Sound Localization and Separation for Teleconferencing Systems

Johannes Feldmaier, Martin Rothbucher, Klaus Diepold

Technical Report

Sound Localization and Separation for Teleconferencing Systems

Johannes Feldmaier, Martin Rothbucher, Klaus Diepold

March 30, 2014

Lehrstuhl für Datenverarbeitung
Technische Universität München

Submitted on March 30, 2014
by Johannes Feldmaier, Martin Rothbucher, Klaus Diepold

Supervised by Prof. Dr.-Ing. K. Diepold

Fakultät für Elektrotechnik und Informationstechnik
Technische Universität München

Abstract

Nowadays, a table-top microphone system, which is used for teleconferencing, is usually installed in conferencing rooms. Using such conference phones, all active speakers are recorded simultaneously and a mixture of the speakers is transmitted. For convenience, it would be great to have a device which acquires remotely a high-quality speech signal for every single participant. So this thesis regards an appliance for remote acquisition of speech signals in common office environments. A combination of a microphone array and signal processing has been applied to localize and separate the speech contributions of the participants. Audio based localization is performed through a Steered Response Power Beamformer and smoothed through particle filtering. The separation process is based on Geometric Source Separation which joins the benefits of Beamforming and Blind Source Separation algorithms.

Performance evaluations have shown that separation quality depends strongly on localization stability and accuracy. The audio-based localization detects the sound sources with a success rate of more than 80 percent the correct position with an accuracy of 4 degree, in an office environment. With these localization data a continuous separation of the speakers can be performed with a mean signal-to-interference ratio of more than 27 dB.

All obtained results presented in this thesis show that reliable remote acquisition of speech signals is possible. In future, devices separating different speakers are entirely conceivable.

Normalerweise wird die Sprache in Telefonkonferenzen mittels Headsets oder speziellen Konferenztelefonen aufgezeichnet. Derzeit kommerziell erhältliche Geräte nehmen alle Sprachsignale gleichzeitig auf und übertragen daher eine Mischung aller aktiven Sprecher. Es wäre wünschenswert ein Gerät zu haben, welches den Komfort eines Tischgeräts mit der Sprachqualität von separaten Ansteckmikrofonen verbindet.

Die vorliegende Arbeit untersucht daher Ansätze zur gerichteten Aufnahme von unterschiedlichen Sprechern in gewöhnlichen Büroräumen. Dazu wird eine Kombination aus Mikrofonarray und Signalverarbeitung verwendet, um die Sprache der Konferenzteilnehmer getrennt aufzunehmen. Die Sprach-Separierung basiert auf einem Algorithmus der *Blind Source Separation* und *Beamforming* vereint. Die dafür notwendigen Lokalisationsdaten der Sprecher werden mittels eines *Steered Response Power* Verfahrens mit anschließender Partikelfilterung ermittelt.

Anschließende Experimente und deren Auswertung haben gezeigt, dass die Qualität der getrennten Aufnahmen stark von der Lokalisierung abhängen. Die rein audiobasierte Ortung der Sprecher erzielt dabei in normalen Büroräumen eine Erkennungsrate von mehr als 80% bei einer Genauigkeit von 4 Grad. Damit kann eine kontinuierliche Separierung mit einem durchschnittlichen Signal-Rausch-Verhältnis von mehr als 27 dB durchgeführt werden.

Alle erzielten Ergebnisse belegen, dass eine verlässliche separierte Aufnahme unterschiedlicher Sprecher durch ein neuartiges Konferenztelefon möglich wäre.

Contents

1. Introduction

Teleconferencing is great because it allows people on the opposite side of the world to talk to each other direct from desk to desk in their own offices. There is no need to do business travels for short meetings. Therefore, teleconferencing saves a lot of money and time in companies.

At a time of rising energy prices and growing environmental awareness, this means to companies that the quality and productivity of teleconferences should be as high as in real meetings. In recent years, the teleconferencing technology has made great progress in quality features like noise reduction and speech processing. New innovative functionalities are rare. There aren't any really new ways to communicate with each other. Traditionally the participants speak into microphones and receive the counterpart stations over headphones or speakers. So in this project, it is tried to invent a new way to interact with the communication system. In preliminary work [17] a system was build to present the participants spatially distributed around the listeners head using ordinary headsets. But, the positions to generate this 3D effect applying Head Related Transfer Functions (HRTFs) were manually obtained. This kind of presentation is useful in conferences where single speakers are connected to each other. But in conferences connecting a meeting room with various speakers to remote single speakers or other conferencing rooms, the individual positions must be detected automatically. This is one aspect studied in this thesis. Another aspect discussed in this thesis is the recording of the speakers in the conference room. The convenience of a teleconference would be diminished if the participants have to use tethered lapel microphones. Recording the voice of each active speaker with a hands-free device would therefore be a convenient new way to enhance a conference system.

Thus, the focus of this thesis lies on the localization and separation of active speakers in a teleconference situation.

1.1. Motivation for Study

Immersive communication is a wide-ranging research topic. Every day people are faced with huge amounts of information of their surrounding world. They have to acquire, select, and process all these informations. The human body provides the ability to decipher, separate and emphasize certain informations within the perceived signals. Depending on the acquired informations the human has to make decisions and execute appropriate reactions.

Scientists all over the world are inspired from nature and try to mimic it. A similar problem should also be tackled in this thesis. Basically the problem can be abstracted from the classic scenario of a cocktail party (also called the Cocktail Party Effect). During a cocktail party many people talk simultaneously on different locations in small groups. Each person of a group can listen to the actual talking person without disruption of other speakers or groups. It's difficult to transfer this ability of selecting and distinguish a speaker into a technical device. Previously

published works have their focus on one aspect of the problem, meaning either localization or separation. Most of the current research is done in the field of robotic. In the context of robotics the applicability of the developed algorithms are not considered for the use in teleconferencing environments. These robotic based approaches are investigated in the scope of this thesis on their strengths and weaknesses with respect to teleconferencing.

1.2. System Overview

This section introduces an immersive teleconferencing system, as developed at the *Institute for Data Processing*, and the different components it consists of. Two of them, the localization process and the separation algorithm, will be examined more detailed in this thesis.

The system considered in this thesis has a modular design. Its architecture can be simplified to a classic communication system, composed of a sender, transmitter and receiver. The sender records a speech signal, the transmitter transfers it to the receiver, the receiver reproduces the signal. In a teleconference scenario, the sender is a conference telephone with the correspondent hardware to record a speech signal. The transmitter is a server structure with the functionality of receiving and sending audio signals. The receiver is a telephone or Voice-over-IP (VoIP) client with the respective hardware to reproduce an audio signal. To evolve this system into a new generation immersive communication system, every block needs to be enhanced and extended. However, the architecture keeps its tripartite and serially ordered design, so it can be easily extended and is compatible to current communication systems.

Every block (sender, transmitter, receiver) contains several modules. Every module adds a certain functionality to the system. So the whole system is not only able to communicate with other external systems, but it can easily be extended by additional modules. Every module has defined interfaces and can be tested and developed independently. In the described scenario, the conference telephone will be extended by speaker recognition, localization and separation features. The server will have the ability to recognize speech, store it centrally and process all the incoming sounds for 3D sound rendering. If compatible, the receiver will be extended by a client software to easily define the spatial position in the virtual environment.

Altogether, the different modules, as described in the following sections, turn a classic teleconferencing system into an immersive audio conferencing system.

Speaker recording and localization

A big issue for conference telephones placed on a table in a room with multiple conference participants is the distant acquisition of the speech signal. It is very vulnerable to interference from concurrent sound sources and noise distortion through reflection. The audio recording module needs to be able to handle all types of noise occurring. Different types of noise can be defined: Additive noise, echo, reverberation, and competing sound sources.

To control noise, reverberation, and competing speech, the speakers must robustly be localized and recorded separately. Both tasks are more precisely considered within this thesis.

Speaker recognition

The speaker recognition is part of the sender block, to enable a separate transmission of every speaker on its own channel, although only one device in the conference room is used for recording. The input of the speaker recognition module ideally is a single channel speech recording or stream. This audio input, with its preprocessing, separation and filtering, is as much as possible free of noise, interference, echo and reverberation.

The result of the speaker recognition is the classification of a speech signal to a defined speaker name and the assignment to a certain output channel corresponding to the speakers name.

The development of this component is not covered by this thesis, but is currently developed at the *Institute for Data Processing*.

Speech recognition

The speech recognition is part of the transmission block, so mainly a module of the server handling the teleconference. It performs speech recognition on each incoming voice stream. As mentioned above, each voice channel is assigned to an individual speaker and therefore, it is possible to create autonomously a transcript of the conference. The quality of this transcript strongly depends on the used speech recognition module. Nowadays, there are commercial solutions available on the market basing on huge training databases. With these professional speech recognizers it is possible to achieve fair recognition rates.

Because of the high development effort generating such large training sets the speech recognition module will be considered as external module.

Speech transmission

Transmissions are handled by a central server, no peer-to-peer technique is used. This is beneficial to create different mixtures for the receiving devices. The mixer processes all incoming signals and creates according to device capabilities appropriate streams presenting the conference in an optimal manner. That means, for single channel devices a mono signal is created and for multichannel room solutions a spatially distributed signal is generated. For this purpose the central transmission component needs efficient sound rendering and management techniques. Additionally, after the mixing process, the transmission component picks an appropriate compression codec according to the receiving device, so that the signals are transmitted with a minimum delay and the highest possible quality, while acquiring a minimal bandwidth.

A further aspect of this component is to ensure the downward compatibility between different devices. For example, it should be possible to connect mobile phones with a VoIP client presenting 3D sound and the new recording device with its separation feature to a conventional desktop phone. Therefore, every device should be handled according to its specific capabilities delivering the best possible conferencing experience.

A first version of this component was already developed, and is currently tested at the institute.

Speech synthesis

As already mentioned, the system is developed to connect conventional telephones and VoIP clients with the new developed devices recording and presenting spatial distributed speakers. It creates for each participating device an appropriate sound signal. In order to exploit the full potential of separately recorded speakers, a client presenting three dimensional sound is needed. Therefore, in previous work [17] a VoIP software client was developed which presents each teleconference participant spatially distributed over a stereo headset. This client uses a set of HRTFs to generate the spatial sound signal.

In future, additional sound synthesis techniques will be investigated to improve the quality of the headphone based sound synthesis and to bring this lifelike multi-party conferencing to full room solutions using multichannel audio systems and techniques like wave field synthesis (WFS).

1.3. Scope of this Thesis

This thesis, as a part of a project at the *Institute for Data Processing*, researches new technologies for teleconferencing systems. The thesis will cover the recording part of the system described above. So an appropriate microphone array needs to be found, that is able to localize and record speakers sitting around a conference table. Therefore different geometrical configurations will be examined and constructed. Furthermore, a fast localization and separation algorithm needs to be selected and implemented. The final system will be tested in an anechoic chamber and under real conditions. For these tests, a complete test sequence will be developed and the results will be evaluated.

In Figure 1.1 an overview of the smart "recording device" as a part of the system proposed in the previous is depicted. The red-coloured parts are covered by this thesis and will be studied further.

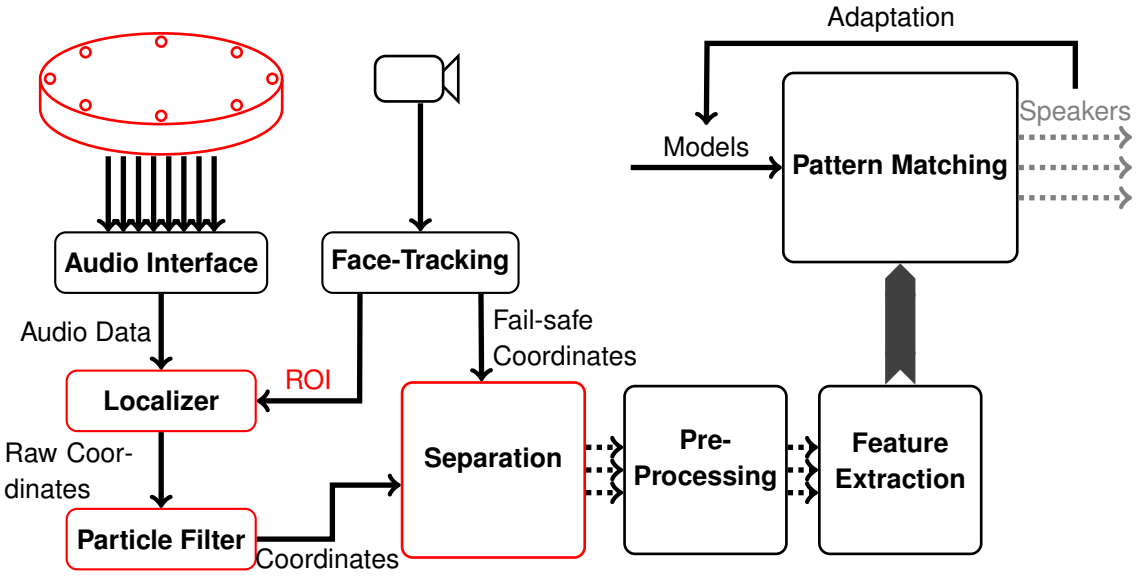


Figure 1.1.: Overview of the conference recording device

2. Background

The speech mixture of two or more speakers is the initial point of this study. In meeting situations, these mixtures arise when speakers change or if a speaker interjects a brief comment. It can be said, that 10 to 15% of words contain overlapping speech during a meeting [43]. These overlapping speech segments distinguish between the different voices, also the speaker positions in the room around the "recording device". Today most commercial conference phones use an arrangement of several microphones, or a so called microphone array, to record the participants. To understand the functional principles of microphone arrays, this section explains some basics of the theory of wave propagation, continuous apertures, and finally discrete sensor arrays.

2.1. Wave Propagation

Acoustic signals can be categorized into different sound fields according to their statistical properties. These statistical properties are mainly influenced by the room acoustics. In reverberant rooms sound waves are reflected by walls and furniture. As a consequence the impinging microphone signals can be divided into direct and indirect components. The direct sound travels directly from the sound source to the microphones, the indirect component results due to multipath propagation, which can still be splitted into early arriving echoes and a diffuse signal of later arriving components. A complete description of the sound field is almost impossible. There are methods like sound ray tracing or the measurement of the Room Impulse Response (RIR) to estimate the sound wave propagation. In general these algorithms are computational intense and complex compared to the statistical description of a sound field.

In this statistical model of the room acoustic, the sound field is described as spatial regions with corresponding acoustic parameters. Two important parameters are reverberation time which characterizes the duration how long acoustic energy remains in a room, and the critical distance stating the distance at which the energy value of the direct signal is equal to the reflected signal.

The reverberation time is given by the *Sabine equation* and was developed in the late 1890s by *Wallace C. Sabine* [11]. It establishes a relation between the RT_{60} (the reverberation time) of a room, its volume, and its total absorption:

$$RT_{60} = \frac{4ln10^6}{c}\frac{V}{Sa} \tag{2.1}$$

where c is the speed of sound, V is the volume of the room in cubic meters, S the total surface area of the room in square meters, and a is the average absorption coefficient of the surfaces.

This equation is useful for simulating sound propagation in rooms with given reverberation times. In this thesis the separation algorithm is partly based on this relation.

The other interesting parameter, the critical distance, is important if speakers are recorded in small rooms with high reverberation. Than the proposed recording device detects the reflection as an additional source. The critical distance is calculated by

$$d_c = \sqrt{\frac{V}{100\pi R T_{60}}}. \tag{2.2}$$

In general, it can be said that high reverberation environments are a major problem for localization and separation algorithms.

Another aspect of wave propagation besides reverberation is intensity distribution in time of sound waves. Sound waves follow the inverse-square law given by

$$I = \frac{P}{4\pi r^2}, \tag{2.3}$$

where I is the sound intensity at the surface of the sphere, and r is the radius of the sphere, and P is the net power radiated by the source. The equation says, that the energy of a point source is distributed as spherical wave, which has to be considered in the development of the recoding device. Because, in most cases the assumption of plane wave fronts is made, which is not always given. For that reason, the far-field assumption is made, so that the wave fronts could be assumed as plane. This is valid for speaker distances r greater than:

$$|r| > \frac{2L^2}{\lambda} \tag{2.4}$$

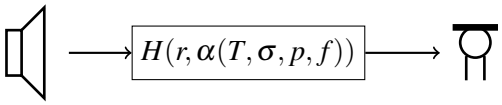
where L represents the complete aperture size (e.g. array length) and λ the wavelength of a specific frequency.

Using equation 2.4 for voice signals with a highest frequency of 3400 Hz and a microphone array with a diameter of 0.24 meters this means far field can be assumed if $r > 1.14m$. So, for an array of 0.24 m diameter, the minimal recording distance corresponds to r.

With the above formal measures a description of the different sound field types is possible. Distinction must be made between diffuse sound fields and directed sources . A diffuse sound field cannot be localized and occurs behind the critical distance, at the points where the direct energy of the sound source and the reflected energy is the same. On the other hand directed sources are recorded within the critical distance and can therefore be localized. A recording device also has to distinguish between noise and speech sources. This can be done by analysing the spectral characteristics and post-filters.

Considering a directed source, the transmission of a speech signal to a microphone can be described by a linear transfer function. A full treatment of this subject is beyond the scope of this thesis but the subject has already been well-researched and the results summarized in ISO 9613-1 (1993). The essence of this ISO standard is that the transfer function of sound is exponentially dependent on the distance r of the sound source, the absorption coefficient α

of the room, which is a function of temperature T, humidity σ, atmospheric pressure p, and frequency f. This can be modelled as:



Besides this influences on the sound wave each signal in a room travels a specific distance to the microphone in a specific amount of time. This duration depends on the speed of sound:

$$c_{air} = (331.3 + (0.606^{\circ} C^{-1} \cdot \theta)) \frac{m}{s} \tag{2.5}$$

where θ is the temperature in degrees Celsius (under the assumption of dry air, 0% humidity). Based on this relation, microphone arrays exploit these different points in time of arriving sound waves to localize or emphasize specific signals.

2.2. Microphone Arrays

It is well known that humans can localize and perceptually segregate different sound sources of their immediate vicinity with only two ears. This human ability is copied to technical approaches like binaural processing. In the book *Computational auditory scene analysis: Principles, algorithms, and applications* by *Richard M. Stern, Guy J. Brown and DeLiang Wang* [52] the various approaches are described. Most technical adaptions of two channel array approaches are limited in the detection of either elevation or azimuth differences and suffering under front-back ambiguities.
So a complete different approach to localize sound sources in contrast to binaural approaches are microphone arrays. Since the computational complexity of array processing is mostly less than of binaural approaches, for this thesis a microphone array was selected.

Microphone arrays are discrete passive apertures. The term *aperture* is used to refer to a spatial region that receives or transmits waves [30]. Receiving apertures are refereed as *passive*, transmitting ones as active. For example, in acoustics, a passive aperture is an electroacoustic transducer that converts acoustic signals into electrical signals (microphone). And finally, *discrete* means that the ideal continuous aperture is transferred to a sampled version. This sampling is necessary because there aren't any continuous sensors. Thus a discrete aperture consists of several discrete sensors building an array, in case of acoustic sensors it is called a microphone array. But the arrangement and number of microphones determines the directivity and frequency dependence of the whole array.

For each application a trade-off between number of microphones, size of the array, and manageable bandwidth has to be found. Localization quality is strongly dependent on the number of microphones. With only two microphones only an estimation of the azimuth is possible, and additionally front-back confusion happens. A minimum of four microphones are necessary for localizing sounds without ambiguities.

The level difference between the main lobe and the side lobes depends also on the number of microphones M. In an ideal case, the level of the main lobe is $10lgM$ higher than the side lobes [31]. Figure 2.2 shows the directivity pattern for an array consisting of 25 microphones steered to $\phi' = 0^\circ$ at a frequency of 1kHz. The main lobe is about $10lgM$ higher than the side lobes.

