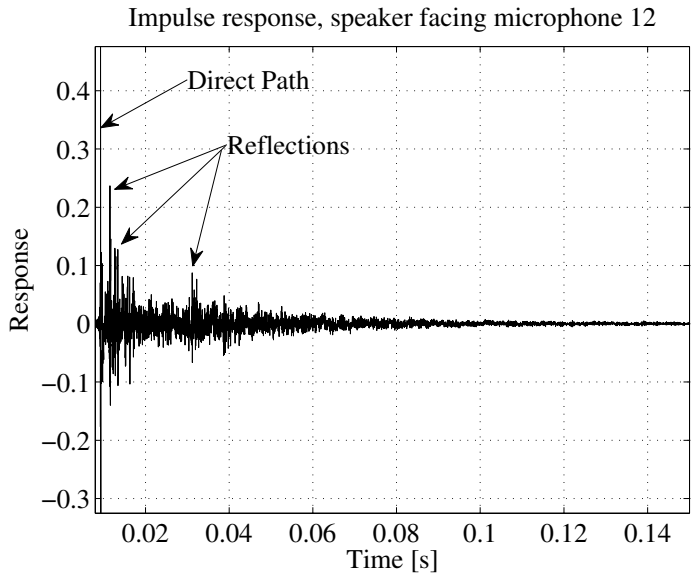



Figure 2.4: A measured impulse response is depicted between microphone 12 (Array 3) and a loudspeaker located on a table at coordinates $[2.5, 3.1, 0.7]^{\mathrm{T}}$. The loudspeaker points towards the microphone. Refer to Fig 2.2 for the room layout.

the shortest propagation time since the speaker is facing the microphone. The measured reverberation time $T_{60}$ of the meeting room[3] is 0.23 s. An example of a person uttering a sentence in the room is displayed in Fig. 2.6(a) in the form of a spectrogram. One processing frame is illustrated in Fig. 2.5.

## 2.4 Simulated Room Environment

A segment of data is simulated with the image source method [All79]. The algorithm constructs the room impulse response between the sound source and a microphone. First, the time sound travels from the source to the microphone is quantized into samples. A value based on distance attenuation is then assigned to the room impulse response indexed by the quantized time delay. The contribution of a source is therefore an impulse with an amplitude. Similarly, for each image source the distance to the microphone determines the impulse index. The impulse value is determined by the distance attenuation and the loss of energy from each reflection, determined by the reflection coefficient. For listening purposes and for mono recordings the impulse delay quantization into samples may be sufficient. Multichannel simulations require more precise time delay between the impulse response components since the quantization of impulse location into samples does not represent a realistic scenario. Peterson [Pet86] presented a version of the

[3] $T_{60}$ was obtained using Schroeder integration of the impulse response [Sch65], $T_{60}$ standard deviation was 0.0087 s over five repetitions.

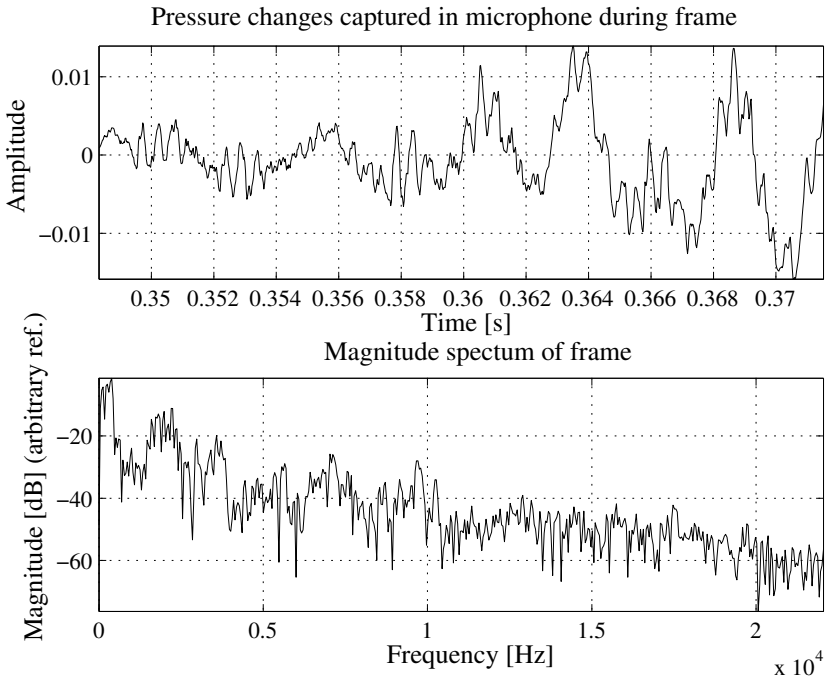



Figure 2.5: Waveform and corresponding amplitude spectrum of the frame outlined in Fig. 2.6(a) is displayed. The sampling frequency is 44100 Hz and the frame length is 23.2 ms.

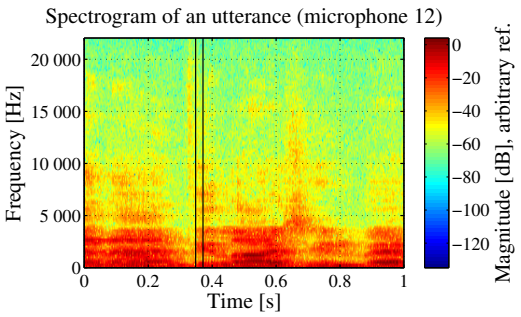
image source method that is suitable for multichannel simulations. Instead of assigning a single value to the room impulse response at the quantized time delay, a lowpass version of the Dirac's delta function is assigned to a window centered on the true delay $t = 0$. The lowpass impulse response values of adjacent quantized time indices are obtained from the Hanning-windowed ideal lowpass function:

$$h(t) = \begin{cases} 0.5\,[1 + \cos(2\pi t/T_w)]\,\mathrm{sinc}(2\pi f_c t), & -T_w/2 < t < T_w/2, \\ 0, & \text{otherwise} \end{cases}, \qquad (2.11)$$

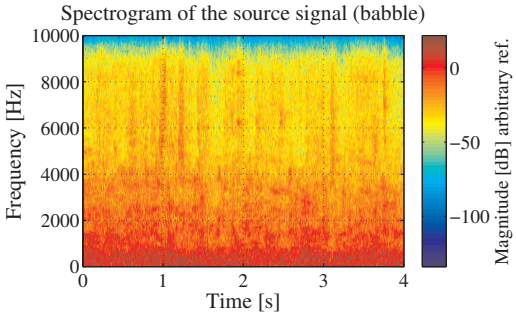
where $h(t)$ is filter response to an impulse at time $t = 0$, $f_c$ is filter cutoff frequency (here $f_s/2$), and $T_w$ is window duration (here 2 ms). The microphone signal is obtained by convolving the source signal with the generated room impulse response (2.6). The source and microphones are assumed omnidirectional.

Figure 2.7 depicts the simulation setup. The room dimensions are equal to the room considered in Fig. 2.2, i.e., $[4.53, 3.69, 2.59]^{\mathrm{T}}$ . The reflection coefficients of the walls $\beta_{wall}$ are varied between 0 and 1, and the ceiling and floor coefficients are obtained by $\sqrt{\beta_{wall}}$ for a more realistic setup. The corresponding $T_{60}$ values are evaluated with the Sabine's equation (1.1) and the absorption coefficient is obtained from the reflection coefficient $\alpha_i = 1-\beta_i^2$, where $i = 1, \ldots, 6$ corresponds to the room surface number [All79].

The simulation includes 32 microphones that are placed in pairs at the heights




(a) Recorded speech from a talker at coordinates $[2.303, 3.000, 1.128]^{\mathrm{T}}$ facing the wall with the microphone 12 at coordinate $[3.517, 0.270, 1.835]^{\mathrm{T}}$ is displayed. The distance between speaker and microphone is 3.1 m.



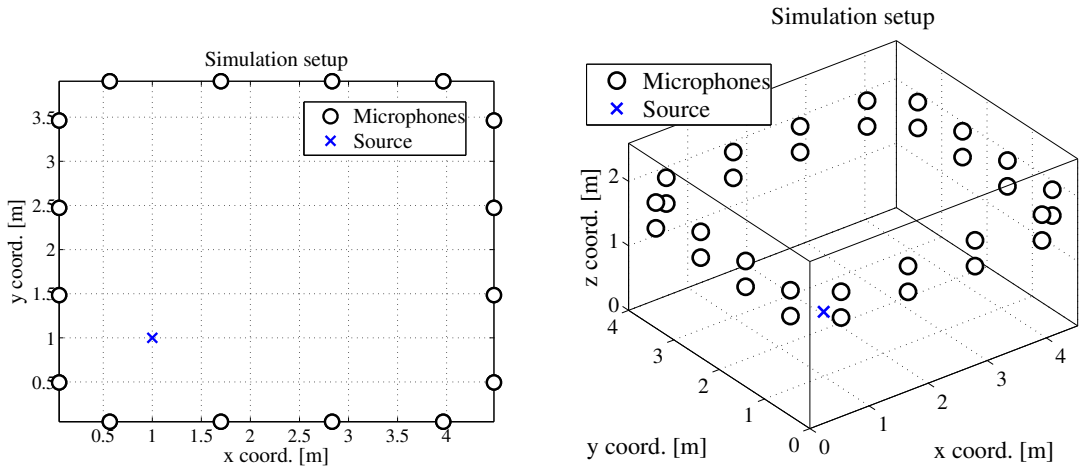

(b) Speech babble segment used in simulations is displayed. The speech is recorded in a canteen with 100 people. Sampling frequency is 19.98 kHz.

Figure 2.6: Signal spectrograms are displayed. The horizontal axis represents time (s), and the vertical axis represents frequency (Hz). Panel (a) displays recorded speech signal (sampling frequency is 44100 Hz). A processing frame of length 23.2 ms is outlined with black vertical lines. Panel (b) displays speech babble used in simulations.

of 1.5 m and 1.9 m and 5 cm out of the wall. The microphones on each wall are equally spaced apart to cover the wall. See Table B.1 (Appendix B, p. 147) for details. The sampling frequency was set to 44100 Hz and 16 bits per sample was used. The source is located at $[1, 1, 1]^{\mathrm{T}}$. The test signal consists of 4 seconds of babble recorded in a canteen with 100 people [IfPT90]. The spectrogram of the speech babble segment used is displayed in Fig. 2.6(b). 14 different reflection coefficients are simulated, which correspond to different room reverberation time $T_{60}$ values specified in Table 2.2.

Table 2.2: Reverberation time $T_{60}$ values for the simulations are presented. The $T_{60}$ values are obtained from the reflection coefficients $\beta_{\mathrm{wall}}$ with Sabine's equation (1.1). For an illustration of the room refer to Fig. 2.7. The first recording setup corresponds to an anechoic room.

| **recording** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $\beta_{\mathrm{wall}}$ | 0 | 0.2 | 0.4 | 0.5 | 0.6 | 0.7 | 0.75 |
| $T_{60}$ [s] | 0.0937 | 0.1055 | 0.1280 | 0.147 | 0.1761 | 0.2255 | 0.2653 |
| **recording** | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| $\beta_{\mathrm{wall}}$ | 0.8 | 0.825 | 0.85 | 0.875 | 0.8875 | 0.9 | 0.925 |
| $T_{60}$ [s] | 0.3253 | 0.3683 | 0.4256 | 0.5060 | 0.5596 | 0.6267 | 0.827 |




(a) Top view

(b) 3D view

Figure 2.7: The simulation setup is depicted. The 32 microphones are marked with circles "$\circ$" and the source ("$\times$") is located at $[1, 1, 1]^{\mathrm{T}}$.

The signal-to-noise ratio (SNR) is here defined as

$$\mathrm{SNR} = 10 \log_{10} \frac{\sum_{t=0}^{T-1} x(t)^2}{\sum_{t=0}^{T-1} w(t)^2} \ \ [\mathrm{dB}], \tag{2.12}$$

where $t$ is discrete index, $T$ number of samples in signal $x$, and $w$ is IID noise generated from the normal distribution with zero mean. Different SNR levels are obtained by adding noise with a specific variance to reach a desired level $\pm 0.1$ dB.

The simulations will be used in the following chapters of this thesis.

## 2.5 Time Difference of Arrival

A source position in space is mapped into a time difference of arrival (TDOA) value between two microphones. A microphone pair $p$ constitutes of microphones $\{l, k\}$ where $l, k \in [1, \ldots, M], k \neq l$. The set of all unique microphone pairs is noted $\Omega$, and the cardinality of pairs is $S = |\Omega| = \binom{M}{2}$. Using the microphone positions of the pair $p$ ($\mathbf{m}_l$ and $\mathbf{m}_k$) and the source position $\mathbf{r}_i$ the TDOA between the pair is written

$$\Delta\tau_{p,\mathbf{r}_i} = (D(\mathbf{r}_i, \mathbf{m}_l) - D(\mathbf{r}_i, \mathbf{m}_k)) \cdot c^{-1}. \tag{2.13}$$

The delay value in discrete time samples is

$$\lceil \Delta\tau_{p,\mathbf{r}_i} \cdot f_s \rfloor,$$

where $\lceil \cdot \rfloor$ represents rounding to nearest integer, and $f_s$ is the sampling frequency. The function (2.13) is a mapping from a three dimensional space (position) to

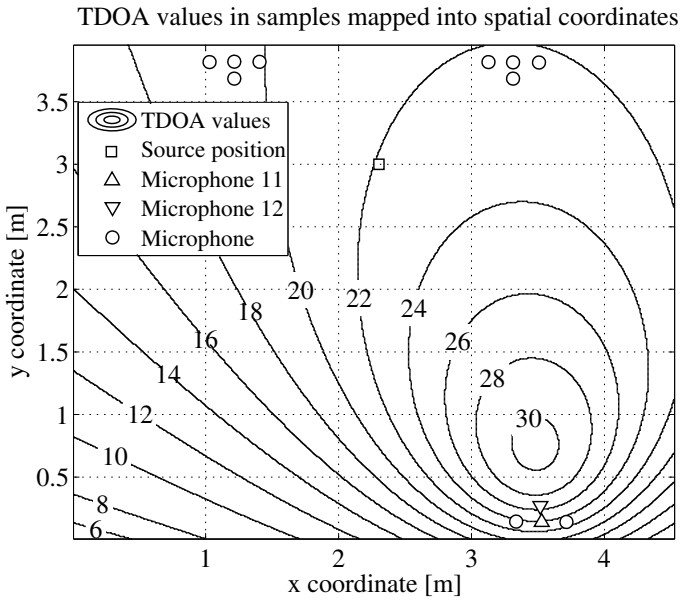



Figure 2.8: The TDOA mapping into spatial coordinates is illustrated. Spatial locations (hyperbolae) related to TDOA values between microphones 11 ("△") and 12 ("▽") are illustrated at the height of 1.128 m in a 2D slice. The real-data microphone setup of Fig. 2.2 is used. Note that the annotated source position ("□") is mapped into TDOA value 22 when sampling frequency ($f_s$) is 44 100 Hz.

a one dimensional space (time) $f : \mathbb{R}^3 \rightarrow \mathbb{R}^1$ and is therefore surjective. No function $f : \mathbb{R}^1 \rightarrow \mathbb{R}^3$ exists that maps a TDOA value into a unique location. Because TDOA values are measurements that are used in the location estimation, the source localization problem violates the second rule of Hadamard's well-posed problem's definition – the solution is not unique. This makes the TDOA-based sound source localization an ill-posed inverse problem[4].

Conversely, a single TDOA value $\Delta\tau_{p,\mathbf{r}_i}$ determines a set of possible source positions (2.13). More specifically, the set of points is a hyperbola, see Fig. 2.8 for an illustration of hyperbolae related to even TDOA values. Multiple spatially separated microphone pairs are often utilized to make the inverse problem solvable.

The microphone pairwise delay is limited by the sensor separation to possible values in range $\Delta\tau_p \in [-\tau_{\max}, \tau_{\max}]$ , where

$$\tau_{\max} = D(\mathbf{m}_l, \mathbf{m}_k) \cdot c^{-1}. \tag{2.14}$$

The unit of delay is one second.

Sometimes partitioning the set of microphones into groups or *arrays* before pairing is justified. The signal coherence between two microphones decreases

[4]The forward problem would be to derive TDOA values when source position is known. This would be a well-posed problem since the source position maps into a unique TDOA value.

when microphone distance increases [Ash05] which favors partitioning the microphones to groups with low sensor distance. Also, the computational complexity of considering all pairs is $\mathcal{O}(M^2)$ which is lower for partitioned arrays.

The sensor separation sets the limits of possible TDOA values (2.14). Selecting a too small sensor separation may lead to over-quantization of the possible TDOA values, where only a few TDOA values exist. Increasing the sampling frequency and/or using interpolation are ways of increasing the amount of available TDOA values.

### Coherence

The coherence function between two wide-sense stationary processes $\mathbf{x}_l$ and $\mathbf{x}_k$ is defined as [Car87]

$$\gamma_{l,k}(f) = \frac{G_{l,k}(f)}{\sqrt{G_{l,l}(f)G_{k,k}(f)}}, \tag{2.15}$$

where $G_{l,k}(f)$ denotes cross power density spectrum of signals $\mathbf{x}_l$ and $\mathbf{x}_k$, $G_{l,l}(f)$ and $G_{k,k}(f)$ are power density spectra of signals $l$ and $k$ respectively, and $f$ denotes frequency (Hz). The cross power density spectrum of the signals is defined as the Fourier transform of the cross correlation function:

$$G_{l,k}(f) = \int_{-\infty}^{\infty} \mathcal{R}_{l,k}(t)e^{-j2\pi ft}dt. \tag{2.16}$$

The cross correlation function is defined as

$$\mathcal{R}_{l,k}(\tau) = E[x_l(t)\, x_k(t-\tau)], \tag{2.17}$$

where $E[\cdot]$ is the expectation value [Car87][5]. The real valued magnitude squared coherence function is written

$$C_{l,k}(f) \triangleq |\gamma_{l,k}(f)|^2, \tag{2.18}$$

where $|\cdot|$ notes absolute value, and

$$0 \leq C_{l,k}(f) \leq 1.$$

In practice, the source is often non-stationary and the cross power density spectrum is estimated from a short time frame of length $L$ samples

$$\hat{G}_{l,k}(n) = X_l(n)\, X_k^*(n), \tag{2.19}$$

where $(\cdot)^*$ notes complex conjugate operation, and $X_i(n), i = l, k$ is defined as the discrete Fourier transform (DFT) of a frame of length $L$ [Che05b, Ch.7]:

$$X_i(n) = \begin{cases} \sum_{t=0}^{L-1} x_i(t)e^{-j2\pi tn/L}, & 0 \leq t \leq L-1 \\ 0, & \text{otherwise} \end{cases}, \tag{2.20}$$

[5]Refer to Appendix D for concepts related to random variables.

where $n$ is a discrete frequency index, and $t$ is discrete time index. When a rectangular window is applied on a frame of data the signal is multiplied with ones inside the frame and zeros outside the frame. In the frequency domain the multiplication is equal to convolution. Therefore, the true spectrum is convolved with the window function in the frequency domain. The window function in the frequency domain consists of a mainlobe and sidelobes. As a result, true spectral peaks may be concealed and false peaks may be generated, which is referred as spectral leakage. A suitable window function may be chosen to have minimal sidelobe levels but this widens the mainlobe causing it to spread into adjacent frequencies. This in turn, results in smearing [Ife93]. Therefore, the window function should be chosen with care. A common choice is the Hanning window.

## 2.6 TDOA Estimation Methods

In this section selected time delay estimation methods are described, including the average magnitude difference function (AMDF) and the widely known generalized cross correlation (GCC). Theory regarding the behavior of the correlation-based TDOA estimation is discussed. In addition, other TDE methods are briefly mentioned. See [Che06] for an overview of TDE methods.

### 2.6.1 Generalized Cross Correlation

The relation between the cross correlation and the cross power spectral density for microphone signal pair $p = \{l, k\}$ is utilized in the generalized cross correlation (GCC) defined by Knapp and Carter [Kna76]

$$\mathcal{R}_p(\tau) = \int_{-\infty}^{\infty} \Psi_p(f) G_{l,k}(f) e^{j2\pi\tau f} df, \tag{2.21}$$

where $\Psi_p(f)$ is a frequency weighting function, and $f$ is frequency. The weighting can be used to incorporate knowledge of the signal and noise statistics. A common weighting function is the PHAT weighting, defined by Knapp and Carter:

$$\Psi_p(f) = \frac{1}{|G_{l,k}(f)|}. \tag{2.22}$$

The PHAT weighting causes unity amplitude for all frequencies of the output. Other weighting schemes include the Roth, SCOT, Eckart, the Hannan-Thomson (maximum likelihood) [Kna76] and the Hassab-Boucher method [Has81]. The maximum likelihood (ML) weighting can be written [Car87]

$$\Psi_p(f)^{\mathrm{ML}} = \frac{1}{|G_{l,k}(f)|} \cdot \frac{C_{l,k}(f)}{[1 - C_{l,k}(f)]}, \tag{2.23}$$

where $C_{l,k}(f)$ is the magnitude squared coherence function (2.18), and $G_{l,k}(f)$ is the cross power spectral density. Substituting the estimated values for these parameters obtained from input signals $\mathbf{x}_l$ and $\mathbf{x}_k$ is a heuristic procedure [Kna76]. Therefore, the ML estimator can only be approximated in practice.

Note that different realizations of the ML TDOA estimator exist, e.g., in [Bra99] the ML estimator weighting has been realized as

$$\hat{\Psi}_p(f)^{\mathrm{ML}} = \frac{|X_k(f)||X_l(f)|}{|N_k(f)|^2|X_l(f)|^2 + |N_l(f)|^2|X_k(f)|^2}, \tag{2.24}$$

where noise power spectra $|N_l(f)|^2$ and $|N_k(f)|^2$ are assumed to be available or estimated from silent segments.

A GCC-based TDOA measurement parametrizes the GCC function by its peak location to estimate the TDOA value

$$\hat{\tau}_p = \arg\max_{\tau} \mathcal{R}_p(\tau), \tag{2.25}$$

where the TDOA value $\hat{\tau}_p$ is the lag sample value associated to the maximum point of the correlation from input signal pair $p$.

### 2.6.2 Average Magnitude Difference Function

Average magnitude difference function (AMDF) for signals from a microphone pair $p$ is defined as [Ros74, Che05a]

$$\mathcal{R}_p^{\mathrm{AMDF}}(\tau) = \frac{1}{L}\sum_{t=0}^{L-1} |x_l(t) - x_k(t+\tau)|, \tag{2.26}$$

where $L$ is the frame length. In [Ros74] AMDF is used as a variation of autocorrelation in pitch detection. Similarly the average magnitude sum function (AMSF) is defined [Che05a]

$$\mathcal{R}_p^{\mathrm{AMSF}}(\tau) = \frac{1}{L}\sum_{t=0}^{L-1} |x_l(t) + x_k(t+\tau)|. \tag{2.27}$$

In [Che05a] (2.26) and (2.27) are combined in to yield the modified average magnitude difference function (MAMDF). In [Che05a] it is shown that AMDF and AMSF do not correlate, and therefore they contain supplementary information that is combined by MAMDF

$$\mathcal{R}_p^{\mathrm{MAMDF}}(\tau) = \frac{\mathcal{R}_p^{\mathrm{AMDF}}(\tau)}{\mathcal{R}_p^{\mathrm{AMSF}}(\tau) + \epsilon}, \tag{2.28}$$

where $\epsilon > 0$ is a small number. The scalar TDOA estimate is selected by minimizing the MAMDF function

$$\hat{\tau} = \arg\min_{\tau} \mathcal{R}_p^{\mathrm{MAMDF}}(\tau). \tag{2.29}$$

The average squared difference function (ASDF) [Jac93] is similar to (2.26), with the exception that the difference is squared. Note that signal prewhitening improves the AMDF estimator performance [Che05a].

### 2.6.3 TDE Function

The MAMDF and GCC-based TDOA estimators parametrize the TDE function by its peak location[6]. This parametrization is often applied by closed-form localization algorithms. Such algorithms, discussed in Section 3.1, are useful when fast analytic solutions are required due to computational restrictions or communication bandwidth limitations. If two (spatially separated) sound sources exist simultaneously the TDE likelihood function $\mathcal{R}_p(\tau)$ contains two peaks if the sources are not correlated [DB03]. If these sources are to be located, finding the just TDOA peaks still leaves the question of data association, i.e., which peak corresponds to which source.

A different approach is to delay the parametrization until the final location estimate is to be derived. In this approach information is retained to the end. This approach is known as the Marr's principle of least commitment and has been applied for correlation-based sound source localization [Bir01]. Section 3.5 discusses near-field localization and Section 4.5 discusses far-field localization methods that utilize this approach, also referred as the *TDE likelihood*-based approach.

An example TDE likelihood function measurement (2.21) is illustrated in Fig. 2.9. The TDOA measurement value $\hat{\tau}$ that maximizes the displayed PHAT weighted GCC function is at sample 22. The microphone and speaker layout are described in Fig. 2.8, where it is seen that this TDOA value corresponds to true source position.

### 2.6.4 Adaptive TDOA Methods

Other TDOA estimation methods include the adaptive eigenvalue decomposition algorithm (AED), which has been proposed for reverberant environments [Ben00]. The method estimates the TDOA values from the source-microphone impulse responses that are the eigenvectors of the signal covariance matrix. The method has been extended to noisy reverberant spaces in [Doc03]. Also, the response between two microphone channels can be modeled as a FIR filter. A method for finding the coefficients of such a filter adaptively has been presented in [Ree81]. In an ideal situation the channels differ only temporally. Therefore the filter response is a Dirac's delta function and the peak is located at the time difference value between channels (in samples).

[6]Note that MAMDF exhibits a minimum at true delay.

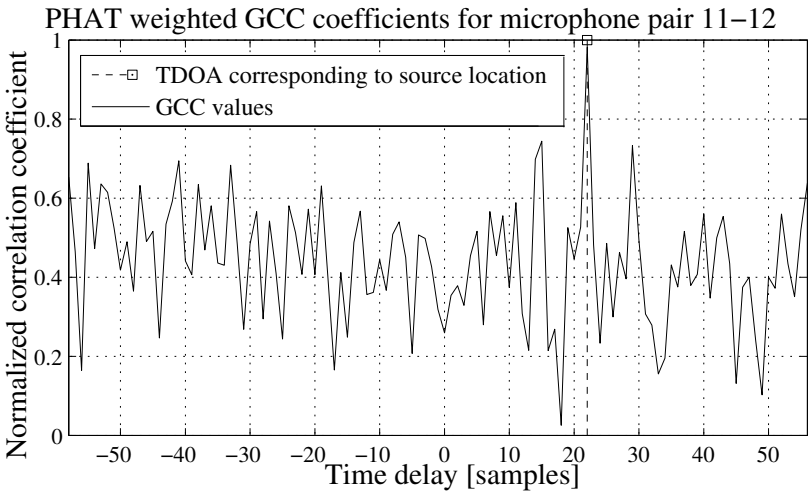



Figure 2.9: An example TDE function is displayed. Specifically, a PHAT-weighted GCC measurement is displayed from a signal pair (scaled between [0,1]). The other signal is displayed in Fig. 2.5. Note that the strongest peak in the GCC function corresponds to the TDOA value that the true source position gets mapped into, evident in Fig. 2.8.

The adaptive filter approach has also been applied for PHAT-based TDOA estimation in [You84] and for the maximum likelihood TDOA estimation [You86, Hah06].

Note that the TDOA methods mentioned above differ from the AMDF and GCC-based methods, since they do not directly provide a TDE likelihood function for different delays. In [War03] the AED was considered as a TDE likelihood function by fitting a Gaussian kernel over the TDOA value. However, such a TDE function contains the same information as the original AED-based TDOA estimate.

### 2.6.5 Source Model-Based TDOA Methods

Source-based modeling methods have also been applied in TDOA estimation and consequently in localization. In [Bra99] a pitch-based weighting for GCC is proposed, where the harmonic structure of speech spectrum is used in the frequency weighting procedure. In this approach the TDOA is estimated from a reduced number of frequency bins instead of the whole spectrum. In [Yeg05] time delay estimation using the voiced excitation caused by the vibrating vocal folds at the glottis is proposed. In practice, the excitation information is extracted from the Hilbert envelope of the linear prediction (LP) residual error signal. The method has been used in speaker localization [Ray05] (with the maximum likelihood TDOA-based approach, discussed in Section 3.1.9).

### 2.6.6 TDOA Interpolation

The sampling frequency quantizes the time resolution of the TDOA estimates to sample accuracy. Interpolation methods can be used to refine the TDOA estimate value. Linear, parabolic, and matched filter interpolation methods have been tested for correlation-based TDOA methods in [Lai99].

## 2.7 TDOA Estimation Bounds

A common statistical performance analysis method of an estimator is to compare its variance to the Cramér-Rao lower bound (CRLB). The CRLB is the minimum achievable variance of any unbiased estimator. If an estimator's variance attains the CRLB and is unbiased the estimator is *efficient*. The probability density function (PDF) of the data must be known and it must satisfy the regularity condition before the CRLB can be determined, see, e.g., [Kay98].

The signal model considered here is (2.8), where the channels differ temporally from each other. The signal is a realization of a Gaussian process, and the noise in each channel is additive Gaussian, if not otherwise stated.

The correlation-based TDOA is defined as the peak location of the GCC-based TDE function [Kna76]. Three distinct SNR ranges (high, low, and the transition range in between) in TDOA estimation accuracy have been identified in a non-reverberant environment [Wei83b, Wei83a]. In the high SNR range the TDOA variance attains the CRLB [Wei83b] and the correlation peak location is related to the true source TDOA value. In the low SNR range the TDE function is dominated by noise, and the peak location is evenly distributed between the TDOA value extremes. In the transition range the TDE peak becomes ambiguous and is not necessary related the correct TDOA value. TDOA estimators fail rapidly when the SNR drops into this transition SNR range [Wei83b]. This is also called the threshold SNR effect. The CRLB fails to predict the behavior at moderate values of SNR. Therefore, tighter bounds have been considered, such as the Barankin bound, Ziv-Zakai bound and the Weiss Weinstein bound. For a summary of different bounds on TDOA estimation, see [Sad06]. Here, the theory related to TDOA CRLB bounds is briefly summarized.

