Cost-effective Spatial Decomposition Method Solution for Impulse Response Capture and Auralization

VICTOR SIMONSSON

Department of Architechture and Civil Engineering

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023
www.chalmers.se

Master's thesis 2023

Cost-effective Spatial Decomposition Method Solution for Impulse Response Capture and Auralization

VICTOR SIMONSSON

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Architechture and Civil Engineering
Division of Applied Acoustics
Audio Technology Group Chalmers University of Technology
Gothenburg, Sweden 2023

Cost-effective Spatial Decomposition Method Solution for Impulse Response Capture and Auralization

VICTOR SIMONSSON

Supervisor: Sergejs Dombrovskis, CEVT
Examiner: Jens Ahrens, Department of Architecture and Civil Engineering

Master's Thesis 2023
Department of Architecture and Civil Engineering
Division Applied Acoustics
Audio Technology Group
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: SDM Array

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2023

Cost-effective Spatial Decomposition Method solution for Impulse Response Capture and Auralization
VICTOR SIMONSSON
Department of Architecture and Civil Engineering
Chalmers University of Technology

Abstract

Spatial decomposition method (SDM) is a method to parameterize a spatial room impulse response into a scalar pressure value and a direction of arrival (DOA) for each time sample. SDM can be therefore be used to simulate acoustics in enclosed spaces and have proven effective when rendering to both headphones and loudspeaker arrays. The aim of this thesis was to design and build a low-cost SDM microphone array using MEMS microphones and investigate how different number of microphones affect auralization performance. All impulse response capture, encoding and decoding was implemented with Python, to be published as an open source library. Python modules were implemented to support rendering as both BRIRs and as virtual loudspeaker signals to be used in external auralization software. Due to hardware limitations the SDM array was limited to 6 microphones, although simulations were carried out to compare a 6 microphone array and a 12 microphone array. These were shown not to have any significant perceptual differences as expected from prior research. Listening tests were conducted where participants compared binaural room impulse responses decoded from SDM with a reference dummy head measurement. Two measurement situations were evaluated, clear line of sight and occlusion between source and receiver. The listening tests showed that neither measurement situation performed better than the other, relative to their respective reference and anchor. No clear difference between the ratings of different configurations of SDM encoding were found. All in all, this implementation proved sufficient at encoding and decoding binaural SDM. With additional post processing the audio quality will get even better.

Keywords: Acoustics, Audio Technology, Spatial Audio, Microphone Array, Binaural Audio

Acknowledgements

I would first like to thank Sergejs Dombrovskijs and CEVT for the opportunity to take on such a fascinating project as this one.

I also want to thank my examiner, Jens Ahrens, for his guidance and support during this project. All the time you have set aside for me to discuss these topics has been much appreciated and have really made me understand these concepts well. Also, a big thank you to Hannes Helmholz, Thomas Deppich and Leon Müller for the discussions and help with hardware, measurement setups and discussions around the theory of this project.

Cheers!

Victor Simonsson, Gothenburg, June 2023

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

BRIR	Binaural Room Impulse Response
DOA	Direction of Arrival
HRIR	Head Related Impulse Response
HRTF	Head Related Transfer Function
IR	Impulse Response
ISM	Image Source Method
LTI	Linear Time Invariant
NLS	Nearest Loudspeaker Synthesis
MEMS	Micro Electro Mechanical System
MUSHRA	Multiple Stimuli Hidden Reference and Anchor
PDM	Pulse Density Modulation
PIV	Pseudo Intensity Vector
RIR	Room Impulse Response
SDM	Spatial Decomposition Method
SRIR	Spatial Room Impulse Response
SOFA	Spatially Oriented Format for Acoustics
TDOA	Time Difference Of Arrival

Contents

List of Figures

List of Tables

1

Introduction

This thesis is in the field of spatial audio capture, encoding and auralization and is done in collaboration with CEVT. Today many different methods are used to capture and process spatial audio which all utilize different processing methods to extract directional and timing information from an IR. This thesis concerns the Spatial Decomposition Method (SDM), its functionality and its advantages and disadvantages compared to other spatial audio capture methods. SDM was developed at Aalto University as a way to decompose a spatial room IR into a set of image sources and scalar pressure values [1].

1.1 Background

Most methods to simulate room acoustics today are quite limited. Most tools used to "add reverb" require specialized microphone that are quite expensive. SDM on the other hand is not limited to a number of channels for output, but instead uses IRs for a number of microphones at known positions and estimate a direction of arrival at each sample in time with essentially unlimited spatial resolution.

For this project, off-the-shelf MEMS (Micro-electro-mechanical system) microphones are used for capture of the spatial data. These are much more compact and cheaper than the dynamic or condenser microphones used in most applications of SDM microphone arrays. Even though MEMS microphones need special hardware to be driven and mostly require some soldering the total cost will be much lower with an implementation like this.

SDM assumes that a sound field can be broken down into a pressure and direction per time sample. The recorded IRs are split into blocks where the calculations of each block determines the Direction of arrival (DOA) at that point in time. This information is then encoded into a directional room IR containing scalar pressure values and averaged direction of incidence.

Decoding and binaurally rendering the measured room IR proves very useful in areas where the acoustic quality of an enclosed space may want to be evaluated remotely, i.e simulating the acoustic character of a space with any excitation signal in headphones whilst keeping the spatial content intact.

As we know from small enclosed spaces, such as a car cabin or a small domestic space, the energy decay time is very short. With such a short IR to extract spatiotemporal information from one could assume that SDM is not well suited for characterizing but previous research [2] shows that SDM in fact can be valid in these smaller spaces given that the window sizes and dimensions of the array are small enough. One aspect that has not been thoroughly researched is occlusion of the source which will be investigated in this thesis.

1.2 Research Questions

The aim of this thesis is to design and build an SDM array and implement SDM encoding and decoding in python. Decoding will be implemented for rendering binaural room IRs as well as loudspeaker signals to be used with head tracking in preexisting audio software.
The thesis will compare how different number of SDM microphones perform as well as the SDM arrays performance against a reference dummy head in a listening experiment. Two different measurement situations are aimed to be evaluated:

- clear line of sight between microphone and source
- some occlusion between microphone and source.

2

Theory

In this chapter, the general theory and concepts of sound and room acoustics, IR capture, SDM and Binaural Auralization will be explained.

2.1 Fundamentals of Sound and Reverberation

Sound is, in essence, fluctuations of the air pressure that is perceived by our ear drum as it moves into our ear. When a sound is excited by a source e.g. a loudspeaker or the vocal cords in our throat, the mass of the exciting body pushes the air in front of it creating an oscillating air pressure that radiates the energy out from the source. This one dimensional wave is called a longitudinal wave and is what we perceive as sound. Figure 2.1 shows how the compression and refraction of air molecules over the one-dimensional plane wave creates pressure fluctuations around the atmospheric air pressure.

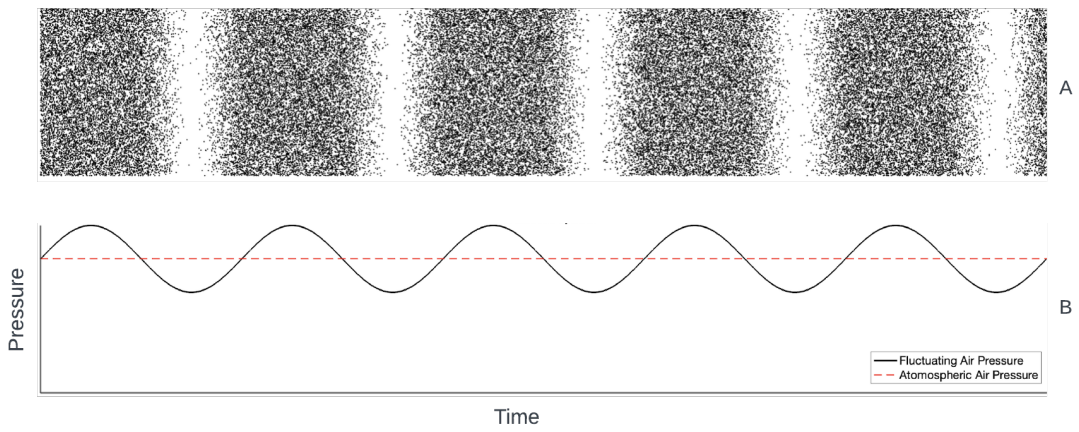

Figure 2.1: Visualization of a planar wave in air. A) The particle density over time. B) The resulting air pressure over time.

This is what is usually referred to as a sound wave. This can be expressed in three dimensions with the wave equation [3]

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

where

$$\nabla^2 p = \frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} + \frac{\partial^2 p}{\partial z^2} \tag{2.2}$$

2.1.1 Reverberation

When exciting a sound in an enclosed space, a listener located in that space will perceive the combination of two sound fields, free field and diffuse field. Because of the fact that sound never only travels in one straight line to the receiver, reflections of the sound appear when the sound interacts with the limiting surfaces of an enclosed space such as walls and other obstructions. All these reflections, combined with the direct sound that the receiver hears, make the sound field of a room. The direct sound is the sound from the source that interacts solely with the propagation medium before reaching the receiver. The early reflections are the sparse copies of the direct sound that appear just after the direct sound. After a short period of time, the reflections in the room start interfering with each other resulting in the diffuse nature of the late reverberation. Figure 2.2 depicts energy decay in a room impulse response (RIR), divided into direct sound, early reflections and late reflections.

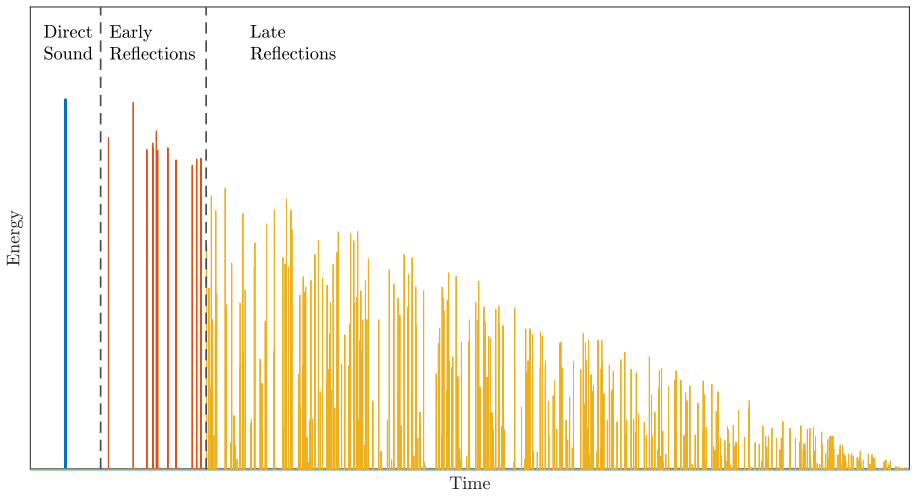

Figure 2.2: Visualization of energy decay in a room divided into direct sound, early and late reflections

Generally, when measuring and evaluating a rooms acoustics a measurement of the IR is taken. Let us consider a room as a linear time-invariant system (LTI). This

gives us a signal flow like the one depicted in figure below. In Figure 2.3 there is an input, x(t), which for this case is a loudspeaker generating a sine sweep. The room, h(t), affects the signal excited by the loudspeaker and the affected signal is captured by the microphone, which is the output of the system, y(t).