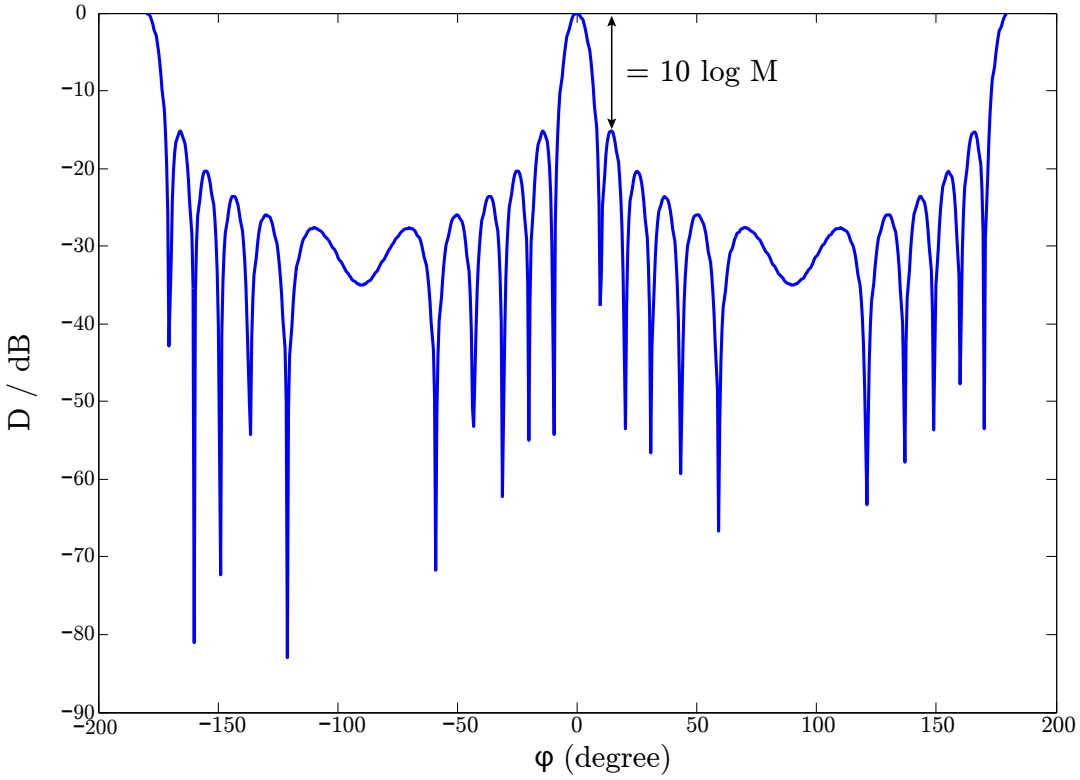


Figure 2.1.: Directivity Pattern of a microphone array with $M = 25$ elements ($\phi' = 0^\circ, f = 1kHz, d = 0.08m$)

The main lobe width is also proportional to $2/M$ hence it decreases slowly with increasing number of microphones. Both the level difference and the main lobe width can be increased with additional filters and different geometrical configurations.

Another important issue of a microphone array is the inter-element distance between each sensor. It determines the highest frequency without spatial aliasing. If spatial aliasing occurs, the directivity pattern shows so called *grating lobes*. This means that the wave front cannot be assigned to the correct direction of arrival due to phase ambiguities. These grating lobes appear if the inter-microphone distance is greater than half of the wavelength of the signal.

Thus the inter-element distance should be chosen to

$$d \leqq \frac{\lambda_{min}}{2} = \frac{c}{2f_{max}}, \tag{2.6}$$

where d is the maximal microphone pair distance, λ_{min} the minimal wave length, and f_{max} the maximum frequency without spatial aliasing.

To overcome this problem so called *harmonically nested subarrays* are used. These arrays do not have one fixed inter-element distance, but rather a combination of arrays with different spacings is used. Figure 2.3 shows a composited array of four different arrays with a frequency range of 0.5 - 8 kHz. One advantage using subarrays is the reduced number of microphones. Compared to four stand alone arrays with 30 microphones the composite array only needs 18 microphones.

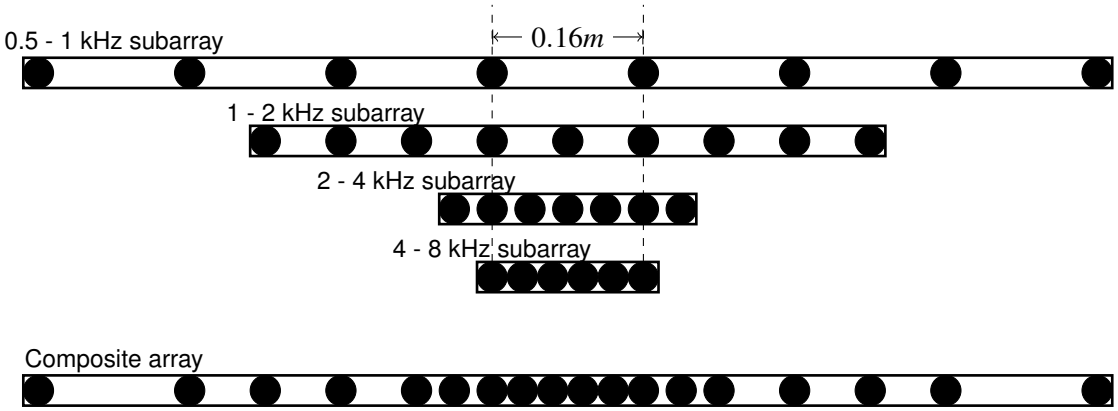


Figure 2.2.: Harmonically nested subarray covering four frequency bands

In harmonically nested subarrays the beamwidth keeps constant over a greater frequency range compared to a fixed distance array. Only the signal processing must be extended with bandpass filters and the microphone signals have to be shared to the corresponding delay elements. This extension of the signal processing is minor compared to the gained quality improvements (further informations in [5] and [31]).

Besides the discussed linear composite arrays alternative geometrical array configurations feature additional properties like increased three dimensional localization or a more homogeneous directivity.

2.3. Geometrical Array Configuration

The geometrical array configuration is one of the most important criterion developing a beamforming application. It influences significantly the system functionality and constrains the maximum of the directivity and the precision of the system. In the following the basic geometrical

configurations, their behaviours and limitations are depicted. Later on, in the experimental part of this thesis, three different configurations are analysed in relation to the use in conferencing environments.
According to the book *Optimum Array Processing* [48], the various array configurations can be divided into three categories:

- Linear
- Planar
- Volumetric (3-D)

To shorten this discussion, a pre-selection was done. Linear arrays are not further considered because their configuration is limited to one angular component (e.g. azimuth). This leads to front/back or left/right ambiguities (according to Chapter 2, pp. 17 et seq. in [48]), which is disadvantageous for the proposed teleconferencing scenario. So the different geometries are splitted into planar and non-planar configurations. Following, the requirements on the array for the intended purpose are stated.

2.3.1. Array Requirements

In this thesis a microphone array should be used to improve the recoding capabilities in a conference scenario. The scenario consists of a round conference table (radius about one meter) with several speakers sitting around the table talking freely to a recoding device which is placed in the middle of the table. Additionally the recording device should only consist of one centralized entity with a form factor comparable to existing table-top conferencing phones. The device should be used in different reverberant environments like common offices, which means the microphone array must be able to record speakers of all directions (all azimuths) around the table with a limitation in distance and aperture angle (elevation). The limited aperture angle is based on the fact, that the device is placed on the table and speakers only can speak from directions in the upper half-plane.

Another limiting factor is the online capability which constrains the maximum number of microphones. Actual available processing power limits the maximum manageable bandwidth which is direct proportional to the number of microphones. Processing complexity is in most algorithms directly related to the number of sensors.

A further aspect to be considered is the frequency range and the sampling frequency. The frequency of voice ranges from 300 Hz to 3400 Hz, resulting in a minimum sampling rate of 8 kHz. Recording with this sampling rate is necessary to reproduce a natural sounding voice, but the sampling rate also influences the localization accuracy, which is discussed in section 3.1.

As seen in this section, the array geometry isn't only influenced by design criteria but also by technical requirements and restrictions. Under these aspects the following sections will give an overview over existing array configurations.

2.3.2. Planar Configurations

A planar array configuration is an array whose elements all lie in one plane. Basically all shapes are imaginable, but in practice shapes with specific symmetrical properties show certain acoustic capabilities. It is possible to subdivide planar configurations into arrays with microphones covering the surfaces and into such arrays which have only elements on their edges. In relation to the teleconferencing scenario, it is always important to consider the processing complexity. With regard to this, the arrays with surfaces fully covered with microphones normally are computationally too demanding. Two examples of this type are shown in Figure 2.4 and 2.5. These array configurations were primary developed for radio communication or radar. But there is one special case, acoustic cameras, in which surface arrays are used in acoustical engineering. Figure 2.6 shows such a acoustic camera build up on a four by four microphone array. In principle these cameras base on rectangular arrays (Figure 2.5) which are orientated like a optical camera towards a scene. Then they receive the radiate sound waves and calculate an acoustical image of the scene. Usually these cameras only have a limited angle of entry which disqualify them for the use on a conference table.

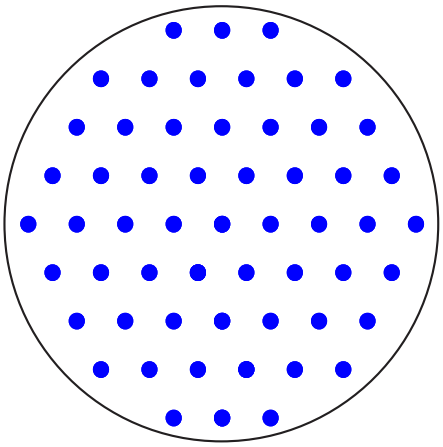


Figure 2.3.: Circular Array with hexagonal microphone distribution

The second type, arrays with microphones only on the edge of the shape, are supposed to be the better choice for teleconferencing. These microphone arrays consist of microphones arbitrary aligned on the boundary of a shape. By intuition, a rotationally symmetrical shape seems logical and in literature these types show the most homogeneous directivity patterns (Chapter 4.2, pp. 285 et seq. in [48] or Chapter 6, pp. 394 et seq. in [31]). The circular or ring array of Figure 2.7 is the simplest form of a rotationally symmetrical array. Other shapes like spiral or star shaped ones only show a increased accuracy detecting the elevation [31], which is not of particular importance for the given scenario.

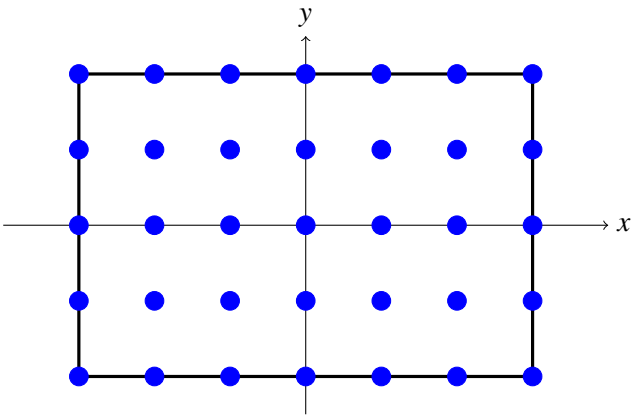


Figure 2.4.: Rectangular Microphone Array

Figure 2.5.: Acoustic Camera (`www.isemcon.com`)

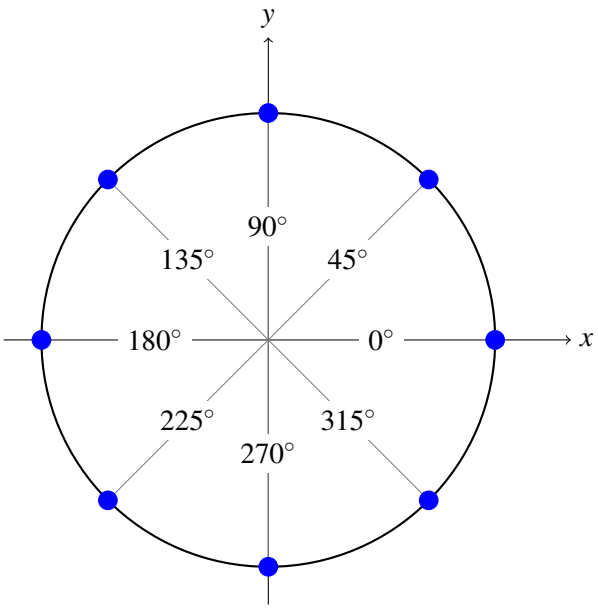


Figure 2.6.: Circular Microphone Array

2.3.3. Nonplanar Configurations

If the sound sources can occur in all spatial directions, then a volumetric array is needed. Those arrays use a third dimension for microphone placement. Possible configurations are three dimensional line arrays or geometrical bodies like spheres, cubes or cylinders. Volumetric bodies, like spherical arrays should either have a structure which not affects the propagating sound field or the influence must be considered in signal processing. This is the case, if the sensors are mounted on surfaces like the body of a robot. In contrast this circumstance is prevented using volumetric line arrays, whose structures only consist of small microphone fixings. The main advantage of a volumetric array is the possibility to detect the correct angular components without the lack of front/back ambiguities. Linear arrays and planar arrays can only detect the correct direction with additionally constrains, like a defined looking direction [31]. Regarding to the proposed scenario the array geometry is established by physical constrains like the table-top design and the positioning on the table. This already constrains the localisation to the upper hemisphere and a real volumetric configuration is not needed. Nevertheless in the experiments two volumetric line arrays are analysed for accuracy improvements.

Consideration of all possible configurations would exceed the scope of this thesis. There are many books about this interesting field of array geometry and it is fascinating to see the various acoustic properties, like the beam and directivity patterns, frequency dependences, and the range of array gains. Reference is therefore made to the book *Optimum Array Processing* of H. Van Trees [48] and to the book *Messtechnik der Akustik* of M. Möser [31].

2.4. Acoustic Beamforming

Beamforming removes interferences introduced by noise and reverberation. It can be considered as multidimensional filtering in space and time. The technique originated in radio astronomy during the 1950's as a way of combining antenna information from collections of antenna dishes. By the 1970's beamforming began to be explored as a generalized method of signal processing for any application involving spatially-distributed sensors [44, 6]. Examples of this expansion includes:

- Radar
- Radio Astronomy
- Sonar
- (Wireless-)Communications
- Sound Source Localization
- Seismology
- Medical diagnosis and treatment

Today beamforming is an active area of research and takes part in a variety of applications. The focus of this thesis is on Acoustic Beamforming, which tries to place a virtual microphone at various positions without physical sensor movement. Those virtual microphones are useful for applications like the human-computer interaction or hands-free telephony. Several beamforming algorithms exist for combining the sensor data, but they all base on delaying the signals and filter them in some way.

Beamforming can be divided broadly in two different approaches:

- **data-independent:** conventional beamformers with a fixed or switched beam

  They only depend on delaying the input signals. In advanced data-independent systems like Filter-and-Sum beamformers, filters additionally form the main lobe and try to suppress side effects.

- **data-dependent:** adaptive beamformers maximizing desired output signals

  The idea of a time-variant spatial-time response stands behind data-dependent beamformers. This response is adapted to maximize the main signal and to suppress interferences.

There are many beamforming algorithms in literature, most of them are only suitable for narrowband signals and inappropriate a wideband speech signal. Human voice covers about four octaves between 500 Hz, 1000 Hz, 2000 Hz and 4000 Hz. This wide range is only covered by special broadband beamformers. Traditionally, a narrowband beamformer has a directivity pattern which broadens towards lower frequencies.

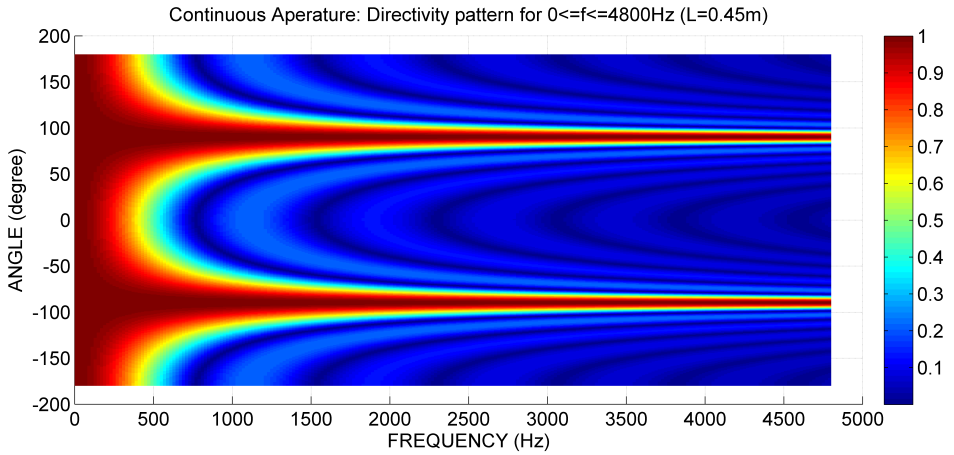


Figure 2.7.: Directivity pattern of a continuous aperture for $0 <= f <= 4800Hz$

Figure 2.8 depicts the directivity pattern of a theoretical continuous linear array without any additionally frequency dependent filtering. The main beam shows a strong frequency dependence and widens towards lower frequencies. This causes interference at low frequencies because every signal besides the main beam will be low-pass filtered rather than uniformly attenuated over its entire frequency range [6]. As described later, the various discrete apertures have all their individual patterns, depending on their dimensions, configurations and inter-element distances.

Signal leakage due to multipath propagation or crosstalk between the microphones decreases also significantly the signal-to-noise ratio in reverberant environments. Crosstalk is minimized automatically with increasing number of microphones. Another approach to overcome leakage effects is post-filtering the output signals with specific multi-channel filters. In

most cases pretty good crosstalk reduction can only achieved either by increasing the number of microphones or advanced filtering which consumes more computational power.

In the following chapters starting with the simplest form of a beamformer going further to more complex array processors, a short introduction in beamforming techniques is given as a theoretical basis for the developed algorithms in this thesis. In the following two different beamformer groups are explained in detail. At first the data-independent delay-and-sum beamformer is explained, than going further to more complex array processors which adapting their responses to the desired signal. A full mathematical description of each approach will be omitted in reference to specialized literature (for example, see *H.L. Van Trees: Optimum Array Processing* [48]).

2.4.1. Data-independent beamforming

Data-independent beamforming means that the algorithms do not adapt their transfer functions to the input signals. The best-known algorithm is the delay-and-sum beamformer (DSB) which has a simple principle. It applies on each incoming microphone signal time-shifts according to the steering angle and microphone array configuration. These short delays compensate the propagation time differences between the source and each microphone. After delaying the signal they are summed up to generate one single output signal. Because of the constructively or destructively superposition of the microphone signals the desired signal at the steering angle is enhanced. In Figure 2.9 the structure of a delay-and-sum beamformer is depicted. It shows the source $s_c(t)$, a variable number of microphones N with corresponding signals $x_n(t)$ and delay elements τ_n. The summation is shown as the addition sign and the output is $y(t)$.