### 2.7.1 CRLB of TDOA Estimation

In the TDOA estimation problem the source signal SNR can be defined as [Wei83b, Sad06]

$$\mathrm{SNR}(\omega) \triangleq \frac{(S(\omega)/N_1(\omega))(S(\omega)/N_2(\omega))}{1 + S(\omega)/N_1(\omega) + S(\omega)/N_2(\omega)}, \tag{2.30}$$

where $S(\omega)$ and $N_i(\omega)$ note signal and noise spectra, and $\omega$ is frequency. The Fisher information matrix (FIM) can be written [Sad06]

$$I(\tau) = \frac{T}{2\pi} \int_{-\infty}^{\infty} \omega^2 \mathrm{SNR}(\omega) d\omega, \tag{2.31}$$

where $T$ is the frame length and the CRLB of the TDOA estimation accuracy is $\sigma_\tau^2 \geq I(\tau)^{-1}$. For wideband signals the SNR threshold region in which the correlation peak location becomes ambiguous is derived in [Wei83a]. The threshold effect assumes that ambiguities in the correlation peak location exist. Such conditions could be encountered when the spacing between receivers is large compared to the half-wavelength of the highest frequency component [Wei83a][7].

The increase in observation time, bandwidth, and SNR improve the achievable variance of the TDOA estimator. The behavior of TDOA estimator has also been verified in practice using outdoor recordings [Ash05]. The effects of imperfect signal coherence between two receivers are discussed in [Koz04].

### 2.7.2 Reverberant Systems

In a reverberant environment, such as a room, the correlation-based TDOA performance is known to rapidly decay when the reverberation time ($T_{60}$) increases [Cha96]. The CRLB of the correlation-based TDOA estimator in the reverberant case is derived in [Gus03] where PHAT weighting is shown to be optimal. In that bound, the signal to noise and reverberation ratio (SNRR) and signal frequency band affect the achievable minimum variance. The SNRR is a function of the acoustic reflection coefficient, noise variance, microphone distance from the source, and the room surface area.

The CRLB of TDOA estimator in a reverberant enclosure is written [Gus03]

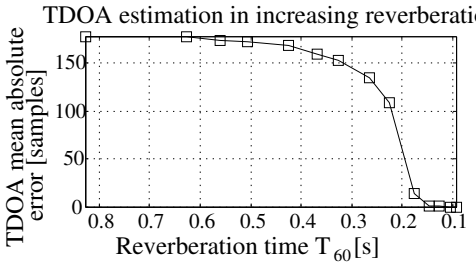
$$\sigma_\tau^2 \geq \left( 2 \sum_{k=k_l}^{k_u} \frac{\mathrm{SNRR}(\omega_k)^2}{1 + 2\,\mathrm{SNRR}(\omega_k)} \omega_k^2 \right)^{-1}, \tag{2.32}$$

where $k_l$, $k_u$ are the lower and upper discrete frequency indices of the signal bandwidth, $\omega_k$ is the angular frequency of band $k$ and

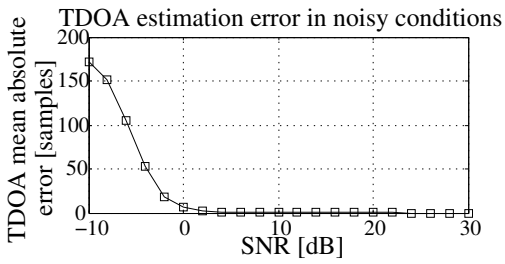
$$\mathrm{SNRR}(\omega_k) = \frac{S(\omega_k)/4\pi r^2}{S(\omega_k) 4\beta^2 / A(1-\beta^2) + \sigma_n^2}, \tag{2.33}$$

where $\sigma_n^2$ is the variance of the additive noise in (2.8), $A$ is the room surface area, $\beta$ is the reflection coefficient of the surfaces, $r$ the distance from the sensor pair midpoint to the source, and $S(\omega_k)$ is the power spectrum of $s(t)$. The bound (2.32) omits the effects of finite measurement time. It is seen from the

[7] E.g., the half wavelenght of a 4000 Hz signal is approximately 4 cm.




(a) Mean absolute error of TDOA estimation with increasing reverberation, SNR (2.12) is +30 dB.



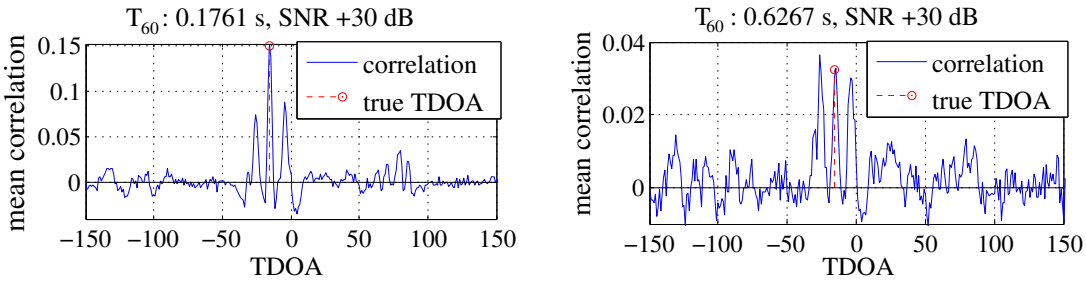

(b) Mean absolute error of TDOA estimation with decreasing SNR (2.12), $T_{60}$ is 0 s (i.e. anechoic room).

Figure 2.10: The threshold effect of TDOA estimation is demonstrated.

bound (2.32), that the increase in bandwidth and room surface area improves the TDOA performance, whereas increase in $r$, $\beta$ and $\sigma_n^2$ lowers the TDOA performance. The relationship between PHAT weighting and the maximum likelihood (ML) approach was considered in a reverberant environment in [Zha08], in which PHAT was found to be an approximate solution of the ML solution. Hereafter, the PHAT weighted GCC is utilized as the TDE weighting function since it is the optimal weighting function for a TDOA estimator in a reverberant environment [Gus03].

### 2.7.3 SNR Threshold in Simulations

The simulated data described in Section 2.4 is used to demonstrate the TDOA estimation threshold effect using reverberation and noise. A frame of length 2048 samples (46.4 ms) is used to calculate the GCC-PHAT (2.25) between all microphone pairs using all 32 microphones. The SNR values correspond to average SNR of each channel evaluated with (2.12). Figure 2.10 depicts the mean absolute error of TDOA estimates. The threshold value where average TDOA estimation error starts to increase rapidly is reached by raising the reverberation time $T_{60}$ from 0.147 to 0.176 s, see Fig. 2.10(a). Figure 2.11 further illustrates the reverberation effect on the PHAT weighted GCC function. The figure details the averaged cross correlation function between microphone pair {1,2} in a high +30 dB SNR case. Without reverberation the only GCC peak is located at the true source position. At $T_{60}$ value 0.176 s the direct path peak still dominates in the microphone pair, see 2.11(a). At larger $T_{60}$ values the peaks due to reverberation start to dominate the GCC maximum, see 2.11(b). The GCC peak locations caused by reverberation are deterministic. If the room shape and dimensions are known this fact could be utilized in localization [Kor08]. However, in this work it is not assumed that the shape and the dimensions of the room are known. In high SNR and reverberant conditions where the dimensions are not known methods for extracting the direct-path TDOA value from a set of candi-




(a) Moderate reverberation

(b) High reverberation

Figure 2.11: Simulated effect of reverberation on averaged GCC-PHAT values between microphones 1 and 2. SNR is the average SNR over all channels and is estimated with (2.12).

date TDOA values (GCC peaks) have been considered in [Sch08]. Since also low SNR conditions are of interest here, only the maximization of GCC for TDOA estimation is considered.

Figure 2.10(b) details the TDOA error as a function of average SNR (2.12). Similarly the TDOA estimation breaks after the SNR threshold value is reached (at negative values). To conclude, the TDOA error behavior in the simulated dataset corresponds to the threshold behavior.

## 2.8 Summary

The problem of time difference of arrival estimation was reviewed along with methods to solve it. The theory related to the accuracy of correlation-based TDOA estimators was reviewed for the basic free-field case and for the reverberant environment with a Gaussian signal model. The correlation-based TDOA estimator exhibits a threshold behavior with respect to the signal parameters, such as reverberation time, SNR, bandwidth, and observation time. Decreasing signal conditions leads to systematic failure of the TDOA estimator when the threshold value is reached. The signal average SNR and reverberation time were two parameters used to demonstrate the threshold phenomena.

Chapter 3

# Time Delay Estimation -Based Localization Methods

This chapter deals with the vast number of time delay estimation (TDE) -based localization methods. These methods assume a near-field scenario where the sound source emitted wavefront is modeled as a spherical wave. In contrast, Chapter 4 discusses methods that assume a far-field scenario where the source wavefront is modeled as a plane wave.

In Section 3.1 the TDOA-based closed-form solutions are grouped and the maximum likelihood solution is reviewed. The dilution of precision (DOP) is discussed in Section 3.2 and the Cramér-Rao lower bound (CRLB) is reviewed in Section 3.3. The state estimation problem and sequential TDOA-based localization methods are briefly discussed in Section 3.4.

Section 3.5 describes the state-of-the-art localization method which utilizes the TDE likelihood function. A general framework for combining the microphone pairwise TDE likelihood values is presented. The effect of the combination operator to the variance of the resulting spatial likelihood function (SLF) is presented. A widely known localization algorithm SRP-PHAT is placed in the framework. Since no obvious closed-form solutions exist for this approach some iterative solutions are reviewed for maximizing the likelihood function in Section 3.6.

Section 3.7 describes the sequential Bayesian approach for time delay estimation based localization and reviews the computationally efficient particle filtering approach. In Section 3.8 simulations are used to compare a TDOA-based method and a TDE likelihood -based method. Section 3.9 reports obtained results. Finally, Section 3.10 summarizes the discussion.

## 3.1 TDOA-Based Closed-Form Localization

In this section a class of closed-form location estimators using TDOA measurements is reviewed. These estimators are typically referred as two-step methods. The first step is to compute TDOA values between microphones, and the second step consists of finding the source location that fits the measured TDOA values. Stoica and Li [Sto06] present a brief summary of various techniques and give a naming convention to the algorithms developed over time.

In the previous chapter the TDOA estimation error threshold effect was discussed. The effect depicts the rapid failure of TDOA estimation after a threshold level of a signal parameter is reached. Such parameters include SNR and reverberation time. Therefore, TDOA-based localization methods are not expected to perform in challenging (noisy, reverberant) environments. However, the closed-form methods represent an active research field with applications outside sound source localization. In addition, some hybrid methods exist that partially utilize the closed-form solutions [Pet05b], and TDOA processing methods for reverberant environments have been studied [Sch08] so the investigation of TDOA-based localization methods is motivated.

Sometimes the TDOA localization problem is treated as the range difference of arrival problem, since time differences can be transferred into range differences by utilizing the knowledge about the speed of sound. However, the availability of sound propagation speed is assumed. Recently, a localization method that also estimates the sound propagation speed was presented [Zhe07] and can therefore be accurately termed a TDOA-based closed-form localization method.

Multiple solutions have been presented for this classical problem. This is mostly due to the non-linear relation between the values of interest (location coordinates) with respect to the measurements (TDOA values) and the large number of different applications. In [Sto06] Stoica and Li summarize some existing methods, give a unified naming convention for the methods, and point out similarities between the methods. In addition to these estimators, some methods not discussed in [Sto06] are also summarized in this section by using a similar naming convention.

In [Mil07] Militello and Buenafuente give a theoretical solution to the TDOA-based localization problem. Gillette and Silverman derived in their work a closed-form solution [Gil08][1] and presented the important extension of the method: the use of multiple reference microphones.

The range difference is defined between two receivers and a source, similarly to the time difference of arrival (2.7). For brevity, one microphone is termed the *reference microphone* and located at the origin $\mathbf{m}_0 = [0, 0, 0]^{\mathrm{T}}$. There are a total of $M+1$ microphones, and $M$ non-reference microphones locations $\mathbf{m}_k$, where $k = 1, 2, \ldots, M$. All microphone locations are assumed to be known. Now, the

[1] which was later proven to be partially previously published by [Wei08]

range difference between the unknown source position $\mathbf{r}$ and microphone $k$ and the reference microphone is written

$$\Delta\tau_k \cdot c = d_k = \|\mathbf{r} - \mathbf{m}_k\| - \|\mathbf{r}\|, \tag{3.1}$$

where $\Delta\tau_k$ can be measured (using a TDOA estimator). Note that the reference microphone $\mathbf{m}_0$ is located at the origin, and therefore is not written. Equation (3.1) is sometimes referred as the *hyperbolic equation*, since the possible source positions form a hyperbolic curve for a fixed range difference value.

From (3.1) follows

$$\|\mathbf{r} - \mathbf{m}_k\|^2 = (d_k + \|\mathbf{r}\|)^2, \tag{3.2}$$

from which

$$d_k\|\mathbf{r}\| + \mathbf{m}_k^{\mathrm{T}}\mathbf{r} = b_k, \tag{3.3}$$

where

$$b_k = \frac{\|\mathbf{m}_k\|^2 - d_k^2}{2}. \tag{3.4}$$

Following the discussion in [Sto06] the objective is to minimize the least squares (LS) criterion

$$J = \sum_{k=1}^{M} \left(d_k\|\mathbf{r}\| + \mathbf{m}_k^{\mathrm{T}}\mathbf{r} - b_k\right)^2. \tag{3.5}$$

### 3.1.1 Unconstrained LS Method

The closed-form solution of TDOA-based source localization originates from [Smi87a, Smi87b]. Here, the closed-form methods are presented using the naming conventions given in [Sto06]. A straightforward minimization to criterion (3.5) can be written by introducing the following notations

$$\mathbf{y}(\mathbf{r}) = \begin{bmatrix} R_0 \\ \mathbf{r} \end{bmatrix}, \boldsymbol{\Phi} = \begin{bmatrix} d_1 & \mathbf{m}_1^{\mathrm{T}} \\ d_2 & \mathbf{m}_2^{\mathrm{T}} \\ \vdots & \vdots \\ d_M & \mathbf{m}_M^{\mathrm{T}} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_M \end{bmatrix}, \tag{3.6}$$

where $R_0$ is the source range to reference sensor. The LS criterion (3.5) can now be written using these definitions

$$J_{\text{U-LS}} = \|\boldsymbol{\Phi}\mathbf{y}(\mathbf{r}) - \mathbf{b}\|^2. \tag{3.7}$$

and the solution can be then written [Sto06]

$$\tilde{\mathbf{y}} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{b}. \tag{3.8}$$

The (unconstrained) location estimate is the lower part of vector $\tilde{\mathbf{y}}$ and the upper part is the distance to the source

$$\hat{\mathbf{r}} = [\mathbf{0} \;\; \mathbf{I}]\,\tilde{\mathbf{y}}, \tag{3.9}$$

where $\mathbf{0}$ is a column vector of zeros, and $\mathbf{I}$ is identity matrix of size 3×3.

Note that the term *unconstrained* is given in [Sto06] to the solution, since the range estimate from reference to source position $\hat{R}_0$ is not constrained onto the location estimate vector $\hat{\mathbf{r}}$, but is essentially removed from the solution vector in (3.9). The spherical interpolation (SI) method [Smi87a, Smi87b] has been shown equal to the unconstrained method [Sto06]. Consequently, a least squares method that minimizes the criterion (3.5) is presented also in [Hua00], where the method is termed the one step least squares method (OSLS) and proven to be equal to SI. It is noted here, that using the unified naming convention of [Sto06] also the OSLS can be called the unconstrained least squares method.

### 3.1.2 Extended Unconstrained LS Method

In [Gil08] the unconstrained LS method was derived using multiple reference microphones. The methods discussed previously do not explicitly consider TDOA measurements between non-reference microphones. However, additional measurements are likely to improve the estimator variance, and should be considered. The augmented source estimate vector can be written

$$\tilde{\mathbf{y}} = (\tilde{\mathbf{\Phi}}^{\mathrm{T}}\tilde{\mathbf{\Phi}})^{-1}\tilde{\mathbf{\Phi}}^{\mathrm{T}}\mathbf{w}. \tag{3.10}$$

For example, using M+2 microphones including two reference microphones 0 and N=(M+2), the matrix $\tilde{\mathbf{\Phi}}$ is written

$$\tilde{\mathbf{\Phi}} = \begin{bmatrix} d_{10} & 0 & \mathbf{m}_1 - \mathbf{m}_0 \\ d_{20} & 0 & \mathbf{m}_2 - \mathbf{m}_0 \\ d_{30} & 0 & \mathbf{m}_3 - \mathbf{m}_0 \\ \vdots & & \\ d_{M0} & 0 & \mathbf{m}_M - \mathbf{m}_0 \\ 0 & d_{1N} & \mathbf{m}_1 - \mathbf{m}_N \\ 0 & d_{2N} & \mathbf{m}_2 - \mathbf{m}_N \\ 0 & d_{3N} & \mathbf{m}_3 - \mathbf{m}_N \\ \vdots & & \\ 0 & d_{MN} & \mathbf{m}_M - \mathbf{m}_N \end{bmatrix}, \tag{3.11}$$

where the range difference is now written explicitly between two microphones

$$d_{ij} = \|\mathbf{r} - \mathbf{m}_i\| - \|\mathbf{r} - \mathbf{m}_j\|, \tag{3.12}$$

(compare to (3.1)) and the extended source position estimate is

$$\mathbf{y}(\mathbf{r}) = \begin{bmatrix} R_0 \\ R_N \\ \mathbf{r} \end{bmatrix} \tag{3.13}$$

and the vector $\mathbf{w}$ is written

$$\mathbf{w} = [w_{10}, w_{20}, w_{30}, \ldots, w_{M0}, w_{1N}, w_{2N}, w_{3N}, \ldots, w_{MN}]^{\mathrm{T}}, \tag{3.14}$$

where

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

The final source coordinate estimate $\hat{\mathbf{r}}$ is selected from the augmented result vector $\tilde{\mathbf{y}}$ (3.10) similarly to (3.9):

$$\hat{\mathbf{r}} = [\mathbf{0}\ \mathbf{0}\ \mathbf{I}]\,\tilde{\mathbf{y}}, \tag{3.15}$$

where $\mathbf{0}$ is column vector of zeros, and $\mathbf{I}$ is identity matrix of size 3×3. All remaining microphones could be similarly used as reference microphones [Gil08].

### 3.1.3 Pre-Multiplying Method

In [Fri87] the equation (3.3) is also derived, not assuming that the reference sensor is located at $[0, 0, 0]^{\mathrm{T}}$. The equation relating source position and range-differences is written here for completeness for microphone pair $\{i, j\}$

$$2(\mathbf{m}_i - \mathbf{m}_j)^{\mathrm{T}}\mathbf{r} = \|\mathbf{m}_i\|^2 - \|\mathbf{m}_j\|^2 - d_{ij}^2 - 2(R_j d_{ij}), \tag{3.16}$$

where $d_{ij}$ is defined in (3.12). Following notations are defined

$$\mathbf{A} = \begin{bmatrix} \mathbf{m}_1 - \mathbf{m}_j \\ \mathbf{m}_2 - \mathbf{m}_j \\ \vdots \\ \mathbf{m}_M - \mathbf{m}_j \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} d_{1j} \\ d_{2j} \\ \vdots \\ d_{Mj} \end{bmatrix}, \mathbf{u} = \frac{1}{2}\begin{bmatrix} \|\mathbf{m}_1\|^2 - \|\mathbf{m}_j\|^2 - d_{1j}^2 \\ \|\mathbf{m}_2\|^2 - \|\mathbf{m}_j\|^2 - d_{2j}^2 \\ \vdots \\ \|\mathbf{m}_M\|^2 - \|\mathbf{m}_j\|^2 - d_{Mj}^2 \end{bmatrix}, \tag{3.17}$$

and the hyperbolic equations (3.16) are written in matrix form

$$\mathbf{A}\mathbf{r} = \mathbf{u} - R_j\mathbf{d}. \tag{3.18}$$

The idea is to pre-multiply (3.18) with a matrix $\mathbf{N}^{\perp}$ that has the range difference vector $\mathbf{d}$ in its nullspace, i.e. [Sto06]

$$\mathbf{N}^{\perp} = \mathbf{I} - \frac{\mathbf{d}\mathbf{d}^{\mathrm{T}}}{\mathbf{d}^{\mathrm{T}}\mathbf{d}}. \tag{3.19}$$

This would eliminate the nuisance parameter $R_j$ that appears on the right hand side. We simplify the notation here by setting the reference sensor to the origin $\mathbf{m}_j = [0, 0, 0]^{\mathrm{T}}$ and $\mathbf{u}$ becomes $\mathbf{b}$ (3.6). After multiplying (3.18) with $\mathbf{N}^{\perp}$ results in

$$\mathbf{N}^{\perp}\mathbf{A}\mathbf{r} = \mathbf{N}^{\perp}\mathbf{b} \tag{3.20}$$

$$\mathbf{N}^{\perp}(\mathbf{A}\mathbf{r} - \mathbf{b}) = \mathbf{0} \tag{3.21}$$

Now the least squares criterion can be written

$$J_{\text{PM}} = \|\mathbf{N}^{\perp}(\mathbf{A}\mathbf{r} - \mathbf{b})\|^2 \tag{3.22}$$

Note the similarity of $J_{\text{PM}}$ (3.22) and $J_{\text{U-LS}}$ (3.7). This similarity is discussed in [Sto06], where it was noticed that the pre-multiplication with the null-spaced matrix does not affect the minimization. Therefore this observation applies also to [Fri87], where the pre-multiplying approach was originally presented. In conclusion, [Fri87] results the unconstrained solution.

### 3.1.4 Constrained LS Method

In the unconstrained solution the range estimate $\hat{R}_0$ and source position estimate $\hat{\mathbf{r}}$ are independently solved. In [Hua01] a method was presented to utilize the dependence between the values. The dependence is here written as

$$\hat{R}_0 = \sqrt{\sum_{i=1}^{3} \hat{r}_i^2}, \tag{3.23}$$

where $\hat{r}_i$ is the $i$th element of the Cartesian source position vector $\hat{\mathbf{r}}$. In [Sto06] the method presented in [Hua01] is derived briefly using the approach followed here. The method is accordingly named the constrained least squares method. This constraint can be written in a matrix form [Sto06]

$$\mathbf{y}^{\text{T}}\mathbf{D}\mathbf{y} = 0,\ \mathbf{D} = \begin{bmatrix} 1 & \mathbf{0}^{\text{T}} \\ \mathbf{0} & -\mathbf{I} \end{bmatrix}, \tag{3.24}$$

where $\mathbf{D}$ is of size $3\times4$. Now the problem is to minimize the unconstrained LS criterion (3.7) together with the additional penalty function (3.24)

$$J_{\text{C-LS}}(\lambda, \mathbf{y}) = \|\mathbf{\Phi}\mathbf{y} - \mathbf{b}\|^2 + \lambda \mathbf{y}^{\text{T}}\mathbf{D}\mathbf{y}, \tag{3.25}$$

where $\lambda$ is a Lagrange multiplier, see, e.g., [Kay98]. The optimization problem is solved by first finding the maximum point of the function (3.25) by deriving with respect to $\mathbf{y}$ and setting the results to zero

$$\begin{aligned} \frac{\partial J_{\text{C-LS}}(\lambda, \mathbf{y})}{\partial \mathbf{y}} &= 0 \\ \Leftrightarrow \frac{\partial \|\mathbf{\Phi}\mathbf{y} - \mathbf{b}\|^2 + \lambda \mathbf{y}^{\text{T}}\mathbf{D}\mathbf{y}}{\partial \mathbf{y}} &= 0 \end{aligned} \tag{3.26}$$

and the source position estimate is solved

$$\hat{\mathbf{y}}(\lambda) = (\mathbf{\Phi}^{\text{T}}\mathbf{\Phi} + \lambda\mathbf{D})^{-1}\mathbf{\Phi}^{\text{T}}\mathbf{b}, \tag{3.27}$$

where the second derivative of $J_{\text{C-LS}}$, i.e., $(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}+\lambda\mathbf{D})$ should be positive semidefinite in order to minimize the function $J_{\text{C-LS}}$. The solution $\hat{\mathbf{y}}(\lambda)$ is then placed into the constraint function (3.24)

$$\hat{\mathbf{y}}(\lambda)^{\mathrm{T}}\mathbf{D}\hat{\mathbf{y}}(\lambda) = 0. \tag{3.28}$$

The problem leads to a sixth order polynomial, which is solved iteratively to find $\hat{\lambda}$. The solution $\hat{\lambda}$ is then placed into (3.27). Once again, the final 3D source position is then acquired by removing the leading source distance parameter from the estimate

$$\hat{\mathbf{r}} = [\mathbf{0}\ \ \mathbf{I}]\,\hat{\mathbf{y}}(\hat{\lambda}). \tag{3.29}$$

The term *constrained* now refers to the final source position estimate, which has been constrained to the range estimate.

### 3.1.5 Approximate LS Method

In [Sto06] a method is introduced, where the least squares criterion function is rewritten using (3.8) and (3.7)

$$J_{\text{A-LS}} = [\mathbf{y}(\mathbf{r}) - \tilde{\mathbf{y}}]^{\mathrm{T}}(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})[\mathbf{y}(\mathbf{r}) - \tilde{\mathbf{y}}]. \tag{3.30}$$

The equation $[\mathbf{y}(\mathbf{r}) - \tilde{\mathbf{y}}]$ is then linearized in the vicinity of the unconstrained location solution $\tilde{\mathbf{r}}$ (3.9) using Taylor series expansion

$$\mathbf{y}(\mathbf{r}) - \tilde{\mathbf{y}} \approx \delta + \mathbf{G}(\mathbf{r} - \tilde{\mathbf{r}}), \tag{3.31}$$

where

$$\mathbf{G} = \begin{bmatrix} \frac{\tilde{\mathbf{r}}^{\mathrm{T}}}{\|\tilde{\mathbf{r}}\|} \\ \mathbf{I} \end{bmatrix},\ \delta = \begin{bmatrix} \|\tilde{\mathbf{r}}\| - \hat{R}_0 \\ \mathbf{0} \end{bmatrix},$$

The result is used to form an approximate quadratic error criterion of (3.7) which is then minimized with a LS estimate

$$\hat{\mathbf{r}} = \tilde{\mathbf{r}} - (\mathbf{G}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{G})^{-1}(\mathbf{G}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\delta). \tag{3.32}$$

### 3.1.6 Two Step Closed-Form Weighted LS Method

In [Cha94] a two step weighted LS solution is presented. The solution is an approximate maximum likelihood (ML) solution, and utilizes a specific implementation of the ML TDOA estimator for a vector of TDOA values, assumed to be asymptotically Gaussian.

The initial LS solution is similar to (3.8) with the additional (unknown) weighting matrix $\mathbf{\Sigma}$, which is the covariance matrix of the TDOA vector. The source distance is also assumed large, i.e., range from source to each sensor is

approximately equal. The initial weighted LS solution can be written, using the previous notations[2]

$$\hat{\mathbf{y}} \approx \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{b}. \tag{3.33}$$

The solution is still nonlinear with respect to the unknown source location. The unknown covariance matrix is approximated using perturbation theory. The relation between the range and the source location is then applied. Using the covariance estimate of the error a weighted version of the source position (constrained solution) is derived. The position values are squared, which leads to an ambiguous position solution. The method is augmented for erroneous receiver positions in [Ho04].

### 3.1.7 Weighted Constrained Least Squares Method

In [So03] a weighted and constrained least squares localization estimate is presented. This approach utilizes the covariance matrix like the method (3.33) but constrains the known variables (location and range) with a Lagrange multiplier as in [Hua01]. The approach can be written as minimizing the WLS criterion

$$(\mathbf{\Phi y} - \mathbf{b})^{\mathrm{T}}\mathbf{W}^{-1}(\mathbf{\Phi y} - \mathbf{b}), \tag{3.34}$$

where $\mathbf{W}$ is a weighting matrix

$$\mathbf{W} = \mathbf{B}\,\mathbf{\Sigma}\,\mathbf{B}^{\mathrm{T}}, \tag{3.35}$$

where $\mathbf{B}$ is a diagonal matrix of form $\mathrm{diag}(d_1^\circ, d_2^\circ, \ldots, d_M^\circ)$, where $d_k^\circ = d_k + \hat{R}_0$ and $\mathbf{\Sigma}$ is the covariance matrix of the range difference error (as in [Cha94]).

The constrained and weighted solution can be written in a similar manner as the constrained solution (3.25)

$$J_{\text{CW-LS}}(\lambda, \mathbf{y}) = (\mathbf{\Phi y} - \mathbf{b})^{\mathrm{T}}\mathbf{W}^{-1}(\mathbf{\Phi y} - \mathbf{b}) + \lambda \mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y}, \tag{3.36}$$

where the only difference compared to (3.25) is the weighting matrix $\mathbf{W}$. The solution can be written [So03] and is very similar to (3.27)

$$\hat{\mathbf{y}}(\lambda) = (\mathbf{\Phi}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{\Phi} + \lambda\mathbf{D})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{b}. \tag{3.37}$$

The unknown parameter $\lambda$ is solved from the equation (3.37) by using the singular value decomposition (SVD) approach, and $\lambda$ is a polynomial function of high order. The unconstrained solution is equal to $\lambda$ being zero. In [So03] the closest root to zero is obtained via Newton's method, from the equation using identity matrix as weighting matrix $\mathbf{W} = \mathbf{I}$. Note that in order to confirm that the obtained root value is actually the minimum of the constraint function and not the maximum, a second derivative test should be taken.

[2] In the original paper of Chan and Ho, $\mathbf{\Phi}$ and $\mathbf{b}$ are multiplied with scalar value (-1) and $\mathbf{\Phi}$ is permuted, but this does not affect the solution.