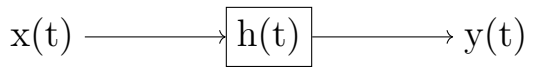

Figure 2.3: Flow chart of an LTI system $h(t)$ with input $x(t)$ and output $y(t)$

With these signals known, one can perform what is called deconvolution. Deconvolution is an inverse convolution which instead of combining two functions, with known input and output of a system, extract the systems transfer function or IR.

With this system identification we can know which frequencies have resonances in the system, how long the reverberation time is and a number of other parameters.

2.2 Spatial Decomposition Method

The Spatial Decomposition Method is a method to parameterize a spatial room impulse response (SRIR) for processing and auralization [1]. It is based on the assumption that at any given sample in time of a room IR, there is an acoustic event with a given DOA $\Omega(t)$, with pressure p(t). The method assumes that with this information at each time frame, the full sound field is characterized.

Typical choices of microphone arrays have been 3D intensity probes, consisting of six omnidirectional capsules at the centers of the faces of an imagined cube, but other configurations have been implemented.

2.2.1 Direction Of Arrival Estimation

Since SDM assumes one single reflection at a specific point in time which, during the encoding, results in a series of DOAs. To estimate the DOA in SDM, one of two methods are used. Time Difference of Arrival Estimation (TDOA) or Pseudo Intensity Vector (PIV). The two analysis methods were developed independently and have been proven to be suitable for different microphone arrays [4]. The following will describe how these work and in what situations each might be applicable.

2.2.1.1 Pseudo Intensity Vector

PIV estimates direction of arrival based on the directional component of sound intensity i.e. a vector in the direction of velocity flow. Sound intensity $\vec{I}$ is defined as

$$\vec{I} = p\vec{v} \tag{2.3}$$

where p is the pressure measured and $\vec{v}$ is the particle velocity.

which involves measuring the pressure and estimating a velocity in three coordinate axes x, y and z [4].

$$\theta(t) = h_w(t)\left[h_x(t)h_y(t)h_z(t)\right] \tag{2.4}$$

This method is most commonly used with closed arrays with microphones with cardioid directivities such as B-Format microphones, like the Core Sound Tetra Mic, or other closed arrays and will therefore not be used to do the SDM encoding in this thesis.

2.2.1.2 Time Difference Of Arrival Estimation

The DOA can also be calculated by estimating the Time Difference of Arrival (TDOA) at each acoustic event. In the paper "Spatial Decomposition Method for Room IRs" [1], Tervo et al, proposes estimating TDOA using the least square solution for localizing the image sources. Using the general correlation method with direct weighting [5] for each step in the window N, the TDOA can be obtained.

For TDOA estimation, the responses of an open array of at least four microphones that are not in the same plane are used to obtain the direction within a block of size N, centered around each sample t.

First the cross correlation between the capsules in a given pair is done with

$$R_{i,j}(n) = \frac{1}{N}\sum_{t=1}^{N} h_i(n)h_j(t+n) \tag{2.5}$$

where N is the window size set, h_i and h_j are the signals for both of the microphones. The timing offsets between the different microphones can then be determined by the index of the max value in $R_{i,j}$.

$$\tau_{i,j} = \mathrm{argmax}R_{i,j}(n) \tag{2.6}$$

where $\tau_{i,j}$ is the time difference of arrival between a pair of microphones. To interpolate the cross correlated time delays with sub-sample accuracy $c_{i,j}$ is computed with 2.7.

$$c_{i,j} = \frac{\mathrm{ln}R_{i,j}(\tau+1) - \mathrm{ln}R_{i,j}(\tau-1)}{4\mathrm{ln}R_{i,j}(\tau) - 2\mathrm{ln}R_{i,j}(\tau-1) - 2\mathrm{ln}R_{i,j}(\tau+1)} \tag{2.7}$$

The TDOA estimates are then defined for each pair of microphones as

$$\boldsymbol{\tau} = [\tau_{1,2} + c_{1,2}, \tau_{1,3} + c_{1,3}, ..., \tau_{M-1,M} + c_{M-1,M}]^T \tag{2.8}$$

where M are the total number of microphones in the array. The position vector of the distances between each pair is described as

$$\mathbf{V} = [\mathbf{r_1} - \mathbf{r_2}, \mathbf{r_1} - \mathbf{r_2}, ..., \mathbf{r_M} - \mathbf{1} - \mathbf{r_M}] \tag{2.9}$$

where r_i is the position of each microphone in cartesian coordinates. Using $\mathbf{V}$ and $\boldsymbol{\emptyset}$ for solving the least square solution gives us slowness vector $\mathbf{k}$ as

$$\mathbf{k} = \mathbf{V}^{+}\boldsymbol{\tau} \tag{2.10}$$

which in turn solves the DOA Ω with

$$\boldsymbol{\Omega} = \frac{\mathbf{k}}{\|\mathbf{k}\|} \tag{2.11}$$

This output, Ω, is then a set of cartesian coordinates of the same length as the impulse response measured.

The minimal block size N_{min} should correspond to at least twice the TDOA for sound incidence from a direction parallel to the larges distance between two microphones d_{max} [4],

$$N \overset{!}{\geq} N_{min} = T_{min} f_s = 2\frac{d_{max}}{c} f_s \tag{2.12}$$

Since TDOA estimates the direction of arrival based on when an acoustic event impinges at a point in space, the usual setup is an open array with omnidirectional microphones. This is the design of the array constructed in this thesis so TDOA is used in this thesis. TDOA has been implemented in Matlab [6] and will be ported to python in this project.

2.2.2 DOA Smoothing

DOA smoothing is a processing step often applied right after the DOAs are estimated. Due to the nature of TDOA estimations being quite chaotic and jumping all over the place, some smoothing of these values have been proven to be quite useful in auralization. By computing a rolling average over a window K these quickly varying angles of incidence the DOAs can become more focused around the source position.

$$SMA_K = \frac{1}{K} \sum_{i=N-K+1}^{M} p_i \tag{2.13}$$

Figure 2.4 displays the impact of adding a simple moving average of order K = 6 and K = 15. As explained, adding smoothing to the DOAs after encoding makes these quickly varying DOAs become more controlled in a sense which can improve the accuracy of the auralization.

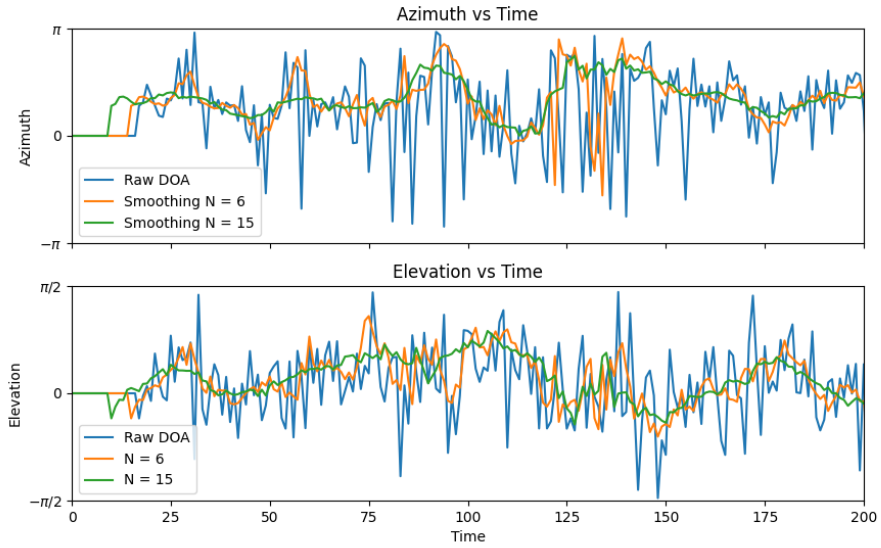

Figure 2.4: Azimuth and Elevation vs time plotted as raw DOAs and with added smoothing factor K = 6 and K = 15

Due to the nature of TDOA estimation, the late reverberation will be scattered and random in nature. DOA smoothing can then mitigate some of this for a better auralization results.

2.3 Binaural Auralization

To auralize a signal is to listen to an audio file as it would sound in a measured or simulated space [7]. To auralize a IR for a listener, two methods can be applied. Either to generate a binaural room IR using Head Related Transfer Functions (HRTFs)

or by rendering audio to loudspeaker arrays where a with known source and listener positions.

2.3.1 Head Related Transfer Function

For headphone auralization, HRTFs are used to act as virtual loudspeakers with a given source coordinate relative to the listener position.

An HRTF is a way to describe how the body, head and outer ears affect the sound waves impinging on a listener from a specific direction. Looking at how humans locate a sound source with aural cues, a few methods seem clear. When locating a sound source to the left or right of a listener, the level difference between the ears is a clear indication to where the source is. This is not the case though for locating a sound behind or in front of the listener which is where the HRTFs make the most difference. Seeing how our ears are not symmetrical over any axis makes it clear how the ear shape impacts the incoming sound in a unique way for each angle of incidence. This together with the placement on our heads, the shape of our facial features and other factors make HRTFs a quite effective way to virtualize directionality of sound in 3D space. The HRTF can also be described in time domain, a so called Head Related Impulse Response (HRIR) but for the purposes of this thesis, the term HRTF will be used throughout. Figure 2.5 displays HRTFs and HRIRs for left and right ear at given source position $\theta = 83°$ and $\phi = -2°$

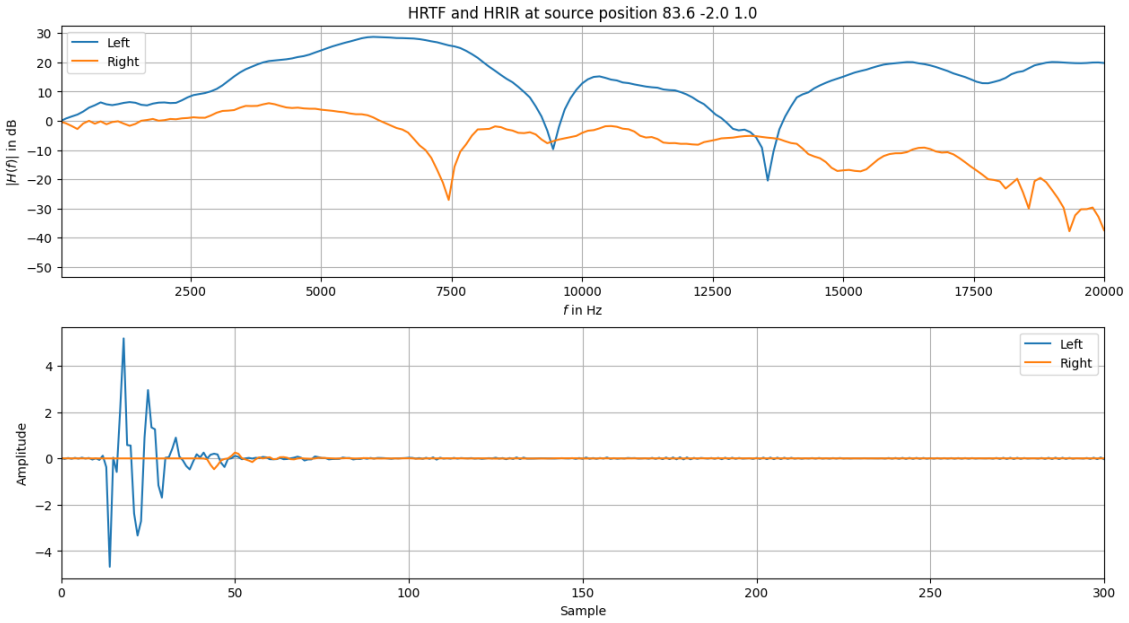

Figure 2.5: HRTF and HRIR for a given source position $\phi = 83.6°$, $\theta = -2°$ and r = 1

As seen in Figure 2.5 above, there is not only a big overall level difference between the two functions for each ear but also the spectral information differs quite a lot with frequency dips in the left ear that does not appear as prominently in the right ear.