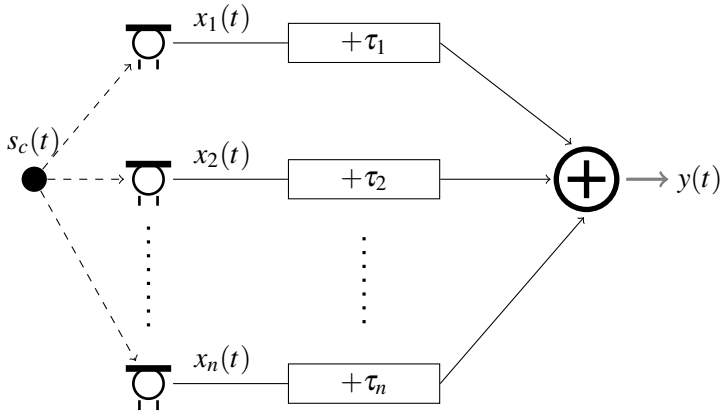


Figure 2.8.: Delay-and-sum beamformer

So in general, the delay-and-sum beamformer output $y(t)$ is computed by:

$$y(t) = \sum_{n=1}^{N} x_n(t - \tau_n). \tag{2.7}$$

This formula represents the most basic version of a delay-and-sum beamformer. In practice it is often extended by a real weighting factor w_n,

$$y(t) = \sum_{n=1}^{N} w_n x_n(t - \tau_n), \tag{2.8}$$

which allows compensation of different microphone gains. As already mentioned above, the DSB strongly depends on the frequency. A beam pattern of a linear array with equidistant microphones and a steering angle of 90 degree is plotted in Figure 2.10. It shows again that the main beam broadens towards lower frequencies and that beyond the maximum frequency strong side lobes occur. That implies that the standard delay-and-sum beamformer has a low-pass characteristic, so that interfering signals are low-pass filtered and added to the output signal. Additionally beyond $f_{max} = \frac{c}{d}$, which is determined by the inter-element distance, spatial aliasing causes ambiguities and so strong interferences in the output. Over the whole frequency

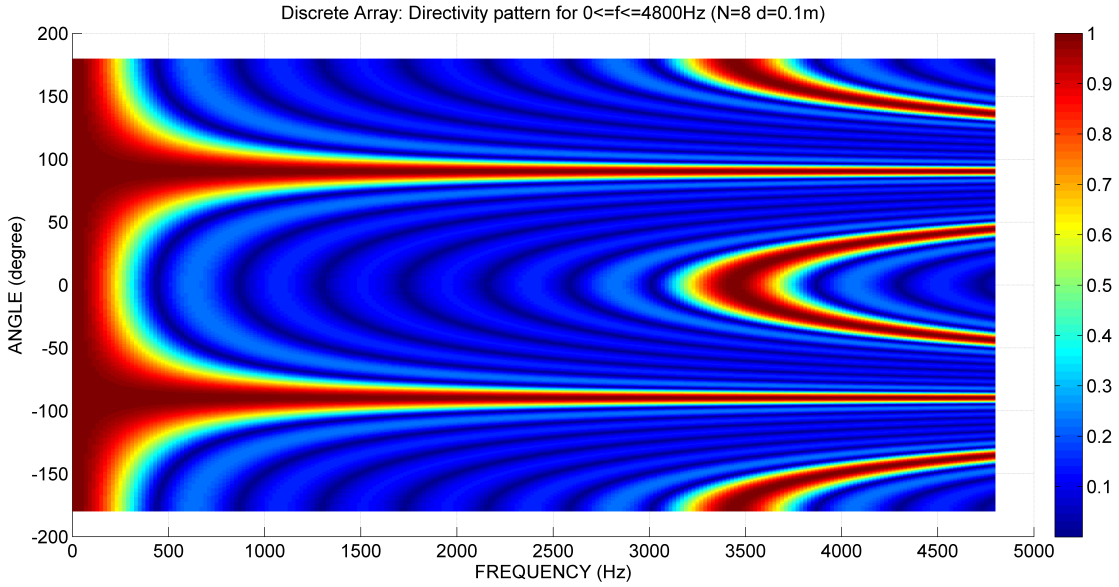


Figure 2.9.: Delay-and-sum beam pattern of a discrete linear array steered to 90 degree

range, the side lobes interfere with the output signal which can be reduced by a increasing number of microphones or through post-filtering.

Filtering the output signal is mostly limited, therefore approaches were developed which apply filters on each microphone signal before summing them up. These sophisticated beamformers with integrated filters are widely known as filter-and-sum beamformers. They try to optimize the beam shape to produce a more homogeneous response for speech acquisition. Further information are denoted in the chapter *Constant Directivity Beamforming* (pp. 3 et seq.) by *Darren B. Ward* in [6]. These methods assume that the nature of the interferences (its statistics in particular) and the signal characteristic is known a priori. According to these assumptions the filters are designed ahead of time for particular applications. But in practice

involving human talkers and a realistic audio scene the prediction of filter responses fail. For these rapidly changing scenarios an adaptive technique would be better. This is the motivation behind the study of data-dependent array processing and is the focus of the next section.

2.4.2. Data-dependent beamforming

An adaptive, or data-dependent beamformer tries to apply weights to each microphone in an optimal sense. This means, in dependence of the source signal all filter weights of a sophisticated beamforming system are adapted accordingly.

A first attempt at inventing such a beamformer was done by Otis Lamont Frost in 1972 [18]. The basic concept behind is the minimization of the output power through an optimal weighting of the beamformer inputs. Several derivations of the algorithm can be found in literature [18, 57, 55] and therefore omitted here.

Simulations of a Frost beamformer show quite narrow main beams [57]. Therefore Frost beamforming does not tolerate errors in the steering vector. Also in reverberant environment the algorithm eliminates parts of the target signal due to correlations of reverberant signals with the target signal. And finally the Frost algorithm depends on precise microphone gains and positioning which is mostly not achievable in practice.

All these drawbacks make the Frost Beamformer not applicable in practical environments, like the proposed teleconferencing scenario. So a more robust approach where considered, the Griffiths-Jim beamformer.

In 1982 L. J. Griffiths and C. W. Jim published their paper about *An Alternative Approach to Linearly Constrained Adaptive Beamforming*, today widely known as the Griffiths-Jim beamformer (GJBF) [19]. The technique belongs to the data-dependent beamformers which according to the source signals adaptively form its directivity patterns. A Griffiths-Jim beamformer (GJBF) or also known as generalized side lobe canceller consists of a fixed beamformer (e.g. filter-and-sum beamformer), a multiple-input canceller and a blocking matrix.

The simple fixed beamformer is steered to a specific direction enhancing the target signal and attenuating as good as possible interfering signals. On the contrary, the blocking matrix generates a signal, which blocks the signal in look direction and let simultaneously pass all other signals besides the target. The simplest realization of a blocking matrix is generally a delay-and-subtract beamformer which delays the microphone signals according to the steering vector and subtracts them from the earliest arriving microphone signal. After this the multiple-input canceller uses multiple adaptive filters which are driven by the blocking matrix outputs to correlate them with previous output signals of the GJBF. Through the correlation the undesirable signals are detected and enhanced and finally subtracted from the fixed beamformer output signal. Through the subtraction all interfering signals besides the look direction are subtracted from the standard fixed beamformer signal yielding a better signal to interference ratio. So the Griffiths-Jim beamformer (GJBF) achieves fair target signal extraction and feasible interference suppression compared to data-independent beamformers [6].

The biggest problem of the Griffiths-Jim beamformer (GJBF) is that it depends strongly on the correctness of the steering vector. The target signal is disproportional attenuated if the

looking direction between two time steps varies too much because the target signal is leaked to the multiple input canceller resulting in target signal attenuation. Steering vector errors are not only caused by a false direction of arrival (DOA) estimation, but also through errors in the microphone positions and those in the microphone characteristics (e.g. inhomogeneous polar plots) [6].

For the particular use case of teleconferencing the localization of the source positions is a difficult task, so that the Griffiths-Jim beamformer (GJBF) is too sensitive and therefore not the best choice for the proposed scenario. There are various techniques to overcome the target-signal leakage problem but most of them induce a additionally delay due to advanced adaptive steps needing previous samples.

2.4.3. Geometric Source Separation

In the previous section a introduction to beamforming solutions was given. Beamforming can be used to enhance a specific signal out of a mixture of signals. Therefore it relies primary on geometric informations and in some cases additionally on statistical properties or special signal characteristics. A completely different approach is Blind Source Separation (BSS) which relies completely on statistical and signal characteristics. Most of them assume that source signals and noise signals are mutually independent or decorrelated. The BSS algorithms try to maximize the statistical independence between the target signals. Best known basic algorithms are Principle Component Analysis, Singular Value Decomposition, or Independent Component Analysis. A good overview of these algorithms is given in the book *Blind speech separation* by *Shoji Makino et al*. [29].

However, Blind Source Separation (BSS) algorithms are mostly computational complex and need a lot of processing time. Often they are also iterative algorithms and their convergence time can be long. So such BSS approaches are in general not suitable for online signal separation.

Recently a new kind of algorithm, the so called Geometric Source Separation (GSS) has been published combining Blind Source Separation (BSS) and beamforming in a beneficial manner. Geometric Source Separation (GSS) were proposed by L. C. Parra and C. V. Alvino in the paper *Geometric Source Separation: Merging Convolutive Source Separation With Geometric Beamforming* [37].

Overcoming the main problem of adaptive beamforming, cross-talk and signal leakage, GSS uses mostly readily available source localization informations. For this purpose, Parra and Alvino combine beamforming with conventional source separation by using cross-power minimization with geometric linear constraints. They assume that the source signals are independent. This causes ambiguities in terms of permutations and scaling because convolutive source separation does not identify the source $s(t)$ and their corresponding frequency bins directly. So it can be formulated a permutation matrix which assigns and scale each frequency to the correct source. This matrix increases with the number of microphones as the number of possible permutations increases.

In the past, most separation approaches have tried to resolve these ambiguities by exploiting the continuity in the signal spectra [8], or the co-modulation of different frequency bins [2]. These methods depend on polyspectra [42, 13], which are in practice for speech signals hard to obtain and are normally computationally very demanding.

Additional ambiguities are introduced by using more sensors than sources. In this overdetermined case conventionally subspace analysis is done to determine the signal and noise subspace. This problem is also handled in the GSS algorithm by constraining the filter based on geometric assumptions.

In general, the separation problem in the discrete time Fourier domain can be denoted as (according to [37]):

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{x}(\omega), \tag{2.9}$$

with

$$\mathbf{x}(\omega) = \mathbf{A}(\omega)\mathbf{s}(\omega), \tag{2.10}$$

where $\mathbf{A}(\omega)$ is the matrix of linear transfer functions between the sources $\mathbf{s}(\omega)$ and the microphones and with $\mathbf{W}(\omega)$ the filters inverting the effect of convolutive source mixing. Equation (2.9) can be rewritten to

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{A}(\omega)\mathbf{s}(\omega) = \mathbf{PS}(\omega)\mathbf{s}(\omega), \tag{2.11}$$

which shows, that the separation problem can be solved except for a arbitrary permutation matrix $\mathbf{P}$ and a arbitrary scaling matrix $\mathbf{S}(\omega)$ per frequency.

Now, this separation problem can be implemented as cross-power minimization to reduce off-diagonal elements of

$$\mathbf{R}_{\mathbf{yy}}(t,\tau) = E[\mathbf{y}(t)\mathbf{y}^H(t+\tau)] \tag{2.12}$$

for different values of t. Minimization of this equation means that the correlation between each channel is minimized and this can be implemented as diagonalization of the cross-power spectra $R_{yy}(t,\tau)$ in time domain. But as well as, for efficiency, this can also be expressed in frequency domain.

Calculating the cross-power spectra in an on-line algorithm would cause an additionally delay, thus only a running estimate of $\mathbf{R}_{\mathbf{yy}}(t,\tau)$ (denoted in frequency domain) directly from the outputs $\mathbf{y}(t)$ is computed as

$$\mathbf{R}_{\mathbf{yy}}(t,\tau) \approx \mathbf{W}(\omega)\mathbf{R}_{\mathbf{xx}}(t,\tau)\mathbf{W}^H(\omega). \tag{2.13}$$

This calculation (according to [16]) rewrites the cross-correlation of the outputs $\mathbf{y}(t)$ using equation (2.9) with the matrix $\mathbf{W}(\omega)$ of the most recent filter coefficients and with a current estimate of the cross-correlation of the recorded signals $\mathbf{x}(\omega)$ as a running estimate. This approximation is only accurate for a filter length Q much shorter than the analysis window length T [24].

For the minimization task of $\mathbf{R}_{\mathbf{yy}}(t,\tau)$ a fast gradient decent algorithm is used. This algorithm minimizes the filter coefficients W which in turn diagonalizes $\mathbf{R}_{\mathbf{yy}}(t,\tau)$. For this minimization task in the paper of *Parra* and *Spence* an algorithm minimizing the sum of squares of the off-diagonal elements under various optimization criteria is proposed [37, 36].

A drawback of this "original" GSS algorithm is the estimation of the correlation matrices $\mathbf{R_{yy}}(t,\tau)$ and $\mathbf{R_{xx}}(t,\tau)$ over several seconds leading to a time delay. A different approach related to the algorithm of *Parra et al.* is proposed by *J. M. Valin* in *Enhanced robot audition based on microphone array source separation with post-filter* [47]. In this paper the focus lies on a simple, fast and robust source separation for the audition of a robot. Therefore, the correlation matrices estimations were simplified to

$$\mathbf{R_{xx}}(t,\tau) = \mathbf{x}(t,\tau)\mathbf{x}(t,\tau)^H \tag{2.14}$$

$$\mathbf{R_{yy}}(t,\tau) = \mathbf{y}(t,\tau)\mathbf{y}(t,\tau)^H \tag{2.15}$$

which is an instantaneous estimation of the correlation. This simplification has not shown any reduction in accuracy and furthermore eases the implementation of an on-line algorithm [47].

Additionally the separation problem of equation (2.9) is reformulated to a form estimating directly the separation matrix $\mathbf{W}(\omega)$ under two constraints. The first constraint (eq. (2.16)) contains the minimization problem of equation (2.12), the second (eq. (2.17)) takes the geometrical informations into account:

$$\mathbf{R_{yy}}(t,\tau) - diag[\mathbf{R_{yy}}(t,\tau)] = 0 \tag{2.16}$$

$$\mathbf{W}(\omega)\mathbf{A}(\omega) = \mathbf{I} \tag{2.17}$$

Equation (2.17) is the geometric constraint, which ensures unity gain (I denotes the identity matrix) in source direction and places zeros in all other directions. Both constraints could be used for separation, the first constraint would minimize correlation between the signals and the second would cancel interference of unwanted directions (e.g. reverberation). Together both constraints are to strong, but can be used as cost functions in a gradient decent algorithm. These cost functions are calculated by:

$$J_1(\mathbf{W}(\omega)) = \|\mathbf{R_{yy}}(t,\tau) - diag[\mathbf{R_{yy}}(t,\tau)\|^2 \tag{2.18}$$

$$J_2(\mathbf{W}(\omega)) = \|\mathbf{W}(\omega)\mathbf{A}(\omega) - \mathbf{I}]\|^2 \tag{2.19}$$

And with this, the corresponding gradient for the cost functions with respect to $\mathbf{W}(\omega)$ are:

$$\frac{\delta J_1(\mathbf{W}(\omega))}{\delta \mathbf{W}^*(\omega))} = 4\mathbf{E}(\omega)\mathbf{W}(\omega)\mathbf{R_{xx}}(t,\tau) \tag{2.20}$$

$$\frac{\delta J_2(\mathbf{W}(\omega))}{\delta \mathbf{W}^*(\omega))} = 2[\mathbf{W}(\omega)\mathbf{A}(\omega) - \mathbf{I}]\mathbf{A}(\omega) \tag{2.21}$$

where $\mathbf{E}(\omega) = \mathbf{R_{yy}}(t,\tau) - diag[\mathbf{R_{yy}}(t,\tau)]$.

With this gradient functions taking decorrelation and geometric properties into account, the separation matrix is then updated as follows:

$$\mathbf{W}^{n+1}(\omega) = \mathbf{W}^n(\omega) - \mu\left[\alpha(\omega)\frac{\delta J_1(\mathbf{W}(\omega))}{\delta \mathbf{W}^*(\omega))} + \frac{\delta J_2(\mathbf{W}(\omega))}{\delta \mathbf{W}^*(\omega))}\right] \tag{2.22}$$

where μ is the adaptation rate and $\alpha(\omega) = \|\mathbf{R_{xx}}(t,\tau)\|^{-2}$ is an energy normalization factor [47]. For calculation, the instantaneous estimations of the cross-correlation matrices of equation

(2.14) and (2.15) are used, which significantly reduces the complexity requiring only matrix-by-vector products:

$$\frac{\delta J_1(\mathbf{W}(\omega))}{\delta \mathbf{W}^*(\omega))} = 4[\mathbf{E}(\omega)\mathbf{W}(\omega)\mathbf{x}(t,\tau)]\mathbf{x}(t,\tau)^H \tag{2.23}$$

and the energy normalization factor reduces to

$$\alpha(\omega) = \left[\|\mathbf{x}(t,\tau)\|^2\right]^{-2}. \tag{2.24}$$

The final question is the initialization of the separation matrix. The paper of the original version of the GSS algorithm denotes various resolutions for the initial values of $\mathbf{W}(\omega)$ [37]. Finally, the best working matrix with acceptable separation quality, contains the filter coefficients of a delay-and-sum beamformer steered into source direction.

With the given algorithm, for each time frame of the captured signal a separation matrix is calculated and then convolved with the input signal according to equation (2.9). More about the implementation aspects can be found in chapter 3.2.

2.5. Audio Based Localization

In this section a short introduction to audio based localization of sound sources is given. Localization of acoustic sources is useful in many practical applications like surveillance systems, video conferencing or for hands-free speech acquisition. The focus will be on locators using microphone arrays, because this is a well known field of research and promises good results. Other solutions like binaural approaches will be omitted and refereed to appropriate literature.

The considered approaches should be adequate for the use in a reverberant environment and should be able to detect at least two simultaneous sources.

Existing source localization procedures basing on microphone arrays may be divided into three groups [6] those based upon maximizing the Steered Response Power (SRP), techniques adopting high-resolution spectral estimation concepts, and approaches exploiting Time Delay of Arrival information.

2.5.1. High-Resolution Subspace Techniques

In the category of the High-Resolution Subspace Techniques or Spectral-Estimation-Based Locators, MUSIC is one of the most popular algorithm [40] besides ESPRIT [38] or MIN-NORM [26]. The term MUSIC means MUltiple SIgnal Classification and in the field of multiple source localization MUSIC shows its advantages. The method bases on the exploitation of the properties of the so called Cross-Sensor Covariance Matrix. This is a correlation matrix calculated across all spatial distributed sensors.

Generally, High-Resolution Subspace Techniques base on the assumption that the considered signal characteristics are known and that only some parameters have to be estimated. The estimation of these parameters is implemented by a Principle Component Analysis (PCA) on the array covariance matrix in order to separate the signal subspace and the noise subspace. The PCA is used in a matrix operation that results in peak responses in the source

directions.

The following gives a short overview of the MUSIC algorithm (as described in [40]). As mentioned MUSIC assumes the signal characteristic (data model) as

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix} = \begin{bmatrix} a(\theta_1) & a(\theta_2) & \cdots & a(\theta_D) \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_D \end{bmatrix} + \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_M \end{bmatrix} \tag{2.25}$$

or

$$\mathbf{X} = \mathbf{AF} + \mathbf{W}, \tag{2.26}$$

where X_M denotes the M microphone signals, F_D the D original source signals, and W_M the noise vector introduced through the recording hardware. The elements of $\mathbf{A}$, $a(\theta_D)$ represent the transfer functions between the microphones and the sources relative to their directions of arrival θ. But this equation is unresolvable due to the ambiguities of source assignment.

Calculating the Cross-Sensor covariance matrix:

$$\mathbf{S} \triangleq \overline{\mathbf{XX}^*} = \mathbf{A}\overline{\mathbf{FF}^*}\mathbf{A}^* + \overline{\mathbf{WW}^*} \tag{2.27}$$

of the microphone signals and hence

$$\mathbf{S} = \mathbf{APA}^* + \lambda\mathbf{S_0} \tag{2.28}$$

where $\mathbf{P}$ is the positive definite matrix of the pair-wise correlation of the source signals and $\mathbf{S_0}$ the metric of $\mathbf{S}$. If there are less sources than microphones $\mathbf{APA}^*$ is singular. Therefore it follows of eq. 2.28

$$|\mathbf{APA}^*| = |\mathbf{S} - \lambda\mathbf{S_0}| = 0 \tag{2.29}$$

and this equation is only satisfied for the minimum eigenvalue λ_{min}. It is not always simple to derive one single λ_{min} as solution to $|\mathbf{S} - \lambda\mathbf{S_0}| = 0$, because there are as much as incident signals. The smallest eigenvalues of eq. 2.29 refer to the eigenvectors which span the signal subspace. The remaining eigenvalues correspond to the eigenvectors spanning the noise subspace. Both subspaces are disjoint. In [40] Schmidt presents an algorithm to calculate these solutions.

There are two main limitations of subspace techniques. First, they are unable to detect more source signals than number of sensors. And for this limiting source number not only real sources count, but also strong interferes like echoes. Second, their computational complexity is too demanding for easy on-line speaker localization. Hence, for the proposed scenario, subspace techniques will not be considered further in this thesis.