### 3.1.8 LS Solution for Source Position, Range, and Propagation Speed

In [Zhe07] a method is presented that does not require a priori knowledge of propagation speed in the medium to solve the source position from TDOA measurements. The procedure to obtain the estimator is to write (3.5) so that the range difference term is expanded as the product of propagation speed of sound in the medium and time difference between the microphone $k$ and reference microphone, i.e., $d_k = c \cdot \Delta\tau_k$. Redefining the notations (3.6) appropriately leads to

$$\mathbf{y}(\mathbf{r}) = \begin{bmatrix} cR_0 \\ c^2 \\ \mathbf{r} \end{bmatrix}, \mathbf{\Phi} = \begin{bmatrix} \tau_1 & \tau_1^2/2 & \mathbf{m}_1^{\mathrm{T}} \\ \tau_2 & \tau_2^2/2 & \mathbf{m}_2^{\mathrm{T}} \\ \vdots & \vdots & \vdots \\ \tau_M & \tau_M^2/2 & \mathbf{m}_M^{\mathrm{T}} \end{bmatrix}, \mathbf{b} = \frac{1}{2} \begin{bmatrix} \|\mathbf{m}_1\|^2 \\ \|\mathbf{m}_2\|^2 \\ \vdots \\ \|\mathbf{m}_M\|^2 \end{bmatrix}. \tag{3.38}$$

The problem is defined as $\mathbf{\Phi}\,\mathbf{y}(\mathbf{r}) = \mathbf{b}$ and solved (unconstrained) with (3.8) and by removing the first two non-coordinate variables from $\tilde{\mathbf{y}}$ similarly to (3.9).

Following the idea of constraining the result variables together [Zhe07] proceeds to do this with a three step procedure, where the first step is to estimate the unconstrained position along with the range to source $R_0$ and speed of sound $c$. The second and third steps consist of applying the TDOA weighting and applying the constraints between $\hat{\mathbf{r}}$, $\hat{R}_0$, and $\hat{c}$. The solution, however, becomes ambiguous.

### 3.1.9 TDOA Maximum Likelihood Approach

Here, it is assumed that TDOA values from $S$ microphone pairs $\mathbf{\Delta t} = \Delta t_1, \Delta t_2, \ldots, \Delta t_S$ are independent and corrupted by additive Gaussian noise with covariance matrix $\mathbf{\Sigma}$. The TDOA values given by source position parameter $\mathbf{r}$ are written $\mathbf{\Delta\tau_r} = \Delta\tau_{1,\mathbf{r}}, \Delta\tau_{2,\mathbf{r}}, \ldots, \Delta\tau_{S,\mathbf{r}}$. TDOA values are non-linearly related to the source position (2.13).

The probability density function (PDF) of the data is then parameterized by the unknown source position [Abe90, Cha94]

$$p(\mathbf{\Delta t}; \mathbf{r}) = \frac{\exp(-\frac{1}{2}[\mathbf{\Delta t} - \mathbf{\Delta\tau_r}]^{\mathrm{T}} \mathbf{\Sigma}^{-1} [\mathbf{\Delta t} - \mathbf{\Delta\tau_r}])}{(2\pi)^{(S/2)} \det(\mathbf{\Sigma})^{1/2}}, \tag{3.39}$$

where $p(\cdot)$ is probability. Taking the logarithm of (3.39) and removing constant terms results in the log-likelihood function

$$P(\mathbf{\Delta t}; \mathbf{r}) = [\mathbf{\Delta t} - \mathbf{\Delta\tau_r}]^{\mathrm{T}} \mathbf{\Sigma}^{-1} [\mathbf{\Delta t} - \mathbf{\Delta\tau_r}]. \tag{3.40}$$

Equation (3.40) is the sum of weighted squared errors between the TDOA values related to the unknown parameter $\mathbf{r}$ and the measured TDOA data values $\mathbf{\Delta t}$.

A maximum likelihood (ML) estimator minimizes (3.40) with respect to source position.

In [Sva97] a ML algorithm for TDOA-based localization is derived for a near-field microphone array. The algorithm minimizes (3.40) using Taylor series expansion in the neighborhood of the hypothetical source position. The source is located using a gradient search method. In [Ray05], (3.40) is directly minimized with the Gauss-Newton optimization method. A localization algorithm called LEMSAlg [Sil05] is derived from a practical viewpoint and includes several heuristics to improve the estimator performance, but essentially minimizes (3.40) using the Simplex optimization method [Lag98]. In [Abe90] Abel presents a divide and conquer based ML estimator as a computationally inexpensive method for a small TDOA error region.

## 3.2 Dilution of Precision

A TDOA value between a microphone pairs maps into a hyperboloid in 3D Cartesian coordinate system. The finite maximum number of available TDOA values therefore causes spatial quantization, since the number of spatial regions is limited. The maximum number of TDOA values is determined by the separation of microphones, the sampling frequency, and possibly interpolation (2.14).

In addition to spatial quantization, the microphone placement affects the accuracy of the localization result. The accuracy analysis of a global positioning system (GPS) includes the dilution of precision (DOP) concept, which is a measure of the geometrical error due to the user and satellite positions [Kuu05]. If a small change in the measurement (TDOA) causes a large change in the location estimate the DOP value is large. In a better geometrical situation a large change in the measurement causes only a minor change in the location estimate. As a result the estimation is more accurate. In [Bar99] the DOP concept was extended to TDOA localization using two microphones, one of which is a reference microphone at the origin. The equations are derived by computing the gradient of the TDOA function (2.13) with respect to the unknown source position, i.e., $\frac{\partial}{\partial \mathbf{r}} \Delta\tau_{p,\mathbf{r}}$. Here, the DOP is derived for the more general case of an arbitrary reference microphone position as follows

$$\begin{aligned} \frac{\partial}{\partial \mathbf{r}} \Delta\tau_{p,\mathbf{r}} &= \frac{\partial}{\partial \mathbf{r}} c^{-1} \left( \|\mathbf{r} - \mathbf{m}_l\| - \|\mathbf{r} - \mathbf{m}_k\| \right) \\ &= c^{-1} \left( \frac{\mathbf{r} - \mathbf{m}_l}{\|\mathbf{r} - \mathbf{m}_l\|} - \frac{\mathbf{r} - \mathbf{m}_k}{\|\mathbf{r} - \mathbf{m}_k\|} \right). \end{aligned} \tag{3.41}$$

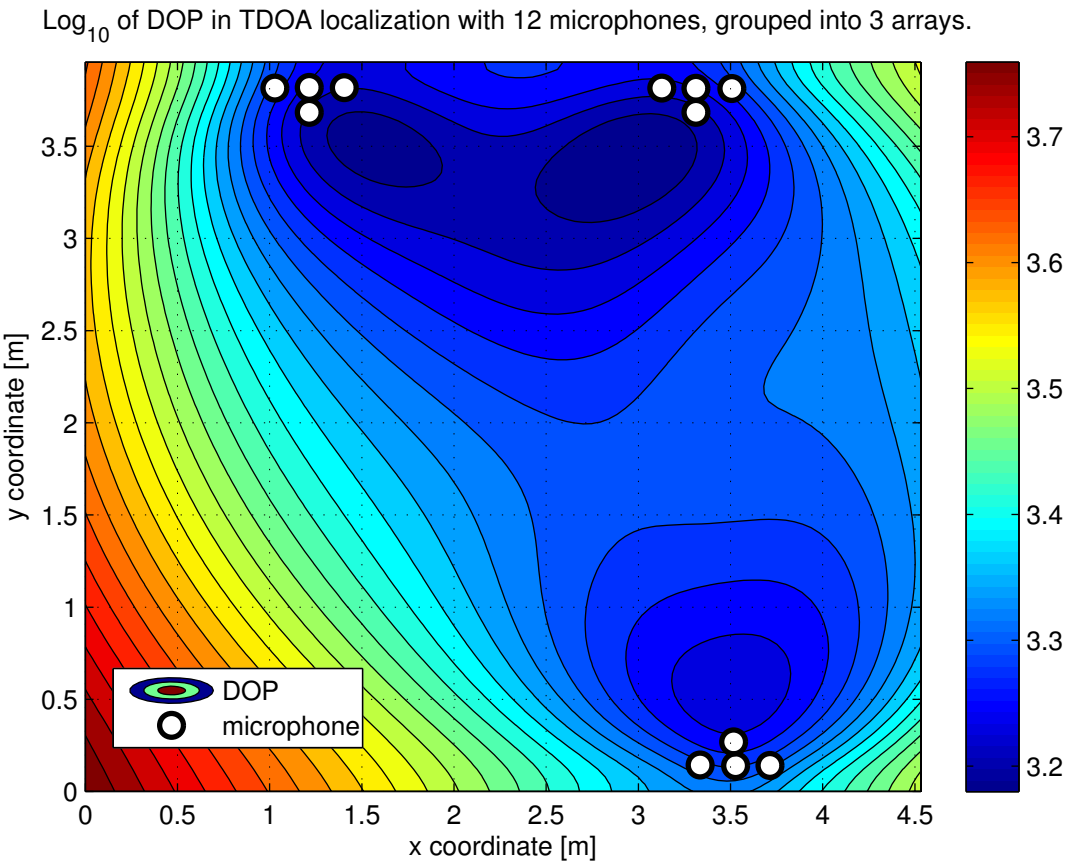



Figure 3.1: An illustration of dilution of precision (DOP) error in the practical recording room with 12 microphones grouped into three arrays. The color indicates the logarithm of DOP value.

A gradient matrix $\mathbf{H}$ is then defined as [Bar99]

$$\mathbf{H} = \begin{bmatrix} \frac{\partial}{\partial \mathbf{r}} \Delta \tau_{1,\mathbf{r}}^{\mathrm{T}} \\ \frac{\partial}{\partial \mathbf{r}} \Delta \tau_{2,\mathbf{r}}^{\mathrm{T}} \\ \vdots \\ \frac{\partial}{\partial \mathbf{r}} \Delta \tau_{S,\mathbf{r}}^{\mathrm{T}} \end{bmatrix}. \tag{3.42}$$

In the GPS discussion the matrix $\mathbf{H}$ is termed as the design matrix. Defining a matrix $\mathbf{\Xi} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}$ the DOP is defined as [Bar99]

$$\mathrm{DOP} = \sqrt{\mathrm{trace}(\mathbf{\Xi})}. \tag{3.43}$$

Using the microphone coordinates described in Table 2.1 the DOP for the recording room is illustrated in Fig. 3.1. The DOP is evaluated by considering only the microphone pairs within the three arrays and omitting inter-array microphone pairs. The DOP values are calculated for the x,y plane at the height

of 1.128 m. The region of smallest DOP values is located near the microphone arrays. The larger the DOP value is, the more sensitive the localization solution is to errors in the time delay estimates.

Note that DOP could be utilized when designing the microphone layout and placement. In [Yan05] the optimal microphone array shape was studied by minimizing (3.43) with respect to microphone placement. In the 2D case the optimal microphone placement is a uniform angular array and in 3D the Platonic solids such as tetrahedron, octahedron, cube, etc. represent optimal microphone geometries. In the example case of the recording room of Fig. 3.1 one should avoid having to locate sound sources near the origin, since the DOP value indicates that the localization is most sensitive to TDOA errors in this region (marked with red color). In [Sch06] practical ultra wide band (UWB) localization accuracy is reported to follow the theoretical values predicted by the DOP.

## 3.3 CRLB of TDOA Localization

In [Cha94] Chan and Ho derive the CRLB of the TDOA based localization assuming that TDOA values are normally distributed with covariance matrix $\mathbf{\Sigma}$. The PDF of the TDOA measurements is $p(\mathbf{\Delta t}; \mathbf{r})$ (3.39). Note that a non-linear transformation of the parameter of interest (here, location) to the measured quantity destroys the efficiency of an estimator [Kay98]. The FIM for the data (3.39) is written as [Cha94]

$$\mathbf{I}(\mathbf{r}) = E\left[\left(\frac{\partial}{\partial \mathbf{r}} \ln p(\mathbf{\Delta t}; \mathbf{r})\right)\left(\frac{\partial}{\partial \mathbf{r}} \ln p(\mathbf{\Delta t}; \mathbf{r})\right)^{\mathrm{T}}\right]_{|\mathbf{r}=\mathbf{r}_0}, \tag{3.44}$$

where the derivative function is evaluated at the true source position ($\mathbf{r}_0$). The diagonal elements of inverse of FIM give the Cramér-Rao lower bound for variance of the estimator. The partial derivative of $\ln p(\mathbf{\Delta t}; \mathbf{r})$ with respect to $\mathbf{r}$ is [Cha94]

$$\frac{\partial}{\partial \mathbf{r}} \ln p(\mathbf{\Delta t}; \mathbf{r}) = -\left(\frac{\partial \mathbf{\Delta \tau_r}}{\partial \mathbf{r}}\right)^{\mathrm{T}} \mathbf{\Sigma}^{-1} \left(\mathbf{\Delta t} - \mathbf{\Delta \tau_r}\right). \tag{3.45}$$

The FIM is then written as [Cha94]

$$\mathbf{I}(\mathbf{r}) = \left(\frac{\partial \mathbf{\Delta \tau_r}}{\partial \mathbf{r}}\right)^{\mathrm{T}} \mathbf{\Sigma}^{-1} \left(\frac{\partial \mathbf{\Delta \tau_r}}{\partial \mathbf{r}}\right). \tag{3.46}$$

Using the matrix $\mathbf{H}$ defined in (3.42) the FIM is here rewritten as

$$\mathbf{I}(\mathbf{r}) = \mathbf{H}^{\mathrm{T}} \mathbf{\Sigma}^{-1} \mathbf{H}. \tag{3.47}$$

It is noted that the DOP (3.43) is the sum of diagonal elements (variances) of $(\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}$, i.e., the covariance matrix is replaced with an identity matrix $\mathbf{\Sigma} = \mathbf{I}$. Refer to [Kap06] for a more complete discussion on GPS accuracy.

## 3.4 TDOA-Based Sequential Localization Methods

The closed-form localization methods discussed in Section 3.1 utilize the TDOA estimates from a single time frame to calculate the source position. A TDOA may be a value that is not directly related to source position via (2.13). This can happen if the TDOA is estimated from the largest TDE function value in the presence of strong background noise or reverberation, as discussed in Section 2.7. As a result the estimation of source location fails since the TDOA values are not directly related to source position.

If the sound source is active between sequential frames and its motion can be modeled, sequential estimation methods could be applied to include the measurement history into the estimation process to improve the overall performance of the localization.

### 3.4.1 State Estimation

So far, the estimation problem has been to estimate the constant parameter – the source position. Another approach is to estimate the unknown *state* of the sound source. The state estimation problem is defined through the measurement equation [Gor93]:

$$\mathbf{z}_t = \mathbf{h}_t(\mathbf{r}_t) + \mathbf{v}_t, \quad t = 0, 1, 2, \dots \quad , \tag{3.48}$$

where $\mathbf{z}_t$ is the measurement, $\mathbf{r}_t$ is the state, $\mathbf{h}_t(\cdot)$ is a known and possibly a time varying and non-linear measurement vector function that maps the state into a measurement, and $\mathbf{v_t}$ is zero mean IID white noise. The state evolution is described with a discrete difference equation

$$\mathbf{r}_{t+1} = \mathbf{f}_t(\mathbf{r}_t) + \mathbf{w}_t \quad t = 0, 1, 2, \dots \quad , \tag{3.49}$$

where $\mathbf{f}_t(\cdot)$ is a vector function, and $\mathbf{w_t}$ is IID zero mean white noise term. It is possible to estimate the marginal PDF $P(\mathbf{r}_t|\mathbf{z}_{1:t})$, i.e., the state of the source given all previous measurements. From this distribution the source state can be extracted. The distribution itself contains more information about the source state than a single point estimate. The Bayesian recursive solution to the state estimation problem is discussed in Section 3.7.

A simple scheme of incorporating temporal information is to smooth the sequential source position estimates. The classical Kalman filter (described, e.g., in [Kay98]) would suffice with a state consisting of position and velocity $\mathbf{r} = [r_x, r_y, r_z, \dot{r}_x, \dot{r}_y, \dot{r}_z]^{\mathrm{T}}$, where $\dot{r}$ denotes velocity. The Kalman filter assumes that the measurements are linearly related to the state (i.e., $\mathbf{h}_t(\cdot)$ can be written as a matrix product), the state transition is a linear process (i.e., $\mathbf{f}_t(\cdot)$ can be written as a matrix product), and the noise of the process ($\mathbf{v}$) and of the measurements ($\mathbf{w}$) are mutually independent zero mean white Gaussian. Under

these assumptions the Kalman filter is an optimal estimator of the state [Kay98]. However, the TDOA estimates are not linearly related to the source position, i.e., $\mathbf{h}_t(\cdot)$ is not linear. This renders the classical Kalman filter unsuitable for the TDOA-based measurement state estimation problem.

The Kalman filter framework can be applied to non-linear problems with the use of extended Kalman filter (EKF). In [Kle06] EKF is applied to source localization using TDOA measurements. However, the EKF is not an optimal estimator since the state and measurement equations are linearized (e.g., using Taylor series expansion). In [Gan06] a recursive form of the Gauss method is derived and compared with two different filters, namely the EKF and the unscented Kalman filter. The unscented Kalman filter avoids the linearization of $\mathbf{h}_t(\cdot)$ and $\mathbf{f}_t(\cdot)$ by estimating the PDF of the random variable from the first two moments of the non-linearly transformed set of points. These methods assume that the problem is non-linear, Gaussian, and allow a non-stationary source.

A method using TDOA estimates and their probability in a sequential Bayesian Monte Carlo framework (particle filter) was proposed in [Ver01]. In [Vog07], particle filtering -based method was used to track the speaker's angle in joint video and audio tracking system. Particle filters are discussed in Section 3.7.1.

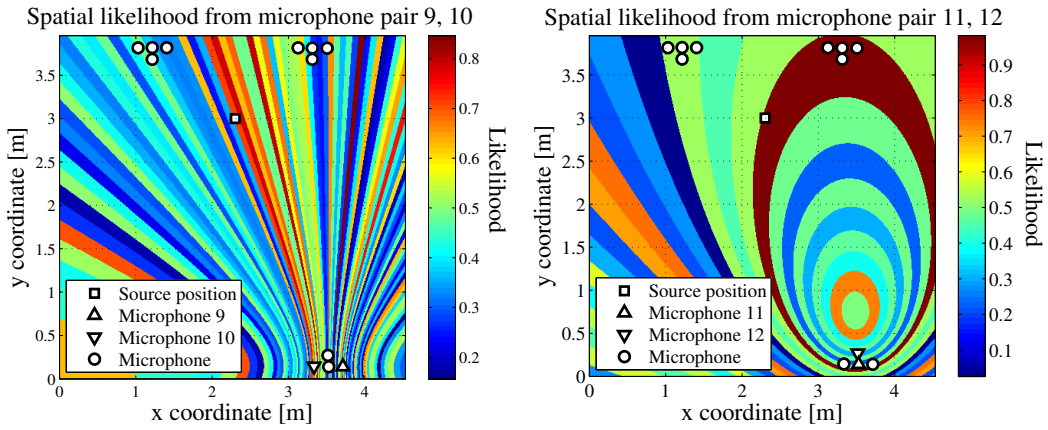
## 3.5 TDE Function -Based Localization

Based on Section 2.7 it is known that below the SNR threshold value the TDOA estimation fails. Similarly if the reverberation becomes strong enough the TDOA values are not directly related to source position anymore. In such challenging conditions the so far discussed localization methods are not applicable directly. This is problematic since such conditions are faced with reverberation times and noise levels of realistic environments.

Motivated by this and the fact that below the SNR threshold value there still exists information in the correlation-based TDE function, the TDE function -based methods are discussed.

In this approach the parametrization of the TDE likelihood function by its peak location is omitted. Instead, the TDE likelihood functions are directly combined to build a spatial likelihood function (SLF) [Aar03]. The SLF represents the likelihood of a sound source at an arbitrary location using all available time delay estimator data. In [Omo94] the term *coherence measure* was utilized for the TDE likelihood term adopted here.

As discussed in Section 2.5 selecting a hypothetical source position $\mathbf{r}$ assigns the microphone pair $p$ a TDOA value $\Delta\tau_{p,\mathbf{r}}$ according to (2.13). From a geometrical viewpoint this means that a source position is mapped into a delay value $f : \mathbb{R}^3 \rightarrow \mathbb{R}^1$, or $f : \mathbf{r} \rightarrow \Delta\tau_{p,\mathbf{r}}$. This mapping is non-injective, since a TDOA value is not inverse-mapped into a unique coordinate. The mapping is also non-




(a) Microphone pair 11-12 (b) Microphone pair 9-11

Figure 3.2: The normalized PHAT-weighted GCC function displayed in Fig. 2.9 is mapped into spatial coordinates using (2.13). The resulting map represents the microphone pairwise spatial likelihood function (SLF) and is displayed in panel 3.2(a). Panel 3.2(b) displays another mapping between neighboring microphones. Note that the two utilized microphone pairs does not give the highest likelihood for the true source location (at the height of 1.128 m).

surjective, since some TDOA values can not be mapped into any location (2.14). In conclusion, the likelihood function value of a single time delay is shared by a set of points (hyperbola) in $\mathbb{R}^3$.

Here, the correlation-based[3] TDE likelihood function (2.21) is utilized. The function indexed with the TDOA value, i.e., $\mathcal{R}_p(\Delta\tau_{p,\mathbf{r}})$, represents the likelihood of the source existing at locations that are specified by the TDOA value, e.g., hyperboloid. The signal pairwise SLF can be written as [P1]

$$P(\mathcal{R}_p|\mathbf{r}) = \mathcal{R}_p(\Delta\tau_{p,\mathbf{r}}) \in [0,1], \tag{3.50}$$

where $P(\cdot|\cdot)$ represents conditional likelihood, scaled between [0,1] in each frame. The scaling can be performed separately or for all pairs. Equation (3.50) can be interpreted as measured likelihood for a given source position $\mathbf{r}$.

An illustration of mapping a microphone pairwise TDE function into spatial coordinate is displayed in Fig. 3.2, where the TDE function values presented in Fig. 2.9 are utilized. The figure represents a two dimensional grid of the SLF at height 1.128 m and the grid cell size is 10 mm $\times$ 10 mm. The z-coordinate value is omitted for clarity. A SLF from a single microphone pair does not offer a unique solution of the source position, as seen from the figure. The source is most likely to lie somewhere on the hyperbola corresponding to the TDE function peak value (marked with dark red).

[3]It is noted, that any TDE likelihood function or a combination of functions could be applied.

A well known localization method, the steered response power using phase transform (SRP-PHAT) [DiB01a][Bra01, Ch.8], is based on the idea to add several pairwise TDE functions from different microphone pairs to reduce the SLF peak location ambiguity. A search for the maximum value can then be performed to locate the source.

More generally, the combination of pairwise SLFs can be written using a combination operator $\otimes$ [P1]:

$$P(\mathcal{R}_{[1:S]}|\mathbf{r}) = \bigotimes_{p=1}^{S} \mathcal{R}_p(\Delta\tau_{p,\mathbf{r}}), \tag{3.51}$$

where $S$ is the number of microphone pairs and $\mathcal{R}_{[1:S]}$ represents corresponding TDE functions. In [P1] rules for the operator $\otimes$ are presented. In short, the operator $\otimes$ is a binary operator combining two likelihoods, and is defined as

$$\otimes : [0,1] \times [0,1] \rightarrow [0,1]. \tag{3.52}$$

In [P1] it is suggested, that operator $\otimes$ is *commutative*, *monotonic*, *associative*, and optionally *bounded* between [0,1]. For likelihoods A,B,C, and D these rules are written as

$$\mathrm{A} \otimes \mathrm{B} = \mathrm{B} \otimes \mathrm{A}, \tag{3.53}$$

$$\mathrm{A} \otimes \mathrm{B} \leq \mathrm{C} \otimes \mathrm{D} \quad \text{if} \quad \mathrm{A} \leq \mathrm{C} \text{ and } \mathrm{B} \leq \mathrm{D}, \tag{3.54}$$

$$\mathrm{A} \otimes (\mathrm{B} \otimes \mathrm{C}) = (\mathrm{A} \otimes \mathrm{B}) \otimes \mathrm{C}. \tag{3.55}$$

Operations such as summation, multiplication, minimum, and maximum follow these rules[4].

### 3.5.1 Correlation Combination with Summation

SRP-PHAT method uses PHAT weighted cross correlation values from microphone signals. The cross correlation values are indexed with a TDOA value from a hypothetical source position [Bra01, Ch.8][Omo98] and then summed

$$P_{\text{SRP-PHAT}}(\mathcal{R}_{[1:S]}|\mathbf{r}) = \sum_{p=1}^{S} \mathcal{R}_p^{\text{GCC-PHAT}}(\Delta\tau_{p,\mathbf{r}}). \tag{3.56}$$

The position $\mathbf{r}$ that maximizes the likelihood function is thought to represent the source position. The SRP-PHAT is a special case of building the likelihood by adding PHAT weighted GCC values. The method in [Che01] sums unweighted GCC values, which is shown equivalent to the steered beamformer. Another GCC combination by summation is presented in [Val07] where precedence weighted GCC values are added together for direction finding.

[4]There rules are followed by S-norm and S-conorm operations [Jan97].

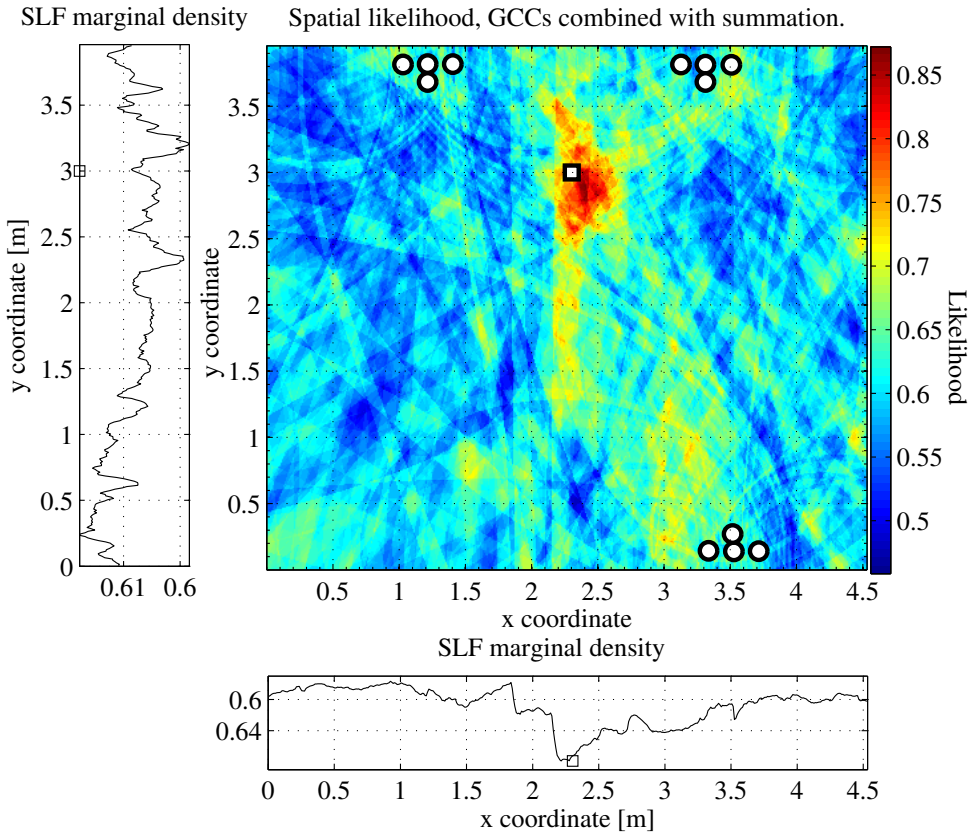



Figure 3.3: An illustration of a two dimensional spatial likelihood function (SLF), generated by adding all microphone pairwise SLFs inside each array. See Fig. 2.8 for an example of two microphone pairwise SLFs. The microphones are marked with circles "$\circ$", and the source with a square "$\square$". The left and bottom panels represent marginal densities of the SLF.

Figure 3.3 illustrates the resulting SLF function which is acquired by adding all pairwise SLFs together. The SLFs are calculated from microphone pairs within each array, marked with circles. The "tails" of the hyperbolae are clearly visible, but the peak is near the annotated location. The annotated source location is marked with "$\square$" and is measured by hand.

### 3.5.2 Correlation Combination with Multiplication

Recently, multiplication has been shown to produce more favorable localization results than the summation, when the SLF function is used by a sequential Bayesian scheme [P1]. Lehmann [Leh04] points out that the combination can be performed by multiplication if the correlation measurements are independent, although it is not clear if they are independent. If the likelihoods are independent, the intersection of sets equals their product. The modified SRP-PHAT algorithm

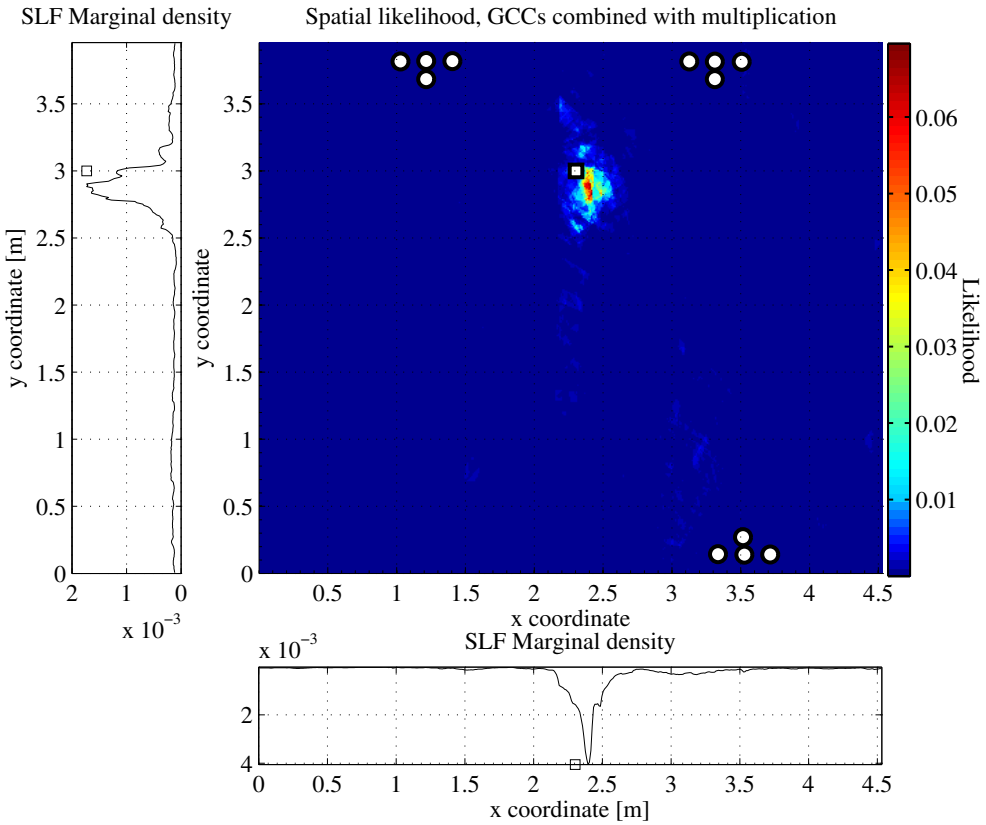



Figure 3.4: An illustration of a two dimensional spatial likelihood function (SLF), generated by multiplying all microphone pairwise SLFs inside each array. See Fig. 2.8 for an example of two microphone pairwise SLFs. The microphones are marked with circles "$\circ$", and the source with a square "$\square$". The left and bottom panels represent marginal densities of the SLF. Note that the marginal densities contain source information, although this is not guaranteed in general.

using the product can be similarly written as a likelihood function of the source position. The method, termed Multi-PHAT in [P1], multiplies the pairwise PHAT weighted GCC values together in contrast to summation [P1][Leh04],

$$P_{\text{Multi-PHAT}}(\mathcal{R}_{[1:S]}|\mathbf{r}) = \prod_{p=1}^{S} \mathcal{R}_p^{\text{GCC-PHAT}}(\Delta\tau_{p,\mathbf{r}}). \tag{3.57}$$

Similarly to the SRP-PHAT, Multi-PHAT (3.57) can be maximized to search the source position.