These differences are the things that give humans information about the location of a source.

2.3.2 Binaural Room Impulse Response

To decode SDM into a static IR containing the spatial information captured in encoding a, BRIR is computed.

HRTFs usually come as a set of transfer functions, each with their corresponding spherical or cartesian coordinates. Utilizing the spatial data captured in a system, one can connect a direction of arrival to a specific HRTF at that point in time. Doing this over an IR results in a sum of weighted HRIRs with appropriate delay. Equation 2.14 shows this computation.

$$s^{l,r}(t) = x(t) * h^{l,r}_{HRIR}(\Omega, t) \tag{2.14}$$

where x is the impulse response, $h^{l,r}$ are the the closest HRIRs for a given DOA Ω and t is the time in samples and $*$ is the symbol for convolution.

For the evaluation of this thesis, BRIRs were rendered for different head rotations to be used in real time with SoundScape Renderer (SSR) [8]. For this an angle offset, representing the negative head rotation angle were added to the DOAs before rendering the BRIR.

2.3.3 Virtual Loudspeaker Array

Another way to decode SDM is to use loudspeakers mounted in a spherical array around the listener. By using the estimated DOAs of a given measurement and its IR the individual gains can be calculated for each loudspeaker of a given loudspeaker array setup. The most straight forward way of doing this is whats called nearest loudspeaker synthesis (nls). Computing the smallest euclidean distance between the DOA and a loudspeaker in the array at each time sample quantizes the DOAs to the loudspeaker grid. This can then be weighted by the instantaneous pressure of the IR to give a set of loudspeaker IRs. Applying these IRs to a source signal with real time convolution auralizes the SDM encoding for a listener positioned in the middle of the loudspeaker array.

These have been proven to work very well but have the downside to be expensive and require an anechoic environment which is rare to have the opportunity to work in. While a physical loudspeaker array might be out of reach virtual loudspeaker arrays can be utilized in a much more accessible manner. By applying HRTFs after rendering the loudspeaker IRs, these virtual loudspeakers can be auralized with

headphones.

This have sparked development of software that render spatial audio to virtual speaker arrays such as the SPARTA binauralizer [9], which routes rendered loudspeaker signals to a set of virtual loudspeakers with set coordinates in 3d space and applies HRTFs in real time, as well as integrate head tracking via OSC (open sound control) in a very straight forward manner. Figure 2.6 shows a screen shot of the SPARTA Binauralizer configured for a 7.4 surround setup.

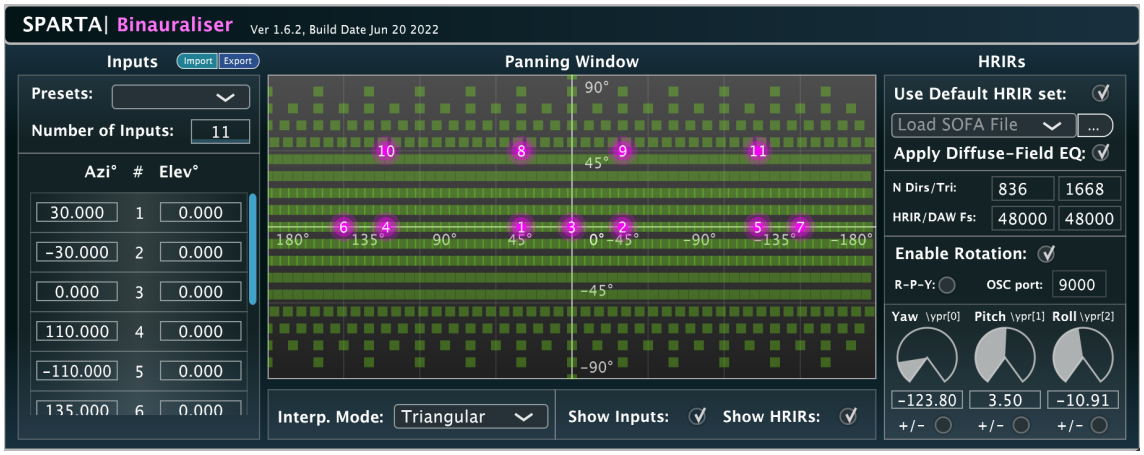

Figure 2.6: SPARTA Binauralizer user interface set up for 7.4 atmos surround. The pink elements show the coordinates of the loudspeakers configured and their channel number. The green elements show the coordinates for the set of HRTFs loaded. [9]

With the SPARTA Binauralizer, head rotation can be controlled with OSC data from a third party head tracker like the Supperware Tracker 1.

3

Implementation

This chapter gives an overview of the development methods and implementation of SDM as well as the listening tests setup.

3.1 Hardware

For the measurement hardware, the whole design is based around two audio interfaces which drive separate microphones. One high quality omnidirectional microphone that capture the IR used for the rendering and one set of cheap microphones used for the spatial data capture.

3.1.1 Microphones and interfacing

The spatial data was captured using MEMS microphones by Infineon Technologies [10]. Infineon supplies a development kit with these capsules mounted on flex boards with solder pads for V_{DD}, ground, PDM Clock and Data. Since these microphones are digital, they do not send an analog voltage as an output but instead send a Pulse Density Modulation (PDM) signal that is decoded by the audio interface to be read into some audio software or other processing unit. PDM is a digital signal protocol that can also be described as an over sampled 1-bit audio stream [11].

Figure 3.2 shows the flex board with the MEMS capsule to the right and the flex flat connector to the left. Marked in the middle is the solder pads used to connect the cable to. As the MEMS can send two microphone signals over one data pin, the select pad (labeled S), is connected to V_{DD} or ground, depending on left or right configuration. The data from these pairs are then read depending on the clock signals rising or falling edge at the audio interface. The flat connector was used with an Infineon Au-

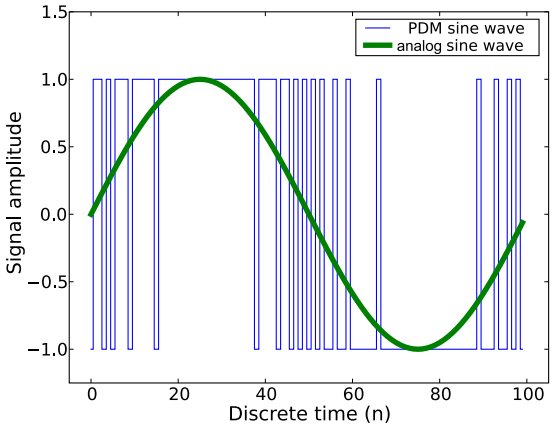

Figure 3.1: Analog sine wave (green)and its resulting digital PDM sine wave [12]

diohub Nano [13] for testing the microphones functionality as well as left and right communication.

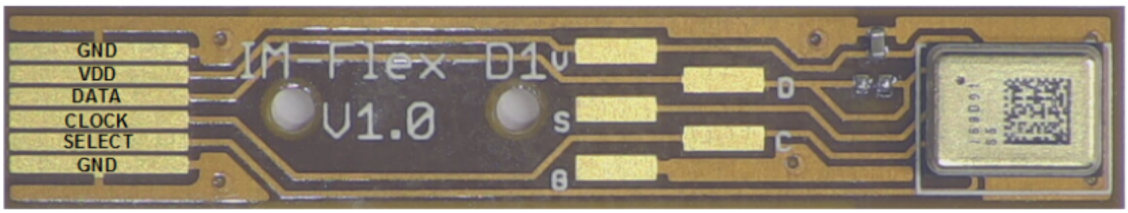

Figure 3.2: Infineon MEMS Flex development board

The signals from the MEMS microphones are sent to a miniDSP MCHStreamer audio interface which can be loaded with a number of different firmware options depending on use case. As stated before these microphones supply a 2 channel PDM signal which can be decoded when using the PDM firmware for the MCHStreamer. Figure 3.3 shows the signal flow of the MCHStreamer when it has the PDM16 firmware installed. For this setup, only the J3 pin header was used for the connections to the microphones

For a simple and patchable setup, two sets of four conductor, shielded cables were acquired. One with a male 3.5 mm connector and one with a female 3.5 mm connector. The cable with the male connector was stripped at the bare end and soldered onto these pads.

These were intended to separate the microphone from the audio interface by building the audio interface, together with a prototype board, into a box. Figure 3.4 shows these cables built into a plastic box with the connectors mounted in the side of the box for easy patching to the microphones.

When testing this setup, some issues arose. Testing the microphones when connected

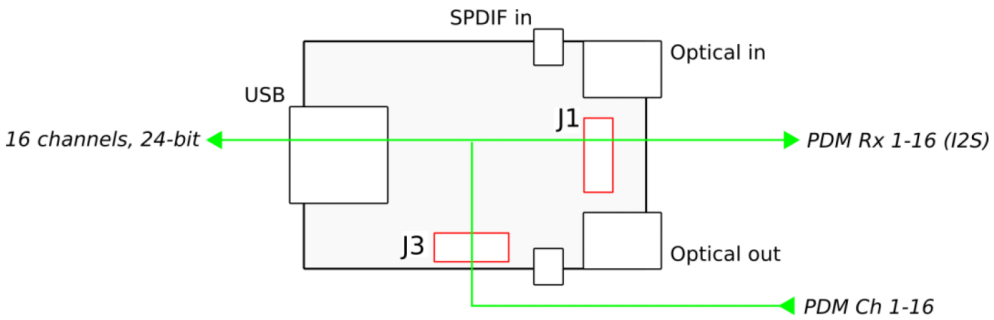

Figure 3.3: PDM Firmware signal flow for MCHStreamer [14]

Figure 3.4: Infineon MEMS Flex development board

showed that one of the two microphone signals for each pin was intact, while the other produced either noise or no signal at all. This caused a problem with the idea of using 12 microphones for SDM encoding. As stated before, each of the 8 PDM pins on the MCHStreamer can take up to 2 microphone signals and these signals are read based on the state of the clock signal i.e. channel 1 is read on rising edge and channel 2 is read on falling edge, there needs to be some clear difference. This was revealed to not be the case when longer cable lengths were connecting the microphones to the MCHStreamer. Shielded cables are constructed in a a way where a number of conductors are each isolated. In the case of these cables these isolated wires carry V_{DD}, clock and data. These are surrounded with twisted strands of wire connected to ground which in turn is enclosed by a thin conductive foil. While these sorts of cables have proven quite effective against electro magnetic interference (EMI), this design can result in a low pass characteristic. The conductive foil and the twisted strands of wire connected to ground act as a capacitance which, with the would make the whole cable act as a first-order RC low-pass filter. Equation 3.1 shows how the ideal transfer function would look for a generic shielded cable.

$$H(j\omega) = \frac{1}{1 + j\omega RC} \tag{3.1}$$

where ω is the angular frequency, R is the resistance in the wire and C is the capacitance between the shielding and ground. Since both the capacitance and and resistance is in this case dependent on the length of the cable, connecting more cables increases the total capacitance of the system and therefore lowers the cut off frequency of the filter.