2.5.2. Estimation of Time Delay of Arrival

Time Delay of Arrival (TDOA) based localization calculates time delay estimations of impinging signals relative to microphone pairs [41, 12]. In combination with the microphone positions and the time delays hyperbolic curves are generated and intersected. These intersecting curves yield a suboptimal position estimation.

In figure 2.11 two examples of regions where the Time Delay of Arrival (TDOA) is equal are shown. Often in literature this is called the cone of confusion. These two cone shaped areas are intersected, resulting in a hyperbolic curve. Every point on this curve is a possible source location. With an additional cone from another pair of microphones the first bearing line is intersected in one point. With more than three microphones several hyperbolic curves intersect in one single point (the source position), under ideal conditions. But under real conditions the bearing lines intersect at different points. These slight displacements of the lines are introduced due to imperfect recording hardware, noise and reverberations. So an estimation algorithm has to find the most likelihood source position. These estimation process differentiates the various TDOA algorithms.

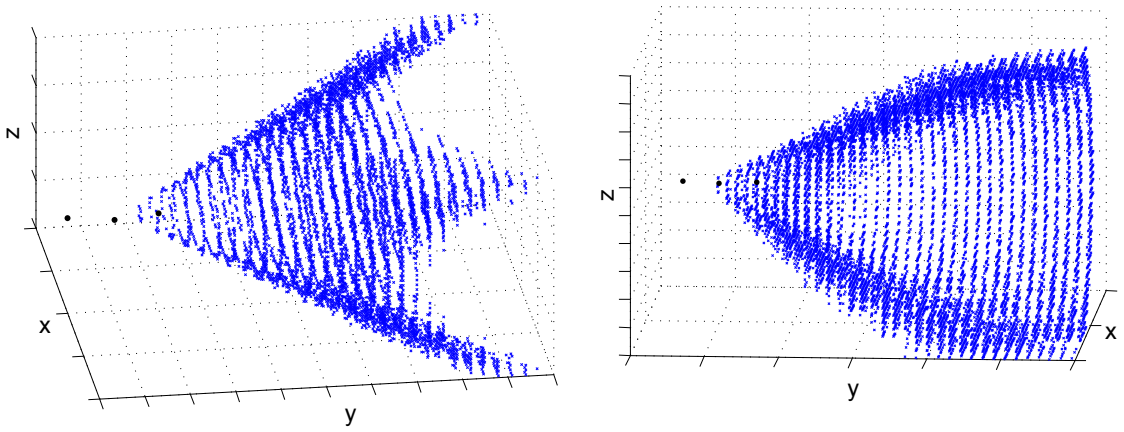


Figure 2.10.: Examples of regions where the Time Delay of Arrival is equal

The multi source localization is a problematic aspect of TDOA based algorithms. Most approaches calculate the time delays by the Generalized Cross-Correlation (GCC) method [25]. Assigning the peaks of the cross-correlation functions to the different sources are a separate task and need special attention [39]. Therefore TDOA localization primarily works well for a single source which is for the intended purpose relatively uninteresting.

2.5.3. Steered Response Power Localization

Beamforming normally is used for capturing signals (e.g. voice) of one specific region. Generally beamforming exploits the knowledge of particular acoustical transfer functions between sources and sensors. Applying this transfer function to captured signals would therefore yield the source signal. Unfortunately, these transfer functions are in most cases only approximations. But this ability to approximate and apply these functions for arbitrary environments results

in a great system flexibility and adaptation capability. Furthermore, the filter function for simple beamformers can be pre-calculated fast and stored in look-up tables.

These efficiency is used in SRP localization techniques. That means, a focused beamformer is steered to various locations and searches for the highest output power. Looking for further peaks in the output extends this approach to a robust and fast multi source localizer.

In time domain the delay-and-sum beamformer provides the simplest type of steered response. The beamformer applies simple time shifts to each microphone signal compensating the propagation delays from the source position to the microphones. The shifted signals are then summed together and the energy of the signal is calculated. In cases, with limited noise, equal source-microphone distances, and low reverberation this simple solution is appropriate. On the other side, this simple solution produces very wide energy peaks which overlap in case of near sources.

Advanced beamformers, as mentioned in previous chapters, like filter-and-sum beamformers apply filters to the microphone signals additional to the delays in order to shape the frequency dependent beams. This beam shaping produces smaller peaks in the energy function, which is required in practical environments with reverberation and varying microphone-source distances. So advanced beamforming solutions with signal filtering and weighting distinguish the different SRP localizers.

One of the most robust localization algorithms is the so called SRP-PHAT (Steered Response Power - Phase Transform) method. Which combine the beneficial GCC and its PHAT implementation with the principle of SRP algorithms.

SRP-PHAT relies on a filter-and-sum beamformer applying a Phase Transform (PHAT) which weights each frequency component equally [14, 25]. This can be formulated in the frequency domain as

$$Y(\omega, \mathbf{p}) = \sum_{n=1}^{N} G_n(\omega) X_n(\omega) e^{j\omega\tau_n}, \tag{2.30}$$

which is the general form of a filter-and-sum beamformer where $\mathbf{q}$ denotes the source position with the corresponding delays τ_n between the source and each microphone. With this, the output power for a specific position $\mathbf{q}$ is defined as

$$P(\mathbf{q}) = \int_{-\infty}^{+\infty} \|Y(\omega)\|^2 d\omega \tag{2.31}$$

and a possible source location is found from

$$\hat{\mathbf{q}}_s = \arg\max_{\mathbf{q}} P(\mathbf{q}) \tag{2.32}$$

evaluated for each considered location $\mathbf{q}$. With this basic formulas the SRP algorithm with additionally PHAT weighting is expressed as

$$P(\mathbf{q}) = \sum_{l=1}^{N} \sum_{k=1}^{N} \int_{-\infty}^{+\infty} \Psi_{lk}(\omega) X_l(\omega) X_k^*(\omega) e^{j\omega(\tau_k - \tau_l)} d\omega, \tag{2.33}$$

where $\Psi_{lk}(\omega)$ corresponds to the multi-channel version of GCC-PHAT weighting and is given by

$$\Psi_{lk}(\omega) = \frac{1}{\|X_l(\omega) X_k^*(\omega)\|}, \tag{2.34}$$

where index k and l denotes the summation over all different microphone pairs. The positions $\mathbf{p}$ are related to points of a search region, like a sphere around the array. For each point of this region equation 2.33 is evaluated and analyzed for peaks. Besides an energy threshold which determines the energy value of a valid source, a value for the maximal number of sources is defined. This value specifies the number of further peaks considered besides the first one, which corresponds to the number of multiple tracked sources.

Normally the SRP-PHAT algorithm is the best trade-off between computationally complexity and localization performance [14]. In [1] SRP-PHAT were compared to a TDOA method (also based on GCC-PHAT) in relation to the speakers head orientation. The results show that the SRP-PHAT algorithm is robust against rotations of the speaker's head. Another comparison of different localization principles is done in [28] with the result that the SRP-PHAT method performs well in a wide variety of different parameters like number of microphones or used frame size. Because of its great robustness against a wide range of interference, its flexibility and efficiency the SRP-PHAT algorithm was selected for the proposed scenario.

2.6. Audio-Visual Tracking

In the last chapters the three basic concepts for audio based localization were presented. There are for each concept a variety of different approaches and modifications. But the limitation to only one type of sensor (in this case microphones) may not be the best solution. Each type of sensor has a specific strength and weakness and a combination of sensor modalities can often achieve better results. For source localization of teleconference participants a combination of video based tracking and acoustic source localization promises a more robust position estimate. So the following chapter presents some basic topics of audio-video based localization.

2.6.1. Face Tracking and Omnidirectional Vision

As mentioned above, a practical example of a multi sensor system may combine a microphone array and a video camera. Either sensor can be used to estimate a speaker position. But only the microphone can detect the voice activity for sure and only the camera can track the speaker in periods of silence. So a combination ensures a continuous and robust position estimation. There are different possibilities for combination of audio based localization and video tracking. Therefore, at first the techniques for camera based face tracking and then the approaches for the sensor fusion are stated.

Several problems arise using camera based object tracking. Generally, camera based tracking is inappropriate for absolute position estimation. Only a calibrated tracking system can perform this task [45]. Normally camera calibration is done in advance using special objects with known dimensions. After calibration the system can recognize the position and distance of an object, under ideal conditions.

In most cases, however, the absolute position of the conference participants is not needed. For steering the beams of a beamformer, the Direction of Arrival (DOA) is sufficient. This task has been researched extensively over the recent years and therefore is well known and can be

used "out of the box". For example the *OpenCV* library[1] delivers various functions to detect faces and their directions. The face tracking components, included in *OpenCV* base on trained classifiers and where originally developed by *Paul Viola* [51] and later improved by *Rainer Lienart* [27]. There are many more approaches for face detection and tracking. Performance evaluations and selection of an approach is beyond the scope of this thesis.

Another aspect which has to be considered in the proposed scenario of teleconferencing is the ability to detect all participants around the table. An omnidirectional camera device mounted on top of the microphone array can perform this. There are approaches using at least two cameras with fish eye lenses facing in opposite directions [35] to acquire an omnidirectional view or systems using a convex paraboloidal shaped mirror to record a full 360 degree image of one hemisphere [32]. Both systems acquire a distorted image and therefore a reconstruction is needed. But it is possible to record a full panoramic view of a teleconference scene and perform on-line face tracking [7, 35]. With this, face tracking can deliver additional informations of present speakers and their estimated directions which can be used for speaker recognition and improvements of the reception quality.

2.6.2. Sensor Fusion and Particle Filtering

To estimate the location of an acoustic source in a room, an approach of chapter 2.5 describing different audio based localizers can be used. Due to interferences, noise, and reverberation, inaccuracies in the position estimate produces noisy localization estimates. These distortions affect the quality of a separation process depending on these position estimates. The noisy instantaneous position estimates can be improved by tracking them over time. For this purpose the particle filter algorithm provides an effective way of modelling a stochastic process with arbitrary probability density functions (pdfs) by approximating it with a cloud (also widely known as a swarm) of points called particles. Particle filters were originally introduced in the computer vision area by *M. Isard* and *A. Blake* [21]. The particles of such filters are described in a process state space at a time t, by a cloud index j, and a particle number $i = 1, ..., N$:

$$s_{j,i}^{(t)} = \begin{bmatrix} \mathbf{x}_{j,i}^{(t)} \\ \dot{\mathbf{x}}_{j,i}^{(t)} \end{bmatrix}, \tag{2.35}$$

where $\mathbf{x}_{j,i}^{(t)}$ is the position and $\dot{\mathbf{x}}_{j,i}^{(t)}$ its derivative of each particle. Corresponding to this, every particle is weighted with a weight $\mathbf{w}_{j,i}^{(t)}$. The particles with its states (position and derivation) and weights represent the probability density function (pdf) for one speaker location. Each particle with the same index j corresponds to one particle swarm and is assigned to one present (and potentially active) voice source. Then, during the on-line location estimation the detected locations are assigned to one swarm. This observation of a source is used to update the particles weights.

After this update stage, the prediction stage of the particle filter follows. The prediction stage uses the system model to predict the pdf of each location from one measurement time to the

[1] `http://opencv.willowgarage.com` (accessed September 13, 2011)

next. This helps to compensate disturbances induced by noisy measurements or interferences [3]. The system model describes the evolution of the particle states, like a model for their possible directions and valid locations. During the next update stage, before the weighting is applied, the particles are propagated in time according their predicted motion model (derivation).

Repeating the update and prediction step continuously and calculating the most likelihood position approximated by the particle swarm delivers a nearly statistical optimal tracking solutions.

Practical implementation of particle filters include additional steps like a resampling stage or advanced predictors. But these various modifications are often made to adjust the particle filter to one specific problem. In the following, therefore, the extension of particle filters to multi-modal sensor fusion algorithms is described, which is also an active research area.

Multi-modality means that informations of different sensors, like acoustic and visual sensors, are used. In case of audio-visual tracking, there are principally three different ways to fuse the audio and video data. First, it is possible to use the audio based localization to steer a camera towards an active speaker, so that the video always shows the current speaker [53]. Second, the camera can be used to steer multiple beams of a beamformer to potential sources such that the reception quality is enhanced [10]. The third possibility is to combine both position estimations to one more reliable source position [49, 33]. But then the problem arises how to combine the two raw measurements. Both localization methods are subjected to measurement errors, which can lead to disagreements between the audio-based and visual tracking. So an adequate way for fusing the sensor data has to be found.

One approach is based on the above mentioned particle filters. In [56] a system is described which combines the results of an audio based localizer with the ones of a visual tracking system. In such a system the particles are additionally weighted by the observation result of the video tracker. That means, if the video tracker and the acoustic source localization detects similar positions, then particles lying near to this position get a quite high weight. Otherwise, if the video and audio position estimates differ considerably the particle swarm widens over time resulting in an uncertain position. This flexible weighting of particles with different sensor data is beneficial for combining multi-modal sensor systems with a particle filter. There are again various approaches in literature combining all possible types of sensors with different methods through a particle filter.

2.6.3. Efficient Video Extension

Besides fusing multi-modal sensor data to one single output, different localization techniques can be used to provide a more robust and efficient localization technique. Combining sensor data with particle filtering adds at least an additional weighting step into the algorithm, which leads to an increased computational complexity. The result is a more robust and accurate estimation with a slightly increased computational time.

Another way of combining two different sensors is to constrain the search region of one sensor with the results of the other one in order to decrease the processing time. In most

cases this limits the accuracy or robustness of the estimation. But in case of audio-visual tracking this can lead to a tracking stabilization and acceleration of the localization algorithm. Especially if the audio-based localization, like SRP-PHAT, uses a predefined search region.

The proposed algorithm therefore, combines the audio-based localization with the video tracker in such a manner that the search region of the SRP-PHAT localizer is limited. In the case that video-based tracking fails or the confidence into the values is too low, the system automatically falls back and uses the default full search region as in the original pure audio-based approach. Figure 2.12 shows the proposed extension.

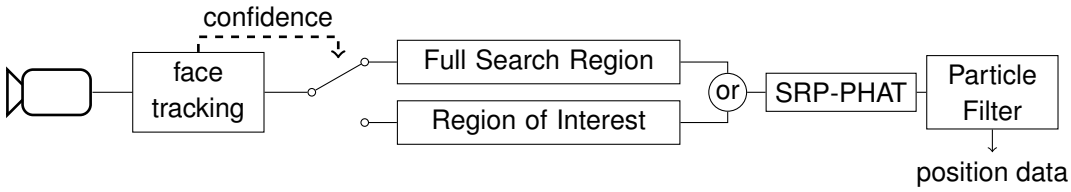


Figure 2.11.: Video-based tracking extension of the SRP-PHAT algorithm

Additionally the face tracking data could also be used for further simple extensions. First, the face tracking algorithm can deliver instantaneous position estimations. The audio based localization normally needs a minimum number of frames to deliver a valid estimation. During that time, the separation algorithm could receive the position data directly from the face tracker. Secondly, if the audio based localization fails, like in strong reverberant and noisy environments, the separation stage could fall back and use also directly the face tracking positions. In future applications it might also be possible to detect the speaker activity according to their mouth movements, which again could improve the confidence value of the face tracker.

The presented extension of Figure 2.12 will be experimentally implemented as a simulation (see for section 3.3), but a full evaluation will be omitted until the real face-tracking extension is ready, which is currently developed at the Institute.

3. Developed Algorithms

Throughout the thesis, a system locating and capturing individual speakers separately were designed. The preceding chapters have given a system overview and an introduction to basic principles needed for this task. In the following sections a selection out of the described theoretical approaches was made and important implementation aspects were stated.

3.1. Selection of most promising Approach

In recent years there were many approaches for localization and separation of multiple speakers in a room. The previous chapters have tried to give complete but still short overview over the different technologies. For the practical implementation, algorithms were selected, which promise the best trade-off between speed and quality in relation to conferencing scenarios.

With some optimizations it has been shown [46], that a Steered Response Power - Phase Transform (SRP-PHAT) localizer can be implemented efficiently. Most speed improvement is made using pre-calculated lookup parameters and a optimized search region. Furthermore, SRP-PHAT algorithms have a good robustness in the presence of room effects and noise [6]. The maximum achieved accuracy depends mainly on the pre-calculated candidate locations and on the sampling frequency. In cases of small arrays the candidate locations normally lie on a sphere or hemisphere around the aperture. Otherwise, if the array consists of microphones spatial distributed around the acoustic scene, the search region is a point cloud inside this scene. The density of these points is in relation to the sampling frequency the theoretical limitation of the obtainable accuracy. That means for example at a sampling frequency of 48 kHz, the periodic time is 21 μs hence the minimal achievable distance between two coordinate locations is about 7 mm. Due to imperfect hardware, in practice, a distance between two grid points of several centimeters is sufficient. Also the number of candidate locations determines directly the processing time of the SRP-PHAT algorithm. For that reasons in the implementation a trade-off between accuracy (number of points) and processing time is chosen.

More precise localization techniques as mentioned above (ESPRIT or MUSIC) are in their broadband implementations computationally too intensive and show comparatively poor robustness to reverberation and low SNR conditions [22]. The TDOA based algorithms, described in section 2.5.2 are limited to one source, which does not fulfil the requirement of multiple source detection.

For the above reasons, the SRP-PHAT algorithm was selected as localizer for the discussed teleconferencing scenario.

Pure SRP-PHAT localization would produce noisy localization results. Therefore post-

filtering is needed to smooth the output. One method performing this kind of post-filtering is the so called Kalman filter [23]. But the original structure of a Kalman filter is limited to one random process and assumes a Gaussian noise distribution. Particle filters overcome these limitations, modelling arbitrary random distributions numerically as a cloud of particles (see section 2.6.2). This enables the efficient handling of multiple sources. Furthermore, for future extension it might be possible integrating video tracking directly into the filter in order to improve detection quality. Therefore this flexible architecture [3] of particle filter was chosen as post-filtering stage.

In summary, for the localization part of the teleconferencing solution, an approach based on a SRP-PHAT localizer with a following particle filter promises the best results. To avoid reinventing the wheel *ManyEars* [46, 47, 48] was selected as a basis, because it bases on a combination of a SRP-PHAT localizer followed by a particle-filter. *ManyEars* was developed during several projects at the *University of Sherbrooke, Canada*. Today it provides an easy to use 'C' library and some corresponding papers. It allows to connect specific sound interfaces and performs on-line tracking of sound sources.

Another part needed for the proposed system, is a separation method which allows to record the localized speakers around the microphone array separated as good as possible. Main interference is introduced by reverberation and competing speakers. Thus, it would be good to have a method recording each active speaker individual. The first idea was to use a beamformer, recording each speaker individually. But, as mentioned above, data-independent beamformers can only deliver a signal distorted through strong crosstalk and reverberation. So one could think, data-dependent beamforming solutions could separate the different speakers. That's correct, but in general adaptive beamforming normally needs some time to adapt the filters until delivering acceptable quality. So in practice, most algorithms induce a high delay, which is impractical during teleconferencing. These delays are also the main reason why Blind Source Separation is not the first choice for the discussed problem. One promising approach is Geometric Source Separation (GSS) [37], which is a combination of Blind Source Separation and Beamforming. It accelerates the long lasting iterative process of Blind Source Separation through geometrical informations. With some additional simplifications and optimal initialization [37], this solution can deliver nearly instantaneous[1] separation results. According to [37], SIR improvements up to 10 dB are possible. Because of the efficiency and the still good separation results, the GSS algorithm was selected for the experiments.

All the algorithms selected above depend on a microphone array. In the imagined teleconferencing scenario several participants sit around a conference table. The "recording device" should be placed in the middle of this table. Thus, the device consisting of microphones should have a omnidirectional characteristic, which fulfills a rotational symmetric microphone array [20]. Such an array has a homogeneous sensitivity in all radial directions. So a circular array was selected as basic array type. In Figure 3.1 the technical drawing of the array is shown. The diameter of 0.24 m was chosen according to other conventional table-top conferencing phones and similar structures. A further requirement was the utilization of standard components instead of professional measurement equipment. Therefore the *Focusrite Saffire PRO*

[1] depending on the adaptation rate and the initialization

40[2] was selected as multi-channel audio interface. The interface already contains preamps with phantom power allowing to connect up to eight cheap electret condenser microphone capsules[3]. These cheap microphones should not be expected to deliver hi-end studio quality in terms of noise and high frequency performance. However, the overall system consisting of off-the-shelf hardware should show that acceptable performance is achievable using advanced signal processing compensating the imperfect hardware. All recordings were made using the open source recording software *Audacity*[4] which runs on a *Windows* PC. A complete overview of the experiment setup is depicted in Table 3.2.

