Acoustic Multi-Microphone Beamforming in a Practical Form Factor

Roope Kiiski

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 23.1.2017

Thesis supervisor:

Prof. Tapio Lokki

Thesis advisors:

M.Sc. Kalle Mäkinen

M.Sc. Kai Saksela

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Roope Kiiski

Title: Acoustic Multi-Microphone Beamforming in a Practical Form Factor

Date: 23.1.2017 Language: English Number of pages: 8+82

Department of Signal Processing and Acoustics

Professorship: Acoustics and Audio Signal Processing Technology

Supervisor: Prof. Tapio Lokki

Advisors: M.Sc. Kalle Mäkinen, M.Sc. Kai Saksela

Signal quality and noiselessness is important for Automatic Speech Recognition and communications. These can be achieved with the help of microphone arrays and beamforming. Multiple microphones can be used to pick sound coming from one direction while attenuating other directions. In this thesis, three arrays and three beamformers are compared. The target is to see how different structures affect the output signal and if one of the arrays is superior to the others.

The results show that the baffled array appears bigger than the reference array or the array on the top surface of a cylinder, but it adds complexity to the beamformers. For Automatic Speech Recognition, all three arrays performed the same, but there were clear differences between the beamforming algorithms. Thus the array to use depends on the use-case and especially on the industrial design of the device.

Keywords: Microphones, Beamforming, Microphone arrays

AALTO-YLIOPISTO
SÄHKÖTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Roope Kiiski

Työn nimi: Akustinen Monimikrofonikeilanmuodostus Käytännöllisessä Laitteessa

Päivämäärä: 23.1.2017 Kieli: Englanti Sivumäärä: 8+82

Signaalinkäsittelyn ja akustiikan laitos

Professuuri: Akustiikka ja äänenkäsittelytekniikka

Työn valvoja: Prof. Tapio Lokki

Työn ohjaajat: DI. Kalle Mäkinen, DI. Kai Saksela

Äänenlaatu ja kohinattomuus ovat tärkeitä automaattisen puheentunnistuksen ja puheviestinnän toimivuuden kannalta. Mikrofoniryhmiä ja keilanmuodostusta voidaan käyttää parantamaan signaalin laatua ja vähentämään kohinaa nauhoittamalla vain yhdestä suunnasta tulevaa ääntä. Tässä työssä vertaillaan kolmea mikrofoniryhmää ja kolmea keilanmuodostusalgoritmia. Tarkoituksena on selvittää miten erilaiset rakenteet, joihin mikrofoniryhmä on kiinnitettynä vaikuttaa nauhoitettuun signaaliin, ja jos jokin ryhmistä on merkittävästi parempi kuin muut.

Tulokset osoittavat, että sylinterin ulkoreunalle kiinnitetty mikrofoniryhmä vaikuttaa akustisesti suuremmalta kuin referenssi mikrofoniryhmä tai sylinterin päälipinnalle kiinnitetty mikrofoniryhmä, mutta se lisää keilanmuodostusalgoritmin monimutkaisuutta. Automaattisessa puheentunnistuksessa kaikki kolme mikrofoniryhmää suoruituivat yhtä hyvin, mutta keilanmuodostusalgoritmien suorituskyvyssä oli selkeät erot. Käytännössä siis mikrofoniryhmä kannattaa valita käyttötarkoitukseen ja etenkin lopullisen tuotteen teollisen muotoiluun sopivaksi.

Avainsanat: Mikrofonit, Keilanmuodostus, Mikrofoniryhmät

Preface

I want to thank Professor Tapio Lokki and advisor Kai Saksela for their guidance and help for this thesis. I'd like to thank my co-workers at Intel Finland, especially my supervisor Tapio Liusvaara who made this work possible. Also the other interns at Intel Finland were a great help, as we shared tips for the writing, enabled venting out frustrations and provided support overall to each others.

In addition, the help of audio people at Intel Finland, especially David Isherwood's, Mikko Kursula's and Seppo Ingalsuo's, was invaluable as their suggestions, comments and know-how helped solve many problems and confusions I had during this process. Lastly, but in no way least, a huge shout-out to Kalle Mäkinen, who as an advisor for this thesis at Intel has helped me in countless ways, such as coming up with the idea for this thesis, helping me through the whole process, overall teaching me invaluable skills and helping me grow as a professional.

Otaniemi, 23.1.2017

Roope Kiiski

Contents

Symbols and Abbreviations

Symbols

$\boldsymbol{a}$	attenuation in time-domain
α	step size of the adaptation algorithm
$\boldsymbol{d}$	delays and attenuations in frequency domain
$\mathbf{\Gamma}$	coherence matrix
E	error function
h	transfer function in time-domain
H	transfer function in frequency-domain
ψ	weighting function for TDoA estimation
Φ	PSD matrix
ϕ	azimuth angle
σ	standard deviation
σ^2	variance
θ	elevation angle
τ	delay in time-domain
$\boldsymbol{w}$	time-domain filter coefficients
$\boldsymbol{W}$	frequency-domain filter coefficients

Abbreviations

AEC	Acoustic Echo Cancellation
AG	Array-Gain
ASIC	Application Specific Integrated Circuit
ASR	Automatic Speech Recognition
CDB	Constant Directivity Beamformer
DSB	Delay-and-Sum Beamformer
DI	Directivity Index
DoA	Direction-of-Arrival
DRR	Direct-to-Reverberant Ratio
DUT	Device Under Test
FBR	Front-to-Back Ratio
GCC	Generalized Cross-Correlation
GJBF	Griffiths-Jim BeamFormer
HATS	Head And Torso Simulator
HRSE	High-Resolution Spectral Estimation techniques
LMS	Least Mean Squares
LUFS	Loudness Unit relative to Full Scale
MEMS	MicroElectroMechanical System
ML	Maximum Likelihood
MSL	Maximum Sidelobe Level
MVDR	Minimum Variance Distortionless Response
PCB	Printed Circuit Board
PHAT	PHAse Transform
PSD	Power Spectral Density
SDB	SuperDirective Beamformer
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
SRP	Steered Response Power technique
TDoA	Time Difference of Arrival
THD	Total Harmonic Distortion
WER	Word Error Rate
WNG	White Noise Gain

1 Introduction

Human auditory system is very complex and it is a product of thousands of years of evolution. Thus it is very adept at differentiating multiple signals, locating signal sources and focusing on single sound events, amongst other awesome things. Even though microphones have improved significantly since their invention, they still perform worse than human ears when it comes to the previously mentioned abilities that human ears have. This is not surprising, considering that microphones are very young invention compared to the human auditory system. Yet there is huge interest in improving the performance of microphones, and to enable them to perform similarly as to the auditory system, especially in source separation, localization and robustness to noise. The increase in the interest can be attributed to the popularity of different kinds of mobile devices, the improvements in the computation capabilities of those devices and new, more natural user interfaces, which include speech-to-machine communication. Overall, different Automatic Speech Recognition applications have become more and more common, with personal assistants such as Apple's Siri and Amazon's Alexa being the prime examples.

Often this is sought to achieve by using an array of microphones and algorithms called beamformers to combine the information from the multiple microphones. Microphone arrays can vary a lot, as they can be different size and shape, have different amounts of microphones, and the acoustic integration can differ. Arrays can be shaped anywhere from a simple line, to random 3D shapes, the sizes can vary from multiple meters to as small as few centimetres and they can have as few as 2 microphones - 1 microphone can't be considered an array after all - and there is no upper limit for the quantity of microphones. Most practical microphone arrays are either linear, circular or spherical arrays, with somewhere between 4 and 100 microphones, and their sizes are often noticeably less than 1 meter. Beamforming algorithms also vary a lot, from simplest Delay-and-Sum beamformers to the more complex Adaptive beamformers. What they all have in common is that they combine the information of the multiple microphones and try to improve the performance compared to a single microphone. These improvements can be noise suppression, dereverberating the input signal, or blocking unwanted sound events, among other things.

1.1 Objectives of this thesis

In this thesis, the target is to figure out what effect the structure the microphones are attached to has for the performance of the beamformer. A single form-factor was chosen, namely a cylinder, to which two arrays are attached: one to the top surface of the cylinder and one around the outer edge of the cylinder. These arrays are then compared to a reference array to figure out how the performance differs between different kinds of arrays. The arrays are measured in both free-field and in diffuse-field, and different beamforming algorithms are used to see if the arrays perform differently with different beamformers. The cylinder was chosen for the form-factor, as it is quite a natural structure for a stand-alone device that would be controlled by voice. This is because such a device can't be hidden in a closet or under a table lest the performance suffers. Because the device needs to be visible and in somewhat central place, it needs to actually look decent or people don't want to have it in their living rooms or elsewhere. An example of devices with microphone beamformers that are designed for consumers are Amazon Echo and Google Home, which both are more or less cylindrical. An added benefit of cylinder as a form factor for this thesis is that it can easily be used to test the previously mentioned two arrays with only a single structure.

The main research question in this thesis is if any of the arrays tested has superior performance compared to the others. The performance is measured with beamwidth, sidelobe-levels, spatial responses and word-error-rate. If there are no huge differences in the performance, what are then the pros and cons of each of the arrays, what kind of phenomena needs to be taken into an account with different arrays. The research is limited to the three chosen arrays and three different beamforming algorithms. In addition, the beamforming is performed for recordings, so it isn't done in real-time and thus the computation requirements for the beamforming aren't considered. Also, array's effect on source localization and separation aren't looked into.

2 Background

This section presents some use cases for microphone beamforming and various challenges for the use cases. Beamforming, in general, is explained and then the basics of microphones - focusing on MEMS microphones - are discussed.

2.1 Speech Recognition Use Case

Generally speaking, communication means exchanging information through a channel between an emitter and a receiver. For communication to work, the emitter and the receiver need to be able to understand one another and the channel used must be able to relay the wanted information. [1] Historically in the case of speech, the channel has been made mostly of air and the emitter and receiver have usually been humans. In the modern world, the channel can also include electricity and media other than air. Also the emitter and receiver are not necessarily humans, for example they can be computers.

Speech communication between humans via devices, or even human-to-machine communication, has become increasingly common and capturing the speech communication source has its own set of challenges. Due to the acoustic environment where the sounds are produced and recorded, a simple microphone is not always enough for good quality communications. This is especially true for human-to-machine communication, as machines are not as adept at recognizing bad quality speech as humans are - after all, mankind has had thousands of years to evolve and adapt to the challenging physical environments we live in. Common problems encountered in real environments are for example echoes, interferences, noises and reverberation. [2, 1]

An example of how well the human auditory system is adapting to the environment is the so-called cocktail party effect. The cocktail party effect is used to describe how a listener can focus on one talker among many other conversations and ignore the other talkers. The effect gets its name from the fact that it can easily be noticed in cocktail parties, where there are plenty of people having different conversations. Basically, human hearing can filter out the unwanted background noise and enhance the wanted signal. If we would replace a human with a microphone in such a scenario, the difference between the two cases would be clearly noticeable. [1]

A single microphone has decent Signal-to-Noise Ratio (SNR) when the source is close enough to it, as then the wanted signal is usually stronger than the noise and reverberation. If the microphone, for some reason is not very close to the speaker, then the situation gets challenging, as the amount of reverberation and noise when compared to the level of the wanted signals increases. [1] Yet with certain use-cases, the microphone may need to be quite far from the speaker, thus decreasing the quality of the recorded signal. These use-cases can include, for example a device which passively listens to the user and with certain keywords performs an action, such as a device that would control the lighting or appliances of a house and other objects by voice commands. With such a device, the user can move around freely while the device stays stationary, and thus at times the user can be several meters away from the device.

For a machine to understand human speech, speech recognition is used. Simply put, speech recognition is the means of finding the most likely word or sequence of words based on the acoustic signal and associated computational models. A speech recognizer usually consists of three parts: a language model, which defines words and if they are likely to occur together, a lexicon, which defines how words are formed from sound units, and an acoustic model, which defines the basic sound units of the language.

Speech recognizer is based on the idea that each phoneme of a certain language, which are the basic speech sounds for the language, has distinct characteristics that can be recognized by analysing the signal. When the phonemes of the signal are recognized, they can be mapped to certain words, for example the phonemes /hə'ləʊ/ in English form the word 'hello'. Usually, in addition to recognizing the words, a speech recognizer also tries to form correct message from the words. In other words, instead of only recognizing the word, it checks what other words it has recognized and if they are likely to occur together linguistically or if the sentence they form makes any grammatical sense.

2.2 Challenges of Speech Recognition

As stated in the section 2.1, human-to-machine communication and various other use cases which require speech recognition are becoming increasingly common. Speech recognition has its own set of challenges, ranging from the changing environments and recording devices to the fact that different languages have different characteristics and that every speaker is unique. For a speech recognition to be robust the input signal needs to be of good quality, otherwise the characteristics of speech might be drowned in noise and other unwanted parts of the signal, making the speech unrecognisable. [2, 1, 3]

The worst-case scenario for speech recognition is most likely the cocktail party-like scenario, where there are multiple speakers and conversations in one space. In addition to the primary speaker possibly being far from the microphone, which decreases quality, a single microphone has a hard time differentiating between different speakers and conversations, making it very hard to recognise the right conversation. [1]

An increasingly common answer to these challenges is microphone array signal processing, and especially its subset called microphone beamforming. The idea behind array processing is that with multiple sensors we can find the Direction-of-Arrival (DoA) of different signals and, in a similar manner to the human auditory system, attenuate unwanted signals and focus on the wanted signal. Array processing also helps with reducing the noise of the signal, and it can be used to dereverberate the signal. [2, 1, 4, 5]

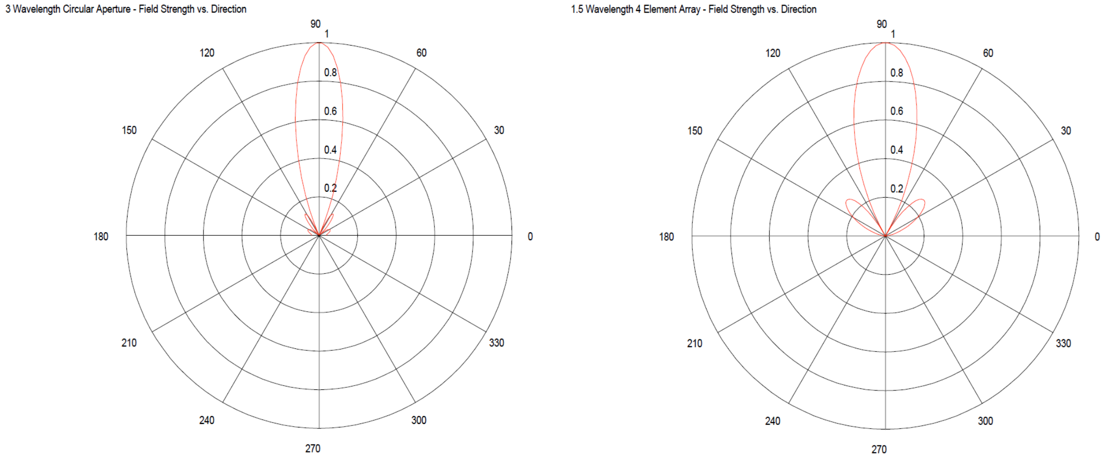
Overall, when considering a device that would utilize automatic speech recognition (ASR), such as a personal assistant, a smart-room or a computer controlled by speech, there are numerous challenges such a system would need to take into an account. For those systems to be commercially viable, they need to recognize speech accurately, and as such devices could be used in many different ways, places and scenarios, there are numerous challenges to overcome. First of all, the user doesn't remain constant. The speaker might change from time to time, so the ASR shouldn't be trained for a single user only, and as different speakers have different accents, vocabularies and other characteristics of speech. The speaker can move around the space in which the ASR is in, meaning that the beamformer needs to somehow move with the speaker. Secondly, with commercial products the environment where the system is in can vary a lot. Some people might use the system in their living room, some in their kitchen and some in somewhere totally different. These differences can be taken into an account with some kind of calibration process, but it needs not to be too cumbersome so that the product would remain easy to use. Thirdly, the scenario where the ASR is used can change a lot. For example, living room can be quiet, with only a single person in it with no other distractions, or it can be filled with people, loud music and overall partying. So the noise levels, types of noise and other distractions can vary a lot. [3, 6]

2.3 Beamforming in General

Beamforming can be generalised as a method of combining signals from many omnidirectional sensors to simulate a large directional sensor [7]. The sensors can be pretty much anything, but most commonly they are antennas, loudspeakers or microphones. In radio applications, the beamforming is used to reduce interference and improve communication quality by pointing the antenna at the signal source and this can be done without physically moving the antenna [7]. Beamforming is used in multiple fields, including radar, sonar, communications, imaging, geophysics, astronomy and biomedicine. Sonar applications include source localization and classification, for communications beamforming is used to get directional transmission and reception. In geophysics, beamforming is used for earth crust mapping and oil exploration and astronomy uses it to get high resolution images of the space. Especially important area is biomedicine, where beamforming is used for example in hearing aids, to improve the quality of the devices. [8]

In radio applications the sensors used are antennas, which are devices that convert electrical signals into electromagnetic waves and vice-versa. Antennas have individual radiation patterns that depend on the physical characteristics of the antenna, and it's usually plotted as field strength versus direction and it is the same for both transmitting and receiving. The radiation pattern of an antenna derives from the fact that each small part of the antenna generates waves with different phases, amplitudes and distances to the measurement location. When the antenna's field is measured far away from the antenna, the field's strength is the sum of the waves generated by each of these small parts. Depending on the direction, some of these different waves are added constructively or destructively, strengthening the signal at certain directions and weakening it at others. [7, 9]

By changing the shape and size of an antenna the radiation pattern can be adjusted to give the wanted directivity. The other way to control the radiation pattern is to use multiple small omnidirectional antennas and feed them all the same



(a) Circular aperture antenna. [7]

(b) Four element linear array. [7]

Figure 1: Example radiation patterns.

signal. In this case the same effect happens, for each antenna radiates the same field but their distances to a point far away from them is different, thus causing destructive and constructive interferences depending on the direction. Examples of directivity patterns for a circular aperture antenna and a four element linear array with size of 1.5 wavelengths are shown in Figure 1. [7]

In the radiation patterns of figure 1 clear examples of so called beams can be noticed. In beamforming, the idea is to steer this beam towards the wanted point in space, thus receiving the signal from that location with maximum level. In addition, unwanted signals from other locations are attenuated, making the output signal have less noise. [7]

When considering the antennas of the figure 1, the circular aperture antenna's beam can only be steered by physically moving the antenna, whereas the array's beam can be steered electronically. Electronical steering in the case of radio beamforming is done by phase-shifting the signals from the antennas before summing them. [7, 9] Steering of the beamformer enables the estimation of DoA of the signal. DoA can be estimated by different algorithms, but one of the simplest methods is to steer the beamformer around the space and find the direction with the most energy. [1]

In the case of microphone beamforming, the sensor array consists of microphones that together measure the sound field at different locations. The information of the sound field's behaviour at different points in space can be used to - among other things -

- implement source localisation and tracking,
- boost the wanted signal,
- attenuate unwanted signals,
- separate different sound sources,
- record signals with spatial cues.

Various parameters have a big impact on how good the result for these usages is. Some of these impacting parameters are the physical size and geometry of the array, the quantity and quality of the microphones, the operation environment and the processing algorithm itself. [4] The effect of these parameters is discussed in more detail in section 3.

Beginning in the 1930s, microphone arrays and beamforming were used for example in stereo recordings, but since then they have gained more applications. The stereo recording arrays usually consisted of two to four microphones, and there were multiple ways to place them. In the end, even though the principle is quite similar to the beamforming and arrays discussed elsewhere in this thesis, the motivation and use case is different. [10]

Broadly speaking, beamformers can be divided into two categories: fixed and adaptive beamformers. Fixed beamformers are common, as they include delay-and-sum, filter-and-sum and the super directive beamformers. Fixed beamformers are such that their processing parameters stay constant and do not change with time. In comparison, adaptive beamformers change their parameters along time. The reason for this is that the noise in real environments doesn't remain constant, instead changing spatially and spectrally, and thus the noise characteristic estimation should adapt to the changes. Adaptive beamformers include, for example, *Frost Algorithm* and *Generalized Sidelobe Canceller*. [1]

In addition, different kinds of beamformers are used for narrowband and broadband signals. A narrowband system can be defined as a system in which the centre frequency is many times higher than the bandwidth of the system. [1, 3] This definition is true for most radio and radar systems, as their signal is usually very high frequency, and the bandwidth is really limited. For example a $2.4\,GHz$ WiFi channel 1 has a centre frequency of $2412\,MHz$ and bandwidth of $20\,MHz$, which by the previous definition is a narrowband, for the bandwidth is less than 1/100th of the centre frequency. For most of the applications of beamforming the *narrowband assumption*, ie. that the beamformer is acquiring a narrowband signal, is true.

Broadband systems on the other hand have bandwidth that is closer to the centre frequency. [1, 3] Examples of wideband systems are sonar and sound systems. For example, consider a microphone beamformer designed to capture music. Let's assume the system is only interested in frequencies between $40\,Hz$ and $20\,kHz$. That makes the bandwidth $19960\,Hz$ and the centre frequency $10020\,Hz$. In this system the bandwidth is roughly twice as big as the centre frequency. Thus, such a system would be a broadband system.

In practice, broadband beamformers can be implemented with multiple narrowband beamformers, so that each narrowband beamformer analyses small part of the signal on its own and the results are combined. This can be done by filtering the input signal into narrow bands which are then fed to each beamformer. This method is considered to be frequency-domain beamforming. Another way to analyse broadband signals with narrowband beamformers is to model the array output as a multidimensional time series. This can be done, for example, with an autoregressive moving average model, and the system's poles can be estimated with for example Yule-Walker equations. By evaluating the estimated spectral density matrix at the system poles, a narrowband techniques can be used. This method is considered to be time-domain beamforming. [3, 11]

A common type of broadband beamformers is called *constant directivity beamformers* (CDB), which are designed to have a constant spatial response over a wide frequency range. Most CDB techniques are based on the idea that different arrays should be used for different frequencies, for the size of the array and the distance between the microphones should change depending on the frequencies the array is used for. A common way to realize this idea is to use nested sub-arrays, so the array consists of multiple equally-spaced arrays, each of which are designed for specific frequency range. The output of each sub-array can be combined by appropriate bandpass filtering. [3]

In addition to the narrowband assumption which, in the case of microphone beamforming is never true, another assumption is commonly made: the *far-field assumption.* The farfield assumption means that the source is located far enough away that the wave arriving to the array can be considered plane waves. This is true in most of the beamforming applications, such as radar, sonar and telecommunications for the distances usually are significantly longer than the wavelengths of the signals. Again, in the microphone beamforming, where the source can be close to the array and the wavelengths can be quite long, this is not always the case. Thus, for microphone beamforming, the far-field assumption can be true or false depending on the scenario. [1, 3, 8, 12]

2.4 Source Localization

As hinted previously, an important part of beamforming is source localization. It is very common that beamformers do some kind of source localization, for example periodically to initialize correct beam direction or continuously to follow moving source. For example radios might only use source localization when they are turned on if it is known that the source doesn't move while it is on. On the other hand, for example in microphone beamforming, the source usually moves, and thus for the beamformer needs to continuously adjust its beam so that the source doesn't move out of it. [3]

The primary measure of source localization's performance is accuracy, and it depends on various factors. The most important of these factors include

- the quality and quantity of the microphones used,
- the geometry of the microphone array,
- the location of the source relative to the microphones,
- amount of ambient noise and reverberation,
- other noise sources, such as other speakers. [3]

In practice in microphone beamforming slight errors in source localization are not too crucial, as the beamwidth of the beamformers in practical arrays can easily be over ten degrees, and thus if the source is located a couple of degrees wrong, it still remains in the beam. Still, if the error is too large, and the beam is pointed towards such a direction that the real source lies in a null of the beampattern, then the wanted signal is attenuated a lot and thus the beamformer is actually degrading the signal.

In addition to finding the wanted signal source, source localization schemes can be used to find locations of unwanted signals. With adaptive beamformers, it is possible to point nulls of beampattern towards certain directions, and thus it is useful to be able to find locations of other sources. When the unwanted signals are attenuated, the performance of the whole beamformer increases. Also, when there are multiple sources present, being able to track all of them is useful when the system wants to distinguish which one of them is the actually wanted signal. In microphone beamforming this could be useful for an ASR system - a phone for example - which is in a space with multiple speakers. If the system notices multiple speakers, it can then separate each of them and recognize what each of them is saying and try to decide which of the speakers is the one the system should be following. [1]

Source localization and tracking for microphone arrays usually consists of three stages. In the first stage, Time Difference of Arrival (TDoA) or Direction of Arrival (DoA) are extracted from the microphone signals. In the second stage, the information extracted in the first stage is used to derive the position of the source in 3D space based on the microphones and the geometry of the array. The third stage consists of tracking the source in the space in the case it moves. The third stage is optional as source tracking is not always needed and implemented. [1]

Most of DoA estimation techniques for microphone arrays assume far-field and narrowband conditions and they can be divided into two categories: Beamforming based methods and subspace based methods. These two categories are also known as steered response power techniques (SRP) and High-Resolution Spectral Estimation techniques (HRSE). In short, SRP techniques work by steering the beamformer - by adjusting the delays - around space and finding the direction with the most energy. As different delays correspond to different locations in space, by finding the set of delays with the most energy the direction with the most energy is also found. Some of the most successful localization systems are based on SRP techniques. HRSE techniques are also quite common and the most well-known technique of this category is the MUSIC algorithm. [1]

TDoA estimation in speech applications is commonly based on cross-correlation. Multi-channel spatial correlation matrix can be calculated with the help of spatial prediction and interpolation, and from the matrix time delays between the microphones can be estimated. Cross-correlation between signals $x_1(t)$ and $x_2(t)$ is defined as

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} X_1(f)X_2^*(f)e^{-j2\pi f\tau}df, \tag{1}$$

where $X_1(f)$ and $X_2(f)$ are Fourier transforms of signals $x_1(t)$ and $x_2(t)$ respectively. In an ideal case, the cross-correlation function has a peak at such a value of τ that corresponds to the lag between the two microphones. Thus by finding the maximum value τ_{12} of the function, the delay between the two signals - and microphones - can be found. Yet in real world this peak can be masked by other factors, such as noise, reverberation and differences in the paths from the source to the microphones. The robustness against these factors can be improved by weighting the cross-correlation function with a frequency dependent function. This is called generalized cross-correlation (GCC) and with it the TDoA can be estimated with

$$\tau_{12} = \max \int_{-\infty}^{\infty} \psi(f) X_1(f) X_2^*(f) e^{-j2\pi f \tau} df. \tag{2}$$

As an interesting note, it can be realized that finding the maximum of the Equation 1 is equivalent to SRP in the case of Delay-and-Sum with only two microphones. [1]

Common weights for the Equation 2 include Maximum Likelihood (ML) and Phase Transform (PHAT) weighting functions. In environments with low reverberation and noise, the usage of weighting of $\psi(f) = 1$ is usually valid. On the other hand, if the noise spectrum for each microphone is known, ML weighting performs better. The problem is that the prior knowledge of noise is seldom available, and ML is not very efficient in reverberant environments. The most common weighting function is the PHAT, as it is very simple and robust to reverberation. PHAT is defined as

$$\psi_{PHAT}(f) = \frac{1}{|X_1(f)||X_2^*(f)|} \tag{3}$$

and by looking at the definition, the benefits are clear: unlike ML, PHAT doesn't depend on the noise spectra. In addition, because each frequency is equally important in PHAT it is robust to reverberation if signal to reverberant ratio is constant for all frequencies, which is a reasonable assumption. [1]

After finding multiple TDoAs - amount of them depending on the pair of microphones - or DoAs, the position of the source can be estimated by minimizing error function. The idea is to find such a position for the source that the theoretical delays corresponds the best with the estimated ones. For example, for P pairs of microphones we have P TDoA estimations and thus the ML error function can be written as:

$$E(\mathbf{x}) = \sum_{p=1}^{P} \frac{1}{\sigma_{\hat{\tau}_p}^2} |\hat{\tau_p} - \tau_p(\mathbf{x})|^2 \tag{4}$$

where $\hat{\tau_p}$ is the estimated TDoA of the p-th pair of microphones, $\hat{\tau_p}(\mathbf{x})$ is the theoretical TDoA from a potential source at position $\mathbf{x}$ and $\sigma_{\hat{\tau}_p}^2$ is the variance of the estimated TDoA. By finding the minima of the error function the likely source position can be found.