This might not affect most applications but since these PDM signals operate at a clock frequency of 3.072 MHz not much capacitance needs to be applied to filter the

clock signal. While these values are not known in the cables used, this is most likely the source of the failure occurring when connecting two MEMS microphones on the same PDM pin.

These issues were not able to be resolved within the time frame of this project so a scaled down version was made to get a working SDM rig. To ensure signal integrity, the cables were cut and soldered directly onto a prototype board. A header cable was soldered onto the prototype board which connected to a pin header on the miniDSP MCHStreamer. Figure 3.5 shows the cables soldered onto a prototype board directly with the intention of connecting the MCHStreamer directly to it. Although crude, this solution is simple and quite robust assuming no short circuits are introduced when soldering the lanes.

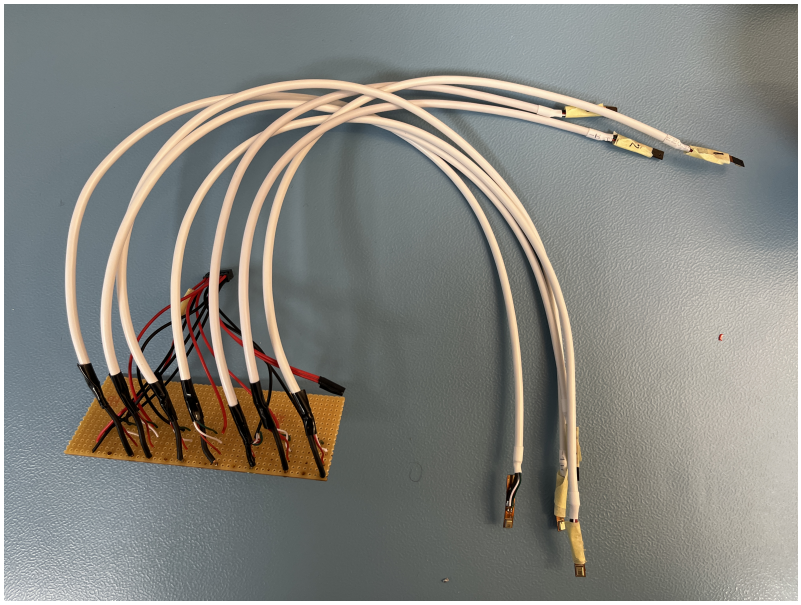

Figure 3.5: MEMS microphones and cables soldered onto the prototype board

Figure 3.6 shows the connectors and how they are connected to the prototype board.

The center microphone was an EarthWorks M23R which easily connected via XLR to another audio interface and just require +48 V phantom power which is readily available on most audio interfaces. For the purposes of this project the two interfaces were combined by making an aggregate device on OSX. While this feature is not native to Windows PCs, third party applications such as ASIO4ALL [15] allow users to combine multiple interfaces into one which can be interfaced with as one singular device.

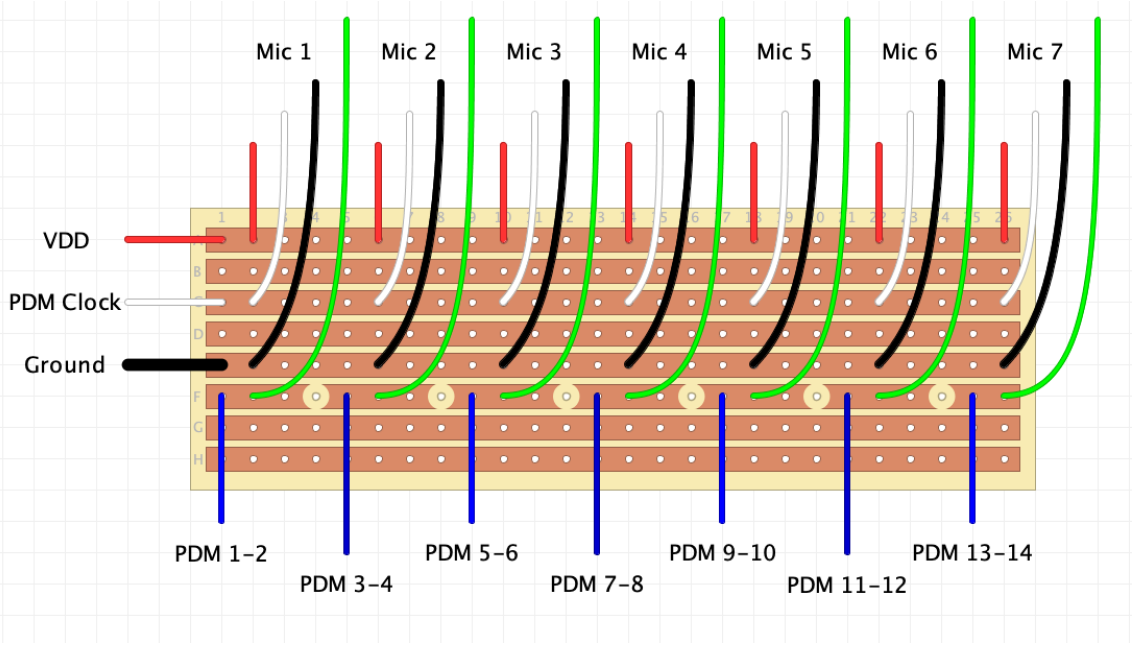

Figure 3.6: Wire connections on the prototype board

3.2 Array Design and dimensions

The design of the array was adapted from Meta Research BinauralSDM project [16] which employs 6 omnidirectional condenser microphones for capture of spatial data and a omnidirectional condenser microphone at the center of the spherical array . A similar mount was designed in Autodesk Fusion 360 but with modifications for the MEMS capsules. Figure 3.7 displays the 3D-model of the microphone array which was designed to be mounted directly onto the EarthWorks M23R microphone. Figure 3.8 shows the array mounted onto the m23r microphone with the MEMS capsules attached to the array.

As stated in chapter 2, the only limiting factor of the dimensions of the array is the maximum distance between two of the satellite microphones, see Equation 2.12. This together with simulation data displayed in [17], validated the diameter of the array to be 5 cm. The paper investigates the DOA error as a function of both diameter and window size and states that window sizes between 36 and 64 samples performed the best for an array with a diameter of 10 cm. Using Equation 2.12 gives us the following:

$$2\frac{2_{max}}{c}fs = 2\frac{0.05\text{m}}{343\text{m}\,\text{s}^{-1}}48\text{kHz} \approx 14 \leq 36 \tag{3.2}$$

This gives a lot of headroom to experiment with different window sizes for the processing.

Figure 3.7: AutoDesk Fusion 360 view of the final array design

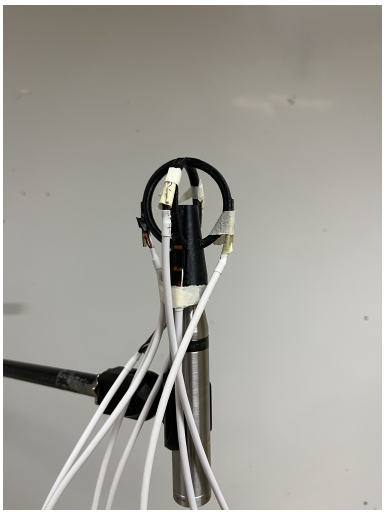

Figure 3.8: Microphone array with 12 microphones, coordinates relative geometrical center of array

3.3 Software

During the development of this, a Python script was built continually to accommodate all the steps of IR capture, SDM encoding and decoding.

Python today is very well equipped for scientific applications with packages such as numpy, scipy and matplotlib which add functionality that mimics a software suite like MatLab. In this project a number of different packages were used to capture the IRs and do all the encoding and processing. While SDM encoding and decoding has been implemented in MatLab with the SDM Toolbox [6] and in Python with spaudipy [18], no Python implementation exists with IR capture, TDOA estimation and BRIR rendering

For the purpose of testing and verification a script was written in Python where encoding and decoding options could be configured for different purposes. This script also makes it easy to perform measurements and rendering loudspeaker signals or BRIRs even with little to no programming knowledge. Figure 3.9 shows a flow chart of this script.

3.3.1 Encoding

As stated in Section 2.2.1.2 TDOA estimation has been implemented in MatLab with SDM Toolbox by Sakari Tervo [6]. In the function `SDMpar.m` which for this thesis was ported to Python to estimate DOAs. During porting of `SDMpar.m` from MatLab to Python, the same set of IRs were used to ensure that always compare

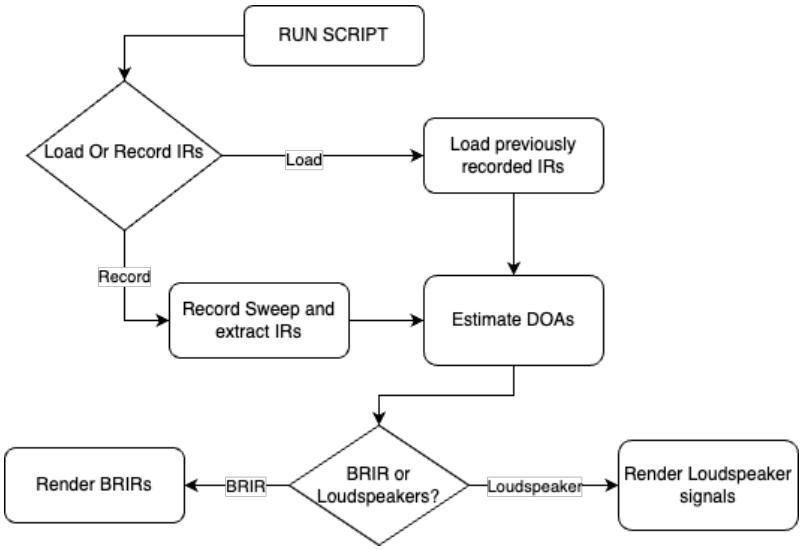

Figure 3.9: Flow chart of main script implementing IR capture or loading, DOA estimation as well as rendering BRIRs our loudspeaker signals

and make sure no miscalculations arose. The DOAs estimated in both Python and Matlab were auralized to ensure that no difference were introduced during encoding.

Based on the processing methods introduced in binauralSDM [16], some extra processing such as DOA alignment and DOA smoothing were implemented to evaluate performance based

For the listening experiment parameters such as window size and DOA smoothing were varied for each stimuli to see how they affect the performance for each measurement situation.

3.3.2 Auralization

The process of generating a BRIR from a set of DOAs and an IR was described in Section 2.3. For the auralization part of the project these methods were implemented in Python. For the validity of the listening tests the HRTF set of the KEMAR mannequin was imported in SOFA format using the `Python-sofa`-library [19]. This was a crucial step due to the fact that comparing a binaural rendering with one set of HRTFs with a measurement using a different dummy will yield some differences. These might be slight since the difference between HRTF sets are very small but if the HRTFs and the dummy are different it is unclear if the difference in perception is due to the different HRTFs or parameter selection for the SRIR encoding or BRIR rendering.

Utilizing the spaudiopy package [18], loudspeaker rendering was also implemented. This to have a straightforward way to record IRs, estimate DOAs and use tools from the open source software such as the SPARTA plugin suite to binauralize SDM with head tracking.

3.4 Simulation

Due to the limitations discovered of the MCHStreamer and the MEMS microphones (described in Section 3.1.1) a simulation was performed to investigate the impact that adding 6 more microphones for the SDM microphones.