Equations (3.56) and (3.57) differ only in the way the microphone pairwise correlation measurement is combined. This affects the shape of the resulting spatial likelihood function. Loosely speaking, the summation operation represents the union of sets, and the multiplication represents the intersection of sets. In the context of TDE likelihood -based source localization the sets correspond to

weighted hyperbolae. Summing weighted hyperbolae leaves the non-overlapping "tails" of high likelihood from pairwise correlation measurements to the combined SLF. This can be seen in Fig. 3.3. The product, on the other hand, keeps only the information all SLFs agree to, see Fig. 3.4.

### 3.5.3 Correlation Combination with Hamacher T-norm

T-norm or triangular norms are often applied in fuzzy logic to combine two values into one. The t-norm is commutative, monotonic, and associative. The Hamacher t-norm [Jan97] is a parametrized norm, and is written for two values $a$ and $b$ as

$$h(a, b, \gamma) = \frac{ab}{\gamma + (1 - \gamma)\,(a + b - ab)}, \tag{3.58}$$

where $\gamma > 0$ is a parameter. The multiplication operation is a special case of (3.58) when $\gamma = 1$. Since the Hamacher t-norm is associative, it can be used to combine pairwise TDE function values in any pair order. The SLF can be written as [P1]

$$P_{\text{Hamacher-PHAT}}(\mathcal{R}_{[1:S]}|\mathbf{r}, \gamma) = h(\ldots h(\mathcal{R}_1(\Delta\tau_{\mathbf{r}}), \mathcal{R}_2(\Delta\tau_{\mathbf{r}}), \gamma), \ldots, \mathcal{R}_S(\Delta\tau_{\mathbf{r}}), \gamma), \tag{3.59}$$

where $\mathcal{R}_S(\Delta\tau_{\mathbf{r}})$ is short for $\mathcal{R}_S^{\text{GCC-PHAT}}(\Delta\tau_{S,\mathbf{r}})$, i.e., the PHAT weighted GCC value from the $S$th microphone pair for location $\mathbf{r}$, $S$ is the number of pairs.

### 3.5.4 Spatial Likelihood Function Variance

The SLF shape changes depending on the TDE likelihood combination operator. It is desirable that the SLF is highly concentrated near the true source position(s). This results in low bias and variance for the location estimate. Figure 3.5 displays two marginal distributions calculated with SRP-PHAT and Multi-PHAT. The likelihood functions are marginalized over time and z-axis (near source true height). The data is collected with the setup described in Section 2.3 from a 26 s dialogue between two speakers, located at the square symbols (□). Refer to [P1] for details. From the figure it is evident that likelihood is centered more to the true speaker positions in the Multi-PHAT approach than with the SRP-PHAT approach.

The simulations presented in Section 2.4 are used to verify the performance of the intersection methods over union based TDE likelihood combination methods quantitatively. For each simulation frame ($T_w = 46.4$ ms) a fixed 2D grid ($\mathbf{G}$) of cell edge length 20 mm was evaluated at true source height inside the room dimensions. A measure of the mass centered on the source position is obtained by a weighted distance error (WDE) [Kor08]

$$\text{WDE} = \frac{1}{T}\sum_{t}^{T-1} \frac{\sum_{\mathbf{p}\in\mathbf{G}} \|\mathbf{r} - \mathbf{p}\| \cdot P(\mathcal{R}_{[1:S]}^{t}|\mathbf{p})}{\sum_{\mathbf{p}\in\mathbf{G}} P(\mathcal{R}_{[1:S]}^{t}|\mathbf{p})}, \tag{3.60}$$

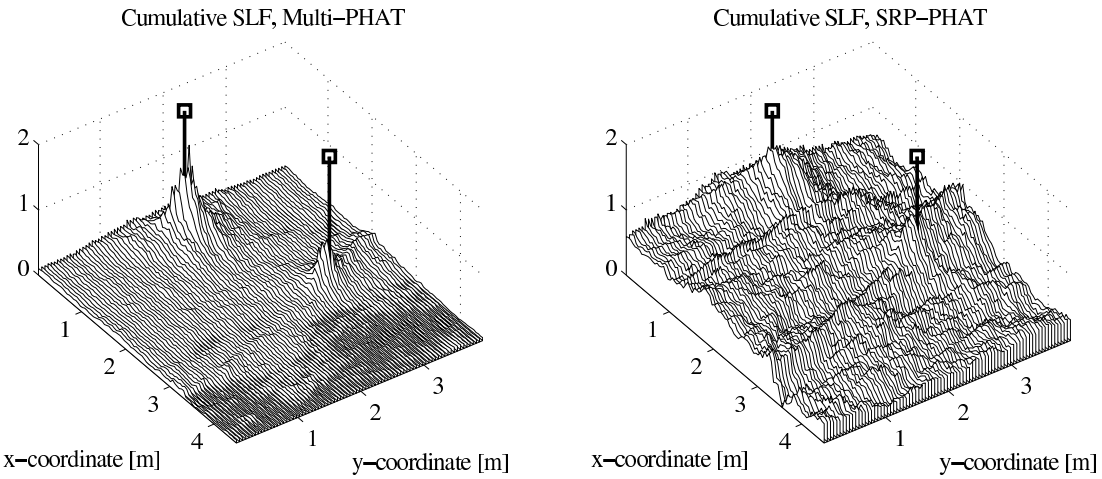



Figure 3.5: The marginal spatial likelihood functions from a real-data recording are displayed. The talker locations are marked with a square symbol ("□"). The z-axis is the marginalized spatial likelihood over the whole recorded two speaker conversation. Source [P1].

where $\mathbf{p}$ loops through the grid locations $\mathbf{G}$ for each frame $t = 0, \ldots, T-1$ and $\mathcal{R}^t_{[1:S]}$ is the measurement from time frame $t$. The WDE is the spatial likelihood value of each grid point $\mathbf{p}$ weighted by the distance from the true source position $\mathbf{r}$. A low WDE means that all likelihood mass is near the source, i.e., the variance is low. Figure 3.6 displays the WDE values for Hamacher-PHAT, Multi-PHAT, and SRP-PHAT methods in rooms with different reverberation times. Note that the SRP-PHAT has the highest WDE, meaning that a relatively large portion of the SLF is outside of the source position. This indicates a larger variance of the SLF compared to the intersection approaches. The shape of the cumulative SLF calculated from the real-data in Fig. 3.5 agrees with this conclusion. The Hamacher-PHAT ($\gamma = 0.75$) and Multi-PHAT are very close to one another, which explains the fine difference in the WDE.

More generally, the intersection of TDE likelihood values results an SLF with lower variance than the union of TDE-likelihood values. The multiplication, Hamacher T-norm, and minimum are examples of intersection operations.

### 3.5.5 TDE Likelihood Function Smoothing and Interpolation

As discussed in Section 2.6.6 the sampling frequency sets temporal quantization step of the time delay values. The temporal quantization maps into spatial coordinates as spatial quantization, visible in Fig. 2.8. In [Cir08] a method for enhancing SRP-PHAT was presented by fitting a Gaussian kernel over the selected number of peaks in TDE function. Such an approach offers a possibility

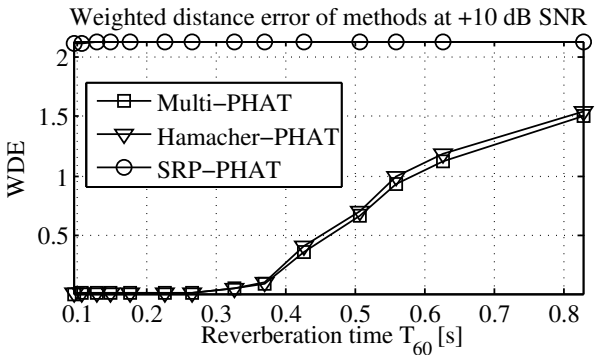



Figure 3.6: The weighted distance error (WDE) of the spatial likelihood function for different methods of combining TDE likelihoods.

for a finer source position estimate than from the quantized TDE likelihood values. Also in [Ter08] the interpolation of TDE values was studied with different interpolation methods for the SRP-PHAT algorithm. It is noted, that such interpolation methods are applicable also for other types of combination operations of TDE likelihoods.

## 3.6 TDE Likelihood-Based Localization by Iteration

A straightforward but computationally expensive approach for source localization is to exhaustively find the maximum value of the SLF

$$\hat{\mathbf{r}} = \underset{\mathbf{r}}{\operatorname{argmax}}\ P(\mathcal{R}_{[1:S]}|\mathbf{r}). \tag{3.61}$$

The SRP-PHAT is perhaps the most common way of building the SLF and therefore several algorithms, including the following ones, have been developed to reduce the computational burden. A stochastic [Ber91, Do07a] and a deterministic [Do07b] way of reducing the number of SLF evaluations have been presented. These methods iteratively reduce the search volume that contains the maximum point until the volume is small enough.

In [Dmo07] the fact that a time delay is inverse-mapped into multiple spatial coordinates was utilized to reduce the number of SLF grid evaluations by considering only the $n$ highest TDE function values and their neighboring values.

In [Zot04] the SLF is maximized initially at low frequencies that correspond to large spatial blocks. The maximum valued SLF block is selected and further divided into smaller blocks by increasing the frequency range. The process is repeated until a desired accuracy is reached. The source must therefore have energy in the lower part of the spectrum for the method to be useful. This approach has been found to be sensitive towards reverberation [Pet05a]. In [Gar07a] a related

two-pass algorithm was described, where first the SRP-PHAT is evaluated with cross correlation components below 9 kHz in 20 cm spacing. The maximum block is selected and a fine search is performed with all frequency content.

Previous examples perform the source location estimation from the current time frame only. If the spatial likelihood function has a peak that is not directly related to true source position, e.g., due to reverberation effects, the estimate will become biased.

Including a priori information into the location estimation process is possible. In [Aar03] some physically impossible sound source positions are encoded into a distribution called the spatial observability function (SOF). This distribution has a low (zero) value at the location of an improbable source position, e.g., the location of a column shaped supporting structure of a building could have minimal weight since no sound source could be inside the column. The SLF can be then multiplied with the SOF to produce a posterior distribution, which is then maximized.

## 3.7 TDE Likelihood-Based Localization with Sequential Bayesian Methods

This section discusses Bayesian methods including particle filtering used in conjunction with the spatial likelihood function (SLF) discussed in Section 3.5. The discussion is based on [Dou01, Aru02].

The key advantage of sequential Bayesian methods is that they estimate the source state $\mathbf{r}_t$ using all previous measurements. The state is here not directly measurable. Instead the TDE-likelihood functions, i.e., $\mathcal{R}^t_{[1:S]}$ represent the measurements at time frame $t$. The SLF therefore acts as the noisy measurement distribution $P(\mathcal{R}^t_{[1:S]}|\mathbf{r}_t)$. In comparison, the traditional approach to maximize the SLF is memoryless and therefore spurious peaks may lead to false location estimates.

The source position is estimated from the posterior distribution $P(\mathbf{r}_{0:t}|\mathcal{R}^{1:t}_{[1:S]})$. The distribution includes all the previous measurements and state information. The initial state $\mathbf{r}_0$ represents a priori information. The first measurement is available at time frame $t = 1$. The posterior distribution is given by Bayes' theorem as [Dou01]

$$P(\mathbf{r}_{0:t}|\mathcal{R}^{1:t}_{[1:S]}) = \frac{P(\mathcal{R}^{1:t}_{[1:S]}|\mathbf{r}_{0:t})P(\mathbf{r}_{0:t})}{\int P(\mathcal{R}^{1:t}_{[1:S]}|\mathbf{r}_{0:t})P(\mathbf{r}_{0:t})d\mathbf{r}_{0:t}}, \tag{3.62}$$

where the nominator is a normalizing constant and $P(\mathbf{r}_{0:t})$ denotes a priori distribution. It is assumed that the hidden states $\mathbf{r}_{0:t}$ can be modeled as a (1st order) Markov process, i.e., all information is contained in the current state $\mathbf{r}_t$, so that the next state depends only on the previous state $P(\mathbf{r}_t|\mathbf{r}_{0:t-1}) = P(\mathbf{r}_t|\mathbf{r}_{t-1})$. It is

possible to estimate the posterior distribution with a recursive formula of (3.62):

$$P(\mathbf{r}_{0:t+1}|\mathcal{R}_{[1:S]}^{1:t+1}) = P(\mathbf{r}_{0:t}|\mathcal{R}_{[1:S]}^{1:t}) \frac{P(\mathcal{R}_{[1:S]}^{t+1}|\mathbf{r}_{t+1}) P(\mathbf{r}_{t+1}|\mathbf{r}_t)}{P(\mathcal{R}_{[1:S]}^{t+1}|\mathcal{R}_{[1:S]}^{1:t})}, \tag{3.63}$$

where $P(\mathbf{r}_{t+1}|\mathbf{r}_t)$ is a transition distribution describing source state evolution in time (also sometimes referred as source motion model). The marginal distribution (also known as the filtering distribution) $P(\mathbf{r}_t|\mathcal{R}_{[1:S]}^{1:t})$ is here of interest. It is from this distribution the source position is derived.

It is possible to estimate the marginal posterior distribution in a recursive manner [Dou01]. This can be done in two steps, termed prediction and update. The prediction of the state distribution is calculated by convolving the posterior distribution with a transition distribution $P(\mathbf{r}_t|\mathbf{r}_{t-1})$ resulting in

$$P(\mathbf{r}_t|\mathcal{R}_{[1:S]}^{1:t-1}) = \int P(\mathbf{r}_t|\mathbf{r}_{t-1}) P(\mathbf{r}_{t-1}|\mathcal{R}_{[1:S]}^{1:t-1}) d\mathbf{r}_{t-1}. \tag{3.64}$$

The new measured SLF, i.e., $P(\mathcal{R}_{[1:S]}^{t}|\mathbf{r}_t)$ is used to correct the prediction distribution:

$$P(\mathbf{r}_t|\mathcal{R}_{[1:S]}^{1:t}) = \frac{P(\mathcal{R}_{[1:S]}^{t}|\mathbf{r}_t) P(\mathbf{r}_t|\mathcal{R}_{[1:S]}^{1:t-1})}{\int P(\mathcal{R}_{[1:S]}^{t}|\mathbf{r}_t) P(\mathbf{r}_t|\mathcal{R}_{[1:S]}^{1:t-1}) d\mathbf{r}_t}, \tag{3.65}$$

where the nominator is a normalizing constant. For each time frame $t$ the two steps, (3.64) and (3.65), are repeated.

The assumption that the previous state is related to the next state (Markov process) is violated if the source (e.g., speaker) becomes silent or another speaker becomes active. In the former case a voice activity detection (VAD) system could be utilized to estimate non-speech frames. In [Leh07] an example system of combining VAD and particle filtering (PF) is discussed.

A straightforward method for source localization is to calculate the prediction and update distributions numerically using a fixed grid [Per07]. This approach requires a large number computations. Therefore, approximations of the integrals have been used to lessen the computational burden. The particle filter is a suitable tool for this purpose, and is described in the following section.

### 3.7.1 Particle Filtering

This section briefly describes the bootstrap method, also known as particle filter (PF), or condensation algorithm. The particle filter has been widely applied in ASL [War03, Leh03, Leh04, Ver01, Val07],[S1],[P1],[P5]. For a more complete presentation of particle filtering methods refer to [Dou01, Aru02, Dju03, Mas02, Can07].

Particle filtering is used to numerically estimate the integrals (3.64)–(3.65). The PF approximates the posterior density with a set of $N_j$ weighted random

samples $\mathcal{X}_t = \{\mathbf{r}_t^j, w_t^j\}_{j=1}^{N_j}$ for each frame $t$ as

$$P(\mathbf{r}_{0:t}|\mathcal{R}_{[1:S]}^{1:t}) \approx \sum_{j=1}^{N_j} w_t^j \delta(\mathbf{r}_{0:t} - \mathbf{r}_{0:t}^j), \tag{3.66}$$

where the scalar weights $w_t^{1,\ldots,N_j}$ sum to unity and $\delta(\cdot)$ is the Dirac's delta function.

Sampling importance resampling (SIR) is an easy to implement particle filter, described as Alg. 1 (Appendix A, p. 141). The algorithm propagates the particles according to the motion model. The motion model should be selected according to the knowledge of the source movement. In [S1] and [P1] a pseudo-stationary speaker is modeled as a Brownian process and in [Ver01] a moving speaker is modeled with Langevin process.

The resampling is applied to avoid the degeneracy problem, where all but one particle have insignificant weight. In the resampling step particles with small weights are replaced with particles of larger weight. For example, the systematic resampling can be applied. It has a favorable resampling quality and low computational complexity [Hol06]. The resampling algorithm is described as Alg. 2 (Appendix A, p. 142).

After estimating the posterior distribution, a point estimate is selected to represent the source position. Point estimation methods include the maximum a posteriori, the conditional expectation

$$\hat{\mathbf{r}}_t^{\mathrm{CE}} = \sum_{j=1}^{N_j} \mathbf{r}_t^j w_t^j, \tag{3.67}$$

and the median particle

$$\hat{\mathbf{r}}_t^{\mathrm{ME}} = \mathrm{median}\{\mathbf{r}_t^1, \mathbf{r}_t^2, \ldots, \mathbf{r}_t^{N_j}\}, \tag{3.68}$$

which is used as the source position estimate in this work. The conditional expectation can results in a biased location estimate if the SLF has multiple peaks, since the mean of the likelihood mass can reside outside of the SLF maxima. The median particle is more robust than the maximum valued particle, since the maximum weighted particle can be easily corrupter by noise.

In the PF framework the changing of the speaker is briefly discussed in [P1], where some of the particles are uniformly spaced in the state space to notice the speaker change. In [Leh06] the concept of importance sampling (IS) is used in the PF framework to include information from the current measurement when redistributing the particles in the resampling step. The use of IS was reported to alleviate the detection of new targets and speaker changes.

## 3.8 Simulations

Simulations presented in Section 2.4 are used to compare the localization performance of a TDOA-based estimation method with a TDE-likelihood based method. Also, the simulations and results presented in [P1] are reviewed.

### 3.8.1 Scoring Metrics

Similar localization estimate scoring metrics to those used in the CLEAR'06 and CLEAR'07 evaluations are applied here [Mos06, Sti08]. The metrics are selected from a larger set of metrics, and are suitable for evaluating a single active source localization method. The metrics are:

- Average Estimate Error (AEE) [mm]: Euclidean mean error of non-missed estimates.
- Miss ratio [%]: Ratio of missed estimates to total number of active sound source annotations.

The AEE metric is defined as an average Euclidean distance for correctly located sources, i.e., points that are closer than 50 cm to ground truth. A point outside the threshold value is classified as a miss. Only frames where a sound source is active are considered.

### 3.8.2 Localization Methods

Two localization methods are compared. The first method is based on TDOA estimates and the second method utilizes the whole GCC function (TDE likelihood -based). Both methods first calculate the microphone pairwise PHAT weighted GCCs. All 496 microphone pairs are used. The largest microphone distance is 5.4 m and sound travels this distance in 15.6 ms. To provide a large overlapping portion of the source signal in each microphone frame the frame length is set to 2048 samples, which corresponds to 46 ms[5]. Microphone positions, room dimensions, and speed of sound are assumed static and known. The localization methods are:

1. Method 1: Iterative ML-TDOA. First, the TDOA value for each microphone pair is extracted from the peak location of the PHAT weighted GCC function of the current frame of data (2.25). Then a rough location estimate is calculated with the extended least squares TDOA-based localization method (3.10). The maximum number of reference microphones (31) is used instead of two as detailed in Section 3.1.2. The final source

[5] An alternative approach is to divide the microphones into groups or to utilize the propagation delay, discussed in Chapter 4

location estimate is obtained by inserting the rough estimate into the ML estimator (3.40) and the solution is calculated iteratively using the Simplex optimization with Matlab's `fminsearch` function[6]. If the rough estimate is outside of room bounds the starting point of the search is randomly selected from a uniform distribution inside the room and between heights of 0.5 to 1.5 m.

2. Method 2: Multi-PHAT+PF. The second localization method is a TDE likelihood -based method: Multi-PHAT with particle filtering. It combines the pairwise TDE likelihoods (PHAT weighted GCC function values) using multiplication to build the spatial likelihood function (3.57). The SLF is estimated with the SIR particle filter described as Algs. 1 and 2, see Appendix A, p. 141. The number of particles ($N_j$) is set to 1000. The motion model was set to Brownian motion, i.e., Alg. 1 line 3 distributes particles according to Gaussian distribution with standard deviation of 0.05[7]. Particles outside the room bounds were randomly distributed into the room. Additionally the possible height of particles was bounded between 0.5 – 1.5 m. In each frame at least 1 % of particles were randomly distributed inside the room to recover from possible convergence to false modes.

For each reverberation value $T_{60}$ the average SNR values (2.12) were varied from +30 dB to -10 dB in steps of 2 dB. White Gaussian IID noise was added to each channel until the desired average SNR was reached. A total of 294 simulations were generated. Each simulation run was repeated 50 times.

### 3.8.3 Simulation Results and Discussion

Table 3.1 presents the average percentage of missed estimates of method 1. The number of missed estimates is near zero at low reverberation time $T_{60}$ values and high SNR values. The increase of $T_{60}$ and the decrease of SNR cause the failure of the TDOA estimation as discussed in Section 2.7.3. As a result the localization estimation deteriorates.

In contrast Table 3.2 presents the percentage of missed estimates of method 2. It is noticed that the miss percentage is lower at higher $T_{60}$ and smaller SNR values compared to method 1. The method 2 therefore results in superior localization performance when the environment becomes more challenging in the form of reverberation and background noise.

Table C.1 (Appendix C, p. 148) displays the accuracy of method 1. The method is accurate in the region with low miss percentages. Similarly Table C.2 (Appendix C, p. 149) gives the accuracy of the method 2. It is noted that the accuracy is not as good as with method 1 in the high SNR and low $T_{60}$

[6]Optimization Toolbox ver. 3.1.1, Matlab ver. 7.4

[7]The standard deviation value was empirically selected from values 0.5, 0.05, and 0.005.

region. The best average accuracy is 12 mm, whereas method 1 achieves best average accuracy of 1 mm. Factors contributing to the inaccuracy of method 2 are that the particle filter takes some time to converge to the true source location. This *burn in period* lasts a few frames and adds to the inaccuracy of the method. Also the PF randomly distributes the particles in the prediction step according to the motion model, causing variance to the point estimate. Large motion model variance increases localization estimate variance and a small motion model variance leads to loss of diversity between particles, i.e., all particles sample the same value.

In [Pet05a] the SRP-PHAT was numerically evaluated by grid evaluation to be more robust than the TDOA based Spherical Interpolation (SI) method, which represents an unconstrained least squares approach discussed in Section 3.1.1.

Here, the Iterative ML-TDOA localization method was numerically evaluated to be less robust than the Multi-PHAT+PF. Although the result was expected the simulations provide insight into the performance difference of TDOA-based and TDE-likelihood based localization methods in different SNR and reverberation conditions. It is concluded, that in a low reverberant environment with good SNR conditions the TDOA-based localization methods lead to more accurate localization estimate than the SLF function evaluated with a particle filter. However, the TDOA-based methods suffer from the threshold effect. Therefore, SLF evaluated with PF is a more robust scheme.

Table 3.1: Percentage of missed estimates in the simulations with method 1 for SNR values 30 to -10 dB with T60 values between 0 and 0.9 s.

| | **Reverberation time $T_{60}$** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNR** | **0.09** | **0.11** | **0.13** | **0.15** | **0.18** | **0.23** | **0.27** | **0.33** | **0.37** | **0.43** | **0.51** | **0.56** | **0.63** | **0.83** |
| **30** | 0.0 | 0.0 | 0.1 | 10.8 | 88.6 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **28** | 0.0 | 0.0 | 0.1 | 20.6 | 94.4 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **26** | 0.0 | 0.0 | 0.7 | 34.4 | 97.5 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **24** | 0.0 | 0.0 | 1.7 | 52.5 | 99.3 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **22** | 0.0 | 0.0 | 6.1 | 71.6 | 99.9 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **20** | 0.0 | 0.0 | 12.7 | 84.7 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **18** | 0.0 | 0.0 | 28.7 | 93.8 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **16** | 0.0 | 0.0 | 51.6 | 98.5 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **14** | 0.0 | 0.6 | 74.7 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **12** | 0.1 | 3.2 | 90.6 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **10** | 0.5 | 11.8 | 98.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **8** | 3.3 | 29.0 | 99.9 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **6** | 13.4 | 59.8 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **4** | 33.4 | 87.8 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **2** | 67.0 | 98.5 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **0** | 92.6 | 99.9 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **-2** | 99.1 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **-4** | 99.7 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **-6** | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **-8** | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **-10** | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

### 3.8.4 TDE Likelihood Combination and PF

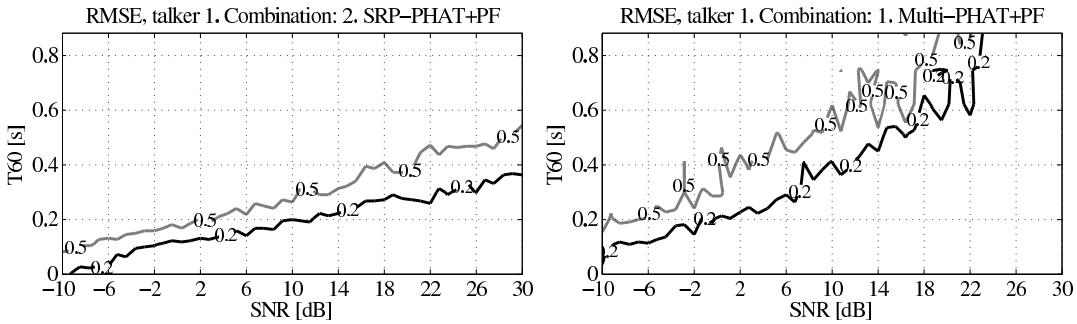
It was observed in Section 3.5.4 that the combination operation of TDE likelihoods affects the variance of the SLF. The multiplication leads to lower variance SLF than the summation. Here, the particle filter behavior on such SLF is evaluated.

In [P1] a similar simulation experiment was conducted[8] with the real-room array geometry described in Section 2.3, i.e., using only 12 microphones, grouped into three arrays listed in Table 2.1. Only microphone pairs inside each array were used, and the inter-array pairs were not considered. This enabled the use of a smaller frame length of 23.2 ms. In [P1] Multi-PHAT, SRP-PHAT, Hamacher-PHAT, and MCCC-PHAT schemes, which are all different ways of combining the microphone pairwise TDE likelihoods to build the SLF, were compared. The source state was estimated with the particle filter from the SLF, see [P1] for details. An identical particle filter scheme to above simulations was applied to estimate the source position. The mean RMS error of the methods was analyzed in similarly varying noise and reverberation conditions. As a result the Multi-PHAT and Hamacher-PHAT ($\gamma = 0.75$) outperformed the other methods. A graphical representation of the mean RMS error between the estimated source position and true source position is displayed in Fig. 3.7. Since the results of

[8] Different source signal and image source implementation was used.