2.5 MEMS-microphones

A device that transforms acoustical power to electrical power is called a microphone. Generally speaking, microphones are either cavity or free-field microphones. While cavity microphones have an output voltage that depends on the acoustic pressure at its diaphragm, free-field microphones are designed to have an output voltage that depends on the acoustic pressure that would be at its diaphragm's position if the microphone wasn't there. In short, the difference is that the free-field microphones try to compensate for the effects caused by the microphone to the sound field, while cavity microphones don't. [13]

Microphones can be divided into categories based on how they function. These different categories include for example condenser and dynamic microphones. Condenser microphones are generally considered to be the best quality microphones, and are used for measurements where accuracy and high-fidelity are required. They typically consist of cylindrical capsule, with one wall consisting of a diaphragm. Inside the capsule, an electrode is placed behind the diaphragm, and the electrode is charged with electrical charge. As the space between the diaphragm and the electrode varies when the diaphragm is vibrated by an acoustic wave, an AC voltage is generated between them. This is basically a capacitor whose other plate moves according to the acoustic wave, thus giving this type its name. Dynamic microphone on the other hand is a very common microphone type, as they are often used by musicians when they perform live. In short, a dynamic microphone works by having a stationary magnet assembly generating a magnetic field and placing a coil in the magnetic field of the magnet. The coil is connected to the diaphragm of the microphone, and thus vibrates with the acoustic waves. As the coil moves in the magnetic field, an AC voltage is generated and thus output of the microphone is generated. [13]

Traditionally condenser microphones have been used in consumer devices. Previously it was due to the fact that there were plenty of low cost electret condenser microphones manufactured, but lately microelectromechanical system (MEMS) condenser microphones have gained popularity. [14, 15, 16] MEMS overall is a term used to describe devices built with equipment and techniques that were developed for integrated circuit manufacturing. Thus, those techniques enable manufacturing of devices with very small and well defined features. [16] MEMS microphones have several advantages over electrets, including surface mounting capability, potentially smaller packaging and not being as sensitive to acceleration effects. [14] Electret microphones can't handle high temperatures and thus can't be soldered using standard surface mount manufacturing flow. This means that to solder an electret microphone to a device, instead of soldering it at the same time as the rest of the components, a separate assembly flow is needed which adds costs to the total system. In comparison, MEMS microphones can tolerate high temperatures and can be assembled with the rest of the components. Electret microphones have become increasingly smaller with time, but as they are quite complex mechanically, it is believed that making them even smaller would be challenging. Lastly, electret microphones are naturally more sensitive to acceleration and its effects than MEMS microphones, meaning that MEMS microphones are suitable for some applications that electrets might struggle with. [14]

A basic MEMS condenser microphone consists of a package, which houses the MEMS chip and an application specific integrated circuit (ASIC) used to read the MEMS, labeld as "Buffer/Amp Chip" in Figures 2-4. Many of the MEMS microphones have a diaphragm made of polysilicon and a perforated backplate, which together form a variable capacitor. [14, 15] As the diaphragm is biased with a constant charge, the voltage change caused by the moving diaphragm can be measured. Both the constant charge biasing and the voltage change measurement are usually done by an ASIC that is in the same package as the MEMS chip itself. This ASIC then outputs the analog or digital signal with the sound information in it. From the design point-of-view, the housing of the microphone is very crucial, as it protects the device, works as an EMI shielding and affects the acoustical characteristics. In practice, there are two types of MEMS microphones, top and bottom port. Top port microphones have the opening from which the sound gets in to the MEMS sensor at the top of the housing, while bottom port microphones have the port at the bottom among the connectors. [15]

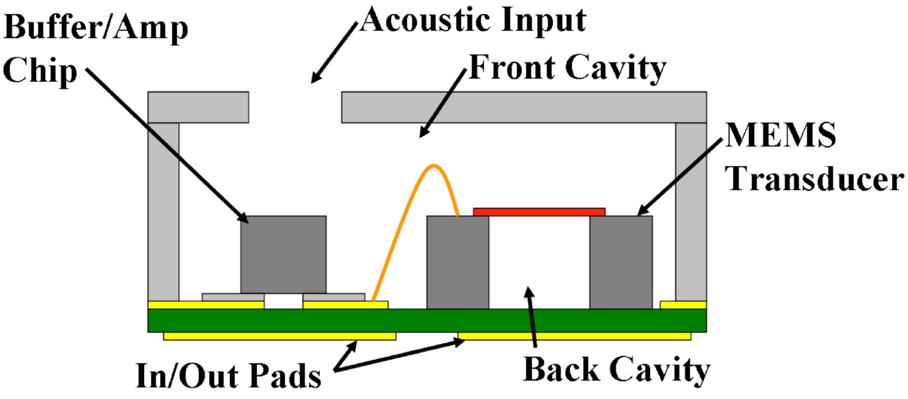


Figure 2: A top port microphone. [16]

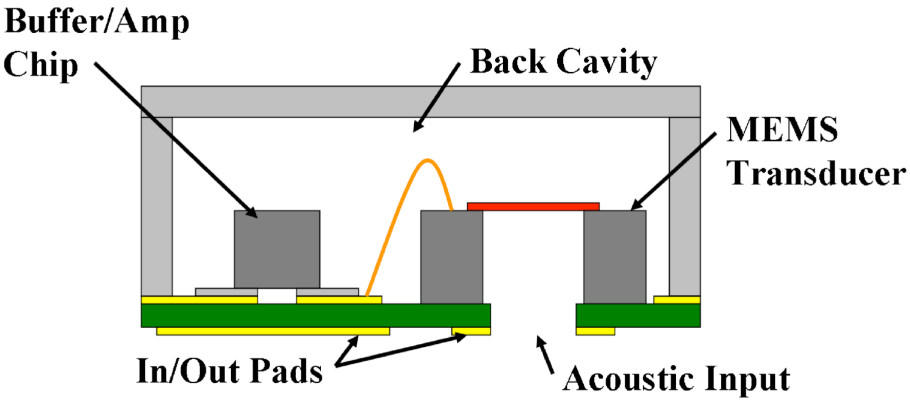


Figure 3: Bottom port microphone. [16]

When comparing top and bottom port microphones, there is a clear difference acoustically in the structures: with top port microphones, the back cavity of the microphone is small, in the Figure 2 it is the area inside the MEMS sensor chip, while in bottom port chip, the back cavity is the whole volume of the module, as seen in the Figure 3. The difference in the back cavity volumes effects the resonance frequency of the microphone, and thus its sensitivity and noise levels. Bigger back cavity volume lowers the resonance frequency, making the bottom port microphone usually a better option from the acoustical performance point-of-view. [15] To increase the

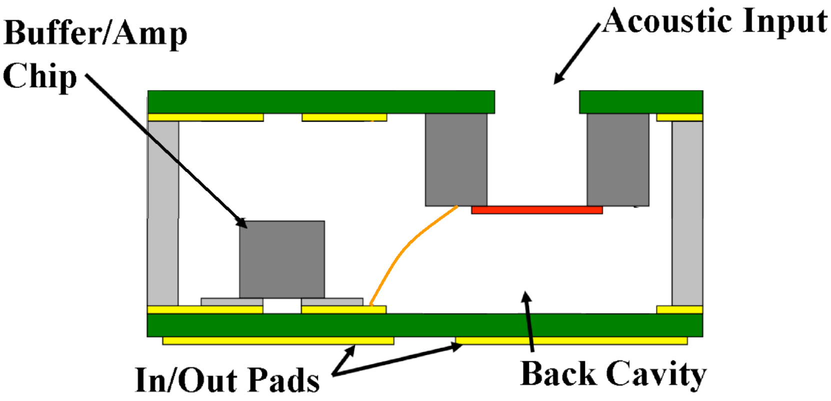


Figure 4: An alternative construct of top port microphone.

performance of the top port microphones, there is a second way to construct them, which more closely resembles a bottom port microphone. In short, the top port can be implemented so that the MEMS sensor is attached to the port at the top, similarly as in bottom port case, but the contacts are on the bottom of the module still. This structure usually leads to longer wirebonds and other leads compared to the more traditional top port structure, which increases inductance to the system which then for example increases noise. This construct is illustrated in the Figure 4.

In addition to condenser microphones, piezoelectric microphones can be constructed with MEMS. Piezoelectric microphones are based on the usage of piezoelectric materials, which are materials that become electrically polarised when mechanical stress is applied. This also works vice-versa, so that when the material is placed in an electric field, a mechanical force is produced. Microphones work by transferring mechanical force - acoustical power - to electrical power, and thus piezoelectric effect can be used to construct them. Piezoelectric microphones have some advantages over capacitive microphones, such as simplicity of fabrication and linearity. In comparison, capacitive microphones traditionally have better sensitivity and noise floor compared to piezoelectric microphones. [16]

The piezoelectric transducer can be made with either diaphragm or cantilever. The diaphragm construct is commonly used, and they produce more normalized output energy if they are well optimized. On the other hand, they are somewhat hard to make and inconsistencies in their parameters lead to inconsistencies in sensitivity and bandwidth and increase the noise floor. The cantilever transducers are somewhat different from the basic diaphragm transducers. In them, the transducer is basically a plate, which is cut so that for example two cantilevers are left in the middle of the plate. When a sound pressure is applied to the plate, the cantilevers vibrate, and the stress caused by the vibration can be transferred to electrical signal with the use of piezoelectric materials. [16]

For this study, condenser MEMS microphones were used, because a device with a form-factor similar to the ones tested in this thesis would most likely use them. In addition, their availability and quality makes them an obvious choice. These quality factors include good SNR, small size and decent sensitivity. Different parameters of the used microphone and better justification for choosing that model are given in the section 4.4

3 Microphone Beamforming

In the light of the problems presented in the section 2.2, microphone arrays are becoming more common. This is due to the fact that they enable de-noising of the received signal and apply spatio-temporal filtering methods, which can be used to avoid, or at least lessen these problems. [2]

In this section microphone arrays for beamforming are presented in more detail, how the architecture of the array affects the result and what kind of different algorithms there are. A solid structure's effect on the performance of the beamformer is discussed, and the effects of the variation in the components' parameters and room reverberation are considered.

3.1 Properties of Beamformers

As the major reasoning behind beamformers is to filter out signals not coming from certain direction or source, most of the different properties of beamformers try to describe how well the beamformer manages to do that. As there are multiple different beamforming algorithms, with various differences in their performance, there are multiple measures that can be used to describe, analyze and compare the beamformers and their performances.

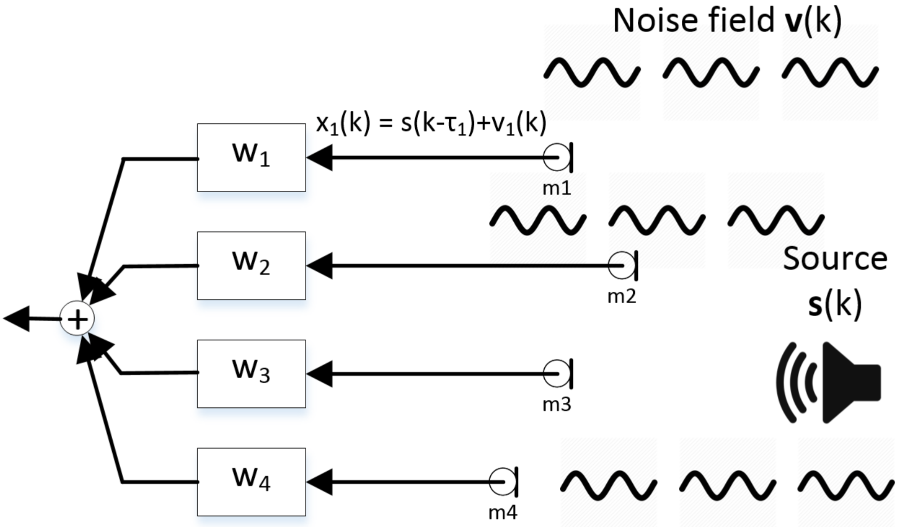


Figure 5: A simple signal model for microphone arrays.

A simple signal model is shown in Figure 5 and it can be defined as:

$$\begin{bmatrix} x_0(k) \\ x_1(k) \\ \vdots \\ x_{N-1}(k) \end{bmatrix} = \begin{bmatrix} a_0 s(k-\tau_0) \\ a_1 s(k-\tau_1) \\ \vdots \\ a_{N-1} s(k-\tau_{N-1}) \end{bmatrix} + \begin{bmatrix} v_o(k) \\ v_1(k) \\ \vdots \\ v_{N-1}(k) \end{bmatrix}. \tag{5}$$

The model is based on the assumption the discrete input $x(k)$ of each n microphones consists of the desired signal that is attenuated and delayed - the $a_i s(k-\tau_n)$

part of the equation - and an arbitrary noise component $v_n(k)$. To make the mathematical definition more compact, it can be written as

$$\boldsymbol{x}(k) = \boldsymbol{a}\boldsymbol{s}(k - \boldsymbol{\tau}) + \boldsymbol{v}(k). \tag{6}$$

As all relevant quantities are dependant on the frequency, this can be transformed to frequency domain with no loss of generality. This Fourier-transform results in

$$\boldsymbol{X}(e^{j\Omega}) = S(e^{j\Omega})\boldsymbol{d} + \boldsymbol{V}(e^{j\Omega}), \tag{7}$$

in which $\boldsymbol{d}$ represents the delays and the attenuation in the frequency domain. The delay and attenuation are dependant on the geometry of the array and the DoA of the source signal. From here, the output signal of the system can be defined as

$$Y_b(e^{j\Omega}) = \sum_{n=0}^{N-1} W_n^*(e^{j\Omega})X_n(e^{j\Omega}) = \boldsymbol{W^H X}, \tag{8}$$

where $W_n(e^{j\Omega})$ are the frequency-domain coefficients of the beamformer's sensors at the frequency Ω and the operator ^{H} is a conjugated transposition. By taking the inverse Fourier-transform of the equation, the result is system's discrete-time output $y_b(k)$. [3]

One of the simpler measures is Array-Gain (AG), which describes the improvement of SNR between one microphone and the array. Thus it is defined mathematically as

$$G = \frac{SNR_{Array}}{SNR_{Sensor}}. \tag{9}$$

[3, 8, 12]

Further, the SNR of a sensor can be calculated by the ratio of the Power Spectral Densities (PSD) of the signal Φ_{SS} and the average noise $\Phi_{V_a V_A}$. The SNR of the array can be calculated from the PSD of the output signal

$$\Phi_{Y_b Y_b} = \boldsymbol{W}^H \boldsymbol{\Phi_{XX}} \boldsymbol{W}, \tag{10}$$

where $\boldsymbol{\Phi_{XX}}$ is a PSD matrix of the array input signals. The output PSD reduces to

$$\Phi_{Y_b Y_b}\Bigg|_{\text{Signal}} = \Phi_{SS}|\boldsymbol{W}^H \boldsymbol{d}|^2, \tag{11}$$

when only the desired signal is present, while if only the noise is present the output reduces to

$$\Phi_{Y_b Y_b}\Bigg|_{\text{Noise}} = \Phi_{V_a V_a} \boldsymbol{W}^H \boldsymbol{\Phi_{VV}} \boldsymbol{W}, \tag{12}$$

where $\boldsymbol{\Phi_{VV}}$ marks the noise's normalized cross PSD matrix. This leads to AG's definition as

$$G = \frac{|\boldsymbol{W}^H\boldsymbol{d}|^2}{\boldsymbol{W}^H\boldsymbol{\Phi_{VV}}\boldsymbol{W}}. \tag{13}$$

[3]

Most of the measures used to evaluate beamformers can be expressed with the help of the coherence matrix

$$\boldsymbol{\Gamma_{VV}} = \begin{bmatrix} 1 & \Gamma_{V_0V_1} & \Gamma_{V_0V_2} & \dots & \Gamma_{V_0V_{N-1}} \\ \Gamma_{V_1V_0} & 1 & \Gamma_{V_1V_2} & \dots & \Gamma_{V_1V_{N-1}} \\ \vdots & \vdots & \ddots & \vdots & \\ \Gamma_{V_{N-1}V_0} & \Gamma_{V_{N-1}V_1} & \Gamma_{V_{N-1}V_2}x & \dots & 1 \end{bmatrix} \tag{14}$$

where

$$\Gamma_{V_nV_m}(e^{j\Omega}) = \frac{\Phi_{V_nV_m}(e^{j\Omega})}{\sqrt{\Phi_{V_nV_n}(e^{j\Omega})\Phi_{V_mV_m}(e^{j\Omega})}} \tag{15}$$

is the coherence function for homogenous noise field. In the case of AG, this leads to

$$G = \frac{|\boldsymbol{W}^H\boldsymbol{d}|^2}{\boldsymbol{W}^H\boldsymbol{\Gamma_{VV}}\boldsymbol{W}}. \tag{16}$$

Theoretically this allows for easy calculations for different kinds of noise fields, as different theoretically defined noise fields have different coherence functions. [3, 17]

Another important measure is Beampattern, which indicates the response of the array to a wavefronts coming from specific angles with specific frequencies. In spherical coordinate system, the beampattern depends on three variables, azimuth and elevation angles and frequency, it can't be displayed in a single plot. Thus, for example in this work, the beampatterns are shown so that beampatterns of different elevation angles or frequencies are shown in their own figures. [3]

While beampattern is very informative and describing measure, it contains lots of information and for layman it can be hard to understand. Thus there are a couple of measures which are used to describe the most important data from the beampatterns. The first is beamwidth, which is a single value to describe how wide the mainlobe of the beampattern is. It is defined as the angle between the $-3\,dB$ points of the mainlobe, so if the lobe's tip is $0\,dB$ and it has $-3\,dB$ points at angles $-10°$ and $10°$, the beamwidth would be $20°$. Another such measure is sidelobe level, which describes how much lower the sidelobes' level is compared to the mainlobe's. It is defined as the difference between the maximum value of the mainlobe and the highest level of the sidelobes.

Beamformer's performance can also be measured with Directivity Index (DI), which describes how well the array suppresses a diffuse noise field. DI is a logarithmic equivalent of directivity factor. Formally DI can be defined as

$$DI(e^{j\Omega}) = 10log_{10}\left(\frac{|H(e^{j\Omega},\phi_0,\theta_0)|^2}{\frac{1}{4\pi}\int_0^{\pi}\int_0^{2\pi}|H(e^{j\Omega},\phi,\theta)|^2\sin(\theta)d\phi d\theta}\right), \tag{17}$$

where $|H(e^{j\Omega}, \phi_0, \theta_0)|^2$ is spatial-temporal transfer function - the beampattern basically -, Ω is frequency, and ϕ and θ are the azimuth and elevation angles respectively. In short, this definition describes the ratio of array's look-direction's transfer function to the spatial integration over all directions of incoming signals. [3]

Alternatively, DI can be defined with the help of the coherence matrix. As the DI describes the beamformer's ability to suppress a diffuse noise, the coherence function of the diffuse noise field needs to be used:

$$\Gamma_{VV}(e^{j\Omega})\Big|_{\text{Diffuse}} = \text{sinc}\left(\frac{\Omega f_s l_{nm}}{c}\right) \tag{18}$$

This results in

$$DI(e^{j\Omega}) = 10log_{10}\left(\frac{|\boldsymbol{W^H d}|^2}{\boldsymbol{W^H \Gamma_{VV}}\Big|_{\text{Diffuse}}\boldsymbol{W}}\right). \tag{19}$$

[3]

Another measure, similar to DI, is Front-to-Back Ratio (FBR), which describes the ratio of signals between front and back of the array. This is useful in cases where there is no clear look-direction, but the sound sources are in front of the array and noise comes from the back. Such applications can be for example video-conferences and orchestra recording. [3]

Lastly, the performance of the beamformer can be measured with White Noise Gain (WNG), which describes the array's ability to suppress spatially uncorrelated noise, for example the self-noise of microphones. The coherence matrix for uncorrelated noise is

$$\boldsymbol{\Gamma_{VV}} = \boldsymbol{I} \tag{20}$$

When this is inserted into the Equation 16, the result is

$$WNG(e^{j\Omega}) = \frac{|\boldsymbol{W^H d}|^2}{\boldsymbol{W^H W}}. \tag{21}$$

[3]

In addition to the basic measures described above, the beamformers can be compared with the help of Word Error Rates (WER). The idea with WER is that as often the use-case of microphone arrays is to improve speech quality, and to help ASR, then it is valuable to measure the beamformer's performance with speech. WER can be measured by recording known spoken sentences in a predefined environment with the array and the beamformers, then feeding the beamformed signal to an ASR. As the sentences are known, the output of the ASR can be compared to the corpus and the WER can be calculated. With the WERs of different arrays with different beamformers a comparison can be made between them to see which array, which beamformer works the best for ASR use-case.

3.2 Architecture of a Beamformer

When designing a microphone array and the beamformer, there are various parameters to consider. In this section, the design choices for the array are discussed, how they effect the beamforming and what kind of phenomena there are that need to be taken into account.

3.2.1 Geometry of the Microphone Array

Practically speaking, there are unlimited number of geometries for microphone arrays. Yet there are several archetypes, which are commonly used and are the most studied ones. One thing to consider is the dimensions of the array: whether it's only a one dimensional line, two dimensional plate or three dimensional structure. Also the way the microphones are placed on two or three dimensional structures is of interest, for example if they are in circular pattern or semi-randomly placed.

The most basic array is the line array, which consists of microphones in a straight line. A simple line array has a directional pattern that is cylindrically symmetrical. In other words, a line array can not differentiate between signals that have the same angle with respect to the axis of the array. So linear arrays can only spot diversity, DoA or other information in one dimension. [1, 12] In addition to this, the beampattern of a linear array is highly dependant of the angle in respect to the array, and it changes with different angles. [18]

An array can also consists of microphones placed in one plane in pretty much any geometry. Having the microphones in a plane instead of a line makes the array able to distinguish sources in second dimension too. Especially circular patterns are commonly used, as they produce nice directivity patterns that can easily be steered towards wanted points in the space. Circular arrays are somewhat limited in that they require quite large area and their locations need to be quite accurate, which can be challenging in for example mobile devices and laptops. Plane arrays have a good resolution in the plane they are in, but they are unable to distinguish if the sound arrives from above or below the plane the array is in. [1, 12] Circular array's beampattern remains constant in the plane the array is in, but when steering them off the plane, the beampattern changes. Another example of 2D-array is an L-shaped four-element array suggested by Microsoft in [19]. The idea in such an array is that it could easily be integrated into a real form-factor device, such as mobile phone, tablet or computer screen. When placing such an array into a corner of the device, the user is unlikely to cover the array with their hands while holding the device.

A three dimensional arrays, of which spherical array has been studied extensively, have the added bonus of being able to steer the beam to any point in space and retaining the same directivity pattern. [18, 20] Yet spherical arrays usually need more microphones than a circular array with similar resolution. [21] Also spherical arrays are challenging from the industrial design point of view, as for a device to have a spherical array, it would need to have a sphere in its design. Most of the consumer devices with microphone arrays - again, computers, mobile phones and screens - don't naturally have a spheres in them, thus making spherical arrays unsuitable for them. Devices that are made for the purpose of having a great sound quality and need beamforming, such as conference phones etc., on the other hand commonly have some kind of spherical part for the array. Yet, it can be argued that for an array whose primary use-case is to record speech of humans, there might not be need for good resolution in more than one plane. [21]

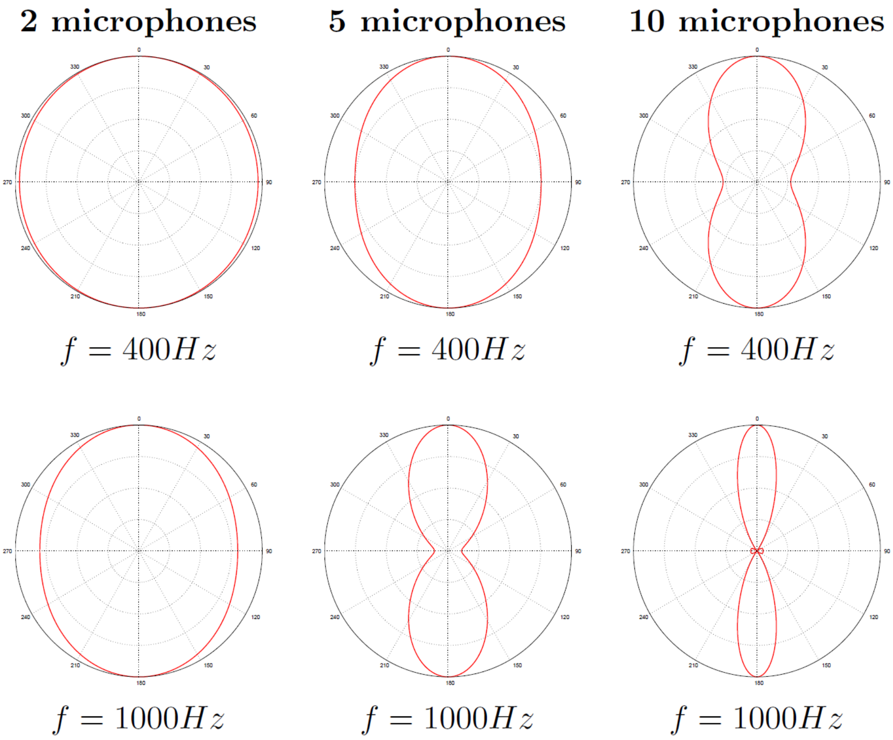
So while line array can only differentiate signals arriving in relation to their angle to the line, plane arrays can only differentiate signals arriving from anywhere but with reduced resolutions and changing directivity patterns, symmetrical three dimensional array's directivity pattern remains constant as it's steered around in space.

3.2.2 Quantity of the Microphones

Another important design decision for microphone arrays is the quantity of the microphones. In practical designs, the amount of microphones is limited by the physical space available and how many microphones can be connected to the system. For example in phones and computers, the microphones are often connected to audio codecs, which have limited amount of microphones inputs, giving an upper limit for the quantity of microphones - though those devices often have size constraints too! The limitations of readily available inputs can be avoided, but it means adding more electrical circuitry, which in turn increase cost of the device. For example, in laptops the microphones could be connected to a integrated sound card specifically designed to have numerous inputs instead of the simple audio codec with only a couple of inputs.

From the performance point-of-view, the amount of microphones is the other big factor in addition to the geometry of the array. As shown by the Figures 6 and 7, if the size of the array remains the same, but the quantity of microphones increases - thus the distance between the individual microphones becomes smaller - the performance at higher frequencies improves. This is due to the fact that with smaller distances between the microphones, the grating lobes caused by spatial aliasing are moved to higher frequencies. On the other hand, if the distance between the microphones remains the same but the quantity changes, thus increasing the overall size of the array, the low frequency performance of the array improves.

In short, the bigger the array is, the better the beamformer's low frequency resolution is. This is due to the fact that as low frequencies have long wavelength. As the beamformers operate by comparing the phases or time delays between the signals recorded by the microphones, with high wavelengths, the signals of two microphones near each other are very close to the same. As there is no big difference in the phases of the signals, the beamformer can't accurately differentiate from which direction the signal is coming, as from its point-of-view, the signal arrives at the same time to both of the microphones. When the microphones are further away, the phase difference between the two microphones increases and thus the beamformer can estimate the direction of the original signal. Consider a case where we are interested in the direction-of-arrival of a $50\,Hz$ sine-wave and we have two microphone arrays; one with two microphones $5\,cm$ apart and the other with two microphones $25\,cm$ apart. With a speed of sound of $343\,m/s$, the first array would notice a phase difference of around 2.6° while the second would notice roughly 13.1°. When considering all the uncertainties, noise and variance in the microphone parameters, a phase shift that is only single digit degrees can easily go unnoticed [22].


2 microphones **5 microphones** **10 microphones**

$f = 400Hz$ $f = 400Hz$ $f = 400Hz$

$f = 1000Hz$ $f = 1000Hz$ $f = 1000Hz$

Figure 6: Directivity patterns of a line arrays with $5\,cm$ separation between microphones. [1]

Another important factor is the distance between the microphones. Simply put, the closer the microphones are to one another, the higher frequencies the grating lobes appear at, due to the spatial aliasing. This is similar to the Nyquist-Shannon sampling theorem which states that in order to prevent aliasing, the sampling rate needs to be at least twice that of the highest frequency of the signal. Similarly in beamforming, signal with wavelength that is higher than twice the distance between the microphones results in spatial aliasing, which can be seen as grating lobes in the directivity pattern. These grating lobes mean that high frequency signals coming from some other directions than the wanted direction aren't attenuated, resulting in worse performance for the beamformer.