The simulation was done with MATLAB and was adapted from an example in the Audio Toolbox [20]. The image source method is a geometric simulation method which treats sound as specular reflections between a source and a receiver. The image source method was implemented in a shoebox room of dimension 5x5x3.5 with source at position 2.5 4 1.8 and receiver centered around 2.5 1 1.8. Figure 3.10 displays the space, the source and the receiver.

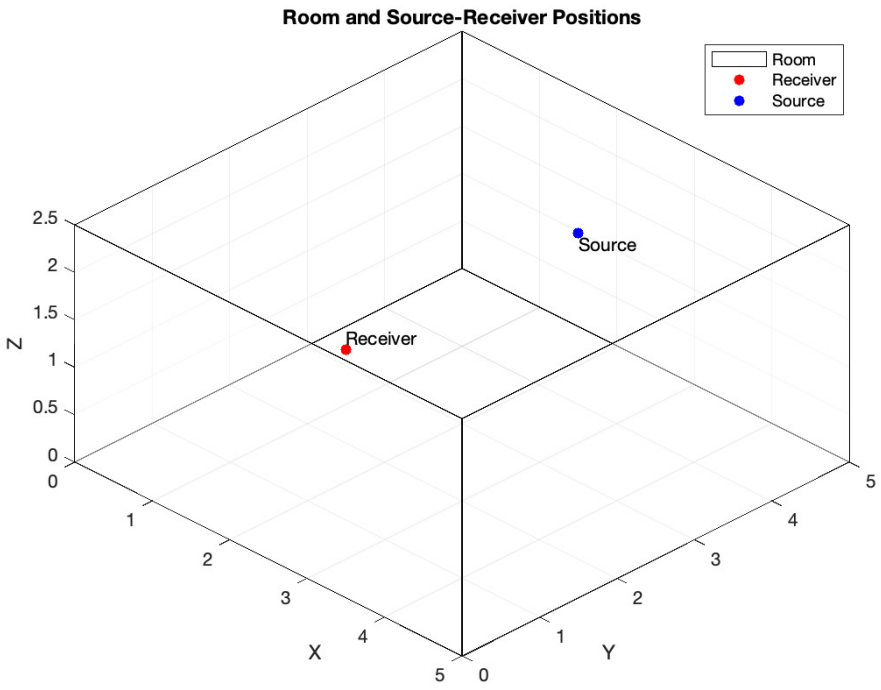

Figure 3.10: Virtual room used for image source method

To achieve the same timing offsets between the different IRs as a real microphone array would produce, a coordinate offset relative the receiver position was applied. With these 7 or 13 receiver positions, ISM was used to compute the IRs for each.

The sets of IRs were then used with the SDM toolbox [6] in MatLab to estimate the directions of arrival and auralize them. In the adapted script, the image source is done for each one of the microphones in each setup. All IRs were then loaded into a modified version of the script `demoBinauralRendering.m` provided with `SDM Toolbox` [6]. The DOAs were analyzed and A-B comparisons were made to evaluate the sonic differences between the two arrays.

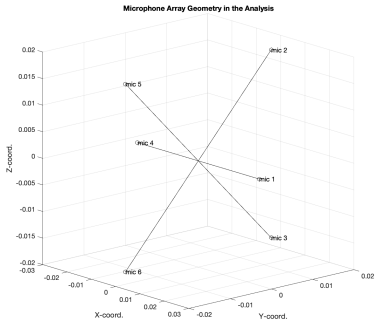

Figure 3.11: Microphone array with 6 microphones, coordinates relative geometrical center of array

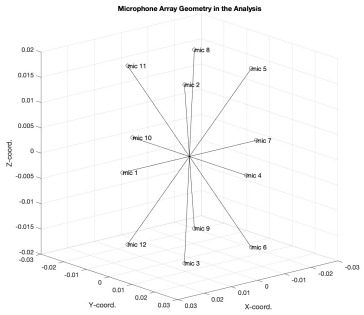

Figure 3.12: Microphone array with 12 microphones, coordinates relative geometrical center of array

3.5 Measurement Setup

For the measurement and evaluation, the microphone array was set up in a corner of the vibro acoustic lab at the division of applied acoustics at Chalmers University of Technology. This to evaluate how the early reflections from a close wall would impact the localization of the direct sound when a lot of early reflections appear from one side of the array. A Genelec 8020D loudspeaker was mounted on a stand at the same height as the array and a Genelec 7050B subwoofer was connected in series with the loudspeaker to fill out the low end where the 8020D lacks power.

The reference measurement used a G.R.A.S 45BB KEMAR dummy head and torso [21] in the same position as the SDM array. A VariSphear turntable was used together with the KEMAR to get a measurement over 360°. For each degree of rotation a set of IRs were saved. Figure 3.13 shows the measurement setup with the KEMAR dummy head. Figure 3.14 displays the measurement layout from a top-down perspective. When measuring with the SDM microphone, the dummy head was removed from the turntable and the SDM array was placed in the center of where the head of the KEMAR would be.

To evaluate how source occlusion affects SDM performance, a 4 cm thick piece of wood was placed in front of the loudspeaker at a distance of 15 cm from the tweeter. This to see how the direct sound would be interpreted when no direct sound is

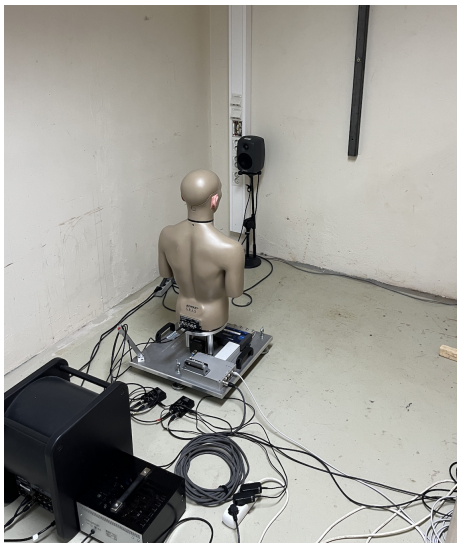

Figure 3.13: Measurement Setup

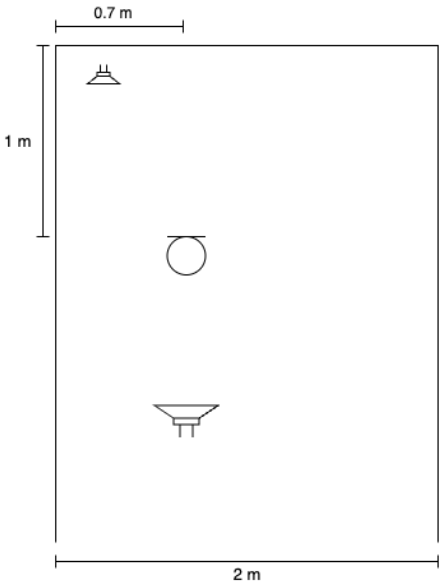

Figure 3.14: Top-down layout of measurement setup

present due to diffraction around the occluding material.

3.6 Listening Experiment

To evaluate the performance of the microphone array and encoding developed in this thesis a MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) listening test was set up with head tracking. MUSHRA is a common method used to conduct listening experiments [22]. In a MUSHRA experiment, a participant evaluates sonic characteristics compared to a known reference. MUSHRA also includes a hidden reference and an anchor among the stimuli. The listening tests were implemented using SoundScape Renderer (SSR). SSR is an open source tool for spatial audio rendering that can accommodate a number of different rendering modules depending on use case [8]. For this implementation binaural room synthesis (BSR) was used to load a series of BRIRs each representing an a separate angle of rotation. These are convolved in real time in SSR with a source signal based on a given head orientation. Head orientation is received from a Polhemus Patriot Head tracker [23] and processed via pure data (pd) to SSR. The Sennheiser HD 650 headphones were used and a measured Headphone Compensation Filter (HpCF) was applied to the output of SSR to ensure a flat frequency response in the headphones. Figure 3.15 shows the signal flow between SSR, the MUSHRA GUI and the various pd patches.

The experiment consisted two parts where the first one was training for the participant to get familiar with the interface before the real study started. The second part was the user study which consisted of 4 trials where all combinations of mea-

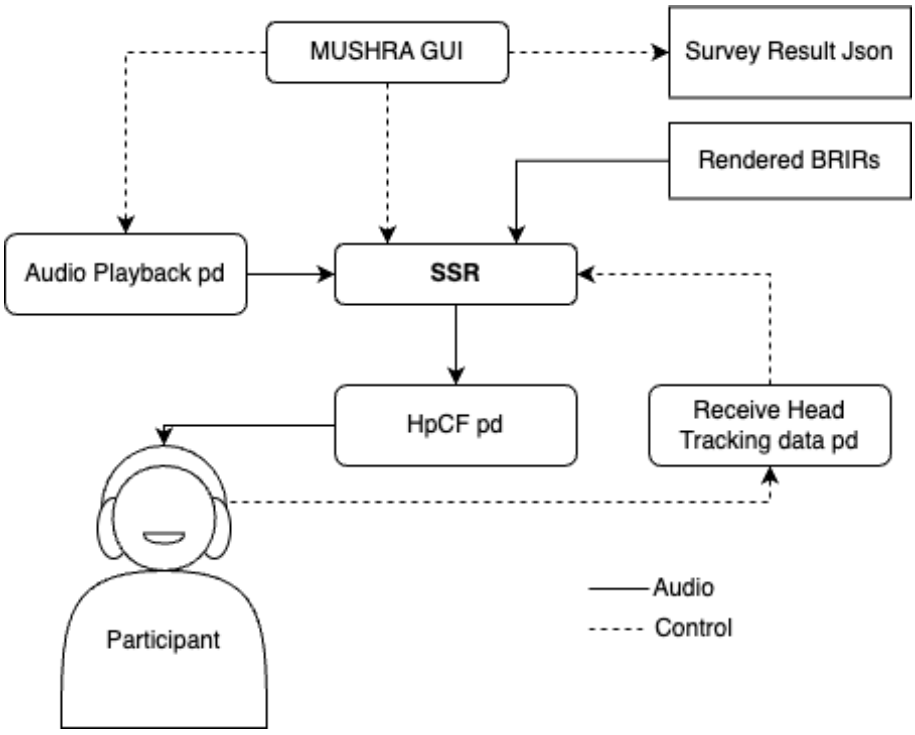

Figure 3.15: Flow chart of Audio and Control data for listening experiment

surements and source signals were included. Table 3.1 shows the trials in order.

Table 3.1: Trials of the listening tests

Trial	Measurement	Source Signal
1	Occluded	Speech
2	Clear	Speech
3	Occluded	Drums
4	Clear	Drums

Each trials stimuli were comprised of a reference, SDM of two different window sizes and DOA smoothing on and off as described in Table 3.2. The Anchor was implemented as pressure + HRTF as described by Ahrens [24]. This was rendered just like the BRIRs but the DOAs were locked to the approximate coordinate of the loudspeaker relative to the microphone and its HRTF was used to render the whole BRIR. This will, according to Ahrens, sound significantly different to the reference and make the participants gain confidence in their performance.

Each measurement was rendered with 4 different conditions as presented in Table 3.2. In cases where the SDM resulted in an angle offset of the direct sound compared to the reference the DOAs were aligned to match as close as possible.

Table 3.2: Stimuli Evaluated in MUSHRA listening test.