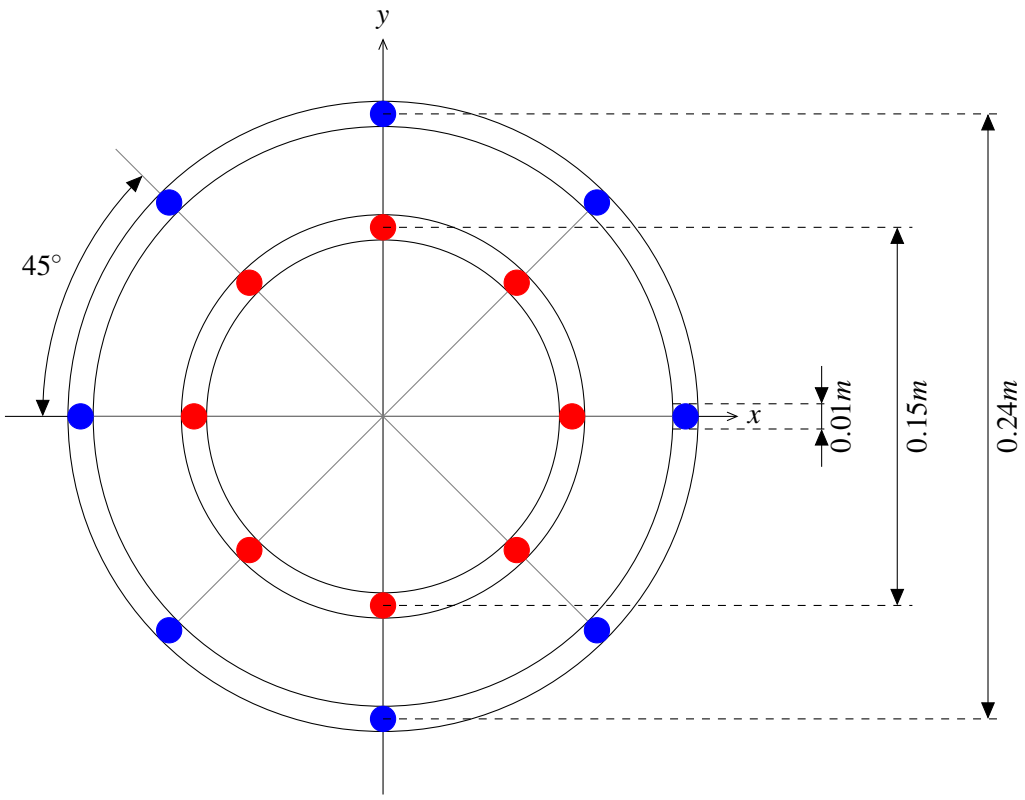


Figure 3.1.: Technical Drawing of the Circular Microphone Array

[2] `http://www.focusrite.com/global/products/audio_interfaces/saffire_pro_40` accessed September 15, 2011
[3] CUI Inc. CMB-6544PF (data sheet attached in the Appendix)
[4] `http://http://audacity.sourceforge.net` accessed September 15, 2011

3.2. ManyEars MATLAB implementation

There is a full 'C' implementation of the *ManyEars* Project[5] available. It contains the localization, tracking, and separation algorithms as proposed by *J.M. Valin* in [46, 47, 48]. In order to extent and evaluate different algorithms, a 'C' implementation is too complex and inflexible. So, for research purposes a complete transfer of the sourcecode to *MATLAB R2010b*[6] was done. This provides an easy way to integrate new features and to try various parameter sets. Also the visualization and evaluation of results is much easier using *MATLAB*. But knowing that the base source code exists as really fast 'C' implementation is a good starting position transferring the results back to practice.

The main drawback of the *MATLAB* implementation is the limitation to off-line processing. It is only possible to analyse pre-recorded data.

For the *MATLAB* implementation the system was divided into two parts. One part is the SRP-PHAT algorithm combined with the particle filter delivering the localization data. The other part is the normalization and separation process. A first implementation of the SRP-PHAT algorithm and the particle filter was already available at the *Institute for Data Processing* and was used as basis for further extensions and studies. The Geometric Source Separation implementation was new developed during this thesis. For both parts convenient control functions, for standalone processing of all experimental data, are provided.

In the following sections the main functions of the localization and separation system is explained in order to give an overview to the implementation architecture.

SRP-PHAT Localizer

Basically, the SRP-PHAT algorithm is very simple to implement. But for a fast version, some aspects need special attention. First, the coordinate candidates need only be calculated once. So, the three dimensional coordinate points of a unit sphere or hemisphere are pre-calculated and stored (Figure 3.2). Secondly, with these points and the known array configuration all delays of arrival between each microphone pair are determined in advance and also cached for further use. After this, the processing of the actual data begins. The input sound file were loaded and each channel is transferred into the frequency domain with preceding windowing (Hamming-Window, 50% overlap). This transformation makes the following energy calculation more efficient and whitening of the signal is easier. With the whitening the energy peaks of the cross-correlation can be narrowed which increases the resolution (according to [34]). In frequency domain, the whitened cross-correlation between the microphone pair ij is computed as:

$$R_{ij}(\tau) \approx \sum_{k=1}^{L-1} \frac{X_i(k)X_j(k)^*}{\|X_i(k)\|\|X_j(k)^*\|} e^{j2\pi k\tau/L}, \tag{3.1}$$

[5] `manyears.sourceforge.net` accessed September 15, 2011

[6] `http://www.mathworks.com/products/matlab/` accessed September 15, 2011

where $X(k)$ is one frame in the frequency domain and L the frame length. While this produces sharper energy peaks it has also one drawback. After the whitening, each frequency bin contributes the same amount of energy to the final correlation. This makes the system less robust against noise, because if some frequency bins are dominated by noise they were also detected beside the voice signals.

To overcome this problem, before calculating the cross-correlation, frequency weighting is applied. This weighting is based on the signal-to-noise ratio according to [15] and the noise estimate is calculated using the Minima-Controlled Recursive Average (MCRA) technique [9]. Additionally to make the system more robust to reverberation a simple reverberation model is also applied as weighting to each frame.

With this enhanced cross-correlation, a pre-defined number of sources is extracted by computing for each point of the search grid (which was previously created, Figure 3.2) the cross-correlation energy. This is done by summing up all cross-correlation values for a specific delay according to a microphone pair. The cross-correlation values must only be selected from the whole cross-correlation function as they were computed previously. So, a complete cross-correlation energy map for the search region is calculated. On this map the highest peaks correspond to the strongest signals. According to the maximum number of assumed sources, the highest peaks were iteratively selected. For each selected peak, the coordinate is stored and forwarded to the particle filter.

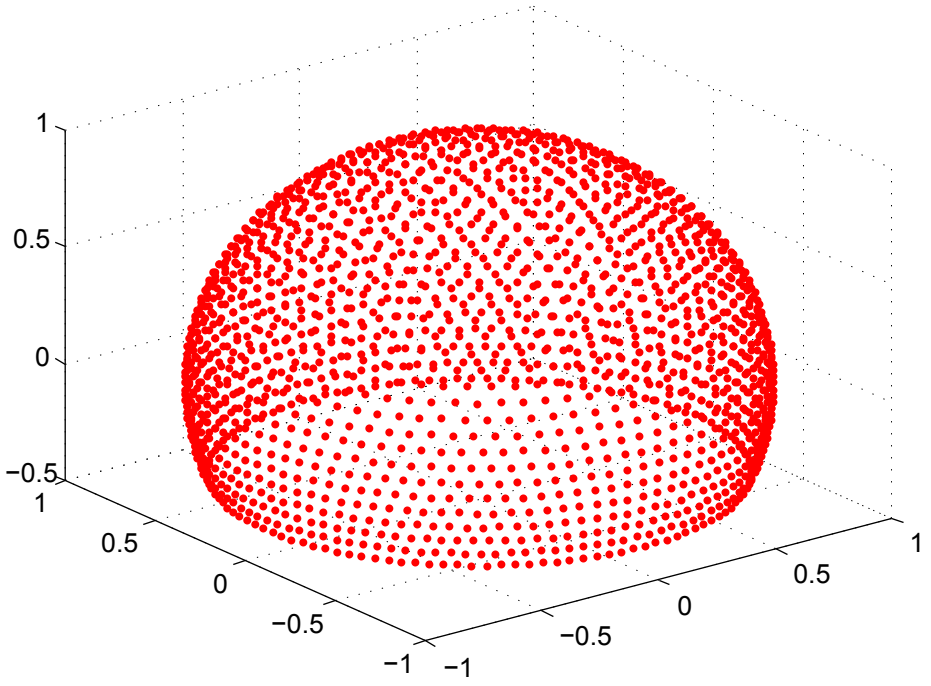


Figure 3.2.: Spherical search region (1861 points)

Particle Filter

In the following the used particle filter is described in detail. The above SRP-PHAT algorithm provides only an instantaneous, noisy measurement of possible source locations without any information about the temporal behaviour of these sources. A probabilistic temporal integration can minimize the lack of reliability delivering a continuously stable source tracking based on the SRP-PHAT estimates.

A particle filter, as mentioned above, can be used for this purpose representing the possible sources as a set of particles. These particles at the beginning are initialized with random values of their position, velocity, and state. The particles have initially uniform weighting with a sum of one over all particles.

Then in the first step for the frame-wise process of the particle filter, the particle prediction is done. Prediction means, that each particle according to its state (stopped, constant, excited) is updated. The amount of particles with a specific state also depends on the assumed state of a tracked source. So some particles stay on their position, some are moved with a constant velocity, and some are moved accelerated. The model behind these movement is called the *excitation-damping* model and is described in [54].

After this prediction step, which is primary completely independent from the measurements, each particle is weighted. Before doing this, some probabilities have to be calculated.

First a confidence value in the beamformer output is calculated depending on the beamformer energy. This probability provides informations about the observation if a potential source is a true source or a false detection. Then for each particle the probability that the observed source is detected at the particle position is calculated according to a normal distribution. This results in a probability density function which represents the beamformer error. After this, the different particle swarms have to be assigned to an observation. For this source-observation assignment problem there are three possible cases: a false detection, a source observation corresponds to a one recently tracked source, or the observation corresponds to new source that is not yet tracked. For the cases of a new source and a false detection an uniform probability density function (pdf) is assumed. Recently tracked sources are approximated by a probability density function calculated by the convolution of the beamformer error pdf and the pdf which is actually approximated by the particles.

With these probabilities and some *a-priori* probabilities, which assume the activity and appearance of speakers, for each observation and detected source corresponding probabilities are calculated on which an assignment is made [46].

After the source-observation assignment problem is solved all particles get a new weighting. This weighting is determined according to particle-source distance (the beamformer error pdf is used) and the previous particle weighting.

The final steps of one cycle of the particle filter loop are the update of the *a-priori* probabilities of the sources. Based on these, decisions for removing old sources or adding new ones are made. Then at the end, the tracked coordinate estimate is obtained through the weighted average of the particle positions, which corresponds to the mean of the approximated probability density function.

Geometric Source Separation

The implementation of the Geometric Source Separation (GSS) bases on an iterative approach which calculates frame-wise a new solution using the previous result and the new audio-frame with corresponding geometric informations. The algorithm corresponds to the theoretical calculations of Section 2.4.3.

The implementation starts with the variable initialization and calculation of the minimum delay which is at least applied to each microphone signal during its way from the source location. Additionally to the original algorithm, the tracking results of the particle filter can optionally be smoothed in case of sudden tracking loss. These missing location data is then filled with previous coordinates. This smoothing works in practice quite good, but is theoretical unnecessary if the particle filter delivers an optimal output.

After these initial steps, the autocorrelation matrix $\mathbf{R}_{\mathbf{yy}}(t,\tau)$ of the previous audio-frame is calculated for each channel in the frequency domain. The matrix $\mathbf{R}_{\mathbf{yy}}(t,\tau) = \mathbf{y}(t,\tau)\mathbf{y}(t,\tau)^H$ is an instantaneous estimation of previous output signals and, as mentioned already, does not significantly reduce the accuracy in practice.

According to equation (2.21) an estimation $\mathbf{A}(\omega)$ of the transfer function between the source and the microphone is needed. The matrix $\mathbf{A}(\omega)$ can be estimated using the coordinates of the SRP-PHAT localizer. Assuming that each microphone has unity gain (which is secured by the normalization) the elements of $\mathbf{A}(\omega)$ can be determined by

$$a_{ij}(k) = e^{-2\pi k \delta_{ij}}, \tag{3.2}$$

where δ_{ij} are the relative time delays between microphone i and source j and k denotes the frequency. This is only a reduced model of the transfer function, but works satisfactorily in practice.

With these matrices the cost functions of equation (2.21) and (2.23) were calculated. Both equation require previous results. So for the first run of the Geometric Source Separation algorithm, an initial version for the separation matrix and the previous output signal is needed. In the original paper [37] of the GSS algorithm several initializations for the separation matrix are discussed. They propose to initialize it with the filter coefficients of corresponding to a delay-and-sum beamformer. But this initialization does only make sense if the direction of the first speaker is known. If the locations are delivered by an automatic source localizer, mostly the first location estimation is not exact and it last several frames for an accurate estimation. During these frames the separation algorithm has enough time to adapt the separation matrix. So in the discussed implementation it is sufficient to initialize the first separation matrix $\mathbf{W}^0(\omega)$ with zeros. For the initial value of the previous output signal also zeros were selected.

Finally the actual separation matrix $\mathbf{W}^{n+1}(\omega)$ is calculated by

$$\mathbf{W}^{n+1}(\omega) = \mathbf{W}^n(\omega) - \mu \left[\alpha(\omega) \frac{\delta J_1(\mathbf{W}(\omega))}{\delta \mathbf{W}^*(\omega))} + \frac{\delta J_2(\mathbf{W}(\omega))}{\delta \mathbf{W}^*(\omega))} \right], \tag{3.3}$$

where μ is the adaptation rate (set to 0.01) and $\alpha(\omega)$ is an energy normalization factor equal to $\left[\|\mathbf{x}(t,\tau)\|^2\right]^{-2}$. This matrix is then multiplied with the actual audio-frame in the frequency domain (equal to a convolution in time domain) delivering the separated signals. For all following frames, this algorithm is repeated using iteratively the previous results. All resulting frames are transformed back into time domain and un-windowing is performed.

Normalization Function

Equal microphone gain is one of the requirements for good separation results. Normally exact gain levels for each input channel of the audio interface can be set digitally. But in case of the *Focusrite Saffire PRO 40* these gain values can only be set manually though analogue controls. This does not allow an equal levelling of each channel. Therefore, these manual gains were set as accurately as possible. After that, a loudspeaker was positioned directly above the array with a distance of one meter. Then artificial noise sounds were played and recorded. With these recordings, the *Normalization Function* calculates a gain factor per channel which must be applied to each following recording.

The determination of these gain factors is done by calculate the signal energy per channel for the different noise types (pink, white and brown noise). The highest energy value of a channel per noise type is selected and a corresponding gain factor for the other channels is calculated and cached. After the calculation of all gain factors for each noise type the mean value of all cached values per channel is determined. The purpose for using different noise types is, that each has its own spectral characteristics and shows specific energy distributions. So, not using only one type ensures to calculate an universal gain factor.

During the experiments, the recording of the noises were repeated at the beginning and the end of each experiment to eliminate possible thermal drifts of the preamps. So, the resulting gain factors also include these thermal effects over time.

After recording the experiments and the test noises, the gain factors were calculated and applied to each channel before performing the source separation.

3.3. Video-Tracking Enhancement

An additional extension to the original localization approach proposed by *J.M. Valin* in [46] was the introduction of a video face-tracking simulation which should assist the pure audio based approach. The reason for using only a simulation was, that the real device was not ready at the time of this thesis and that with a simulation additional errors introduced by real device can be excluded.

The implementation structure corresponds to the Figure 2.12 in section 2.6.3. This approach combines the additional video-tracking with the SRP-PHAT algorithm in a sequential manner to accelerate the localization.

For the simulation the conditional switch depending on the video confidence values was omitted, because the confidence of the generated data is always known in advance, hence the switch can be set manually.

During the simulation, the virtual video tracking device generates localization data for each recording. The data is generated based on the ground truth, which is known for each recorded experiment. These coordinates are used to generate the search region for the SRP-PHAT algorithm as shown in Figure 3.3.

In Figure 3.3 an example of a generated region of interest is shown, the two blue regions

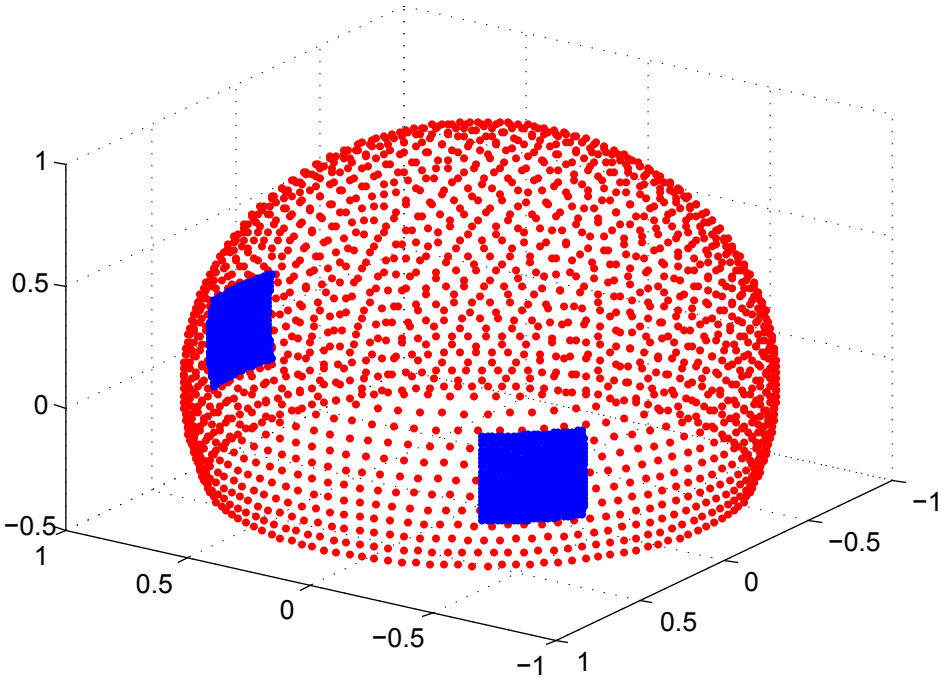


Figure 3.3.: SRP-PHAT Region of Interest constrained by the video face tracking

correspond to detected faces of the video-tracking. Only these two regions with candidate points are used in the SRP-PHAT algorithm as search region for possible sound sources. For clarification, the red points illustrate the complete search region. The point density can be chosen freely, but it is limited by the minimum detectable delay which in turn depends on the sampling rate. So, in general a smaller number of points is generated for the region of interest, as in case of the full search region. Because of the smaller regions, the particle filter has to be adapted for this configuration. For the smaller regions of interest only a reduced number of particles is needed to cover the search region. So both, the smaller search region and the reduced number of particles accelerate the localization process.

Preliminary experiments were conducted and it has been shown, that the limited search region could accelerate and stabilize the location process. But due to some adaption problems of the particle filter, no full analysis of the video extension was made. During the time this thesis took place a parallel work dealt with the real implementation of the video-tracking, so further studies and results could be found there.

3.4. Experiments and Analysis

Experiments were performed in two different environments. The first studies were done in an anechoic room, to exclude external interferences and to show that the algorithms work correctly under ideal conditions. Then the same experiments were repeated in an instrumented office room. Details of the different environments are given in Table 3.1.

Table 3.1.: Room Configurations

	Anechoic	Office
Dimensions	4.7m x 3.7 m x 2.84 m	5.41 x 3.48 m x 3.1 m
Noise Level	< 30 dBA	$\approx$ 35 - 40 dBA
Reverb. Time RT_{60}	0.08s	0.16s

In order to simulate moving sound sources, the microphone array was mounted on a turntable rotating in the azimuthal plane. For the elevation adjustment, in the anechoic room, the first loudspeaker was mounted on a metallic arch in the elevation plane allowing free adjustment of elevation. The second loudspeaker was mounted on a fixed elevation of 0°. In the office environment, both loudspeakers were mounted on tripods adjustable in height, allowing an elevation adjustment from 0° to 45°.