3.3 Beamforming algorithms

3.3.1 Delay-and-Sum

Delay-and-Sum is one of the most basic beamforming methods, in which the signals from each microphone are individually delayed and then summed together - thus gaining its name. The idea is that by delaying the microphone signals by certain amounts, the signal originating from a certain point or direction will be emphasized,

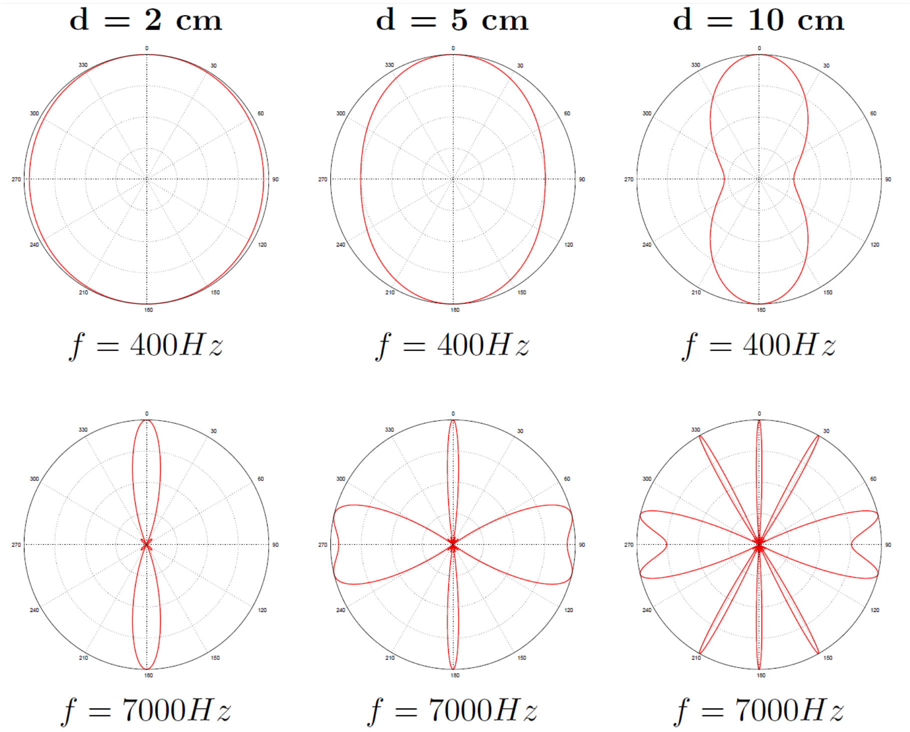


Figure 7: Directivity patterns of a line arrays with 5 microphones with different distances between them. [1]

whereas signals from other directions and points will not be. The amount of delay for a microphone's signal is directly related to the propagation time between the microphone and the reference point. In addition to strengthening the wanted signal, the self-noise of the microphones and their connections will be reduced as they are highly uncorrelated between the microphones. [23, 12, 8]

Even though the Delay-and-Sum is an old and simple method, it remains a powerful option and is still widely used today. The popularity is due to the simplicity of the implementation, low computing cost of the algorithm, and the fact that Delay-and-Sum is widely researched and its limitations and behaviour is well known. Mathematically the output of a Delay-and-Sum beamformer consisting of M microphones can be described with

$$z(t) = \sum_{m=0}^{M-1} w_m y_m(t - \Delta_m) \tag{22}$$

where y_m is the waveform captured by the mth microphone, Δ_m is the delay that corresponds to the mth microphone and w_m is an amplitude weighting for each microphone's output. [12] The amplitude weighting could be such that each microphone is given the same weight, usually either

$$w_m = 1 \quad \text{or} \quad w_m = 1/M, \quad m = 1, 2, ..., M, \tag{23}$$

or then the weighting function can be a window function. Possible tapering windows include triangular window, which is defined as

$$w_m = \frac{M - m + 1}{M}, \quad m = 1, 2, ..., M, \tag{24}$$

Hamming window, defined as

$$w_m = 0.54 - 0.46 \cos\left(\frac{\pi(m - M - 2)}{M + 1}\right), \quad m = 1, 2, ..., M, \tag{25}$$

Hann window, which is very similar to the Hamming window, defined as

$$w_m = 0.5 - 0.5 \cos\left(\frac{\pi(m - M - 2)}{M + 1}\right), \quad m = 1, 2, ..., M, \tag{26}$$

and Blackman window, defined as

$$\begin{aligned} w_m = 0.42 - 0.5 \cos\left(\frac{\pi(m - M - 2)}{M + 1}\right) + \\ 0.08 \cos\left(\frac{2\pi(m - M - 2)}{M + 1}\right), \quad m = 1, 2, ..., M. \end{aligned} \tag{27}$$

[24] Different weighting functions change the behaviour of the directivity pattern of the beamformer, usually so that if the sidelobes attenuate, then the beamwidth increases and vice versa. [8, 24]

The most simple Delay-and-Sum only takes into an account the sound pressures measured by the microphones and calculates the plane wave's directions of arrival from them. This method can only estimate the direction, but if more advanced, called *focused*, methods are used then the beamformer can listen to a points instead of directions. Thus with focused beamforming, different sources in the same direction can be separated. [25] Simply put, to detect signals from certain points in space, we need to assume the source is in *nearfield* instead of *farfield* and thus in addition to the delay between microphones, we also have an attenuation between them. The attenuation is caused by the spherical spreading of the sound wave. In addition, in *nearfield* scenario, the delays are different than in the *farfield* case, for the sound wave is shaped differently: in the former case, the wave is spherical and in the latter it is a plane. [12] Focused methods require the array to be big, or the source to be near for them to work, for otherwise the attenuation of the sound and the difference between a plane-wave and spherical-wave are very small. [25]

Consider the setup visualized in Figure 8; due to the differences in the distance of the microphones from the speaker, the sound from the speaker doesn't arrive at the same time to all of the microphones, instead arriving slightly earlier to the microphones closer to the speaker. Let the time difference between the signals arriving to the microphone n and 4 be Δ_n: if we delay the signal of m1 by Δ_1, then the signals of the microphone m1 and m4 will be in the same phase, if we delay the signal of m2 by Δ_2 then the signals of the microphone m2 and m4 will be in the same phase etc. As the delayed signals are in the same phase, their summing results in constructive interference, and if there are some components in the signals that are out-of-phase, such as uncorrelated noise or signals coming from wrong directions, they will experience destructive interference.

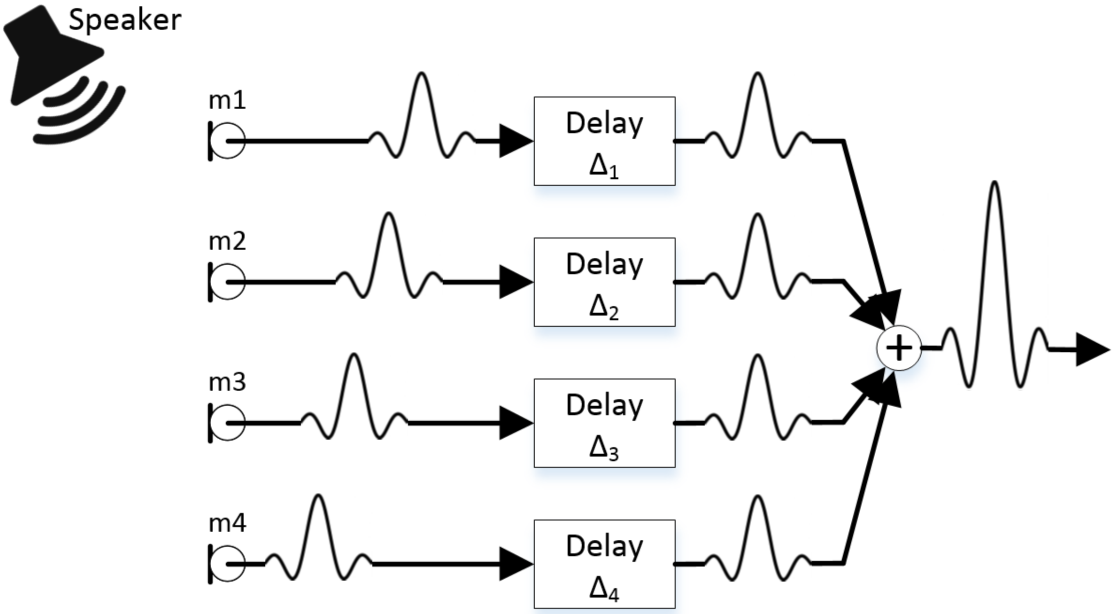


Figure 8: A setup used to illustrate Delay-and-Sum.

For Delay-and-Sum beamforming to work, one of the two variables needs to be known: the speed or the direction of the propagating signal. In microphone beamforming, the speed of the signal is usually known, as the speed of sound is quite stable, ie. it doesn't suddenly change.

3.3.2 Filter-and-Sum

Filter-and-Sum is considered a generalisation of the Delay-and-Sum algorithm. The fundamental idea in Filter-and-Sum is that filtering the outputs of the microphones with linear filters helps to remove unwanted disturbances from the signals. For example, if it is known that the array is going to be used in an application where the interesting frequency range is from $200\,Hz$ to $10\,kHz$, then the outputs of the microphones could be filtered so that only that range is left, to remove noise that could affect the performance of the beamformer. Filter-and-Sum can be defined similarly as the Delay-and-Sum, with the following equation:

$$z(t) = \sum_{m=0}^{M-1} w_m y_m(t - \Delta_m) \tag{28}$$

where again Δ_m is the delay that corresponds to the mth microphone and w_m is an amplitude weighting for each microphone's output. This time though, the y_m is the filtered waveform captured by the mth microphone, which can be defined as

$$y_m(t) = \int_0^\infty h_m(\tau) f(\boldsymbol{x}_m, t - \tau) d\tau \tag{29}$$

where h_m is the impulse response of the filter for mth microphone, $f(\boldsymbol{x}, t)$ is the measured wavefield, and $\boldsymbol{x}_m$ is the vector describing the location of the mth microphone. [12]

3.3.3 Superdirective Beamforming

Superdirective beamformers are beamformers whose main benefit compared to standard Delay-and-Sum beamformers is higher directivity. This suggests that if directivity is the desired parameter of the beamformer, then just summing the signals isn't the best choice. As directivity means beamformer's ability to suppress noise without affecting a desired signal, the main design criterion of Superdirective beamformer is the ability to suppress noise.

For Superdirective Beamformers, to achieve the design criterion the power of the output signal $y_b(k)$ of the array needs to be minimized. The output Power Spectral Density (PSD), as defined by the Equation 10, depends on the input signal and the coefficients of the beamformer. As there is possibility of a trivial solution at $W_n = 0$, the minimization needs to be constrained with

$$\boldsymbol{W}^H \boldsymbol{d} = 1. \tag{30}$$

Thus the minimization problem to be solved is defined as

$$\min_{\boldsymbol{W}} \boldsymbol{W}^H \boldsymbol{\Phi}_{\boldsymbol{XX}} \boldsymbol{W} \quad \text{subject to} \quad \boldsymbol{W}^H \boldsymbol{d} = 1. \tag{31}$$

[3]

For Equation 31 there is solution called Minimum Variance Distortionless Response (MVDR) beamformer. If a homogeneous noise field is assumed, then the solution can be defined as a function of the coherence matrix $\boldsymbol{\Gamma}$:

$$\boldsymbol{W} = \frac{\boldsymbol{\Gamma}_{\boldsymbol{VV}}{}^{-1} \boldsymbol{d}}{\boldsymbol{d}^H \boldsymbol{\Gamma}_{\boldsymbol{VV}}{}^{-1} \boldsymbol{d}}. \tag{32}$$

This equation is basically a spatial decorrelation process with a matched filter for the desired signal. The signal response towards the look direction is unity thanks to the normalization in the denominator. This design procedure means that to get optimal designs for different applications, correct noise-fields need to be chosen. In addition, the desired signal can be modelled differently to get designs for farfield and nearfield scenarios. [3]

Common well-behaving noise-fields used for Superdirective beamformers are spherical isotropic noise-field - ie. diffuse-field - and cylindrical isotropic noise-field. Cylindrical noise-field is similar to diffuse-field, but instead of noise arriving from every direction in three dimensional space, it arrives from everywhere in two dimensional space. This corresponds well to the noise that is present when a lot of people speak in a large room with well damped floor and ceiling or in the free-field. So basically cylindrical isotropic noise-field assumption works well with the cocktail-party scenario, and that is the reason why it is often used for applications that need speech enhancement, such as hearing-aids. [3]

The classic Superdirective beamformer (SDB), is designed by solving the Equation 32 with the help of the coherence matrix of the diffuse noise field, which is given in the Equation 18. The problem with this kind of SDB is that it actually boosts white noise at low frequencies, making it unsuitable for any real-world application. The boosting is caused by the coefficients of the beamformer, which force the phase of the noise between microphones to be π, which compensates the correlated part of the noise, but also reduces the desired signal. Because one requirement for the beamformer was that it shouldn't affect the desired signal, the input signal needs to be boosted. This results in boosting of uncorrelated noise. [3]

The self-noise amplification problem of SDB can be improved by adding a small scalar μ to the main diagonal of the normalized PSD or coherence matrix:

$$\boldsymbol{W} = \frac{(\boldsymbol{\Gamma}_{\boldsymbol{VV}} + \mu \boldsymbol{I})^{-1} \boldsymbol{d}}{\boldsymbol{d}^H (\boldsymbol{\Gamma}_{\boldsymbol{VV}} + \mu \boldsymbol{I})^{-1} \boldsymbol{d}}. \tag{33}$$

There is an alternative form, which preserves the interpretation as a coherence matrix whose elements are smaller than one: This alternative form differs so, that instead of adding the scalar to the main diagonal, each non-diagonal element is divided by $1 + \mu$. This leads to the interpretation of μ that is represents the ratio of the sensor noise σ^2 to the ambient noise power Φ_{VV}. So, the non-diagonal elements of a diffuse field can be calculated as

$$\Gamma_{V_n V_m}(e^{j\Omega}) = \frac{\operatorname{sinc}\left(\dfrac{\Omega f_s l_{nm}}{c}\right)}{1 + \dfrac{\sigma^2}{\Phi_{VV}}} \tag{34}$$

This factor μ can be anything between zero and infinity, but typically it has values between $-10\,dB$ and $-30\,dB$. [3]

3.3.4 Adaptive Methods

The previously described algorithms are all fixed beamformers, which generally speaking are quite simple methods. The problem with fixed beamformers is that the real world doesn't remain fixed, so the noise characteristics and other disturbances vary with time and place. This results in the performance deteriorating due to the beamformers inability to adapt to the changes, instead trusting on the original assumptions of the noise. To avoid this, there are adaptive beamformers which are designed to take the changing environment into an account, and change the beamformer's parameters accordingly. Overall, the interference suppression capabilities of adaptive beamformers are generally better than that of fixed beamformers. [3, 12] On the other hand, adaptive methods are generally more sensitive to sensor calibration errors than fixed methods. These errors include sensor location and transfer characteristics of the sensor. The term *adaptive* can be interpreted as meaning any method whose characteristics depend on the input data, or more restrictedly as only methods that update themselves after each observation is obtained. In beamforming literature, "adaptive" often refers to the former definition. [12]

Overall, there are numerous different adaptive beamformers, with different kind of designs, strengths, weaknesses and target applications. Usually adaptive beamformers are much more complex than fixed beamformers, but as their performance is, generally speaking, better there is much interest in researching them and using them in real-world applications. Different adaptive methods include Capon, MUSIC, Maximum Likelihood and Griffiths-Jim Beamformer.

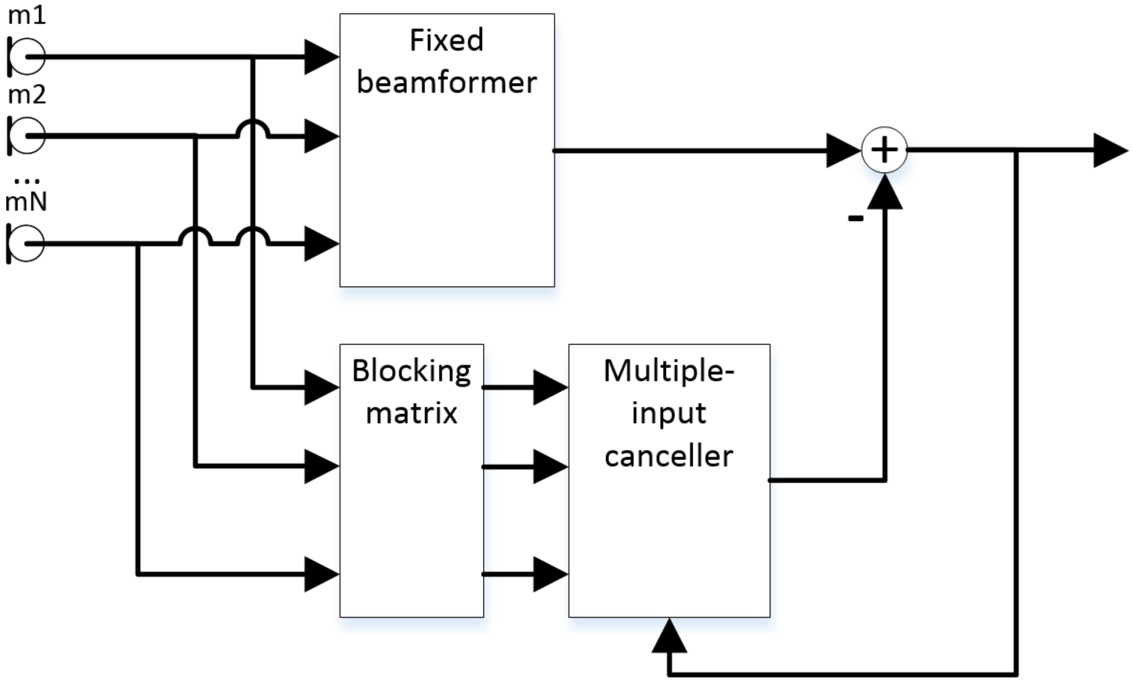


Figure 9: A block diagram of Griffiths-Jim Beamformer.

An example of an adaptive beamformer is Griffiths-Jim Beamformer (GJBF), which is the Adaptive beamforming algorithm used in this thesis. GJBF consists of three parts: of a fixed beamformer, a multiple-input canceller and a blocking matrix, as shown in Figure 9. In the simplest implementation, the fixed beamformer can be for example a Delay-and-Sum beamformer. The output of the microphones is fed to the fixed beamformer and the blocking matrix. The blocking matrix is basically inverse of a beamformer: it’s task is to block signal coming from a certain direction. One simple method of implementing a blocking matrix is to use Delay-and-Subtract, which is roughly the same idea as Delay-and-Sum but instead of summing, subtracting the delayed signals. In GJBF, the blocking matrix blocks the signal coming from the same direction where the fixed beamformer is looking, ie. the direction of the wanted signal. Thus the output of the blocking matrix is the measured signal without the wanted signal. [3] The output of the blocking matrix is then fed to the multiple-input canceller, which generates copies of the interferences’ correlated components with adaptive filters. These copies are then subtracted from the output of the fixed beamformer. This results in the system’s output signal having enhanced version of the desired signal, thanks to the fixed beamformer, and suppressed noise, thanks to the blocking matrix and multiple-input canceller. The output of the system is used to control the adaptive filters, thus making GJBF an adaptive beamformer. [3]

In this thesis, a simple GJBF was used: the fixed beamformer was Delay-and-Sum beamformer, while the blocking matrix was done with Delay-and-Subtract. Delay-and-Subtract is implemented so that the signal of microphone N is delayed by the same amount as in the Delay-and-Sum beamformer, and then subtracted from the delayed signal of microphone $N-1$ etc., resulting in $N-1$ outputs. These $N-1$ outputs are then fed into Multiple-Input Canceller, which consists of $N-1$ adaptive filters. These filters are adapted using Least Mean Squares (LMS) algorithm such that

$$\mathbf{w}_i(n+1) = \mathbf{w}_i(n) + \alpha \frac{y(n)}{\mathbf{z}_i^T(n)\mathbf{z}_i(n)} \mathbf{z}_i(n), \quad \alpha \in [0,2], \quad i \in [1, N-1] \tag{35}$$

where $\mathbf{w}_i$ stands for the coefficient vector of the ith filter, $y(n)$ is the output of the whole system, $\mathbf{z_i}$ are the outputs of the blocking matrix and α is the step size of the adaptation algorithm. Output of each of these filters is summed together into one signal, which is then subtracted from the output of the fixed beamformer. The resulting output is then used as the error signal for the adaptation algorithm.

Even though GJBF is an adaptive beamformer and its performance theoretically is good, in practice it has some problems, such as lack of robustness. For example, if there is a slight mismatch between the DoA-estimate and the actual DoA, then the GJBF will also attenuate the wanted signal. There are many reasons for mismatch between the estimate and the real DoA, including differences between microphone gains and positions, and reverberation of the space. Because the mismatch occurs so easily, the target-signal cancellation can't be avoided, making GJBF's lack of robustness a serious problem. To improve GJBF's robustness, different techniques have been proposed and they are often called *robust adaptive beamforming*, as their target is to be robust against errors. These techniques can include improvements in either blocking matrix or multiple-input canceller. [3, 26] For example in their paper, Hoshuyama et al [26] proposed using adaptive blocking matrix, which uses the fixed beamformer's output to adapt the matrix, and norm-constrained adaptive filters in the multiple-input canceller.

3.4 Effect of a Structure on a Beamformer

The structure effects the signals recorded by microphone due to various effects, including scattering [18, 27, 28, 29] and changing the directivity of the microphones [30, 18]. Generally speaking, for spherical arrays, the open design performs worse when compared to an array attached to a solid structure. The error of the open array increases at certain frequencies depending on the physical characteristics of the array, ie. the diameter. [21] Yet as the solid sphere interferes with the sound field the array is trying to capture, there are some additional considerations. For example, the sound waves that scatters from the solid sphere can be reflected back from the walls or other objects in the room, and thus creating additional incident waves. [28] In addition, the diffraction of the sound caused by the solid sphere improves the SNR of the beamformer, especially at low frequencies. [18] For Delay-and-Sum Beamformers in an open and solid spheres, it can be said that the solid sphere is more directive, and its White Noise Gain (WNG) is better. [20] For dereverberation purposes, it was found out that Delay-and-Sum Beamformers perform $1 - 2\,dB$ better in a solid sphere than in an open sphere configuration. [31]

When comparing circular baffled and unbaffled circular arrays, the phenomena are similar to those of baffled versus unbaffled spherical arrays. In their study, Tiana-Roig et al [29] found that with Delay-and-Sum beamformer, the Maximum Sidelobe Level (MSL) and resolution are better for baffled arrays than for unbaffled. The effect was such that a $10\,cm$ baffled circular array had similar MSL and resolution as a $20\,cm$ unbaffled circular array. Thus, mounting the array on a solid structure makes the array seem like a "larger" unbaffled array.

When designing a device that is targeted to consumers, such as mobile phone, laptop or television, the microphones are usually attached to a rigid structure. Due to the aesthetics of the devices, the microphones are usually to be hidden somehow inside the structure, with only a small hole leading to the microphone. Instead, if the aesthetics were not a factor, the microphones could be just left visible on the device, sticking slightly out of the structure.

The fact that there is a small cavity between the microphone and the outside of the device affects the directivity of the microphone. Commonly the microphone components used in consumer devices are omnidirectional, but with the cavity the microphone's directivity increases, especially at higher frequencies. For low frequencies, the change is practically non-existing, because the cavities are so small compared to the wave lengths, but when the wavelength approaches the size of the cavity, the effect becomes clear. For example, if the cavity's diameter is $4\,cm$ and depth is $2\,cm$, the directivity is changed already at frequencies above $2\,kHz$, and at frequencies over $10\,kHz$ the change is very noticeable. [30]. Yet in most practical form-factors, the cavities would be a lot smaller, the diameter and depth being in the range of few millimetres. With such small cavities, the effect is mostly non-existing in the audio range, as a $5\,mm$ wavelength would correspond to $67\,kHz$, which is out of range for basically any microphone beamformer. Nevertheless, the mechanical integration of the microphone components affects the directivity of the microphones, thus even though the components themselves are omnidirectional, the finished product's microphone might not be.

3.5 Variation of Microphones' Parameters Effect on a Beamformer

Microphones have multiple parameters that all affect their performance. Most often cited parameters include sensitivity, frequency response, Signal-to-Noise Ratio (SNR) and Total Harmonic Distortion (THD). When multiple microphones are used for beamforming, changes in those parameters change the performance of the whole beamformer, and for the best quality the microphones should be matched as closely as possible. Yet perfect matching is often not possible or at least not viable, so knowledge how the variance in the parameters affect the beamformer is important.

Generally speaking, variance in the phase responses of the microphones causes degrading performance mainly in lower frequencies. This is because the acoustic phase difference between two microphones is lower at low frequencies, and even small variance in the phase response can then be significant compared to it. For example, for two microphones with distance of 20 cm, a sinusoidal wave with frequency of $20\,Hz$ would have maximum acoustical phase variation of 4.2°, while $200\,Hz$ would have 42°. [22] As differences between the phase responses of two similar microphones remains quite constant across frequencies the error is more pronounced at low frequencies.

Sensitivity of the microphones describes the level of the output relative to the air pressure at the microphone's diaphragm. Variance in the sensitivity of the microphones thus leads to the same air pressure generating different voltages at the output of the microphones. Lets consider two microphones, one with sensitivity of $X\,dBV/Pa$ and the other with sensitivity of $0.9X\,dBV/Pa$. For a simple Delay-and-Sum beamformer this results in summation that is slightly weighted, even though the algorithm itself would have equal weights: the second microphone appears to be weighted by 0.9 already. Yet good quality microphones are pretty well matched, so if an array is constructed out of the same microphones, this isn't usually a problem. For example the microphones used in this thesis have sensitivity of $-38 \pm 1\,dBV/Pa$, so the error isn't huge [32]. Overall it can be said that the error caused by variances in the microphone parameters decreases, as the quantity of microphone increases.

3.6 Effect of Room Reverberation and Echoes on a Beamformer

As microphone arrays are often used in spaces that are not anechoic, the room's reverberation and echoes affect the performance of the beamformer. The reverberation and echoes can mask the original DoA of the desired signal and at least cause attenuated versions of the desired signal to appear from different directions. This can cause for example some simple adaptive beamformer to adapt so that the desired signal is also slightly attenuated, as it appears among the "noise" in addition to the DoA. One benefit of using microphone arrays instead of single microphone is that the array can perform echo cancellation and dereverberation on incoming signals, and basically beamforming itself is a way to dereverberate and denoise signals.

Acoustic Echo Cancellation (AEC) is especially important for applications with natural human-machine interactions, such as Apple's Siri, Amazon's Alexa/Echo, Google's Assistant/Home and Microsoft's Cortana, for there are scenarios where the device needs to communicate and listen at the same time, but also for conference phones etc. In this kind of scenario, the echo to cancel is known and thus only the room and it's effect on the signal needs to be considered. To simplify, AEC aims to remove the signal $v(n)$ from the microphone signal $x(n)$. For a simple example, lets consider the structure shown in Figure 10. To achieve its target, AEC tries to produce a copy $\boldsymbol{v}(n)$ of the signal $v(n)$ which is the device's output signal $u(n)$ echoed to the microphone. In addition to the signal $v(n)$, the input signal $x(n)$ contains the desired sources $s(n)$ and interferences of the space $r(n)$. If we define residual echo as $e(n) = v(n) - \boldsymbol{v}(n)$, then the desired signal $\boldsymbol{s}(n)$ becomes

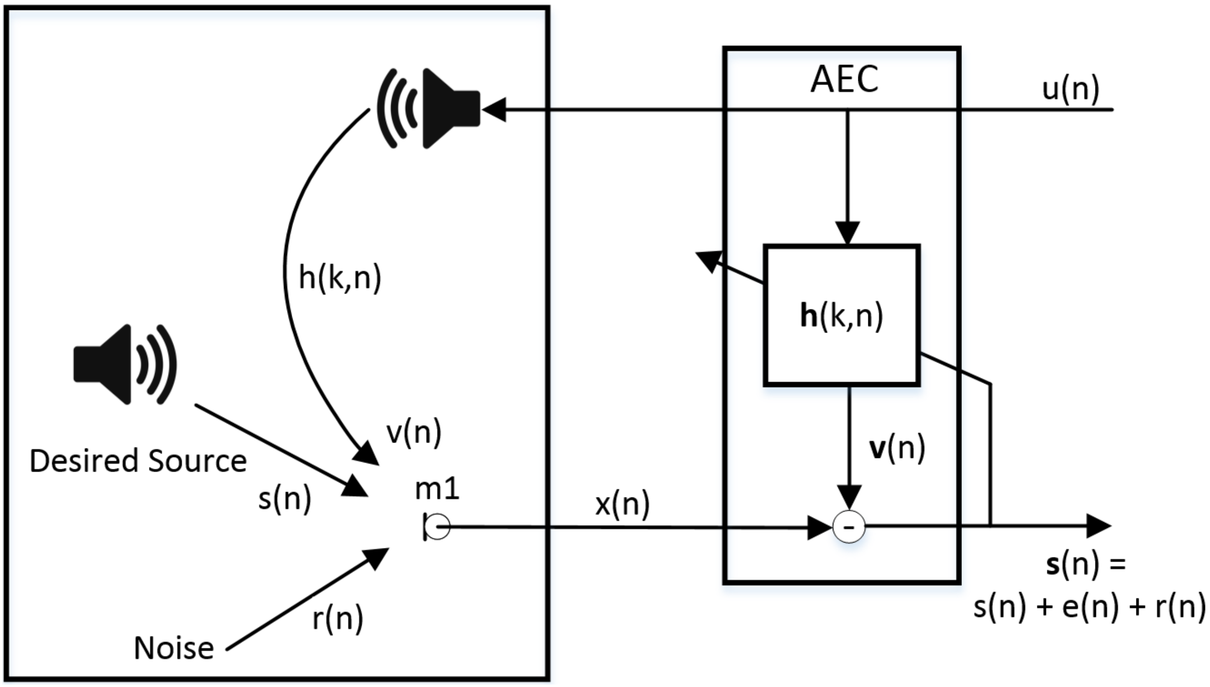


Figure 10: Diagram of single-channel AEC.

$$\boldsymbol{s}(n) = x(n) - \boldsymbol{v}(n) = s(n) + e(n) + r(n). \tag{36}$$

So the problem of AEC is basically measuring, modelling or otherwise getting the impulse response of the loudspeaker-room-microphone-system $h(k, n)$. It should be noted that the impulse response varies over time as objects in the space are moved and the environment varies, for example its temperature changes. [3]

Room's reverberation, on the other hand, is a problem practically always when the microphones are not close to the speaker. As hands-free operation is increasingly common, be it for mobile devices, haring aids or other devices, it is quite common that the microphones are not next to the speaker. To improve the signal quality in such an environment, the disturbances caused by the reverberation, ambient noise and acoustic echo need to be filtered out. Filtering out the ambient noise and dereverberating the signal are somewhat connected problem, or at least often they can be performed simultaneously. Overall dereverberating is not a trivial problem, as the environment can vary very much over time for a single device, and for different devices the environments can be totally different, thus the filtering needs to work well without precise knowledge of the acoustic characteristics of the space. [2, 33, 34]

Even a simple Delay-and-Sum beamformer can perform dereverberation and denoising quite well but yet again, the performance at lower frequencies is highly dependant on the size of the array. To be able to use smaller arrays, other dereverberation techniques need to be used. One common technique is the MVDR beamformer, which dereverberates and denoises the signal at the same time, but it can be argued that by performing them both at the same time decreases the overall performance of the beamformer. [2]

Beamformer's dereverberation performance can be measured by Direct-to-Reverberant Ratio (DRR). DRR is defined as the ratio of the energy of the direct signal component to the energy of the reflected signal components. [1, 31]

Overall, room reverberation has a big impact on source localization performance, as in some cases the direct sound might actually be weaker than for example the first reflected sound. Such a case could be for example if the speaker is speaking away from the array, and as the directivity pattern of head attenuates sound that is going backwards, the sound that is reflected from a wall or other object nearby can be stronger. If the object from which the sound reflects is in different direction than the speaker, this can confuse the source localization algorithm. [1]