Stimuli	Method	Window Length	Smoothing Order	Label
1	SDM	32	6	SDM_32_sma
2	SDM	64	6	SDM_64_sma
3	SDM	32	1	SDM_32
4	SDM	64	1	SDM_64
5	p + HRTF	n/a	n/a	Anchor
Ref	DH	n/a	n/a	Reference

3.6.1 Survey

The MUSHRA user interface was used by the participant to select which rendering to listen to and rank their difference relative the reference rendering. Each participant got a short introduction to the project and the experiment. The experiment contained two parts, training and experiment. During the training part the participants familiarized themselves with the interface and the different sources, stimuli and measurement setups. Figure 3.16 shows the MUSHRA GUI set up for the training part of the experiment. No results were acquired from the training part but results were only obtained from the second part. The participants were encouraged to use the entire scale of the sliders.

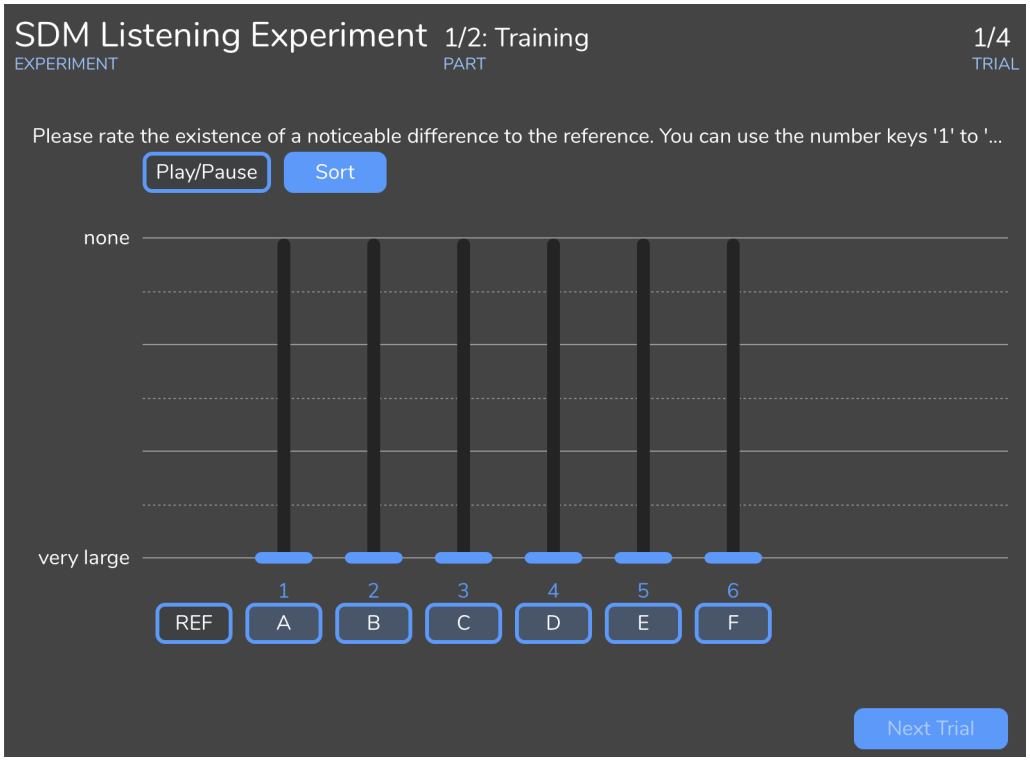

Figure 3.16: Screenshot of the MUSHRA User Interface used in the listening experiment

The participants were instructed to evaluate the difference based on coloration,

reverberation and externalization. Due to errors inherent in the estimation the direct sound the DOAs were aligned so that the direct sound aligned with that of the reference and anchor.

4

Results

In this chapter, the results of the listening tests as well as the simulations will be presented together with some short explanations.

4.1 Simulation

As explained in chapter 3, issues with hardware resulted in simulations comparing SDM performance between a 6 microphone array and a 12 microphone array. The auralized signals were saved and compared in an A-B comparison.

Figure 4.1 and 4.2 displays directions of arrival and pressure for SDM simulated using IRs generated by the image source method. These are plotted over azimuth and elevation and the size of each element is the normalized pressure of the center microphone.

4.2 Direction Of Arrival Estimation

Figure 4.3 through 4.10 presents the DOAs of the 4 rendered stimuli shown in Table 3.2 for both the clear and occluded measurement setup. The y-axis shows the elevation and x-axis shows azimuth. DOAs were aligned as described in 3.6.

4.3 Listening Tests

The following figures displays box plots of participants rating of perceived audio quality and externalization relative to a reference dummy head measurement. 10 participants performed the listening experiment consisting of 3 women and 7 men.

Table 4.1 shows the average rating of each stimuli and each measurement situation.