Table 3.2: Percentage of missed estimates in the simulations with method 2 for SNR values 30 to -10 dB with T60 values between 0 and 0.9 s.

| | **Reverberation time $T_{60}$** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNR** | **0.09** | **0.11** | **0.13** | **0.15** | **0.18** | **0.23** | **0.27** | **0.33** | **0.37** | **0.43** | **0.51** | **0.56** | **0.63** | **0.83** |
| **30** | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.1 | 0.1 | 2.1 | 2.2 | 2.2 | 0.1 | 5.9 | 7.8 | 14.5 |
| **28** | 0.0 | 0.3 | 0.0 | 0.1 | 0.0 | 0.0 | 2.1 | 0.4 | 0.6 | 1.0 | 2.1 | 10.0 | 4.7 | 9.3 |
| **26** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 3.6 | 0.2 | 0.1 | 3.0 | 3.1 | 6.8 | 14.7 |
| **24** | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.2 | 0.6 | 2.4 | 6.2 | 4.6 | 8.1 |
| **22** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.1 | 0.1 | 0.1 | 1.4 | 3.6 | 6.3 | 16.9 |
| **20** | 0.1 | 0.0 | 0.1 | 0.0 | 0.1 | 0.4 | 0.0 | 0.2 | 1.6 | 0.2 | 8.8 | 6.7 | 12.2 | 16.8 |
| **18** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.8 | 0.1 | 2.0 | 2.6 | 4.3 | 7.7 | 23.5 |
| **16** | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 1.2 | 2.1 | 4.8 | 4.6 | 9.7 | 16.7 |
| **14** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 3.3 | 7.3 | 1.5 | 13.8 | 27.8 |
| **12** | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 1.3 | 0.0 | 0.3 | 2.3 | 1.3 | 11.8 | 13.3 | 25.1 |
| **10** | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 1.2 | 1.6 | 6.5 | 12.5 | 14.7 | 20.3 | 42.3 |
| **8** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.4 | 1.3 | 3.2 | 14.3 | 17.7 | 20.2 | 51.9 |
| **6** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.0 | 2.5 | 5.7 | 14.6 | 22.7 | 32.0 | 57.0 |
| **4** | 0.1 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 1.0 | 3.4 | 6.8 | 10.7 | 24.9 | 30.6 | 43.6 | 68.1 |
| **2** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 2.0 | 6.8 | 13.6 | 36.8 | 56.6 | 52.4 | 71.4 |
| **0** | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 1.1 | 2.6 | 5.8 | 6.6 | 29.1 | 54.6 | 46.7 | 57.0 | 82.5 |
| **-2** | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.0 | 0.1 | 5.3 | 23.6 | 28.8 | 60.0 | 74.3 | 71.7 | 95.7 |
| **-4** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.4 | 5.3 | 19.2 | 36.1 | 49.0 | 72.6 | 78.0 | 82.0 | 93.9 |
| **-6** | 0.1 | 0.0 | 0.0 | 0.1 | 0.0 | 1.7 | 14.7 | 30.6 | 45.3 | 64.5 | 86.4 | 87.1 | 90.0 | 98.6 |
| **-8** | 0.1 | 0.0 | 0.1 | 0.0 | 0.3 | 9.4 | 16.0 | 48.5 | 69.0 | 86.2 | 86.7 | 93.7 | 96.9 | 94.7 |
| **-10** | 0.0 | 0.0 | 0.1 | 0.1 | 1.5 | 14.4 | 42.7 | 64.8 | 77.3 | 88.0 | 91.0 | 93.6 | 97.3 | 95.7 |




(a) SRP-PHAT + PF

(b) Multi-PHAT + PF

Figure 3.7: The figure presents localization RMS error for talker 1 location $\mathbf{r}_1 = [2.406, 2.970, 1.118]^{\mathrm{T}}$. The signals SNR values range from -10 to 30 dB, with reverberation time T60 between 0 and 0.9 s. The contour lines represent RMS error values at steps [0.2,0.5] m. The 12 microphone array geometry detailed in Table 2.1 was used. Source [P1].

Multi-PHAT were similar to the results Hamacher-PHAT, and the results of MCCC-PHAT were similar to the results of SRP-PHAT, only Multi-PHAT and SRP-PHAT are detailed.

Based on 3.7 the Multi-PHAT using PF is more robust towards reverberation and noise than SRP-PHAT with PF. The behavior can be explained by the variance of the SLF. The SRP-PHAT uses summation which results in a larger variance in the SLF than the multiplication of values Multi-PHAT. The PF converges faster into a lower variance peak in the SLF. This leads to improved localization accuracy and robustness towards noise and reverberation.

## 3.9 Results with Speech Data

In [P1] a 26 s segment of real-data was utilized to confirm the proposed behavior of the different combination operations to the localization accuracy, observed from the simulations. The room and microphone geometry is described in Section 2.3. During the recording two speakers were engaged in a dialogue. The localization estimator mean RMS error was 0.14 m for both Multi-PHAT and Hamacher-PHAT, and the percentage of estimates inside a sphere with 25 cm radius was 92.4% and 93.1%, respectively. A PF with Brownian motion model with 5000 particles was applied with processing frame length of 23.2 ms.

### 3.9.1 CLEAR'07 Dataset Description

The CLEAR'07 dataset is a collection of multimodal data described in [Sti08]. The CLEAR'07 contains two parts: *development set* (166 min) and *evaluation*

*set* (200 min). The sets contain recordings from five separate smart rooms of varying sizes. Three to seven upside down T-shaped microphone arrays placed on room walls were used to gather data. The number of arrays depends on the specific room. The T-shaped array consists of three microphones on a horizontal line 20 cm apart. The fourth microphone was located 30 cm above the middle microphone[9]. The sampling frequency was 44.1 kHz and 24 bits per sample was used.

During the recordings multiple people were discussing in a meeting situation. Therefore, the active speaker location changed frequently and the speaker had different orientations and characteristics. The annotations of speaker locations were provided in 1 s intervals.

### 3.9.2 Results with CLEAR'07 Dataset

The length of the processing frame $T_w$ was set equal to the annotation rate (1 s) since raising the integration time improves the correlation function peak strength, as discussed in Section 2.7. In addition, 50 % overlapping was applied. This is an extremely large frame length, and can be justified by assuming that people do not move (rapidly) in meetings which is also the justification for the Brownian motion model. The localization result was not used as a front end of any other system that requires a smaller frame size, e.g., an automatic speech recognition (ASR) system.

[S1] details the CLEAR'07 evaluation dataset results with the Multi-PHAT+PF approach (3.57). A PF with Brownian motion model and 50000 particles was used to estimate the speaker position. Here, the localization system [S1] is analyzed in addition with the Hamacher-PHAT ($\gamma = 0.75$) and SRP-PHAT methods. The summary results for different TDE likelihood combination operations with particle filtering are presented in Table 3.3.

The results of the *development set* are in general better compared to the *evaluation set*. The Hamacher-PHAT+PF performs with the lowest miss ratio of 7.67 % close to Multi-PHAT+PF 8.14 % miss ratio. In contrast the SLF obtained with SRP-PHAT analyzed with particle filtering resulted in the miss of 15.99 % of estimates, double the amount compared to the other approaches tested. Considering the estimates that were not classified as misses an average accuracy of 17 cm was obtained for the intersection-based methods (Hamacher-PHAT and Multi-PHAT), and 20 cm accuracy was obtained with the union based method (SRP-PHAT).

In the *evaluation set* the results are similar, except the miss ratio is higher in all methods. The Multi-PHAT+PF has a slightly smaller miss percentage of estimates (14.96 %) than the Hamacher-PHAT+PF approach (15.25 %). The SRP-PHAT+PF now has a miss ratio of 29.50 % which is again the largest. The

[9]In one of the rooms the dimensions were 26 cm and 40 cm

average accuracy was 15 cm for the intersection methods and 20 cm for the SRP-PHAT+PF. To conclude, the real-data results agree with the simulations. In the CLEAR'07 dataset the difference in miss percentages between Hamacher-PHAT and Multi-PHAT was not found statistically significant when the calculations were repeated 50 times with 5000 particles.

## 3.10 Summary

The problem of localization using TDOA values was discussed. A review of the TDOA-based closed-form methods was presented. The maximum likelihood (ML) approach of TDOA localization was then reviewed along with several ML estimators. The estimation accuracy was considered with the known CRLB analysis and the dilution of precision (DOP) concept. The use of the measurement history to improve localization accuracy was considered with the Bayesian sequential state estimation concept. The extended Kalman filter and particle filtering are such approaches and have been previously applied for the problem. The TDOA based localization methods, especially the closed-form solutions, are computationally light, but require adequate signal conditions for the TDOA estimation to succeed.

An alternative localization approach is to combine the microphone pairwise TDE likelihoods into a spatial likelihood function (SLF). The function gives a likelihood for each source position and is traditionally maximized to locate the source. Such an approach has no obvious closed-form solutions and the SLF must be evaluated at several positions to obtain the highest likelihood position. The Bayesian sequential state estimation can also be applied to this problem. This means that the source location estimate is obtained by using all previous measurements and a model for the state transition and measurement observation. The particle filter (PF) is a computationally efficient method for sequential Bayesian

Table 3.3: The localization performance of different schemes to combine the TDE-likelihoods using particle filtering is presented with the CLEAR'07 *evaluation* and *development* datasets. Miss percent is the number of estimates more than 50 cm away from true source position divided by the number of active source annotations. Accuracy depicts the 3D mean Euclidean error of non-missed estimates.

| Dataset | *Development set* | | |
|---|---|---|---|
| Method | Hamacher-PHAT+PF | Multi-PHAT+PF | SRP-PHAT+PF |
| Miss percent [%] | 7.67 | 8.14 | 15.99 |
| Accuracy (AEE) [mm] | 166 | 165 | 198 |
| Dataset | *Evaluation set* | | |
| Method | Hamacher-PHAT+PF | Multi-PHAT+PF | SRP-PHAT+PF |
| Miss percent [%] | 15.25 | 14.96 | 29.50 |
| Accuracy (AEE) [mm] | 154 | 152 | 198 |

estimation. It represents the marginal posterior distribution of source location with a set of weighted particles. The source location estimate can be extracted from this distribution. This scheme improves the performance of localization compared to the TDOA-based localization methods.

The SRP-PHAT is a well known TDE likelihood -based localization method. It combines the TDE functions by summation and is usually maximized to estimate the source position but PF can also be applied. Different TDE combination schemes were studied and a framework for combining TDE values was presented. It was observed that by adding the TDE likelihoods the resulting SLF has a higher variance than the SLF obtained by multiplying TDE likelihoods. The PF converges faster to the lower variance distribution. Therefore, the multiplication of TDE likelihoods (referred as Multi-PHAT) results in a more accurate localization method than the SRP-PHAT when estimated with PF. This was observed from simulations with varying reverberation time and noise conditions [P1] as well as from two real-data scenarios: from a 26 s short recording [P1], and from the CLEAR'07 development and evaluation datasets. In addition the parameterized multiplication of TDE likelihoods using Hamacher T-norm was proposed. The multiplication requires that all TDE likelihoods are high in order for the output SLF value to be high. The parameterization allows some freedom to control the amount of disagreement between the measurements.

The simulated data was used to compare a ML-TDOA method with the combination of Multi-PHAT and PF. In the simulations the amount of room reverberation was controlled between 0 and 0.9 seconds. For each reverberation time the average SNR was varied from +30 dB to -10 dB. After a sufficient amount of noise and/or reverberation is added the TDE function peak is not directly related to the true source position anymore. As a result, the TDOA-based estimation fails at relatively low reverberation and in moderate noise conditions. In contrast, the Multi-PHAT using PF localization scheme can produce meaningful results in a far more reverberant and noisy environment.

Based on real-data analysis using the CLEAR'07 evaluation dataset in 85 % of speech activity the speaker was located with 15 cm average accuracy. In the CLEAR'07 development dataset the active speaker was located in 92 % of the time with 17 cm average accuracy.

Chapter 4

# Direction of Arrival -Based Localization

Source direction estimation is an important branch of signal processing. Its applications include underwater surveillance, ground vehicle tracking, and automated camera management. Multiple spatially separated sensors can be used to estimate the direction of arrival (DOA) of a plane wave. Approaches for DOA estimation include beamforming techniques [Joh93, Mor07, Vee88], TDOA-based closed form DOA estimators [YH96, YH99], TDE-based steered response methods [Joh02], and other methods such as acoustic vector sensors [Haw03]. Popular beamforming schemes include the delay-and-sum beamformer and frequency covariance estimation based approaches such as MVDR and MUSIC [Joh93].

The first part of this chapter focuses on DOA-based localization which uses the azimuth and elevation estimates of source direction to estimate source location [Blu00, Haw03, Kap01, Dom87]. The bearings-only problem is a special case that utilizes only the azimuth estimates. The latter part of this chapter discusses the use of TDE-based array steered response in localization.

The chapter outline is the following. The DOA-based localization problem and its special case the bearings-only localization are described in Section 4.1. A closed-form solution of DOA-based localization is reviewed in Section 4.2. A robust DOA-based localization is then presented in Section 4.3. The problem of limited sound propagation speed and its relation to localization is discussed in Section 4.4, where a novel DOA-based localization method is presented that models the propagation delay. Section 4.5 presents the TDE-based array steered response source localization scheme in the far-field case. Section 4.6 then augments the sound propagation speed into the array steered response based localization. The performance of the proposed estimator is presented using simulations. Section 4.7 summarizes the discussion.

## 4.1 DOA-Based Localization Problem

Let $\mathbf{p}_i$ represent the $i$th DOA sensor station location, and $i \in 1, \ldots, N_s$. The sensor station consists of multiple microphones. The station location is given as a three dimensional vector $\mathbf{p}_i = [p_i^x, p_i^y, p_i^z]^{\mathrm{T}}$ in a Cartesian coordinate system. The direction from the station to the source is also a 3D vector $\mathbf{k}_i = [k_i^x, k_i^y, k_i^z]^{\mathrm{T}}$. It is assumed, that the sensor station locations and orientations are known. The DOA vector is estimated from the impinging planar wavefront using an arbitrary DOA estimation method. Figure 4.1 illustrates the DOA-based localization problem. The estimated DOA vectors are noted as $\hat{\mathbf{k}}_i$ and are assumed of unity length, i.e., $\|\hat{\mathbf{k}}_i\| = 1$.

Additionally, errors in array geometry, location, and orientation are omitted. It is also assumed that the effects of possible source movement during the time window required by DOA estimation are insignificant. For simplicity, it is further assumed that all microphones are synchronized. The precision of DOA-based localization is assumed to depend only on the accuracy of individual DOA estimates.

### 4.1.1 Bearings-Only Source Localization

The bearings-only source localization problem is a special case of the DOA-based localization, since only one dimensional observations and 2D source location is of interest. Consider a scenario, where source bearing measurements are defined

$$\alpha_i = \mathbf{g}_i(\mathbf{r}) + w, \tag{4.1}$$

where $\mathbf{g}_i(\mathbf{r}) = \arctan((r_i^y - p_i^y)/(r_i^x - p_i^x))$ for $i = 1, \ldots, N_s$, and $w$ is IID Gaussian zero mean noise with covariance matrix $\mathbf{\Sigma}$, and $\mathbf{r}$ is source position. Refer to Fig. 4.1 for illustration. The FIM is written as [Gav92]

$$I(\mathbf{r}) = (\frac{\partial \mathbf{g}}{\partial \mathbf{r}})^{\mathrm{T}} \Sigma^{-1} (\frac{\partial \mathbf{g}}{\partial \mathbf{r}}), \tag{4.2}$$

where

$$\frac{\partial \mathbf{g}}{\partial \mathbf{r}} = \begin{bmatrix} -\frac{(r_1^y - p_1^y)}{\|\mathbf{r} - \mathbf{p}_1\|^2} & -\frac{r_2^y - p_2^y}{\|\mathbf{r} - \mathbf{p}_2\|^2} & \cdots & -\frac{r_{N_s}^y - p_{N_s}^y}{\|\mathbf{r} - \mathbf{p}_{N_s}\|^2} \\ \frac{r_1^x - p_1^x}{\|\mathbf{r} - \mathbf{p}_1\|^2} & \frac{r_2^x - p_2^x}{\|\mathbf{r} - \mathbf{p}_2\|^2} & \cdots & \frac{r_{N_s}^x - p_{N_s}^x}{\|\mathbf{r} - \mathbf{p}_{N_s}\|^2} \end{bmatrix}.$$

The minimum variance of an unbiased bearings-only location estimator is obtained from the diagonal elements of $I(\mathbf{r})^{-1}$ which is the CRLB. The ML estimator is summarized in [Gav92]. The location is, again, non-linear with respect to measurements. Therefore the ML estimator is analyzed iteratively.

The Stansfield's estimator is an approximation of the ML estimator, and is described, e.g., in [Gav92]. The Stansfield's estimator minimizes the sinusoid of

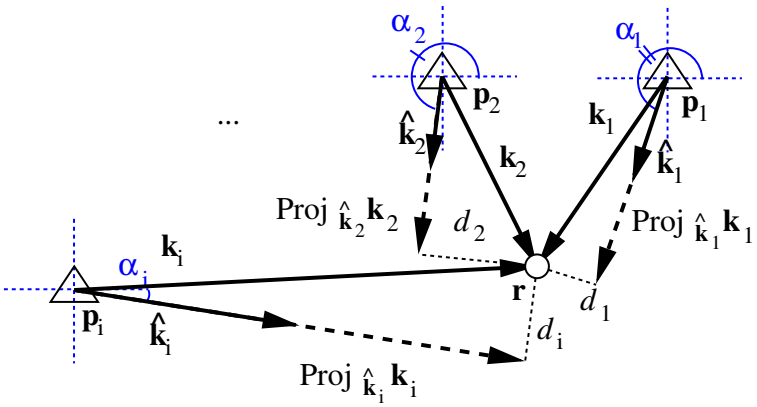



Figure 4.1: The DOA-based source localization problem is illustrated. The source position $\mathbf{r}$ is unknown and sensor station locations $\mathbf{p}_i$ are assumed known and the direction of arrival (DOA) vectors $\hat{\mathbf{k}}_i$ are measured, where $i \in 1, \ldots, N_s$. In the bearings-only localization case, the angles $\alpha_i$ are the measurements (marked with blue).

angular error and assumes that the station-to-source ranges are known. Rough range estimates can be used without significantly affecting the solution, since the cost function is a weak function of the ranges [Gav92]. The estimator can be expressed in a linear closed-form unlike the ML estimator which is a nonlinear LS minimization task. In [Tor84] a statistical method for multiple bearing measurement based localization is given.

Bearings only tracking (BOT) estimates the source state (e.g., position, heading, and velocity) based on bearing estimates made at a single or multiple stations. The single station problem is described in [Kar05] and recursive Bayesian methods are applied to estimate the source state. [Aru04] discusses also a two station problem with recursive Bayesian methods.

## 4.2 DOA-Based Closed-Form Localization

This work focuses on 3D localization, and therefore the 2D bearings-only estimators are not discussed in detail. The weighted least squares solution to the DOA-based 3D localization problem is presented in [Haw03]. The solution is derived here according to the original presentation. Let the sensor to source vector be defined as

$$\mathbf{k}_i = \mathbf{r} - \mathbf{p}_i. \tag{4.3}$$

The closest point from the line defined by the measurement vector $\hat{\mathbf{k}}_i$ to the source position is written as

$$\mathbf{p}_i + \mathrm{Proj}_{\hat{\mathbf{k}}_i} \mathbf{k}_i, \tag{4.4}$$

where the projection vector is defined as

$$\mathrm{Proj}_{\hat{\mathbf{k}}_i} \mathbf{k}_i = \hat{\mathbf{k}}_i^{\mathrm{T}} (\mathbf{r} - \mathbf{p}_i) \hat{\mathbf{k}}_i. \tag{4.5}$$

Note that in (4.5) the hypothetical source direction vector $\mathbf{k}_i$ is projected onto the actual DOA measurement vector $\hat{\mathbf{k}}_i$. The distance from the source position $\mathbf{r}$ to this closest point is

$$d_i = \|\mathbf{p}_i + \mathrm{Proj}_{\hat{\mathbf{k}}_i}\mathbf{k}_i - \mathbf{r}\|, \tag{4.6}$$

which is squared and summed to form the weighted least squares estimate

$$\hat{\mathbf{r}} = \underset{\mathbf{r}}{\operatorname{argmin}} \sum_{i=1}^{N_s} d_i^2 w_i \tag{4.7}$$

$$= \underset{\mathbf{r}}{\operatorname{argmin}} \sum_{i=1}^{N_s} \|\mathbf{p}_i + \hat{\mathbf{k}}_i^{\mathrm{T}}(\mathbf{r} - \mathbf{p}_i)\hat{\mathbf{k}}_i - \mathbf{r}\|^2 w_i, \tag{4.8}$$

where $w_i$ is a weight associated with the accuracy of each DOA estimate vector $\hat{\mathbf{k}}_i$. The localization criterion (4.6) is illustrated in Fig. 4.1. Equation (4.8) can be written as

$$\hat{\mathbf{r}} = \underset{\mathbf{r}}{\operatorname{argmin}} \sum_{i=1}^{N_s} \left(-2\mathbf{p}_i^{\mathrm{T}}(\mathbf{I} - \hat{\mathbf{k}}_i\hat{\mathbf{k}}_i^{\mathrm{T}})\mathbf{r} + \mathbf{r}^{\mathrm{T}}(\mathbf{I} - \hat{\mathbf{k}}_i\hat{\mathbf{k}}_i^{\mathrm{T}})\mathbf{r}\right) w_i, \tag{4.9}$$

where terms not related to $\mathbf{r}$ are omitted. Differentiating (4.9) with respect to the source position $\mathbf{r}$ and setting the result to zero results in

$$2\sum_{i=1}^{N_s}(\mathbf{I} - \hat{\mathbf{k}}_i\hat{\mathbf{k}}_i^{\mathrm{T}})(\mathbf{p}_i - \hat{\mathbf{r}})w_i = 0. \tag{4.10}$$

It is noted here that the second derivative should be confirmed positive semidefinite to ensure that the point is actually the minimum and not the maximum. The closed-form solution can be then written [Haw03]

$$\hat{\mathbf{r}} = \left[\left(\sum_{i=1}^{N_s} w_i\right)\mathbf{I} - \hat{\mathbf{K}}\mathbf{W}\hat{\mathbf{K}}^{\mathrm{T}}\right]^{-1}\mathbf{A}\mathbf{w}, \tag{4.11}$$

where $\mathbf{w} = [w_1, w_2, \ldots, w_{N_s}]^{\mathrm{T}}$, $\mathbf{W} = \mathrm{diag}(\mathbf{w})$, $\hat{\mathbf{K}} = [\hat{\mathbf{k}}_1, \hat{\mathbf{k}}_2, \ldots, \hat{\mathbf{k}}_{N_s}]^{\mathrm{T}}$, and

$$\mathbf{A} = \left[(\mathbf{I} - \hat{\mathbf{k}}_1\hat{\mathbf{k}}_1^{\mathrm{T}})\mathbf{p}_1, (\mathbf{I} - \hat{\mathbf{k}}_2\hat{\mathbf{k}}_2^{\mathrm{T}})\mathbf{p}_2, \ldots, (\mathbf{I} - \hat{\mathbf{k}}_{N_s}\hat{\mathbf{k}}_{N_s}^{\mathrm{T}})\mathbf{p}_{N_s}\right]. \tag{4.12}$$

Different weighting selections are discussed in the original paper [Haw03], including a range based reweighting method. This reweighting scheme extends the Stansfield's estimator to 3D source localization [Haw03].

Some DOA estimators are capable of producing a reliability measure of the DOA estimate. A reliability measure could be used to weight the DOA measurement. Such DOA estimators include the TDOA-based LS DOA estimator [YH96], of which there exists a reliability measure [Pir03].

## 4.3 Robust DOA-Based Localization

This section is based on [P2], where a robust DOA-based source localization method is presented. In localization of a point source, as discussed above, some localization criterion is minimized over the space, e.g., the shortest distance from the observed direction lines to the source [Haw03] or angle deviation between measured directions and array-to-source directions [P4].

In [P2] the source distance deviation is selected as the criterion to be minimized. The distance deviation is defined in (4.6), see Fig. 4.1 for illustration.

DOA measurements made at several spatially separated stations might not originate from the same source. When such DOA estimates are used in the closed-form solution (4.11) the result will become biased. Even if multitude of sensor stations exist, a single outlier may reduce the efficiency of the method. In some GPS and navigation applications receiver autonomous integrity monitoring (RAIM) is performed to detect faults and exclude them in overdetermined solutions. In [Kuu05] criteria for measurement exclusion were discussed. A similar approach can be adopted for DOA-based source localization.

Let the set of sensor stations be $\Omega = \{1, 2, \ldots, N_s\}$, where $N_s$ is the number of stations that provide DOA measurements. The power set

$$\mathcal{P}(\Omega) = \{\{\emptyset\}, \{1\}, \ldots, \{N_s\}, \{1,2\}, \ldots, \{N_s - 1, N_s\}, \{1,2,3\}, \ldots\} \tag{4.13}$$

is a set of all subsets of $\Omega$. Two DOA stations are sufficient for localization[1] purposes but three stations guarantee redundancy. A subset of the powerset with three or more sensor stations (and their corresponding DOA measurements) is noted by $\Omega_n$. There are $N_{3^+}$ such subsets, where $N_{3^+} = \sum_{s=3}^{N_s} \binom{N_s}{s}$.

The average distance criterion (ADC) from a specific subset $n$ can be written [P2]

$$\mathrm{ADC}(n) = \frac{1}{|\Omega_n|} \sum_{k \in \Omega_n} d_{k,n}, \tag{4.14}$$

where $n = 1, \ldots, N_{3^+}$ is the subset index, $|\Omega_n|$ is cardinality of set $\Omega_n$, and $d_{k,n}$ the distance criterion contributed by the $k$th array in the subset $n$, see (4.6). The subset that has the minimum (ADC) is selected to produce the least squares location solution that is assigned as the final source location estimate $\hat{\mathbf{r}}_{\mathrm{ADC}}$:

$$\tilde{n} = \operatorname*{argmin}_n \mathrm{ADC}(n), \tag{4.15}$$

$$\hat{\mathbf{r}}_{\mathrm{ADC}} = \hat{\mathbf{r}}_{\tilde{n}}. \tag{4.16}$$

The method is a special case of observation subset testing, where a maximal subset with the smallest acceptable test statistics is chosen with respect to a predefined threshold [Kuu05]. The threshold selection is omitted. The method is described as Alg. 3 (Appendix A, p. 143).

[1] A pathological case is the line connecting stations.

The method (Alg. 3) was tested in [P2] with simulations and real-data, and compared to the weighted[2] least squares solution (4.11). In the real-data analysis the DOA was obtained from a closed-form TDOA based least squares method, presented in [YH96], where the TDOA estimates were obtained from PHAT-weighted correlation functions.

### 4.3.1 Simulations

The two localization performance metrics discussed in Section 3.8.1 are used, i.e., the AEE which is the average estimate error and miss percentage.

Simulations are used to compare the localization performance of the ADC (Algorithm 3) and traditional least squares (4.8) methods. The DOA estimates are simulated with varying amounts of two different types of disturbances. The simulation scenario consists of a continuous sound source and five microphone arrays. The arrays are located on the walls of an anechoic room of dimension $6 \times 6$ meters. The exact simulation setup is described in Table 4.1. The simulations are performed in 2D for simplicity. The DOA measurement model for each array is given in polar coordinates $(\alpha, r)$:

$$\alpha_i(m) = \begin{cases} \alpha_{i,\boldsymbol{\theta}} + \mathcal{N}(0, \sigma^2) & p > p_{outlier} \\ \mathcal{U}(0, 2\pi) & p \leq p_{outlier} \end{cases}, r_i = 1; \tag{4.17}$$

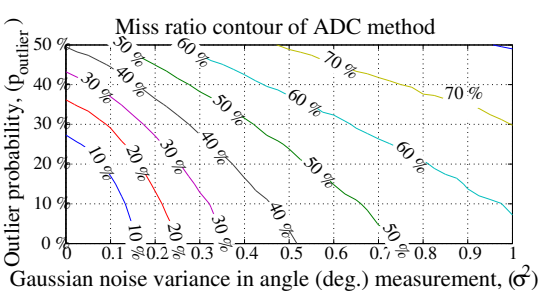
where $m$ is a repetition number $m = 1, ..., 50000$, and $\alpha_{i,\boldsymbol{\theta}}$ is the true source direction at array $i$. The disturbances are IID Gaussian noise with zero mean and variance $\sigma^2$. $p_{outlier}$ is probability of an outlier. An outlier is modeled as an uniformly distributed random direction estimate. In the simulation the degree of disturbances was varied to find out the typical behavior of the methods. The results for both methods are given in Fig. 4.2, and are originally presented in [P2].

It is seen from Fig. 4.2 that the advantage of the ADC method is clear when the outlier probability increases. The miss ratio is better in the ADC method when a certain amount of outliers are present in the direction estimates. For example, if the outlier probability $p_{outlier} = 25$ % and Gaussian angle variance $\sigma^2 = 0.2$ the ADC method has a miss ratio of less than 30 % and the LS miss ratio is close to 60 %.

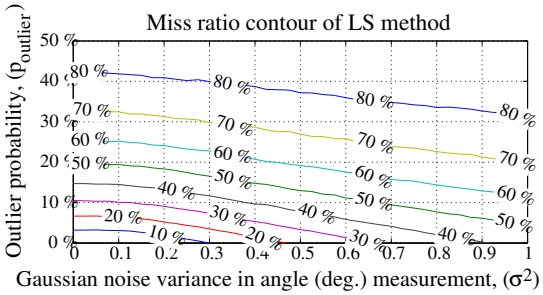
[2] All weights set to one.

Table 4.1: Simulation geometry is given. Five microphone arrays are located at $\mathbf{p}_1, \ldots, \mathbf{p}_5$ and the source location is $\mathbf{r}$.

| | $\mathbf{p}_1$ | $\mathbf{p}_2$ | $\mathbf{p}_3$ | $\mathbf{p}_4$ | $\mathbf{p}_5$ | $\mathbf{r}$ |
|---|---|---|---|---|---|---|
| $x$-coordinate [m] | 0 | 2 | 6 | 6 | 2 | 1.5 |
| $y$-coordinate [m] | 3 | 0 | 0 | 6 | 6 | 1.5 |




(a) Results for the localization miss ratio of the ADC method.




(b) Results for the localization miss ratio of the LS method.

Figure 4.2: Localization results in terms of miss ratio is given for the simulation data. The simulation geometry is described in Table 4.1. The x-axis is the variance of Gaussian noise present in a single DOA estimate $\sigma^2$. The y-axis is the probability that a single DOA is an outlier. It is noted that the surface area under the same miss ratio is larger in the average distance criterion (ADC) method than in the least squares (LS) method. Source [P2].

The simulation results show, that the ADC method is more robust against outliers in the angular data. When no outliers exist, the LS solution produces better results, since removal of estimates always increases variance.

### 4.3.2 Results with Speech Data

In the real-data evaluation, presented in [P2], a subset of evaluation data from the CLEAR'06 evaluation was used. The 3.2 h evaluation dataset contains multimodal data (audio, video, orientation) of talkers in a room environment. In the utilized subset (120 min), the audio data is gathered with four microphone arrays and each array consists of four microphones. The array is upside down T-shaped, as described in Section 3.9. The arrays are located on the walls of the room in the University of Karlsruhe (UKA). The 3D reference location of the active speaker is provided in 1 s intervals. For a more complete description of the data and evaluation, see [Mos06]. In the results the average error of the ADC method was 263 mm, and the average error of the LS method was 275 mm. A more significant result is that the percentage of estimates with error larger than 50 cm was reduced from 62 % to 48 % of total estimates. Note that the sequential information could be used to improve the actual localization accuracy.

The ADC localization method was tested in the CLEAR'06 speaker localization evaluation contest. In this approach, the location estimate provided by the ADC method was tracked with a particle filter to utilize the sequential nature of the data. The results are presented in [S2], where the average estimate accuracy within 50 cm of the speaker (AEE) was 245 mm, 72 % of the time of speech activity.

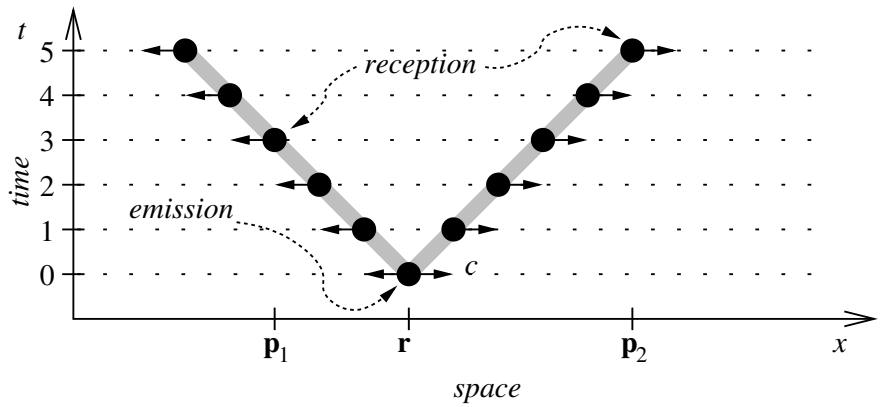



Figure 4.3: The propagation path from source $\mathbf{r}$ to sensor stations $\mathbf{p}_1$ and $\mathbf{p}_2$ is illustrated. The y-axis represents time, and the x-axis represents the space or distance. The emitted wavefront reaches $\mathbf{p}_1$ at time $t = 3$ and station $\mathbf{p}_2$ is reached when $t = 5$. The sensor station $\mathbf{p}_1$ is located closer to the source and therefore receives the source wavefront earlier than station $\mathbf{p}_2$. Note that the angle represents the speed of sound.

## 4.4 DOA Vector-Based Localization Using Propagation Delay

The problem setting discussed in Section 4.1 has assumptions that do not hold in some applications. When the source is located at moderate or large distances the sound propagation time from source to sensor is much greater than the analysis frame length. Omitting the propagation time from source to sensor biases the source position estimate, since the DOA vectors of different sensor stations are not related to the same time instant of source wavefront emission. In a homogeneous environment, the source emitted signal wavefront travels to all directions at the speed of sound. The DOA information is processed by using a frame of $L$ samples, or length $T_w$ seconds.

The observation of acoustic events can be depicted with space-time diagrams [Sea68], which are often used to illustrate the electromagnetic propagation model. Figure 4.3 illustrates an example scenario, where the source at $\mathbf{r}$ emits a wavefront at time $t = 0$. The wavefront then travels at the speed of sound $c$ to all directions and reaches station $\mathbf{p}_1$ at time $t = 3$, and station $\mathbf{p}_2$ is reached at time $t = 5$. Station $\mathbf{p}_1$ is closer to the source, and therefore receives the wavefront before $\mathbf{p}_2$. Now, if at time $t = 3$ the source direction information at both stations is used for localization, $\mathbf{p}_1$ would use information from a wavefront that has not yet arrived at $\mathbf{p}_2$. In fact, using these two sensor stations $\mathbf{p}_{1,2}$ the source can be localized at time $t = 5$ at earliest, since only then the wavefront has reached all sensors and the same information is available. The farther apart the stations are from the source, the more significant this propagation time becomes [P3].

The proposed localization solution including the propagation delay is illus-

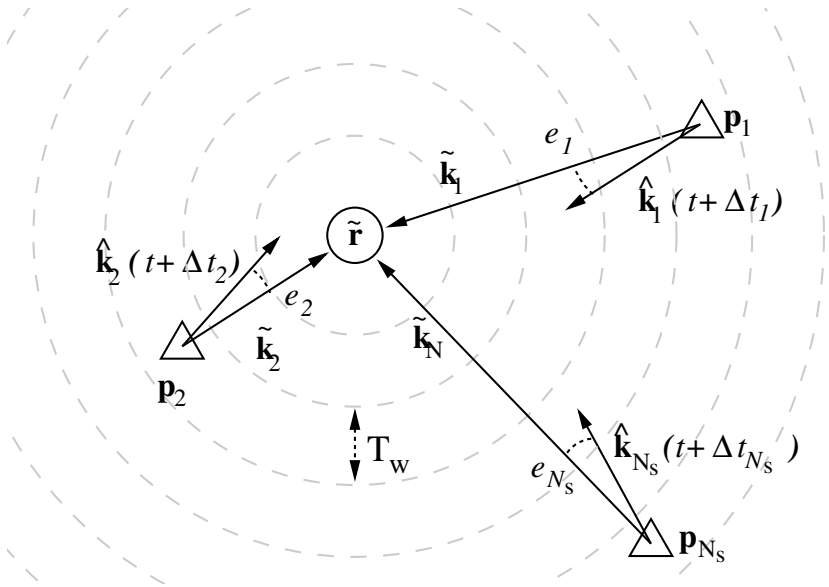



Figure 4.4: The source localization problem with propagation time is illustrated. The source position $\tilde{\mathbf{r}}$ is unknown, sensor station locations $\mathbf{p}_i$ are assumed known, and the direction of arrival (DOA) vectors $\hat{\mathbf{k}}_i$ are estimated, where $i \in 1, \ldots, N_s$. The circles represent the boundaries of window length ($T_w$), i.e., quantized propagation time. The hypothetical DOA vectors are marked $\tilde{\mathbf{k}}_i$.

trated in Fig. 4.4. In order to calculate the likelihood of hypothetical source position $\tilde{\mathbf{r}}$ the propagation delay ($\Delta t_i$) from $\tilde{\mathbf{r}}$ to each sensor $\mathbf{p}_i$ is calculated and used to retrieve a DOA measurement that is time-aligned to the source position $\hat{\mathbf{k}}_i(t + \Delta t_i)$. These DOA vectors are then used to locate the source. In addition to the previous illustration of the DOA based localization problem (Fig. 4.1) the angular error between measurements and reference ($e_i$) is also illustrated. Furthermore, the DOA vectors are indexed with time to emphasize that the propagation time from a hypothetical source position to the sensor station is included in the localization model.

In [Blu00] an iterative solution was presented and used for moderate source (aircraft) distances in outdoors. The initial location is calculated from DOA measurements of each array via triangulation at time $t$ and $t + 1$. The method interpolates new DOA estimates for time $t$ based on the propagation delay and DOA measurements from time $t+1$. The final source estimate is calculated from the interpolated DOA vectors. In [Dom87] a localization scheme is presented, where a single DOA observation line is searched to find the source location with propagation delay corrections for each sensor station.

The propagation delay from a hypothetical source position $\tilde{\mathbf{r}}$ to the observing sensor station at $\mathbf{p}_i$, can be written as

$$\Delta \mathrm{t}_i = \frac{\| \tilde{\mathbf{r}} - \mathbf{p}_i \|}{c} = \frac{\| \tilde{\mathbf{k}}_i \|}{c}, \tag{4.18}$$

where the hypothetical DOA vector or the station-to-source vector $\tilde{\mathbf{k}}_i$ is defined $\tilde{\mathbf{k}}_i = \tilde{\mathbf{r}} - \mathbf{p}_i$, similarly to (4.3). In practice, the data is processed in sequential frames of length $L$ and the time delay is written as number of processing frames:

$$\Delta t_i = \lfloor f_s \cdot \parallel \tilde{\mathbf{k}}_i \parallel \cdot (c \cdot L)^{-1} \rfloor. \tag{4.19}$$

Including the propagation delay into a localization solution means that the spatial likelihood function depends not only on the most recent set of DOA observation from all stations (like the closed-form solution), but also on the future DOA observations.

Several likelihood criteria for a source position can be adopted in the localization problem. The closed-form solution (4.11) uses distances from the observation lines to the hypothetical source position (4.6). Another criterion is the angular error between observations and the selected hypothetical source position [P3], illustrated in Fig. 4.4.

The likelihood for a hypothetical source location $\tilde{\mathbf{r}}$ at time $t$ given the DOA measurement vectors $\hat{\mathbf{K}}(\mathbf{t})$ from all necessary time[3] frames

$$\mathbf{t} \triangleq \{t, t+1, \ldots, t + \max_i(\Delta t_i)\}, \tag{4.20}$$

can be written as

$$\begin{aligned} P(\hat{\mathbf{K}}(\mathbf{t})|\tilde{\mathbf{r}}) &= \textstyle\sum_{i=1}^{N_s} \dfrac{\hat{\mathbf{k}}_i(t+\Delta t_i)^{\mathrm{T}}(\tilde{\mathbf{r}} - \mathbf{p}_i)}{\|\hat{\mathbf{k}}_i(t+\Delta t_i)^{\mathrm{T}}(\tilde{\mathbf{r}} - \mathbf{p}_i)\|} \\ &= \textstyle\sum_{i=1}^{N_s} \dfrac{\hat{\mathbf{k}}_i(t+\Delta t_i)^{\mathrm{T}}\tilde{\mathbf{k}}_i}{\|\hat{\mathbf{k}}_i(t+\Delta t_i)^{\mathrm{T}}\tilde{\mathbf{k}}_i\|} \\ &= \textstyle\sum_{i=1}^{N_s} e_i, \end{aligned} \tag{4.21}$$

where dot product is implicitly used to measure the deviation between the measurement and hypothetical directions. The possible deviation values are $e_i \in [-1, 1]$. Value 1 of $e_i$ indicates that the hypothesis matches exactly the measurement made at station $i$.

The DOA-based source localization algorithm using the propagation time correction is displayed in Alg. 4 (Appendix A, p. 144). In [P3] the source location is estimated by maximizing the function (4.21):

$$\hat{\mathbf{r}}(t) = \underset{\tilde{\mathbf{r}}}{\operatorname{argmax}}\, P(\hat{\mathbf{K}}(\mathbf{t})|\tilde{\mathbf{r}}). \tag{4.22}$$

It is noted, that the function (4.21) is not continuous with respect to the source position. The function utilizes several independent DOA estimates form different time frames from multiple sensor stations. Therefore, no exact closed-form

[3]Note that future measurements are required.

solution exists. However, the function can be minimized, e.g., iteratively using the closed-form solution as an initial guess. The method is directly applicable to scenarios with moving sound sources [P4]. Also sequential estimation methods could be applied to solve the source location from (4.21). Real-world experiments have been used to verify that the introduction of propagation delay to the DOA vector-based localization is feasible [S3]. In a related work [Guo08], a recursive method for predicting the current source position including the propagation delay is considered.

### 4.4.1 Simulation Results

Simulations presented in [P3] consider a transient sound source switching between active and silent modes in 500 ms periods. The signal travels at the speed of sound and reaches the separated DOA stations according to their distances from the source. If the received signal originates from an active source mode the DOA points towards the source. Otherwise the DOA is drawn from a random distribution. The DOA is simulated in 20 ms segments. The source location is then estimated by two separate methods from the simulated DOA values. The methods estimate the source position with and without considering the propagation delay. The source is located at the center of an area and the area size is varied from a room scale environment to a 1000 m $\times$ 1000 m area. When the area size increases the observed DOA values start to originate from both silent and active source time instants. Combining such estimates in the localization process without considering the propagation delay leads to erroneous results.

In [P4] DOA observations of a moving source are simulated for three spatially separated sensor stations. The results indicate that the source location estimate contains bias if the sound propagation delay is not considered. This bias is due to combining DOA observations that originate from different source locations. The bias increases as a function of source velocity since the source locations from which the used DOA observations originate from are increasingly farther apart from each other. The bias increase as a function of source velocity is removed by using the propagation delay (4.21). The proposed method therefore improves the localization of distant moving sources.

To conclude, when spatially separated sensor stations receive direction observations from a distant source the observations originate from different source emission instants. If the source is distant and also moving the observations originate from different source emission times and also different source positions due to movement. The proposed method uses propagation delay to time-align the observations so that they originate from the same source time instant. Without using the time-alignment the source location estimate will become biased.

## 4.5 Localization Using TDE-Based Array Steered Responses

Section 4.2 discusses a method to solve the source localization problem using the DOA vectors obtained from spatially separated microphone arrays. The specific method for obtaining the source direction was not discussed. Possible methods for DOA estimation include beamforming [Joh93], TDE-based array steered response [Joh02] and TDOA-based closed form methods, such as the least squares DOA method [YH96]. Traditionally the DOA is obtained by maximizing the microphone array steered response. If the maximum is a result of a spurious peak the DOA estimate will become biased, and possibly corrupt the location estimate. This section proposes direct utilization of TDE-based array steered responses for source location estimation.

The extension of the TDE likelihood function concept to DOA based localization involves a formulation of a function that maps the microphone pairwise TDE responses into an array response. Combining responses from several such arrays then results in the spatial likelihood function (SLF) of source position.

The TDOA between two microphones $p = \{l, k\}$ is now obtained from the planar wavefront[4]

$$\begin{aligned}\Delta\tau_p(\mathbf{k}) &= \frac{(\mathbf{m}_l - \mathbf{m}_k)^{\mathrm{T}}\mathbf{k}}{\|\mathbf{m}_l - \mathbf{m}_k\| \cdot \|\mathbf{k}\|} \cdot \|\mathbf{m}_l - \mathbf{m}_k\| \cdot \frac{1}{c} \\ &= \frac{(\mathbf{m}_l - \mathbf{m}_k)^{\mathrm{T}}\mathbf{k}}{\|\mathbf{k}\| \cdot c},\end{aligned} \tag{4.23}$$

where $\mathbf{m}$ is a symbol for microphone location.

The TDE-based likelihood function value for a hypothetical direction vector $\mathbf{k}$ is

$$P_{\mathrm{DOA}}(\mathcal{R}_p|\mathbf{k}) = \mathcal{R}_p(\Delta\tau_p(\mathbf{k})), \tag{4.24}$$

where $P_{\mathrm{DOA}}(\cdot|\cdot)$ represents conditional likelihood for the measurement $\mathcal{R}_p$ between a microphone pair $p$ given the source direction $\mathbf{k}$. Measurement likelihood for a source direction is illustrated in Fig. 4.5, where the TDE function between two microphones is mapped into different source directions using (4.24).

Combining microphone pairwise TDE-based likelihoods in array $i$ is written[5]

$$P_{\mathrm{DOA}}(\mathcal{R}_{[1:S_i]}|\mathbf{k}_i) = \bigotimes_{p=1}^{S_i} \mathcal{R}_p(\Delta\tau_p(\mathbf{k}_i)), \tag{4.25}$$

where $\mathcal{R}_{[1:S_i]}$ denotes the pairwise TDE measurements at array $i$, $\otimes$ is a combination operator, and $S_i$ is the number of microphone pairs in array $i$. Function (4.25) represents the array steered response. The SRP-PHAT approach for

[4] Instead of spherical wavefront as in (2.13).

[5] Note again the similarity to (3.51).

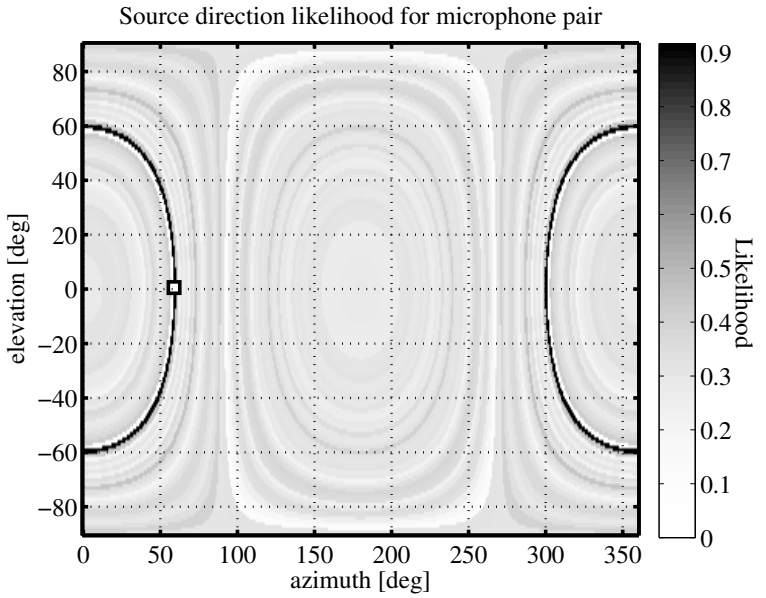



Figure 4.5: Figure illustrates sound source direction of arrival (DOA) likelihood obtained using TDE based likelihood function between a microphone pair. The source azimuth and elevation are $[58.9, 0.59]^{\mathrm{T}}$ degrees and range is 19 m. The source direction is illustrated with a square symbol "□". The microphones are on the x-axis with a separation of 30 cm. The sampling frequency is 48 kHz, and frame length is 21.3 ms. The source signal is white noise.

direction of arrival estimation (or far-field SRP-PHAT [Joh02]) uses summation operation in (4.25) to construct the steered response.

Figure 4.6 illustrates a steered response of a four microphone 3D array in a real scenario. The panels display the summation (left) and multiplication (right) of six microphone pairwise TDE likelihood functions for each direction in space. Note the false intersection of the pairwise likelihood e.g. $[328, 59]^{\mathrm{T}}$ degrees in the summation approach.

The (hypothetical) DOA vector can be obtained from a hypothetical source position $\mathbf{k}_i = \mathbf{r} - \mathbf{p}_i$ as defined previously (4.3). The source position measurement is ambiguous, since the likelihood changes only when the direction of the source changes and is therefore invariant to source range as long as the source is in far-field of the array. The spatial likelihood function can be constructed by combining steered responses (4.25) from arrays $i = 1, \ldots, N_s$. The resulting likelihood function for a source position $\mathbf{r}$ is written

$$P(\mathcal{R}_{[1:S_1,\ldots,1:S_{N_s}]}|\mathbf{r}) = \bigoplus_{i=1}^{N_s} \bigotimes_{p=1}^{S_i} \mathcal{R}_p(\Delta\tau_p(\mathbf{r} - \mathbf{p}_i)), \tag{4.26}$$

where $\mathcal{R}_{[1:S_1,\ldots,1:S_{N_s}]}$ denotes TDE-based likelihood functions from arrays $1, 2, \ldots, N_s$, $\oplus$ and $\otimes$ are operations that follow the likelihood combination rules

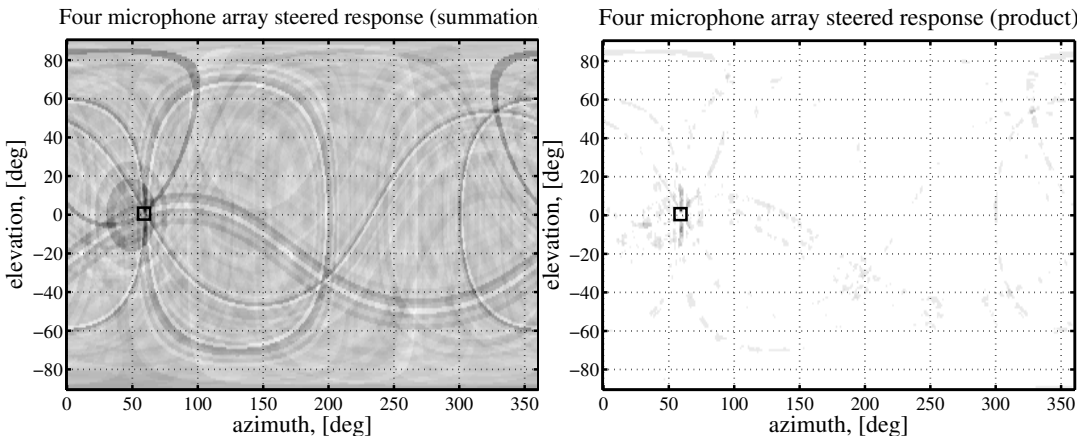



Figure 4.6: Figure illustrates the TDE-based array directional steered response. The source is located at direction (azimuth, elevation) $[58.9, 0.59]^{\mathrm{T}}$ degrees, illustrated with a square symbol "□". The response in the left panel is obtained by adding the pairwise TDE likelihoods. The response in the right panel is obtained by multiplying the likelihoods. Note the difference in the response peak shape and location.

presented in [P1], and also discussed Section 3.5. The selected combination operation affects the resulting spatial likelihood function (SLF) shape. In the spherical wavefront case, the intersection (e.g. multiplication) of the values was found more suitable than the union (e.g. summation) of values using simulations and real-data [P1] and in the CLEAR'07 evaluation dataset as discussed in Section 3.9.2. This performance increase is due to lower variance in the source position distribution as discussed in Section 3.5.4. A similar trend can be observed from Figure 4.6.

Note that the far-field scenario is very similar to the near-field scenario and the only difference is that the time difference of arrival value is obtained by assuming a planar wavefront instead of the spherical wavefront. This affects the construction of the SLF and its geometrical interpretation. More precisely, the calculation of the hypothetical time difference value between the microphones is different in Eqs. (2.13) and (4.23). Also, in the near-field the SLF represents the source position likelihood whereas in the far-field the SLF is the directional likelihood or array directional response. These directional responses are combined to result the spatial likelihood function of source location. The techniques to estimate the source position from the SLF discussed in Sections 3.6 and 3.7 are applicable also here, i.e., maximum search and sequential Bayesian estimation.

The algorithm describing the approach is written as Alg. 5 (Appendix A, p. 145), where $\otimes$ is used to combine the pairwise likelihoods inside each array $i$ at line 16 to form the array steered response. Operator $\oplus$ then combines the array responses for each source position at line 20. Note that in the far-field generalization of SRP-PHAT [Joh02] summation of the TDE likelihood values at line 16 is used.

A similar approach has been presented in [Ali07], where a beamforming based array steered responses are combined in to a spatial likelihood function. The SLF is obtained by summing the log-likelihoods of each array to each hypothetical source direction. In [Ali07] the source emits a narrowband acoustic signal[6], and beamforming is therefore a natural selection for the array response algorithm.

To conclude, the combination of directional responses of TDE-based microphone arrays was considered for constructing the source spatial likelihood function in a similar framework to the near-field localization framework discussed in Section 3.5.

## 4.6 Sound Propagation Delay in Directional Steered Response Localization

Section 4.5 discussed the use of array directional response (such as far-field SRP-PHAT) in spatially separated array based sound source localization. However, as previously motivated in Section 4.4 the problem of localizing distant sources with acoustical observations requires compensation of the acoustic propagation delay, since at moderate and large distances the acoustic propagation delay from a source to a receiving array is larger than the processing frame length.

Therefore, in this section the array directional response -based localization is extended to include the propagation delay, based on [P5]. This results in a novel source localization estimator.

As discussion in Section 1.2.2 the sound is attenuated by the geometrical spreading attenuation, atmospheric attenuation, ground attenuation. To focus specifically on the propagation delay the received signal is modeled as a delayed and scaled version of the emitted signal:

$$x_l(r_l, t) = A \cdot s(t - r_l/c) + v_l(t), \tag{4.27}$$

where $r_l$ is Euclidean distance from microphone $l$ to source, i.e., $r_l = \|\mathbf{m}_l - \mathbf{r}\|$, $A$ is amplitude, $s(t)$ is source signal, $c$ is speed of sound, and $v_l$ is white IID noise. The received signal can be written in the form:

$$x_l(r_l, t + r_l/c) = A \cdot s(t) + v_l(t + r_l/c). \tag{4.28}$$

In (4.28) the received signal is time-aligned at the source emission location (time $t$) for all microphones. Causality prevents an online solution since samples from future time instants related to source distance are required.

It is assumed that the source signal's statistical properties remain unchanged during a frame of length $L$ samples. If the source is moving, the propagation delay is time varying. If a sound source is moving away from the observer, the emitted

[6] Bird song.

samples spread in time. Conversely, for an approaching source the samples are temporally clustered. This is known as the Doppler effect. The propagation delay difference between two microphones is approximated as a constant during a time frame period $t \in [0, T_w]$:

$$\tau_l(t) - \tau_k(t) \approx \Delta\tau_p, \tag{4.29}$$

where $\tau_l(t) = r_l(t)/c$ is the propagation delay from the (moving) source to microphone $l$. Equation (4.29) holds exactly in the special case the source is moving on a hyperboloid defined by the source and microphone positions. For long processing frames the Doppler effect becomes significant and could be compensated [Fer99]. In [Tea07] the Doppler shift is modeled and used for estimating the source position. In practice, the approximation (4.29) can be justified if the microphones are located close to each other, located in the far-field of the source, and $T_w$ is kept short. Here, we restrict the frame length and group the microphones into arrays to satisfy this assumption. In addition, in the atmosphere, the coherence between two microphones decreases when the microphone separation increases, due to different acoustic paths and their different properties [Ash05]. This observation also suggests that microphones should be grouped into arrays, with relatively small inter-microphone distances.

The processing of data in frames results in the quantization of the propagation delay into frame lengths. For example, if the distance from the source to receiver is 200 m the propagation delay is 0.58 seconds. For a 20 ms frame length this corresponds to 29 frames, see (4.19).

The time aligned directional spatial response for array $i$ is written as

$$P_{\mathrm{DOA}}(\mathcal{R}_{[1:S_i]}^{t+\Delta t_i}|\mathbf{r}) = \bigotimes_{p=1}^{S_i} \mathcal{R}_p^{t+\Delta t_i}(\Delta\tau_p(\mathbf{r} - \mathbf{p}_i)), \tag{4.30}$$

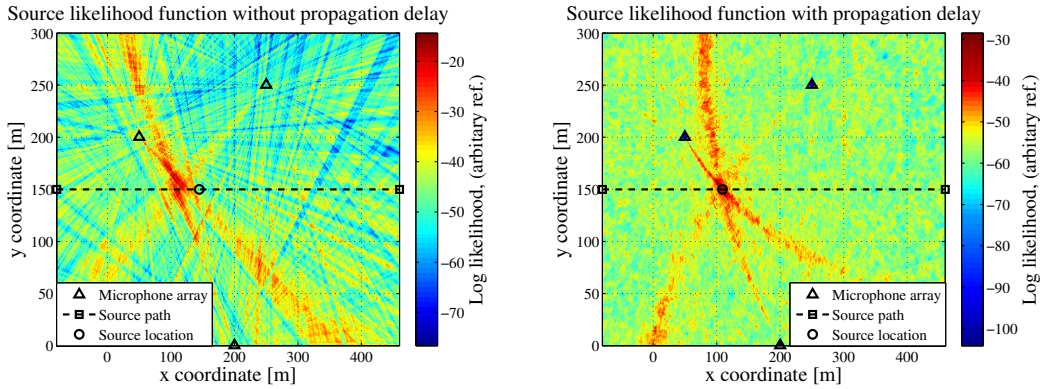
where $\Delta t_i$ is the propagation delay in frames (4.19). Note that the time-aligned spatial response (4.30) requires a source position hypothesis in order to include the propagation delay, unlike the traditional steered array response (4.25). In (4.25) the time index was omitted. Combining directional likelihoods (4.30) from arrays is then written as [P5]

$$P(\mathcal{R}_{\{[1:S_1],\ldots,[1:S_{N_s}]\}}^{\mathbf{t}}|\mathbf{r}) = \bigoplus_{i=1}^{N_s} \bigotimes_{p=1}^{S_i} \mathcal{R}_p^{t+\Delta t_i}(\Delta\tau_p(\mathbf{r} - \mathbf{p}_i)), \tag{4.31}$$

where $\mathcal{R}_{\{[1:S_1],\ldots,[1:S_{N_s}]\}}^{\mathbf{t}}$ denotes TDE-based steered responses from arrays $1, 2, \ldots, N_s$ from time frames $\mathbf{t}$ (4.20).

### 4.6.1 Implementation Issues

The far-field TDE-based localization function using the propagation delay is detailed in Alg. 6 (Appendix A, p. 146). The method is implemented using a




(a) Traditional far-field SRP-PHAT based SLF.

(b) Far-field SRP-PHAT based SLF with time-alignment.

Figure 4.7: An illustration of the 2D spatial likelihood function (SLF) using steered array responses directly (left) and with the proposed time-alignment (right). The average SNR is set to -5 dB in each channel. The endpoints of source path ("□"), source location ("○"), and microphone array locations ("▽") are illustrated. The source velocity is 66.7 m/s in the positive x-axis direction. Source [P5].

grid-based maximum search method. At line 9 a loop is added to calculate the future signal correlation values. The maximum future time instant is determined by the maximum propagation time from any source location to any array. Note that this renders the algorithm offline, and the resulting source location answers the question "where was the source", rather than "where is the source now". The latter question cannot be directly answered, since no information from the current source position is available due to the limited propagation speed of sound. Only predictions, e.g., based on source state could be given to estimate the current position.

### 4.6.2 Simulations

In [P5] the proposed approach is tested with simulations. A source is set to emit pseudo-random noise with a flat spectrum at frequencies in the range [20, 10k] Hz. The source moves at a constant speed of 66.7 m/s along a line defined by coordinates $[-80, 150]^{\mathrm{T}}$ and $[460, 150]^{\mathrm{T}}$. The simulation models the geometrical and atmospheric attenuation according to [IE93] using 20 °C temperature ($T$) and 70 % humidity ($H$) as parameters.

Three arrays each consisting of four microphones are used with a sampling frequency of 48 kHz. The 2D arrays are located at $[200, 0]^{\mathrm{T}}$, $[50, 200]^{\mathrm{T}}$, and $[250, 250]^{\mathrm{T}}$. The arrays are similar in shape; three array microphones are placed in a triangle with 1 m edge length and the fourth microphone located at center of the triangle.

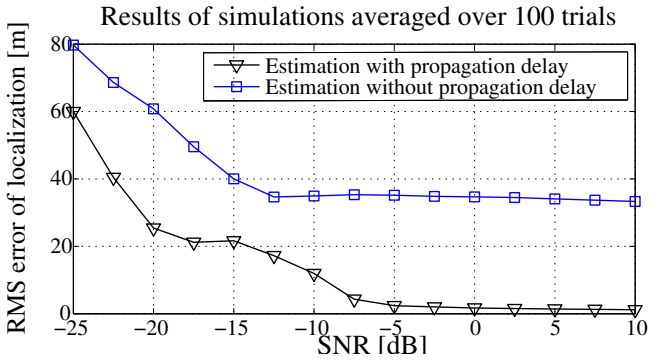



Figure 4.8: The RMS error of localization with simulations is displayed. The x-axis indicates signal SNR and y-axis is RMS error. The method with propagation delay (▽) increases localization accuracy, compared to not using propagation delay to time-align observations (□). Source [P5].

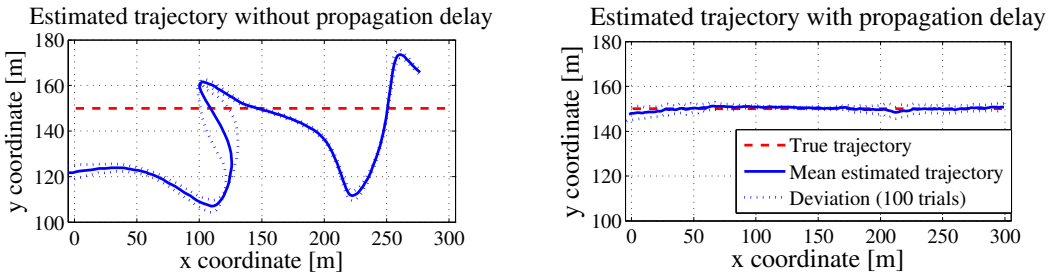
Fig. 4.7 shows the output for the traditional method (4.26) in panel 4.7(a) and the time-aligned method (4.31) is displayed in panel 4.7(b). The output is calculated for a fixed grid of cell size 1 $m^2$ using Alg. 6. The array response was obtained with the far-field SRP-PHAT and the array responses were combined with multiplication to estimate the SLF. In panel 4.7(a) the maximum response direction from each array points behind the moving source due to propagation delay. In panel 4.7(b) data from adjacent time frames is used based on hypothetical source distance. Therefore, the observations seem to curve behind the source position.

### 4.6.3 Results

The traditional approach of not utilizing a propagation delay (4.26) was compared with the proposed model that includes the propagation delay (4.31). A particle filter with 5000 particles was used to estimate the source state. The state consisted of (x,y) velocity and location components, and the Langevin motion model was used.

The results are given for the mean RMS error as a function of microphone signal average SNR in Fig. 4.8. The x-axis indicates the average SNR for each microphone signal for the whole recording, and y-axis displays the RMS error. A small amount of data from the beginning and end of simulation is omitted to lessen errors due to filter convergence and end of data. Additionally, the particles were centered near the true source position in the beginning of calculation. Since the PF is a stochastic method the results of 100 trial runs were averaged. For details refer to [P5].

The results show a significant decrease in the RMS error when the propagation delay is included in the estimation. With every SNR value the proposed method outperforms the traditional approach. When the propagation delay is not in-




(a) Average source trajectory without propagation delay.

(b) Average source trajectory with propagation delay.

Figure 4.9: The location estimates of -5 dB SNR case with the traditional method and the proposed time-aligned method is detailed. The x-axis corresponds to x coordinate, and the y-axis is y coordinate. The standard deviation of the 100 simulations is illustrated with the mean location estimate.

cluded a constant bias exists in the RMS error. This is removed by the proposed approach. Figure 4.9 details one simulation with SNR -5 dB. The source location deviation is quite small, so the shape of the source trajectory bias is revealed in the panel 4.9(a), which is calculated without including the propagation delay. As seen from Fig. 4.9(b) the time-alignment of signal frames based on propagation delay removes the bias.

## 4.7 Summary

The DOA-based and bearings only localization problem was reviewed in Section 4.1 with the CRLB of the bearings-only estimator. In Section 4.2 a closed-form 3D localization estimator was reviewed, and a robust version with DOA discrimination was proposed for speaker localization in Section 4.3 based on [P2]. The method was shown to improve source localization in the case of DOA outlier values.

The problem of limited speed of sound and delayed observations was discussed for localization in Section 4.4, where a DOA-based localization method was presented based on [P3][P4]. The method calculates the propagation delay between the hypothetical source position and receiver array, and uses this propagation delay to index a proper temporal DOA measurement. The approach is suitable for locating moving sources.

As discussed in Section 2.7 TDOA estimates are known to fail rapidly due to the threshold effect when conditions become challenging. To extend the operation conditions of DOA based localization, the mapping of array steered responses into the spatial likelihood function of source position was discussed in Section 4.5. To the knowledge of the author this is the first time that TDE likelihood -based steered array responses are used to build the spatial likelihood function. As

discussed in Section 4.4 the propagation delay presents a problem for locating a moving or a transient sound source. The steered array response localization was then extended to include the use of the propagation delay in Section 4.6. This novel estimator was tested with simulations and shown to be more accurate than the basic method of directly combining the array steered responses.

Chapter 5

# Conclusions, Discussion, and Future Work

THIS thesis discussed acoustic source localization methods in reverberant rooms and in large outdoor spaces. The sound sources of interest in this work have a wide spectrum, such as speech, different vehicles, and jet aircraft. The time delay estimation methods are suitable for wideband sources, and methods for estimating the time difference of arrival (TDOA) between two microphones were summarized. The TDOA methods exhibit a threshold behavior where the TDOA error suddenly increases after some parameter value falls outside an acceptable region. Such parameters include reverberation time, coherence, and SNR.

The TDOA based closed-form localization solutions were summarized. The closed-form solutions are computationally light but are sensitive to the TDOA threshold effect. To enable localization in more challenging environments the TDE likelihood -based methods were discussed since they are not similarly degraded by the threshold effect. The SRP-PHAT has been used for speaker localization in many practical environments. In this thesis the SRP-PHAT was examined as a function to combine TDE-based likelihood values. A general framework for this purpose was described and other possible combination operations such as multiplication (Multi-PHAT) and the Hamacher t-norm (Hamacher-PHAT) were considered. The Multi-PHAT and Hamacher-PHAT outperform SRP-PHAT in a noisy and reverberant room environment when combined with the efficient and computationally light particle filtering techniques [P1]. The Multi-PHAT, SRP-PHAT, and Hamacher-PHAT -based localization with a particle filtering was also tested with the CLEAR'07 technology evaluation database for indoor talker localization. Results agree with the performance difference predicted by the simulations.

The direction of arrival -based localization was examined from the viewpoint of using the direction observations of multiple acoustical arrays. A robust outlier removal method was presented and evaluated for indoor sound source localiza-

tion [P2]. When the distance between source and array becomes moderate the propagation delay from source to sensor is multiple times longer than the processing frame length. Combining such directional estimates results in a biased location estimate if the source is transient [P3] or moving [P4]. A method for using temporally different direction estimates from multiple arrays was then presented based on the conducted research.

Motivated by the performance of TDE-based likelihood localization in indoors the propagation delay compensation method was included in this model. The proposed method combines the state-of-the-art indoor sound source localization method with the propagation delay model for outdoor sound source localization [P5]. The simulation results show that the method can remove the bias from the moving source location estimate obtained with the traditional approach.

# Discussion and Future Work

The first part of this thesis focuses on near-field localization methods with a similar geometrical model of the problem setting. The observation that TDE likelihood function -based localization is more robust than the closed-form solution using TDE likelihood parameterization is rather obvious, and the result is probably known to many researchers. Still the degree of performance difference was somewhat surprising.

The sound source localization was studied in the reverberant room environment, where the room impulse response determines the contributions of the reflections to the microphone pairwise TDE function. By assuming the shape and dimensions of the room is available the reverberation could be harnessed for localization. However, in this thesis such information was not assumed available and the effect of reverberation is examined by the error it generates into the location estimator.

In the outdoor sound source localization estimation the speed of wave propagation was assumed constant. The effect of error in propagation speed $c$ was not evaluated. By not including the propagation delay is equal to calculating the propagation delay with $c$ set to infinite. Therefore, small error in $c$ can be thought not to deteriorate the location estimation. The air is not homogeneous and the speed of sound is not constant. Also wind moves the medium in which sound travels therefore affecting sound travel path. Wind speeds could be easily estimated with affordable sensors. The use of wind speed maps in the localization estimation process therefore represents a promising direction of future research. Also other information available on sound propagation based on meteorological conditions presents an interesting future research topic. The sound source position likelihood conditioned on the measurements and speed of sound could be conditioned on a more complex model of the propagation of path.

Chapter 6

# Errata

In [P1] p. 3 Section 3:

- "Equation (7) can be interpreted as a likelihood of a source having location $\mathbf{r}$ given the measurement $P(\mathcal{R}_p|\mathbf{r})$." should read "Equation (7) can be interpreted as measured likelihood given source position $\mathbf{r}$."

# Bibliography


[Aar03] P. Aarabi, The Fusion of Distributed Microphone Arrays for Sound Localization. *EURASIP J. Appl. Signal Process.*, 4:pp. 338 – 347, 2003.

[Abe90] J. Abel, A Divide and Conquer Approach to Least-Squares Estimation. *IEEE Trans. Aerospace and Electronic Systems*, 26(2):pp. 423 – 427, 1990.

[Ali07] A. Ali, T. Collier, L. Girod, K. Yao, C. Taylor, and D. Blumstein, An Empirical Study of Collaborative Acoustic Source Localization. In *Proceedings of the 6th international conference on Information processing in sensor networks*, pp. 41–50, 2007.

[All79] J. Allen and D. Berkley, Image Method for Efficiently Simulating Small-Room Acoustics. *J. Acoust. Soc. Am.*, 65(4):pp. 943 – 950, 1979.

[Aru02] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, A Tutorial on Particle Filters for On-line Non-linear/Non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):pp. 174 – 188, 2002.

[Aru04] M. S. Arulampalam, B. Ristic, N. Gordon, and T. Mansell, Bearings-Only Tracking of Manoeuvring Targets Using Particle Filters. *EURASIP Journal on Applied Signal Processing*, 2004(15):pp. 2351 – 2365, 2004.

[Ash05] J. Ash and R. Moses, Acoustic Time Delay Estimation and Sensor Network Self-Localization: Experimental Results. *J. Acoust. Soc. of Am.*, 118(2):pp. 841–850, Aug. 2005.

[Bar99] J. Bard and F. Ham, Time Difference of Arrival Dilution of Precision and Applications. *IEEE Trans. on Signal Process.*, 47(2):pp. 521–523, 1999.

[Ben00] J. Benesty, Adaptive Eigenvalue Decomposition Algorithm for Passive Acoustic Source Localization. *J. Acoust. Soc. Am*, 107(1):pp. 384 – 391, 2000.

[Ber86] L. Beranek, *Acoustics*. American Institute of Physics, 1986.

[Ber91] M. Berger and H. Silverman, Microphone array optimization by stochastic region contraction. *IEEE Transactions on Signal Processing*, 39(11):pp. 2377 – 2386, 1991.

[Bir01] S. Birchfield and D. Gillmor, Acoustic source direction by hemisphere sampling. In *Proc. Acoust., Speech, and Signal Processing (ICASSP'01)*, vol. 5, pp. 3053 – 3056, 2001.

[Blu00] R. Blumrich and J. Altmann, Medium-Range Localisation of Aircraft via Triangulation. *Applied Acoustics*, 61(1):pp. 65 – 82, 2000.

[Bra95] M. Brandstein, *Framework for Speech Source Localization Using Sensor Arrays*. Ph.D. thesis, Brown University, Providence, RI, USA, 1995.

[Bra99] M. Brandstein, Time-delay estimation of reverberated speech exploiting harmonic structure. *The Journal of the Acoustical Society of America*, 105(5):pp. 2914 – 2919, 1999.

[Bra01] M. Brandstein and D. Ward, (Eds.) *Microphone Arrays*. Springer-Verlag, 2001.

[Bru07] A. Brutti, *Distributed Microphone Networks for sound source localization in smart rooms*. Ph.D. thesis, DIT - University of Trento, Italy, 2007.

[Can07] J. Candy, Bootstrap particle filtering. *IEEE Signal Processing Magazine*, 24(4):pp. 73 – 85, July 2007.

[Car81] G. Carter, Guest editorial time delay estimation. *IEEE Trans. on Acoust., Speech, and Signal Process.*, 29(3):pp. 461–462, 1981.

[Car87] G. Carter, Coherence and time delay estimation. *Proceedings of the IEEE*, 75(2):pp. 236 – 255, 1987.

[Cha94] Y. T. Chan and K. C. Ho, A simple and efficient estimator for hyperbolic location. *IEEE Transactions on Signal Processing*, 42(8):pp. 1905 – 1915, 1994.

[Cha96] B. Champagne, S. Bedard, and A. Stephenne, Performance of time-delay estimation in the presence of room reverberation. *IEEE Transactions on Speech and Audio Processing*, 4(2):pp. 148 – 152, 1996.

[Che01] J. Chen, R. Hudson, and K. Yao, A maximum-likelihood parametric approach to source localizations. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP*, vol. 5, pp. 3013 – 3016, 2001.

[Che05a] J. Chen, J. Benesty, and Y. Huang, Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments. *EURASIP J. Applied Signal Process.*, 2005(1):pp. 25 – 36, 2005.

[Che05b] W.-K. Chen, (Ed.) *The Electrical Engineering Handbook*. Elsevier Academic Press, 200 Wheeler Road, Burlington, MA 01803, USA, 2005, iSBN: 0-12-170960-4.

[Che06] J. Chen, J. Benesty, and Y. Huang, Time delay estimation in room acoustic environments: An overview. *EURASIP J. Applied Signal Process.*, 2006:pp. 1–9, 2006.

[Cho81] S. Chow and P. Schultheiss, Delay estimation using narrow-band processes. *IEEE Trans. Acoust., Speech, and Signal Process.*, 29(3):pp. 478 – 484, 1981.

[Cir08] A. Cirillo, R. Parisi, and A. Uncini, Sound mapping in reverberant rooms by a robust direct method. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, pp. 285 – 288, 2008.

[cle07] CLEAR 2007 Evaluation and Workshop. `http://www.clear-evaluation.org/`, April 2007, last checked 7.11.2008.

[Cro97] M. Crocker, (Ed.) *Encyclopedia of Acoustics*. John Wiley & Sons, INC, 1997.

[DB03] K. K. D. Bechler, Considering the second peak in the gcc function for multi-source tdoa estimation with a microphone array. In *Proc. 8th Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 315 – 318, 2003.

[DiB01a] J. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*. Ph.D. thesis, Brown University, Providence, RI, USA, May 2001.

[DiB01b] J. DiBiase, H. Silverman, and M. Brandstein, *Microphone Arrays*. Springer-Verlag, 2001.

[Dju03] P. Djuric, J. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. Bugallo, and J. Miguez, Particle filtering. *Signal Processing Magazine, IEEE*, 20(5):pp. 19 – 38, Sep. 2003.

[Dmo07] J. Dmochowski, J. Benesty, and S. Affes, A generalized steered response power method for computationally viable source localization. *IEEE Trans. Audio, Speech, Language Processing,*, 15:pp. 2510 – 2526, 2007.

[Do07a] H. Do and H. Silverman, A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC). In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'07)*, pp. 295 – 298, 2007.

[Do07b] H. Do and H. Silverman, A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, pp. 121–124, 2007.

[Doc03] S. Doclo and M. Moonen, Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments. *EURASIP Journal on Applied Signal Processing*, 2003(11):pp. 1110 – 1124, 2003.

[Dom87] F. M. Dommermuth, A Simple Procedure for Tracking Fast Maneuvering Aircraft Using Spatially Distributed Acoustic Sensors. *J. Acoust. Soc. Am.*, 82:pp. 1418 – 1424, 1987.

[Dou01] A. Doucet, N. de Freitas, and N. Gordon, (Eds.) *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science, New York, NY: Springer-Verlag, 2001.

[Dun39] H. K. Dunn and D. W. Farnsworth, Exploration of pressure field around the human head during speech. *The Journal of the Acoustical Society of America*, 10(3):pp. 184 – 199, 1939.

[Emb96] T. Embleton, Tutorial on sound propagation outdoors. *J. Acoust. Soc. Am.*, 100(1):pp. 31–48, 1996.

[Far00] A. Farina, Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *108th AES Convention*, 2000.

[Fer99] B. Ferguson and K. Lo, Passive wideband cross correlation with differential Doppler compensation using the continuous wavelet transform. *The Journal of the Acoustical Society of America*, 106(6):pp. 3434 – 3444, 1999.

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

[Gan06] S. Gannot and G. Dvorkind, Microphone array speaker localizers using spatial-temporal information. *EURASIP Journal on Applied Signal Processing*, 2006:pp. 1–17, 2006, article ID 59625.

[Gar07a] A. Gareta, *A multi-microphone approach to speech processing in a smart-room environment*. Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, 2007.

[Gar07b] J. Garofolo, R. Rose, and R. Stiefelhagen, Eval-ware: Multimodal interaction [best of the web]. *Signal Processing Magazine, IEEE*, 24(2):pp. 154 – 155, 2007.

[Gav92] M. Gavish and A. Weiss, Performance analysis of bearing-only target location algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 28(3):pp. 817 – 828, 1992.

[Gil08] M. Gillette and H. Silverman, A linear closed-form algorithm for source localization from time-differences of arrival. *IEEE Signal Processing Letters*, 15(1):pp. 1–4, 2008.

[Gor93] N. Gordon, D. Salmond, and A. Smith, Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F, Radar and Signal Processing*, 140(2):pp. 107 – 113, 1993.

[Guo08] Y. Guo, A. Xuea, and D. Penga, A recursive algorithm for bearings-only tracking with signal time delay. *Signal Processing*, 88(6):pp. 1539 – 1552, June 2008.

[Gus03] T. Gustafsson, B. Rao, and M. Trivedi, Source localization in reverberant environments: Modeling and statistical analysis. *IEEE Trans. Speech and Audio Process.*, 11(6):pp. 791–803, 2003.

[Hah06] P. Hahn, V. Mathews, and T. Tran, Adaptive realization of a maximum likelihood time delay estimator. In *IEEE Conf. Acoust., Speech and Signal Process. ICASSP'06*, vol. 6, pp. 3121–3124, 2006.

[Has81] J. Hassab and R. Boucher, Performance of the generalized cross correlator in the presence of a strong spectral peak in the signal. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):pp. 549 – 555, 1981.

[Haw03] M. Hawkes and A. Nehorai, Wideband Source Localization Using a Distributed Acoustic Vector-Sensor Array. *IEEE Transactions on Signal Processing*, 51(6):pp. 1479 – 1491, 2003.

[Ho04] K. Ho, L. Kovavisaruch, and H. Parikh, Source Localization Using TDOA with Erroneous Receiver Positions. In *Proceedings of the 2004 International Symposium on Circuits and Systems, ISCAS '04*, vol. 3, pp. 453 – 456, 2004.

[Hol06] J. Hol, T. Schön, and F. Gustafsson, On resampling algorithms for particle filters. In *Proceedings of the Nonlinear Statistical Signal Processing Workshop*, pp. 79 – 82, 2006.

[Hua00] Y. Huang, J. Benesty, and G. W. Elko, Passive acoustic source localization for video camera steering. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, vol. 2, pp. 909 – 912, 2000.

[Hua01] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, Real-time passive source localization: A practical linear-correction least-squares approach. *IEEE Transactions on Speech and Audio Processing*, 9(8):pp. 943 – 956, 2001.

[IE93] ISO:9613-1:1993(E), International Standard 9613-1: Acoustics – Attenuation of sound during propagation outdoors – Part 1: Calculation of the absorption of sound by the atmosphere. 1993.

[Ife93] E. Ifeachor and B. Jerwis, *Digital Signal Processing*. Addison–Wesley, 1993.

[IfPT90] T. N. Institute for Perception-TNO, Speech babble measurement signal. `http://www.spib.rice.edu/spib/data/signals/noise/babble.html`, 1990, sampling rate: 19.98 KHz, A/D: 16 bit.

[Jac93] G. Jacovitti and G. Scarano, Discrete time techniques for time delay estimation. *IEEE Transactions on Signal Processing*, 41(2):pp. 525 – 533, 1993.

[Jan97] J.-S. Jang, C.-T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing - A Computational Approach to Learning and Machine Intelligence*, chap. 2, pp. 35 – 42. Upper Saddle River, NJ: Prentice Hall, 1997.

[Joh93] D. Johnson and D. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Signal Processing Series, Prentice Hall, Upper Saddle River, NJ 07458, 1993.

[Joh02] A. Johansson, N. Grbic, and S. Nordholm, Speaker localisation using the far-field SRP-PHAT in conference telephony. In *International Symposium on Intelligent Signal Processing and Communication Systems, Kaohsiung, Taiwan ROC*, 2002.

[Kap01] L. M. Kaplan, Q. Le, and P. Molnár, Maximum Likelihood Methods for Bearings-Only Target Localization. In *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 5, pp. 3001–3004, 2001.

[Kap06] E. D. Kaplan and C. Hegarty, (Eds.), *Understanding GPS: Principles and Applications*, chap. 7, p. 726. Artech House Publishers, 2nd edn., 2006.

[Kar05] R. Karlsson and F. Gustafsson, Recursive Bayesian Estimation: Bearings-only Applications. *Radar, Sonar and Navigation, IEE Proceedings*, 152(5):pp. 305 – 313, 2005.

[Kay98] S. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall Signal Processing Series, Prentice Hall, 1998.

[Kle06] U. Klee, T. Gehrig, and J. McDonough, Kalman filters for time delay of arrival-based source localization. *EURASIP Journal on Applied Signal Processing*, 2006:pp. 1–15, 2006.

[Kna76] C. Knapp and G. Carter, The Generalized Correlation Method for Estimation of Time Delay. *IEEE Trans. on Acoust., Speech, and Signal Process.*, 24(4):pp. 320 – 327, Aug 1976.

[Kor08] T. Korhonen, Acoustic localization using reverberation with virtual microphones. In *Proceedings of the 11th International Workshop on Acoustic Echo and Noise Control*, 2008.

[Koz04] R. Kozick and B. Sadler, Source localization with distributed sensor arrays and partial spatial coherence. *IEEE Transactions on Signal Processing*, 52(3):pp. 601 – 616, 2004.

[Kuu05] H. Kuusniemi, *User-Level Reliability and Quality Monitoring in Satellite-Based Personal Navigation*. Ph.D. thesis, Tampere University of Technology, Tampere, Finland, 2005.

[Lag98] J. Lagarias, J. Reeds, M. Wright, and P. Wright, Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9(1):pp. 112–147, 1998.

[Lai99] X. Lai and H. Torp, Interpolation methods for time-delay estimation using cross-correlation method for blood velocity measurement. *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, 46(2):pp. 277 – 290, 1999.

[Leh03] E. Lehmann, D. Ward, and R.Williamson, Experimental comparison of particle filtering algorithms for acoustic source localization in a reverberant room. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, pp. 177 – 180, 2003.

[Leh04] E. A. Lehmann, *Particle Filtering Methods for Acoustic Source Localisation and Tracking*. Ph.D. thesis, Australian National University (ANU), Canberra, Australia, July 2004.

[Leh06] E. Lehmann and R. Williamson, Particle filter design using importance sampling for acoustic source localisation and tracking in reverberant environments. *EURASIP Journal on Applied Signal Processing, special issue on Advances in Multi-Microphone Speech Processing*, 2006:p. 9, 2006.

[Leh07] E. Lehmann and A. Johansson, Particle filter with integrated voice activity detection for acoustic source tracking. *EURASIP Journal on Applied Signal Processing*, 2007(1):pp. 1–11, 2007.

[Liu01] Q. Liu, Y. Rui, A. Gupta, and J. Cadiz, Automating camera management for lecture room environments. In *Proc. of the SIGCHI conf. on Human factors in computing systems*, pp. 442 – 449, ACM Press New York, NY, USA, 2001.

[Mas02] S. Maskell and N. Gordon, A tutorial on particle filters for on-line nonlinear/non-gaussian bayesian tracking. In *Proceedings of IEE Colloquium on Tracking*, 2002.

[Mil07] C. Militello and S. Buenafuente, An Exact Noniterative Linear Method for Locating Sources Based on Measuring Receiver Arrival Times. *J. Acoust. Soc. Am.*, 121(6):pp. 3595 – 3601, June 2007.

[Moh08] S. Mohan, M. Lockwood, M. Kramer, and D. Jones, Localization of multiple acoustic sources with small arrays using a coherence test. *J. Acoust. Soc. Am.*, 123(4):pp. 2136–2147, 2008.

[Mor07] M. Moran, R. Greenfield, and D. Wilson, Acoustic array tracking performance under moderately complex environmental conditions. *Applied Acoustics*, 68:pp. 1241–1262, 2007.

[Mos06] D. Mostefa, M.-N. Garcia, K. Bernardin, R. Stiefelhagen, J. McDonough, M. Voit, M. Omologo, F. Marques, H. Ekenel, and A. Pnevmatikakis, Clear evaluation plan v.1.1. Feb 2006, `http://isl.ira.uka.de/clear06/downloads/chil-clear-v1.1-2006-02-21.pdf` last checked Nov. 2008.

[Mos07] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, M. Chu, J. Casas, J. Turmo, L. Cristoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet, The CHIL Audiovisual Corpus for Lecture and Meeting analysis inside Smart Rooms. *Language Resources and Evaluation*, 41(3 – 4):pp. 389 – 407, 2007, `http://www.springerlink.com/content/70h381g7qv721547/fulltext.pdf`.

[Omo94] M. Omologo and P. Svaizer, Acoustic event localization using a crosspower-spectrum phase based technique. In *Proc. Acoust., Speech, and Signal Process., ICASSP'94*, vol. 2, pp. 19 – 22, 1994.

[Omo98] M. Omologo, P. Svaizer, and R. D. Mori, *Spoken Dialogs with Computers*. Elsevier Science & Technology, 1998.

[Pee93] P. Z. Peebles, Jr., *Probability, Random Variables, and Random Signal Processing*. McGraw-Hill, 3rd edn., 1993.

[Per07] P. Pertilä, Sound Source Localization in a Bayesian Framework. In P. Koivisto, (Ed.) *Digest of TISE Seminar 2007*, pp. 64 – 69, Tampere Graduate School in Information Science and Engineering (TISE), Tampere University of Technology, P.O. Box 553, FI-33101 Tampere, Finland, June 2007, iSSN 1458-8463, ISBN 978-952-15-1768-6.

[Pet86] P. Peterson, Simulating the response of multiple microphones to a single acoustic source in a reverberant room. *J. Acoust. Soc. Am.*, 80(5), 1986.

[Pet05a] J. Peterson and C. Kyriakakis, Analysis of fast localization algorithms for acoustical environments. In *Conference Record of the Thirty-Ninth Asilomar Conference on Signals, Systems and Computers*, pp. 1385 – 1389, 2005.

[Pet05b] J. Peterson and C. Kyriakakis, Hybrid algorithm for robust, real-time source localization in reverberant environments. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, pp. 1053 – 1056, 2005.

[Pir03] T. Pirinen, P. Pertilä, and A. Visa, Toward intelligent sensors - reliability for time delay based direction of arrival estimates. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*, vol. 5, pp. 197 – 200, 2003.

[Ray05] V. Raykar, B. Yegnanarayana, S. Mahadeva Prasanna, and R. Duraiswami, Speaker localization using excitation source information. *IEEE Trans. Speech and Audio Processing*, 13(5):pp. 751 – 761, 2005.

[Ree81] F. Reed, P. Feintuch, and N. Bershad, Time delay estimation using the LMS adaptive filter – static behavior. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):pp. 561 – 571, 1981.

[Ros74] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, Average magnitude difference function pitch extractor. *IEEE Trans. Acoust., Speech, and Signal Process.*, 22(5):pp. 353 – 362, 1974.

[Ros90] T. Rossing, *The Science of Sound*. Addison-Wesley, 2nd edn., 1990.

[Sad06] B. Sadler and R. Kozick, A survey of time delay estimation performance bounds. In *Proc. Sensor Array and Multichannel Signal Processing, IEEE Workshop*, 2006.

[Sch65] M. Schroeder, New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(3):pp. 409 – 412, 1965.

[Sch06] J. Schroeder, S. Galler, K. Kyamakya, and K. Jobmann, Practical considerations of optimal three-dimensional indoor localization. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, 2006*, pp. 439 – 443, 2006.

[Sch08] J. Scheuing and B. Yang, *Speech and Audio Processing in Adverse Environments*, chap. 11: Correlation-Based TDOA-Estimation for Multiple Sources in Reverberant Environments. Springer Berlin Heidelberg, 2008.

[Sea68] F. W. Sears and R. W. Brehme, *Introduction to the Theory of Relativity*, chap. 3, p. 46. Addison-Wesley, 1968.

[She05] X. Sheng and Y.-H. Hu, Maximum Likelihood Multiple-Source Localization Using Acoustic Energy Measurements with Wireless Sensor Networks. *IEEE Trans. Signal Process.*, 53:pp. 44 – 53, Jan 2005.

[Sil05] H. Silverman, Y. Ying, J. Sachar, and W. Patterson, Performance of Real-Time Source-Location Estimators for a Large-Aperture Microphone Array. *IEEE Trans. Speech Audio Process.*, 13:pp. 593 – 606, July 2005.

[Smi87a] J. Smith and J. Abel, Closed-form least squares source location estimation from range-difference measurements. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-35(12):pp. 1661 – 1669, Dec. 1987.

[Smi87b] J. O. Smith and J. S. Abel, The spherical interpolation method of source localization. *IEEE Journal of Oceanic Engineering*, 12(1):pp. 246 – 252, 1987.

[So03] H. So and S. Hui, Constrained localization algorithm using tdoa measurements. *IEICE Trans. Fundamentals*, E86-A(12):pp. 3291 – 3293, December 2003.

[Sti08] R. Stiefelhagen, K. Bernardin, R. Bowers, R. Rose, M. Michel, and J. Garofolo, The CLEAR 2007 Evaluation. In R. Stiefelhagen, R. Bowers, and J. Fiscus, (Eds.) *Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR 2007 and RT 2007. Revised Selected Papers*, vol. 4625 of *Series: Lecture Notes in Computer Science*, pp. 3 – 34, Springer, 2008, `http://www.springerlink.com/content/l04734212898863t/?p=6c4518fc3614495eab7cb7c3710eba32\&pi=0`.

[Sto06] P. Stoica and J. Li, Source Localization from Range-Difference Measurements. *IEEE Signal Processing Mag.*, 23:pp. 63 – 66, November 2006.

[Sva97] P. Svaizer, M. Matassoni, and M. Omologo, Acoustic source location in a three-dimensional space using crosspower spectrum phase. In *Proc. Acoust., Speech, and Signal Process., ICASSP'97*, vol. 1, pp. 231 – 234, 1997.

[Tar] A. Tarantola, Inverse problem. From MathWorld–A Wolfram Web Resource, created by Eric W. Weisstein., `http://mathworld.wolfram.com/InverseProblem.html` last checked 9.7.2008.

[Tea07] P. Teal, Tracking wide-band targets having significant Doppler shift. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):pp. 489 – 497, 2007.

[Ter08] S. Tervo and T. Lokki, Interpolation methods for the SRP-PHAT algorithm. In *Proceedings of the 11th International Workshop on Acoustic Echo and Noise Control*, 2008.

[Tor84] D. J. Torrieri, Statistical theory of passive location systems. *IEEE Transactions on Aerospace and Electronic Systems*, 20(2):pp. 183 – 198, 1984.

[Tre02] H. V. Trees, *Detection, Estimation, and Modulation Theory*. Part IV, Optimum Array Processing, John Wiley & Sons, 2002.

[Val07] J. Valin, F. Michaud, and J. Rouat, Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems Journal (Elsevier)*, 55(3):pp. 216 – 228, 2007.

[Vee88] B. D. V. Veen and K. M. Buckley, Beamforming: A Versatile Approach to Spatial Filtering. *IEEE ASSP Magazine*, 5(2):pp. 4 – 24, 1988.

[Ver01] J. Vermaak and A. Blake, Nonlinear filtering for speaker tracking in noisy and reverberant environments. In *Int. Conf. Acoust., Speech, and Signal Process., (ICASSP'01)*., vol. 5, pp. 3021 – 3024, 2001.

[Vog07] C. Voges, P. Bauer, and T. Fingscheidt, A particle filtering algorithm for audiovisual speaker localisation. In *Workshop on Positioning, Navigation and Tracking (WPNC)*, 2007.

[War03] D. Ward, E. Lehmann, and R. Williamson, Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Transactions on Speech and Audio Processing*, 11(6):pp. 826 – 836, 2003.

[Wei83a] E. Weinstein and A. Weiss, Fundamental limitations in passive time delay estimation – part II: Wide-band systems. *IEEE Trans. Acoust., Speech and Signal Process.*, 31(2):pp. 472 – 486, 1983.

[Wei83b] A. Weiss and E. Weinstein, Fundamental limitations in passive time delay estimation – part I: Narrow-band systems. *IEEE Trans. Acoust., Speech and Signal Process.*, 31(2):pp. 472 – 486, 1983.

[Wei08] H.-W. Wei and S. Ye, Comments on "A Linear Closed-Form Algorithm for Source Localization from Time-Differences of Arrival". *IEEE Signal Processing Letters*, 15:p. 895, 2008.

[Yan03] S. Yan and Y. Ma, High-resolution broadband beamforming and detection methods with real data. *Acoust. Sci. & Tech.*, 25(1):pp. 73 – 76, 2003.

[Yan05] B. Yang and J. Scheuing, Cramér-Rao bound and optimum sensor array for source localization from time differences of arrival. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. (ICASSP '05)*, vol. 4, pp. 961 – 964, 2005.

[Yeg05] B. Yegnanarayana, S. Mahadeva Prasanna, R. Duraiswami, and D. Zotkin, Processing Reverberant Speech for Time-Delay Estimation. *IEEE Trans. on Speech and Audio Processing*, 13(6):pp. 1110 – 1118, 2005.

[YH96] J. Yli-Hietanen, K. Kalliojärvi, and J. Astola, Robust time-delay based angle of arrival estimation. In *Proceedings of the 1996 IEEE Nordic Signal Processing Symposium (NORSIG 96)*, pp. 219 – 222, 1996.

[YH99] J. Yli-Hietanen, K. Koppinen, and J. Astola, Time-delay selection for robust angle of arrival estimation. In *Proc. of IASTED Int. Conf., Signal and Image Processing (SIP'99)*, pp. 81 – 83, 1999.

[You84] D. Youn and V. Mathews, Adaptive realization of the maximum likelihood processor for time delay estimation. *IEEE Trans. Acoust., Speech, and Signal Proc.*, 32(4):p. 1984, August 1984.

[You86] D. Youn, S. Chiou, and V. Mathews, Adaptive phase transform processors for time delay estimation. *J. Acoust. Soc. America*, 80(1):pp. 188 – 194, 1986.

[Zha08] C. Zhang, D. Florêncio, and Z. Zhang, Why does PHAT work well in low noise, reverberative environments? In *Proc. Acoust., Speech, and Signal Processing (ICASSP'08)*, pp. 2565 – 2568, 2008.

[Zhe07] J. Zheng, K. Lui, and H. So, Accurate three-step algorithm for joint source position and propagation speed estimation. *Signal Processing*, 87(12):pp. 3096 – 3100, 2007.

[Zot04] D. Zotkin and R. Duraiswami, Accelerated speech source localization via a hierarchical search of steered response power. *IEEE Transactions on Speech and Audio Processing*, 12(5):pp. 499 – 508, 2004.

# Publication P1

Pasi Pertilä, Teemu Korhonen, and Ari Visa, Measurement Combination for Acoustic Source Localization in a Room Environment. *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2008, Article ID 278185, 14 pages, 2008. doi:10.1155/2008/278185.

# Publication P2

Pasi Pertilä, and Mikko Parviainen, Robust Speaker Localization in Meeting Room Domain. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'07)*, vol. 4, pages 497 – 500, 2007.

Publication P3

Pasi Pertilä, Mikko Parviainen, Teemu Korhonen, and Ari Visa, A Spatiotemporal Approach to Passive Sound Source Localization. In *Proceedings of International Symposium on Communications and Information Technologies 2004 (ISCIT'04)*, pages 1150–1154, 2004.

# Publication P4

Pasi Pertilä, Mikko Parviainen, Teemu Korhonen, and Ari Visa, Moving Sound Source Localization in Large Areas. In *2005 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2005)*, pages 745–748, 2005.

Publication P5

Pasi Pertilä, Array Steered Response Time-Alignment for Propagation Delay Compensation for Acoustic Localization. In *42nd Asilomar Conference on Signals, Systems, and Computers*, In Press, 2008

Appendix A

# Algorithm Descriptions

The Sampling Importance Resampling algorithm:

**function** $\mathcal{X}_t = \text{SIR}\{\ \mathcal{X}_{t-1}, \mathcal{R}^t_{[1:S]}\ \}$ ;
**for** $j = 1$ *to* $N_j$ **do**
    $\mathbf{r}^j_t \sim P(\mathbf{r}_t|\mathbf{r}^j_{t-1})$ ;
    Calculate $w^j_t = P(\mathcal{R}^t_{[1:S]}|\mathbf{r}^j_t)$ ;
**end**
Normalize weights, $w^{1:N_j}_t / \sum_{j=1}^{N_j} w^j_t$ ;
**return** $\mathcal{X}_t = \text{RESAMPLE}\{\ \mathcal{X}_t\ \}$ ;

Algorithm 1: SIR algorithm for particle filtering [Aru02].

The systematic resampling algorithm:

**function** $\mathcal{X}_t$ = RESAMPLE$\{\mathcal{X}_t\}$ ;
Copy states: $\mathbf{r}_t^{(1:N_j)} = \mathbf{r}_t^{1:N_j}$ ;
$c_1 = 0$ ;
**for** $j = 2$ *to* $N_j$ **do**
    Calculate CDF: $c_j = c_{j-1} + w_t^j$ ;
**end**
$j = 1$;
Draw a starting point: $u_1 \sim \mathcal{U}(0, N_j^{-1})$ ;
**for** $k = 1$ *to* $N_j$ **do**
    $u_k = u_1 + (k-1) \cdot N_j^{-1}$ ;
    **while** $u_k > c_j$ **do**
        $j = j + 1$ ;
    **end**
    Set new state $\mathbf{r}_t^{(k)} = \mathbf{r}_t^j$ ;
    Set weight $w_t^k = N_j^{-1}$ ;
**end**
**return** $\mathcal{X}_t = \{\mathbf{r}_t^{(j)}, w_t^j\}_{j=1}^{N_j}$ ;

Algorithm 2: The systematic resampling algorithm [Aru02].

The Average Distance Criterion (ADC) algorithm for DOA-based source localization.

**Data**: DOA measurement vectors $\hat{\mathbf{K}}$
**function** $\hat{\mathbf{r}}_{\text{ADC}}$ = LOCATE_ADC($\hat{\mathbf{K}}$, $\mathbf{P}$);
**for** *n = 1 to* $N_{3^+}$ **do**
    ADC($n$) $\leftarrow$ 0;
    $\hat{\mathbf{r}}_n \leftarrow$ calculate the LS solution of set $\Omega_n$ using (4.11);
    **for** *array* $k$ *= 1 to* $|\Omega_n|$ **do**
        `// Calculate array-to-source vector;`
        $\mathbf{k}_k \leftarrow (\hat{\mathbf{r}}_n - \mathbf{p}_k)/\|\mathbf{p}_k - \hat{\mathbf{r}}_n\|$;
        $\text{Proj}_{\hat{\mathbf{k}}_k}\mathbf{k}_k \leftarrow$ calculate using (4.4);
        $d_{k,n} \leftarrow$ calculate using (4.6);
        ADC($n$) $\leftarrow$ ADC($n$) + $d_{k,n}$ ;
    **end**
    `// Calculate average of distance criterion for set` $n$;
    ADC($n$) $\leftarrow$ ADC($n$) / $|\Omega_n|$;
**end**
$\tilde{n} = \underset{n}{\text{argmin}}\ \text{ADC}(n)$ (4.15);
$\hat{\mathbf{r}}_{\text{ADC}} = \hat{\mathbf{r}}_{\tilde{n}}$ (4.16);
**return** $\hat{\mathbf{r}}_{\text{ADC}}$;

Algorithm 3: ADC method for Speaker localization [P2].

Algorithm for DOA based source location estimation by using the propagation delay.

**Data**: DOA measurement vectors $\hat{\mathbf{K}}(\mathbf{t})$
**function** $\hat{\mathbf{r}}$ = DOA_LOCATE_PROP_DELAY($\hat{\mathbf{K}}(\mathbf{t})$,$\mathbf{P}$,$t$,$f_s$,$c$,L);
`// Initialize a uniform grid of search area locations;`
$\mathbf{g} \leftarrow$ uniform grid;
`// Initialize grid likelihoods;`
$P(\mathbf{g}) = \mathbf{0}$ ;
**for** *each* $\tilde{\mathbf{r}}$ *in* $\mathbf{g}$ **do**
    **for** *station* $i$ *= 1 to* $N_s$ **do**
        calculate hypothetical DOA: $\tilde{\mathbf{k}}_i = \tilde{\mathbf{r}} - \mathbf{p}_i$ ;
        `// Propagation delay to hypothetical source position;`
        $\Delta t_i \leftarrow$ (4.19) ;
        `// retrieve DOA measurement` $\hat{\mathbf{k}}_i(t + \Delta t_i)$ ;
        $e_i \leftarrow$ dot product of DOA measurement and hypothesis (4.21);
        `// update spatial likelihood;`
        $P(\tilde{\mathbf{r}}) = P(\tilde{\mathbf{r}}) + e_i$ ;
    **end**
**end**
**return** $\hat{\mathbf{r}} = \underset{\tilde{\mathbf{r}}}{\operatorname{argmax}} P(\tilde{\mathbf{r}})$ ;

Algorithm 4: DOA vector-based localization with propagation delay [P3].

Algorithm for calculating the far-field sound source location estimate from TDE-likelihood based observations.

**Data**: Signals $\mathbf{X}_{[1:N_s]}$
**function** $\hat{\mathbf{r}}$=TDE_FAR_LOCATE($\mathbf{X}_{[1:N_s]}, \mathbf{M}_{[1:N_s]}, f_s, t$,L);
`// Initialize a uniform grid;`
$\mathbf{g} \leftarrow$ uniform grid;
`// Clear source position likelihood grid;`
Clear $P(\mathbf{g})$;
**for** *Array* $i = 1$ *to* $N_s$ **do**
    Clear $P_i(\mathbf{g})$;
    **for** *Microphone* $k = 1$ *to* $N_i$ **do**
        $X_k^i = \mathrm{DFT}\{\mathbf{x}_k^i(t)\}$;
    **end**
    **for** *Microphone* $l = 1$ *to* $N_i - 1$ **do**
        **for** *Microphone* $k = l$ *to* $N_i$ **do**
            $p = \{l, k\}$ ;
            `// Correlate signals` (2.21);
            $\mathcal{R}_p(\tau) = \mathrm{IDFT}\{\Psi_p(f) X_l^i(f) X_k^{i*}(f)\}$ ;
            **for** *each location* $\mathbf{r}$ *in grid* $\mathbf{g}$ **do**
                `// Hypothetical station to source vector;`
                $\mathbf{k} = \mathbf{r} - \mathbf{p}_i$;
                `// TDOA of source direction` (4.23);
                $\Delta\tau_p(\mathbf{k}) = c^{-1} \cdot (\mathbf{m}_k - \mathbf{m}_l)^{\mathrm{T}} \mathbf{k} / \|\mathbf{k}\|$ ;
                `// Combine likelihoods for position` $\mathbf{r}$;
                $P_i(\mathbf{r}) = P_i(\mathbf{r}) \otimes \mathcal{R}_p(\lceil \Delta\tau_p(\mathbf{k}) \cdot f_s \rfloor)$ ;
            **end**
        **end**
    **end**
    `// Combine array likelihoods;`
    $P(\mathbf{r}) = P(\mathbf{r}) \oplus P_i(\mathbf{r})$;
**end**
**return**: $\hat{\mathbf{r}} = \underset{\mathbf{r}}{\operatorname{argmax}}\, P(\mathbf{r})$;

Algorithm 5: TDE-based directional likelihood for far-field source localization.

Algorithm for calculating the far-field sound source location estimate from TDE-likelihood based observations using propagation delay.

**Data**: Signals $\mathbf{X}_{[1:N_s]}(\mathbf{t})$
**function** $\hat{\mathbf{r}}$=TDE_FAR_PROP_LOC($\mathbf{X}_{[1:N_s]}(\mathbf{t}), \mathbf{M}_{[1:N_s]}, f_s$,t,L,$c$);
$\mathbf{g} \leftarrow$ grid of search area locations ;
Reset $P(\mathbf{g})$ ;
**for** *Array* $i$ *= 1 to* $N_s$ **do**
    Reset $P_i(\mathbf{g})$;
    **for** *Microphone* $l$ *= 1 to* $N_i - 1$ **do**
        **for** *Microphone* $k$ *=* $l$ *to* $N_i$ **do**
            $p = \{l, k\}$ ;
            **for** $\Delta t_i = 0$ *to* $\Delta t_{\max}$ **do**
                $X_l^i = \mathrm{DFT}\{\mathbf{x}_l^i(t + \Delta t_i)\}$;
                $X_k^i = \mathrm{DFT}\{\mathbf{x}_k^i(t + \Delta t_i)\}$ ;
                `// Correlate signals` (2.21)
                $\mathcal{R}_p^{t+\Delta t_i}(\tau) = \mathrm{IDFT}\{\Psi_p^i(f) X_l^i(f) X_k^{i*}(f)\}$ ;
            **end**
            **for** *each location* $\mathbf{r}$ *in grid* $\mathbf{g}$ **do**
                `// Hypothetical station to source vector;`
                $\mathbf{k}_i = \mathbf{r} - \mathbf{p}_i$ ;
                `// Calculate the propagation delay in frames;`
                $\Delta t_i \leftarrow$ (4.19);
                `// Calculate time difference of arrival;`
                $\Delta\tau_p(\mathbf{k}_i) = c^{-1} \cdot (\mathbf{m}_k - \mathbf{m}_l)^{\mathrm{T}} \mathbf{k}_i / \|\mathbf{k}_i\|$ ;
                `// Combine pairwise likelihoods for position;`
                $P_i(\mathbf{r}) = P_i(\mathbf{r}) \otimes \mathcal{R}_p^{t+\Delta t_i}(\lceil \Delta\tau_p(\mathbf{k}_i) \cdot f_s \rfloor)$;
            **end**
        **end**
    **end**
    `// Combine array likelihoods for position;`
    $P(\mathbf{r}) = P(\mathbf{r}) \oplus P_i(\mathbf{r})$;
**end**
**return**: $\hat{\mathbf{r}} = \underset{\mathbf{r}}{\operatorname{argmax}}\, P(\mathbf{r})$;

Algorithm 6: TDE-based directional likelihood for far-field source localization with propagation delay according to [P5].

Appendix B

# Simulation Setup

The geometry of the microphone setup discussed in Section 2.4 is listed in Table B.1.

Table B.1: Microphone coordinates.

| Mic. | x-coord. | y-coord | z-coord |
|---|---|---|---|
| 1 | 0.567 | 0.050 | 1.500 |
| 2 | 1.700 | 0.050 | 1.500 |
| 3 | 2.833 | 0.050 | 1.500 |
| 4 | 3.966 | 0.050 | 1.500 |
| 5 | 0.567 | 0.050 | 1.900 |
| 6 | 1.700 | 0.050 | 1.900 |
| 7 | 2.833 | 0.050 | 1.900 |
| 8 | 3.966 | 0.050 | 1.900 |
| 9 | 0.567 | 3.907 | 1.500 |
| 10 | 1.700 | 3.907 | 1.500 |
| 11 | 2.833 | 3.907 | 1.500 |
| 12 | 3.966 | 3.907 | 1.500 |
| 13 | 0.567 | 3.907 | 1.900 |
| 14 | 1.700 | 3.907 | 1.900 |
| 15 | 2.833 | 3.907 | 1.900 |
| 16 | 3.966 | 3.907 | 1.900 |
| 17 | 0.050 | 0.495 | 1.500 |
| 18 | 0.050 | 1.484 | 1.500 |
| 19 | 0.050 | 2.473 | 1.500 |
| 20 | 0.050 | 3.462 | 1.500 |
| 21 | 0.050 | 0.495 | 1.900 |
| 22 | 0.050 | 1.484 | 1.900 |
| 23 | 0.050 | 2.473 | 1.900 |
| 24 | 0.050 | 3.462 | 1.900 |
| 25 | 4.482 | 0.495 | 1.500 |
| 26 | 4.482 | 1.484 | 1.500 |
| 27 | 4.482 | 2.473 | 1.500 |
| 28 | 4.482 | 3.462 | 1.500 |
| 29 | 4.482 | 0.495 | 1.900 |
| 30 | 4.482 | 1.484 | 1.900 |
| 31 | 4.482 | 2.473 | 1.900 |
| 32 | 4.482 | 3.462 | 1.900 |

Appendix C

# Simulation Results

This section presents the simulation results described in Section 3.8.3.

Table C.1: Accuracy [mm] of non-missed estimates for method 1 for SNR values 30 to -10 dB with T60 values between 0 and 0.9 s.

| | **Reverberation time $T_{60}$** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNR** | **0.09** | **0.11** | **0.13** | **0.15** | **0.18** | **0.23** | **0.27** | **0.33** | **0.37** | **0.43** | **0.51** | **0.56** | **0.63** | **0.83** |
| **30** | 1.0 | 1.0 | 26.0 | 165.0 | 362.0 | . | . | . | . | . | . | . | . | . |
| **28** | 1.0 | 1.0 | 37.0 | 198.0 | 374.0 | . | . | . | . | . | . | . | . | . |
| **26** | 1.0 | 1.0 | 58.0 | 232.0 | 405.0 | . | . | . | . | . | . | . | . | . |
| **24** | 1.0 | 2.0 | 86.0 | 271.0 | 417.0 | . | . | . | . | . | . | . | . | . |
| **22** | 2.0 | 3.0 | 125.0 | 295.0 | 443.0 | . | . | . | . | . | . | . | . | . |
| **20** | 2.0 | 5.0 | 182.0 | 330.0 | . | . | . | . | . | . | . | . | . | . |
| **18** | 2.0 | 10.0 | 232.0 | 368.0 | . | . | . | . | . | . | . | . | . | . |
| **16** | 4.0 | 23.0 | 278.0 | 428.0 | . | . | . | . | . | . | . | . | . | . |
| **14** | 8.0 | 53.0 | 321.0 | 485.0 | . | . | . | . | . | . | . | . | . | . |
| **12** | 21.0 | 102.0 | 372.0 | . | . | . | . | . | . | . | . | . | . | . |
| **10** | 49.0 | 168.0 | 434.0 | . | . | . | . | . | . | . | . | . | . | . |
| **8** | 100.0 | 244.0 | 481.0 | . | . | . | . | . | . | . | . | . | . | . |
| **6** | 168.0 | 323.0 | . | . | . | . | . | . | . | . | . | . | . | . |
| **4** | 261.0 | 375.0 | . | . | . | . | . | . | . | . | . | . | . | . |
| **2** | 333.0 | 380.0 | . | . | . | . | . | . | . | . | . | . | . | . |
| **0** | 380.0 | 415.0 | . | . | . | . | . | . | . | . | . | . | . | . |
| **2** | 346.0 | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **4** | 459.0 | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **6** | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **8** | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **10** | . | . | . | . | . | . | . | . | . | . | . | . | . | . |

Table C.2: Accuracy [mm] of non-missed estimates for method 2 for SNR values 30 to -10 dB with T60 values between 0 and 0.9 s.

| | **Reverberation time $T_{60}$** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNR** | **0.09** | **0.11** | **0.13** | **0.15** | **0.18** | **0.23** | **0.27** | **0.33** | **0.37** | **0.43** | **0.51** | **0.56** | **0.63** | **0.83** |
| **30** | 14.0 | 12.0 | 14.0 | 15.0 | 17.0 | 15.0 | 19.0 | 19.0 | 25.0 | 37.0 | 30.0 | 53.0 | 68.0 | 67.0 |
| **28** | 13.0 | 14.0 | 13.0 | 14.0 | 14.0 | 16.0 | 18.0 | 22.0 | 28.0 | 29.0 | 31.0 | 69.0 | 48.0 | 70.0 |
| **26** | 13.0 | 14.0 | 14.0 | 14.0 | 15.0 | 16.0 | 18.0 | 35.0 | 33.0 | 29.0 | 44.0 | 42.0 | 63.0 | 94.0 |
| **24** | 13.0 | 14.0 | 14.0 | 14.0 | 15.0 | 17.0 | 23.0 | 26.0 | 32.0 | 32.0 | 46.0 | 51.0 | 47.0 | 84.0 |
| **22** | 13.0 | 14.0 | 14.0 | 15.0 | 15.0 | 18.0 | 20.0 | 27.0 | 28.0 | 32.0 | 38.0 | 59.0 | 59.0 | 96.0 |
| **20** | 13.0 | 13.0 | 14.0 | 16.0 | 17.0 | 22.0 | 23.0 | 29.0 | 35.0 | 36.0 | 43.0 | 68.0 | 99.0 | 115.0 |
| **18** | 14.0 | 14.0 | 15.0 | 15.0 | 16.0 | 22.0 | 27.0 | 31.0 | 37.0 | 45.0 | 56.0 | 65.0 | 85.0 | 139.0 |
| **16** | 13.0 | 14.0 | 15.0 | 15.0 | 17.0 | 22.0 | 25.0 | 32.0 | 34.0 | 48.0 | 63.0 | 65.0 | 115.0 | 130.0 |
| **14** | 13.0 | 15.0 | 15.0 | 15.0 | 18.0 | 25.0 | 29.0 | 37.0 | 43.0 | 56.0 | 80.0 | 79.0 | 113.0 | 180.0 |
| **12** | 14.0 | 16.0 | 17.0 | 19.0 | 20.0 | 25.0 | 33.0 | 37.0 | 45.0 | 67.0 | 79.0 | 117.0 | 140.0 | 198.0 |
| **10** | 14.0 | 13.0 | 18.0 | 20.0 | 22.0 | 29.0 | 35.0 | 53.0 | 54.0 | 85.0 | 114.0 | 121.0 | 160.0 | 217.0 |
| **8** | 14.0 | 16.0 | 18.0 | 21.0 | 24.0 | 31.0 | 39.0 | 59.0 | 61.0 | 91.0 | 125.0 | 156.0 | 158.0 | 280.0 |
| **6** | 15.0 | 18.0 | 18.0 | 21.0 | 27.0 | 35.0 | 46.0 | 61.0 | 84.0 | 104.0 | 149.0 | 173.0 | 217.0 | 257.0 |
| **4** | 16.0 | 18.0 | 19.0 | 24.0 | 32.0 | 42.0 | 55.0 | 74.0 | 106.0 | 122.0 | 191.0 | 202.0 | 203.0 | 255.0 |
| **2** | 18.0 | 18.0 | 22.0 | 28.0 | 35.0 | 47.0 | 62.0 | 86.0 | 118.0 | 138.0 | 197.0 | 284.0 | 260.0 | 305.0 |
| **0** | 20.0 | 19.0 | 28.0 | 29.0 | 40.0 | 61.0 | 77.0 | 104.0 | 140.0 | 192.0 | 239.0 | 254.0 | 269.0 | 299.0 |
| **-2** | 20.0 | 24.0 | 28.0 | 37.0 | 52.0 | 72.0 | 95.0 | 140.0 | 193.0 | 215.0 | 276.0 | 315.0 | 286.0 | 359.0 |
| **-4** | 22.0 | 27.0 | 34.0 | 39.0 | 58.0 | 87.0 | 117.0 | 170.0 | 212.0 | 235.0 | 284.0 | 328.0 | 308.0 | 404.0 |
| **-6** | 25.0 | 33.0 | 42.0 | 49.0 | 65.0 | 93.0 | 143.0 | 208.0 | 253.0 | 271.0 | 266.0 | 351.0 | 359.0 | 374.0 |
| **-8** | 31.0 | 34.0 | 47.0 | 62.0 | 74.0 | 136.0 | 184.0 | 268.0 | 289.0 | 332.0 | 342.0 | 378.0 | 367.0 | 339.0 |
| **-10** | 33.0 | 46.0 | 58.0 | 68.0 | 100.0 | 161.0 | 219.0 | 271.0 | 287.0 | 297.0 | 323.0 | 382.0 | 421.0 | 418.0 |

Appendix D

# Concepts Related to Random Processes

This section is based on [Pee93] and defines properties related to random variables and random sequences. The expected value of any random variable $X$ is defined by:

$$E[X] = \int_{-\infty}^{\infty} x f_x(x) dx, \tag{D.1}$$

where $f_x(x)$ is the probability density function.

The expected value of a discrete random variable $X$ having discrete values $x_i$ occurring with probabilities $P(x_i)$ is defined as

$$E[x] = \sum_{i=1}^{N} x_i P(x_i). \tag{D.2}$$

A random process $X(t)$ is *wide-sense stationary (WSS) process* if

$$E[X(t)] = \bar{X} = \text{constant} \tag{D.3}$$

$$E[X(t)X(t+\tau)] = \mathcal{R}_{X,X}(\tau), \tag{D.4}$$

where $\tau$ is the difference between absolute sequence time, and $\mathcal{R}_{X,X}(\tau)$ is called autocorrelation, and it does not depend on absolute time. Two random processes $X$ and $Y$ are jointly wide-sense stationary if they satisfy (D.3)– (D.4) and if their cross correlation function is a function of time difference $\tau$ (not absolute times):

$$\mathcal{R}_{X,Y}(\tau) = E[X(t)Y(t+\tau)]. \tag{D.5}$$

The time autocorrelation $\mathscr{R}_{x,x}(t)$ of a sample function $x(t)$ and the time average $\bar{x}$ are defined

$$\bar{x} = A[x(t)] = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t) dt \tag{D.6}$$

$$\begin{aligned} \mathscr{R}_{x,x}(\tau) =& A[x(t)x(t+\tau)] \\ =& \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)x(t+\tau) dt \end{aligned} \tag{D.7}$$

Values $\bar{x}$ and $\mathscr{R}_{x,x}(\tau)$ are themselves random numbers, and their expectation values are

$$E[\bar{x}] = \bar{X} \tag{D.8}$$

$$E[\mathscr{R}_{x,x}(\tau)] = \mathcal{R}_{X,X}(\tau). \tag{D.9}$$

The *ergodic theorem* allows the time average $\bar{x}$ to be equal to the statistical average $\bar{X}$ and the time autocorrelation $\mathscr{R}_{x,x}(\tau)$ to be equal to the statistical autocorrelation $\mathcal{R}_{X,X}(\tau)$. Such a process in called an *ergodic process*. Two random processes $X$ and $Y$ are jointly ergodic if both processes are individually ergodic and their time cross correlation function equals their statistical cross correlation function

$$\mathscr{R}_{x,y}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)y(t+\tau)dt = \mathcal{R}_{X,Y}(\tau). \tag{D.10}$$

Another important concept is the *cross covariance* function of two processes $X(t)$ and $Y(t)$ defined for a random process as

$$C_{X,Y}(t, t+\tau) = E[\{X(t) - E[X(t)]\}\{Y(t+\tau) - E[Y(t+\tau)]\}] \tag{D.11}$$

and covariance of the wide-sense stationary process does not depend on absolute time and is written

$$C_{X,Y}(\tau) = \mathcal{R}_{X,Y}(\tau) - \bar{X}\bar{Y} \tag{D.12}$$

The measurement frames of duration $T_w$ are available for estimating the cross correlation between channels. The statistical cross correlation is approximated from the measured data frame:

$$\mathcal{R}_{x,y}(\tau) = \frac{1}{T_w} \int_{-T_w/2}^{T_w/2} x(t)y(t+\tau)dt \approx \mathscr{R}_{x,y}(\tau) = \mathcal{R}_{X,Y}(\tau). \tag{D.13}$$