4 Research Material and Methods

4.1 Tested Structures

4.1.1 Reference Circular Array

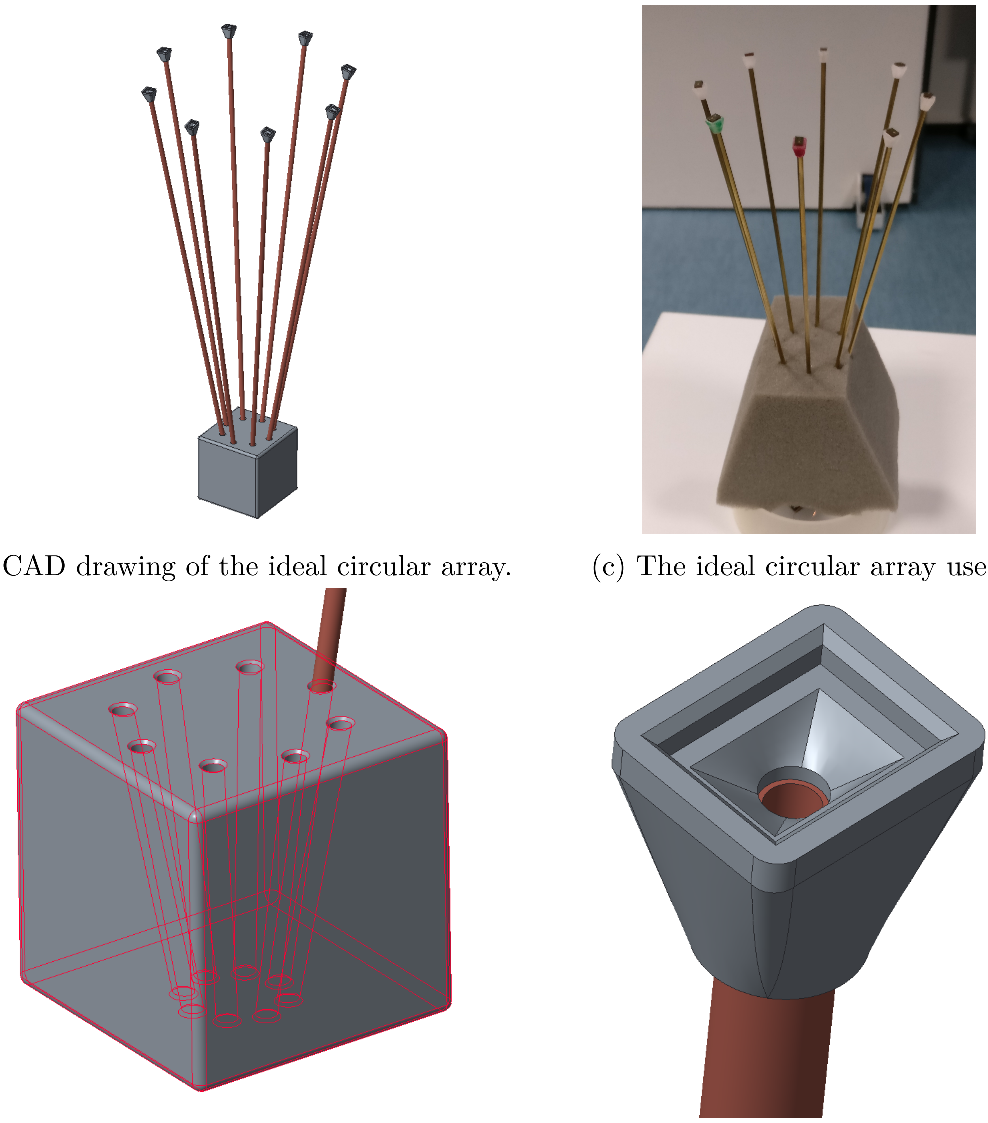
The first array to be tested is an "ideal" circular array, meaning that it is basically eight microphones held in free space. This array is used as a reference for the other arrays, so that the structures' effects on the results and performance can be easily noticed. In this case, this was implemented by attaching the microphones to the ends of $2\,mm$ thin brass tubes. The microphones form a circle with a diameter of $83\,mm$. The reason for the microphones to be attached to the ends of tubes is to raise them away from the control electronics. As the MEMS-microphones used need a power supply and a connection to a sound card, there is a Printed Circuit Board (PCB) to which the tubes are attached. If the microphones would be too close to the PCB, it could cause reflections which would affect the performance of the beamformer. As the microphones are raised some $200\,mm$ from the PCB, the effect of possible reflections is minimized. To further minimise the possible reflections and other effects caused by the block, it is hidden inside a foam block. The foam is some basic open cell-type packaging foam used when posting devices, which is then cut to shape and size that would not be overly big.

CAD drawings and a picture of this array are shown in Figure 11. In the figures, it can be noticed that the array consists of three parts: a cube, tubes and microphone holders. The cube, whose size is $30 \times 30 \times 30\,mm$, is illustrated in Figure 11b. The cube is there to hold the tubes in place and make sure their angles are correct. PCB to which the microphones are connected is attached to the bottom of the cube. The holders, illustrated in figure 11d are then used to attach the microphones to the tubes. As the microphones are sized $2.65 \times 3.5\,mm$, the tubes' outer diameter being $2\,mm$ and the inner diameter $1\,mm$, there needs to be this kind of holder to keep the microphones in place and enable routing of the wires.

This array is used to obtain a reference for the two other arrays, to assess how the mechanical structures affect the performance of the beamformers. Yet of all the structures, this is the one with most uncertainties, as the locations of the microphones is somewhat hard to accurately check. Each of the microphones is at the end of a tube, all of which can be moved up and down, and the tubes might bend slightly. This results in bigger placement uncertainty for the microphones, even though the distances between the microphones have been measured and they are placed as close to the correct locations as possible. When comparing to the next two arrays, their microphones' locations are much more accurate as there are rigid mechanical guides to position them at exactly correct places.

4.1.2 Array on a Solid Cylinder's Top Surface

The second array to be tested is a circular array of eight microphones on the top surface of a cylinder. In this array, the microphones also form a circle with a diameter of $83\,mm$. The cylinder to which the microphones are attached has a diameter of



(a) CAD drawing of the ideal circular array.

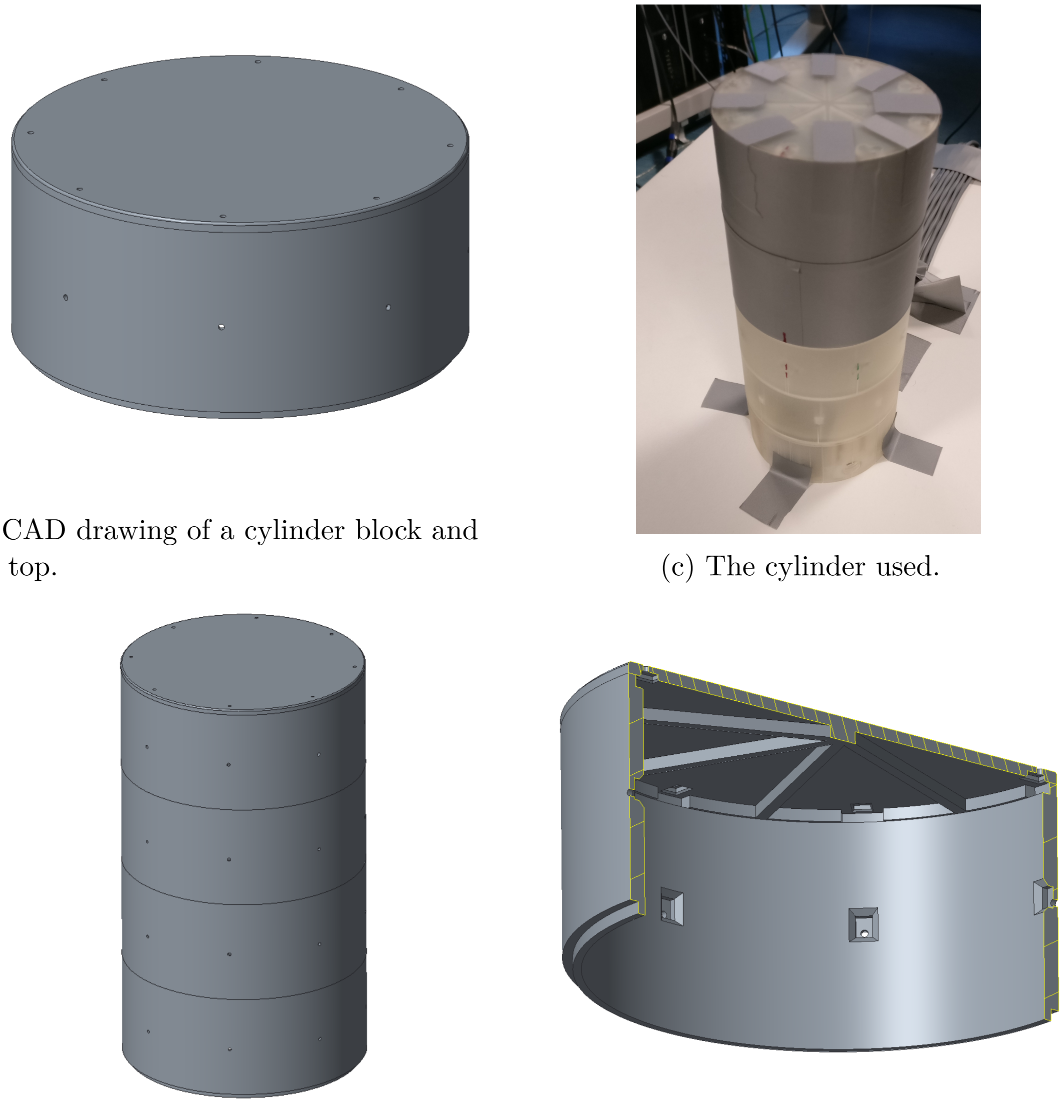
(c) The ideal circular array used.

(b) CAD drawing of the cube.

(d) CAD drawing of the microphone holder.

Figure 11: Illustrations of the array

$90\,mm$, so the microphones are very close to the edges of the structure. The cylinder and its top can be seen in Figure 12. In the figure 12a a single block and the top of the cylinder are illustrated. The idea is that multiple blocks can be stacked on the top of each other, as shown in Figure 12c. With multiple blocks stacked atop each other, the height of the cylinder can be adjusted, which opens interesting possibilities in testing the structure's effects. The inside of a block and the top can be seen in the cross-sectional view in Figure 12d, showing the places for the microphones and the structure overall.



(a) CAD drawing of a cylinder block and the top.

(c) The cylinder used.

(b) CAD drawing of multiple stacked cylinder blocks.

(d) Cross-section view of the structure.

Figure 12: Illustrations of the arrays

4.1.3 Baffled Circular Microphone Array

A circular microphone array with eight microphones that is mounted around a cylinder surface, referred as baffled circular microphone array [35], is the third array to be measured. This array also has microphones in a circle with 83 *mm* diameter, but the structure they are attached to has a diameter of 90 *mm*. The structure used is the same one as with the previous array (see figure 12). The difference with the previous array is that while it had the microphones attached to the top, this time the microphones are around the cylinder surface.

With this particular array and the way the structures are constructed, the effect of the height of the cylinder can be tested. A single cylinder is $42\,mm$ high, so by stacking them together, a wide range of different heights can be achieved. As a rule of thumb, if the cylinder is 2.8 times higher than the radius of the cylinder, then it closely approximates an infinitely high cylinder [35]. Thus in this thesis, the default configuration consists of 5 stacked cylinders, making the whole structure roughly $210\,mm$ high, which corresponds to $46.7 \times 45\,0m$, so it should approximate an infinitely high cylinder.

These cylinders are then placed on top of a similarly sized cylinder which is designed to enable attaching the stack of cylinders into a holder. The holes for the holders are in the centre of the cylinder and at the bottom edge, so that the whole array system can be attached either vertically or horizontally. The bottom part is shown in Figure 13.

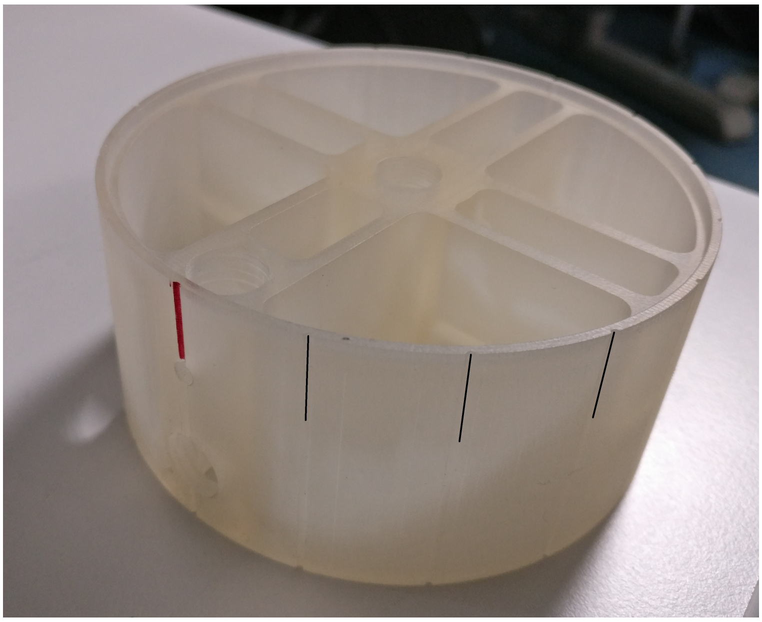


Figure 13: A photo of the bottom part of the stack, which is used to attach the cylinder to the turntable. 22.5° markings are visible as are the holes to which the holders are attached.

4.2 Measurement Methods

4.2.1 Measurements in Free-field

The performance of the above described arrays is measured in an anechoic chamber. In brief, the idea is to measure impulse responses of arrays' microphones at different angles in relation to the sound source. From the impulse responses, the parameters of a single microphone can be analyzed, and by beamforming the impulse responses, directivities and other measures of the arrays as a system can be found.

The basic setup consists of a computer, a soundcard, an amplifier, a loudspeaker, a microphone array under test and the anechoic chamber. In this study, the computer was connected to APx525 Audio Analyzer [36], which was used to generate the test signals. The USB soundcard used for microphone signal acquisition is RME Fireface UFX [37], the amplifier is Lab.Gruppen fP 2400Q [38] and the speaker is a full-range loudspeaker with 4.5" element, specifically Css FR125SR [39]. The volume of the loudspeaker's cabinet is $5\,dm^3$ ($5\,liters$). This setup is shown in Figure 14.

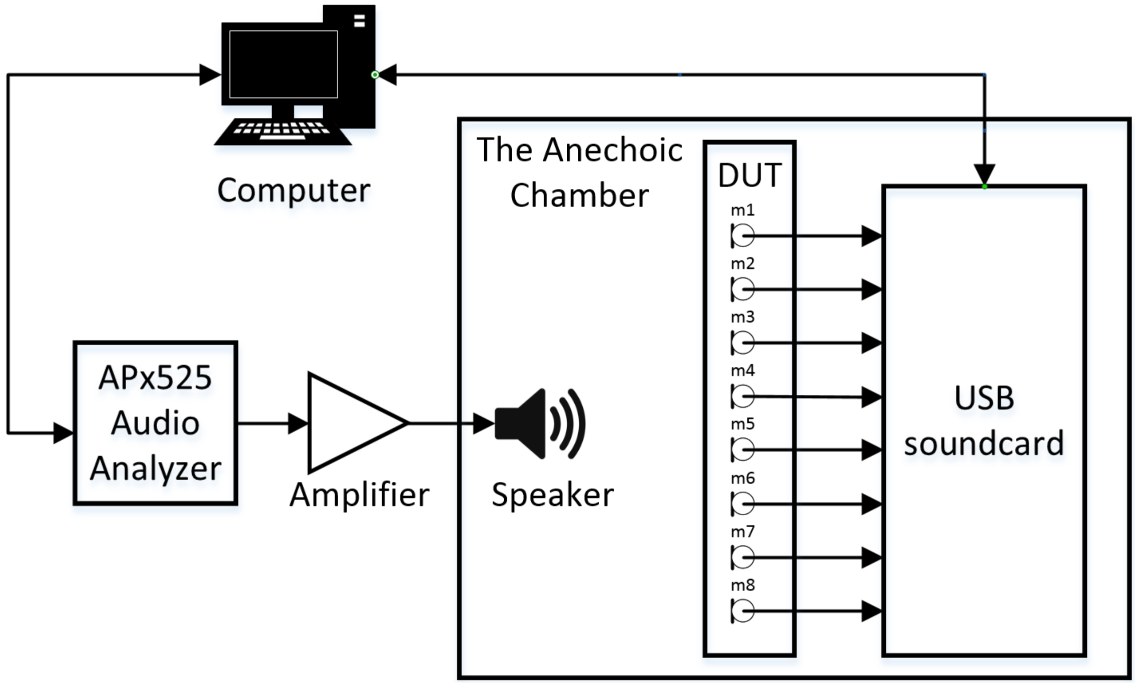


Figure 14: A block diagram of the measurement setup used in this thesis.

The anechoic chamber is located inside an office building, which limits the dimensions of the chamber. As the dimensions are limited, so are the frequency range, volume and measurement area within which the free field criterion can be fulfilled. The inner dimensions of the anechoic chamber's rigid outer shell are $3.5 \times 3.7 \times 2.8\,m\,(L \times W \times H)$, while the dimensions between the wedges' tips are $2.25 \times 2.75 \times 1.5\,m$. These measures limit the free field conditions to frequencies above $200\,Hz$, which for speech audio is sufficient. Overall, the anechoic chamber meets the requirements of ISO-3745:2003 [40] and partially the recommendations of 3GPP TS 26.132 Release 10 (2011-09) [41].

In the test setup, the Device Under Test (DUT) is placed to the middle of the anechoic chamber. The loudspeaker that generates the test signals is in the corner of the chamber, roughly $180\,cm$ away from the DUT. The DUT is attached to a turntable, which rotates the DUT with respect to one axis in steps at user defined step sizes.

For elevation angle of 0°, the idea is to rotate the DUT in 1° steps with respect to the azimuth angle and measure the impulse responses of the microphones at each step. Then the impulse responses of each step are beamformed with a Delay-and-Sum beamformer fixed towards 0°. From the impulse responses, the frequency responses of each step can be calculated and thus the beampattern, directivity etc. of the DUT with the Delay-and-Sum can be analyzed. The impulse response itself is measured using the APx Signal Analyzer using a continuous sine sweep. The measurement is done with 192 kHz sampling rate, and the sweep begins from 50 Hz and ends at 20 kHz. The high sample rate and large frequency range are used to ensure that all the possible details are captured, and to prevent measurement artefacts within the frequency range of interest caused by beginning and ending of the sweep. A photo of an array in the measurement system is shown in the Figure 15.

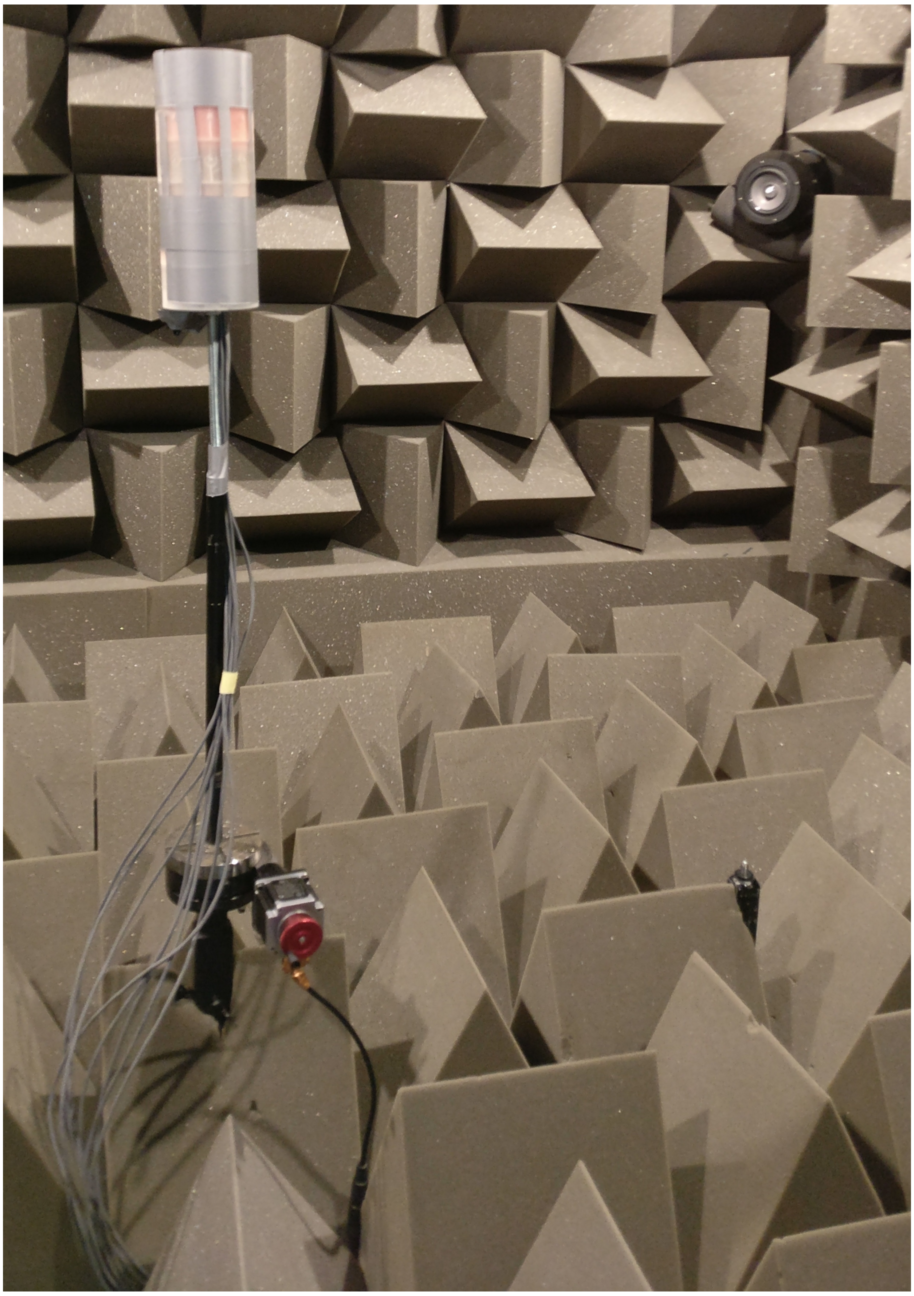


Figure 15: A photo of the baffled circular array in the measurement system.

For other elevation angles, the setup is slightly different. The setup is shown in the

Figure 16 and, as can be noticed, the DUT now rotates so that the turntable's steps correspond to different elevation angles, while the azimuth angle remains constant. When the DUT is measured at different elevation angles, it is rotated along the azimuth angle by hand, and the test is repeated until the whole circle is measured. For this measurement, the step size for elevation angle is 5° and the azimuth angle's step size is 11.25°. These step sizes were chosen so that the amount of measurements remains sane, but that the resolution is good enough. For azimuth angle, the odd step size was used because it was decided that the measurements steps should be fractional of the microphones' intervals - ie. of 45°. As the azimuth angle was changed by hand, the placement error with smaller step size would have been more pronounced, and because the bottom part of the DUT had markings at 22.5° steps, it was natural to use half of that. The step-size markings and overall the bottom part of the cylinder which is used to attach the arrays to the turntable can be seen in the Figure 13.

These assessments thus allows us to measure accurate beampatterns and other metrics of the arrays at elevation of 0° and also 3D beampatterns and other measures with reduced accuracy.

In this test setup, the effect of the loudspeaker is removed by first measuring the loudspeaker's impulse response with a measurement microphone, in this case B&K Type 4190-L-001 half-inch Free-field Microphone [42], at the centre of the chamber. As the measurement microphone is low-noise and high quality with flat frequency response, the resulting impulse response can be removed from the arrays' microphones' impulse responses with a simple convolution in time domain, or in the frequency domain with a division. In addition, the USB soundcard is within the anechoic chamber, hidden beneath the wedges, to minimize the length of the analog signal cables from the array to the soundcard. As the soundcard is within the anechoic chamber, the cables are slightly over $1\,m$ long, whereas if the soundcard would be outside of the chamber, the cables would need to be at least three meters long, which would increase noise levels.

4.2.2 Measurements in Diffuse-field

To compare the performance of Superdirective and Adaptive beamformers with these arrays, the arrays are also measured in a control room. In the control room, a diffuse noise field is generated with multiple loudspeakers playing noise around the room while one speaker is used to play the test signal. The fundamental idea is then to record the resulting sound field with the arrays, and using different beamformers to see how the performance changes. The beamformers are pointed towards the test signal source, and measures such as SNR, Word Error Rate (WER) and Beampattern can be used to describe the differences between the arrays and algorithms. In this particular case, the loudspeakers used to generate the soundfield are eight Genelec 8240 APM -loudspeakers. These loudspeakers are pointed towards the walls and corners of the room to increase the diffuseness of the resulting sound field. In addition, they are at different heights, such that the few are slightly lower than the array, few are at roughly the same level, while some are higher than the array, to increase the

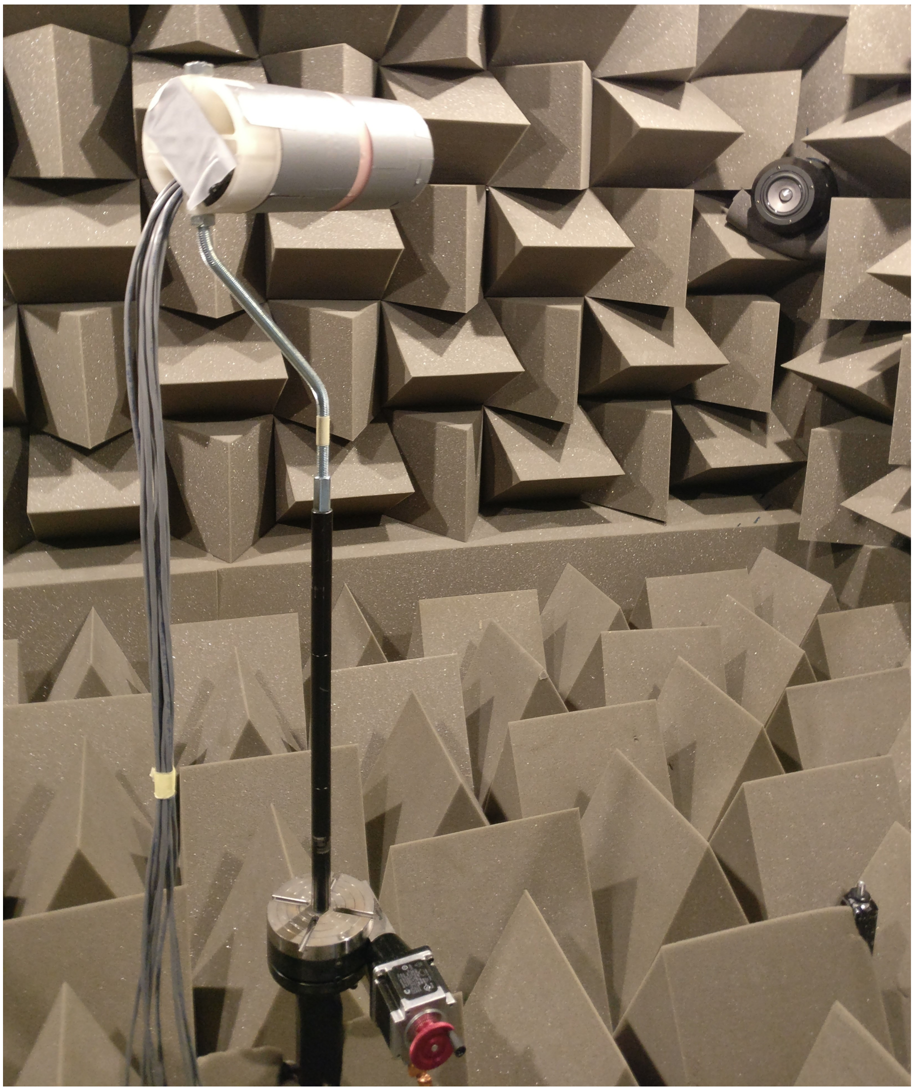


Figure 16: A photo of the baffled circular array in the measurement system when measuring the responses at different elevation angles.

diffusiveness of the field. The test signal itself is generated with Brüel & Kjær Head and Torso Simulator Type 4128C (HATS) [43], such that the directivity and other characteristics of the sound source resemble those of a real human speaker. The HATS is placed $1\,m$ away from the array with the mouth slightly higher than the top of the array. The basic idea of this setup is shown in Figure 17. As can be noticed by comparing the figures 14 and 17, the main differences between the setups are the amount of loudspeakers and that the DUT is in the control room instead of the anechoic chamber.

Superdirective and Adaptive beamformers can't be measured in the anechoic chamber because either they are designed to remove noise, in the case of superdirective, or to adapt to the noise, in the case of adaptive beamformers, and it is practically impossible to produce a diffuse noise field in an anechoic chamber. If the Superdirective beamformer is measured in an anechoic chamber as described previously, the result will be practically the same as for a Delay-and-Sum beamformer.