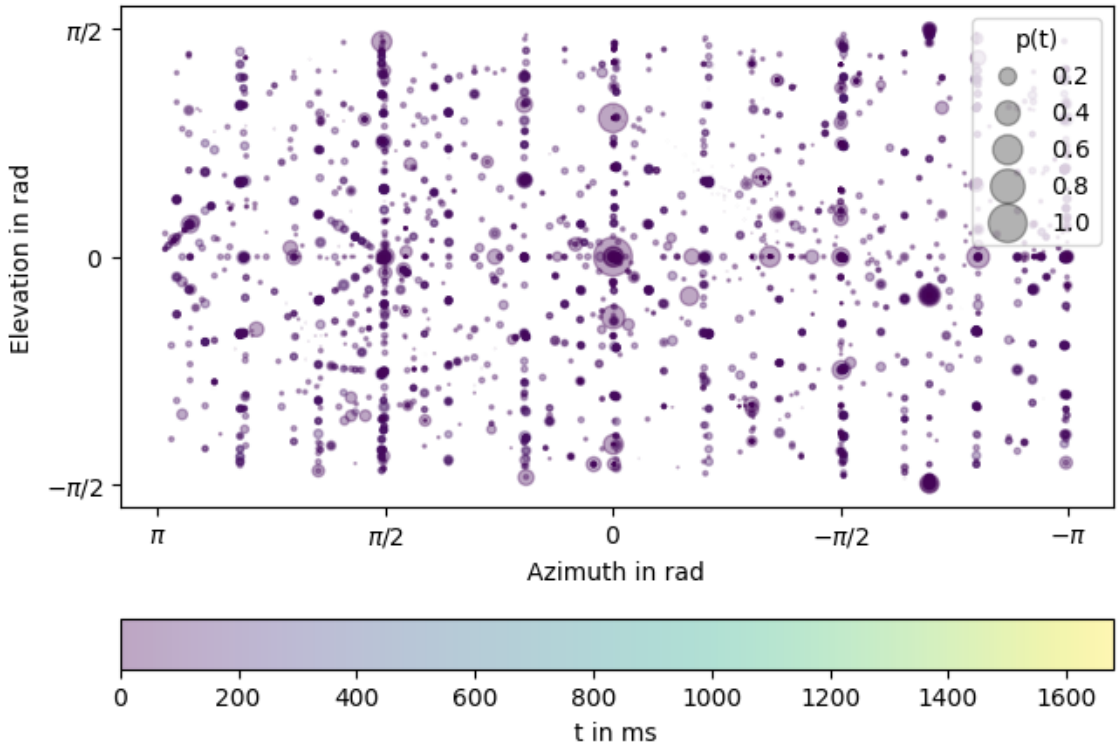

Figure 4.1: Directions of arrival for a simulated 6 microphone array

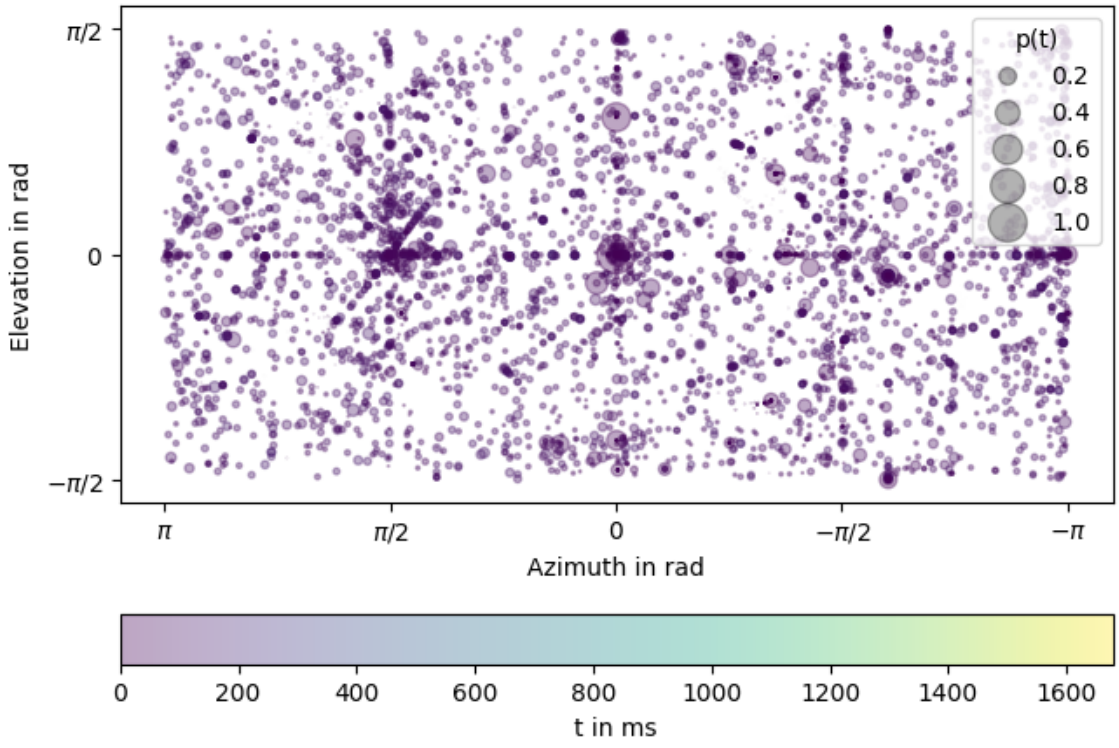

Figure 4.2: Directions of arrival for a simulated 12 microphone array

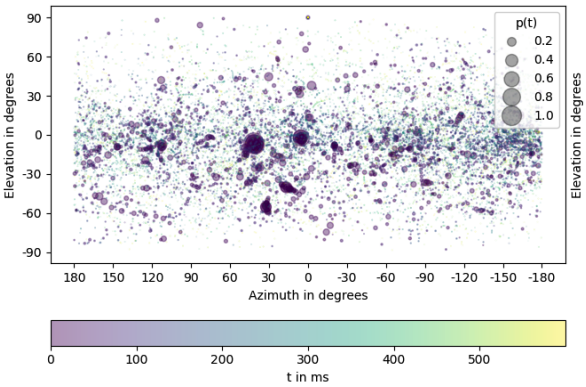

Figure 4.3: DOAs for SDM_32_sma_clear

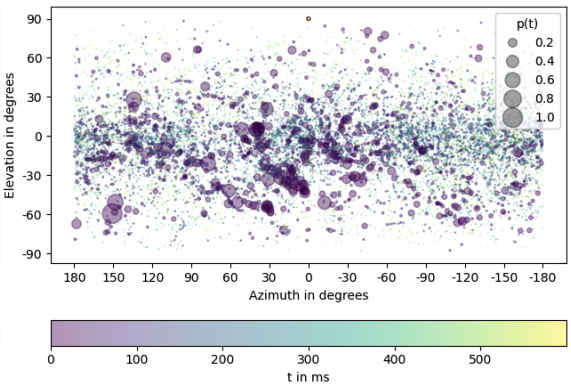

Figure 4.4: DOAs for SDM_32_sma_occluded

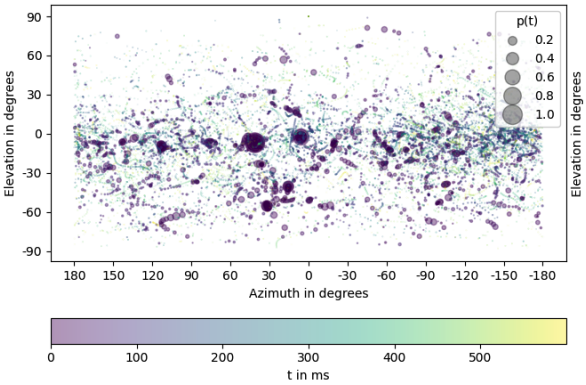

Figure 4.5: DOAs for SDM_64_sma_clear

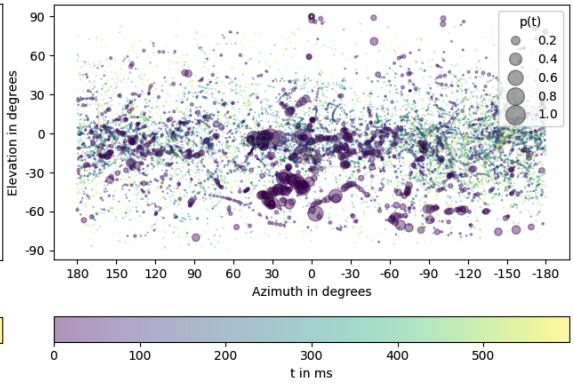

Figure 4.6: DOAs for SDM_64_sma_occluded

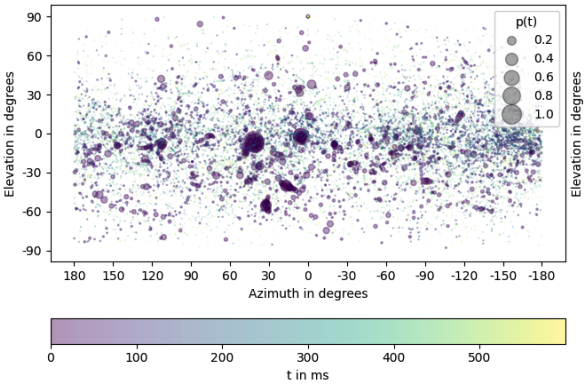

Figure 4.7: DOAs for SDM_32_clear

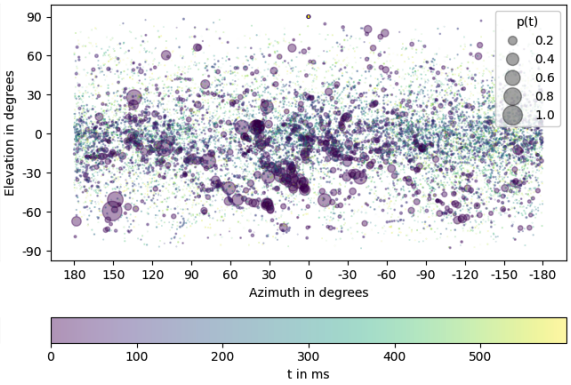

Figure 4.8: DOAs for SDM_32_occluded

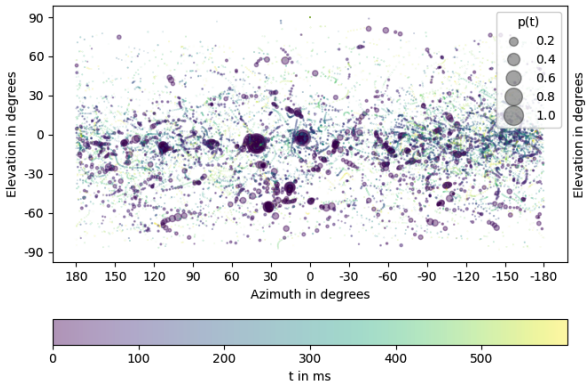

Figure 4.9: DOAs for SDM_64_clear

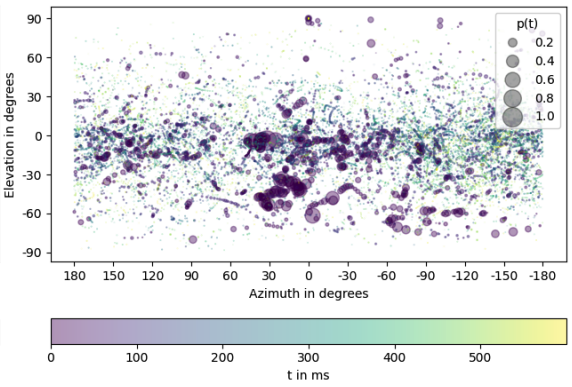

Figure 4.10: DOAs for SDM_64_occ

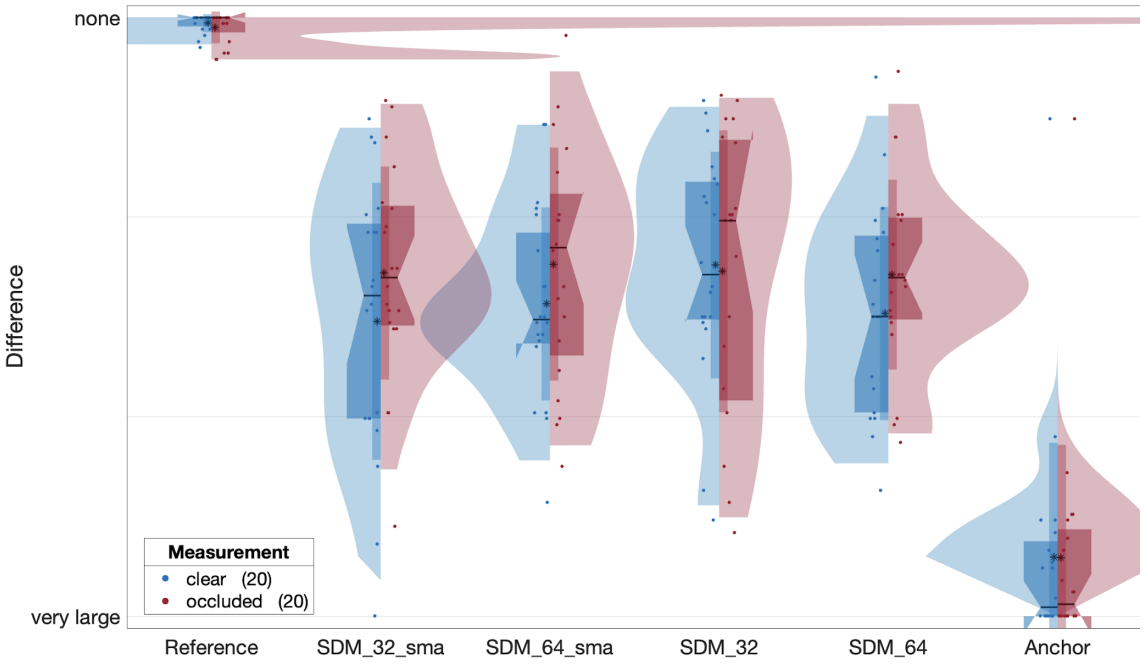

Figure 4.11: Combined box plot and density plot compared by measurement situation

Table 4.1: Average rating for each stimuli by each measurement and by all data.

Stimuli	Label	Average Rating Clear	Average Rating Occluded	Average by all
1	SDM_32_sma	-50.8	-42.75	-46.775
2	SDM_64_sma	-47.9	-41.25	-44.575
3	SDM_32	-41.4	-42.45	-41.925
4	SDM_64	-49.5	-43	-46.25

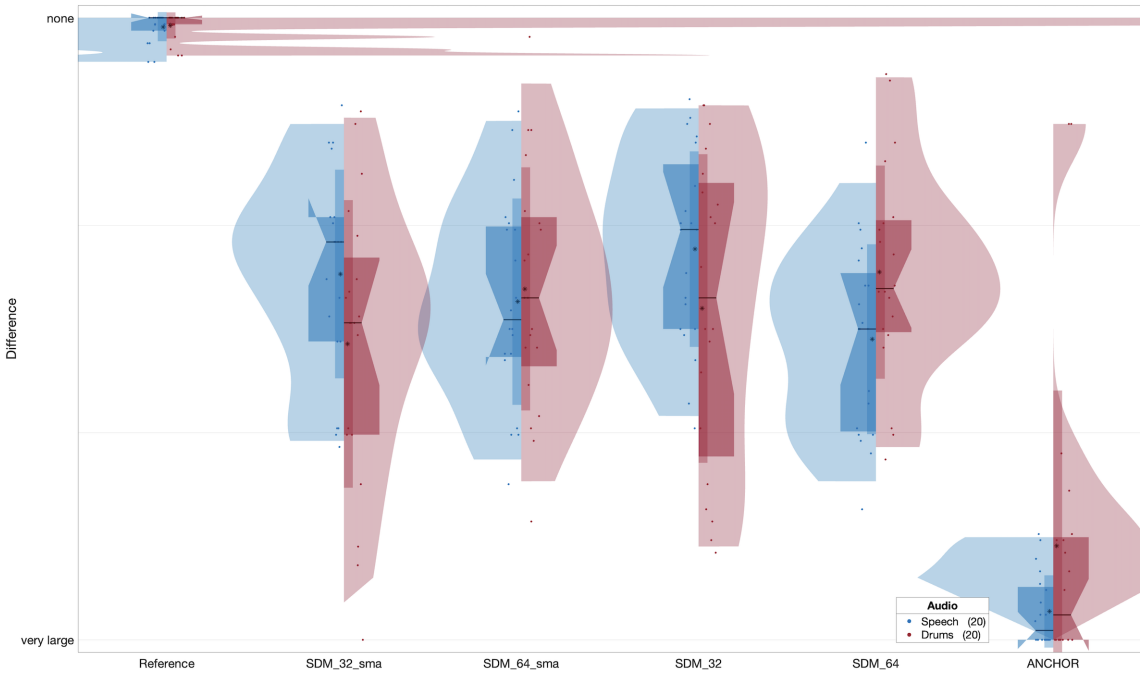

Figure 4.12: Combined box plot and density plot compared by source signal

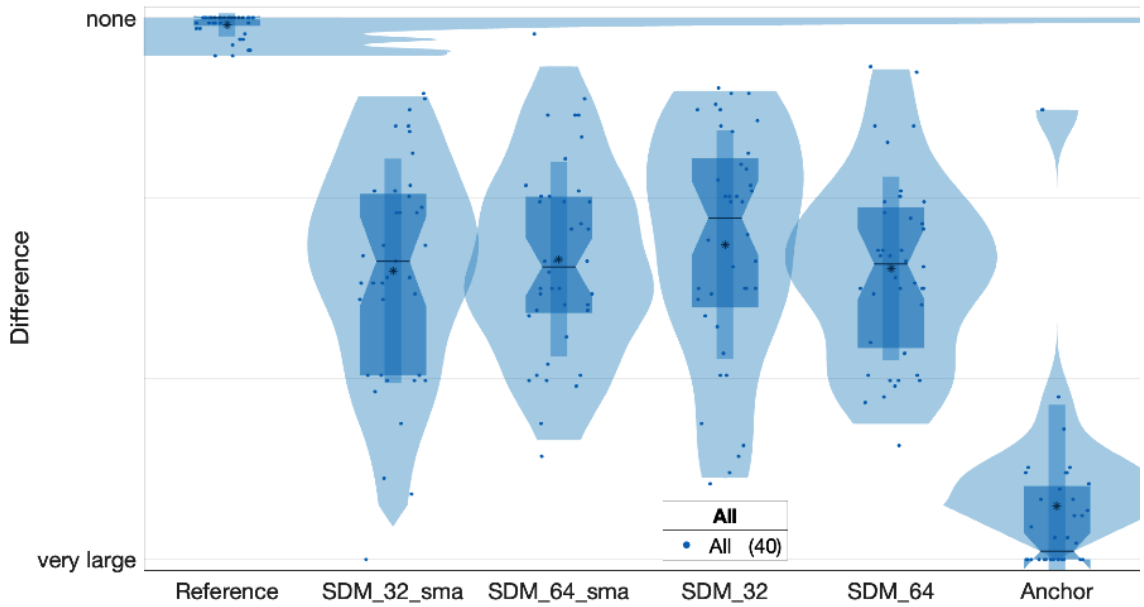

Figure 4.13: Combined box plot and density plot for each stimuli over all measurement situations and source signals