The same configuration of the microphone array and the loudspeakers was used in both environments. For this, the array was placed in the room surrounded by the two loudspeakers. The array position in the anechoic room was exactly in the center of the room. In the office room an almost centered position was selected. The distance r between the loudspeakers and

the array was measured for each environment. Both loudspeakers were placed with an angular distance of 135° in the azimuth plane. This basic configuration is depicted in Figure 3.4.

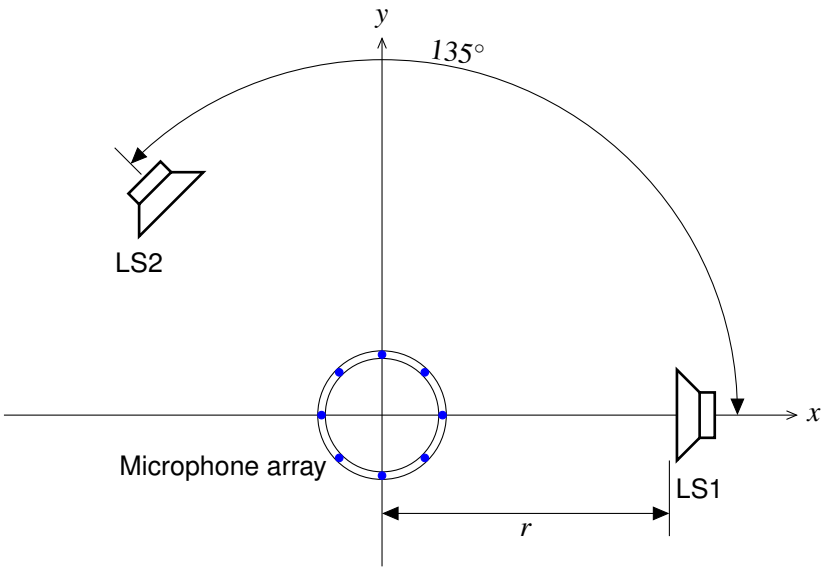


Figure 3.4.: Microphone array and Loudspeaker Configuration

With this setup, recordings with one or two simultaneously active sound sources were performed. First, three different speakers were played one after the other through loudspeaker LS1 for five azimuth values (0°, 45°, 90°, 135°, 180°). Then, during the turntable rotates, the same speakers were recorded again. The rotation between two azimuth values should simulate moving sound sources.

Similar recordings were performed for two simultaneously played sound files. This should simulate two competing sound sources. On each loudspeaker (LS1 and LS2) consecutively four different voice recordings were played. This is also repeated for five azimuth positions. The second loudspeaker (LS2) is always shifted by 135° in the azimuth plane. Besides these stationary recordings, also recordings while the turntable rotates are made. Again, as with one single source, the turntable rotates a specific angular distance in the azimuthal plane while both loudspeakers play different sound files.

The last recordings which were made using again both loudspeakers playing a simulated conversation of two speakers. The simulation consists of alternately played short sound samples (with an overlapping part) as shown in Figure 3.5. Each channel is assigned to one loudspeaker, which are positioned as described previously (Figure 3.4).

The recordings were repeated again for the five different azimuth values and the constant angular difference of 135° between the two loudspeakers.

As afore mentioned, the whole system should be evaluated in terms of the usability in conferencing environments. Therefore not only the localization of the azimuth direction is

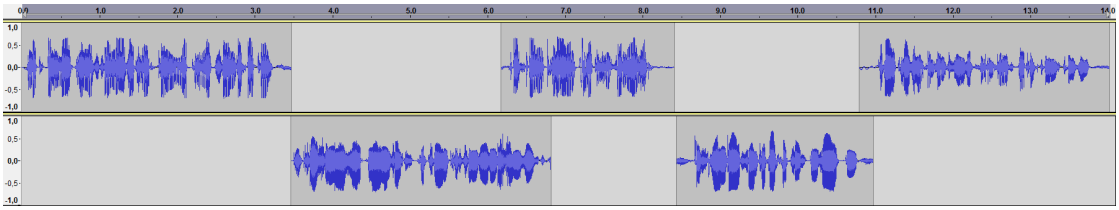


Figure 3.5.: Simulated Conversation

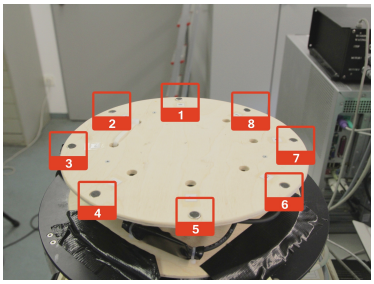


Figure 3.6.: Array Setup 1 (A1)

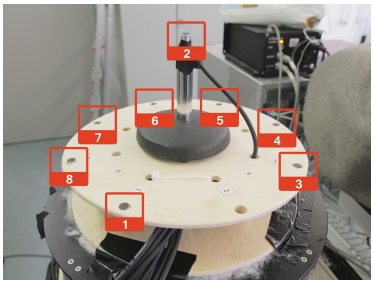


Figure 3.7.: Array Setup 2 (A2)

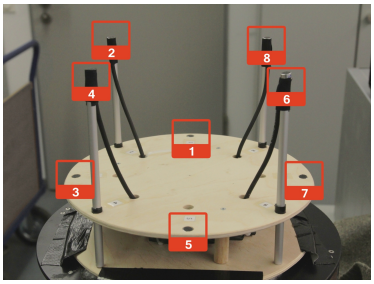


Figure 3.8.: Array Setup 3 (A3)

important, but also the accuracy of the detected elevation has to be considered. So, additional experiments were performed by repeating all above described recordings at three different elevation levels (0°, 20°, 45°). In the office room, where the loudspeakers were mounted on tripods, both loudspeakers were set to the same elevation level. In case of the anechoic room, only one loudspeaker (LS1) was altered in its elevation level, while the second loudspeaker (LS2) remains at 0° elevation.

Furthermore, to evaluate the influence of a planar array (as most commercial available devices) on the elevation localization accuracy, two more array configurations were studied. The basic planar circular array is labelled as *A1* (Figure 3.6). Figure 3.7 and 3.8 show the two modifications. In the first modification (A2), one microphone was removed from the circle of the eight microphones and were placed on top of a spacer directly in the middle of the array. The second modification (A3) places every second microphone of the previously planar circular array on its original position on top a spacer. This modification builds a symmetrical volumetric array in contrast to the second modification which is not any more rotational symmetric.

Briefly summarized, three different array configurations were used to perform recordings at five distinct azimuth positions, each on three different elevation levels. Additionally, moving sound sources were recorded using the microphone array mounted on a rotating turntable, again at the three different elevations.

All these recordings were performed using the recording hardware already described in the previous section, but for clarification consolidated in Table 3.2.

Table 3.2.: Details of Recording Hardware

Basic Array	
Number of microphones	8
Shape of array	Circular
Diameter of array	24 cm
Microphone type	CMB-6544PF
Recording Hardware	
Interface	Focusrite Saffire Pro 40
Input Channels	8
Interface	Firewire
Preamps	build in Focusrite preamps
Recording details	
Software	Audacity 1.3.13 (Beta)
Operating System	Windows XP
Sampling Rate	48 kHz
Bits per sample	16
Format	PCM WAV

4. Discussion and Conclusion

In this chapter the previously described experiments were evaluated. For that, quantitative and qualitative values from all experimental data were calculated and presented. In the conclusion these results are discussed and the solution for problem statement is evaluated. Finally, in the last section for still existing deficiencies, potential solutions and future prospects are given.

4.1. Results

For the two parts of this thesis - the localization and the separation - corresponding quality values were calculated. For the localization process the qualitative accuracy according to the ground truth and the quantitative detection rate in relation to a specified tolerance range were determined. On the other hand, for the signal separation process figures like the Signal-to-Interference Ratios were calculated describing the objective separation quality. Additionally the subjective auditory impression is briefly described.

All results are expressed in terms of azimuth and elevation of the sources relative to the microphone array. Degree was used as angular unit instead of radian. All signal-to-interference ratio values are expressed in decibels [dB].

The qualitative accuracy is given in terms of the average value over all frames with valid localization data. So only frames where the particle filter deliver valid localization data are taken into account to compute a qualitative figure for the localization.

In contrast to that, the quantitative analysis of the localization algorithm is more important for the proposed teleconferencing scenario, because frequent loss of correct tracking causes bad separation quality. So the localization success rate is given in terms of percent of time in relation to a given tolerance (it is determined for each time frame of a recording). It can be calculated for both angular planes - azimuth and elevation. But while generating the results it has become apparent, that the deviations of the azimuth values correlates with those of the elevation values. So, for clarification purposes, only the localization success ratio for the different azimuth positions were denoted.

4.1.1. Localization Results

At first, the localization success rates will be evaluated for all stationary recordings made in the anechoic room. These results were then compared to the real world recordings made in a reverberant office environment.

Figure 4.1 shows the mean values for the localization success rates for all recordings with a

single source made in the anechoic room and Figure 4.2 shows them for the office room. The mean value is calculated over 45 recordings with about 1000 frames respectively, and for each array.

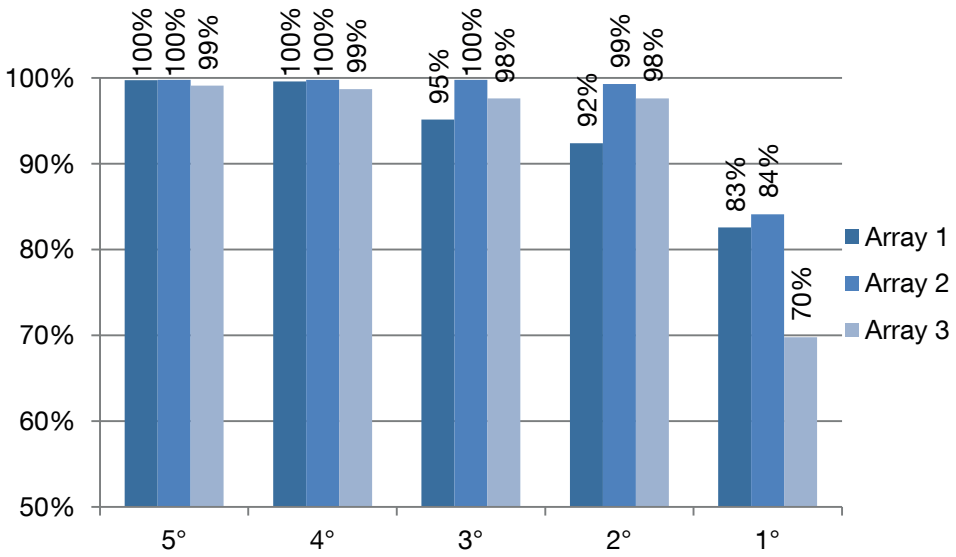


Figure 4.1.: Location Success Rates at given tolerance for a single sound source in the anechoic case. Each bar corresponds to a different array configuration

It is clearly evident, that the most robust localization is achieved under the ideal conditions of the anechoic room. But also the results for the reverberant environment are good. A localization of one single active speaker is in more than 90 percent of time with an accuracy of 2 degrees possible.

After this, the possibility detecting two simultaneous sound sources is evaluated. This test was done to show that the system is capable to handle crosstalk during conferences. Again, in the first Graph 4.3 the results for the anechoic environment show that localization of two competing sound sources is possible. In all cases, one source is localized more robust than the second one. This can be attributed to the SRP-PHAT localizer, which detects the sound sources successive according to the energy peaks of the cross-correlation function. Overall, the results are worse than for one single sound source. The difference between the source one and two is caused due to the delayed detection of the second source. The first source is detected after a short delay, then after some frames the second source is localized. This delay, which counts as false detections, is the main reason for the decreased success rate of the second source. The qualitative accuracy however is only slightly affected through competing sound sources.

However, the localization quantity is under the ideal conditions of the anechoic chamber high enough to deliver great separation results. Slightly weaker are the localization success rates for the echoic cases. There are only good rates for one source, the second source is localized

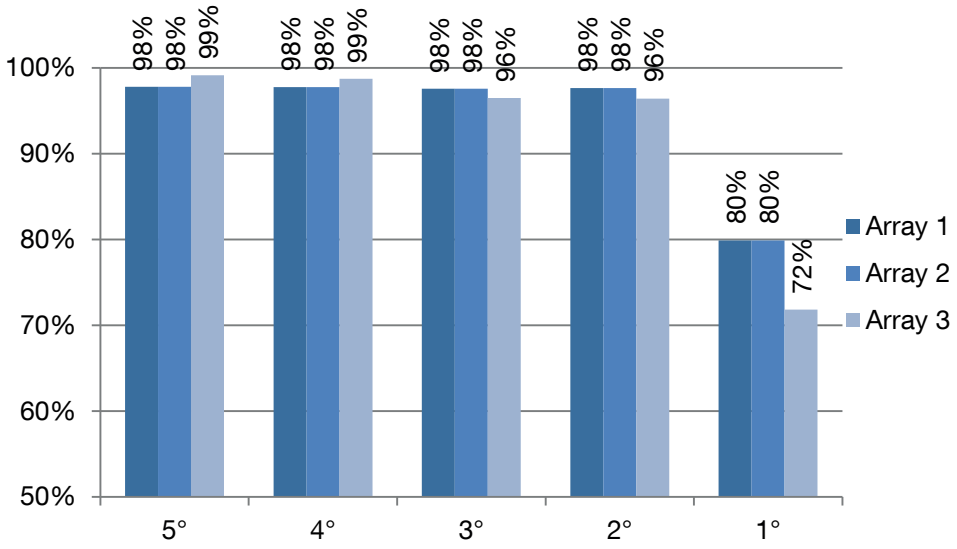


Figure 4.2.: Location Success Rates for a single source in the reverberant office room

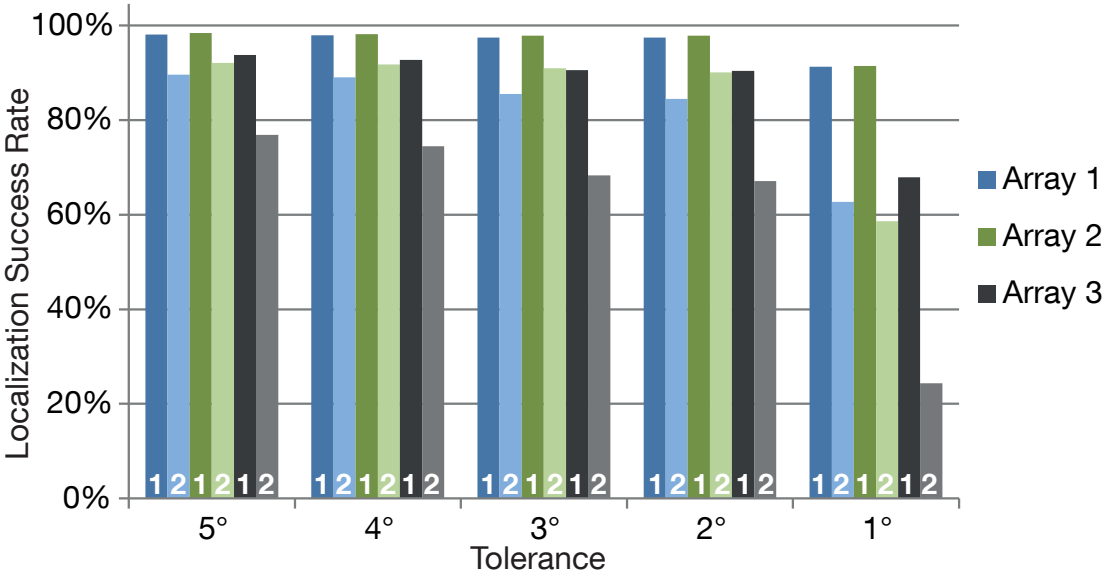


Figure 4.3.: Location Success Rates at a given tolerance of two simultaneously active sound sources in the **anechoic** case. Each light blue bar corresponds to the first detected source and the dark blue to the second

less robust and the accuracy decreases faster. But these results still suitable for acceptable separation results. Especially array two (A2) performs well and achieves success rates above 80% with an accuracy of 4 degree.

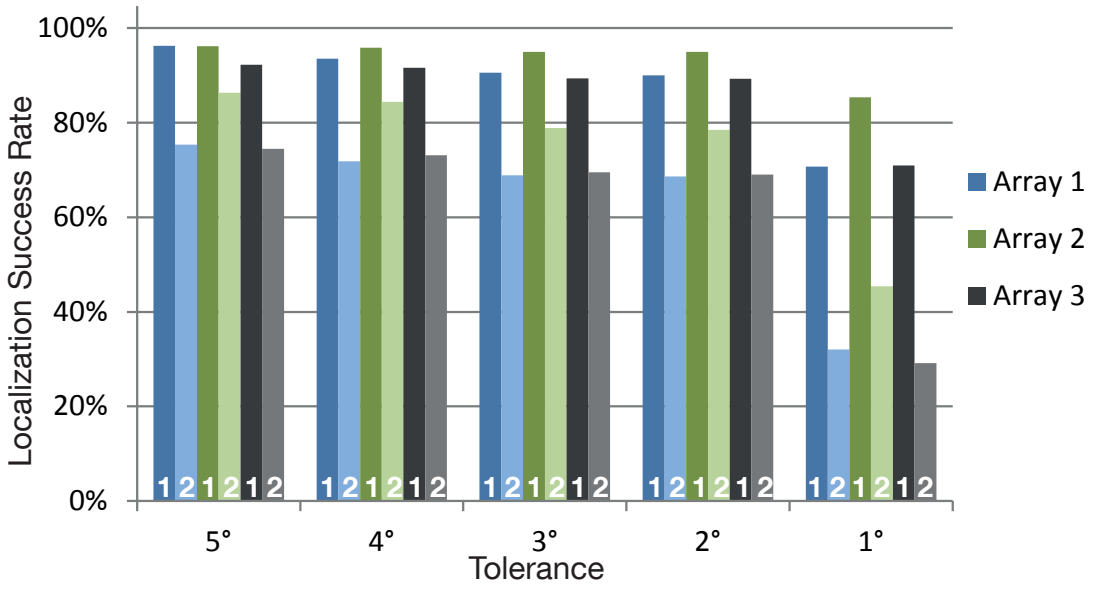


Figure 4.4.: Location Success Rates at a given tolerance of two simultaneously active sound sources in the **echoic** case. Each light blue bar corresponds to the first detected source and the dark blue to the second

Taking an overall view, the results show that a localization of the azimuth angle with an accuracy of 5 degrees is in most cases possible (more than 80 percent for two simultaneously active sources). It should be noted that the above accuracy is depicted for the azimuth value. In case of the elevation the following results show that the qualitative accuracy is not always reached and that it strongly depends on the array type.

As previously, the following figures show first the results of the ideal anechoic chamber, then the results for office room are shown. In Figure 4.5 the statistical distribution of the measured absolute elevation values is figured for the three array configurations in case of one single sound source. The statistical figures take all experiments at the same elevation level with varying azimuth positions into account. The figure show for each array configuration the results at three elevation levels, 0°, 20° and 45°. With the first array (A1), which was the planar circular array, small angles of incidence can not be detected correctly. Only the highest elevation value (45°) was correctly detected. With the volumetric arrays (A2 and A3) the accuracy clearly increases and the variances are limited. The best values are measured with the Array 3 with a maximum variance of 4 degree.

Similar results were achieved detecting two simultaneous sources. In Figure 4.6 the statisti-

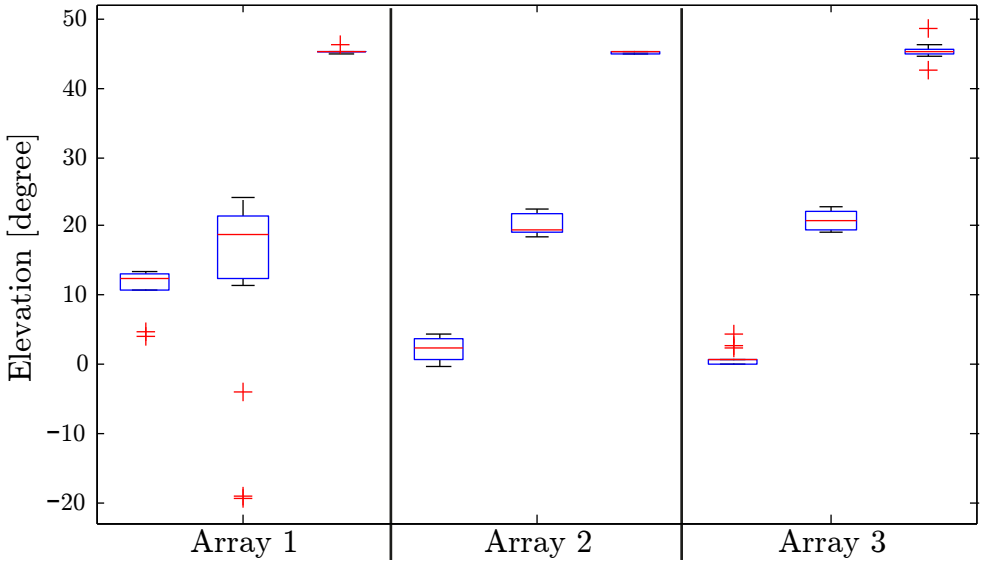


Figure 4.5.: Detected elevation for one single source at 0°, 20° and 45° (Anechoic Room)

cal analysis for the three arrays in the anechoic room are shown. Each array was tested with three different source positions. In the first experiment both sources were at 0° elevation, in the second and third, one source was set to 20° or 45° the other one remains on 0°. It can be seen, that the accuracy slightly decreases as compared to one single sound source. But the results of Array 2 and 3 are adequate enough for the separation process.

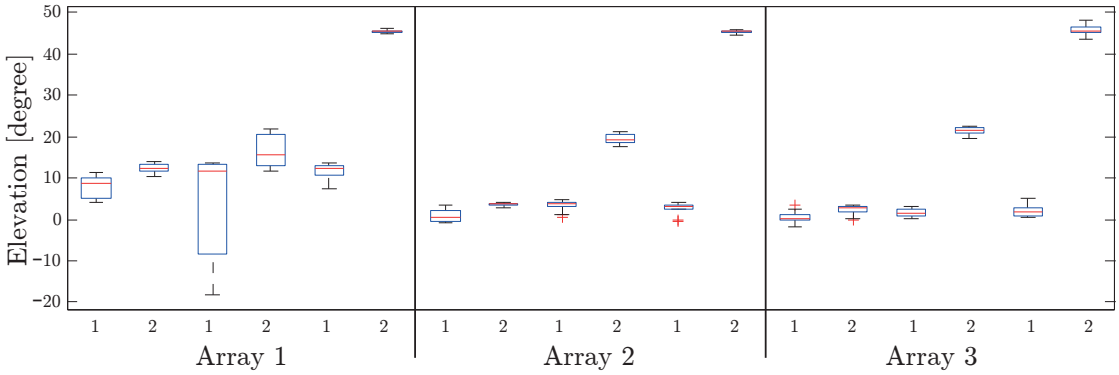


Figure 4.6.: Detected elevation values for two simultaneously played sources played in the anechoic room (ground truth: first source always at 0° the second at 0°, 20° and 45°)

In the following, these results are again compared to the real world studies. Figure 4.7 shows the results for one single active source at three different elevation levels (0°, 20° and 45°) and Figure 4.8 depicts a similar study with two simultaneously active sound sources. The

accuracy is in relation to the anechoic case slightly decreased and the values are a little bit more distributed. Together with the azimuth errors these slightly decreased accuracy in the echoic case affects the separation quality, which can be seen by the results of the separation quality analysis.

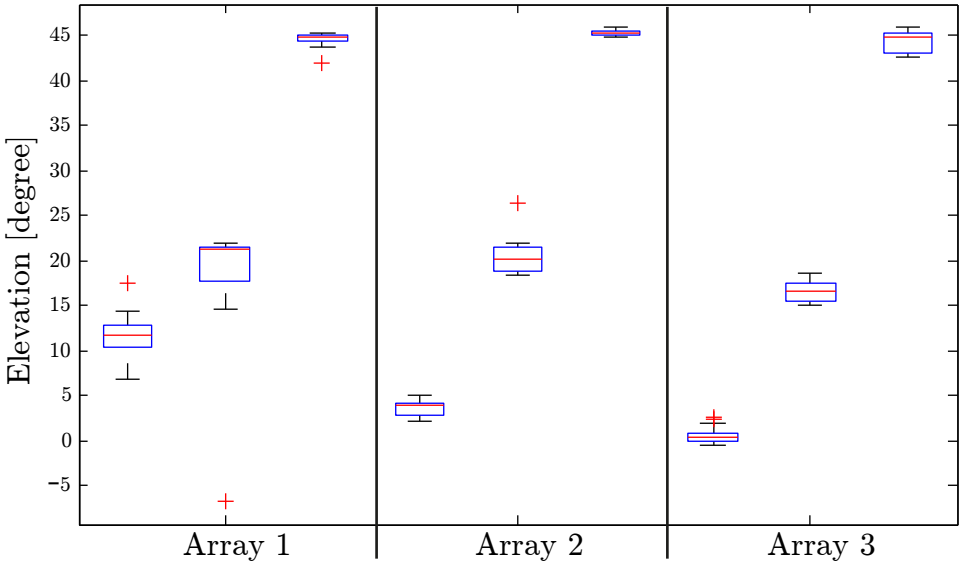


Figure 4.7.: Detected elevation for one single source at 0°, 20° and 45° (Anechoic Room)

Finally, with the above results it can be said, that a robust audio source localization with an adequate accuracy is possible. It is evident, that the various arrays perform differently and that the planar configuration of Array 1 suffers from low detection rates at the elevation plane. The volumetric arrays perform in both dimensions quite well. Altogether, Array 2 seems to be the best configuration delivering the best localization success rates and a high qualitative accuracy. To resolve the question if these results are also applicable to a teleconferencing scenario, additional experiments were performed. The interesting facts are the possibility to detect moving sources, the dynamic detection of various speakers, and the speed of the localizing process. Due to the realistic recording scenario, it was too difficult determining an exact ground truth for the automatic analysis of all recordings. Therefore the recordings with moving sound sources and dialogue simulations were exemplarily analysed. The results should show that these cases were considered through the proposed system.

At first, rotating sources were considered. As it was impossible to simulate realistically a moving human speaker, a moving sound source was simulated through the rotation of the microphone array. Therefore, the array was placed on a turntable which then rotates slowly a specific angular distance while one or two sound sources were recorded. For the echoic case, Figure 4.9 shows the results of an experiment with two simultaneously active sources while the array rotates 90 degrees. The grey dotted line indicates the ground truth. Similar results were

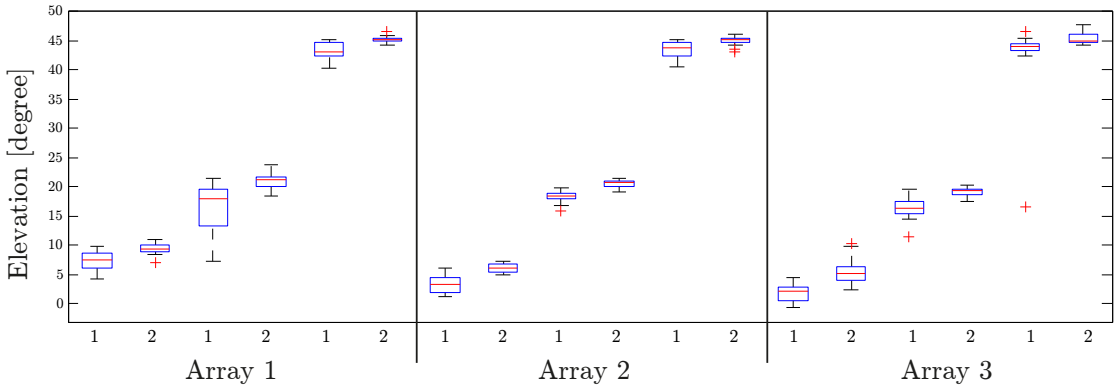


Figure 4.8.: Detected elevation values for two simultaneously played sources in the office room (ground truth: both sources were set to 0°, 20° and finally to 45°)

achieved at different elevations, angles and rotation speeds. More robust results were achieved in the anechoic chamber or with single sound sources.

It can be seen, that the localizer is able to track moving sound sources. The accuracy of this tracking is accordingly to the stationary measurements. The robustness of the localization is slightly decreased (it should be denoted, that additional noise is induced by the turntable). Single sources are tracked robustly, but similar to the stationary cases above, the second competing source is detected after several time frames. For appropriate operation of the

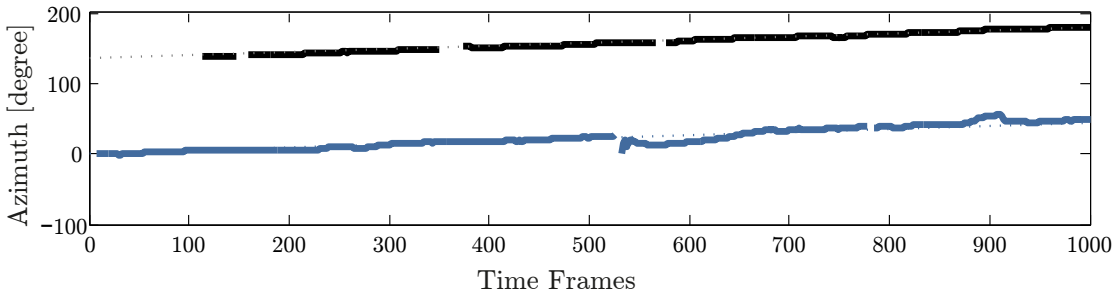


Figure 4.9.: Tracked azimuth values for two simultaneously active sound sources moving from 0° to 45° and from 135° to 180° (grey dots depict the ground truth)

conferencing device these delays are problematic because during that time, the separation process do not get localization data and is not able to separate the competing speakers. Therefore additional investigations are performed to determine the time which is needed to detect simultaneously active speakers during a conference.

The afore mentioned simulated dialogue recordings can be used to study the performance of the system during overlapping speech. As it can be seen in Figure 4.10 two overlapping parts occur during the simulation. In the ideal case a first tracked location can be delivered

after a minimum time of three frames (equal to 64 ms at a sampling frequency of 48000 kHz and a frame length of 1024 samples). Manual analysis of several tracking results reveal a delay between 3 and 7 frames using real recordings of the anechoic room. Overlapping speech extends this delay up to 10 frames. In Figure 4.11, another dialogue simulation with extended

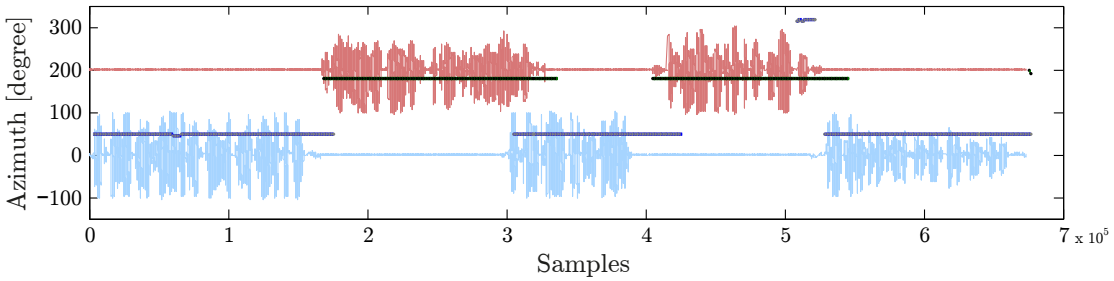


Figure 4.10.: Tracked azimuth values for the dialogue simulation recorded in the anechoic chamber (in the background the original waveforms are displayed)

overlapping parts were recorded in the office environment. Due to the strong reverberation and other room effects, the localization delay increases to 37 frames (about 0.75 s). But the detection robustness and accuracy performs well in these dialogue situations, the false detection rate remains limited and the accuracy is similar to the results above. Further investigations on additional recordings have shown corresponding results. This means for the use in a teleconferencing scenario, that overlapping parts below one second can not detected perfectly, but compared to a system without such localization and separation techniques this is a great improvement.

4.1.2. Separation Quality

With the localization results above the recordings were separated using the proposed Geometric Source Separation (GSS) algorithm. In case of one single active source, the separation process is used to perform dereverberation and noise cancellation. In situation with two simultaneously active speakers the GSS algorithm separates the signals according to the given coordinates. In accordance with the BSS EVAL toolbox, the Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR) and the Signal to Artifacts Ratio (SAR) are computed to evaluate the separation performance. To take the effects of a real test environment into account, the BSS EVAL toolbox [50] is enhanced by a gain-shift decomposition function, which allows a gain factor and a time-shift of the original signal. The BSS EVAL toolbox calculates with this decomposition function a target signal, an interference signal and an artefacts signal using a certain ground truth. This ground truth determines mainly the results and therefore must be selected carefully. For this reason, two references were selected. Firstly, the original played sound file was used, so the results contain all effects starting with the audio interface, and the room effects, and finally the separation process. In the second case, it was tried to neglect room effects and recording hardware imperfections. Therefore, for each ground truth signal a recording with only one simultaneously active speaker were separated with the GSS algorithm,

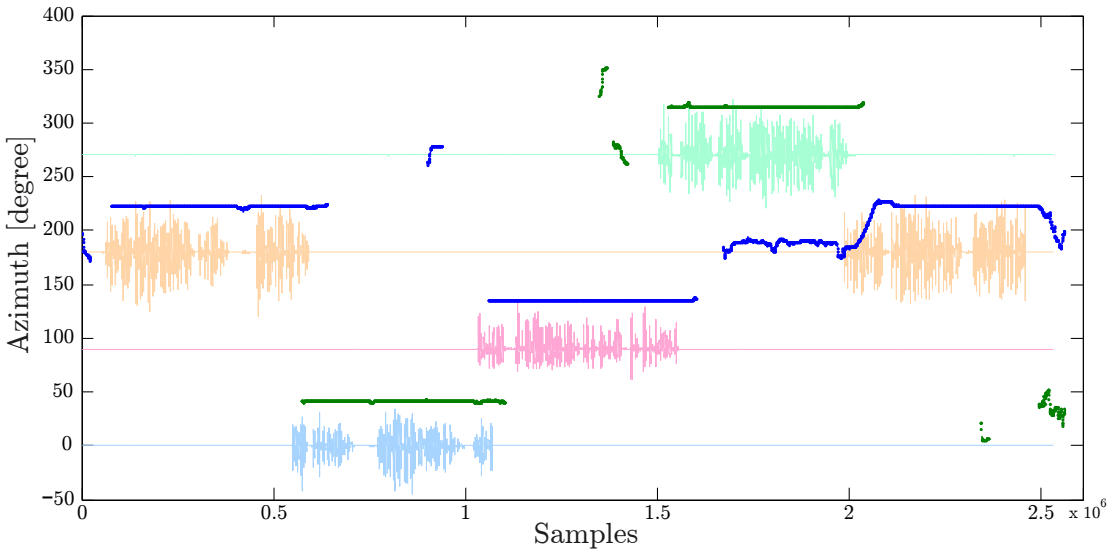


Figure 4.11.: Tracked azimuth values for the dialogue simulation recorded the office environment (in the background the original waveforms are displayed)

delivering a signal which was used to compare only the pure separation quality of two competing sound sources. With these ground truth signals the decomposition function creates the interference signals which are compared to the target signal. The results are the three different ratios – SDR, SIR and SAR – which can be used to evaluate the separation quality according to the specific ground truth signals.

Because of the circumstance that each recording was acquired manually, the start and end points of each recording vary widely. Therefore, the automatic time-shift compensation of the enhanced BSS EVAL toolbox was not able to compensate them. This has prevented a broad evaluation of all experiments. So, each recording has to be manually time aligned and evaluated, which was done for six recordings of each environment. Nevertheless, the results still depend on the correct alignment, which could not ideally achieved by hand. So the results must be interpreted with caution. In Table 4.1 the ground truth was set to the original sound file which was previously used for the recordings. Then, Table 4.2 show the different ratios for the anechoic case and the reverberant office room in relation to the individual recorded and separated signal.

Table 4.1.: Separation quality in relation to the original sound file

	Anechoic Room			Office Room		
	SDR	SIR	SAR	SDR	SIR	SAR
Source 1	-5.8628	21.1287	-5.7378	-13.2820	15.8023	-13.0311
Source 2	-6.7627	27.9194	-6.7520	-10.2631	27.5823	-10.2072

Table 4.2.: Separation quality in relation to the individual recorded speakers

	Anechoic Room			Office Room		
	SDR	SIR	SAR	SDR	SIR	SAR
Source 1	7.2185	39.8634	7.2480	-1.7768	29.2546	-1.7453
Source 2	9.7685	40.8791	9.7807	-5.8601	20.9444	-5.6647

The Signal to Interference Ratio values are quite good for the anechoic chamber as well as for the office room. Generally, the results of the anechoic environment can be seen as the ideal case and the SDR and SAR values are compared to other separation methods fair enough [50]. During the analysis, it has become apparent that the these performance figures are not suitable evaluating dialogue recordings. The influence of different starting points of each speaker in the recordings and the original sound files degenerate these values enormously. So further evaluations with more suitable recordings has to be done. But the subjective auditory impression was quite good and in case of single sources the dereverberation and noise reduction were audible. Further subjective evaluation and studies were not performed within this thesis.

4.2. Conclusion and Future Work

The goal of this thesis was to study and evaluate localization and separation technologies for immersive teleconferencing systems. All experiments and analyses of the previous chapters have yielded the following conclusions:

- Based on the shape of commercial conferencing solutions, a circular planar array consisting of eight microphone were selected as basis array configuration. Then two modifications of this shape were designed. During the experimental phase of this thesis all three types were tested equally resulting in the clear statement, that the planar array achieves weak localization accuracy in both angular directions. In contrast to that, both volumetric arrays detect both azimuth and elevation in an optimal manner. So in an innovative teleconferencing system, the recording device must consist of a volumetric array like the proposed configuration of Array 2.

- With a suitable array configuration a dynamic localization of various speakers is possible. The localization also works for more than two simultaneously active speakers. In the anechoic chamber, nearly perfect localization results were achieved using the proposed SRP-PHAT localizer. In previous works [46] already a live application using a fast implementation of a SRP-PHAT localizer was proposed. So, such a localizer can be used for an immersive conferencing solution.

- The localizer gains its robustness through the combination with a particle filter. Such a probabilistic temporal integration of the raw localization data is necessary for a robust localization in echoic environments. Only with this temporal filtering the detection of multiple sources and the dynamic detection of dialogues with vanishing sources is possible.

In addition the particle filter could be used for future extensions improving the detection quality.

- For the processing of the multiple inputs of the microphone array a Geometric Source Separation (GSS) algorithm was used. This enables separative recording of multiple speakers, and additionally through its directional characteristic echo and interferences were eliminated to a certain degree. So, during situations with multiple competing speakers this allows a better intelligibility and additional post-processing could be applied on each speech signal.

As can be seen the results of the thesis are numerous and substantial. Although the results have covered many aspects of direction finding and separation, there are areas that require further research. Additional post-filtering and a parameter estimation for the SRP-PHAT localizer show promise for improved results. Outside of the ideal conditions of the anechoic environment, the GSS approach produces still audible crosstalk which could be further suppressed through multichannel post-filtering [9]. Also the dereverberation performance could be increased through additional multichannel acoustic echo cancellation [4]. But all these enhancing filtering stages have to be evaluated in relation to resource expenses and induced delays. All techniques used this thesis could be implemented as a fast near real-time system (corresponding approaches were proposed in [47, 46]).

During the experiments it was found, that an adaption of the SRP-PHAT parameters between the echoic and anechoic achieves great accuracy improvements. Although a single parameter set of the SRP-PHAT algorithm is applicable to various environments, a fine-tuned parameter set could lead to optimal localization results. Therefore a parameter database for various environments would be useful. Also a self-calibrating parameter estimation using short autonomously initial measurements is imaginable.

A last, already discussed briefly, extension of the system would be an active generation and limitation of the region of interest. The proposed localizer uses generally a default region of interest, therefore an adaptive or video-based limitation of this search region would accelerate the detection and reduces false-detections.

It's worth emphasizing one more time that a new generation of immersive teleconferencing devices are possible. The results of this thesis show that recording devices with enhanced features like localization can improve the speech quality far more than conventional post-processing techniques improving the signal-to-noise ratio.

A. Appendix

A.1. Audio Processing Parameters

Table A.1.: Parameters of audio processing

General Audio Processing Parameters	
Sampling frequency	48 kHz
FFT length	1024
Window overlap	512
Window type	Hamming
SRP-PHAT Parameters	
Number of assumed sources	2
Search region	Hemisphere
Number of candidate coordinates	1861
Particle Filter Parameters	
Number of Particle Clouds	2
Number of Particles	800
Minimum number of frames to count a source as existing	5
Maximum number of frames while the source has not been tracked in order to delete it	30
Geometric Source Separation Parameters	
FFT length	1024
Window overlap	512
Initial separation matrix	Zeros

A.2. MATLAB Functions

A digital copy of the Matlab source code is provided alongside the thesis. The attached DVD contains the following MATLAB functions:

[analysis_and_plot_functions\]	
dialog_1_all_locs_pf.mat	Localization data used for Figure 4.10
dialog_1.wav	Wave file used for Figure 4.10
plot_figure_4_10.m	Plot function for Figure 4.10
part_3_array_2_for_figure_4_11.wav	Localization data used for Figure 4.11
part_3_all_locs_pf_for_figure_4_11.mat	Wave file used for Figure 4.11
plot_figure_4_11.m	Plot function for Figure 4.11
figure_4_5.m	Plot function for Figure 4.5
figure_4_9.m	Plot function for Figure 4.9
plot_loc_results.m	Plots directly the output of the localization algorithm
sdr_sir_sar_dir_analysis.m	Performs the SDR/SAR/SIR analysis using the BSS evaluation toolbox and calculates the results of Table 4.1 tab:sep_qual2 and 4.2 tab:sep_qual1
recorded_single_speakers_after_sep.mat	Reference data for the SDR/SAR/SIR calculation (single speakers after the separation process recorded in the anechoic room)
recorded_single_speakers_org.mat	Reference data for the SDR/SAR/SIR calculation (original recordings of each speaker)
recorded_single_speakers_z940.mat	Reference data for the SDR/SAR/SIR calculation (single speakers after the separation process recorded in the echoic environment)
loc_results_comp.m	Compares and plots the azimuth values for the given files and localizer results according to the ground truth
loc_results_comp_single.m	Compares and plots the azimuth values for a single file according to the ground truth
loc_results_comp_elev.m	Compares and plots the elevation values for the given files and localizer results according to the ground truth
loc_results_comp_elev2.m	Enhanced version of the 'loc_results_comp_elev.m' function, room type and manual elevation value can be specified
plot_all_locs_pf.m	Plots the raw localization data
plot_all_locs_pf_sig.m	Plots the raw localization data and additionally the waveform of the recorded signal

[BSS_eval_toolbox\]	
bss_eval_SiSec2008.m	Calculates the SDR/SIR/SAR values to given reference signals
[localization_and_separation\]	
srp_phat_pf.m	SRP-PHAT localizer and particle filter
process_data.m	Separates the given sound file for pre-calculated localization results and a specified array configuration
gss_sep.m	Geometric Source Separation implementation
process_data_dir.m	Separates all recordings of a given folder using pre-calculated coordinates
analyze_data.m	Performs the SRP-PHAT localization on full recording sets
process_data_file.m	Performs localization and separation on a given file for a specified microphone array configuration
create_region_of_interest_video.m	Creates the region of interest for the SRP-PHAT localizer in case of video assisted localization (uses simulated video coordinates)
analyze_data_video.m	Performs the SRP-PHAT localization on full recording sets additionally using the video tracking simulation
mics_array_1.mat	Configuration data for the circular microphone array A1
mics_array_2.mat	Configuration data for the circular microphone array A2
mics_array_3.mat	Configuration data for the circular microphone array A3
mics_array_2_hamza.mat	Configuration data for the circular microphone array A2 (used in the experiments)
mics_array_3_hamza.mat	Configuration data for the circular microphone array A3 (used in the experiments)
create_region_of_interest.m	Creates the region of interest for the SRP-PHAT localizer for given fixed coordinates
create_half_sphere.m	Creates the default hemisphere shaped search region for the SRP-PHAT localizer
apply_gain_dir.m	Applies the calculated gain factors on the recordings
calc_gains.m	Calculates the gain factors for the given noise measurements
apply_gain.m	Applies the calculated gain factors on a single recording

inv_st_fft.m	Inverses the Short Time Fourier Transformation
short_time_fft.m	Performs the Short Time Fourier Transformation
fillup_mirrored_complex_conjugate.m	Completes the second half of a complex spectrum
maxfilt1D.m	Maximum filter (for one dimension)
maxfilt1.m	Maximum filter
delays.m	Pre-calculates all delays between the points of the search region and the microphones
multictransp.m	Transposing arrays of complex matrices
multitransp.m	Transposing arrays of matrices
randomVector.m	Generates a 3D random vector
linframe.m	Generates a matrix with column number equal to the frame-size and rows according to the length of the input signal
linunframe.m	Recompose frames back to a signal
gen_window.m	Genrates a Hamming Window of a specified length
create_sphere.m	Creates full spherical search region for the SRP-PHAT localizer
coord.txt	Pre-calculated sphere coordinates
multiprod.m	Multiplying 1-D or 2-D subarrays contained in two N-D arrays
[miscellaneous\]	
directivity_pattern_tutorial_disc.m	Plots a directivity pattern for a discrete aperture
directivity_pattern_tutorial_cont.m	Plots a directivity pattern for a continuous aperture
beam_pattern3D_circular_array.m	Plots a 3D beam pattern for a circular array
directivity_pattern3D_tutorial_disc.m	Plots a 3D beam pattern for a discrete array
plot_directivity_pattern.m	Plots a directivity pattern for a discrete linear array
dsbeam.m	Performs simple delay-and-sum beamforming for a specific input signal
beam_pattern_circular_array.m	Plots a beam pattern for a circular array
directivity_pattern_circular_array.m	Plots a directivity pattern for a circular array
uniform_lin_array_polar_plot.m	Polar plot of a linear uniform array
plot_directivitiy_pattern_3D.m	Plots a 3D directivity pattern for a discrete linear array
beampattern.m	Plots a beam pattern for a discrete linear array
mmpolar.m	Polar Plot with Settable Properties
polar3d.m	Plots a 3D polar surface

A.3. DVD Content

The attached DVD contains besides the MATLAB functions described in the previous section, also the complete datasets which were used for creating the figures and tables within this thesis.

For the two different experiments, the data is clearly structured within the both Excel files:

- dataset_anechoic.xlsx
- dataset_office.xlsx

Furthermore, the complete papers stated in the bibliography are provided digitally (All rights remain with the publisher).

A.4. Microphone Capsule Data Sheet

CUI INC

page 1 of 4
date 06/2008

PART NUMBER: CMB-6544PF **DESCRIPTION:** electret condenser microphone

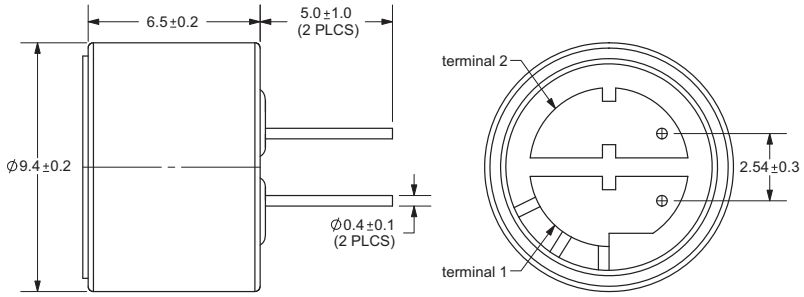
SPECIFICATIONS

directivity	omnidirectional	
sensitivity (S)	-44 ±3 db	f = 1KHz, 1Pa 0dB = 1V/Pa
sensitivity reduction (ΔS-Vs)	-3 dB	f = 1KHz, 1Pa Vs = 4.5 ~ 1.5 V dc
operating voltage	4.5 V dc (standard), 10 V dc (max.)	
output impedance (Zout)	1 KΩ	f = 1KHz, 1Pa
operating frequency (f)	20 ~ 20,000 Hz	
current consumption (I_{DSS})	0.5 mA max.	Vs = 4.5 V dc RL = 1KΩ
signal to noise ratio (S/N)	60 dBA	f = 1KHz, 1Pa A-weighted
operating temperature	-20 ~ +70° C	
storage temperature	-20 ~ +70° C	
dimensions	ø9.4 x 6.5 mm	
weight	0.7 g max.	
material	Al	
terminal	pin type (hand soldering only)	
RoHS	yes	

note: We use the “Pascal (Pa)” indication of sensitivity as per the recomendation of I.E.C. (International Electrotechnical Commission). The sensitivity of “Pa” will increase 20dB compared to the “ubar” indication. Example: -60dB (0dB = 1V/ubar) = -40dB (1V/Pa)

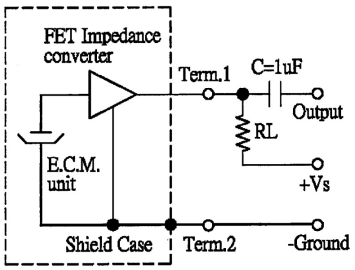
APPEARANCE DRAWING

tolerances not shown: ±0.3mm



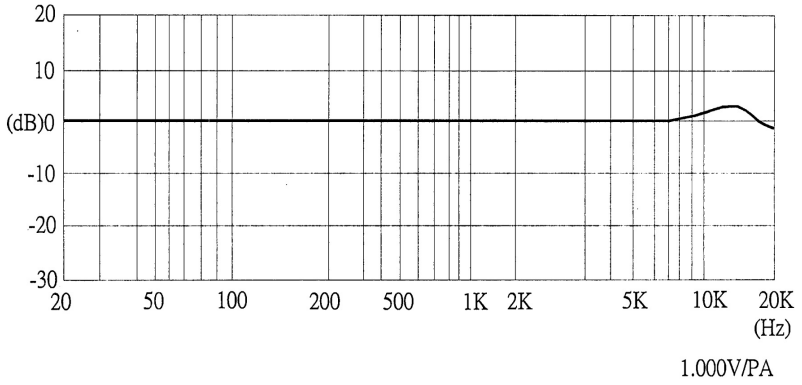
PART NUMBER: CMB-6544PF **DESCRIPTION:** electret condenser microphone

MEASUREMENT CIRCUIT



Schematic Diagram **RL=1.0KΩ**

FREQUENCY RESPONSE CURVE



List of Acronyms

DSB delay-and-sum beamformer 24
TDOA Time Delay of Arrival 32
MUSIC MUltiple SIgnal Classification
SRP Steered Response Power 30
HRTF Head Related Transfer Function 7
RIR Room Impulse Response 13
GJBF Griffiths-Jim beamformer 26
BSS Blind Source Separation 27
GSS Geometric Source Separation 27
GCC Generalized Cross-Correlation 32
PCA Principle Component Analysis 30
DOA Direction of Arrival 34
pdf probability density function 35
MCRA Minima-Controlled Recursive Average 43
SDR Signal to Distortion Ratio 60
SIR Signal to Interference Ratio 60
SAR Signal to Artifacts Ratio 60
VoIP Voice-over-IP 8

Bibliography

[1] A. Abad, D. Macho, C. Segura, J. Hernando, and C. Nadeu. Effect of head orientation on the speaker localization performance in smart-room environment. In *Proceedings of the 9th European Conference on Speech Communication and Technology*. 2005.

[2] J. Anemüller and B. Kollmeier. Amplitude modulation decorrelation for convolutive blind source separation. In *Proceddings of the 2nd international workshop on independent component analysis and blind signal separation*, pp. 215 – 220. jun. 2000.

[3] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. In *IEEE Transactions on Signal Processing*, 50(2), pp. 174 – 188, 2002.

[4] J. Benesty, S. Makino, and J. Chen. *Speech Enhancement (Signals and Communication Technology)*. Springer, 2005.

[5] J. Benesty, M.M. Sondhi, and Y. Huang (eds.). *Springer Handbook of Speech Processing*. 1st edition. Springer, 12 2007.

[6] M. Brandstein and D. Ward. *Microphone arrays: signal processing techniques and applications*. Springer Verlag, 2001.

[7] C. Busso, P. Georgiou, and S. Narayanan. Real-time monitoring of participants' interaction in a meeting using audio-visual sensors. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pp. 685 – 688. 2007.

[8] V. Capdevielle, C. Serviere, and J. Lacoume. Blind separation of wide-band sources in the frequency domain. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2080 – 2083. 1995.

[9] I. Cohen and B. Berdugo. Speech enhancement for non-stationary noise environments. In *Signal processing*, 81(11), pp. 2403 – 2418, 2001.

[10] M. Collobert, R. Feraud, G. Le Tourneur, O. Bernier, J. Viallet, Y. Mahieux, and D. Collobert. Listen: a system for locating and tracking individual speakers. In *Proceedings of the Second IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 283 – 288. 1996.

[11] D. Davis and C. Davis. *Sound System Engineering*. 2nd edition. Focal Press, 1987.

[12] J. Delosme, M. Morf, and B. Friedlander. Source location from time differences of arrival: Identifiability and estimation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pp. 818 – 824. apr. 1980.

[13] K. Diamantaras, A. Petropulu, and B. Chen. Blind two-input-two-output fir channel identification based on frequency domain second-order statistics. In *IEEE Transactions on Signal Processing*, 48(2), pp. 534 – 542, 2000.

[14] J. DiBiase, H. Silverman, and M. Brandstein. *Microphone arrays: Signal processing techniques and applications*, chapter 8, pp. 157 – 180. Springer Verlag, 2001.

[15] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(6), pp. 1109 – 1121, 1984.

[16] C. Fancourt and L. Parra. The coherence function in blind source separation of convolutive mixtures of non-stationary signals. In *Proceedings of the IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing*, pp. 303 – 312. 2001.

[17] J. Feldmaier. *Erstellung eines Telefonkonferenzsystem-Framework mit räumlich lokalisierbaren Teilnehmern*. Bachelor thesis, Technische Universität München, 2010.

[18] O. Frost. An algorithm for linearly constrained adaptive array processing. In *Proceedings of the IEEE*, 60(8), pp. 926 – 935, aug. 1972.

[19] L. Griffiths and C. Jim. An alternative approach to linearly constrained adaptive beamforming. In *IEEE Transactions on Antennas and Propagation*, 30(1), pp. 27 – 34, jan. 1982.

[20] E. Hulsebos, T. Schuurmans, D. de Vries, and R. Boone. Circular microphone array for discrete multichannel audio recording. In *Preprints of the Audio Engineering Society*, 2003.

[21] M. Isard and A. Blake. Condensation—conditional density propagation for visual tracking. In *International journal of computer vision*, 29(1), pp. 5 – 28, 1998.

[22] A. Johansson, G. Cook, and S. Nordholm. Acoustic direction of arrival estimation, a comparison between root-music and srp-phat. In *IEEE Region*, volume 10.

[23] R. Kalman et al.. A new approach to linear filtering and prediction problems. In *Journal of basic Engineering*, 82(1), pp. 35 – 45, 1960.

[24] M. Kawamoto, K. Matsuoka, and N. Ohnishi. A method of blind separation for convolved non-stationary signals. In *Neurocomputing*, 22(1 – 3), pp. 157 – 171, 1998.

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

[26] R. Kumaresan and D. Tufts. Estimating the angles of arrival of multiple plane waves. In *IEEE Transactions on Aerospace and Electronic Systems*, AES-19(1), pp. 134 – 139, jan. 1983.

[27] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Proceedings of IEEE International Conference on Image Processing*, volume 1, pp. 900 – 904. 2002.

[28] B. Loesch and B. Yang. Comparison of different algorithms for acoustic source localization. In *ITG Fachtagung Sprachkommunikation*, 2010.

[29] S. Makino, T. Lee, and H. Sawada. *Blind speech separation*. Signals and Communication Technology. Springer, 2007.

[30] I. McCowan. *Microphone arrays: A tutorial*. Technical Report, Queensland University, Australia, 2001.

[31] M. Möser. *Messtechnik der Akustik*. 1st edition. Springer, 2010.

[32] S. Nayar. Catadioptric omnidirectional camera. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 482 – 488. 1997.

[33] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough. A joint particle filter for audio-visual speaker tracking. In *Proceedings of the 7th international conference on Multimodal interfaces*, pp. 61 – 68. ACM, 2005.

[34] M. Omologo and P. Svaizer. Acoustic event localization using a crosspower-spectrum phase based technique. In *IEEE Proceedings on Acoustics, Speech, and Signal Processing*, volume 2, pp. 273 – 276. apr. 1994.

[35] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato. A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization. In *Proceedings of the 10th international conference on Multimodal interfaces*, pp. 257 – 264. ACM, 2008.

[36] L. Parra and C. Spence. Convolutive blind separation of non-stationary sources. In *IEEE Transactions on Speech and Audio Processing*, 8(3), pp. 320 – 327, may 2000.

[37] L. Parra and C. Alvino. Geometric source separation: Merging convolutive source separation with geometric beamforming. In *IEEE Transactions on Speech and Audio Processing*, 10(6), pp. 352 – 362, 2002.

[38] R. Roy and T. Kailath. Esprit-estimation of signal parameters via rotational invariance techniques. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(7), pp. 984 – 995, jul. 1989.

[39] J. Scheuing and B. Yang. Disambiguation of tdoa estimates in multi-path multi-source environments (datemm). In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4. may 2006.

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

[41] R. Schmidt. A new approach to geometry of range difference location. In *IEEE Transactions on Aerospace and Electronic Systems*, AES-8(6), pp. 821 – 835, nov. 1972.

[42] S. Shamsunder and G. Giannakis. Multichannel blind signal separation and reconstruction. In *IEEE Transactions on Speech and Audio Processing*, 5(6), pp. 515 – 528, nov. 1997.

[43] E. Shriberg, A. Stolcke, and D. Baron. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Proceedings of Eurospeech*, volume 2, pp. 1359 – 1362. 2001.

[44] P. Townsend. *Enhancements to the Generalized Sidelobe Canceller for Audio Beamforming in an Immersive Environment*. Master's thesis, University of Kentucky, 2009.

[45] R. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. In *IEEE Journal of Robotics and Automation*, 3(4), pp. 323 – 344, 1987.

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

[47] J. Valin, J. Rouat, and F. Michaud. Enhanced robot audition based on microphone array source separation with post-filter. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pp. 2123 – 2128. 2004.

[48] H. Van Trees and J. Wiley. *Optimum array processing*. Wiley Online Library, 2002.

[49] J. Vermaak, M. Gangnet, A. Blake, and P. Perez. Sequential monte carlo fusion of sound and vision for speaker tracking. In *Proceedings of 18th IEEE International Conference on Computer Vision*, volume 1, pp. 741 – 746. 2001.

[50] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca. First stereo audio source separation evaluation campaign: Data, algorithms and results. In *In Proceedings of 7th International Conference on Independent Component Analysis and Signal Separation*, pp. 552 – 559. 2007.

[51] P. Viola and M. Jones. Robust real-time face detection. In *International journal of computer vision*, 57(2), pp. 137 – 154, 2004.

[52] D. Wang and G. Brown. *Computational auditory scene analysis: Principles, algorithms, and applications*. IEEE Press, 2006.

[53] H. Wang and P. Chu. Voice source localization for automatic camera pointing system in videoconferencing. In *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 187 – 190. 1997.

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

[55] B. Widrow and S. Stearns. *Adaptive signal processing*. Prentice-Hall signal processing series. 1985.

[56] D. Zotkin, R. Duraiswami, and L. Davis. Joint audio-visual tracking using particle filters. In *EURASIP Journal on Applied Signal Processing*, 2002(1), pp. 1154 – 1164, 2002.

[57] Q. Zou, Z.L. Yu, and Z. Lin. A robust algorithm for linearly constrained adaptive beamforming. In *Signal Processing Letters, IEEE*, 11(1), pp. 26 – 29, jan. 2004.