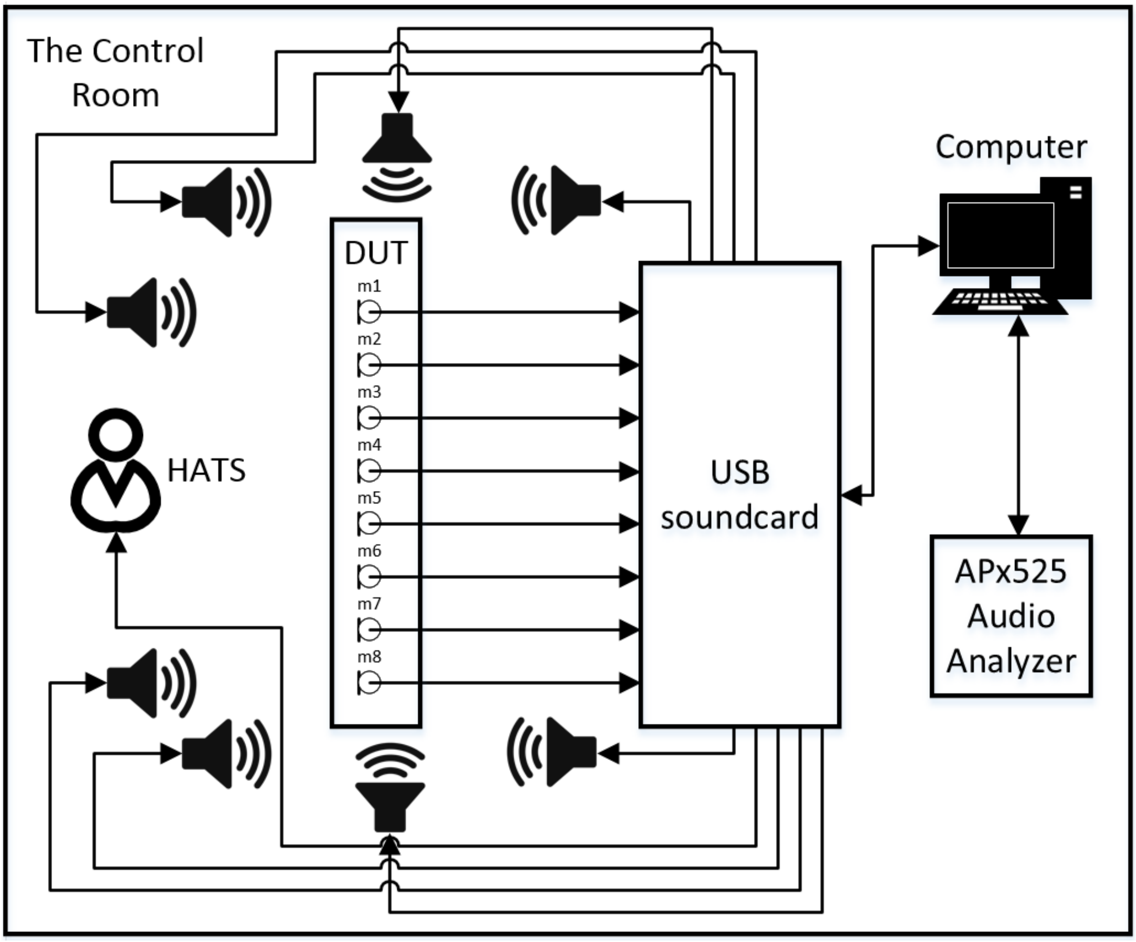


Figure 17: A block diagram of the measurement setup used in this thesis for diffuse field measurements.

The control room is acoustically treated and has a floor area of $22\,m^2$. The room has absorbing material in the ceiling, and heavy-weight absorption curtains and diffusers on the walls. The room follows the recommendations of ITU-R BS.1116-1 [44] where applicable. It's reverberation time is within the tolerance limit times set by ITU-R BS.1116-1, meaning in short that for frequencies over $160\,Hz$, the reverberation time T_{20} is between $0,15\,s$ and $0,25\,s$. The room has different kinds of equipment in it, such as tables, computers and monitors, oscilloscopes, a rack with amplifiers and other devices. The background noise level of the room is approximately $L_{Aeq} = 29dB$, which is similar to an average quiet living-room.

For the diffuse field measurements, the speech signal was repeated five times with different background noises. Once with only the room's background noise, second time with the Genelecs playing white noise at equal levels, so that the Sound Pressure Level (SPL) at the array location is $40\,dB(A)$, third time with white noise resulting in $50\,dB(A)$ SPL at the array location. Fourth and fifth time the background noise used was a babbling noises recorded at cafe, obtained from DEMAND: Diverse

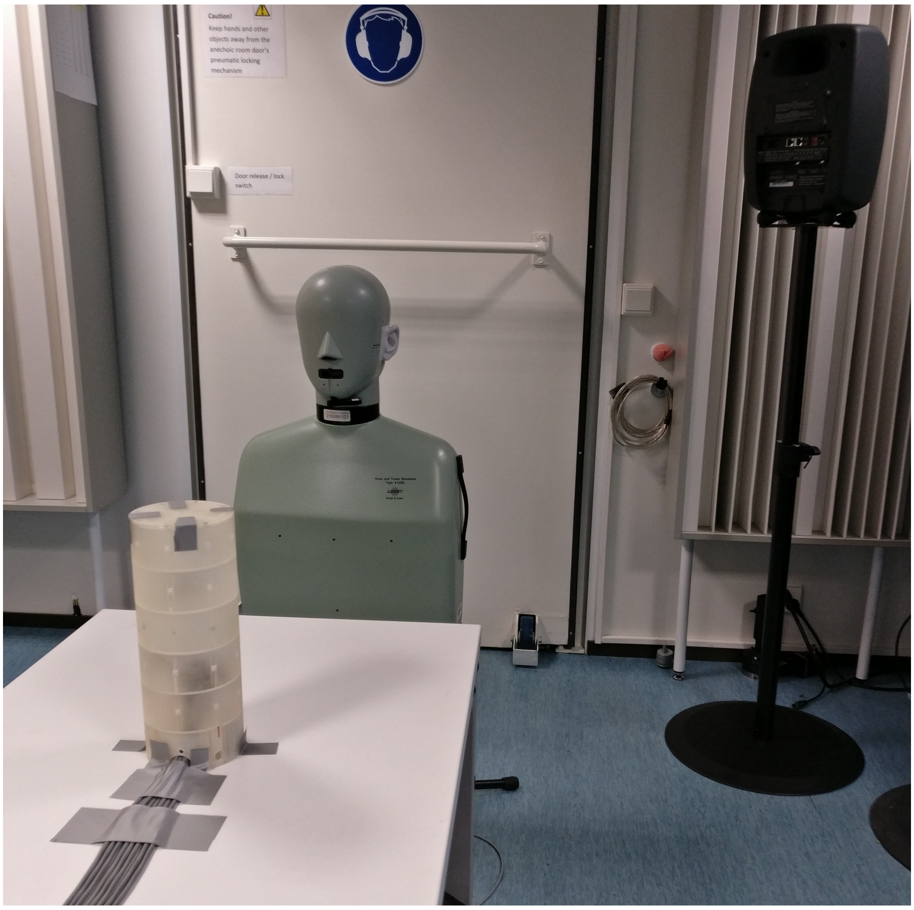


Figure 18: A photo showing part of the measurement setup: the array, the Head and Torso Simulator and a loudspeaker used to generate the noise field.

Environments Multichannel Acoustic Noise Database [45]. The babbling noises were set so that their recorded digital long-term-integrated signal level measured in Loudness Unit relative to Full Scale (LUFS) according to EBU R128 [46] and ITU-R BS.1770-4 [47] were $10\,LUFS$ and $20\,LUFS$ lower than the speech signal's loudness. The idea of using different background noises is to compare how the ASR behaves with different kinds of noises, and if certain beamformers perform better with certain kind of noise.

The sound source was equalized in the anechoic chamber, so that the frequency response of the HATS was flat at the measurement point specified in the ITU-T P.58 [48] standard - ie. $25\,mm$ at the front of the mouth. The level of the HATS was calibrated so that the HATS was placed at the measurement setup's source location and a B&K Type 4190-L-001 microphone placed at the spot of the array. A calibration signal of $94\,dB$ is recorded with the microphone, and then a speech sample. The speech sample's level is measured so that it's loudness is calculated according to the ITU-R BS.1770-4, and then compared to a calibrator signal, ie. $94\,dB\,@\,1\,kHz$, and the loudness is set to be $-34\,LUFS$ compared to the calibrator signal. Thus the speech's level is roughly $60\,dB$, which corresponds with speech levels of normal conversations [49, 50, 51]. This leads to the measurement having three different SNRs acoustically: the first, with no background noise, having SNR of approximately $30\,dB$, the first white-noise and babbling noise measurements having roughly $20\,dB$ and the second white-noise and babbling noise measurements with around $10\,dB$. The spectra of the different noises are shown in Figure 19. The three spectra labeled as "Silent room", "WN Low" and "WN High" are measured with B&K measurement microphone, while the three others are measured with the microphone 1 of the reference array. The effect of the microphone 1's frequency response can be seen especially at high frequencies where the spectra differ quite clearly. Also it should be noted that due to the self-noise of the microphones, the microphone 1's noise level for the silent room is clearly higher than for the measurement microphone, and thus the SNR for the silent room isn't actually that $30\,dB$.

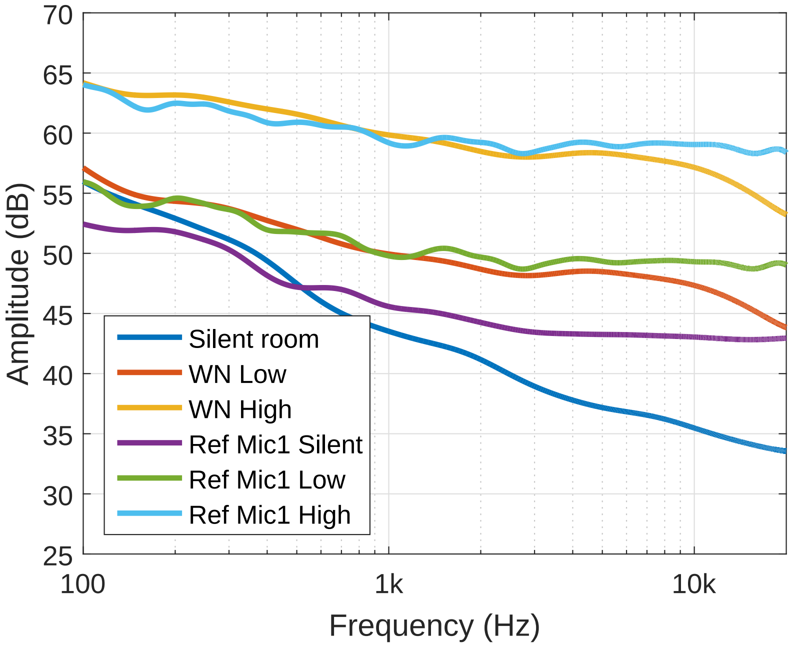


Figure 19: Spectra of the different noises as captured by a B&K microphone and the microphone 1 of the reference array. The spectra are $1/3rd$ octave smoothed.

The speech signal itself is constructed using speech samples from Intel's library. The different speech samples are level calibrated so that they all have equal loudness according to EBU R128 and ITU-R BS.1770-4. The speech signal has 80 samples in it, each about sentence or two long, resulting in 1000 words. This speech signal is played through the HATS and recorded with the different arrays. The recordings are then beamformed with the different algorithms, which result in a total of 45 recordings: three arrays, with three different beamformers, with five different background noises. For the beamforming, a constant steering direction is used, so that each beamformer points towards 0° - ie. the sound source - as the source doesn't move and as to have really apples-to-apples comparison between the beamformers. These recordings are then passed through Intel's own ASR-system, whose output can then be compared to the transcriptions of the speech signal using sclite-tool [52]. Thus the WER for different arrays, beamformers and noises can be derived and compared.

4.3 Compensation of the Structure's Effect on Beamformer

The three arrays described in the section 4.1 have different kinds of structures, and each of those structures will affect the array and beamformers differently. In this section, a brief explanation of the differences - or assumed differences - between the arrays are given.

To begin with, the first array is the reference array, which will act as the anchor to which compare the two other structures. As the microphones practically just hover in the air there is no shadowing caused by any structure, and the delays between the microphones is assumed to only dependant on the incoming angle of the sound. In this array, the microphones are still omnidirectional.

For the second array, the one with the microphones on the top of the cylinder, the delays between the microphones are assumed to be the same as for the reference array. They might differ in the case of negative elevation angles, as then there will be some shadowing caused by the structure, but for elevation angle of zero or above, the delays should be the same. The major difference is the directivity and SNR of the microphones themselves: as they are attached to the structure their directivity pattern isn't omnidirectional any longer and as they are in half-space, their SNR should be better than reference array's. The directivity of the microphones should still be roughly identical, as they are all attached to the same plane.

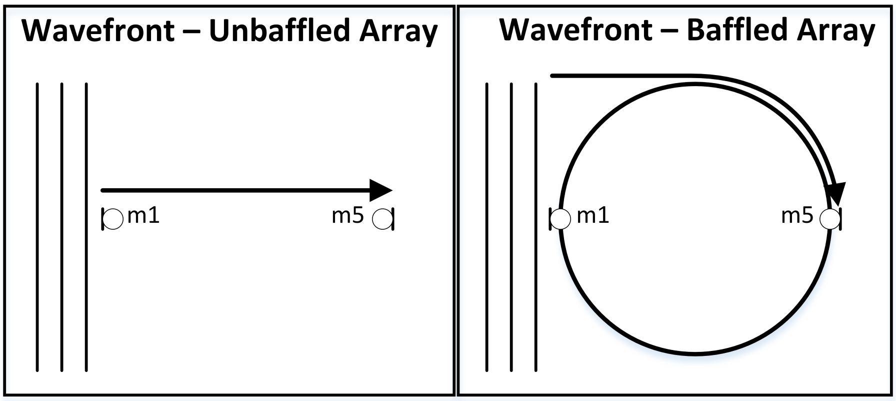


Figure 20: Sound's route to microphones with unbaffled and baffled circular arrays.

Lastly, the baffled circular array should differ quite a lot from the reference array: Firstly, the delays between the microphones are assumed to be bigger for a signal with same DoA. This is due to the fact that the radius of this array's outer structure is 90 mm, and it alone should increase the delays slightly. In addition to that, the fact that there is structure between the microphones should increase the delays, making the array appear larger. This increase in delay is assumed to be the such that for a sound with DoA of 0°, the delay between microphones 1 and 5 (microphone 1 being the one closest to the source, and 5 being the one on the opposite side of the cylinder) is proportional to $\frac{\pi r}{2} - r$, ie. to the distance difference between the radius and one quarter of the circumference of the cylinder. The reasoning behind is illustrated in the Figure 20. In short, instead of the sound travelling straight from microphone 1 to microphone 5, the wavefront travels to the edge of the cylinder, from where it travels along the edge of the cylinder to the microphone 5.

Secondly, again the microphones themselves aren't omnidirectional. In addition to this, all microphones have different directivities: the patterns should be identical but to different directions. So for example, the microphone 1 should be sensitive to sounds coming from azimuth angle of zero, the microphone 5 should be less sensitive to those same sounds, as the structure between the source and the microphone attenuates the sound. Lastly, the SNR for this array should also be better, as the mechanical integration of the microphones means that they are not in the free-field any longer.

These differences between the arrays mean that their performances differ, but also that at different metrics different arrays can be better than the others.

4.4 Electronics

For the described structures to do anything, they need various electrical components. The most important component is naturally the microphones themselves. For this thesis Knowles' SPH0642HT5H-1 MEMS-microphones were used. They are top-port microphones with matched sensitivity, low noise, flat frequency response, small packaging and other desirable features [32]. They are analog microphones, and their sensitivity is $-38 \pm 1\, dBV/Pa$, SNR is $65\, dB(A)$ and THD $0.25\,\%$ according to the datasheet. The microphones need three connectors; one for supply voltage, one for the output signal and one for ground. The mechanical dimensions of the microphones are $3.5 \times 2.65 \times 1\, mm\, (L \times W \times H)$. [32] Picture of the microphone can be seen in Figure 21.

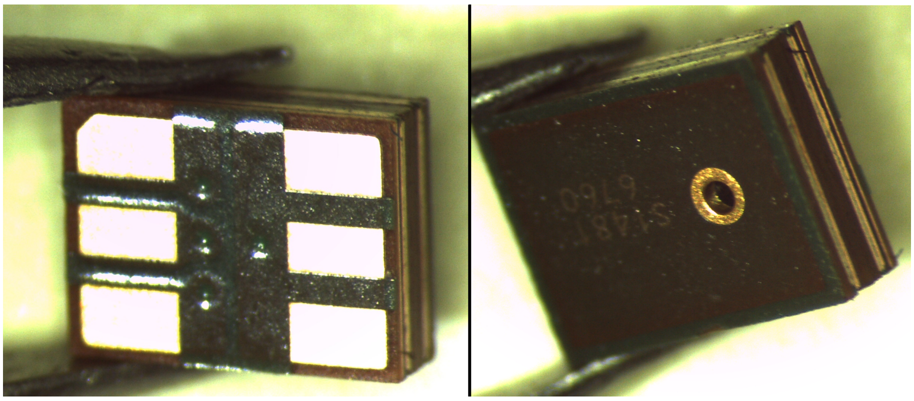


Figure 21: A picture taken of the microphone used in the arrays. On the left, the bottom of the microphone and on the right the top view of the microphone.

In addition to the microphone, the arrays need a power source for the microphones, which in this case is two AA (IEC R6) batteries in series, producing a voltage source of $3\,VDC$. The microphones use a voltage of $1.5...3.6\,VDC$. To make sure that the voltage supply is stable, it is filtered with a capacitor of $1\,mF$ between the supply and the ground. Further, to reduce noise in the output, the microphones' output is transformed into a pseudo-differential output with a simple electronic circuit. This is done by placing a capacitor into the signal path of the microphone's output, and then placing a similar capacitor between the power supply and a second output pin. This is illustrated in Figure 22. In the figure, the output pins T and R form a pseudo-differential output, meaning that both should have similar noise characteristics, coming from the capacitor and from the cable attached to the output. When the device the microphone is connected to has a balanced input, the pseudo-differential output can be used to reduce the noise induced in the cable and from the capacitor.

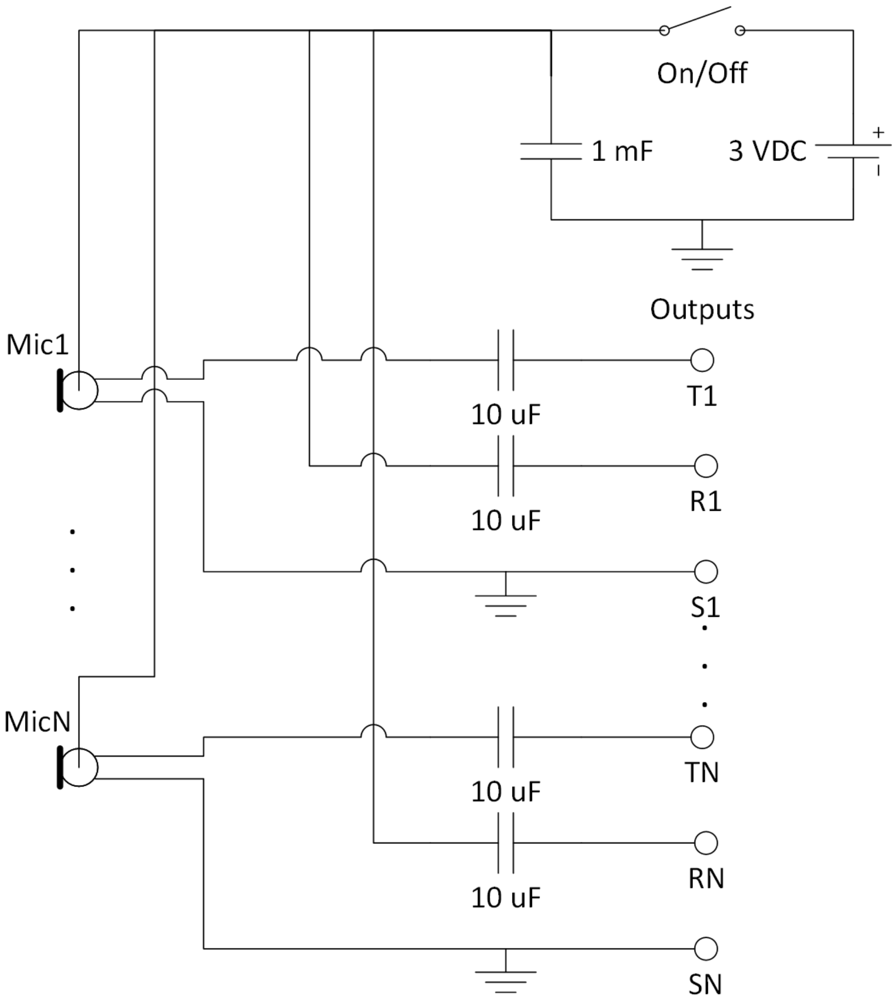


Figure 22: Schematic for the microphone and its connections.

When considering this pseudo-differential generating circuit, it can be seen that when the output is connected to a device with input impedance R, it forms a high-pass filter. This high-pass filter has a corner frequency f_c which can be calculated as

$$f_c = \frac{1}{2\pi RC} \tag{37}$$

In this particular case, the C is $10\,\mu F$ and R is the input impedance of the attached sound card, which for the RME Fireface UFX used in these measurements is $10\,k\Omega$, according to the datasheet [37]. Thus the corner frequency is $\sim 1.59\,Hz$, which is important for the behaviour of the beamformer, because this kind of first-order high-pass filter has a phase shift of 0° when the frequency is decade below the corner frequency, and of 180° when the frequency is a decade above the corner frequency. For the beamformer to work correctly, the phase response of the microphones should be as close to each other as possible. Because each microphones has its own filter, even small changes in the component values could affect the phase responses by a noticeable amount. Thus, when the corner frequency is around $1.59\,Hz$, the phase shift is a constant 180° for frequencies above $15.9\,Hz$. As we are mostly interested in frequencies over $200\,Hz$, it means that the phase-shift caused by this high-pass filter is not an issue.

5 Results

5.1 Microphone and Array parameters

5.1.1 Delays Between the Microphones

From the measurements done in the anechoic chamber the delays between different microphones were extracted. The delays were found out by taking cross-correlation between two microphones and finding the maxima of the results. This corresponds to the delay in samples between the microphones, which can be converted to time delays and distance between the microphones as the sample rate and speed of sound are known. In the tables below, the delays when the sound arrives from 0°, ie. from the direction of microphone 1, are shown for each array. For these measurements, the sampling rate was $192\,kHz$ and the impulse responses were oversampled by factor of 10 using piecewise cubic interpolation to increase the resolution of the results.

By comparing the Tables 1 and 2, it can be noticed that the delays of the reference and top array are basically the same. Some microphones have small delay differences, but that is most likely due to the placement error of the reference array, and because the responses were measured at 1° intervals, it is possible that there is slight misplacement in the angle. Placement error in the microphone locations can be noticed by comparing the delays between the microphones: if the microphones are located ideally, the delays for microphone pairs of 4 and 6, 3 and 7, and 2 and 8 should be identical, as can be seen in the case of theoretical delays for $83\,mm$ circular array, given in Table 4. It can be noticed that for the reference and top arrays, they are very close to being identical, but there are slight differences of less than 1 sample. As one sample corresponds to $\sim 1.8\,mm$ in distance,as the sample rate is $192\,kHz$, the difference in locations are under $\sim 1.8\,mm$. In short, it can be concluded that the reference array and the array on the top of the cylinder are similarly sized acoustically.

The baffled array, on the other hand, differs quite a lot from the two other arrays, as can be noticed by comparing delays in Table 3 to the delays of the two other arrays. As explained in the section 4.3, this array was assumed to be acoustically bigger than the others, and according to the delays that is the case. While for the two other arrays, the maximum size was 45.5 samples, corresponding to $81.9 \pm 0.09\,mm$ for the baffled array, the biggest delay was 68.0 samples, meaning $122.1 \pm 0.09\,mm$. This makes the baffled array roughly 50% bigger acoustically than the two other arrays. When comparing this with the assumption that the size increase is proportional to $\frac{\pi r}{2} - r$, it appears that the difference is bigger: that equation equals $25.69\,mm$, meaning that an array with diameter of $90\,mm$ would appear $115.69\,mm$ which is still some $6\,mm$ smaller. Another interesting note is that the baffled array doesn't appear to be equally spaced: for the two other arrays the delays between microphones 2 and 3, 3 and 4, 1 and 3, and 3 and 5 were equal, for the baffled array they are not. For the baffled array the distance between microphones 3 and 4, and 3 and 5 is clearly larger than that between microphones 2 and 3, and 1 and 3. This is something that is related to the phenomenon referred in the Figure 20.

For the remaining results, for Delay-and-Sum beamformer the delays were based

Table 1: Delays between the microphones of the reference array, in samples.

Mic	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	0.0	-6.8	-22.7	-38.5	-45.5	-38.5	-22.4	-6.8
2	6.8	0.0	-15.8	-31.7	-38.7	-31.6	-15.6	0.0
3	22.7	15.8	0.0	-15.8	-22.8	-15.8	0.2	15.9
4	38.5	31.7	15.8	0.0	-7.0	0.0	16.1	31.7
5	45.5	38.7	22.8	7.0	0.0	7.0	23.1	38.7
6	38.5	31.6	15.8	0.0	-7.0	0.0	16.1	31.7
7	22.4	15.6	-0.2	-16.1	-23.1	-16.1	0.0	15.6
8	6.8	0.0	-15.9	-31.7	-38.7	-31.7	-15.6	0.0

Table 2: Delays between the microphones of the array on the top of the cylinder, in samples.

Mic	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	0.0	-6.5	-23.2	-38.8	-45.5	-38.8	-23.1	-7.1
2	6.5	0.0	-16.6	-32.2	-39.0	-32.3	-16.6	-0.5
3	23.2	16.6	0.0	-15.6	-22.3	-15.6	0.0	16.1
4	38.8	32.2	15.6	0.0	-6.8	0.0	15.6	31.7
5	45.5	39.0	22.3	6.8	0.0	6.7	22.4	38.4
6	38.8	32.3	15.6	0.0	-6.7	0.0	15.7	31.7
7	23.1	16.6	0.0	-15.6	-22.4	-15.7	0.0	16.0
8	7.1	0.5	-16.1	-31.7	-38.4	-31.7	-16.0	0.0

Table 3: Delays between the microphones of the baffled array, in samples.

Mic	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	0.0	-8.0	-25.1	-46.7	-68.0	-46.9	-26.6	-7.6
2	8.0	0.0	-17.1	-38.6	-60.0	-38.9	-18.7	0.3
3	25.1	17.1	0.0	-21.5	-42.9	-21.7	-1.6	17.5
4	46.7	38.6	21.5	0.0	-21.3	-0.1	19.9	39.0
5	68.0	60.0	42.9	21.3	0.0	21.1	41.4	60.4
6	46.9	38.9	21.7	0.1	-21.1	0.0	20.2	39.2
7	26.6	18.7	1.6	-19.9	-41.4	-20.2	0.0	19.0
8	7.6	-0.3	-17.5	-39.0	-60.4	-39.2	-19.0	0.0

on these results: each microphone's signal was delayed by the corresponding delay for the correct array. This results in pretty ideal Delay-and-Sum beamformer, as the delays are based on the measurements instead of calculations based on theoretical locations of the microphones.

Table 4: Theoretical delays for 83 mm circular array.

Mic	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	0.0	-6.8	-23.1	-39.5	-46.2	-39.5	-23.1	-6.8
2	6.8	0.0	-16.3	-32.7	-39.5	-32.7	-16.3	0.0
3	23.1	16.3	0.0	-16.3	-23.1	-16.3	0.0	16.3
4	39.5	32.7	16.3	0.0	-6.8	0.0	16.3	32.7
5	46.2	39.5	23.1	6.8	0.0	6.8	23.1	39.5
6	39.5	32.7	16.3	0.0	-6.8	0.0	16.3	32.7
7	23.1	16.3	0.0	-16.3	-23.1	-16.3	0.0	16.3
8	6.8	0.0	-16.3	-32.7	-39.5	-32.7	-16.3	0.0

5.1.2 Acoustic Locations of the Microphones

Based on the delays introduced in the previous section the acoustic locations of the microphones were derived. By finding the delays with different DoAs, the distances between the microphones along the x-axis at different angles can be found. The basic idea was to find the delays for DoAs between 0° and 45°, as then the microphone 1 has moved to the location of microphone 2 etc. The delays from different DoAs were normalized so that the origin is in the middle of the array. The microphones' locations along the y-axis were found by using the measurements with DoA of 90° and $-90°$ compared to the one from which the x-axis delays were found. The idea was that by combining the x-axis and y-axis delays, the result would be the location of the microphone at Cartesian coordinates with certain DoA. These results are shown in the Figures 23-25. In those figures, the sound arrives along the x-axis from $-\infty$, the array is rotated and the location of each microphone is stored. Then those locations are used to plot how each microphone's location changes.

In the Figures 23 and 24 it can be noticed that both the reference and top array, the microphone locations basically behave as circles, as expected. There are small discontinuations between the microphones due to the placement errors. As explained in the 4.1 the placement error is bigger in the reference array, and it is visible in these figures. The Figure 25, on the other hand, shows very interesting behaviour of the baffled array. First of all, it can be noticed that the acoustic locations of the microphones don't behave circle at all, instead they form more oval-like shape. Secondly, what happens behind the array is intriguing, for the shape isn't oval as it has that kind of a bump at the tip. All these shapes are shown in the Figure 26, and the similarities between the reference array and the top array can be noticed, as can be the fact that the baffled array is bigger - 90 mm instead of 83 mm - than the two others.

As discussed previously, the fact that the sound needs to travel along the surface of the cylinder explains the oval shape in some manner, but not fully: especially the

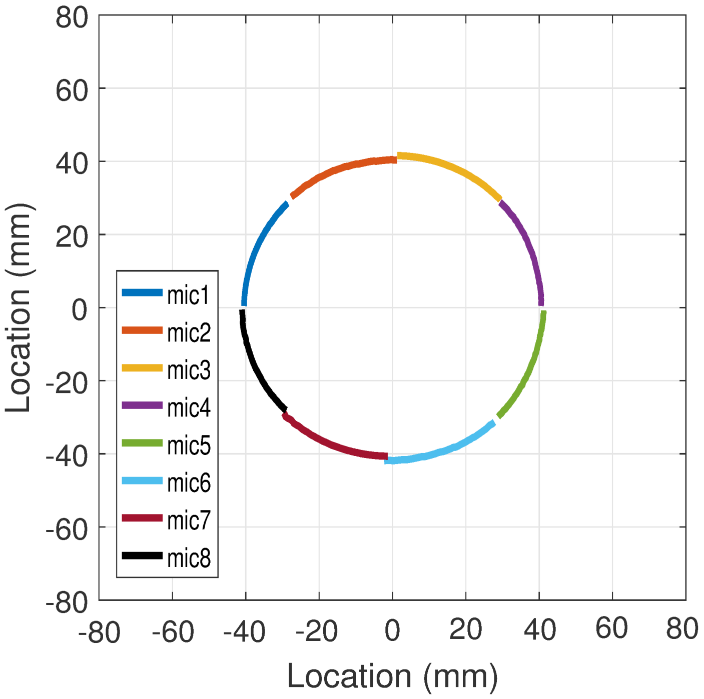


Figure 23: Pattern that the reference array's microphones draw when it is rotated 45°.

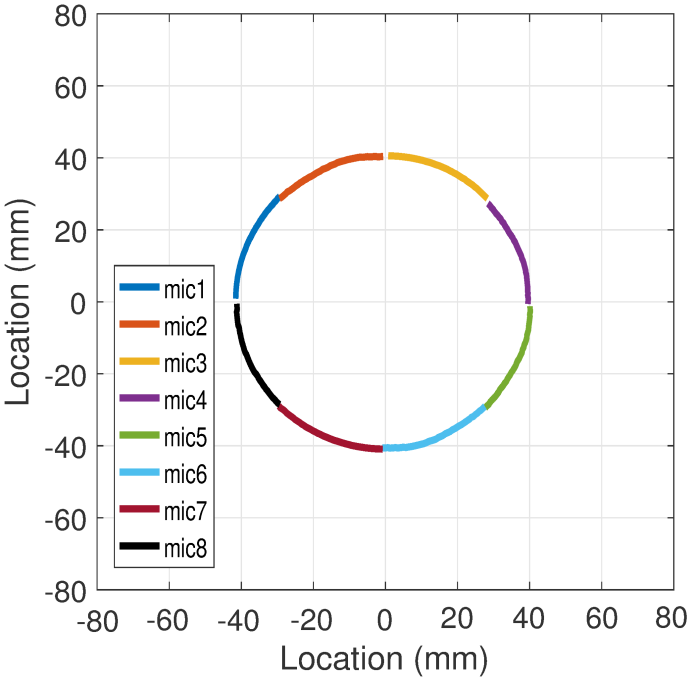


Figure 24: Pattern that the top array's microphones draw when it is rotated 45°.

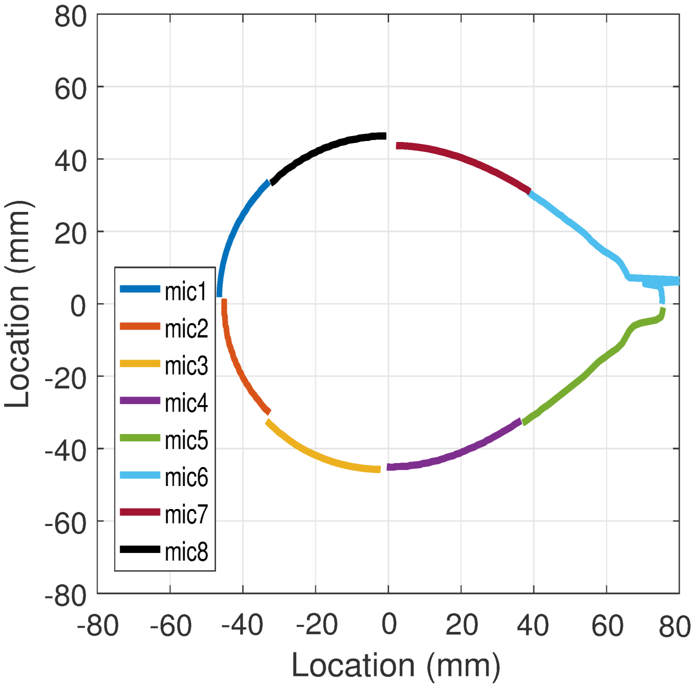


Figure 25: Pattern that the baffled array's microphones draw when it is rotated 45°.

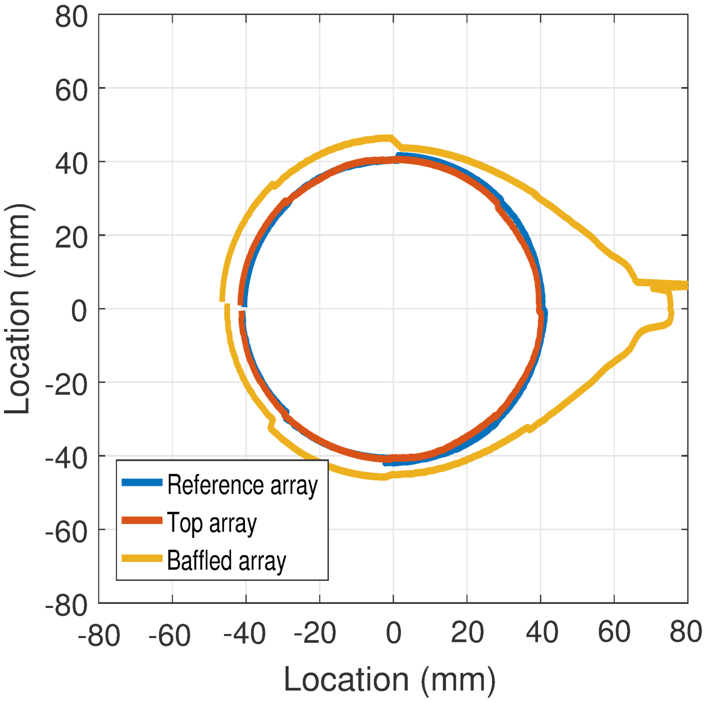


Figure 26: Comparison of the patterns of the arrays.

bump at the end isn't explained by that as is. As the microphones 1 and 5 appear to be further from one another, in the case that the DoA is 0°, than the difference between the distance of the microphones and the distance along the surface of the cylinder suggests that the sound travels slower near the surface of the structure. The slower speed wouldn't explain the bump at the back though. By looking at the figure 25 more closely, it can be noticed that before the tip, the locations become slightly wavy with increasing amplitude towards the tip. This combined with the single outlier in the microphone 6's location suggests that this issue could be caused by shadowing, which causes the impulse responses of the microphones at the front and at the back to differ.

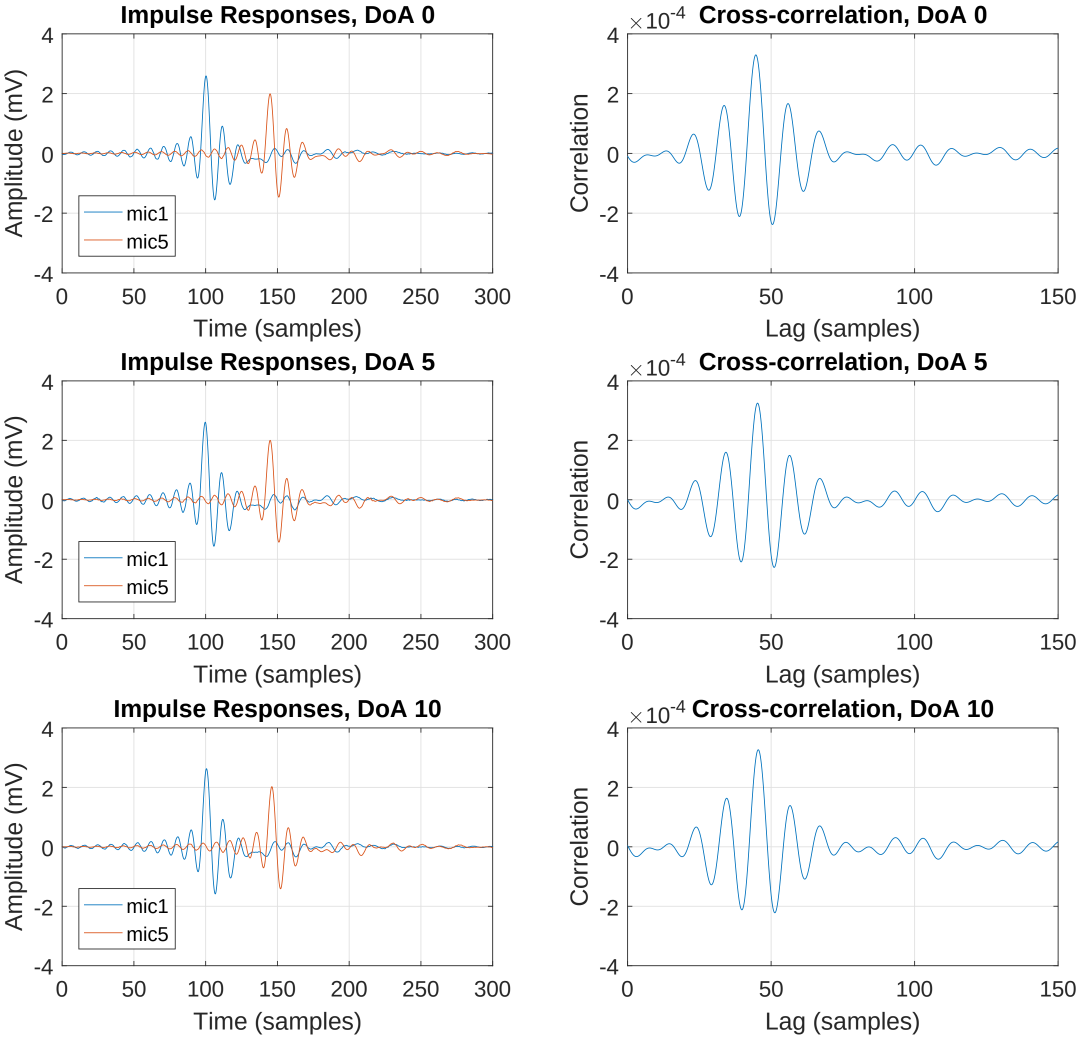


Figure 27: Impulse responses of microphones 1 and 5 and their cross-correlations for the top array.

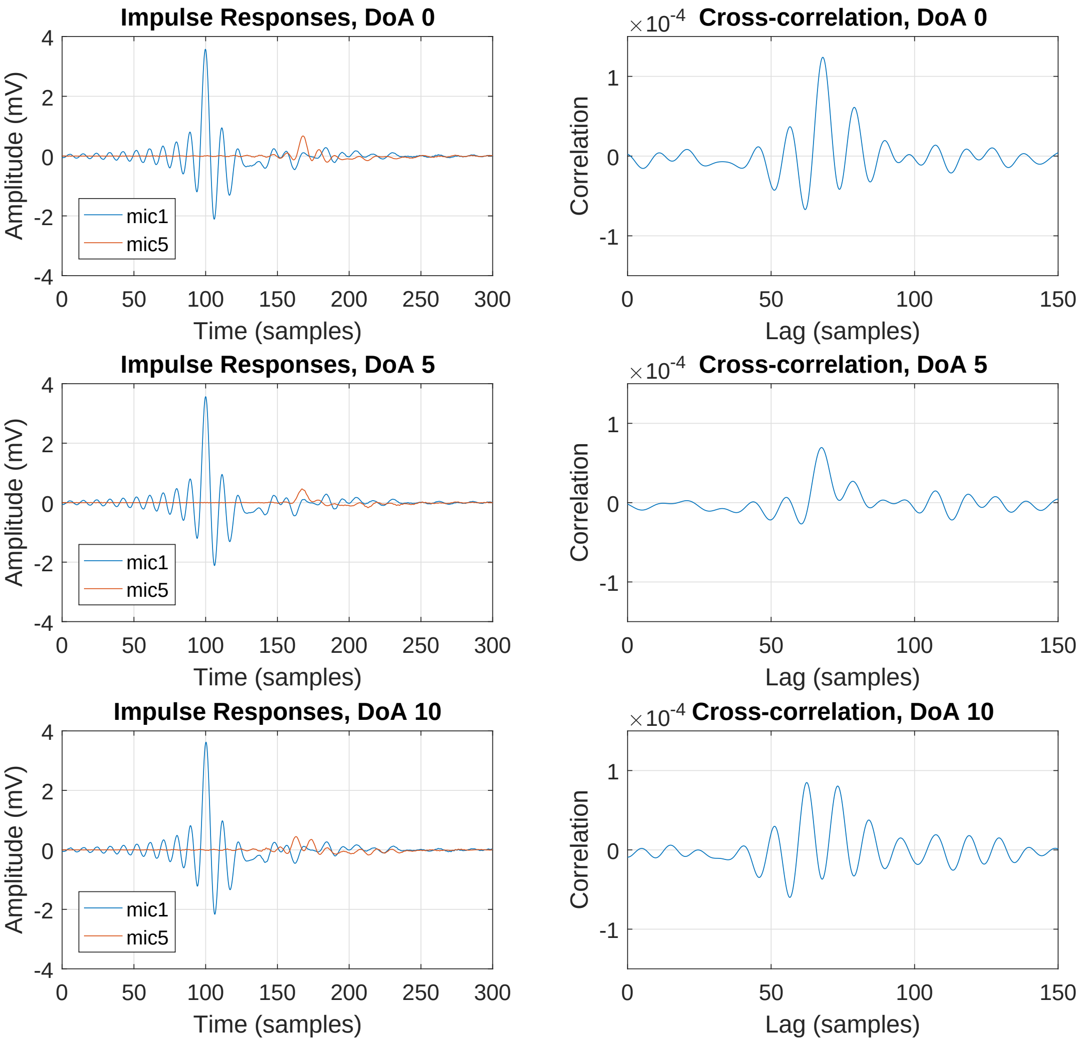


Figure 28: Impulse responses of microphones 1 and 5 and their cross-correlations for the baffled array.

The differing of the impulses and their cross-correlations are compared in Figures 27 and 28. In the Figure 27 the impulses and cross-correlations of the top arrays microphone 1 and 5 with different DoAs are shown, and the Figure 28 shows the impulses and their cross-correlations for the baffled array for the same DoAs. Comparing those two figures, it can be noticed that for the top array, the impulses don't change that much with different DoAs and the impulses of microphone 1 and 5 are quite similar. This results in the cross-correlations for top array to remain practically the same with different DoAs. For the baffeled array, on the other hand, the impulses of the microphone 1 and 5 differ noticeably due to the shadowing. In addition, especially the impulse of microphone 5 changes a lot with different DoAs, meaning that the cross-correlations change with the DoAs, as can be seen in the

Figure 28.

By filtering the impulse responses with a 1/3rd octave filter bank, the acoustic locations of the microphones at different frequency bands can be found. The results are shown in Figure 29, and it can be noticed that for low frequencies - ie. $500 Hz$ 1/3rd octave band - the array looks circular. What happens with higher frequencies is very intriguing: for $1000\,Hz$ the array appears clearly the biggest and it's shape is thickest, and for higher frequencies, the oval-like shape gets sharper and the length of the oval gets smaller. For high frequencies, the bump at the tip of the oval also becomes apparent, further suggesting that the bump is in fact caused by the shadowing which mostly happens at high frequencies.

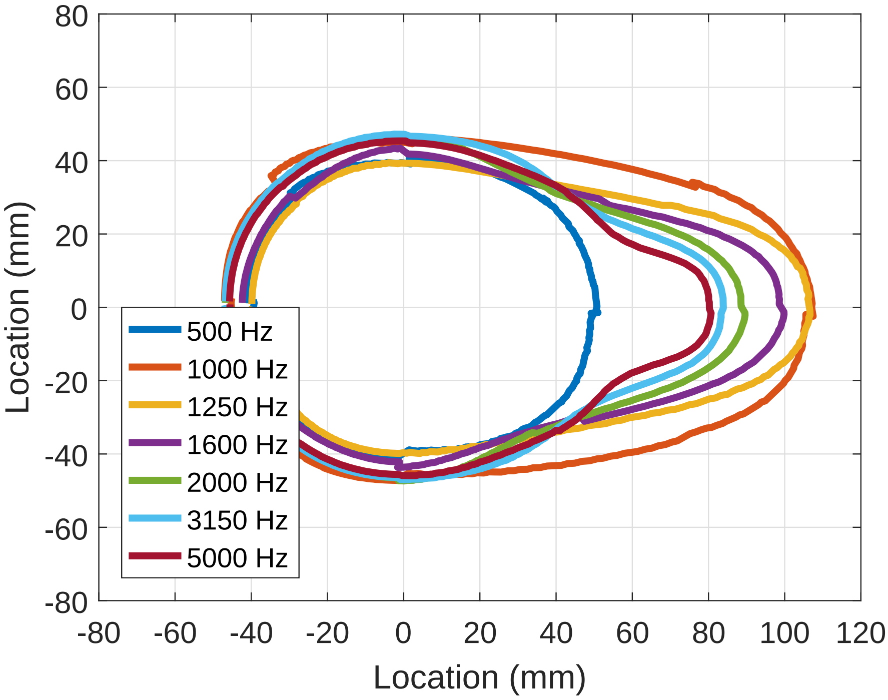


Figure 29: Comparison of the baffled array's mic location patterns for different frequency-bands.

Overall, the microphone's acoustic location's raises intriguing questions for beamformers that require the locations of the microphones when initializing the algorithm. For example, in this thesis the superdirective and the adaptive beamformer use the locations when they are initialized. For such an algorithm, what locations should be used? First of all, the locations are dependant on the DoA, but for the baffled array the locations are not so trivial to figure as for the top and reference arrays. Secondly, in the diffuse-field, should these acoustic locations be used at all, as then the sound arrives from everywhere and the target of the beamformer can be to reduce noise. So how does using these oval-like locations for the baffled array affect the performance of the beamformer in the diffuse field? Lastly, would frequency-based beamforming work better for the baffled array, such that the input is first filtered to frequency-bands and each frequency-band has beamformer with band-dependant microphone locations?

5.1.3 Sensitivity, Frequency Response and Directivity of the Microphones

From the measurements done in the anechoic chamber, the sensitivity for each microphone in each array was found. As the sound source was calibrated to produce $74\,dB$ sound pressure level at $1\,kHz$ - meaning $0.1\,Pascals$ - at the location of the array, the sensitivity can be derived. It should be noted that these sensitivities also include the effect of the structures and electronics, thus for some microphones the sensitivities are outside the range given in the datasheet [32], ie. $37 - 39\,dBV/Pa$. The sensitivities are listed in Table 5

Table 5: Sensitivities of the microphones in different arrays in dbV/Pa.

Mic	1	2	3	4	5	6	7	8	σ
Reference Array	-38.5	-39.1	-38.4	-38.4	-39.4	-38.3	-37.7	-38.3	0.52
Top Array	-38.5	-37.9	-40.1	-39.5	-39.3	-40.3	-39.2	-39.1	0.78
Baffled Array	-38.1	-39.0	-37.6	-38.4	-38.2	-38.1	-38.8	-38.3	0.44

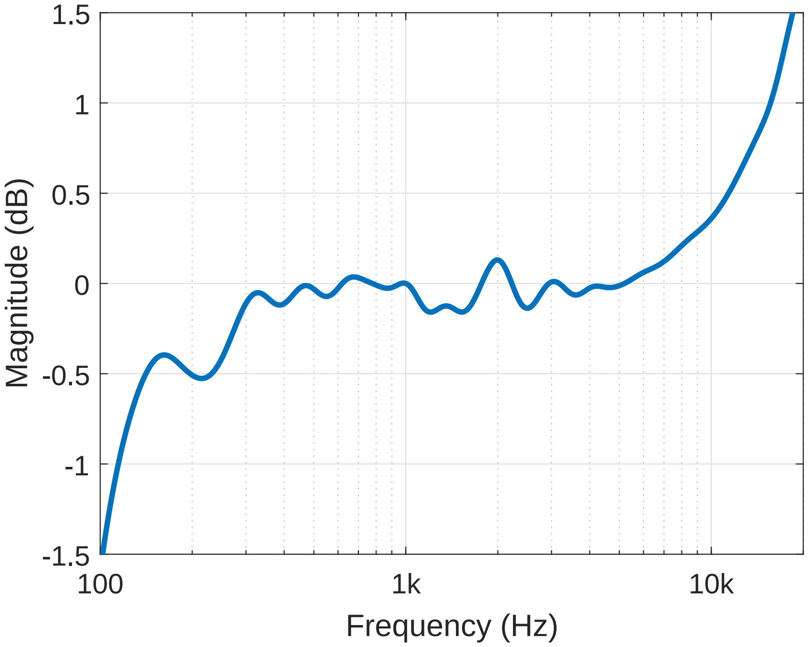


Figure 30: Frequency response of the reference array's microphones averaged and $1/3rd$ octave smoothed.

The average of the frequency responses of all the microphone in the reference array is shown in Figure 30. The sound source's response is removed from the frequency response and the result is $1/3rd$ octave smoothed. It can be noticed that the response fits within $\pm 0.25\,dB$ between $250\,Hz$ and $8\,kHz$. The resonance, which is the peak around $23\,kHz$ - which isn't visible in the plot - can be noticed in the figure as the response's magnitude starts to noticeably increase for frequencies above $8\,kHz$.

Directivity patterns for microphones at each array were found out by plotting the frequency responses of a single microphone at different azimuth angles. These patterns are shown in Figures 31-33. In the figures, it can be noticed that the microphones in the reference array are practically omnidirectional, while the top array's microphone has slight attenuation at angles over 100° or less than $-100°$. As expected, the baffled array's microphones have very clear directivity, as when the structure is between the microphone and the sound source, the structure blocks most of the incoming signal.

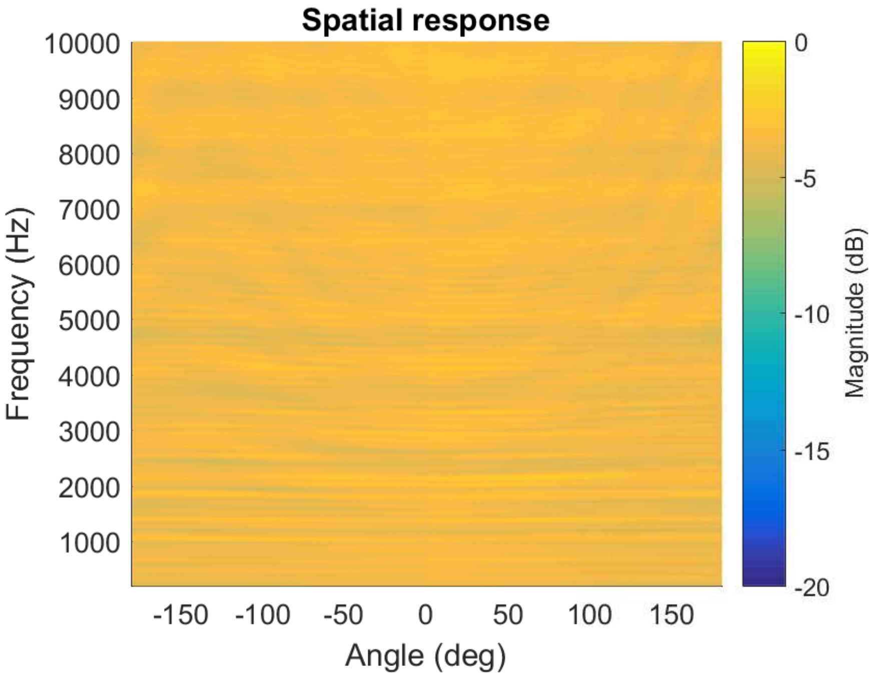


Figure 31: Directivity pattern of the microphone 1 of the reference array.

5.2 Free-field Measurement Results

5.2.1 Spatial Responses

From the impulse responses at different angles, the spatial responses of the different arrays and beamformers can be plotted. Spatial response shows the system's frequency response at different angles, so that the x-axis is the angle, y-axis is the frequency and colour describes the magnitude with that frequency at that angle. The spatial responses in Figure 34 show that the Delay-and-Sum and Superdirective beamformer behave practically the same. It can also be noticed that the reference and the

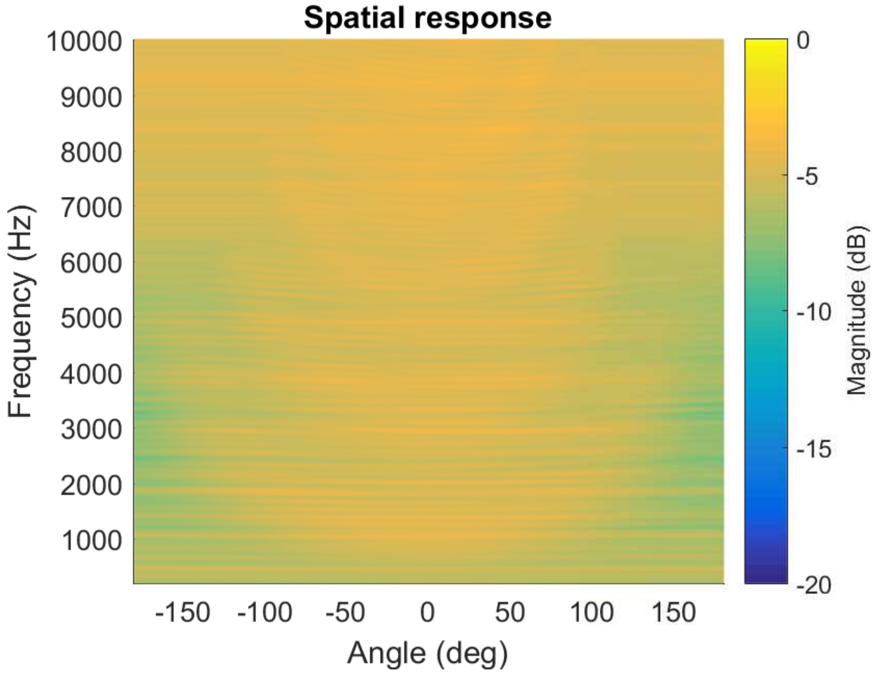


Figure 32: Directivity pattern of the microphone 1 of the top array.

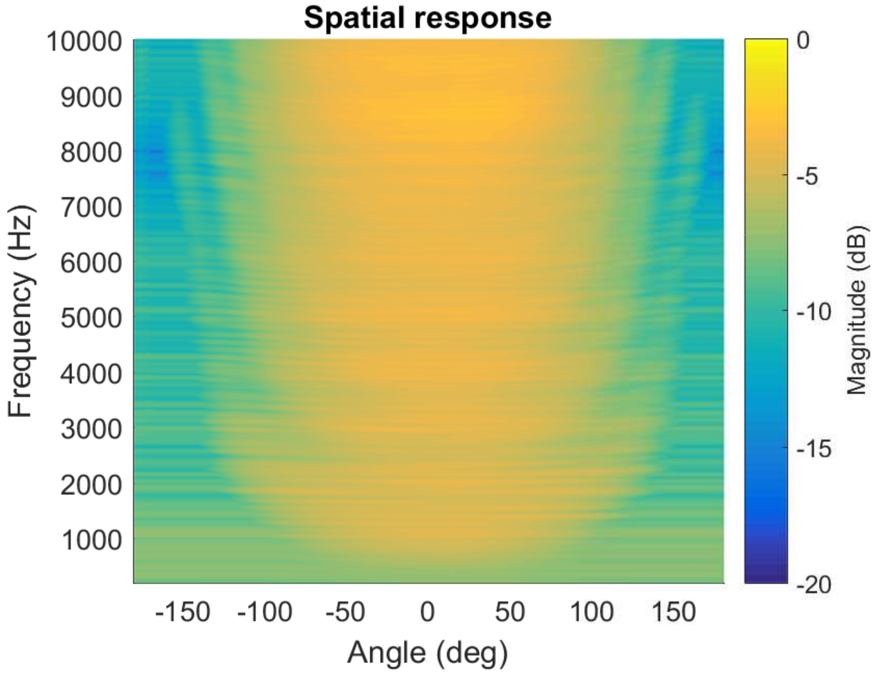
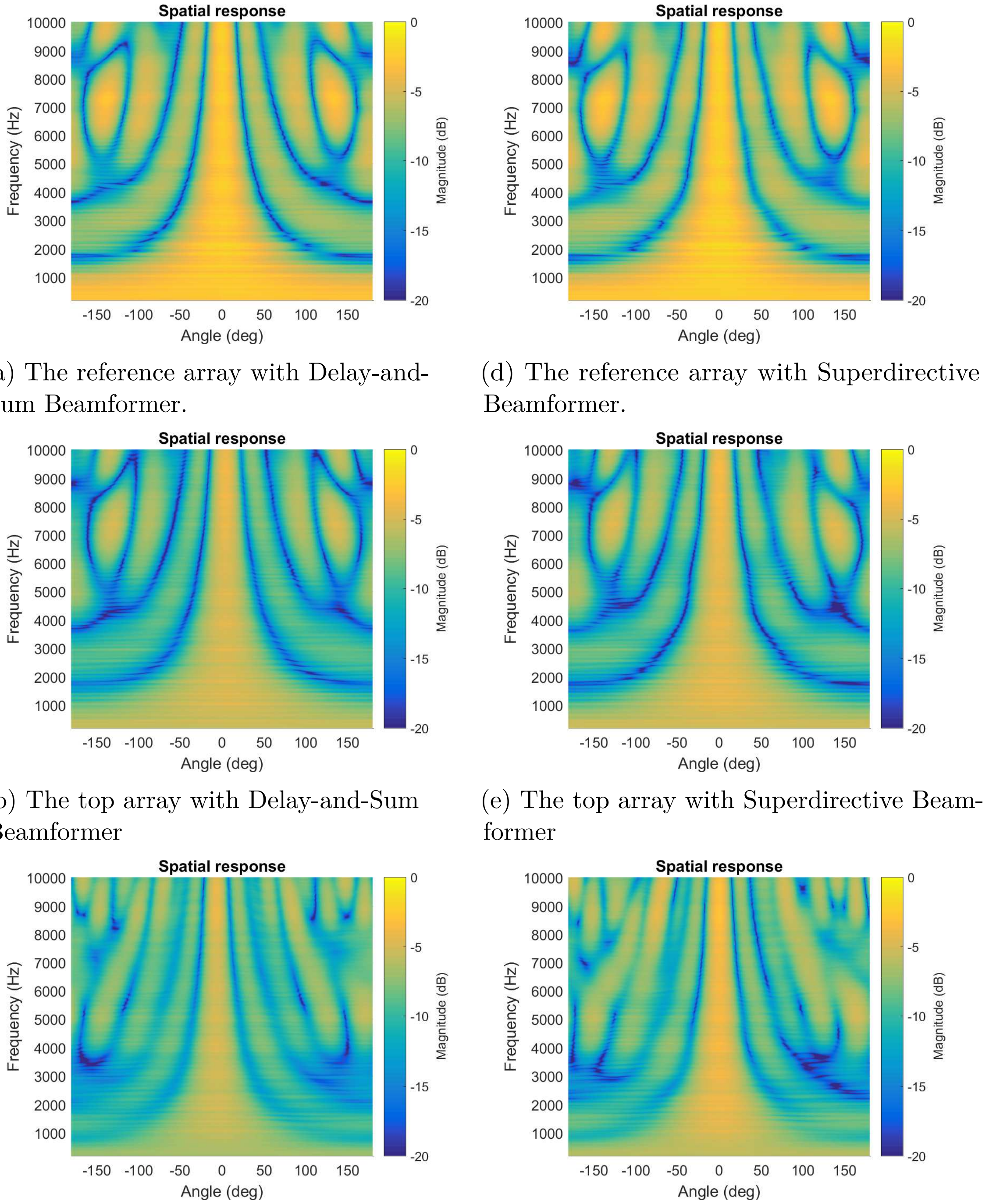


Figure 33: Directivity pattern of the microphone 1 of the baffled array.

top array perform very similarly, with the top array having maybe slightly lower magnitude overall - ie. all the colours are slightly shifted towards blue compared to the reference array's colours. The baffled array on the other hand differs from the two other arrays: first of all, notches appear at clearly lower frequencies, around 1000 Hz while for the other arrays they appear around 2000 Hz. This is due to the baffled array being acoustically bigger than the two other arrays. Secondly, the notches don’t appear to be as deep, but the areas between notches - sidelobes - appear to have lower magnitude than for the two other arrays. The adaptive beamformer, shown in Figure 35, behaves similarly to the two other beamformers.



(a) The reference array with Delay-and-Sum Beamformer.

(d) The reference array with Superdirective Beamformer.

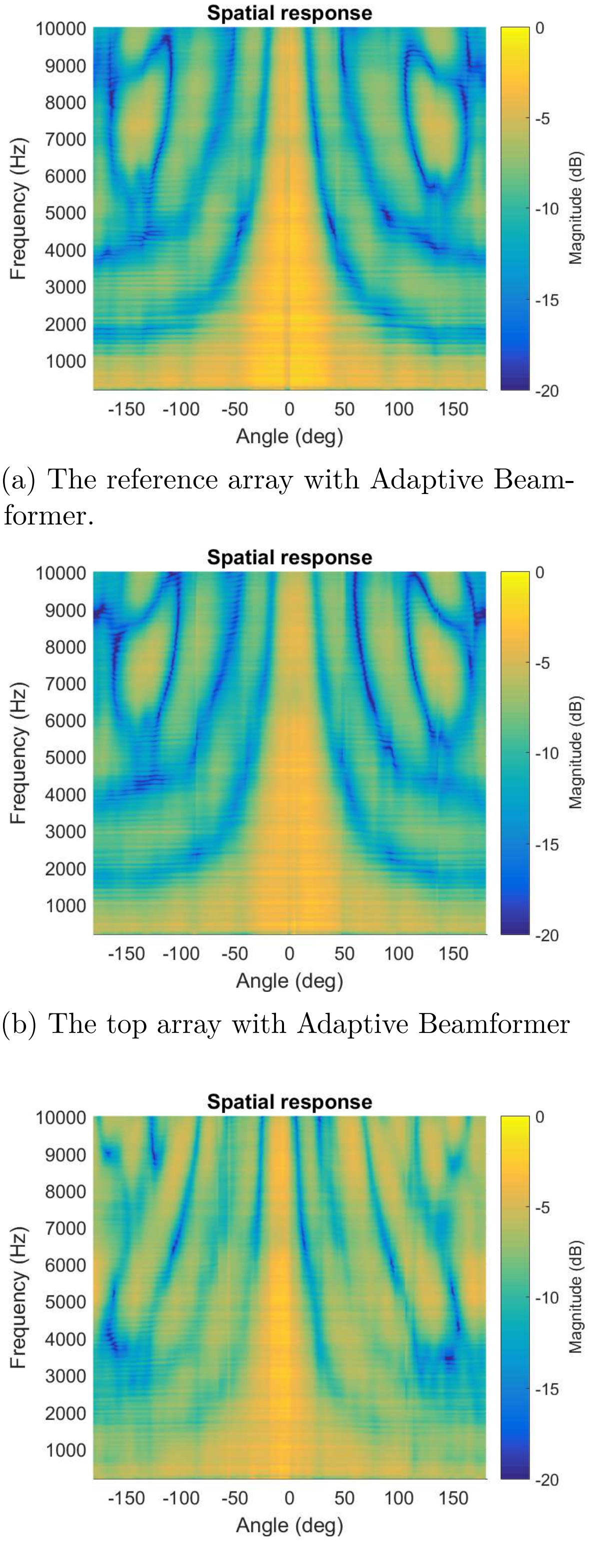
(b) The top array with Delay-and-Sum Beamformer

(e) The top array with Superdirective Beamformer

(c) The baffled array with Delay-and-Sum Beamformer

(f) The baffled array with Superdirective Beamformer

Figure 34: Spatial Responses of the arrays with Delay-and-Sum and Superdirective Beamformers.



(a) The reference array with Adaptive Beamformer.

(b) The top array with Adaptive Beamformer

(c) The baffled array with Adaptive Beamformer

Figure 35: Spatial Responses of the arrays with the Adaptive Beamformer.

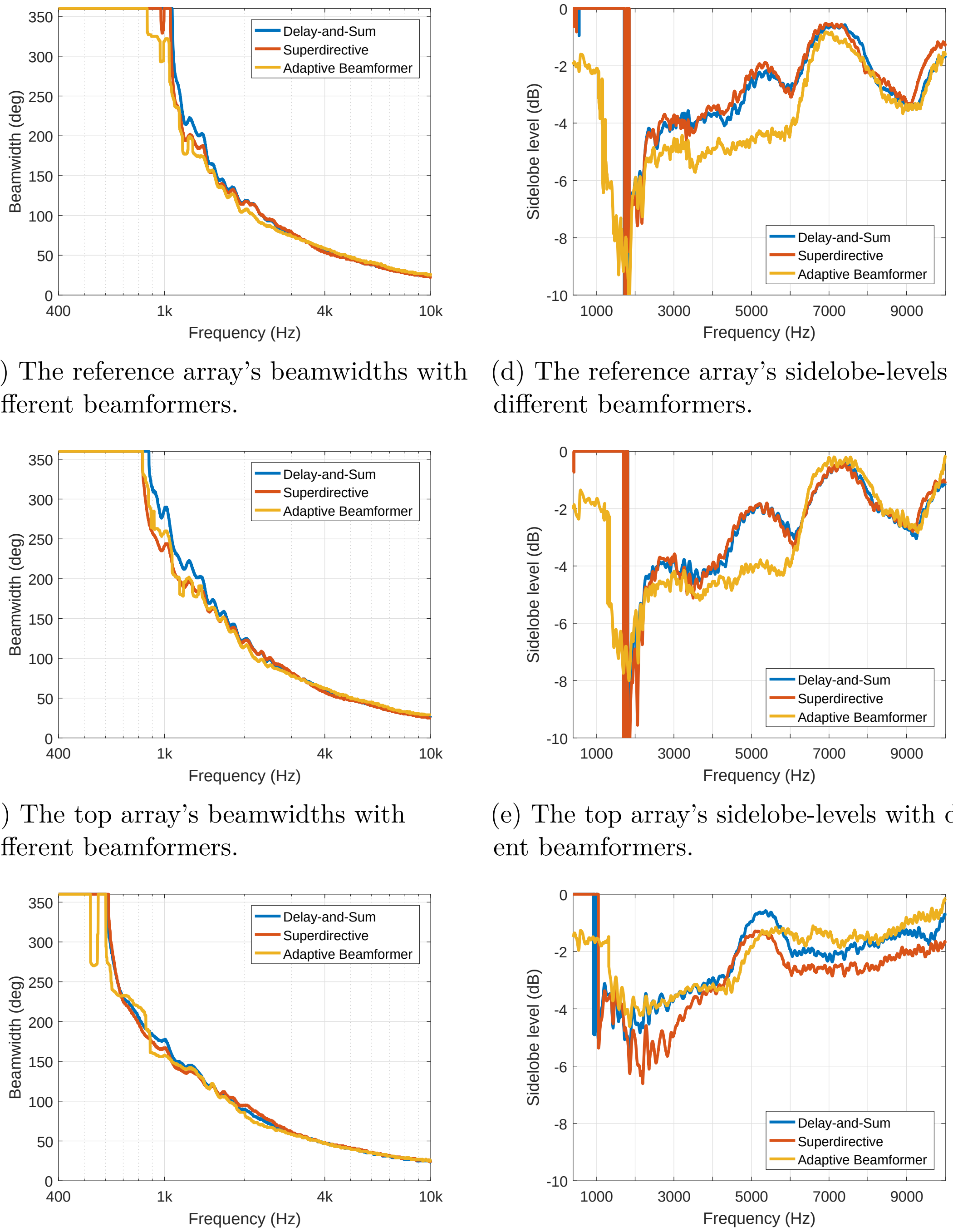
5.2.2 Beamwidths and Sidelobe-levels

Beamwidths and sidelobe-levels can be seen in the spatial responses presented in the section 5.2.1, but the spatial response plots are not so detailed, and it is hard to read from them the exact width of the beam or to see the level of the sidelobes. Thus in this section, those are presented in more readable format.

The beamwidths and sidelobe-levels of the arrays with different beamformers are shown in Figure 36. The beamwidths presented in the Figures 36a-36c show that for each array, the beamwidth doesn't differ much between the beamformers. In comparison, the beamwidth changes clearly when comparing the baffled array with the two others: for the reference and the top arrays, the beamwidth is 360° - ie. the array is practically omnidirectional - for frequencies under around $1000\,Hz$, while the baffled array is omnidirectional only to around $600\,Hz$. In short, if small beam for lower frequencies is critical, baffled array has a clear advantage over the top or reference arrays, due to it being acoustically bigger..

Sidelobe-level describes the difference between the mainlobe's level and the maximum level of sidelobes. Sidelobe-levels show bit more differing between them: for the reference and the top arrays, the Delay-and-Sum and Superdirective beamformers behave and perform practically the same, while the Adaptive beamformer performs better for frequencies under $6000\,Hz$, and for frequencies above that it behaves the same. An interesting difference is also the fact that while for frequencies under roughly $1800\,Hz$, the Delay-and-Sum and Superdirective beamformers have sidelobe-levels of $0\,dB$, meaning that they don't have sidelobes, while the Adaptive beamformer's level's are all over the place. The lack of sidelobes comes from the fact that the first zero appears to the response of the beamformers around that $1800\,Hz$ mark, while the Adaptive beamformer's response appears to have sidelobes even at low frequencies, as is apparent when looking at Figure 37, which shows the magnitude of the beamformer at certain frequency at different angles. For the baffled array, the Delay-and-Sum and Superdirective beamformers differ slightly, with the Superdirective having roughly $0.5\,dB$ better sidelobe-levels. In addition, as with the beamwidth, the sidelobes appear nearly an octave lower for the baffled array than for the two other arrays.

Lastly, the Figure 37 shows that the beamformers behave similarly to one another, having the notches at the same locations, generally speaking having the same levels etc. An interesting difference between them though is how the notch depths change: at $2000\,Hz$ the Superdirective has deeper notches, while for $3000\,Hz$ the Delay-and-Sum has deeper notches. In addition, the Figure 37 shows why the Adaptive beamformer is impractical to measure in the free-field: its response is not so well behaving as the response of the two other beamformers, and it is assumed to be due to the fact that each angle corresponds to different impulse response, and the beamformer adapts differently each time. Similar phenomena can be noticed with the similar figures of the other arrays, shown in figures 38 and 39.



(a) The reference array's beamwidths with different beamformers.

(d) The reference array's sidelobe-levels with different beamformers.

(b) The top array's beamwidths with different beamformers.

(e) The top array's sidelobe-levels with different beamformers.

(c) The baffled array's beamwidths with different beamformers.

(f) The baffled array's sidelobe-levels with different beamformers.

Figure 36: Beamwidths and sidelobe levels of the arrays with different beamformers.

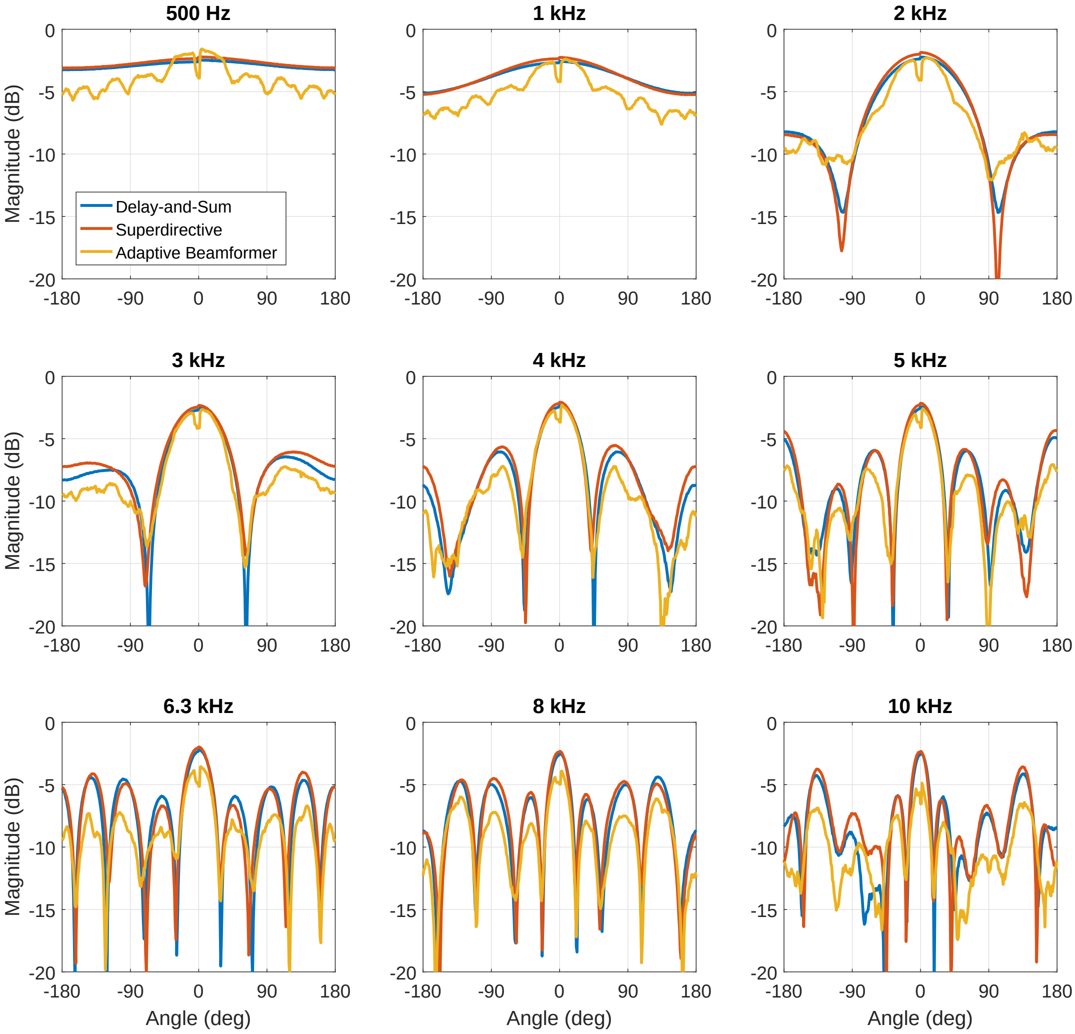


Figure 37: Plots of reference array's magnitude at different frequencies for different angles, with different beamformers. X-axis represent angle in degrees while the y-axis is the magnitude in dB.

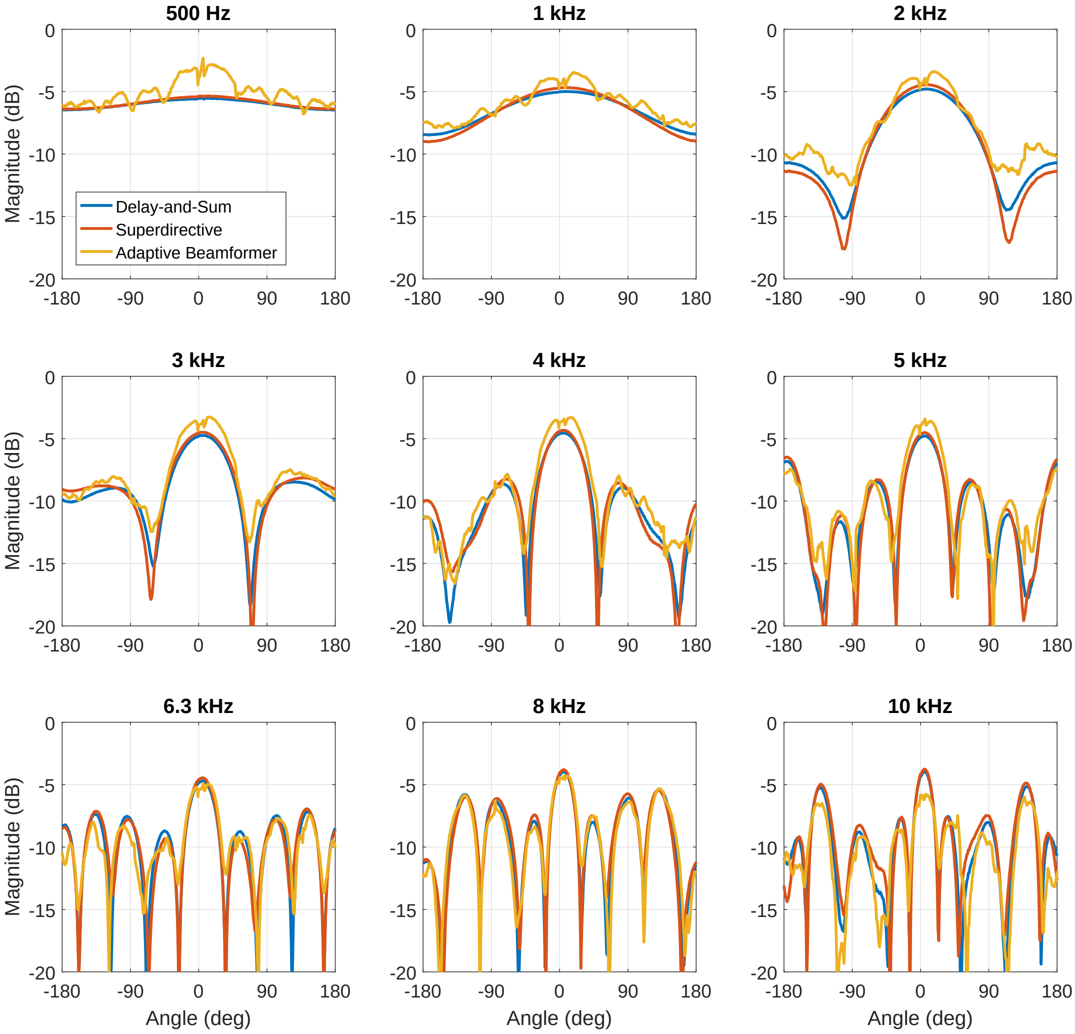


Figure 38: Plots of top array's magnitude at different frequencies for different angles, with different beamformers. X-axis represent angle in degrees while the y-axis is the magnitude in dB.

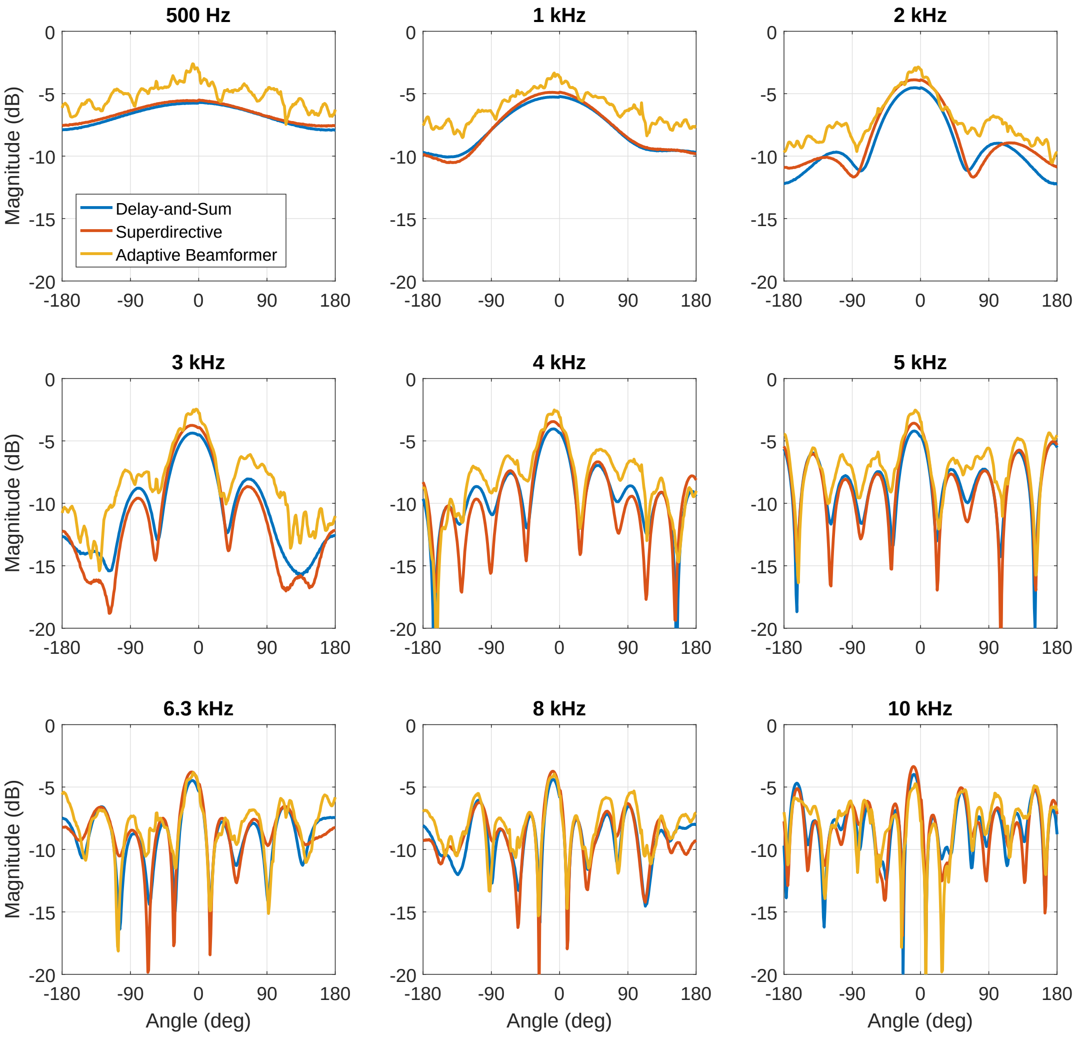


Figure 39: Plots of baffled array's magnitude at different frequencies for different angles, with different beamformers. X-axis represent angle in degrees while the y-axis is the magnitude in dB.

5.3 Diffuse-field Measurements

As explained in the section 4.2.2, the diffuse-field measurements are mostly evaluated using Word-Error-Rate (WER), which describes how well the Automatic Speech Recognition (ASR) system recognizes the correct words. WER is defined as the amount of deleted, substituted or inserted words in the ASR's output compared to the reference text, divided by the length of the reference text. [53] WER is also often used as a metric to evaluate ASRs themselves, and often 10% is considered an acceptable WER for an ASR system. The best systems at the time of writing can reach WER of 6.3% [54]. In addition, ASR performance is often optimized so that the WER is minimized, yet this is not always the optimal target [55]. Nevertheless that is not the point of using WERs in this case: here, they are used just to see how the different arrays and beamformers perform in speech recognition use-case. And as the recordings weren't made in similar fashion as in [54], they aren't even comparable.

First, the WER of the ASR was tested by analysing the original speech samples with the ARS and finding its WER to be 2.4%. Then the signals from different arrays were passed through the same ASR, and the results are given in Table 6. For all the arrays the WER of the microphone 1 - the one closest to the sound source - and the three different beamformers are given. For the baffled array the WER of the microphone 5 is also given, to show how the performance deteriorates when the single microphone is pointing away from the sound source. For the top and reference arrays, the results for microphone 5s aren't shown, as they don't differ noticeably from the microphone 1s results.

By looking at the words the ASR outputs and comparing them to the reference texts, some interesting errors can be noticed: if the system outputs "I'm" while the reference text has "I am", it is considered an error, even though clearly they mean the same. Another similar is where one sentence includes "... watching TV in...", the ASR outputs "... watching T V in...", which is counted as an error. It can be debated whether these should be errors or not, but as long as they are systematically counted the same they don't affect the results of this work greatly. One particular error that is more debatable why it shouldn't be considered error is the sentence "How about now, what's the result, what do you want to do", and the ASR detects it as "How about now, what's the result, what do you wanna do", where the ASR detects the sentence practically correctly, but uses spoken language for one term. But as this work isn't about the WER itself nor how the performance of ASRs should be measured, these errors aren't so crucial.

The WER values in the Table 6 are result of running the same files through the ASR ten times and then taking the average of the results, to see how much there is variance in the results of the ASR. The WERs slightly differ with each measurement, and it can be seen in Figure 40, where the data of Superdirective beamformer of the top array are plotted in a box plot. The results are similar in their variance and distribution for other arrays and beamformers also, so their similar figures aren't shown here. Overall, the differences with different noise levels are clearly bigger than the variance of the ASR. The outliers visible in the Figure 40 were included in the

Table 6: WERs of single microphones, different arrays and beamformers. WN means White Noise in the background, while BN is for Babbling Noise in the background. The WER of the system was 2.4% for the speech samples themselves.

	Silence	WN 40 dB	WN 50 dB	BN 40 dB	BN 50 dB
Reference array					
Mic 1	7.1%	10.1%	20.0%	15.3%	28.5%
Delay-and-Sum	7.6%	10.2%	21.7%	14.7%	28.6%
Superdirective	4.5%	7.0%	11.4%	8.4%	14.2%
Adaptive	6.2%	9.4%	13.6%	11.4%	17.0%
Top array					
Mic 1	6.7%	11.4%	20.0%	14.4%	27.5%
Delay-and-Sum	7.7%	10.9%	22.1%	14.7%	24.2%
Superdirective	3.4%	6.2%	11.8%	9.7%	15.6%
Adaptive	5.1%	7.6%	14.5%	10.4%	16.5%
Baffled array					
Mic 1	4.8%	7.9%	15.1%	13.9%	18.9%
Mic 5	12.6%	22.0%	42.0%	29.1%	57.8%
Delay-and-Sum	6.8%	10.4%	21.3%	15.1%	23.3%
Superdirective	4.1%	6.2%	10.8%	10.0%	16.0%
Adaptive	5.2%	9.1%	17.3%	9.4%	16.6%

average WERs.

In the Table 6 and Figure 41, several trends can be seen. First of all, overall the WERs increase as the background noise levels rise, as expected. Secondly, generally the babbling noise results in worse WERs than similar level white noise, as expected. The only exception for this is the baffled array and adaptive beamformer, as it's WERs with babbling noise are roughly the same as with white noise. In addition, it can be noticed that systematically the Superdirective beamformer increases the performance by around $20 - 50\%$ compared to the microphone 1. The Adaptive beamformer performs slightly worse than the Superdirective beamformer on average, but as discussed in the Section 5.2, this might be due to not using optimal parameters during initializing the beamformer. The performance of the Delay-and-Sum is somewhat surprising, as it performs the same or even worse than the microphone 1 of each array. This could suggest that there is some problem in the implementation of the beamformer, even though it isn't visible in the free-field measurement results. Finally, the baffled array's micorphone 1 performs very well compared to the other arrays, thanks to its directivity. On the other hand, microphone 5 of the baffled array performs significantly worse than any other microphone or beamformer, again thanks to its directivity. This clearly shows one of the advantages beamformers have: with single microphone, the microphone needs to point towards the sound source, while a beamformer can be steered towards the sound source, thus resulting in good

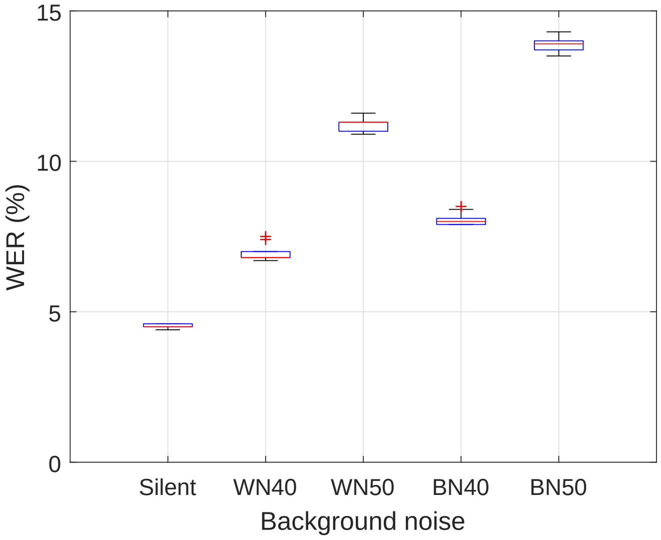


Figure 40: Boxplot of the data for Superdirective beamformer of the top array.

performance regardless of the sound source's position in the plane.

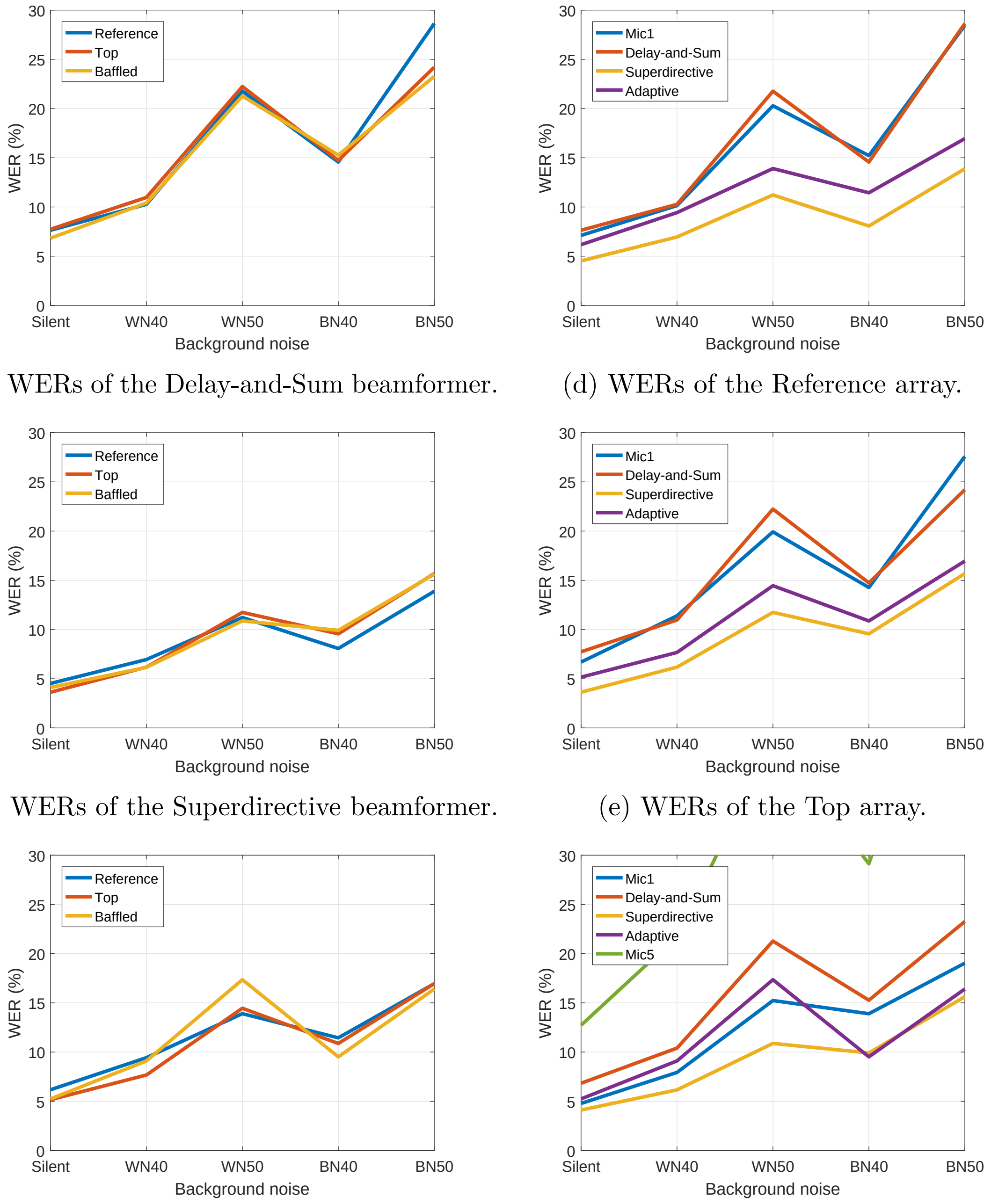
From the WERs, it can also be noticed that the arrays perform approximately the same: for example the Superdirective beamformer's performance in each of the arrays is practically the same, as can be seen in Figure 41b. Similarly, the Delay-and-Sum performs quite consistently with different arrays, as shown in Figure 41a. Slight exception for this is the Adaptive beamformer, which performs slightly differently in the Baffled array compared to the Reference and Top arrays, as easily visible in Figure 41c.

As discussed in the Section 5.1, the physical locations and acoustic locations of the microphones for the baffled array differ from each others. In the free-field results, the difference when initializing the beamformers with the physical or acoustic locations of the microphones was noticeable, and so it was also tested for the diffuse-field measurements. The WERs for the baffled array's microphone 1, Superdirective beamformer with physical locations and Superdirective with acoustic locations of the microphones are shown in Figure 42, and it can be noticed that the performance is slightly better for the acoustic locations case. The difference isn't huge, but it is clear, especially when background noise is white noise. The differences for the other beamformers are similar and thus not presented here.

5.4 Result Analysis

From the previously discussed results, it can be said that in the free-field, the differences between the beamformers are very small, especially when considering the Beamwidths and Sidelobe-levels. In this case, the adaptive beamformer had slightly better Sidelobe-levels than the two other beamformers, but the difference isn't huge, only a decibel or two in the range from 4000 Hz to 6000 Hz.

Yet, the arrays had clear differences between them, or more precisely, the baffled



(a) WERs of the Delay-and-Sum beamformer.

(d) WERs of the Reference array.

(b) WERs of the Superdirective beamformer.

(e) WERs of the Top array.

(c) WERs of the Adaptive beamformer.

(f) WERs of the Baffled array.

Figure 41: WERs of the different arrays and beamformers.

array differed from the reference and top arrays. While the top and reference arrays had practically identical Beamwidths and Sidelobe-levels, the baffled array's beam was clearly narrower than that of the other arrays at the same frequencies. In addition, the Sidelobe-level of the baffled array performed more cleanly than that of the other arrays: instead of it having multiple clear notches and peaks, it only has one notch and otherwise is quite linear. These results show that the baffled array appears bigger than the top and reference arrays - thus improving the performance

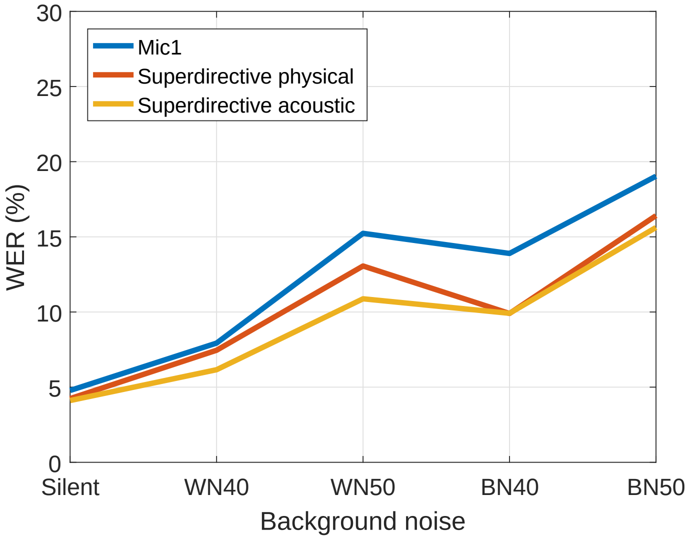


Figure 42: WERs for the baffled array's microphone 1, Superdirective initialized with physical locations and Superdirective initialized with acoustic locations.

at lower frequencies. The bad side-effect of the increased size is then apparent when comparing Figures 37 - 39: as the size increases, the grating lobes appear at lower frequencies as the distance between the microphones increases. For the top and reference arrays, the main lobe is clearly distinguishable at $4\,kHz$, even at $5\,kHz$ - though at that frequency the grating lobe at 180° is nearly as strong as the main lobe - for frequencies above that, the directivity pattern becomes a mess. The same happens for the baffled array, but at roughly $1\,kHz$ earlier.

This leads to an interesting result: if we have an application whose form-factor is a fixed sized cylinder, the choice between the top and baffled array depends on the target use-case. If the use-case needs to maximise the performance at lower frequencies, then the baffled array should be used, but if the performance at higher frequencies is more important, then the top array is the better choice. For this particular size, the decision would be which frequency range is more important: $500 - 3000\,Hz$ or $1000 - 4000\,Hz$.

To actually measure if these measured differences have any practical effect, the WERs of the systems were measured in a diffuse-field. WERs are a nice measure, as they clearly show the performance of the particular array and beamformer in a real use-case, ie. ASR, but on the other hand, they don't measure the performance of the arrays for other use-cases!

Overall, the results were that for ASR, the arrays didn't differ very much. The Delay-and-Sum and Superdirective beamformers performed practically identically for each array. For the Adaptive beamformer, the baffled array performed slightly worse at with high-level White Noise as the background noise, while otherwise performing approximately the same or slightly better as the others. For WERs, the differences between the beamformer algorithms are clear: the Delay-and-Sum performs similarly to a single microphone or worse, but the Superdirective or the

Adaptive beamformers perform clearly better. Adaptive beamformer performs slightly worse than the Superdirective, but as this Adaptive beamformer was quite a simple GJBF, the results could be different if more advanced algorithm was used. In addition, tuning the parameters of the Superdirective and Adaptive beamformers changes their performance, and it could be that the parameters used for these weren't the most optimal ones. Thus, for ASR use-case, it can be said that the choice of the array for a fixed sized cylinder doesn't really matter, but the beamformer has more effect.

Lastly the problem with the baffled array needs to be considered: what is the correct location of the microphones when initializing the beamformers. Some more advanced beamformers don't need to be initialized with the locations, but the implementation of the Superdirective and the Adaptive beamformers in this thesis needed - and Delay-and-Sum used the delays acquired from the impulse responses, so it also practically uses locations. As the baffled array's microphones acoustic locations aren't the same as their physical locations for a wave coming from a certain direction, this needs to be taken into an account. It was confirmed that using the physical locations of the microphones as the initializing values, the performance deteriorates compared to using the acoustic locations extracted from the free-field measurements. The challenge is that the acoustic locations depend on the DoA of the sound event, and how to take this into an account. This makes correctly estimating DoA extremely important for the baffled array. In this sense, beamformers for top array is simpler to implement than for the baffled array. In addition, it is possible to add ninth microphone to the top array, to the centre of the circle, which would increase the directivity of the beamformer. For the baffled array it is not possible.

6 Conclusions

Human auditory system has evolved for thousands of years and is naturally able to perform well even in very challenging environments, such as in babbling noise. This is achieved thanks to the ability to differentiate multiple signals and focus on single sound events, enabling humans to "block out" unwanted noises. Lately, Automatic Speech Recognition applications has become increasingly common, and thus the need for human-to-machine communication by speech. This is especially driven by personal assistants, such as Apple's Siri and Amazon's Alexa, but overall as speech is the natural way of communication for humans, it is not surprising that machines would be controlled with it. To enable human-like auditory performance for machines, microphone arrays and beamformers are used. Multiple microphones combined with beamformers enable complex directivity patterns which can be steered towards the sound source(s), while simultaneously blocking noises and other sound sources, thus enabling similar features as the human auditory system has. Even though the human-to-machine communication is a hot topic at the moment, it is not the only use-case for microphone arrays and beamformers, as they are used for telephony also: they are used in phones - be it mobile or conference phone, or hands-free system in a car - to increase the Signal-to-Noise Ratio and the quality overall of the transmitted speech signal. In addition they can be used to enable devices and use-cases that are not possible with single microphones, such as acoustic camera.

To combine the signals coming from multiple microphones a beamformer is used. Generally speaking, beamforming is a way to combine signals from multiple spatially different sensors - be it microphones, antennas or something else - into a single signal in such a way that the wanted signal is left unchanged while noise and other unwanted signals are attenuated. This can be done in various ways, with different levels of complexity and performance. Generally speaking, the more complex beamforming algorithms have better performance, such as more accurate source localization or better attenuation of unwanted sources, or they add new features, such as an ability to track multiple sources or steer nulls in the beampattern in addition to steering the mainlobe. The most well-known and simplest beamformer is called Delay-and-Sum beamformer, whose idea is just to delay the signals from the microphones, and then sum them. The delays for each signal is chosen so that the wanted signals are in phase and thus summing the signals causes constructive interference, while for signals coming from other directions, the summing results in destructive interference. For this thesis, in addition to Delay-and-Sum, Superdirective and Griffiths-Jim Adaptive Beamformers were used. Superdirective beamformer is designed to remove noise and to be more directive than Delay-and-Sum, while Griffiths-Jim Beamformer adapts to the environment in a way that produces the best output signal quality.

Beamforming is used in various fields, including radar, sonar, communications, imaging, geophysics, astronomy and biomedicine. The basic idea of beamforming for each field is the same, but the sensors, the signal or the medium in which the signal travels differs. The applications also differ from the requirement point-of-view: while radio applications usually have narrowband signal in far-field, audio signals are often wideband and they can be either in the near-field or far-field. This thesis focused on microphone beamforming, in which the sensors are microphones, the signal is sound and it travels through air.

In addition to the different beamformers, the arrays have an effect on the resulting output signal. The array can be anything from a simple line array with a couple of microphones to a huge spherical array with tens or hundreds of microphones and the sizes can vary from less than a few centimetres to over a meter. The array basically defines the performance of the beamformer: the size of the array determines the low-frequency performance, while the distance between the microphones determines the high-frequency performance. The amount of microphones ties these two together: in some cases the amount of microphones is predetermined, and thus a trade-off between high- and low-frequency performance needs to be made. In addition, the geometry of the array limits the beampattern: a simple line array can only detect the Direction-of-Arrival in one dimension, a two dimensional array can detect Direction-of-Arrival from any direction but can't differentiate between signals coming from up or down, while symmetrical three-dimensional arrays have a constant beampattern to all directions. When designing a device with microphone array, the size and geometry of the array are usually limited by the industrial design of the device. As an example, a high-end laptops nowadays often have multiple microphones in them, but due to the form-factor - a thin slab basically - it is practically impossible to use three-dimensional arrays in them, and as motherboards have limited amount of microphone inputs, the amount of microphones used is limited without the use of an additional circuitry.

The purpose of this thesis is to figure out how specific practical structure the microphones are attached to affect the performance of the beamformers. A cylinder was chosen as the form-factor, based on the fact that most of the practical devices that use microphone arrays and beamformers for Automatic Speech Recognition use-case are cylindrical. A simple circular array can be attached to either the top surface of the cylinder or around the surface of the cylinder. These arrays are compared to a reference array, which consists of microphones at the end of thing tubes, such that they basically hover in the air. The reference array and the top array both have a diameter of $83\,mm$ while the cylinder - and thus the openings for the microphones of the baffled array - has a diameter of $90\,mm$. In this thesis the arrays have eight microphones as commercially available audio codecs have maximum of eight microphone inputs for the time being. The microphones used were analog MEMS-microphones, to enable easy connection to a sound card for measurement.

The different arrays were measured in a free-field in an anechoic chamber and in a diffuse field. In the free-field, the impulse responses of the microphones of the different arrays were measured at different azimuth and elevation angles. From the impulse responses, different measures can be derived, such as beamwidths, side-lobe levels and spatial responses of the arrays and beamformers. From the measurements, it becomes apparent that the beamformers performed quite similarly in the free-field, but that was expected for there is no noise to cancel or adapt to. On the other hand, the differences between the arrays were clear, especially the difference between the baffled array and the other arrays. The baffled array appeared to be practically

50% bigger than the other arrays - even though physically it was less than 10% bigger - resulting in better low-frequency performance but poorer high-frequency performance. Based on the delays between the microphones, the acoustic locations of the microphones for different arrays were calculated. For the reference and the top arrays they were as expected: approximately 83 *mm* circles. For the baffled array, the acoustic locations differed greatly from that. Instead of the locations forming a circle, the result is more oval-like. Analysing the locations for different frequency-bands suggests that the shape is caused by the shadowing caused by the structure.

The diffuse-field measurements were done with a specific use-case in mind: Automatic Speech Recognition. The idea was to measure how the different arrays and beamformers perform with different background noise-levels when used for Speech Recognition. The measurements were done in an acoustically well behaving room, in which the background noises were generated by eight loudspeakers, while the measurement signal was generated by a Head-and-Torso-Simulator to simulate a real human speaker. The measurement signal was speech samples consisting of 1000 words. The signal was played through the Head-and-Torso-Simulator and recorded by the arrays, then beamformed with different algorithms, and the results were analyzed with Intel's Automatic Speech Recognizer. The performance used was measured using Word-Error-Rate, which basically describes how big percentage of the words the recognizer got wrong. The results show that for this use-case using arrays and advanced beamformers clearly improves the performance compared to using a single microphone, but that the arrays themselves have very little impact on the performance. On the other hand, the beamformers differ quite clearly so that the Superdirective beamformer was the best, the Adaptive beamformer was slightly worse yet better than a single microphone and the Delay-and-Sum performed similarly to a single microphone.

The measurements done in the free-field rose the question about the acoustic locations of the microphones in the baffled array, and should the acoustic locations of the microphones be used when initializing the beamformers instead of the physical locations. When comparing the results from the free-field measurements with the beamformer initialized using acoustic locations and physical locations, the one with the acoustic locations performs better. In the diffuse-field measurements, when comparing the Word-Error-Rates, the beamformer initialized with the acoustic locations performed better than the one initialized with the physical locations, which further strengthens the idea that the acoustic locations should be used for initializing the beamformers. The problem with the locations is only relevant for the beamformers which use the locations of the microphones to calculate delays or some other parameters of the beamformer, while some more advanced beamformers don't need the location information at all.

This thesis leaves some interesting questions open such as:

- Is the size increase for baffled array always approximately 50%?
- How much would performance of the beamformer increase if the inputs were filtered to frequency bands and each band would have its own acoustic location, the bands would be beamformed and then the wideband signal would be

reconstructed? Ie. how the acoustic locations should be used for frequency domain beamforming?

- Is there way to use only four microphones without much of a performance hit by smartly switching between the microphones in use? This is especially interesting in the case of the baffled array as the signals from the microphones behind the structure are deteriorated due to the shadowing. For real devices, this could enable using cheaper audio codecs for example.

- How would the performance differ if the microphones weren't in a plane? For example, what kind of results would combining the top and the baffled arrays lead to? Or two this kind of baffled arrays so that the second one is few centimetres above the other?

- How does the baffled array affect the Direction-of-Arrival estimation and its accuracy?

The data recorded for this thesis enables further research on the second and third open questions listed above, but as they are out of the scope of this thesis, they are not reported here. Furthermore, the measurement setups could be changed slightly to enable answering the first, the fourth and the fifth questions, allowing to answer all these questions.

All the hyperlinks in references were checked to work on 16.1.2017.

References

[1] Alberto Abad Gareta. *A Multi-Microphone Approach to Speech Processing in a Smart-Room Environment.* PhD thesis, Speech Processing Group, Department of Signal Theory and Communications, Universitat Politecnica de Catalunya, 2007.

[2] Israel Cohen, Jacob Benesty, and Sharon Gannot. *Speech processing in modern communication: challenges and perspectives*, volume 3. Springer Science & Business Media, 2009.

[3] Michael Brandstein and Darren Ward. *Microphone arrays: signal processing techniques and applications.* Springer Science & Business Media, 2013.

[4] John McDonough Matthias Wölfel. *Distant Speech Recognition.* Wiley, 2009.

[5] Kenichi Kumatani, John McDonough, and Bhiksha Raj. Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. *Signal Processing Magazine, IEEE*, 29(6):127–140, 2012.

[6] Jacob Benesty and Chen Jingdong. *Study and design of differential microphone arrays*, volume 6. Springer Science & Business Media, 2012.

[7] Toby Haynes. A primer on digital beamforming. *Spectrum Signal Processing*, 11, 1998.

[8] Daniel Jackson Allred. Evaluation and comparison of beamforming algorithms for microphone array speech processing. Master's thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology, August 2006.

[9] Antti Räisänen and Arto Lehto. *Radiotekniikan perusteet.* Otatieto, 2011.

[10] Ron Streicher and Wes Dooley. Basic stereo microphone perspectives: A review. In *Audio Engineering Society Conference: 2nd International Conference: The Art and Technology of Recording*, May 1984. URL `http://www.aes.org/e-lib/browse.cfm?elib=11662`.

[11] Hamid Krim and Mats Viberg. Two decades of array signal processing research: the parametric approach. *Signal Processing Magazine, IEEE*, 13(4):67–94, 1996.

[12] Don E Dudgeon and Don H Johnson. Array signal processing: concepts and techniques. *PRT Prentice Hall, Englewood Cliffs, NJ*, 1993.

[13] W Marshall Leach. *Introduction to electroacoustics and audio amplifier design.* Kendall/Hunt Publishing Company, 2003.

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

[15] Alfons Dehé, Martin Wurzer, Marc Füldner, and Ulrich Krumbein. A4. 3-the infineon silicon mems microphone. *Proceedings SENSOR 2013*, pages 95–99, 2013.

[16] Robert John Littrell. *High performance piezoelectric MEMS microphones.* PhD thesis, The University of Michigan, 2010.

[17] Simon Doclo and Marc Moonen. Superdirective beamforming robust against microphone mismatch. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(2):617–631, 2007.

[18] Jens Meyer and Gary Elko. A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II–1781. IEEE, 2002.

[19] *Microphone Array Support in Windows.* Microsoft, 04 2014. `http://download.microsoft.com/download/9/c/5/9c5b2167-8017-4bae-9fde-d599bac8184a/MicArrays.doc`.

[20] Mingsian R. Bai, Yueh Hua Yao, Chang-Sheng Lai, and Yi-Yang Lo. Design and implementation of a space domain spherical microphone array with application to source localization and separation. *The Journal of the Acoustical Society of America*, 139(3):1058–1070, 2016. doi: http://dx.doi.org/10.1121/1.4942639. URL `http://scitation.aip.org/content/asa/journal/jasa/139/3/10.1121/1.4942639`.

[21] Abhaya Parthy, Craig Jin, and Andrévan Schaik. Measured and theoretical performance comparison of a co-centred rigid and open spherical microphone array. In *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on*, pages 1289–1294. IEEE, 2008.

[22] Mark A Poletti. Effect of noise and transducer variability on the performance of circular microphone arrays. *Journal of the Audio Engineering Society*, 53(5): 371–384, 2005.

[23] Kazunori Kobayashi, Ken-ichi Furuya, and Akitoshi Kataoka. A talker-tracking microphone array for teleconferencing systems. In *Audio Engineering Society Convention 113*, Oct 2002. URL `http://www.aes.org/e-lib/browse.cfm?elib=11295`.

[24] Mostafa Nofal, Sultan Aljahdali, and Yasser Albagory. Tapered beamforming for concentric ring arrays. *AEU-International Journal of Electronics and Communications*, 67(1):58–63, 2013.

[25] Philippe-Aubert Gauthier, Éric Chambatte, Cédric Camier, Yann Pasco, and Alain Berry. Beamforming regularization, scaling matrices, and inverse problems for sound field extrapolation and characterization: Part i – theory. *J. Audio Eng. Soc*, 62(3):77–98, 2014. URL `http://www.aes.org/e-lib/browse.cfm?elib=17125`.

[26] Osamu Hoshuyama, Akihiko Sugiyama, and Akihiro Hirano. A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. *IEEE Transactions on signal processing*, 47(10):2677–2684, 1999.

[27] Jens Meyer. Beamforming for a circular microphone array mounted on spherically shaped objects. *The Journal of the Acoustical Society of America*, 109(1):185–193, 2001.

[28] Boaz Rafaely. Analysis and design of spherical microphone arrays. *Speech and Audio Processing, IEEE Transactions on*, 13(1):135–143, 2005.

[29] Elisabet Tiana-Roig, Finn Jacobsen, and Efrén Fernández Grande. Beamforming with a circular microphone array for localization of environmental noise sources. *The Journal of the Acoustical Society of America*, 128(6):3535–3542, 2010. doi: http://dx.doi.org/10.1121/1.3500669. URL `http://scitation.aip.org/content/asa/journal/jasa/128/6/10.1121/1.3500669`.

[30] Jerome Daniel and Nicolas Epain. Improving spherical microphone arrays. In *Audio Engineering Society Convention 124*, May 2008. URL `http://www.aes.org/e-lib/browse.cfm?elib=14609`.

[31] Daniel P Jarrett, Emanuël AP Habets, Mark RP Thomas, Nikolay D Gaubitch, and Patrick A Naylor. Dereverberation performance of rigid and open spherical microphone arrays: Theory & simulation. In *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on*, pages 145–150. IEEE, 2011.

[32] *SPH0642HT5H-1, Wide Bandwidth, Low Noise, Precision Top Port SiSonic^{TM} Microphone*. Knowles, 04 2014. Rev. A `http://www.mouser.com/ds/2/218/-746181.pdf`.

[33] Masato Miyoshi Marc Delcroix, Takafumi Hikichi. Precise dereverberation using multichannel linear prediction. *IEEE Transactions on Audio, Speech and Language Processing*, 15(2), February 2007.

[34] Emanuël Habets Reinhold Haeb-Umbach Armin Sehr Walter Kellermann Sharon Gannot Bhiksra Raj Keisuke Kinoshitaa, Marc Delcroix. The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In *IEEE Workshop on Applicatoin of Signal Processing to Audio and Acoustics*, 2013.

[35] Craig Jin, Abhaya Parthy, and André van Schaik. Measured and theoretical performance comparison of a broadband circular microphone array. In *Audio Engineering Society Conference: 31st International Conference: New Directions in High Resolution Audio*, Jun 2007. URL `http://www.aes.org/e-lib/browse.cfm?elib=13955`.

[36] *APx525 Audio Analyzer.* Audio Precision. `http://www.alava-ing.es/repositorio/0067/pdf/1067/2/analizador-de-audio-audio-precision-apx525-caracteristicas.pdf`.

[37] *User's Guide Fireface UFX.* RME, 03 2016. Version 2.4 `https://www.rme-audio.de/download/fface_ufx_e.pdf`.

[38] *fP 2400Q User Manual.* Lab.Gruppen, 10 2003. Version 0.8 `http://www.marketing.labgruppen.com/webservices/dh.ashx?t=qv&v=6114`.

[39] *FR125SR Full Range.* Creative Sound Solutions. `http://diyaudioprojects.com/Drivers/CSS-FR125SR/CSS-FR125SR-Speaker-Datasheet.pdf`.

[40] ISO. 3745:2003, determination of sound power levels of noise sources using sound pressure, 2003.

[41] 3GPP. 26.132, speech and video telephony terminal acoustic test specification, 2011.

[42] *TEDS Microphones.* Brüel & Kjær, . `https://www.bksv.com/~/media/literature/Product%20Data/bp2225.ashx`.

[43] *Head and Torso Simulator Types 4128-C and 4128-D.* Brüel & Kjær, . `https://www.bksv.com/~/media/literature/Product%20Data/bp0521.ashx`.

[44] ITU. Recommendation ITU-R BS.1116-1, 1994-1997.

[45] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings. *The Journal of the Acoustical Society of America*, 133(5):3591–3591, 2013.

[46] EBU. R128, loudness normalisation and permitted maximum level of audio signals, 2014.

[47] ITU. Recommendation ITU-R BS.1770-4, 2015.

[48] ITU. Recommendation ITU-T P.58, 2013.

[49] Karl S Pearsons and R Horonjeff. Measurement of speech levels in the presence of time varying background noise. 1982.

[50] Karl S Pearsons, Ricarda L Bennett, and Sanford Fidell. *Speech levels in various noise environments.* Office of Health and Ecological Effects, Office of Research and Development, US EPA, 1977.

[51] Wayne O Olsen. Average speech levels and spectra in various speaking/listening conditionsa summary of the pearson, bennett, & fidell (1977) report. *American Journal of Audiology*, 7(2):21–25, 1998.

[52] J Fiscus. Sclite scoring package, version 1.5. us national institute of standard technology (nist), 2015.

[53] Benoit Favre, Kyla Cheung, Siavash Kazemian, Adam Lee, Yang Liu, Cosmin Munteanu, Ani Nenkova, Dennis Ochei, Gerald Penn, Stephen Tratz, et al. Automatic human utility evaluation of asr systems: does wer really predict performance? In *INTERSPEECH*, pages 3463–3467, 2013.

[54] W Xiong, J Droppo, X Huang, F Seide, M Seltzer, A Stolcke, D Yu, and G Zweig. The microsoft 2016 conversational speech recognition system. *arXiv preprint arXiv:1609.03528*, 2016.

[55] Xiaodong He, Li Deng, and Alex Acero. Why word error rate is not a good metric for speech recognizer training for the speech translation task? In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5632–5635. IEEE, 2011.