5

Discussion

Evaluating the different array types in the simulation did not seem to show any big difference. While figure 4.2 shows a lot more detail and precision in its DOAs this does not necessarily constitute better results when auralizing. According to Amengual Gari et. al. [16], lower spatial resolution does not necessarily impair the listeners perception of the rendering. A spatial resolution as low as 14 DOAs has been proven to render the same plausibility as a rendering with up to 50 DOAs. When rendering the two array types a simple A-B comparison was done to see if there was any significant difference in sound quality. Although there were sonic differences in terms of coloration and reverberation these were minuscule and no version sounded better than the other. DOA estimation of different types of arrays will always sound different since the geometry of the microphone positions will result in different sets of DOAs but this does not necessarily impact the sound quality for better or worse.

With these 4 SDM configurations were selected based on previous research it might be why no configuration showed a clear performance difference to the rest. In this range of smoothing and window sizes, the TDoA estimation might not be as sensitive as in other ranges. For this microphone array and these measurement conditions where a lot of early reflections could make the encoding non-ideal, the window sizes were kept in a range which has been proven effective in most applications of SDM. Of course, without the limitations of the MCHStreamers sampling rate in this PDM16 firmware being maximum 48 kHz, these longer window lengths could have been used. If smoothing orders had been tested in a wider range, a clearer difference might have been detected of the different stimuli, but with this set of stimuli no difference was discerned between smoothing on and off.

Looking at the results presented in figure 4.11 and 4.12 shows clearly that the MUSHRA test was successful. While no clear difference can be discerned when either comparing the results by measurement or source signal the MUSHRA worked as intended where the majority identified the reference and the anchor correctly. For this MUSHRA test the two measurement situations were evaluated separately on their own pages. This might be a cause to the similar results in rating between the

two different measurement setups. To further investigate a listening test could be set up where no reference is involved and just compare the two situations, clear and occluded, directly for different settings. By not conducting a MUSHRA test, participants could rank the perceived sound quality without a reference and comparing the two measurement situations and their respective encoding settings.

6

Conclusions and Future Work

From this project a few conclusions can be drawn.

Using MEMS microphones are a good way to get more microphones into rigs like this where their audio quality is not critical. While cheap, these microphones are quite robust and well suited for this application. Improving the hardware with a custom PCB and improve the microphone mounts on the microphone array will result in a rig that will be more reliable in terms of durability. With regards to the cable choice, for shorter runs of cables generic flat cables should do the trick. These would be more flexible and easier to work with as well as not apply any low pass filtering. Although the filtering characteristics of the cables did not affect the signals in this setup, the PDM signals might become less reliable when having longer cables that would be easier to handle.

Regarding the number of microphones, simulations show that more microphones does results in higher spatial resolution but does not improve perceived sound quality. SDM already has essentially infinite spatial resolution which is not necessary for better auralization performance.

SDM does not seem to be sensitive to these small parameter differences. Looking the ratings for the clear and occluded measurements individually shows that both measurements perform similarly when comparing their different SDM configurations. Expanding the ranges of window sizes and smoothing orders could have yielded bigger difference in performance between the stimuli but this can not be concluded with this listening experiment. DOA smoothing was proposed by Amengual et. al. [17] but no indication of its effectiveness has been presented.

Occlusion does not impact the SDM performance significantly compared to a clear line of sight between source and microphone. While these listening experiments could have been set up in a more comparative manner for these two conditions these results does indicate that the differences are not significant.

There are obviously some improvements to be made in the hardware setup. Due to the weak clock signals and the low-pass filtering occurring in the shielded cables some modifications could be done to make this more robust. While the implementation of MEMS microphones still yielded good quality estimations, the hardware choices were not ideal. Most applications using digital MEMS microphones usually have them mounted onto a printed circuit board and physically close to decoding interfaces which of course negates the effects of the cable on the signals. With the MCHStreamer being such a compact and simple device designing a case for it that can mount on a microphone stand close to the array would be ideal. And with the standard pin header connector a printed circuit board could be designed to be connected directly to the streamer. Running short cables from this PCB to the microphones to supply the needed signals as well as receive data could be a perfect solution that is robust and reliable.

While the hardware used in this project lowers the cost of an SDM measurement setup significantly, optional hardware can be used to cut costs even more. By using, for example a miniDSP UMIK-1 and modifying the array mount for it would cut make this whole rig a lot cheaper since it connects directly to the computer via USB. Connecting the loudspeaker via the analog out of the computer would result in a similar setup and similar performance.

Since research in SDM is very much active at institutions all over, new advancements in SDM performance are made constantly. Many post processing methods and improvements are being developed to the estimation to make SDM more robust and can deliver better sound quality. With the spectral whitening generated when rendering BRIRs with SDM being a big factor in impairing the sound quality of SDM, applying post equalization to flatten the frequency response should yield better results. With the implementations done with `binauralSDM` [25] for Matlab, post equalization for SDM could be implemented in a python project like this one. Amengual et. al. have proposed more methods to enhance the rendering quality which all should be considered in future work on this method and its implementations.

Bibliography

[1] Sakari Tervo et al. “Spatial Decomposition Method for Room Impulse Responses”. In: *Journal of the Audio Engineering Society* 61 (Jan. 2013), pp. 16–27.

[2] Neofytos Kaplanis. “Perception of Reverberation in Domestic and Automotive Environments - PhD Thesis”. PhD thesis. Dec. 2016. DOI: 10.13140/RG.2.2.35180.59528.

[3] Mendel Kleiner. *Acoustics and audio technology.* J. Ross Pub., 2012. ISBN: 9781604270525. URL: https://search.ebscohost.com/login.aspx?direct=true&db=cat09075a&AN=clpc.oai.edge.chalmers.folio.ebsco.com.fs00001000.04cbafc9.d815.49fa.a9f6.60fda2627371&site=eds-live&scope=site&authtype=guest&custid=s3911979&groupid=main&profile=eds.

[4] Nils Meyer-Kahlen, Sebastià V. Amengual, and Tapio Lokki. “What the Spatial Decomposition Method can and cannot do”. English. In: (2022). | openaire: EC/HE/812719/EU//VRACE; International Congress on Acoustics, ICA2022 ; Conference date: 24-10-2022 Through 28-10-2022. URL: https://ica2022korea.org.

[5] 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 (1976), pp. 320–327. DOI: 10.1109/TASSP.1976.1162830.

[6] Sakari Tervo. *SDM Toolbox*. Natick, Massachusetts, United States, 2018. URL: https://www.mathworks.com/matlabcentral/fileexchange/56663-sdm-toolbox.

[7] *Auralization*. en-GB. URL: https://odeon.dk/learn/articles/auralization/ (visited on 05/22/2023).

[8] *SoundScape Renderer User Manual — SoundScape Renderer 0.6.0 documentation*. en. ssr. URL: https://ssr.readthedocs.io/en/0.6.0/index.html (visited on 05/11/2023).

[9] *SPARTA suite*. en-us. Aug. 2021. URL: https://leomccormack.github.io/sparta-site/docs/plugins/sparta-suite/ (visited on 05/24/2023).

[10] *High performance digital XENSIVTM MEMS microphone*. IM69D130. Rev. 1.0. Infineon Technologies AG. Dec. 2017.

[11] Thomas Kite. “Understanding PDM Digital Audio”. en. In: (2012).

[12] This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.

org/licenses/by/4.0/. URL: https://upload.wikimedia.org/wikipedia/commons/e/e7/Pulse_density_modulation.svg.

[13] *Infineon Audiohub Nano*. IM69D130. Infineon Technologies AG. Oct. 2019.

[14] *MCHStreamer Manual*. LTC3600. Rev. D. miniDSP. 2022. URL: https://www.minidsp.com/images/documents/Product%%20Brief-MCHStreamer.pdf.

[15] *ASIO4ALL Official Home – Universal Windows ASIO Driver*. URL: https://asio4all.org/ (visited on 06/02/2023).

[16] Sebastia V. Amengual Gari et al. "Optimizations of the Spatial Decomposition Method for Binaural Reproduction". In: *Journal of the Audio Engineering Society* 68.12 (2020), pp. 959 –976. DOI: https://doi.org/10.17743/jaes.2020.0063.

[17] sebastià v. amengual garí et al. "optimizations of the spatial decomposition method for binaural reproduction". In: *journal of the audio engineering society* 68.12 (2021), pp. 959–976. DOI: https://doi.org/10.17743/jaes.2020.0063.

[18] Chris Hold. *spaudiopy*. original-date: 2019-02-04T16:45:51Z. June 2023. URL: https://github.com/chris-hld/spaudiopy (visited on 06/09/2023).

[19] *Spatially Oriented Format for Acoustics (SOFA) API for Python*. URL: https://python-sofa.readthedocs.io/en/latest/index.html.

[20] The MathWorks Inc. *Room Impulse Response Simulation with the Image-Source Method and HRTF Interpolation*. Natick, Massachusetts, United States, 2022. URL: https://se.mathworks.com/help/audio/ug/room-impulse-response-simulation-with-image-source-method-and-hrtf-interpolation.html.

[21] *GRAS 45BB KEMAR Head & Torso, Non-configured*. da-dk. URL: https://www.grasacoustics.com/products/head-torso-simulators-kemar/product/733-45bb (visited on 05/15/2023).

[22] S. R. M. Prasanna and Martin P. Hollier. "MUSHRA: A new method for the evaluation of sound quality". In: *The Journal of the Acoustical Society of America* 105.4 (1999), pp. 2465–2465. DOI: 10.1121/1.426366.

[23] *Polhemus Patriot*. URL: https://polhemus.com/motion-tracking/all-trackers/patriot/ (visited on 06/22/2023).

[24] Jens Ahrens. "Auralization of Omnidirectional Room Impulse Responses Based on the Spatial Decomposition Method and Synthetic Spatial Data". en. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, May 2019, pp. 146–150. ISBN: 978-1-4799-8131-1. DOI: 10.1109/ICASSP.2019.8683661. URL: https://ieeexplore.ieee.org/document/8683661/ (visited on 05/25/2023).

[25] *BinauralSDM*. original-date: 2020-09-11T17:49:16Z. June 2023. URL: https://github.com/facebookresearch/BinauralSDM (visited on 06/19/2023).

Department of Civil Engineering and Architechture
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY