UiO : Department of Informatics
University of Oslo

Microphone array processing using reflections constructively

Anders Ueland

Master's Thesis Spring 2014

Abstract

A method for speech enhancement in a reverberating environment using microphone arrays is presented. It is defined in the time-domain, based on an acoustic signal model formulated using image sources.

The method uses the output of a delay and sum (DAS) beamformer steered towards a source together with the outputs of multiple DAS beamformers steered towards the reflections. The outputs are delayed and added together so that the direct signal and the reflections are added coherently. A minimum power/distortionless response (MVDR)-formulation for weighting of the reflections is also presented.

The method requires that the directions of arrival (DOA) and time differences of arrival (TDOA) for the reflections are known. In addition the MVDR formulation requires that the relative reflection strengths are known approximately.

Simulations and measurements in a real room show that the method suppresses reverberations, interfering sources, and microphone self noise. The direct signal is not distorted. The method has improved performance compared to the conventional DAS beamformer given that the reflections used in the beamformer are strong.

Acknowledgements

This thesis is submitted as a part of the master's degree in informatics: Technical and Scientific Applications. It has been very interesting to work with acoustics and microphone arrays, but also demanding and time consuming.

I would like to thank my supervisor Carl-Inge C. Nilsen for his generous use of time and knowledge. I have learned a lot about signal processing and research in general from him.

Thanks to my internal supervisor Sverre Holm, and the rest of the employees at the DSB group[1], for good help and for inviting us (master students) to become a part of the group. Also I would like to give a special thanks to senior engineer Svein Bøe for helping me with using the Condor cluster. Thanks to Ines Hafizovic for guidance and Squarehead Technology AS for providing me with a microphone array for the measurements.

Also I would like to thank my fellow students for good technical discussions and much needed coffee breaks. Finally, thanks to Asbjørn Ueland and my girlfriend Una Kristin Waldeland for helping me correct errors in the report.

[1]Digital Signal Processing and Image Analysis group at the department of Informatics (IFI), University of Oslo

Contents

Chapter 1

Introduction

1.1 Beamforming using microphone arrays

This study investigates a new beamforming method for microphone arrays to increase speech intelligibility.

A microphone array is a collection of spatially distributed microphones that are measuring the same sound[1] field. Consider a sound source in a given position in space; the sound will reach each microphone at slightly different times due to the differences in the length of travel paths. The microphone outputs can be processed to give information about how the wave field varies in both time and space.

Beamforming is the process of using an array to listen to sound waves from only one direction at the time. This is also called "spatial filtering". It means that the sound originating from some part of space is kept undistorted, while sound originating from other places is suppressed. Some use the word "beamforming" for the process of using microphone arrays to determine the directions of arrival (DOA) of a source. This is not what is meant by "beamforming" in this thesis.

The most common beamforming method is the delay and sum (DAS) beamformer. Chapter 2 offers a more detailed and precise definition, while this is just a short introduction. Given a direction where the beamformer should be steered towards, each sensor output is *delayed* so that sound waves from this direction are aligned in time. The delayed outputs are then *summed* together. The sound waves are added coherently[2]/constructively and are therefore amplified. Sound from other directions will not be aligned by the sensor delays, and will therefore be added incoherently[3]/destructively and be suppressed. Figure 1.1 shows how the amplification for a DAS beamformer varies depending on the direction of the incoming sound. This is

[1] *Sound* means pressure waves that are detectable by the human hearing system.[1]

[2] *Coherently* means that the signals have no relative delay to each other. This is visible when the signals are plotted on top of each other; the wave peaks and bottoms are aligned.

[3] *Incoherently* means that the signals are delayed relative to each other. When the signals are plotted on top of each other, the wave peaks and bottoms are not aligned.

called the *beampattern*. The *beam* is some part around the steered direction which amplifies the signal by the same factor, within some tolerance (often 3dB). By adjusting the delays, the beam can be electronically steered towards any direction. This is the main difference to a highly directional microphone (i.e. a shotgun mic), which must be steered physically. While the directional microphone can only be steered towards one position at the time, one single array can be steered towards different positions simultaneous using multiple beamformers.

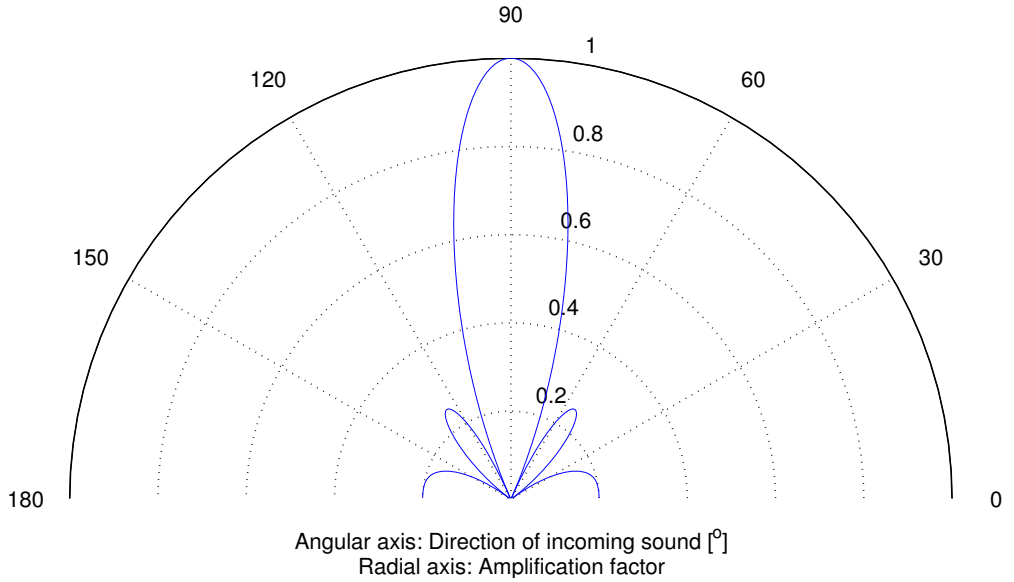


Figure 1.1: Beampattern for a 5-element array when steered towards 90°*. The signal has a wavelength equal two times the distance between the elements.*

Beamforming with microphone arrays has many applications.

- In the industry beamforming is often used for noise analysis. The beam can be steered towards all positions in a grid. In each point the power of the beamformer output is mapped, creating an *acoustic map*. This can tell the engineers where the noise is coming from and its strength. Such analysis is used when designings cars, trains and aircrafts [2].
- Microphone arrays are used in auditoriums or in videoconference systems to amplify the speech of a talking person, while other sound sources are suppressed. This can e.g. be useful as an alternative to passing a microphone around at Q&A sessions [3].
- A microphone array can be used in combination with a video camera for surveillance applications. When a scene is recorded, a visual inspection or automatic video analysis of the video can give information about which region of space that is of interest. The beam can then be steered towards that location. The beamformer will suppress other interfering sources, so that the speech of the persons of interest becomes more intelligible [3].

Ethical aspects of beamforming with microphone arrays The possibility to steer beamformers poses an ethical challenge with respect to privacy. Especially the possibility to steer the beam on recorded data makes it possible to do extensive surveillance. Audio surveillance is in some ways more intrusive than regular video surveillance. Microphone arrays could give governments the possibility to get solid evidence of criminal actions that would not be possible to achieve in any other way. At the same time they pose a threat to privacy. It is important that users of this technology discuss these issues. For the authority it is important to establish legal regulation for the use of such equipment.

Norwegian law [4] has the same restrictions for audio and video surveillance. In questions on whether public survailiance is legal or not, it states that there should be put emphasis of whether the surveillance prevents personal injuries and repeated or severe criminal actions[4]. In addition there should be written warnings that specify that audio is being recorded[5]. Otherwise the affected persons must give their approval to the operators of the array.

1.1.1 The history of beamforming with microphone arrays

In the first world war (1914-1918), acoustical arrays were used by the French military forces to detect enemy aircraft [5]. The arrays were purely analog. They consisted of tracts acting as microphones. The sound was lead from the tracts, through tubes, to the ears of an operator. By adjusting the angle and position of the tracts the operator could detect where the aircraft were coming from.

The first electrical microphone array was proposed by Billingsley and Kinns in 1976 [6, 7]. It was used for noise analysis of Rolls Royce aircraft engines. The array consisted of 14 microphones positioned along a straight line. The computer analysing the signals used the DAS algorithm and frequency filters. It produced spectrograms of frequency and spatial position.

Microphone arrays were in the 70´s and 80´s used mainly to analyze noise from trains [8]. Later they were used to track and explore the noise of flying aircrafts [9]. The array processing was still done with the conventional DAS algorithm. Much of the research was focused on the effect of the digital sampling rate and array geometries. The results from Piet, Michel, and Böhning [10] show that the up-sampling of the data and the used interpolation method had a great effect on the signal to noise ratio of the beamforming.

A detailed narrative of the microphone array history is presented in [11].

[4]From [4], §37 *"... for kameraovervåking legges vesentlig vekt på om overvåkingen bidrar til å verne om liv eller helse eller forebygger gjentatte eller alvorlige straffbare handlinger. Kameraovervåking skal kun anses som behandling av sensitive personopplysninger der slike utgjør en vesentlig del av opplysningene som overvåkingen omfatter."*

[5]From [4], § 40: *Ved kameraovervåking på offentlig sted eller sted hvor en begrenset krets av personer ferdes jevnlig, skal det ved skilting eller på annen måte gjøres tydelig oppmerksom på at stedet blir overvåket, at overvåkingen eventuelt inkluderer lydopptak og hvem som er behandlingsansvarlig.* "

Adaptive beamformers are based on weighting the different array microphones in an optimal way. This is depending on the characteristics of the observed wave field. *Optimal* may mean to achieve minimization of the reverberation, interference or noise while the signal is kept undistorted. Some techniques worth mentioning are the Capon beamformer (1969) [12], the Frost beamformer (1972) [13] and the generalized sidelobe canceller (GSC) (1976) [14]. They have not necessarily been developed for microphone arrays, but they have all been used for this purpose at some point.

1.1.2 Microphone arrays in reverberating environments

For most purposes microphone arrays are used in a reverberating environment. This is an environment with non-absorbing materials, causing the sound to be reflected from surfaces. The reflections may go back and forth, creating a slowly attenuating sound field. A room is an example of a reverberating environment. Here the sound is reflected back and forth between the walls, floor, ceiling and other obstacles. Figure 1.2 shows how the sound travels from the source and via some of the reflections to the microphones. The DAS beamformer can be steered towards a source, see

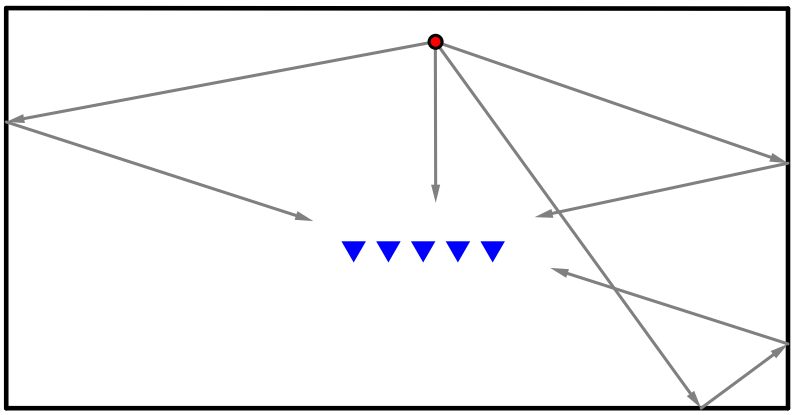


Figure 1.2: Propagating sound (grey arrows) emitted from a source (red dot) received by array microphones (blue triangles) in a two dimensional room (black square). The figure only shows some of the reflections.

figure 1.1. Reflections reach the array from other directions, outside the beam in the beampattern. The beamformer suppresses the reflections, as it would do for other sound sources located in the reflection positions.

1.2 Research question

The goal of the work presented in this thesis is to create an algorithm for better speech enhancement using microphone arrays. The new method, called constructive reflections method (CRM), will be using beams steered towards the reflections in combination with the beam steered directly at the source. The outputs of the beams will be delayed according to the differences in travel path between the direct sound and the reflections, and the summed together. The research question for this thesis then becomes:

- Will the new method improve the speech intelligibility compared to the DAS beamformer?
  - If so, how much will the improvement be compared to DAS beamformer?
  - Under which conditions will the new method be a better choice of beamformer than DAS?
  - Under which conditions will DAS be the better beamformer?

1.2.1 Definitions for the signal model

The recorded sound from a microphone consists of several components. These components are defined as following:

- "Signal" means the direct sound from a source of interest. It appears on the recording as an attenuated and delayed version of the emitted signal.
- "Reverberation" means the sound emitted from the same source but reflected from walls, ceiling, floor or other reflecting surfaces. Other names for this are *echoes*, *reflections* and *room effects*. The key feature of the reverberation is that it reaches the microphones at a different time than the direct signal.
- "Interference" means the sound from other sources than the source of interest, including the reflections of these sources.
- "Noise" mean sound that is spatially white, i.e. it is uncorrelated for spatially separated microphones.

1.2.2 Metrics of performance

The *signal-ratios*; signal to reverberation ratio (SRR), signal to interference ratio (SIR) and signal to noise ratio (SNR), are chosen as metrics of performance for this thesis. In research on microphone arrays these are the most common metrics. Although the signal-ratios are important for speech intelligibility, there are other metrics that also analyze the "shape" of the reverberation and frequency content

of the noise and interference. These are more suitable for measuring the speech intelligibility in a concert hall, PA system or other scenarios.

1.2.3 Hypotheses

As for the DAS algorithm, the reflection beams will be delayed so that the reflected signals are aligned with the direct signal. This will cause constructive interference. The reflection beam's output will be a mix of signal, reverberation, interference, and noise. It is expected that:

1. The proposed method has higher signal-ratios than DAS.
2. The proposed method will remove the room effects from the emitted source signal. This will be apparent in a smaller deviation from a flat frequency response compared to DAS beamformer.

It is also expected that the selection of reflections used in CRM greatly affects the performance. The hypotheses are that:

3. Some reflections will contribute to higher SRR, some will contribute to a lower SRR.
4. Some reflections will contribute to higher SIR, some will contribute to a lower SIR.
5. All reflections will contribute to a lower SNR

To deal with the problem of choosing the right reflections, an adaptive method for weighting the reflections will be proposed. The algorithm will be based on a minimization of the output power with a distortionless response criterion. This method is called the minimum power/distortionless response CRM (MVDR-CRM). It is expected that:

6. The adaptive algorithm will give better and more stable signal-ratios than the non-adaptive one.

1.3 Research methods

The main research method selected for this theses is simulations. This allows for extensive testing of many different scenarios compared to measurements which that each scenario is constructed.

1.3.1 Literature search

A literature search was conducted to see how other methods deals with reverberation. The searches where done in all the *IEEE* journals and the *Journal of Acoustical Society of America*. Speech enhancement with microphone arrays is a fairly new concept. The period of interest stretches from the 90's up to today.

Most of the methods regard the reverberation as stationary interference. The reverberation is partly removed from the array output by suppressing it using a beamformer. There are some who study the effect of it, and others who cancel it out using inverse filtering or echo cancellation. Only one reference was found to a method that uses the reflections constructively: The SCENIC-project state "*The acoustic Rake receiver ... aims to add coherently the reflected signals to the direct-path signal in order to improve the output signal-to-noise ratio (SNR).*"[15] There is a reference to a paper: "Method for dereverberation and noise reduction using spherical microphone arrays" by Peled and Rafaely. Here they do this in the spherical harmonics domain and finds that the method has a "*significant dereverberation and noise reduction*" [16, 17]. White noise is added to the microphone outputs.

This thesis presents a similar method but formulated in the time domain. Also this study investigates how a broader range of parameters affects the performance of the algorithm.

1.3.2 Simulations

Simulation software allows room impulse responses for rectangular rooms to be calculated. The simulation parts of the study will investigate how important parameters like the array size and reverberation time affects the proposed method. The main part of the research is done with simulations.

1.3.3 Measurements

Measurements are performed to validate the simulations. An array provided by Squarehead Technology AS [3] consisting of 256 elements is used in an empty rectangular room.

1.3.4 Listening tests

It is possible to perform listening tests on the data acquired from the simulations or measurements. It requires many observers to produce a statistical significant answer, which is demanding to realize. For this thesis it has been chosen not to perform listening tests.

1.4 Outline of thesis

The structure of the thesis is a slightly modified version of the IMRaD[6] outline (Introduction, Method, Results [and] Discussion). There is one extra chapter presenting the mathematical definition and derivations for the proposed methods. The results are extensive and discussion is interleaved with these results. To avoid repetition or extensive use of references, the results and the discussions are combined in one chapter.

Chapter 2: ***Background*** lays the mathematical foundation for the coordinate system, the DAS beamformer, and the signal model. At the end of the chapter, there is a discussion of the differences between common speech enhancement methods.

Chapter 3: ***Method*** is divided into four sections. It starts with a presentation go the relevant assumptions. The performance metrics are described and precisely formulated. After that there are two sections containing information about the simulation experiment and the real-world experiment.

Chapter 4: ***The proposed beamformers*** presents the two developed methods: constructive reflections method (CRM) and minimum power/distortionless response CRM (MVDR-CRM). The theoretical array gain for DAS and CRM beamformers are presented and compared.

Chapter 5: ***Results and Discussion*** presents and discuss the important findings from the simulations and the measurements. The beamformers are tested under various conditions to get a good overview of their strengths and weaknesses. An audio sample from the different beamformers is available online.

Chapter 6: ***Review, Summary and Further Work*** sums up the results and discussion in light of the hypothesis (presented in chapter 1) and the assumptions (presented in chapter 3).

Appendix A: ***MATLAB source code*** shows the MATLAB code for the implementation of CRM and MVDR-CRM including supporting functions.

Appendix B: ***Position of sources*** offers a description of 500 random source and interference positions used in some of the simulations. An URL to a MATLAB file containing the positions is available.

[6]More can be read in the Wikipedia article on IMRaD; wikipedia.org/wiki/IMRaD.

Chapter 2

Background

2.1 Coordinate systems

Both a room-centered and an array-centered coordinate system is needed. The simulations are done using a room-centered Cartesian coordinate system. Important array features, like beamwidth and beampatterns, are often expressed using spherical coordinates centered in the phase center of the array. This section provides a definition of the two coordinate systems and the translation between them. First the general spherical and Cartesian coordinate system is explored.

Figure 2.1 shows a point defined in the two coordinate systems.

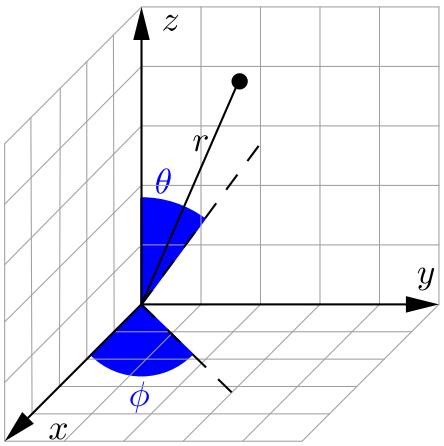


Figure 2.1: A point defined in the Cartesian and spherical coordinate system.

The translation between Cartesian ($\vec{x} = (x, y, z)$) and spherical ($\vec{\Omega} = (r, \theta, \phi)$) coordinates is in defined in *ISO 80000-3:2006 Quantities and units Part 3: Space and time*. The transformation is done with the following operations:

$$r = \sqrt{x^2 + y^2 + x^2} \tag{2.1}$$

$$\theta = \cos^{-1}\left(\frac{z}{r}\right) \tag{2.2}$$

$$\phi = \tan^{-1}\left(\frac{y}{x}\right) \tag{2.3}$$

The transitions from spherical to Cartesian coordinates are:

$$x = r\sin\theta\cos\phi \tag{2.4}$$

$$y = r\sin\theta\sin\phi \tag{2.5}$$

$$z = r\cos\theta \tag{2.6}$$

2.1.1 Room-centered coordinate system

For box-shaped rooms, the origin is placed in one of the corners. The *west*, *east*, *south* and *north* walls and *ceiling* and *floor* are defined with this origin as the basis. Figure 2.2 shows how this is defined.

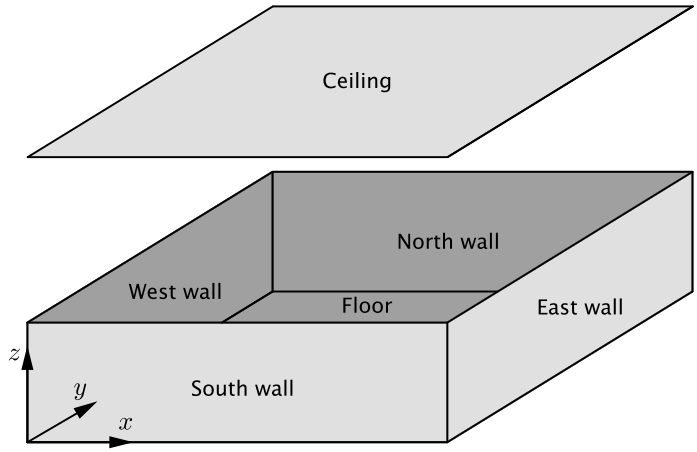


Figure 2.2: The different surfaces in a box-shape room using a room-centered Cartesian coordinate system.

The room-centered coordinates is denoted by:

$$\vec{x} = (x,\, y,\, z),\, \vec{\Omega} = (r,\, \theta,\, \phi)$$

For other room geometries, the origin and orientation can be arbitrarily chosen.

2.1.2 array-centered coordinate system

An array must have a defined *phase center* ($\vec{x}_{phase\,center}$), *front side direction*, and *top side direction* in order to define an array-centered coordinate system. The top side direction must be normal to the front side direction. Figure 2.3 shows an array positioned in the two coordinate systems.

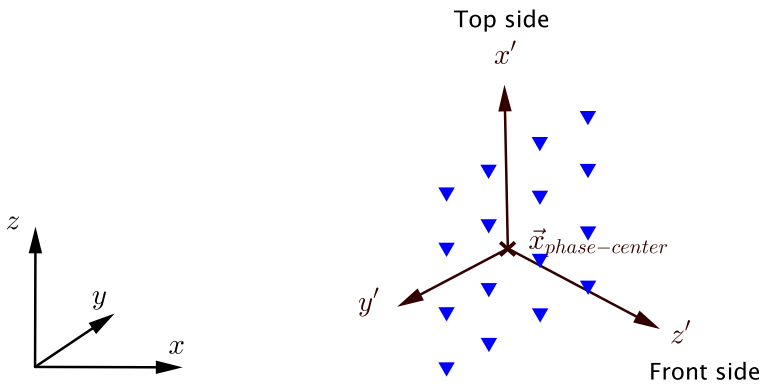


Figure 2.3: An array (blue triangles) in the room-centered coordinate system (x, y, z) and the array-centered coordinate system (x', y', z').

The *front side direction* is parallel to the z'-axis and the *top side direction* is parallel to x'-axis. The array-centered coordinates are marked with $'$:

$$\vec{x}' = (x',\, y',\, z'),\, \vec{\Omega}' = (r',\, \theta',\, \phi')$$

The origin is positioned in the phase center of the array:

$$\vec{x'}_{phase\,center} \triangleq (0, 0, 0) \tag{2.7}$$

The room-centered coordinate for the phase center is the mean position vector for the microphones in the array:

$$\vec{x}_{phase\,center} = \frac{1}{M} \sum_{m=0}^{M-1} \vec{x}_m \tag{2.8}$$

2.1.3 Transformation between the coordinate systems

The transformation from $\vec{x}$ (room-centered) to $\vec{x}'$(array-centered) consists of

- a rotation around the x axis with the rotation γ
- a rotation around the y axis with the rotation β
- a rotation around the z axis with the rotation α
- translation from $(0, 0, 0)$ to $\vec{x}_{phase\,center}$

The array-centered coordinates as a function of the room-centered:

$$\vec{x}' = \vec{x}\,\boldsymbol{R}_x(\gamma)\,\boldsymbol{R}_y(\beta)\,\boldsymbol{R}_z(\alpha)\,\boldsymbol{T}(\vec{x}_{phase\,center}) \tag{2.9}$$

And the room-centered coordinates as a function of the array-centered:

$$\vec{x} = \vec{x}'\ (\boldsymbol{R}_x(\gamma)\,\boldsymbol{R}_y(\beta)\,\boldsymbol{R}_z(\alpha)\,\boldsymbol{T}(\vec{x}_{phase\,center}))^{-1} \tag{2.10}$$

where

$$\boldsymbol{R}_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}$$

$$\boldsymbol{R}_y(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}$$

$$\boldsymbol{R}_x(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{T}(\vec{x}) = \begin{bmatrix} 1 & 0 & 0 & \vec{x}_x \\ 1 & 0 & 0 & \vec{x}_y \\ 1 & 0 & 0 & \vec{x}_z \end{bmatrix}$$

2.2 Wave propagation

This section gives a brief overview of some selected parts of wave physics that are important for sound waves.

2.2.1 Wave equation

The behavior of waves is governed by the wave equation:

$$\frac{\partial^2 p}{\partial \vec{x}^2} = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{2.11}$$

where p is the pressure and c is the wave speed. A general solution[1] to the wave equation is the monochromatic wave[2]:

$$p(\vec{x}, t) = A\exp\left\{i(\omega t - \vec{k}\vec{x})\right\} \tag{2.12}$$

where A is the amplitude of the wave, i the imaginary unit, ω the (temporal) frequency and $\vec{k}$ the wave number.

2.2.2 Reflection and transmission

When a sound wave in one medium travels to another medium, both reflection and transmission can occur, see figure 2.4. The angle (θ_1) of the incoming wave is equal to the angle of the reflected wave. The angle of the transmitted wave (θ_2) follows Descartes's law of sines:

$$\frac{\sin\theta_1}{c_1} = \frac{\sin\theta_2}{c_2} \tag{2.13}$$

where c_1 and c_2 are the wave speeds in the two mediums.

The amount of reflection and transmission is dependent on the incident angle of the sound wave and the properties of the two mediums. The reflection coefficient is defined by the Fresnel equation:

$$R = \left(\frac{Z_2\,\cos\theta_1 - Z_1\,\cos\theta_1}{Z_2\,\cos\theta_2 + Z_1\,\cos\theta_2}\right)^2 \tag{2.14}$$

where $Z = \rho c$, θ is the angles (see figure 2.4) and the subscripts $_1$ and $_2$ denotes the two media. ρ is the density of the material and c is the wave speed. If the next material is denser than the previous ($c_2 > c_1$), the reflected wave is phase shifted by 180°.

[1] Fourier series shows that any continuous signal can be constructed by a sum of monochromatic waves with different frequencies and phases.

[2] A *monochromatic wave* is a wave consisting of only one frequency

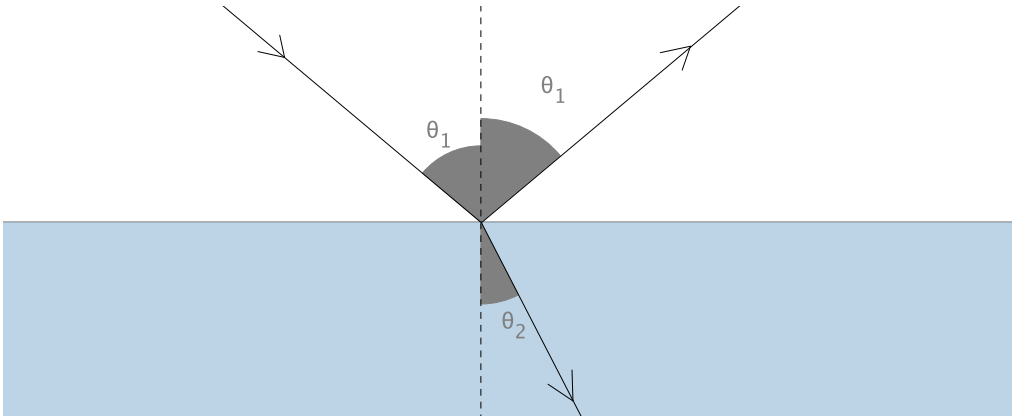


Figure 2.4: The angle of an incoming, reflected and transmitted wave.

2.2.3 Diffraction

Waves that encounter obstacles can bend around them. Wright gives a more precise definition in "Fundamentals of Diffraction": "*Diffraction is the change in direction of propagation of a wave front due to the presence of an obstacle or discontinuity (with no change in velocity)*" [19]. Diffraction only occurs when the size of the object is at the same order of size as the wavelength. Figure 2.5 shows the diffraction of a wave encountering a slit.

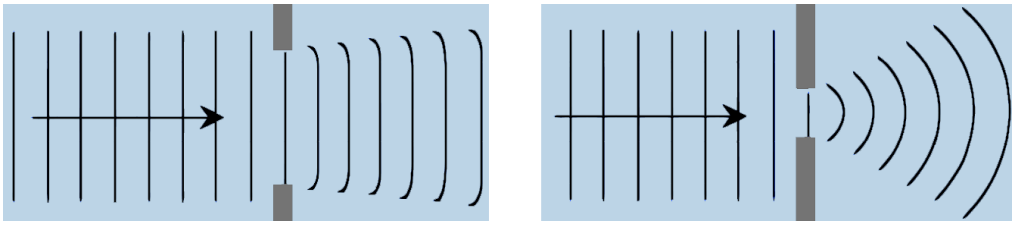


Figure 2.5: Waves encountering a slit. Left: the slit is large compared to the wavelength. Right: the slit is small compared to the wavelength. Source: SMK Bandar Utama Damansara

Diffraction is a "low frequency" phenomenon. If we assume that a signal only contains waves with small wavelength compared to the geometries of the environment, we can neglect the effect of diffraction.

2.3 Sound and acoustics

The earth's atmosphere holds a normal pressure around 101.325 kPa. The pressure is depending on temperature, height above sea level and the composition of the air [20]. Audible sound waves are small perturbations in this pressure with frequencies from 20 Hz up to 20 kHz.

2.3.1 The decibel scale

The decibel scale is often used to describe the amplitude of sound pressure.

$$L_p = 10 \log \left(\frac{p_{rms}^2}{p_0^2} \right) \tag{2.15}$$

The decibel scale is a relative scale. A reference level always needs to be specified. In this case the reference value is $p_0 = 20\mu$Pa. p_{rms} is the root mean squared pressure. This is calculated using the observed pressure $p(t)$ for a time period of T:

$$p_{rms}(t) = \sqrt{\frac{1}{T} \int_{t-T/2}^{t+T/2} p(t)^2 dt} \tag{2.16}$$

The lowest audible sounds are around 0 dB (20μ Pa), while exposure to sound at a 140 dB (200 Pa) will give immediate damage to the ear [21].

2.3.2 Room acoustics

Room acoustics is the theory of how sound behaves in a room. When sound is travelling from a source to a receiver, the effect of a room can be viewed as a finite impulse response (FIR) filter. The direct sound from the source to the receiver will be a delayed and attenuated version of the emitted signal, while the reflections from walls and other obstacles are delayed and attenuated versions of the direct signal.

2.3.3 Room impulse response

As for any filter, the room filter is characterized by its impulse response (IR) called the room impulse response (RIR). For a source q and a receiver m we define the impulse response as $h_{q,m}(t)$. Figure 2.6 shows a simple RIR.

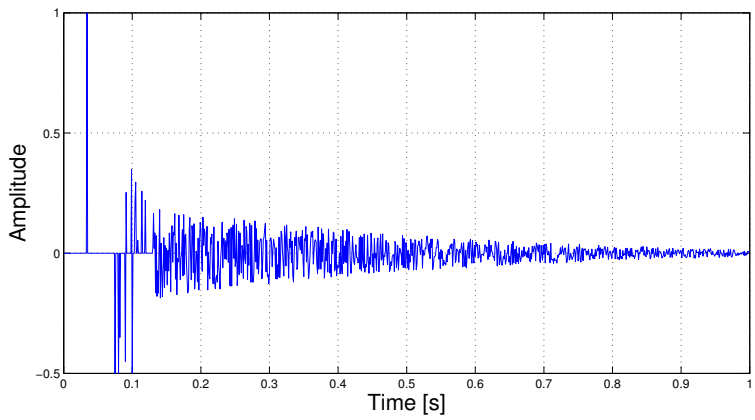


Figure 2.6: A simple sketch of a RIR.

The different parts of the RIR are:

- The first spike in the RIR is the direct sound. It is delayed by the time the sound uses to travel from the source to the receiver.

- The first spikes following the direct sound are the early reflections. Plotting the RIR with a high enough sampling rate, they can often be visually separated from the rest of the RIR. They originate from sound traveling via one or more reflecting surfaces to the receiver. They will have lower amplitudes than the direct sound due to geometrical spreading, absorption at the reflecting surfaces and attenuation in the air [22]. The reflected sound has traveled longer than the direct sound, so the spikes will appear after the direct sound. Reflections at hard surfaces gives a phase shift. If a phase-shifted signal encounters a second reflection at a hard surface, the signal is shifted back again. The early reflections can therefore be seen as both negative and positive spikes in the RIR

- When the sound is reflected many times, the spikes can not be separated from each other. This is the reverberation field. There is no hard limit between the early reflections and the reverberation field. The reverberation field will be decaying due to the loss of energy at each reflection. To describe this decay the reverberation time is used. It is defined as the time it takes for the reverberation sound to drop to -60dB compared to the direct sound, T_{60}.

Dereverberation The process of removing the effect of the RIR is often called *dereverberation*. A completly dereverberated RIR will only consist of single impulse.

Order of reflection is the number of reflections that have occurred before the sound reaches the microphone.

2.3.4 Absorbing materials

The materials of the walls have a great impact on the amount of sound that is reflected. This will directly affect the reverberation time. To control the acoustics in a room it is common to install absorbing materials on the walls and ceiling. A hard surface, made of concrete, metal or thick glass, reflects almost 100% of the sound energy. A soft material, like a curtain or a padded chair, would absorb some of the sound energy. The absorption is in most cases frequency dependent. A common measure of the absorption is the absorption coefficient, defined as:

$$\alpha = \sqrt{1 - R^2} \tag{2.17}$$

where R is the amplitude of the reflected sound compared to the amplitude of the incoming sound. Lab measured absorption values for some common building materials are listed in table 2.1.

	Absorbtion value					
Materials/Frequency band	**125**	**250**	**500**	**100**	**2000**	**4000**
Glass, window pane	0.35	0.25	0.18	0.12	0.07	0.04
Wooden walls	0.14	0.10	0.07	0.05	0.05	0.05
Concrete	0.01	0.01	0.02	0.02	0.02	0.02
Plywood	0.60	0.30	0.10	0.09	0.09	0.09
Acoustic tile in ceiling	0.10	0.25	0.55	0.65	0.65	0.60
Seat with cloth-cover	0.20	0.35	0.55	0.65	0.60	0.60

Table 2.1: Some selected absorption values for common building materials [23].

2.4 Speech and hearing

2.4.1 Speech

The human speech is generated by air flowing from the lungs and to the mouth, see figure 2.7. The body uses muscles in the chest, diaphragm and stomach to push the air out from the lungs. The air travels past the larynx and vocal chords. These organs modulate the flowing air with vibrations. This is where the body regulates the pitch or tone of the voice. The air goes through the mouth and the nasal cavity. These organs can be opened and closed, changing their internal volume or shape, and thereby creating an acoustical filter. The different vowels are made here. The tongue, teeth, and lips generate the consonants by closing and opening the mouth, creating turbulence through the teeth or tongue. The result is sound waves propagating from the mouth that we observe as speech [23]. The speech spectrum, see figure 2.8, is

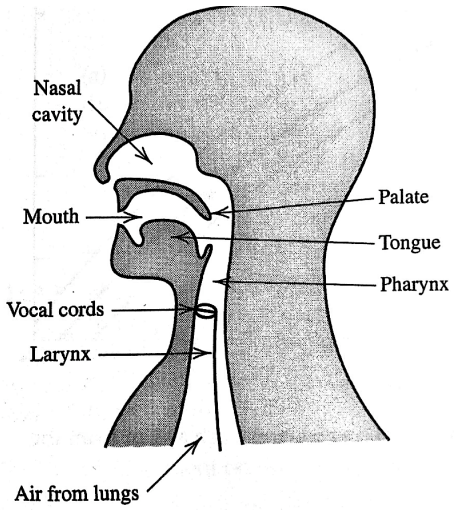


Figure 2.7: A sectional view of the head, showing the important elements of the voice mechanism. Source: Fundamentals of Acoustics *[23].*

mainly between 100Hz to 3500 Hz with the peak frequency around 300 Hz. The female voice has higher pitch than the male voice [23].

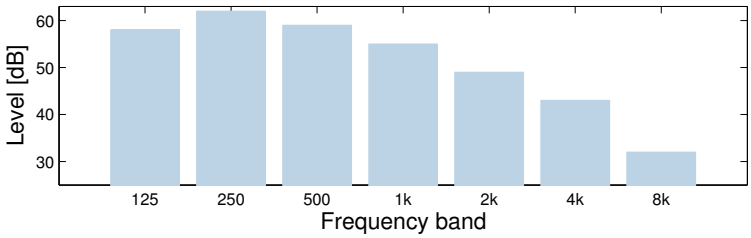


Figure 2.8: Long-term male speech spectrum. The plot is reproduced from a plot found in Speech Intelligibility - A JBL professional Technial Note *[24].*

2.4.2 Hearing

The human perception of sound happens in the inner and outer ear. It is a complex process and more details can be found in [23]. The head has two ears and the human brain uses these to achieve directional hearing. Shadowing from the head and the relative time difference between each ear are used to locate where sounds come from in the horizontal plane.

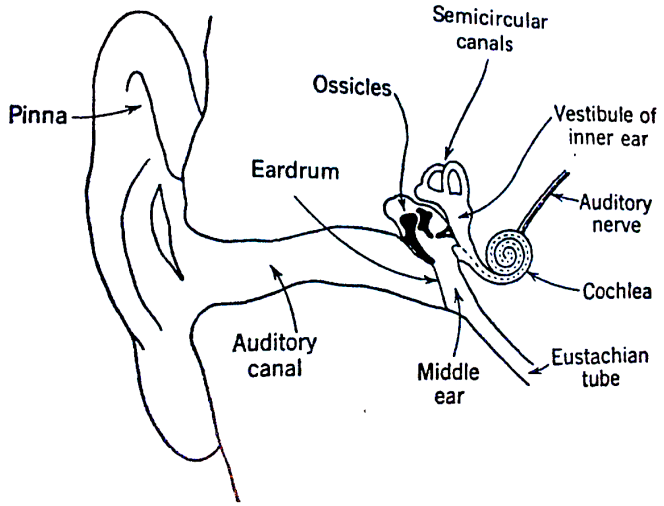


Figure 2.9: Sketch of the ear. Source: Fundamentals of Acoustics *[23].*

In each ear the sound roughly goes through three stages, see figure 2.9:

- The sound is diffracted and reflected in the outer ear (pinna and the auditory canal). The sound changes frequency content depending on the direction of the sound. This makes it possible to detect where the sound is coming from in the vertical plane.

- In the middle ear (eardrum and ossicles) the sound energy is transformed to mechanical vibrations.

- In the inner ear the vibrations are lead to the basilar membrane in the cochlea. This membrane resonates at different spatial positions according to the frequency content of the sound. Nerves detect the vibrations at the different locations on the membrane and send nerve signals to the brain.

To compare how sound levels affect human hearing, it is common to use weighting filters where the frequency components are weighted according to the sensitivity of the ear. A weighting is the most common psychophysical weighting for audio.

2.4.3 Speech intelligibility

Speech intelligibility is defined as the percentage of correctly observed words or sentences. Figure 2.10 shows how the intelligibility increases when the A-weighted SNR increases. Figure 2.11 shows which frequency bands are most important for intelligibility. The speech spectrum consists of a considerable amount of low frequency components that do not contribute much to increased speech intelligibility, see figure 2.8.

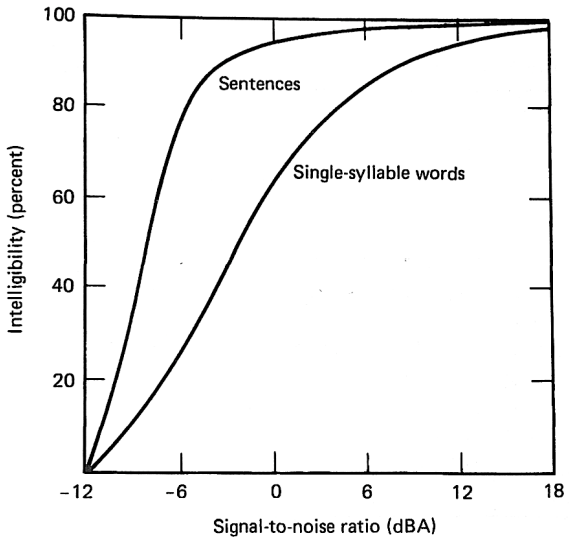


Figure 2.10: Speech intelligibility as a function of signal-to-noise ratios. Source: Fundamentals of Acoustics *[23].*

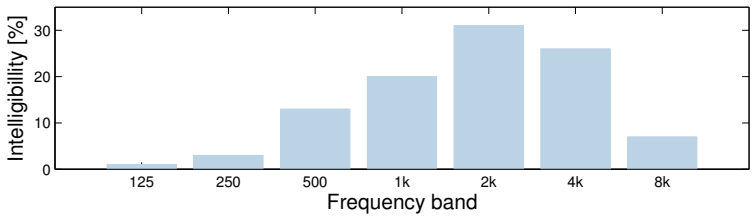


Figure 2.11: Octave-band contributions to speech intelligibility. The plot is reproduced from a plot found in Speech Intelligibility - A JBL professional Technial Note *[24].*

2.5 Delay and sum (DAS) beamformer

The signal emitted from a source will arrive at each (m) element in the array with a slightly different time delay τ_m (figure 2.12).

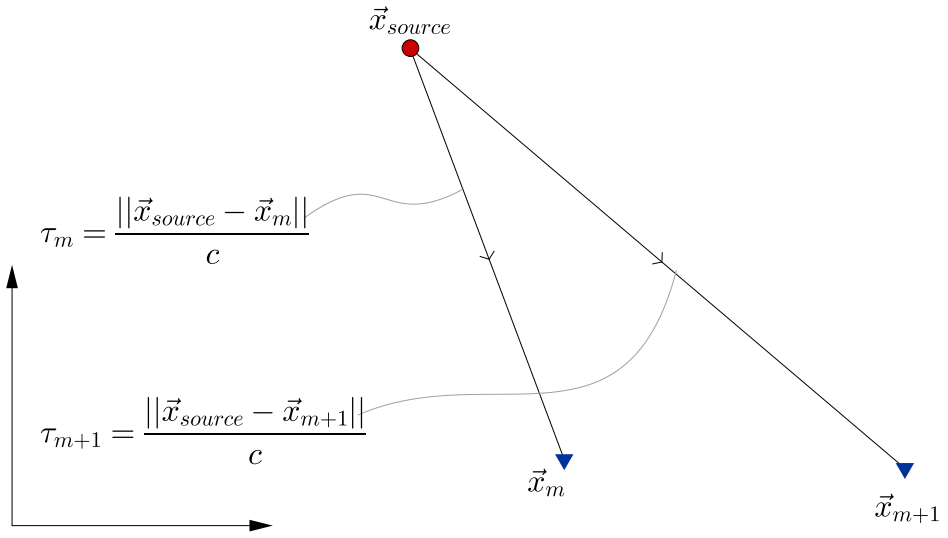


Figure 2.12: The time the sound uses from a source (x_{source}) to a receiver (x_m) based on the length of the travel path where c is the wave speed.

This corresponds to the distance between each sensor and the source. If the source emits a signal $s(t)$, assuming that there is no air absorption and both the source and sensors are *omnidirectional*[3], the output at the m-th sensor will be:

$$y_m(t) = \frac{1}{||\vec{x}_m - \vec{x}_{source}||^2} s(t - \tau_m) \tag{2.18}$$

where $\vec{x}_m$ is the position of the element and $\vec{x}_{source}$ is the position of the source.

2.5.1 Mathematical formulation

The DAS beamformer first delays the signal at each sensor to contract the propagation delay:

$$\tau_m = \frac{||\vec{x}_m - \vec{x}_{source}||}{c}$$

This makes the signal from the source align on all the microphone outputs. The beamformer sums the outputs, and the signals are added coherently. Signals emitted from sources at different locations will be added incoherently. Each sensor could

[3] *Omnidirectional* means that the sensors have the same sensitivity to waves arriving from all directions. When the word, *omnidirectional*, is used about sound sources, it means that the radiate the sound equally in all directions.

also be weighted differently, with the weights w_m. The output of the weighted DAS beamformer is defined as:

$$z(t) = \sum_{m=0}^{M-1} w_m y_m(t + \tau_m) \tag{2.19}$$

Normalized weights The amplitude of the signal component on the DAS output should be equal to the amplitude of the signal component on a single microphone. This is what is called *distortionless response* and makes the outputs directly comparable. To achieve this, the weights are normalized by dividing each weight by the sum of all the weights so that $\sum_m w_m = 1$.

Front and top side directions For a general array the front side and top side directions can be freely chosen, under the constraint that they are normal to each other. For some common array geometries it makes sense to define them. Figure 2.13 shows how these array properties are defined for a uniform linear array (ULA) and a uniform rectangular array (URA).

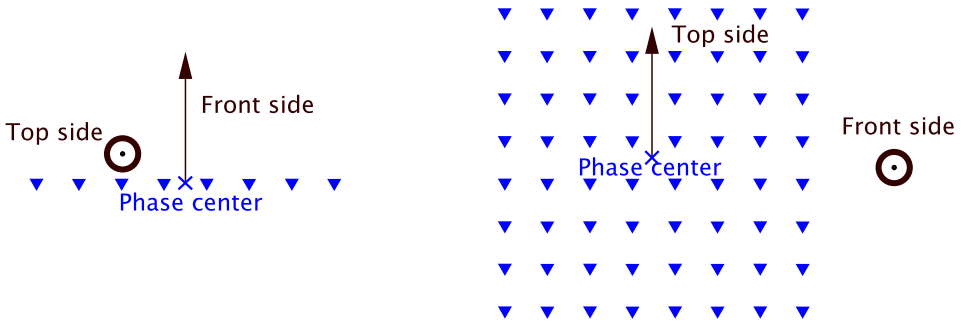


Figure 2.13: The front side and top side directions for a ULA (left) and URA (right). The microphone elements are marked with blue.

Array size The array size, D, determines many important features of the beamformer. For a ULA the array size is simply the length of the array. For arrays in 2D/3D, the length and geometries in all the directions are important.

Near-field far-field limit It is common to separate between near-field and far-field. In the near-field the wave fronts are observed as curved over the array, see figure 2.14. This happens when the source is close to the array. When the source is far away from the array, the wave fronts become more and more planar over the

array. The far-field is when the waves are approximately planar within a tolerance depending on the application. The far-field limit is given by:

$$r > \frac{D^2}{n\lambda} \tag{2.20}$$

where $n = 1, 2, 4$ and D is the array size. The variable n depends on the chosen tolerance: $\frac{1}{16}\lambda$, $\frac{1}{6}\lambda$, or $\frac{1}{4}\lambda$.

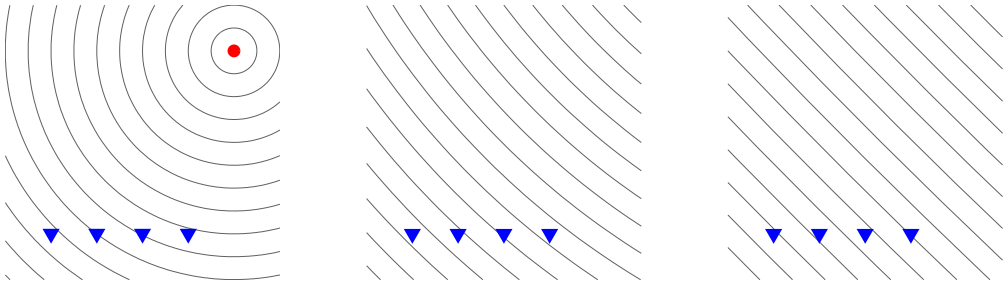


Figure 2.14: A source (red dot) radiates the signal, the wave fronts are visualized with the gray lines, observed by an array (blue triangles). Left: the source is in the near-field. Middle: the source is somewhere between the near- and far-field. Right: the source is in the far-field.

2.5.2 Beampattern

The beampattern describes how the signal is amplified as a function of the direction of the incoming signal. It is calculated by the formula:

$$W(\vec{k}) = \sum_m w_m e^{i\vec{k}\cdot\vec{x}_m} \tag{2.21}$$

where $\vec{k}$ is the wavenumber vector, m is the sensor index and $\vec{x}_m$ is the distance from each sensor to the origin of an arbitrarily chosen coordinate system. The beamformer creates a spatial *filter*, and the weights can be used to manipulate the beampattern in the same manner as filter coefficients are used to manipulate an ordinary filter. The beampattern for a ULA is a sinc, see figure 2.15.

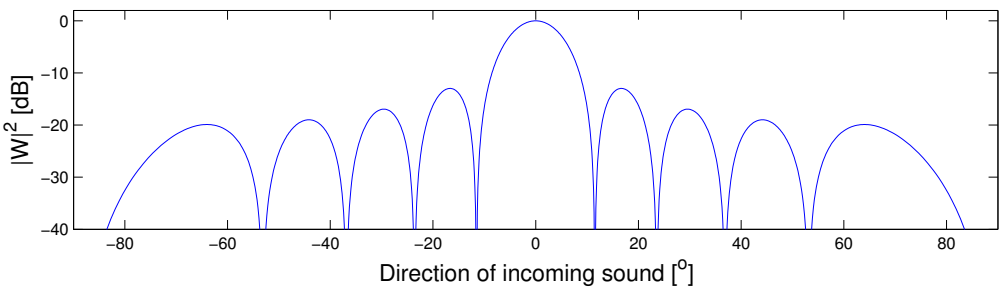


Figure 2.15: The beampattern for a DAS beamformer on a ULA with 10 mic. $d = \frac{\lambda}{2}$

Zeros There are dips in the beampattern. At these locations there is full cancellation of the incoming signal. This is called "zeros".

Mainlobe - sidelobe ratio There are two important properties of a beampattern; one is the mainlobe - sidelobe ratio. This tells us how much the system amplifies the source of interest compared to interfering signals. Figure 2.16 shows how the mainlobe - sidelobe ratio can be found from the beampattern.

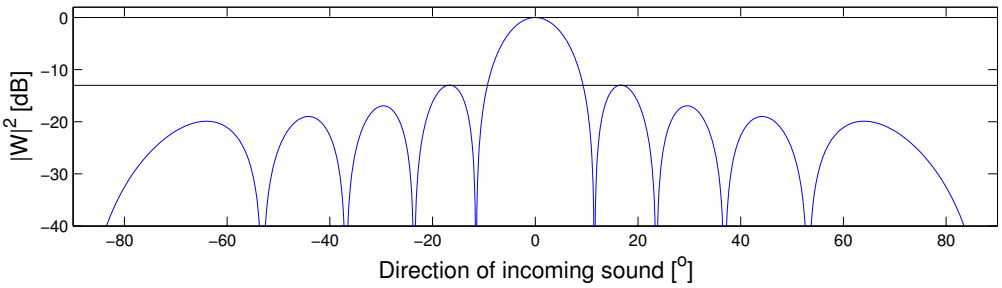


Figure 2.16: The main-lobe - side-lobe ratio is visualized by the distance between the two black lines, here 13 dB.

Beamwidth The other important feature of the beampattern is the width of the main lobe at -3db (half maximum in non dB-scale). This gives the spatial resolution of the beamformer, as shown in figure 2.17. The resolution depends on the geometry of the array, but for regular arrays the 3dB beamwidth can be approximated by:

$$\sin(\theta) \approx \frac{\lambda}{D} \tag{2.22}$$

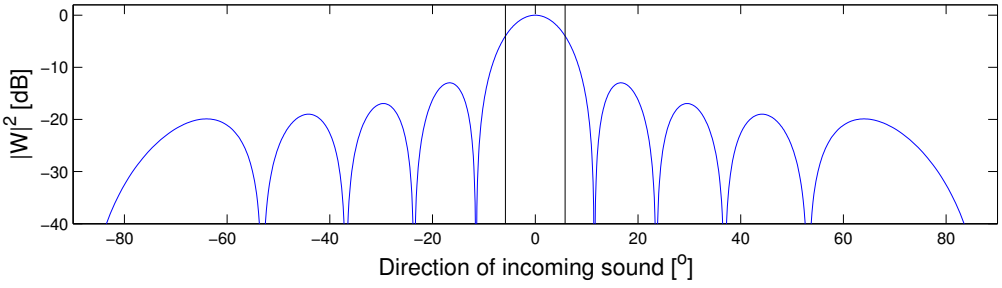


Figure 2.17: The -3dB beamwidth is the distance between the two black lines.

Grating lobes If the wavelength is less than twice the spacing between the array elements, spatial aliasing will occur. This can be observed as grating lobes in the beampattern, see figure 2.18. It means that the sound arriving from the location of a grating lobe will be "leaking" into the output unsuppressed. The system has

to follow the Nyquist sampling criterion to avoid grating lobes in the beampattern, where λ_{min} is the smallest wavelength in the observed signal:

$$\lambda_{min} \geq 2d \tag{2.23}$$

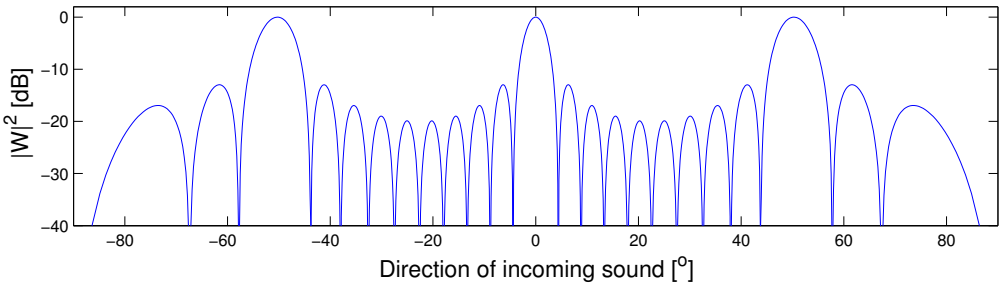


Figure 2.18: Grating lobes in the beampattern. $d = 1.3\lambda$

2.5.3 Vector notation

Consider the output of the DAS beamformer, equation 2.19. When the sensor outputs y_m are properly delayed, the beamformer can be written using vector notation:

$$z(t) = \boldsymbol{w}^H \boldsymbol{y}(t) \tag{2.24}$$

where

$$\boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{M-1} \end{bmatrix}, \qquad \boldsymbol{y}(t) = \begin{bmatrix} y_0(t + \tilde{\tau}_0) \\ y_m 1(t + \tilde{\tau}_1) \\ \vdots \\ y_{M-1}(t + \tilde{\tau}_{M-1}) \end{bmatrix}$$

The power output of the DAS beamformer is the absolute value squared:

$$P = |z|^2 = zz^* = \boldsymbol{w}^H \boldsymbol{y}\boldsymbol{y}^* \boldsymbol{w} = \boldsymbol{w}^H \boldsymbol{R}\boldsymbol{w} \tag{2.25}$$

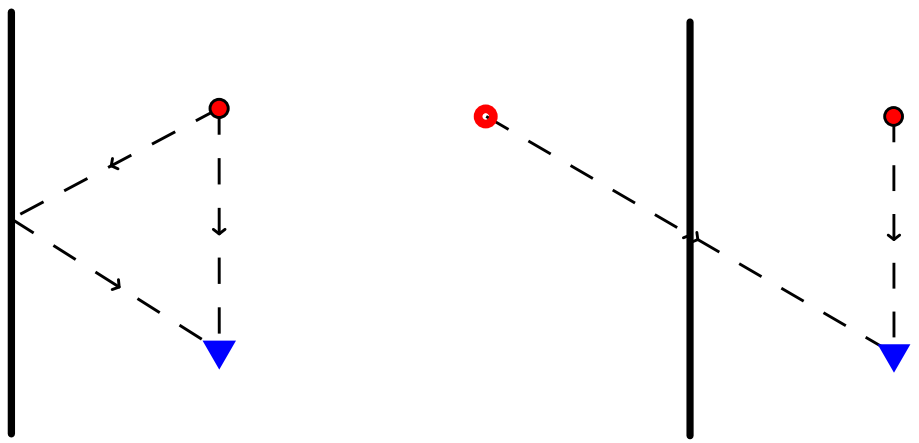
where R is the covariance matrix of $\boldsymbol{y}$.

2.6 The reverberant signal model

2.6.1 Image source method

The image source method for room acoustics was proposed by Allen and Berkley in April 1979 [25]. It is also widely used in other areas of physics. It provides a conceptually easy and computationally fast method for simulating room acoustics. The drawback is that it does not model diffraction, i.e. it is a high frequency approximation.

The main idea is that a source and a reflection can be seen as a source and its mirror image source, see figure 2.19. The source mirrored over the reflecting surface is the image source. It emits a 180° phase shifted version of the signal from the original source. The signal is attenuated by the absorption in the air and the reflecting surface. In addition the image source will be placed further away from the receiver, and experience more attenuation due to geometrical spreading.



(a) A receiver (blue triangle) observing the direct and reflected signal from the source (red dot).

(b) A receiver (blue triangle) observing a source (red dot) and the first order image source (hollow red dot).

Figure 2.19: A simple image source method example.

If there are multiple reflecting surfaces, there will be correspondingly many first order reflections. Each reflection can be described by an image source of the source mirrored over the reflecting surfaces. Second order reflections are modeled by image sources of the first-order image sources. Third order reflections are image sources of the second order image sources, and so on. The signal is phase shifted 180° for each reflection. The observed signal $y_m(t)$ will be the sum of all the J image sources (including the original source), emitting the signal $s(t)$:

$$y_m(t) = \sum_{0}^{J-1} p_j A_{j,m}\, s(t - \tau_{m,j}) \tag{2.26}$$

where $\tau_{m,j}$ is the delay between the m-th sensor and the j-th image source. p_j is the phase shift associated with the j-th image source and the phase shift p_j will either be 1 or -1. $A_{j,m}$ is the amplitudes of the signal observed at the microphones, it consist of the attenuations in the reflections R_j and the loss due to geometrical spreading. We assume that there is no air absorption. The amplitude factors becomes:

$$A_{j,m} = \frac{R_j}{\left(||\vec{x_m} - \vec{x_j}||\right)^2} \tag{2.27}$$

If the image sources is far away from the array, compared to the array size, the amplitudes can be considered as equal on all elements, $A_{j,m} \approx A_j$. We define the collection of J image sources to be unordered except for $j = 0$ which is the original source. Theoretically $J = \infty$. Since the reflection amplitudes drop, we can consider only a finite number of image sources, while neglecting the rest.

A visual impression of the first and second order image sources in a 2D room are shown in figure 2.20. The figure shows that as the order of reflections increases, the image sources are spread around in the room.

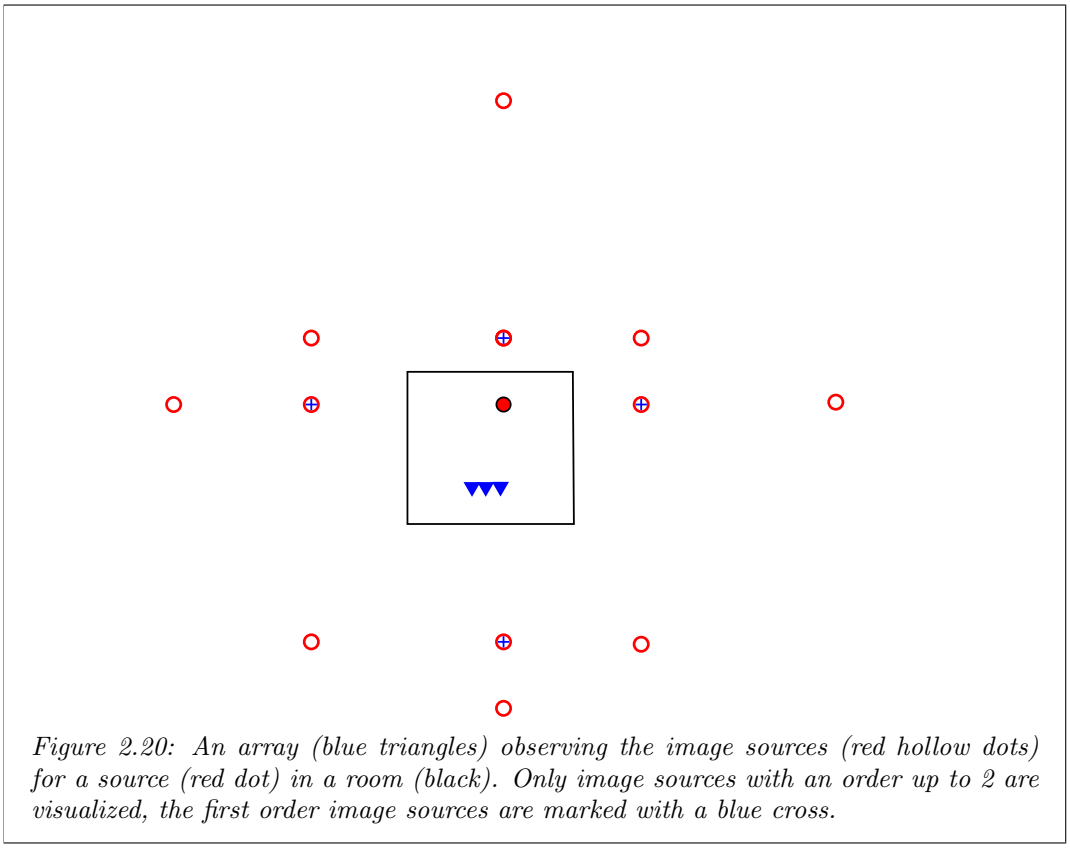


Figure 2.20: An array (blue triangles) observing the image sources (red hollow dots) for a source (red dot) in a room (black). Only image sources with an order up to 2 are visualized, the first order image sources are marked with a blue cross.

2.6.2 Signal model

Recall the definitions of the different components of the sound-field offered in section 1.2.1. The output y_m of a microphone element m consists of four components:

- signal (y_m^s) - the direct sound from the source of interest
- reverberation (y_m^r) - the reflections of the source of interest (the image sources)
- interference (y_m^i) - interfering sources, including their reflections
- noise (y_m^n) - spatially white noise

$$y_m(t) = y_m^s(t) + y_m^r(t) + y_m^i(t) + y_m^n(t) \tag{2.28}$$

The signal is the direct sound:

$$y_m^s(t) = A_0 s_0(t - \tau_m) \tag{2.29}$$

The reverberation can be expressed using the image source method (equation 2.26):

$$y_m^r(t) = \sum_{j=1}^{J-1} A_j \, p_j \, s_0(t - \tau_{m,j}) \tag{2.30}$$

and the sum of the signal and reverberation is the emitted sound $s_0(t)$ convolved with the RIR for the source of interest and the m-th element, $h_{0,m}(t)$:

$$y_m^s(t) + y_m^r(t) = s_0(t) * h_{0,m}(t) = \sum_{j=0}^{J-1} A_j \, p_j \, s_0(t - \tau_{m,j}) \tag{2.31}$$

The interference is the sum over all the other sources, convolved with their respective RIRs:

$$y_m^i(t) = \sum_{q=1}^{Q-1} s_q(t) * h_{q,m}(t) \tag{2.32}$$

2.6.3 Vector notation

By delaying the outputs with the delay $\tilde{\tau}_m$ and dropping the time-notation, the signal model can be expressed using vector notation:

$$\boldsymbol{y} = \boldsymbol{y}^s + \boldsymbol{y}^r + \boldsymbol{y}^i + \boldsymbol{y}^n \tag{2.33}$$

where $\boldsymbol{y} = \begin{bmatrix} y_0(t - \tilde{\tau}_m) \\ y_1(t - \tilde{\tau}_m) \\ \vdots \\ y_{M-1}(t - \tilde{\tau}_m) \end{bmatrix}$.

2.7 Concepts for speech enhancement

This section will give a short overview of common techniques for speech enhancement based on the signal model. The overview will be given on a conceptual level, but some references to more in-depth studies is presented. The perfect method will keep the signal and suppress everything else, so that the output becomes:

$$z = y^s + B(y^r + y^i + y^n) \tag{2.34}$$

where $0 \leq B << 1$.

2.7.1 Conventional beamforming

The first method is the DAS beamformer. It adds the signal from each output constructively and "smears" out the other components.

$$z = y^s + \boldsymbol{w}^H(\boldsymbol{y}^r + \boldsymbol{y}^i + \boldsymbol{y}^n) \tag{2.35}$$

The weighting function could be uniform or some sort of window function. Conventional (DAS) beamforming has proved to be a good and robust method for dereverberation [26, 27].

2.7.2 Adaptive beamforming

In adaptive beamforming the weights are used to create the filter that minimizes the reflections, interference and noise while keeping the signal intact. The beampattern is adapted to the particular sound field. The Capon and GSC beamformers are both variants of this technique [12, 14]. These methods achieve a large increase in the signal-ratios when there is no correlated sound present. Reflections from walls are correlated with the direct sound, causing trouble for the adaptive beamformers. There are several ways of improving the beamformers for correlated signals.

2.7.3 Inverse filtering the room impulse response

Inverse filtering is based on the fact that the signal and reflection are the emitted signal convolved with a filter, the RIR. If the RIR is known or is possible to estimate, then an inverse filter ($\hat{h}^{-1}_{0,m}$) can be created for each sensor to remove the effect of the RIR.

$$z = (y^s_m + y^r_m + y^i_m + y^n_m) * \hat{h}^{-1}_{0,m} \tag{2.36}$$

Full de-reverberation is hard and in some cases impossible to achieve with inverse filtering. Inverse filtering with respect to the RIR may have unknown effects on the noise and interference. If there is a high level of noise and interference, compared to the signal and reflections, this must be considered in the design of the filter.

Inverse filtering + beamforming It is possible to use an inverse filter on each element in a microphone array before the beamforming is applied. The output becomes:

$$z = \boldsymbol{w}^H \left((\boldsymbol{y}^s + \boldsymbol{y}^r + \boldsymbol{y}^i + \boldsymbol{y}^n) * \hat{\boldsymbol{h}}^{-1} \right) \tag{2.37}$$

where $\hat{\boldsymbol{h}}^{-1}$ is a vector containing all $\hat{h}^{-1}_{0,m}$. This has the advantage of removing some of the reverberation by inverse filtering, and then suppressing the remaining reverberation along with the interference and noise. A pre-requisition is that the inverse filter does not change the relative time-delays between the signals.

Beamforming + inverse filtering Another option is to use a filter on the beamformer output. The inverse filter for the RIR viewed through the beamformer, $\hat{h}^{-1}$ is applied:

$$z = \boldsymbol{w}^H (\boldsymbol{y}^s + \boldsymbol{y}^r + \boldsymbol{y}^i + \boldsymbol{y}^n) * \hat{h}^{-1} \tag{2.38}$$

The advantage of doing beamforming prior to inverse filtering is that there is only one filter operation needed, which is more computationally efficient.

2.7.4 Echo cancellation

In most cases it is not possible to estimate the full RIR without doing measurements, and even then it is hard to create a good inverse filter. A possibility is to just estimate the most prominent reflections and create a filter that removes these. An echo is a delayed and attenuated copy of the original signal. If this delay and attenuation is known, a simple filter could remove this echo. This is called *echo cancellation* and is a variant of inverse filtering. If multiple echoes should be removed, then a more advanced filter is needed. There also exist adaptive filters that automatically estimate the delay and attenuation factors [28].

Echo cancellation + beamforming As for the inverse filtering, echo cancellation is more effective before beamforming is applied to the microphone outputs. This is because the reflections are more prominent before beamforming.

Beamforming + echo cancellation After the beamforming, the prominent echoes have been added incoherently by the beamformer. This makes it harder to achieve the same effect as removing the echoes before beamforming since they are less prominent.

2.7.5 Noise filtering

In noise cancellation a filter h^n_m is created to remove the noise. h^n_m is created such that $y^n_m * h^n_m = 0$. The filter could be based on some a priori knowledge about the noise characteristics or an adaptive filter. The output would become:

$$z = (y^s_m + y^r_m + y^i_m + y^n_m) * h^n_m \tag{2.39}$$

Noise filtering + beamforming Noise filtering in advance of beamforming can be done:

$$z = \boldsymbol{w}^H \left((\boldsymbol{y}^s + \boldsymbol{y}^r + \boldsymbol{y}^i + \boldsymbol{y}^n) * \boldsymbol{h}^n \right) \tag{2.40}$$

where $\boldsymbol{h}^n$ is a vector containing all the m cancellation filters h_m^n. The DAS beamformer has a factor of M higher SNR relative to the single sensor, assuming that the noise is uncorrelated in time. If the remaining noise after filtering is correlated in time, the SNR gain of the beamformer would decrease. A better option would be to apply the noise filter after beamforming.

Beamforming + noise filtering If beamforming were applied before the noise filtering, some noise reduction would occur in the beamformer. The noise filter would then work on the remaining noise:

$$z = \boldsymbol{w}^H \left(\boldsymbol{y}^s + \boldsymbol{y}^r + \boldsymbol{y}^i + \boldsymbol{y}^n \right) * h^n \tag{2.41}$$

h^n is created such that $\boldsymbol{w}^H \boldsymbol{y}^n * h^n = 0$.

2.7.6 Interference filtering

Interference filtering for a single microphone is similar to noise filtering. If the interference is stationary and the emitted signal is possible to estimate, a filter h_m^i can be created such that $y_m^i * h_m^i = 0$. The difference between the interference cancellation and the noise cancellation is that the interference includes the reverberation, which is room dependent.

$$z = (y_m^s + y_m^r + y_m^i + y_m^n) * h_m^i \tag{2.42}$$

Interference filtering + beamforming Some filter could be applied in advance of beamforming. Beamforming, however, is very capable for removing interference. A zero in the beampattern could be placed at the location of the interfering source and nearly complete interference cancellation for a narrow band could be achieved.

Chapter 3

Method

3.1 Assumptions

Some assumptions have been made prior to the experiments. These are as follows:

Scope

1. We want to enhance speech which is band limited between 100Hz and 3500Hz.
2. The goal is to listen to a source, not determine DOA.

Sound sources

3. The sources are stationary, i.e. they do not move within the timeframe in which they are observed.
4. There is no air absorption present. (This is a fair assumption for the selected frequency range [22].)
5. The sound emitted from the source of interest is uncorrelated with the sound emitted from interfering sources.
6. The source is omnidirectional.
7. Spatial position of the source of interest is known.

Environment

8. The environment is stationary, meaning that reflecting surfaces do not move and the wave speed is constant.
9. The spatial positions of the image sources of the source of interest are known.
10. The absorption coefficients for the room are frequency independent.

Microphone array

11. The array has omnidirectional microphone elements.
12. The array is working and there are no defective elements in the array.
13. The array is well sampled ($d \leq \frac{\lambda_{min}}{2}$).

Other

14. The sampling rate is high enough to get a precise delay, $Fs \geq 10 f_{max}$. (More on this in [29, 30]) .

3.2 Metrics of performance

The goal of the beamformer is to increase the speech intelligibility of a source of interest. In this section several metrics of performance are presented and discussed.

3.2.1 Signal to noise, interference and reflections ratio

A commonly used set of metrics for the performance of a beamformer is the signal-ratios. For the signal model used in this thesis it makes sense to have one signal-ratio for each component of unwanted sound: signal to noise ratio (SNR), signal to interference ratio (SIR) and signal to reverberation ratio (SRR). They are defined as the power ratios between the signal and the other component:

$$SRR = \frac{P_{signal}}{P_{reverberation}} = \frac{\mathbb{E}\{|z_{signal}|^2\}}{\mathbb{E}\{|z_{reverberation}|^2\}} \tag{3.1}$$

$$SIR = \frac{P_{signal}}{P_{interference}} = \frac{\mathbb{E}\{|z_{signal}|^2\}}{\mathbb{E}\{|z_{interference}|^2\}} \tag{3.2}$$

$$SNR = \frac{P_{signal}}{P_{noise}} = \frac{\mathbb{E}\{|z_{signal}|^2\}}{\mathbb{E}\{|z_{noise}|^2\}} \tag{3.3}$$

where $\mathbb{E}\{\}$ is the statistical expectation operator.

The minimum power/distortionless response CRM (MVDR-CRM) algorithm minimizes the unwanted sounds while keeping the signal undistorted. This is equivalent of maximizing the signal to reverberation, interference and noise ratio (SRINR):

$$SRINR = \frac{P_{signal}}{P_{noise}} = \frac{\mathbb{E}\{|z_{signal}|^2\}}{\mathbb{E}\{|z_{noise} + z_{interference} + z_{reverberation}|^2\}} \tag{3.4}$$

Separation of signal, reverberation, interference and noise In order to calculate the signal-ratios, the different components need to be separated.

Simulated data In a simulated setting, the source, interference and noise components can simply be turned off. These sound components can be used in the beamformer separately, see figure 3.1. To separate the signal and reverberation from each other, the RIRs from the source to the array microphones are needed. First the simulation software generates the RIRs for a given source. The RIRs are used in the selected beamformer. The output is a single beamformed RIR. The highest peak in the RIR is the signal and the other peaks are the reverberation (It would also makes sense to define the first peak in the RIR as the signal since the direct sound is the first to reach the array. However, on the beamformed output this is not necessarily the case. This is why the highest peak is chosen rather than the first peak). The signal and reverberation can be separated into two RIRs, see figure 3.2. The last step is to convolve the RIRs with the source signal. Note that the adaptive

method, MVDR-CRM, always calculates its weights on the entire recorded sound. When beamforming the different components of the sound, these weights are used.

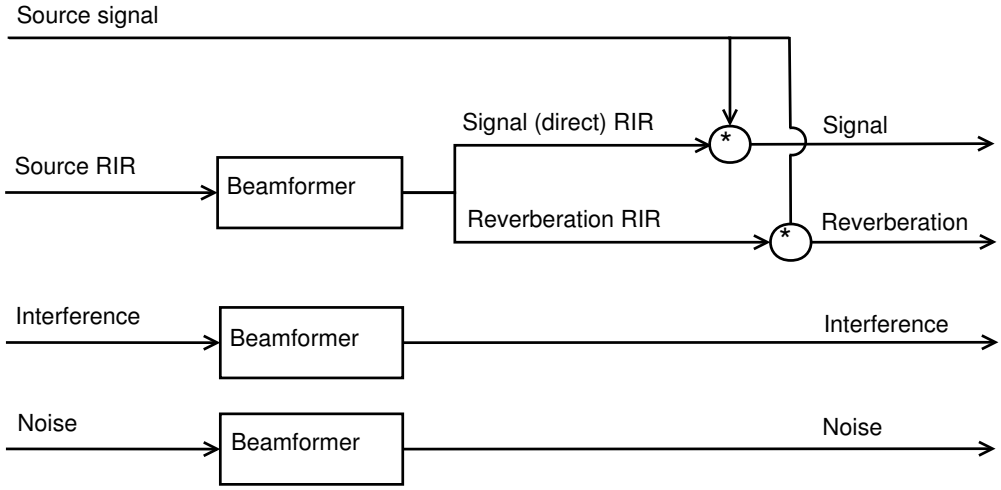


Figure 3.1: Separation of signal, reflection, interference and noise in a simulated setting.

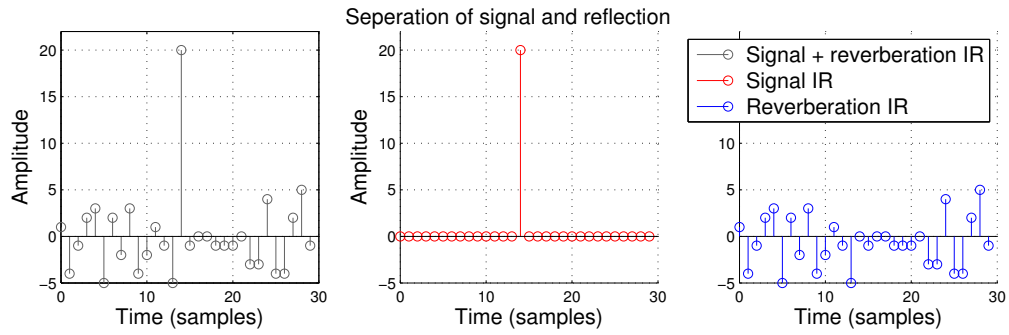


Figure 3.2: An impulse response recorded using a beamformer. The peak (marked with red) represents the signal. The rest are reverberation.

Measured data In the lab-measurements it is not possible to turn off the noise, since it is generated by the microphones. The interference (which is unwanted sound sources positioned in space) partly consists of sound arriving from ventilation systems, which is also not possible to turn off. The solution is to:

- Record the source RIR by using the swept sine method[1] The noise will degrade the swept sine analysis. This introduces a small error.
- Record the noise and interference together, and treat them as compound noise/interference.

[1]The swept sine method for measuring IRs where proposed by [31]. A MATLAB implementation made by [32] where used.

- Separate the signal from the reverberation by cutting out the highest peak of the RIR and convolving with a source signal (as for the simulated case).

This makes it possible to calculate the SRINR metric for the measurements. The SNR and SIR components are not possible to calculate for the chosen test setup.

3.2.2 Measuring speech intelligibility

Speech intelligibility is defined as the percentage of words or sentences that are correctly perceived by a listener [23]. This is can be evaluated using listening tests, but it requires a large number of test persons. There are many metrics available to objectively determine the speech intelligibility [33]. Meyer-Sound gives an brief overview at their webpage [34]. They state that there are two main categories of objective metrics for speech intelligibility.

1. Analysis of the reverberation.
2. Analysis of the signal to noise ratio.

The speech transmission index for public address systems (STIPA) method is of the latter category and is the most popular. It fits the subjective tests well, but assumes that the noise and interference is white. Experiments show that the high energy in the early part of the reverberation, about the first 50ms, increases the speech intelligibility [35]. The early to late ratio (ELR) metrics includes this effect. It divides the RIR into a *early* part, and a *late* part. The early part is the first 50ms of the impulse response and the late part is the rest. Like for the signal-ratios, it is possible to define the *early-ratios* for the three types of unwanted sound:

$$ELR = \frac{P_{early}}{P_{late}} = \frac{\mathbb{E}\{|z_{early}|^2\}}{\mathbb{E}\{|z_{late}|^2\}} \tag{3.5}$$

$$EIR = \frac{P_{early}}{P_{interference}} = \frac{\mathbb{E}\{|z_{early}|^2\}}{\mathbb{E}\{|z_{interference}|^2\}} \tag{3.6}$$

$$ENR = \frac{P_{early}}{P_{noise}} = \frac{\mathbb{E}\{|z_{early}|^2\}}{\mathbb{E}\{|z_{noise}|^2\}} \tag{3.7}$$

The early to interference ratio (EIR) and early to noise ratio (ENR) needs to be verified against listening tests to be certain that they do measure the speech intelligibility. Still they will give a good measure of the performance of the beamformers relative to each other. Figure 3.3 shows how the early part of the RIR is separated in the same manner as for the signal and reverberation.

While the signal-ratios are closer connected to the methods (beamformers), the early-ratios are closer connected to the applications (speech enhancement).

For this thesis it was chosen to use the signal-ratios as the main metric since it is the most commonly used for microphone array beamforming. There is one exception

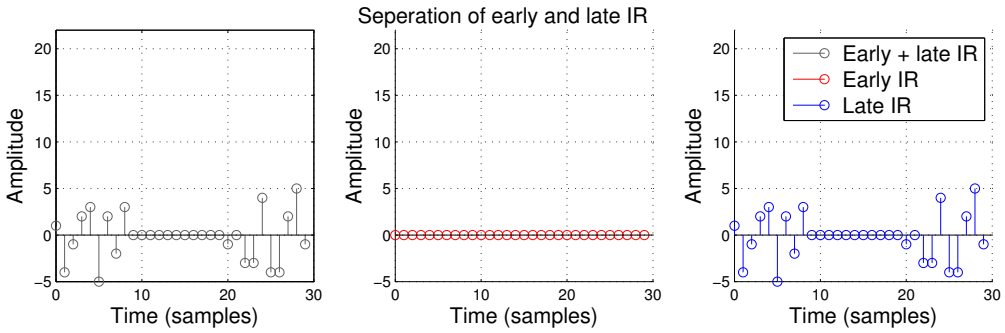


Figure 3.3: A impulse response recorded using a beamformer. The peak (marked with red) represents the early part. This is the signal peak $\pm 25ms$. The rest is reflections.

in the discussion chapter where the early-ratios are used as a complement to the signal-ratios.

3.2.3 Reference level

For power estimates it is common to use the decibel scale. This requires a reference level. For this report the reference level is always the microphone in the array that is positioned closest to the source. For example:

$$SNR_{increase,\,dB} = SNR_{beamformer,\,dB} - SNR_{element,\,dB}$$

This gives an impression of how much better the array beamformer is relative to using a single microphone.

3.2.4 Reference method

The new method is compared to the DAS method. This is the most widely used method and is clearly defined. Since CRM is using DAS beams it can be seen as an extension of DAS.

The DAS beamformer use uniform weighting for the outputs. It would, however, be possible to use a better weighting in the DAS beamformer. Either some sort of windowing function making the sidelobe levels lower or by positioning zeros in the position of strong reflections and interference. The reflections beams used in CRM could also position a zero in the position of the direct sound and other strong reflections and interferences. Both of these methods would increase the performance for both CRM and DAS substantially.

Another option would be to compare against an adaptive beamformer. Many adaptive beamformers have problems with correlated signals [5]. Reflections are correlated with the direct signal, this makes the use of adaptive beamformers complicated.

Since the same beams are used in both DAS and CRM it is expected that the performance of DAS directly affects the CRM beamformer. constructive reflections method (CRM) can be used with any steerable beamformer. This makes the choice of reference method less important since it directly affects the performance of CRM.

3.3 Simulations

There is an unlimited number of possible source and array positions, room sizes and absorption coefficients that can be chosen. The test setup is chosen to resemble a realistic use-case for a microphone array.

Two test cases have been selected, one with a single reflecting surface and one with a more realistic full room. A third room has been simulated to compare with the real measurements. This room is presented in the following *Measurements* section 3.4.

All the simulated results in chapter 5 have been created by these three simulations. The parameters used in the simulations are those that are listed here. If there is a deviation from this, it is stated in the figure text.

3.3.1 Test setup 1: Single reflection

The single reflection is the simplest case. This setup is chosen to be able to investigate how fundamental parameters, like the array size and reflection strength, affect the proposed algorithm.

The RIRs are analytically calculated using the time delays associated with the travel length. The reflection strength observed at the array is set to $A_1 = 0.68A_0$, where A_0 is the strength of the direct arrival. All the microphones observe the same signal. Figure 3.4 shows a visualization of the test setup.

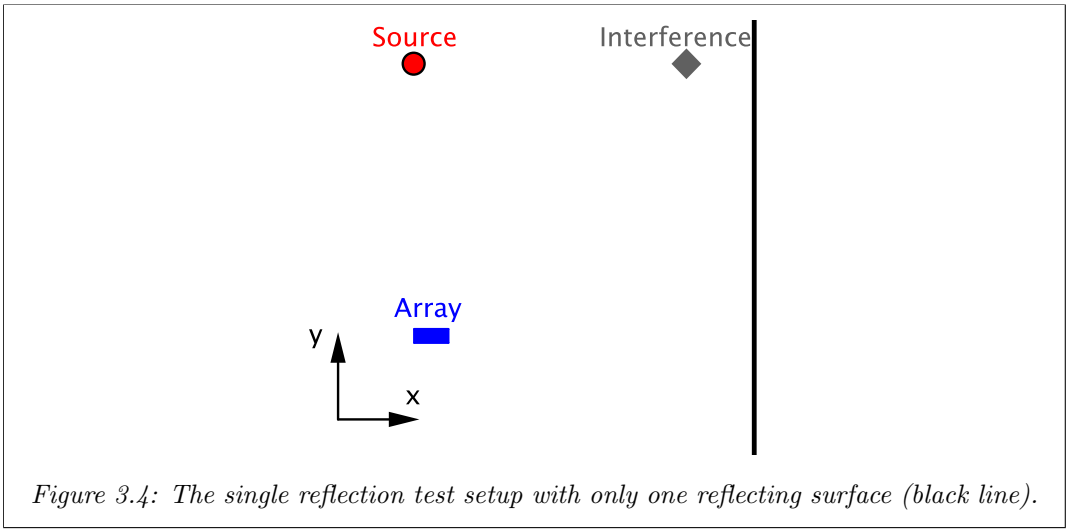


Figure 3.4: The single reflection test setup with only one reflecting surface (black line).

Room The room is a large room where all the walls have total absorption, except for the east wall. The width of the room is $50m$. Due to restrictions in the simulation software, the simulations are performed in the $z = 0.5m$ plane. This has no practical impact on the results.

The image sources used in CRM and MVDR-CRM are the direct sound and the single reflection.

Source and interference positions

- Source position: $\vec{x}_{source}$ = (0 m, 40 m, 0.5 m).
  This is 3.6° from the array normal.
- Interference position: $\vec{x}_{interference}$ = (40 m, 40 m, 0.5 m).
  This is −43.2° from the array normal.

The array is a ULA with 100 elements positioned along the x-axis at $y = 0.05$m, $z =$ 0.5 m. The phase center is positioned at $\vec{x}_{phase\,center}$ = (2.5 m, 0.05 m, 0.5 m) and the element spacing is $d = 0.05$ m, making the array size $D = 4.95$ m. This is an unrealistically large microphone array. Still it serves its purpose, to investigate the basic features of the beamformers.

The emitted signals are monochromatic waves:

$$s_0(t) = A_0 cos(-\frac{c}{\lambda} 2\pi t) \tag{3.8}$$

and the emitted interfering signal is

$$s_1(t) = A_1 cos(-0.99\frac{c}{\lambda} 2\pi t) \tag{3.9}$$

where the factor 0.99 is added to ensure that the two signals are uncorrelated.

The noise is Gaussian white noise, uncorrelated in time and from element to element. In MATLAB this is created with the command:

```
1 noise = wgn(M, N, power);
```

where M is the number of elements and N is the number of samples in time. The SNR observed at the microphones is set to 63 dB.

Simulation settings

- Sample rate: $Fs = 44100$ Hz
- Wave speed: $c = 343$ m/s
- Simulation length: $T = 5$ s
- Lambda: $\lambda = 2d$

3.3.2 Test setup 2: A small room

The second test case is designed to imitate a more realistic scenario for a microphone array. This setup offers an infinite number of reflections. Most of the experiments presented are performed with this setup. The room dimensions and absorption coefficients are common for a meeting room or small classroom. The array is placed on one of the walls, see figure 3.5.

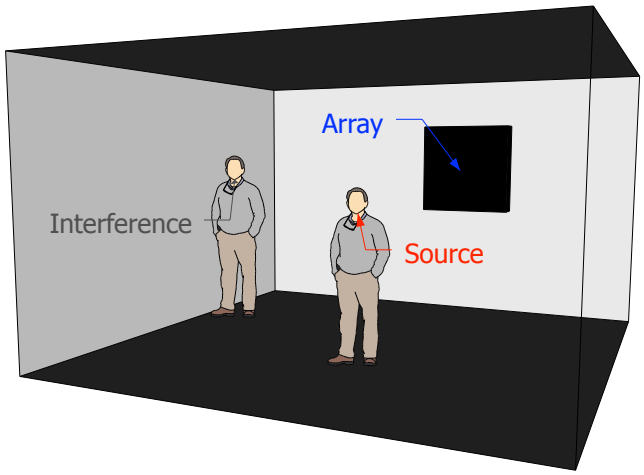


Figure 3.5: A sketch of test setup 2: A small room.

Room The walls have absorption coefficients that could be used in a normal classroom. To fulfill the assumption about frequency independent absorption, the same α values are used for the entire frequency range.

- west wall, windows ($\alpha = 0.07$)
- east wall, gypsum ($\alpha = 0.07$)
- south wall, gypsum ($\alpha = 0.07$)
- north wall, blackboard and gypsum ($\alpha = 0.07$)
- floor with linoleum and chairs ($\alpha = 0.5$)
- ceiling with absorbing tiles ($\alpha = 0.65$).

All the alpha values are chosen from [23]. The room dimensions are

- Width - 5.00m
- Length - 5.00m
- Height - 3.00m

The image sources used in CRM and MVDR-CRM are the 63 first. This is the direct source and all the 1st, 2nd and 3rd order reflections. They are sorted to have decreasing signal strength.

Source and interference positions

- Source position: $\vec{x}_{source} = (3.00 \text{ m}, 3.50 \text{ m}, 1.50 \text{ m})$
  This is $-8.2°$ from the array normal in the x, y-plane and $0°$ in the y, z-plane.
- Interference position: $\vec{x}_{interference} = (1.50 \text{ m}, 3.00 \text{ m}, 2.00 \text{ m})$
  This is $18.7°$ from the array normal in the x, y-plane and $-26.6°$ in the y, z-plane.

The results are heavily dependent on the source and interference positions due to the shape of the beampattern. To get more general knowledge of the behavior of the algorithm, 500 random source and interference positions are used to create an average. The positions are available in Appendix 6.4.3.

The array is a uniform rectangular array (URA) positioned at the south wall.

- Element count: $M = 10 \times 10$
- Element spacing: $d = 5$ cm
- Phase center: $\vec{x}_{phase\,center} = (2.50 \text{ m}, 0.05 \text{ m}, 1.50 \text{ m})$
- Extended in the x, z-plane

Lowpass filtering is applied to the microphone outputs with a cutoff frequency at 3430 Hz to avoid aliasing. The filter has an order of 100 and is created by MATLABs *fir1* function. Figure 3.6 shows the frequency response and the group delay.

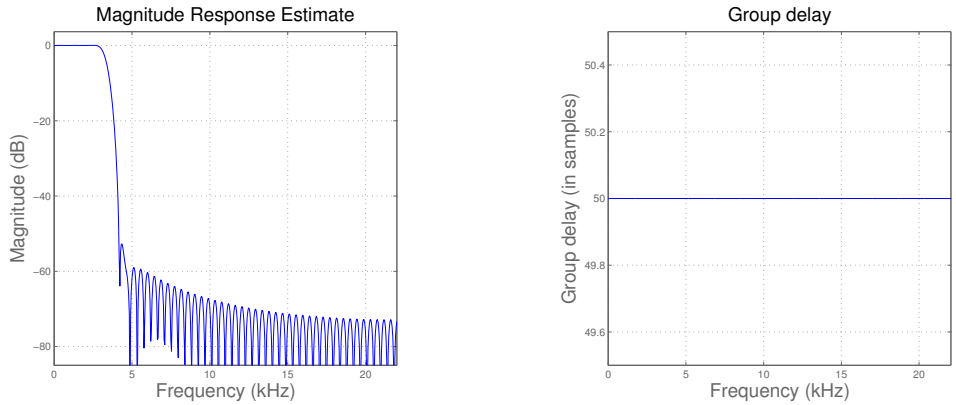


Figure 3.6: The filter used to avoid aliasing in the DAS beamformer.

The emitted signals are provided by the International Telecommunication Union (ITU). They have recommended some test signals for testing of telecommunication equipment. The signals mimic human speech and have the same frequency spectrum as regular voice. It sounds like a speaking person where the sound is distorted so that the words are unrecognizable. They are available online at [36].

The emitted signal from the source is the *p50m.wav* from ITU, resampled to 44100Hz using MATLAB's *resample*-function. The emitted signal from the interfering source is the *art_ v_ M.wav*, also resampled to 44100Hz. The average SIR for the array elements is 0 dB.

The noise is the same as for test setup 1.

Simulation settings

- Sample rate: $Fs = 44100Hz$
- Wave speed: $c = 343m/s$
- Simulation length: $T = 5s$

3.3.3 Simulating room impulse responses

A MATLAB [37] implementation of the image source method for calculation RIRs for rectangular rooms was created by Lehmann and Johansson [38].

The software simulates one RIR for each source-receiver pair. By convolving these RIRs with a signal of choice, the element outputs ($y_m(t)$) are found and beamforming can be applied.

Simulation speed-up The generation of the RIRs is time consuming (A 256-element array took 2 hours on a modern laptop running on a 2 GHz Intel Core i7 CPU for each source position). To speed-up the simulation, it was run in parallel at the Condor-cluster at the Department of Informatics, University of Oslo.

3.4 Measurements

Measurements in a real room with a real array where performed to validate the simulations.

3.4.1 Test setup 3: A real room

The room is a small room located in the basement at *Kristen Nygaards hus* at the University of Oslo, see figure 3.7, using an array produced by Squarehead Technology AS. This room was chosen because of its box-like geometry, so that it would be possible to simulate it with the image source method (ISM) implementation.

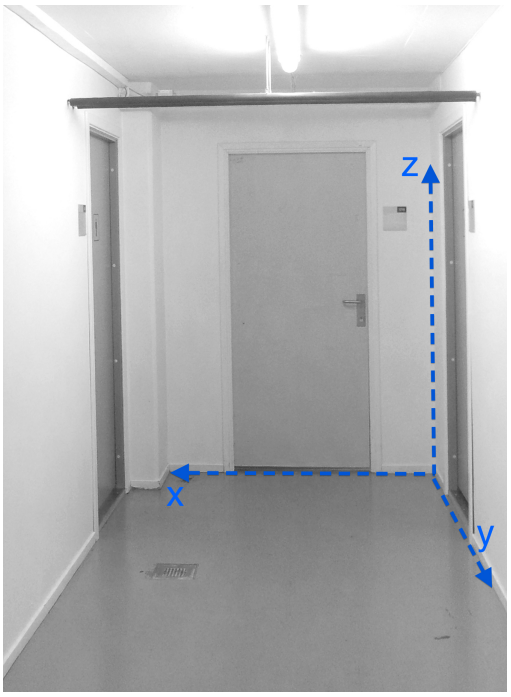


Figure 3.7: A panorama view of the room for the measurements. The room-centered coordinate system is visualized with the overlaying blue arrows.

Room The room dimensions are

- Width - 1.96m
- Length - 5.04m
- Height - 2.54m

The absorption values where estimated based on the materials of the wall:

- West wall - plywood ($\alpha = 0.09$)
- East wall - plywood ($\alpha = 0.09$)
- South wall - plywood ($\alpha = 0.09$)
- North wall - metal door ($\alpha = 0.03$)
- Floor - painted concrete ($\alpha = 0.07$)
- Ceiling - painted concrete ($\alpha = 0.07$)

The number of image sources used in in CRM and MVDR-CRM is 11. These are the direct source and the 1st, 2nd and 3rd order reflections that are positioned within a 90° angle from the front side direction of the array. The array has a closed casing making it "blind" to sound coming from the back of the array.

Sound sources The sources are small speakers, see figure 3.8 a. The source is playing a monochromatic wave, as for the *one reflection* test setup, with the frequency $f = 6860$Hz. The interfering source will play the *p50m* test signal from ITU. The source positions are:

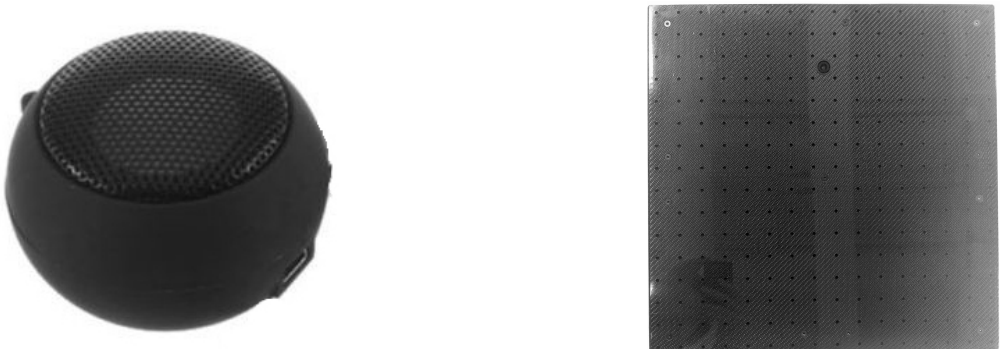
- Source position: $\vec{x}_{source} = (3.00m, 3.50m, 1.50m)$
  This is $-8.2°$ from array normal in the x, y-plane and 0° in the y, z-plane.
- Interference position: $\vec{x}_{interference} = (1.50m, 3.00m, 2.00m)$
  This is $-8.2°$ from array normal in the x, y-plane and 0° in the y, z-plane.¨å

The microphone array is the Squarehead CAPsure array. This is a uniform rectangular array (URA) consisting of 16×16 elements, see figure 3.8 b. The microphones are positioned with a 2.5 cm spacing, making the array $37.5cm \times 37.5cm$ large (plus casing).

The array is positioned on the floor and tilted 45° around the y-axis, see figure 3.9. The phase center is located at $\vec{x}_{phase\,center} = (0.16m, 2.51m, 0.16m)$.

Lowpass filtering is applied to the microphone outputs with a cutoff frequency at 6860 Hz to avoid spatial aliasing. It is created using MATLAB's *fir1* functions with an order of 100, see figure 3.10.

Squarehead provides a software solution that captures the raw data with a sampling frequency of $Fs = 44100Hz$. The data was analyzed using MATLAB.



(a) (b)

Figure 3.8: a. The speakers used for the source and interference. b. The Squarehead CAPsure microphone array.

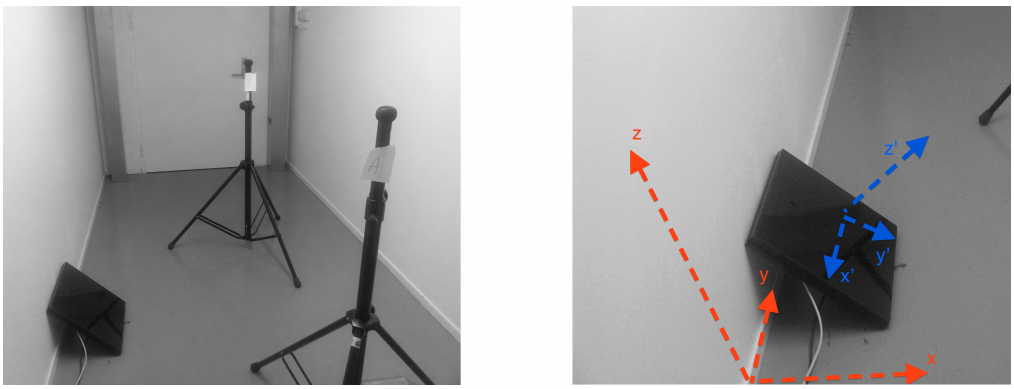


Figure 3.9: Left: The equipment setup in the room. Right: The room centric (red) and array centric (blue) coordinate systems.

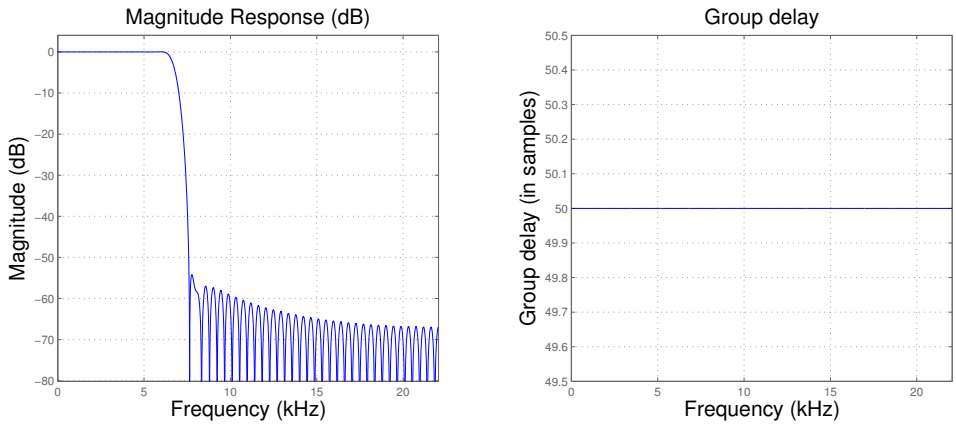


Figure 3.10: The filter used to avoid ailiasing in the DAS beamformer.

3.4.2 Challenges with real measurements

Some of the assumptions listed in section 3.1 can not be met in the real room experiment:

6. *The source is omnidirectional.*

The speakers are not omnidirectional. This will probably give rise to inaccurate estimation of the reflection amplitudes.

9. *The spatial positions of the image sources of the source of interest are known.*

The room is not a perfect rectangular box, meaning that the estimate of the image sources may be inaccurate.

11. *The array has omnidirectional microphone elements.*

The microphones in the experiment are positioned in a casing. This makes the array much more sensitive to sound coming from the front than to sound coming from the back.

12. *The array is working and there are no defective elements in the array.*

This it hard to validate, some elements may be more sensitive than others. Some elements may be broken.

The solution is to:

- **Assumption 6:** This is hard to do anything about without using expensive loudspeakers and the speakers are therefore assumed to be omnidirectional.
- **Assumption 9:** The error is expected to be relatively small compared to the wavelength. Therefore it is ignored.
- **Assumption 11:** Only use the image sources that are in front of the array.
- **Assumption 12:** The array will be treated as a fully working array.

Chapter 4

The proposed beamformers

4.1 Constructive reflections method (CRM)

When sound from a source is reflected, each microphone in an array observes the same signal twice, first the direct sound, then the reflected, see figure 4.1. The idea of the constructive reflections method (CRM) beamformer is to use two DAS beams, one steered towards the source and one towards the reflection. The output of each steered DAS is considered as an individual sensor observing the same signal, but delayed according to the travel path. CRM then delays the outputs to reverse the time delays so that the signals are aligned. Finally the outputs are added together.

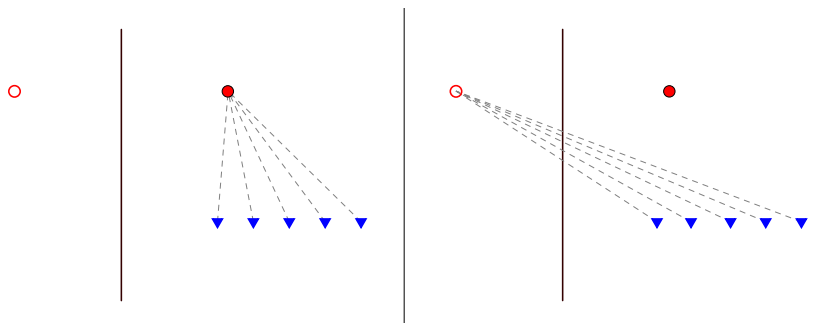


Figure 4.1: Left: An array observing an source. Right: An array observing the image source.

4.1.1 Mathematical formulation

Recall the DAS beamformer, equation 2.19:

$$z(t) = \sum_{m=0}^{M-1} w_m \, y_m(t + \tau_m)$$

This is a special case where DAS is steered at the source itself ($j = 0$). A more general expression for DAS is:

$$z_j(t) = \sum_{m=0}^{M-1} w_{m,j}\, y_m(t + \tau_{m,j}) \tag{4.1}$$

where j is the index denoting which image source the beamformer is steered at and $\tau_{m,j}$ is the time delay between the image source j and the microphone m:

$$\tau_{m,j} = \frac{||\vec{x}_m - \vec{x}_j||}{c}$$

CRM performs a second *delay and sum* operation on the DAS beams. A sub-set of the J sources are included in CRM. The image sources used in the beamformer are denoted using the index l counting from 0 to $J-1$. The CRM beamformer becomes:

$$z'(t) = \sum_{l=0}^{L-1} w'_l\, p_l\, z_l(t) \tag{4.2}$$

where L is the number of image sources used in the CRM beamformer. p_l is the phase shift associated with each image source being either 1 or -1. w'_l is the weight for each DAS output. (Note that there are no extra delays necessary in the CRM formulation, since the signals in the reflection beams already are aligned with the signal in the direct sound beam.) Then by replacing the z_l term with the DAS definition (equation 4.1), we get:

$$z'(t) = \sum_{l=0}^{L-1} w'_l\, p_l \sum_{m=0}^{M-1} w_{m,l}\, y_m(t + \tau_{m,l}) \tag{4.3}$$

4.1.2 Vector notation

Recall the vector notation for the DAS beamformer:

$$z(t) = \boldsymbol{w}^H \boldsymbol{y}(t)$$

This can also be generalized for all j image sources:

$$z_j(t) = \boldsymbol{w}_j^H \boldsymbol{y}_j(t)$$

where $\boldsymbol{w_j} = \begin{bmatrix} w_{0,j} \\ w_{1,j} \\ \vdots \\ w_{M-1,j} \end{bmatrix}$ and $\boldsymbol{y}_j(t) = \begin{bmatrix} y_0(t + \tau_{0,j}) \\ y_m 1(t + \tau_{1,j}) \\ \vdots \\ y_{M-1}(t + \tau_{M-1,j}) \end{bmatrix}$.

Inserted into the CRM formulation (eq. 4.2) the CRM beamformer becomes:

$$z'(t) = \sum_{l=0}^{L-1} w'_l\, p_l\, \boldsymbol{w}_l^H \boldsymbol{y}_l(t)$$

The CRM part of the beamformer expression can also be written on vector form:

$$z'(t) = (\boldsymbol{w'})^H \, \boldsymbol{z}(t) \tag{4.4}$$

where $\boldsymbol{w'} = \begin{bmatrix} w'_0 \\ w'_1 \\ \vdots \\ w'_{L-1} \end{bmatrix}$ and $\boldsymbol{z}(t) = \begin{bmatrix} p_0 \, z_0(t) \\ p_1 \, z_1(t) \\ \vdots \\ p_l \, z_{L-1}(t) \end{bmatrix}$

4.1.3 Behavior on signal model

The signal model (equation 2.28) can be inserted into DAS and CRM beamformers. This makes it possible to find the analytical array gain for each beamformer.

DAS
The signal model is inserted into the DAS beamformer (equation 4.1) when steered at the source ($j = 0$). The output of the beamformer is divided into four categories:

Signal:

$$z_0^s(t) = \sum_{m=0}^{M-1} w_{m,0} \, A_0 \, s_0(t - \tau_{m,0} + \tau_{m,0}) \tag{4.5}$$

Reverberation:

$$z_0^r(t) = \sum_{m=0}^{M-1} w_{m,0} \sum_{j=1}^{J-1} A_j \, p_j \, s_0(t - \tau_{m,j} + \tau_{m,0}) \tag{4.6}$$

Interference:

$$z_0^i(t) = \sum_{m=0}^{M-1} w_{m,0} \sum_{q=1}^{Q-1} s_q(t + \tau_{m,0}) * h_{q,m}(t + \tau_{m,0}) \tag{4.7}$$

Noise:

$$z_0^n(t) = \sum_{m=0}^{M-1} w_{m,0} \, y_m^n(t + \tau_{m,0}) \tag{4.8}$$

CRM
Like for DAS the signal model is inserted into the CRM definition (equation 4.3) and the output is divided into the same four components:

Signal:

$$z'^{\,s}(t) = \sum_{l=0}^{L-1} w'_l \, p_l \sum_{m=0}^{M-1} w_{m,l} \sum_{j=0}^{J-1} A_j \, p_j \, s_0(t - \tau_{m,j} + \tau_{m,l}) \quad \Big| \; l = j \tag{4.9}$$

Reverberation:

$$z'^{\,r}(t) = \sum_{l=0}^{L-1} w'_l \, p_l \sum_{m=0}^{M-1} w_{m,l} \sum_{j=0}^{J-1} A_j \, p_j \, s_0(t - \tau_{m,j} + \tau_{m,l}) \quad \Big| \; l \neq j \tag{4.10}$$

Interference:

$$z'^{\,i}(t) = \sum_{l=0}^{L-1} w'_l \, p_l \sum_{m=0}^{M-1} w_{m,l} \sum_{q=1}^{Q-1} s_q(t + \tau_{m,l}) * h_{q,m}(t + \tau_{m,l}) \tag{4.11}$$

Noise:

$$z'^{\,n}(t) = \sum_{l=0}^{L-1} w'_l \, p_l \sum_{m=0}^{M-1} w_{m,l} \, y^n_m(t + \tau_{m,l}) \tag{4.12}$$

Reverberation, DAS vs CRM

The expressions for the reverberation outputs are quite complicated. It is hard to say something general about them, but some assumptions could be made to investigate a simple scenario. Let the room be anechoic (full absorption at all surfaces) except for one reflecting surcface. The reverberation components of the beamformer output becomes for DAS:

$$z^r_0(t) = \sum_{m=0}^{M-1} w_m \; - A_1 \, s_0(t - \tau_{m,1} + \tau_{m,0}) \tag{4.13}$$

and CRM:

$$z'^{\,r}(t) = \sum_{l=0}^{1} w'_l \, p_l \sum_{m=0}^{M-1} w_{m,l} \; - A_1 \, s_0(t - \tau_{m,1} + \tau_{m,l}) \quad \Big| \; l \neq 1 \tag{4.14}$$

If the reflection is not too close to the source the delays applied by the beamformer ($\tau_{m,0}$) and the delay associated with length of travel path ($\tau_{m,1}$) will be different from each other. It will cause destructive interference. This is also the case for the reflection beam used in CRM. The output of DAS will be M incoherently added versions of the reflection, while for CRM it would be $2M$ ($L = 2$).

Interference, DAS vs CRM

As for the reverberation, some assumptions must be made. Let the room be fully anechoic with one reflecting surface. Only one interfering source is present. The interference components of the beamformer output becomes for DAS:

$$z^i_0(t) = \sum_{m=0}^{M-1} w_m \left(A_1 s_1(t - \tau^*_{1,m} + \tau_{m,0}) + A_2 s_1(t - \tau^*_{2,m} + \tau_{m,0}) \right) \tag{4.15}$$

and CRM:

$$z'^{\,i}(t) = \sum_{l=0}^{L} w'_l \, p_l \sum_{m=0}^{M-1} w_{m,l} \left(A_1 s_1(t - \tau^*_{1,m} + \tau_{m,0}) + A_2 s_1(t - \tau^*_{2,m} + \tau_{m,0}) \right) \tag{4.16}$$

where A_k is the attenuation of the source due to the travel path and $\tau^*_{k,m}$ is the delay to each m sensor via the direct path ($k = 1$) and reflection ($k = 2$). Let the source and interference be positioned apart from each other. The DAS beamformer (direct beam) will observe the interference through a sidelobe or zero in the beampattern and added incoherently. For the reflection beam used in CRM it is not certain that the interference is not positioned in the main lobe (the location of the reflection). This makes the DAS output consist of M incoherently added versions of the interference. CRM could in the best case consist of $2M$ incoherently added versions of the interference, but in the worst case M incoherently and M coherently added versions.

Array gain, DAS vs CRM

The array gains tell us how effectivly the beamformer can remove noise from the output relative to a single microphone. It is defined as the ratio of SNR between the beamformer output and a single element [5]:

$$G \equiv \frac{SNR_{array}}{SNR_{element}} \tag{4.17}$$

(This is the same as the SNR metric presented in chapter 3 with the single element as reference.) We assume that the signal and noise have *zero mean*:

$$E\left\{y^s_m(t)\right\} = E\left\{y^n_m(t)\right\} = 0 \tag{4.18}$$

We also assume that the noise is uncorrelated in time and uncorrelated from sensor to sensor:

$$E\left\{y^n_m(t)y^n_{m'}(t')^*\right\} = \sigma^2_{noise}\delta(m-m')\delta(t-t') \tag{4.19}$$

where δ is the Dirac delta function and t', m' are lags in time and sensors, σ^2_{signal} is the variance in the noise component.

The different SNR levels can be found analytically. We assume uniform weighting for both DAS and CRM beamformers and let σ^2_{noise} be the variance of the signal.

$$SNR_{element} = \frac{E\left\{|y^s_m(t)|^2\right\}}{E\left\{|y^n_m(t)|^2\right\}} = \frac{A_0^2\,\sigma^2_{signal}}{\sigma^2_{noise}} \tag{4.20}$$

$$SNR_{DAS} = \frac{E\left\{|z^s(t)|^2\right\}}{E\left\{|z^n(t)|^2\right\}}\frac{MA_0^2\,\delta^2_{signal}}{\delta^2_{noise}} \tag{4.21}$$

$$SNR_{CRM,best} = \frac{E\left\{|z'^s(t)|^2\right\}}{E\left\{|z'^n(t)|^2\right\}} = \frac{(\sum_{l=0}^{L-1} A_l)^2 M\sigma^2_{signal}}{L\sigma^2_{noise}} \tag{4.22}$$

$$\Big|\ \tau_{m,l} \neq \tau_{m',l'}\ \forall\ m,l,m',l'\ \Big|\ m \neq m', l \neq l'$$

It is assumed that the delays for each microphone are ambiguous for all the image sources. If some of the image sources are positioned at equal distances from the array it is possible that the microphones are delayed equally. The theoretical worst case would be a situation where the array and the source are positioned so that all the

microphones are delayed the same for each source. In a rectangular room this is only possible if the both the source and array are positioned in the center. Two and two reflection beams would have full correlation in the noise components. The SNR for this case is:

$$SNR_{CRM,worst} = \frac{E\left\{|z'^{s}(t)|^2\right\}}{E\left\{|z'^{n}(t)|^2\right\}} = \frac{(\sum_{l=0}^{L-1} A_l)^2 M \sigma^2_{signal}}{(L\%2 + 2(L - L\%2))\sigma^2_{noise}} \tag{4.23}$$

$$\Big| \ \tau_{m,l} = \tau_{m,l'} \ \forall \ m, l, l' \ , \quad \tau_{m,l} \neq \tau_{m',l} \ \forall \ m, l, m' \ \Big| \ l \neq l'$$

The array gain for DAS becomes:

$$G_{DAS} = \frac{SNR_{DAS}}{SNR_{element}} = M \tag{4.24}$$

It shows that the array gain for DAS is only dependent on the number of microphones. The CRM array gain becomes:

$$G_{CRM,best} = \frac{SNR_{CRM}}{SNR_{element}} = \frac{(\sum_{l=0}^{L-1} A_l)^2 M}{L\, A_0^2} \tag{4.25}$$

$$G_{CRM,worst} = \frac{SNR_{CRM}}{SNR_{element}} = \frac{(\sum_{l=0}^{L-1} A_l)^2 M}{(L\%2 + 2(L - L\%2)\, A_0^2} \tag{4.26}$$

The CRM array gain depends on the strength of the reflections used in the beamformer. It is most likely that the result would be closest to the best case. In most cases there would be either no equal delays, or one element with equal delays in two beams. For this reason only the best case is considered for the following analysis.

Figure 4.2 shows how the array gain of the CRM beamformer behaves when an increasing number of reflections is used in the algorithm. The amplitude values (A_l) are analytically calculated for the room used in test setup 2.

The (best-case) CRM has a rapidly increasing array gain for the first 30 reflections. After this the array gain decreases slowly, and eventually drops below the DAS levels. The 63 first reflections that are used in the simulations only give a 0.5 dB lower array gain than at the optimal number of reflections.

The decrease in array gain happens because the signal component in the reflections gets weaker. It gets weaker because the travel path gets longer (more geometrical spreading) and there are more reflections (absorption at each reflection). While the signal component decreases, the noise component stays the same, creating an optimum number of sources.

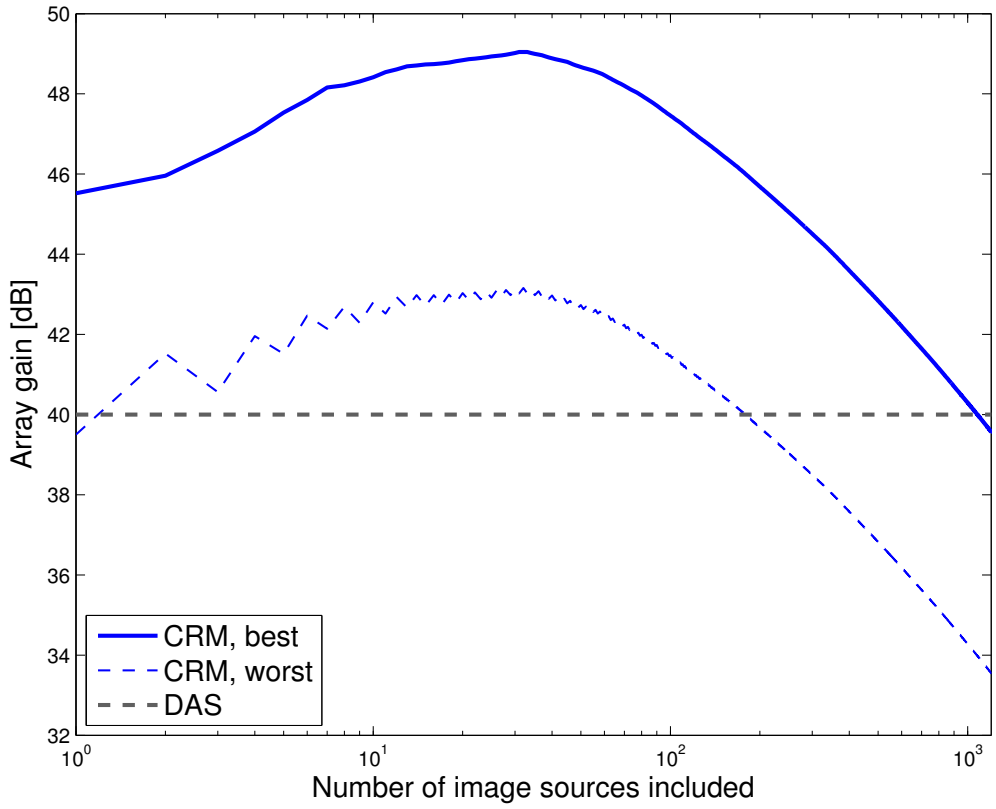


Figure 4.2: The array gain for CRM compared to DAS. The sources are added one by one, beginning with the strongest reflections first.

4.2 MVDR-weighting for CRM (MVDR-CRM)

As the different reflection beams contribute to either an increase or a decrease in the SNR it is reasonable to assume that they do so for the SRR and SIR. One solution is to only use a small number of reflections. Another option is to weight the beams differently. This section presents an adaptive method for determining the weighting of the beams.

The minimum power/distortionless response (MVDR) criterion is commonly used to find the optimal weights for a DAS beamformer. It uses the weights to minimize the beamformer's output power, while having the constraint that the amplification in the steering direction should be 1. This section will provide an MVDR formulation for the weighting of image sources in CRM.

Recall the vector notation for CRM (equation 4.4):

$$z'(t) = (\boldsymbol{w'})^H \, \boldsymbol{z}(t)$$

The power is defined as:

$$P = |z'|^2 = z' z'^* = (\boldsymbol{w'})^H \, \boldsymbol{z}(t) \, \boldsymbol{z}(t)^* \, (\boldsymbol{w'}) = (\boldsymbol{w'})^H \boldsymbol{R'} (\boldsymbol{w'}) \tag{4.27}$$

where $\boldsymbol{R'}$ is the covariance matrix for$\boldsymbol{z}(t)$

The MVDR formulation becomes:

$$\underset{\boldsymbol{w'}}{min} \left((\boldsymbol{w'})^H \boldsymbol{R'} (\boldsymbol{w'}) \right), \quad s.t \; \boldsymbol{w}^H \boldsymbol{A} = 1 \tag{4.28}$$

where $\boldsymbol{A} = \begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_{L-1} \end{bmatrix}$. The constraint, $\boldsymbol{w}^H \boldsymbol{A} = 1$, forces the MVDR algorithm to keep the signal level undistorted. The solution to the MVDR problem is:

$$\boldsymbol{w'} = \frac{\boldsymbol{R'}^{-1} \boldsymbol{A}}{\boldsymbol{A}^H \boldsymbol{R'}^{-1} \boldsymbol{A}} \tag{4.29}$$

The classic MVDR formulation for DAS (also know as minimum power/minimum variance (MP/MV)/Capon) breaks down when there are correlated signals present. The algorithm uses the correlated signal to cancel out the direct signal. MVDR-CRM is performed on the DAS beams. The outputs ($\boldsymbol{z}(t)$) are already phase shifted so that the signals have the same phase in all the beams. The constraint in the MVDR-CRM algorithm prevents it from cancelling the signal part. There are, however, scenarios where MVDR-CRM could cause signal cancellation, namely when the $\boldsymbol{A}$ estimate is inaccurate or there are interfering sources emitting a correlated signal.

4.2.1 Diagonal loading for robustness

All the techniques used in order to make MVDR formulations for DAS robust, can be applied to MVDR-CRM as well. In general all the techniques have an adjustable parameter that controls how aggressive the MVDR algorithm should be.

To illustrate this, the method of diagonal loading has been chosen. Here the covariance matrix ($\boldsymbol{R}$) is loaded on the diagonal. This mimics the effect of adding white noise to the microphone outputs. To perform diagonal loading on $\boldsymbol{R'}$ we replace it by $\epsilon\boldsymbol{I} + \boldsymbol{R'}$ making the solution:

$$\boldsymbol{w'} = \frac{\left(\epsilon\boldsymbol{I} + \boldsymbol{R'}\right)^{-1}\boldsymbol{A}}{\boldsymbol{A}^H\left(\epsilon\boldsymbol{I} + \boldsymbol{R'}\right)^{-1}\boldsymbol{A}} \tag{4.30}$$

where ϵ is a adjustable parameter and $\boldsymbol{I}$ is the identity matrix. When $\epsilon \to \infty$ ($\epsilon\boldsymbol{I} + \boldsymbol{R'} \to \epsilon\boldsymbol{I}$) the weighting becomes a function of the amplitudes from the image source model:

$$\boldsymbol{w'} = \frac{\left(\epsilon\boldsymbol{I}\right)^{-1}\boldsymbol{A}}{\boldsymbol{A}^H\left(\epsilon\boldsymbol{I}\right)^{-1}\boldsymbol{A}} = \frac{\boldsymbol{A}}{\boldsymbol{A}^H\boldsymbol{A}} \tag{4.31}$$

where $\boldsymbol{A}^H\boldsymbol{A}$ is a scalar value. In the implementation of the beamformers, the weights are normalized. The value of $\boldsymbol{A}^H\boldsymbol{A}$ becomes irrelevant and the weights could be expressed by as the amplitudes:

$$\boldsymbol{w'} = \boldsymbol{A} \tag{4.32}$$

The value of the ϵ parameter is discussed in [39] and is set to be:

$$\epsilon = \Delta tr\left\{\boldsymbol{R'}\right\} \tag{4.33}$$

with suitable Δ values ranging from $\Delta = \frac{10}{L}$ to $\Delta = \frac{1}{100L}$. L is the dimension of $\boldsymbol{R'}$ (the number of image sources used) and $tr\{\}$ is the trace operator (summing the diagonal of a input matrix).

Chapter 5

Results and Discussion

5.1 Simulations: beamformer behavior

The simulations are performed to investigate how the two algorithms perform under various circumstances. Due to the amount of results, the discussion is interleaved with the results to avoid repetition and extensive references. Each plot represents a result of one of the tests. For each plot there is a paragraph about the how the test is conducted and what the test can tell us. Then the results for DAS, CRM and MVDR are highlighted and discussed. Finally the key findings are presented. To make it easier for the reader to identify the key findings, they are highlighted using **bold face**.

5.1.1 Array size, single reflection

The array size test is performed by adding elements to an array, starting with a center element and adding one element at the time. The elements are added from the phase center, alternating from the left and the right side.

Figure 5.1 shows that DAS is improving its performance relative to the single microphone on SRR and SNR gradually as the array size increases. This is what is expected, as the resolution gets better and the array gain increases. The rapid oscillations in the graphs are generated by zeros moving around as more microphones are added (each new microphone adds a new zero to the beampattern). On the SIR metric we observe a large dip around $M = 38$. This may be due to a sidelobe being shifted across the path of the interfering signal.

CRM performs better than DAS on both the SRR and SNR metrics. The performance stays respectively 9 dB and 3 dB over DAS levels. CRM is consequently performing below DAS levels on the SIR metric. The signal is component is increased more than the reverberation and noise components. Both the beam steered directly at the source and the one steered at the reflection contain energy from the interfering source. It is non-destructive, which gives the increased SIR level.

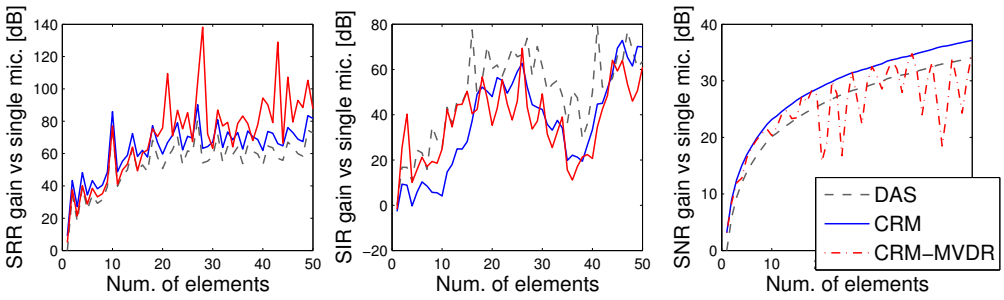


Figure 5.1: The beamformer's performance on an increasing number of elements. This result is generated with test setup 1: Single reflection.

The adaptive MVDR-CRM beamformer weights are visualized in figure 5.2. It seems that the weightings are fairly stable up to 15 elements where the direct sound gets most of the weight. For larger array sizes the reflection is weighted highest, and the weights are less stable. From figure 5.1, we see that for some element sizes, the SRR-curve for MVDR-CRM has high peaks ($M = 27, 48$). For these array sizes the direct signal comes close to a zero in the beampattern of the beam steered at the reflection. This causes the SRR gain for the reflection to become really high, and MVDR puts emphasis on the reflection. This is the reason for the oscillating weights in figure 5.2.

The test shows that the improvement on the SRR and SNR for CRM, compared to DAS is not affected by the array size. The behavior of MVDR is unpredictable for the one reflection scenario because the positioning of the zeros has a large impact on the amount of reflection and interference in the beams.

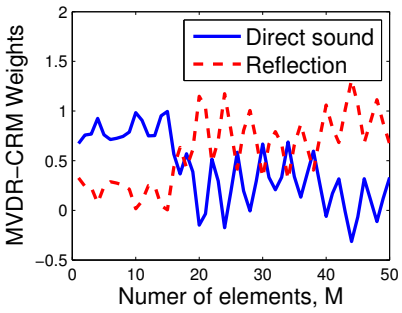


Figure 5.2: The MVDR weights on increasing number of elements. This result is generated with test setup 1: Single reflection.

5.1.2 Array size, full room

The new method can be affected by the source and interference positions. To find a general trend in the performance, 500 different random source and interference positions (listed in Appendix 6.4.3) where used. They where selected from the areas of the room where it would be sensible to listen for speech signals with an array itself. This excludes some areas close to the wall, floor, ceiling and the array. The median performance on the metrics gives an impression of the improvement of the algorithm. A mean SRINR over the 500 source positions with a 95% confidence interval gives a general .

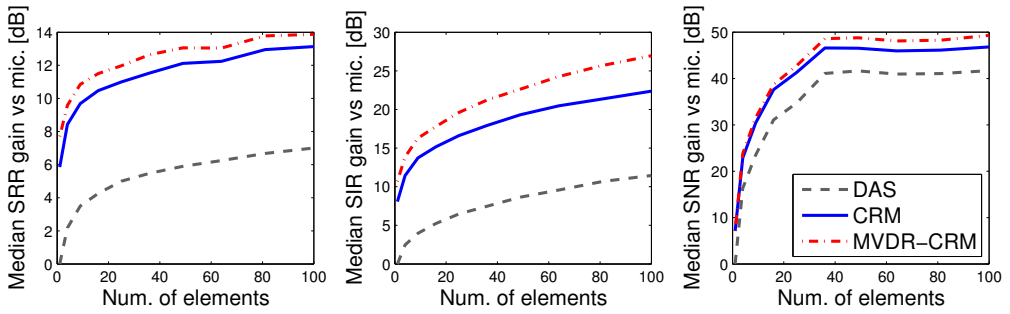


Figure 5.3: Median performance on 500 random source and interference positions. This result is generated with test setup 2: A small room.

DAS is performing like expected, see figure 5.3, the performance increases as the number of elements in the array increases.

CRM is performing better than DAS, having at least 6 dB higher SRR, 10 dB higher SIR and 5 dB higher SNR. It is interesting to see that the increase relative to DAS is independent of the array size. This is what was observed in the single reflection test. This confirms that the performance of CRM is directly dependent on the performance of the beamformer used to create the beams. This was discussed in section 3.2.4.

MVDR-CRM uses the weights to increase the SIR. This is what is expected since the interference has the highest sound level of all the components of unwanted sound. MVDR-CRM is still performing well above the DAS beamformer on all metrics. The increase is stable at 7 dB on SRR and SNR. On SIR the improvement increases from 10 dB to 15 dB as the resolution improves. This may be because some of the reflection beams improve their performance. The adaptive beamformer weights the beams with good performance higher.

An interesting result is that both CRM and MVDR-CRM have increased performance on the one-element array. At this point the algorithms only use the signal from one microphone. The output is the summation of multiple versions of this one signal, but with different delays. This shows that the proposed algorithm can be used to improve the performance of a single microphone. As a contrast to inverse filtering or echo cancellation the reflections are not removed but used to increase the signal. The reflecting surfaces make the same microphone observe the source from different

angles, creating a virtual array.

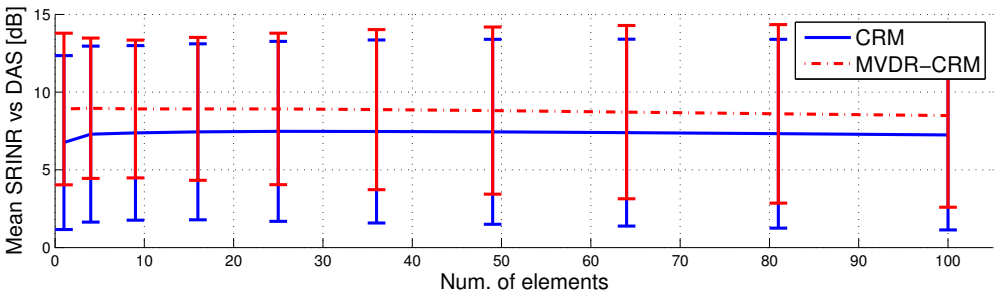


Figure 5.4: Mean performance compared to DAS on 500 random source and interference positions. The error bars mark 2σ, a 95% confidence interval. This result is generated with test setup 2: A small room.

The mean performance on the SRINR metric, see figure 5.4, shows that both CRM and MVDR-CRM have better performance than DAS with 95% confidence. The results also show that the improvement are nearly un-affected by the array size. The mean improvement is 6 dB and 8 dB for CRM and MVDR-CRM.

The key findings of this tests is that both CRM and MVDR-CRM has increased performance compared to DAS on all metrics. The improvement is between 5 dB and 15 dB.

5.1.3 Reflection strength, single reflection

The reflection strength test shows how the strength of the reflection affects the beamformers. In a real-life situation there will be many reflections. This test cannot tell us if a reflection is too weak to include in CRM, but it will give insight into how much better a strong reflection would be than a weaker one. (MVDR-CRM is not evaluated in this test.)

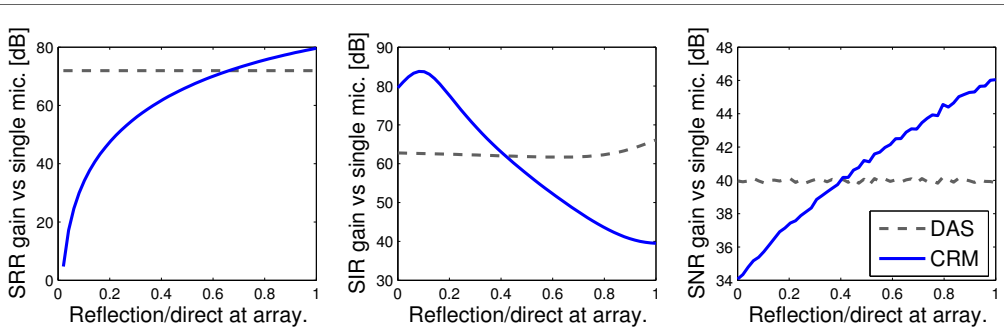


Figure 5.5: Reflection strength. This result generated with test setup 1: Single reflection, while the reflections reflection strength are increased. The reflection strength is measured at the array elements.

Figure 5.5 shows that DAS is performing fairly stable with respect to the single

element on all performance metrics. The high SRR level indicates that the reflection of the source arrives close to a zero in the beampattern. The SIR level increases, which indicates that the reflection of the interference arrives in a sidelobe. As the reflection becomes stronger, the reflection of the interfering source also gets stronger.

As expected CRM is improving its performance on SRR and SNR. Higher reflections give rise to a higher signal contribution from the beam steered at the reflection. CRM crosses DAS on the SRR metric at 0.65. Here the signal contribution of the reflection beam equals the reverberation contribution. On SIR the performance is dropping as the reflection strength is increasing. This happens as a result of the reflected interference getting higher, CRM has four sources for interference, the direct and the reflected interference leaking into the output via side lobes, for the two CRM beams. The analytical SNR curve (figure 4.2 in section 4.1.3) show that as the reflections gets weaker, they stop contributing to better SNR. The single reflection test confirms this.

The test shows that a strong reflection directly increases the SRR and SNR for CRM. If the reflection is too low compared to the direct sound, DAS will be better. The SIR improvement is dependent on how much stronger the reflections of the interfering source become.

5.1.4 Reverberation time

The reverberation time test tells us how the room acoustics affect the method. It may give upper or lower limits for the reverberation time for where the CRM beamformer will function. The reverberation time also directly affects the reflection strengths.

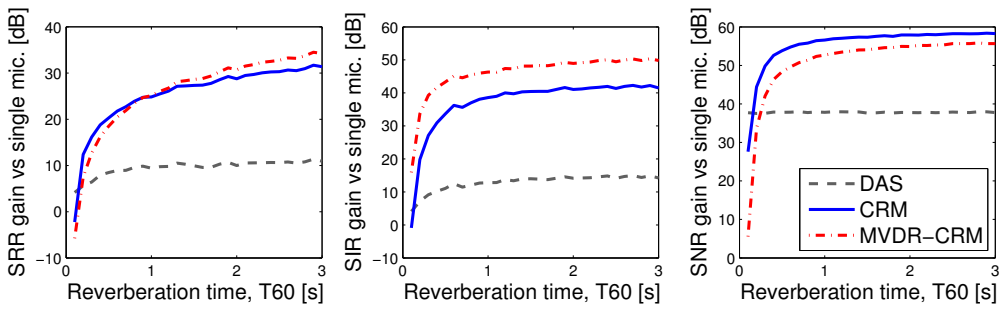


Figure 5.6: Reflection strength. This result is generated with test setup 2: A small room. The absorption coefficients are the same on all the surfaces and are adjusted according to the wanted reverberation time.

Figure 5.6 shows that the reverberation time has a small impact on DAS. Its performance on SRR and SIR increases slightly when the reverberation time increases. Increasing reverberation time gives higher reverberation and interference, both for the single microphone and DAS. While the single microphone is omnidirectional and increases the interference faster according to the reverberation time, the DAS beam is not omnidirectional and therefore does not increase the interference part this much.

The SNR is independent of reverberation time.

CRMs performance increases rapidly for increasing reverberation time. For reverberation times below 0.2s CRM performed worse than DAS. This is when the signal in the reflections is too low to compensate for the increased noise/interference in the reflections beam. After this the performance for CRM is much higher than the performance for DAS. The signals in the reflection beams are so high that they contribute to higher signal-ratios.

MVDR-CRM behaves similarly as CRM. The only difference is that it suppresses some reverberation and interference like we have seen on the other tests.

The test shows that CRM and MVDR-CRM are better than DAS as long as there is a small amount of reverberation.

5.1.5 Number of reflections

The number of reflections used in CRM is a fundamental parameter. This is also a useful parameter to vary in order to investigate and verify the adaptive MVDR-CRM-method. The reflections are included in two different orders. The first is by ordering the reflections by increasing time-delays. The second is by ordering the reflections by the angle between the reflection and the front-side-direction.

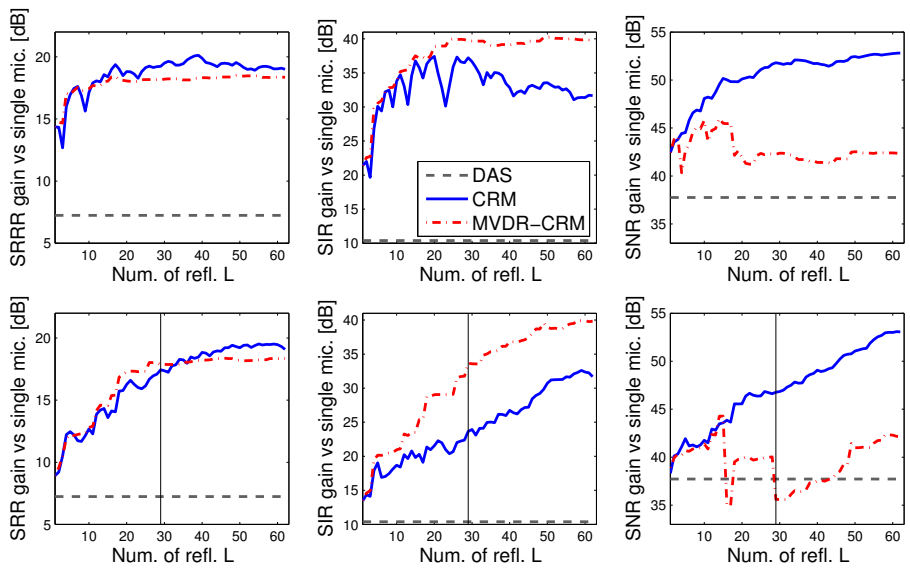


Figure 5.7: The beamformers performance on increasing number of reflections. Top: reflections are ordered by increasing time delays. Bottom: reflections are ordered by angle. The black line marks where the reflections at the back of the array are starting to be added. This result is generated with test setup 2: A small room.

Figure 5.7 show the results of this test. The CRM beamformer has increasing performance on all metrics. It is somewhat unstable, jumping up and down, as new image sources are included. This happens because the image sources have varying amounts of signal, reverberation, interference, and noise.

MVDR-CRM will ignore reflection beams that make the performance drop. If possible it rearranges the weights to suppress the unwanted sound. This is visible from sudden drops and risings in the graphs. Here the beamformer increases the performance on one metrics by decreasing it at another.

It is interesting to compare the two ways of ordering. The beamformers end at the same levels for the two orderings, but sorting by distance reaches a higher level faster than if the reflections are sorted on angle. This happens since the strongest reflections have their image sources closest to the array. It tells us that the new methods works best for arrays with the possibility to steer the beamformer in all directions. Also, there is a substantial increase for CRM ($\approx$ 15 dB) on all the metrics, for a small number of sources ($L = 10$) when the image sources are ordered by distance. This shows that the proposed method could work well with a small number of reflections if it is used with an open array. If the array is in a casing, more reflections are needed to achieve the same performance.

The test shows that an open array is preferred for the proposed methods. It shows that even a small number of image sources can give an significant increase on SRR and SIR, and that MVDR-CRM makes the performance more stable.

5.1.6 In-accurate estimation of reflection strength

It is predicted in section 4.2 that the MVDR-CRM beamformer breaks down for inaccurate estimates of the signal amplitudes. To test this, a percent wise error is added to all the amplitudes in the $\boldsymbol{A}$-matrix, except for the direct amplitude:

$$\boldsymbol{A} = \begin{bmatrix} A_0 \\ A_1(1-\delta) \\ A_2(1-\delta) \\ \vdots \\ A_{L-1}(1-\delta) \end{bmatrix}$$

where δ goes from 0 (perfect estimate of $\boldsymbol{A}$) to 1 (no estimate of $\boldsymbol{A}$).

Figure 5.8 shows how the SRINR decreases for MVDR-CRM as δ goes to 1. It is interesting to see that the MVDR-CRM beamformer performs better or equally good as CRM for fairly high δ-values ($\delta < 0.7$), and DAS for even higher values ($\delta < 0.97$). The results show that the adaptive beamformer is very robust against bad amplitude estimates. One weakness with the test is that the amplitude errors are percent-wise the same on all the reflections.

Diagonal loading can be applied. The analytical calculations show that when the loading gets infinitely high, MVDR-CRM goes to CRM with the weight vector $\boldsymbol{w'} = \boldsymbol{A}$. Figure 5.8 verifies this. For this case MVDR-CRM is performing better than the weighted CRM for almost all δ-values. It makes little sense to use diagonal loading for this case, but it proves that diagonal loading can be applied to make MVDR-CRM more robust against inaccurate estimates of the amplitudes.

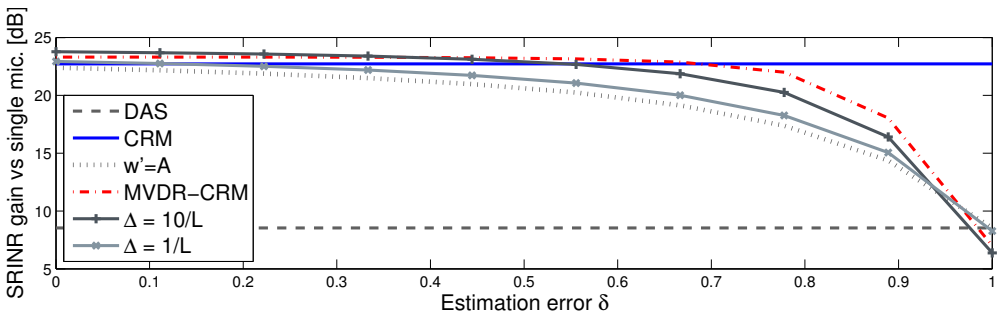


Figure 5.8: The MVDR-CRM beamformer performance on increasing error in the estimation of $\boldsymbol{A}$, using diagonal loading. This result is generated with test setup 2: A small room. Only the first 19 reflections (when sorted on angle) are used.

The test shows that an inaccurate estimate of the amplitudes degrades the performance of MVDR-CRM. The method could be made more robust by using diagonal loading, but it is not given that the performance will increase.

5.1.7 Correlated sources

Correlated sources is a familiar problem for MVDR algorithms. The weights are used so that the sound from the interfering source cancels out the direct sound. The correlation test shows how MVDR-CRM can be made robust against correlation using diagonal loading.

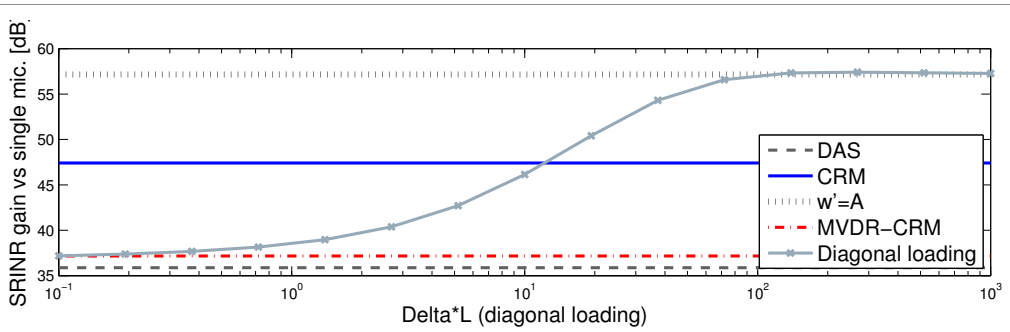


Figure 5.9: Correlated signals. This result is generated with test setup 2: A small room, the sources are emitting the same monochromatic wave signal.

Figure 5.9 shows that MVDR-CRM breaks down in the presence of correlated sources, performing much worse than CRM, and at the same level as DAS, on the SRINR metric. By applying diagonal loading, the MVDR weighting gradually goes towards $\boldsymbol{w'} = \boldsymbol{A}$.

The test shows that correlated sources degrade the performance of MVDR-CRM while diagonal loading makes it robust against this.

5.1.8 High interference and noise levels

This test investigates how the beamformers behave for increased noise or interference levels.

The level of the interference (figure 5.10) or the noise (figure 5.11) seems to have the no effect on DAS and CRM compared to the single microphone.

MVDR-CRM rearranges the weights in order to suppress the increasing interference/noise. This makes the performance on the other metrics drop as expected, since the effort is put on suppressing the interference/noise. On the interference level test, the performance on SNR is affected by the rearranging of the weights. When the weights are used to maintain a low interference level, the noise level is increased.

The test shows that the noise and interference level do not affect the beamformers, except that MVDR-CRM rearranges the weights.

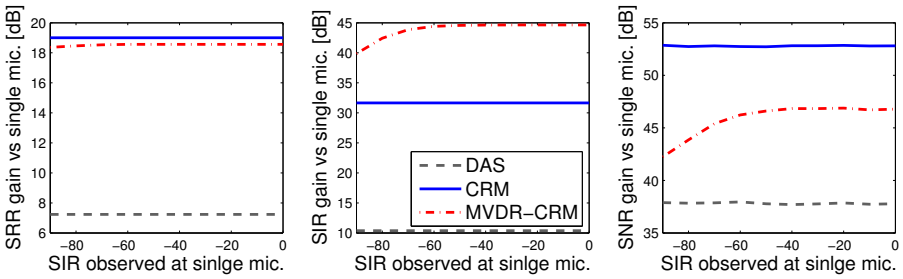


Figure 5.10: Interference level test. Note that the x-axis goes from high to low interference level. This result is generated with test setup 2: A small room.

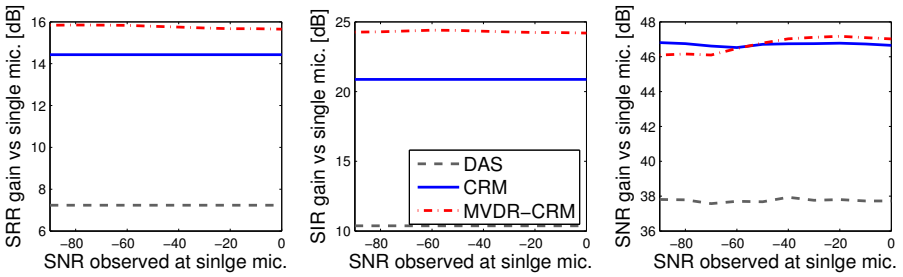


Figure 5.11: Noise level test. Note that the x-axis goes from high to low noise level. This result is generated with test setup 2: A small room.

5.1.9 Many interfering sources

By increasing the number of interfering sources, not only the interference level becomes higher, but also the spatial spectrum of the interference becomes wider. The new beamformers use DAS beams steered towards different locations in the room. While the spatial spectrum of the interference gets wider, more of these beams will have higher interference. The test shows how this affects the beamformers.

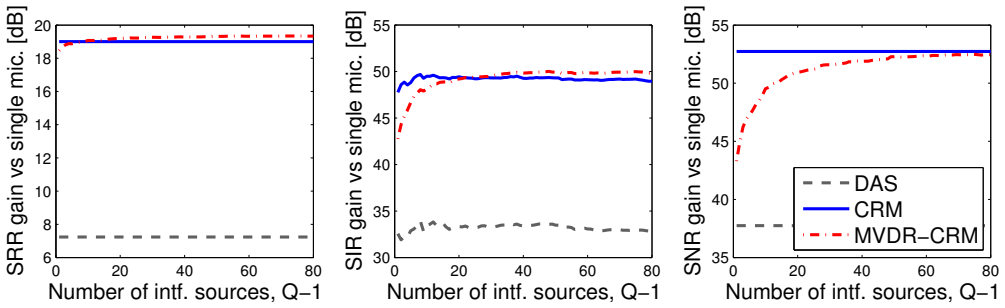


Figure 5.12: Increasing number of randomly distributed interfering sources. This result is generated with test setup 2: A small room. The interfering sources used are listed in appendix 6.4.3 and emit white noise, uncorrelated from source to source.

Figure 5.12 shows that the DAS and CRM beamformers do not loose performance relative to the single microphone. The MVDR-CRM beamformer rearranges the weights to suppress the interference. It is interesting that the noise suppression also improves. This is because the interfering sources (which emit white noise) become more spatially white (like the noise component).

The test shows that the performance of CRM is stable compared to a single microphone on a spatial "whitening" of the interference.

5.1.10 In-accurate estimation of source position

By inserting an error in the information given to the beamformer about the source location, the sensitivity near the center of the beam can be explored. A narrow sensitivity can be both good and bad depending on the situation. If we have perfect knowledge about the image source positions, then a narrow sensitivity means good suppression of nearby sources. If the knowledge about the source position is uncertain, then a wider sensitivity is required to ensure that the source of interest is amplified.

Figure 5.13 shows the results for this test. The first interesting finding is that the DAS and CRM curves have almost the same shape within the signal-ratios. This tells us that it the signal part of the sound is very sensitive to an inaccurate estimation of the source position. The unwanted sound in the output, reverberation, interference and noise, is much less sensitive tthis.

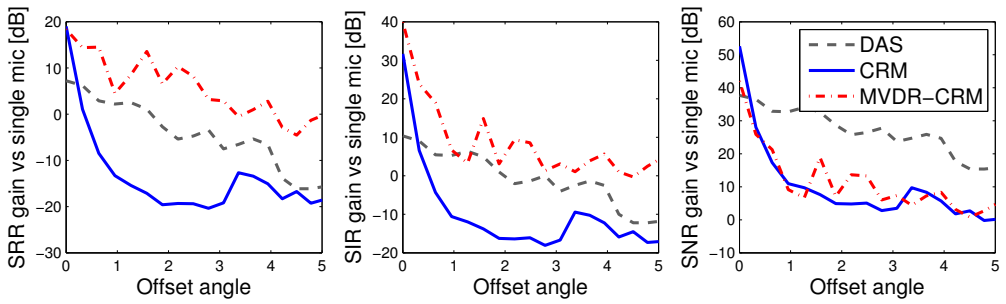


Figure 5.13: Error in source position information. This result is generated with test setup 2: A small room. The beamformers are gradually steered away from the actual source position.

The DAS beamformer gradually decrease on the signal-ratios. This indicates that the sensitivity for DAS is wide.

CRM drops significantly on the signal-metrics, 40dB, when the error is larger than 2°. Considering the beamformed RIRs in each reflection beam. When the reflection beam is steered in a slightly wrong direction, the signal spikes in the RIRs do not get aligned. This makes the output RIR consist of several spikes placed closely together. The signal part of the output is defined as the highest spike, and this is what is causing the dramatic drop in the SRR, SIR and SNR curves. (A comment to this: The early-ratios, see figure 5.14, is less sensitive to this since the early-part of the RIR is the largest spike including the spikes within 25ms at each side of it.)

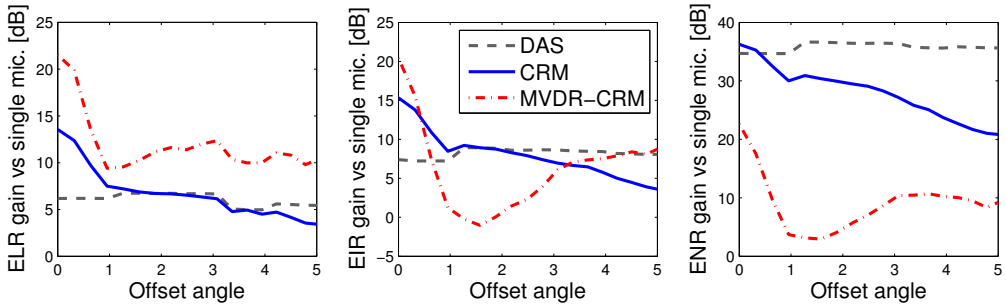


Figure 5.14: The same test as showed in figure 5.13 but using the early metrics.

MVDR-CRM again suppresses the low leveled noise to increase the SRR and SIR. This makes the sensitivity wider than CRM on SRR and SIR, but narrower for the SNR.

The test shows that CRM suppresses the sources near the beam-center more than DAS, while MVDR-CRM suppresses them less than DAS.

5.2 Simulations: beamformer output

5.2.1 Impulse response and frequency spectrum

As we see from the plots, in figure 5.15, there is a substantial de-reverberation effect on the beamformers. While CRM and offers a flatter frequency spectrum, MVDR has a large impact on the frequency content of the signal compared to DAS.

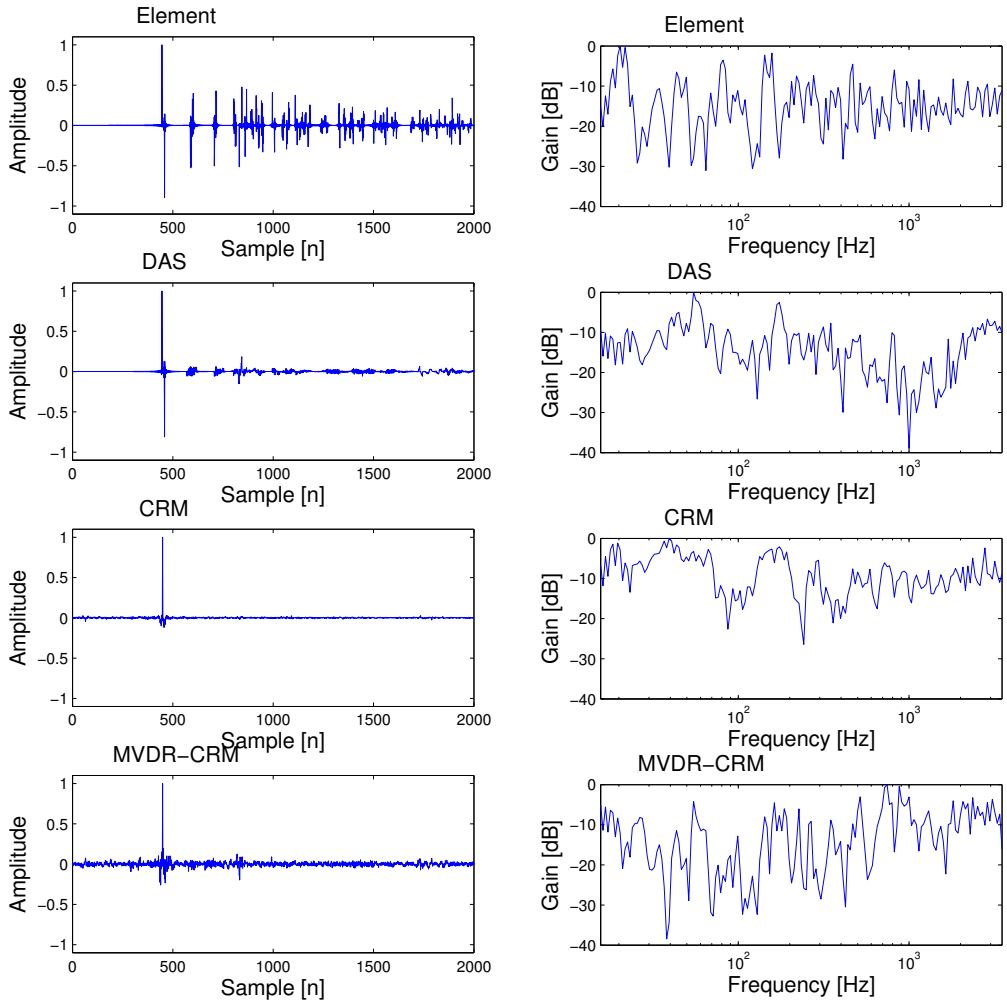


Figure 5.15: The beamformers impulse responses (left) from a source when no noise or interference is present. (It looks like MVDR has a higher power on its output than CRM but the amplitude decreases dramatically for the samples that is outside the plot.) The plots to the right are the frequency spectra of the impulse responses using a 1/9-th octave smoothing. The plot is created using 70 image sources + the original source in CRM.

5.2.2 Sound samples

Some listening samples from the simulations with test setup 2 are available in a zip-file at: *tiny.cc/og8wcx*.

The recordings are generated by

- Convolving two different talk signal with the source and the interfering source RIRs generated with ISM.
- White Gaussian noise is added.
- The components are gained until the mean SIR and SNR on the array elements is -10 dB.
- The source, interference source, and noise are added together.
- Lowpass filtering the total sound with a cutoff frequency at 3500 Hz to avoid aliasing.
- The beamformers are applied on the data using 63 reflections. This corresponds to all the 1st, 2nd and 3rd order reflections.

This file is 7.5 MB and contains 5 wav-files with 20 seconds of sound sampled at 41000Hz:

- *simulation_emitted_signal.wav* - the signal emitted from the source of interest
- *simulation_single_mic.wav* - the sound recorded at the closest array element
- *simulation_DAS.wav* - the DAS output
- *simulation_CRM.wav* - the CRM output
- *simulation_MVDR_CRM.wav* - the MVDR-CRM output

No controlled listening test has been performed. In the author's subjective opinion, the recordings show that the speech intelligibility is increased on CRM and MVDR-CRM outputs compared to DAS. The reader is encouraged to listen to the recordings to verify of falsify this statement.

5.3 Measurements: Verification of simulations

The measurements are conducted to verify that the simulated results are possible to reproduce on real data. The analysis will not be as complete as for the simulations. Only some of the tests will be performed.

5.3.1 Noise analysis

The noise analysis on the recorded data, see figure 5.16 (top), shows that the noise is white in temporal frequency, with an exception at high frequencies. The autocorrelation of the recorded noise on element 5 (middle) shows that the noise is uncorrelated in time. This is what was used for the noise components in the simulation. On the cross-correlation between two channels (bottom), we see that there are some sign of correlation between the two signals. This is not the case for the noise used in the simulations. It indicates that there are some interference (maybe from heat, ventilation, air conditioning (HVAC)-systems) in addition to the controlled interfering source.

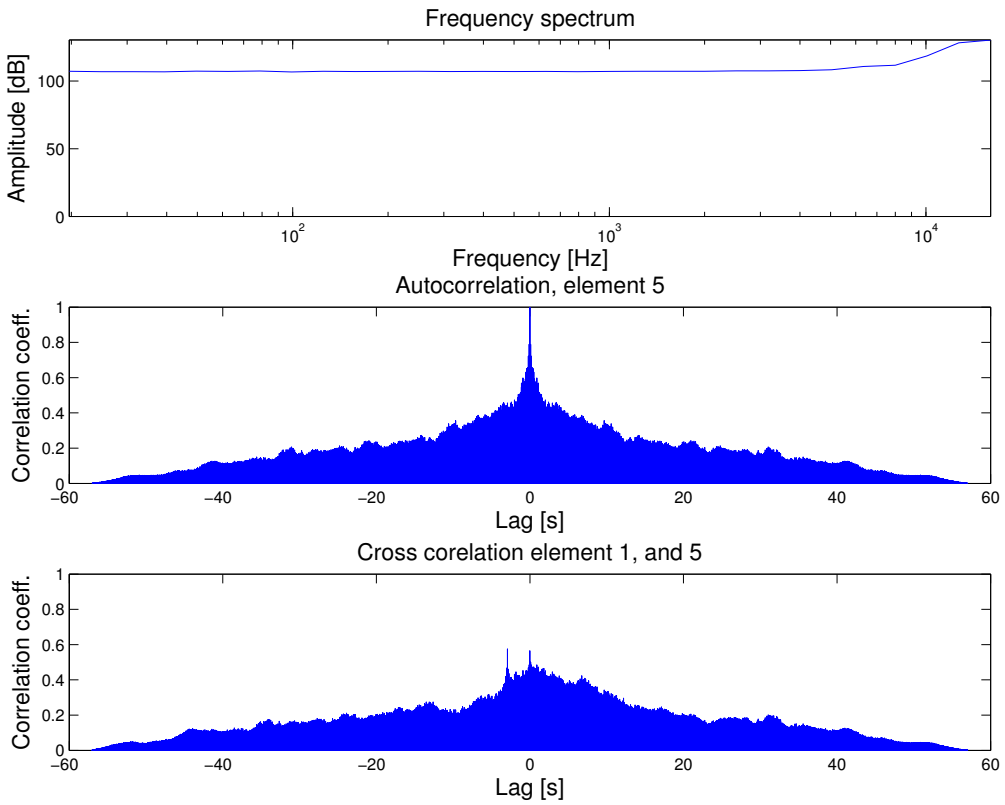


Figure 5.16: Noise analysis

5.3.2 Number of reflections

As for the simulations, a test where CRM and MVDR-CRM are using an increasing number for reflections is done for the measured data. Figure 5.17 shows the SRINR for both the simulated and real data. Only reflections in front of the array are used since the array used in the simulations has a closed casing.

We see that the DAS beamformer is performing worse on the simulated case than for the measured case. This can be due to two things. It is possible that the comparison element has lower sensitivity in the real array than the other elements. The reason can be that the simulations are done with an open array, meaning that the DAS beamformer is more sensitive to unwanted sound coming from the back of the array.

CRM is mostly performing worse than DAS on the both the data sets. On the real data it is performing lower than on the measured data. This may be due to the array casing which may introduce shadowing and resonances. Another explanation is the correlated noise on the array elements, and therefore added coherently in the CRM beamformer.

MVDR-CRM shows a drastic increase when the 8-th reflection is included in the beamformer. The fact that the same effect occurs on both the simulated and the measured data indicates that the method is behaving as expected on the measured data.

The test shows that the beamformers have similar trends on the measured and simulated data, although there are some significant deviations between them.

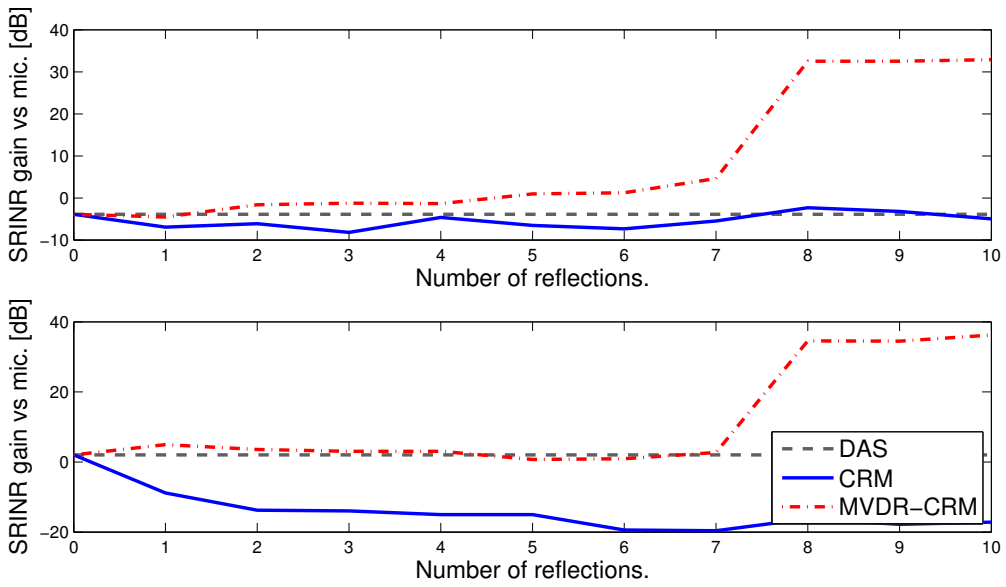


Figure 5.17: The beamformers performance as a function of increasing number of reflections. Simulations on top, measurements on bottom.

5.3.3 Output comparison

The output comparison gives us a possibility to visually inspect the performance of the beamformers. For For the next sections the experiment performed with test setup 3: a real room is analyzed. The 256-element array were used and the source were emitting a monochromatic sine. The noise component was gained fairly high. 10 reflections were used in the CRM algorithms.

The beamformer outputs are visualized in figure 5.18. Note that the y and x ax3s are different for the source-plots and the other plots to be able to see the signal and reverberation. The signal and reverberation are visualized together as the RIR measurement where inaccurate and following made it hard to separate the two.

The plots show that the DAS beamformer does not suppress the noise/interference components at all. The signal/reverberation component is smaller than for the single mic but also less distorted. This shows the dereverberating effect of DAS.

CRM increases its signal component, but also increases the noise/interference component.

The MVDR-CRM method does not increase the signal-component relative to DAS but manages to cancel out the noise and interference components. This is an interesting finding since it shows that the **MVDR uses the weights to cancel unwanted sound**. The next section investigates how MVDR-CRM does this.

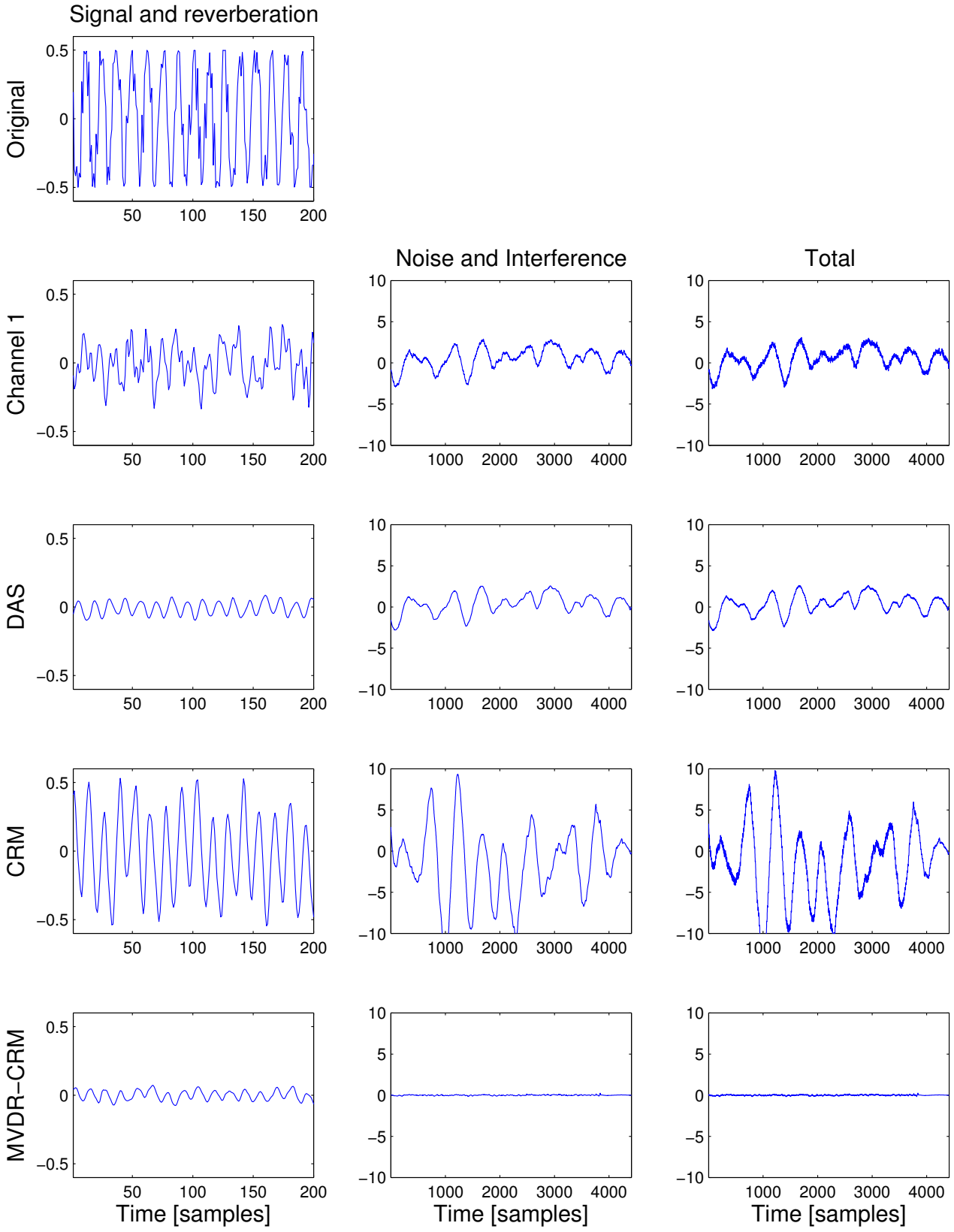


Figure 5.18: The beamformer output on the measured data when 11 reflections are used.

5.3.4 Investigating the MVDR-CRM weights

The two components that were possible to seperate were signal/reflection and noise/interference. By plotting these two multiplied by the MVDR weights, we can observe how MVDR-CRM uses two beams to cancel out unwanted sound.

Figure 5.19 shows that there are mainly five reflections that are used: the 3th, 7th, 8th, 9th and 10th (the direct beam is weighted down). Reflection 3 and 8 have approximately the same interference/noise component, but phase shifted by 180° for the 8th reflection. This is how MVDR cancels out the interference/noise component. Reflection 3 has more signal than the other reflections. This is where the main part of the signal/reverberation originates from is kept. This explains the large increase in SRINR when the 8th reflection is included in figure 5.17.

A small increase is visible when the 10th reflection is included. Here we observe the same phenomenom, the 7-th and 9-th reflection cancels out the noise in the 11th reflection, leaving the signal part.

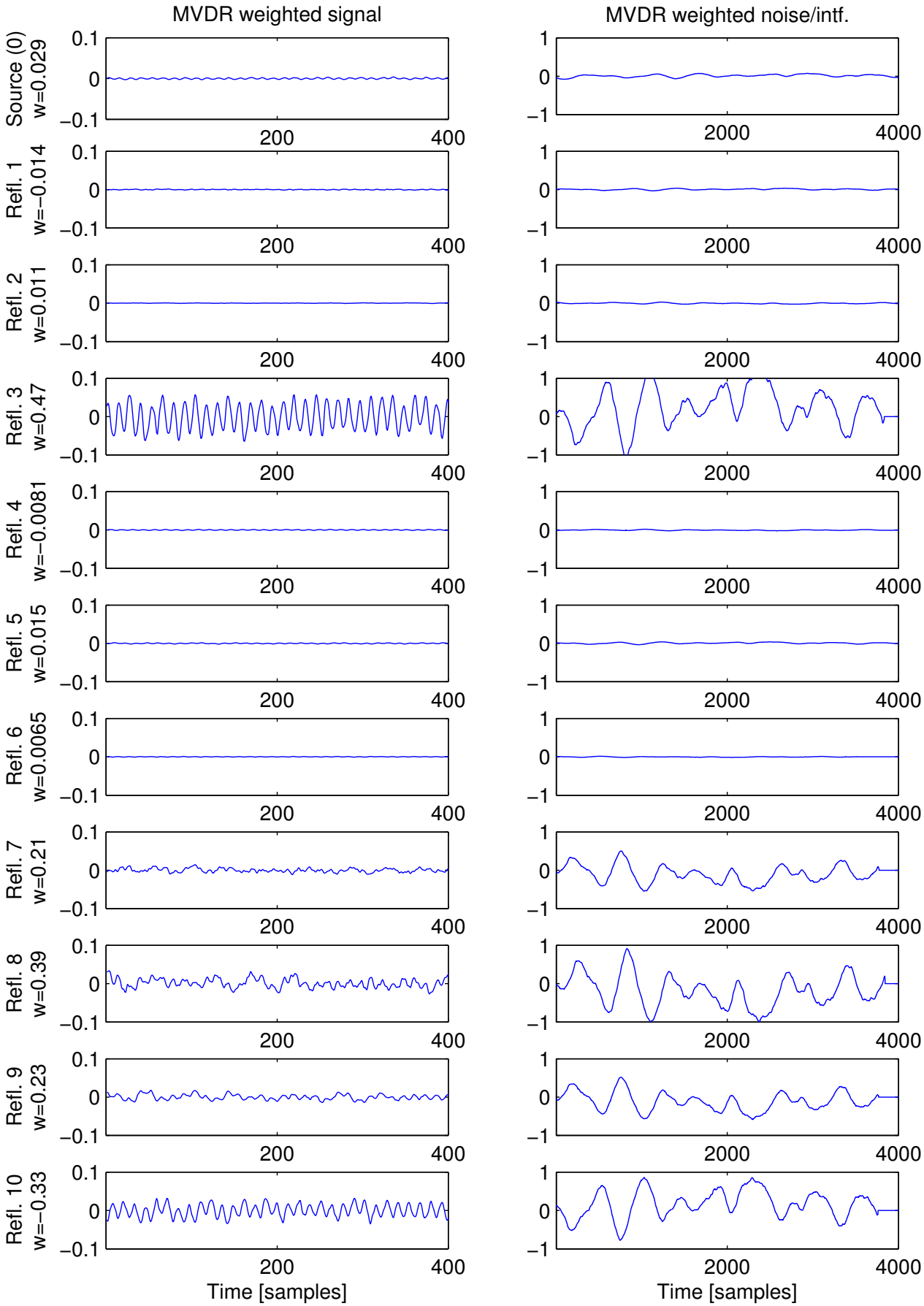


Figure 5.19: The MVDR-CRM weights used on the beamformer output in figure 5.18.

5.3.5 Uncertainties in the measurements

Separation of signal and reflection To find the increase in the signal-ratios the signal had to be separated from the reverberation. This was done by performing a RIR-estimation using the swept sine method. This introduces several uncertainties. The speaker and microphone responses are not separable from the room response. When the signal-peak is "cut out" from the RIR it is assumed that the rest of the RIR is reverberation. Some of this is not reverberation but originate from speaker and microphone effects.

The effect of this is unknown since the speaker and microphone responses are unknown. The speakers are cheap and are therefore assumed to have distorting effects. They may also be non-linear, meaning that the distortion may be different when playing the swept sine for the RIR-estimation and for the sine-signal used in the beamforming algorithm. The microphones and array case may also introduce errors in the RIR-estimation.

Room The room is fairly rectangular, but there are some pipes in the ceiling and door cases. This causes some small errors in the estimation of the image source positions. These errors are considered small with respect to the 50ms "early" definition, but for the signal-measure, only the sound from the reflections that are perfectly aligned with the signal-spike will be considered as signal.

Chapter 6

Review, Summary and Further Work

6.1 Reviewing the hypotheses

The hypotheses for the new method were presented in chapter 1. In this section they are reviewed is in light of the results. The hypotheses are marked with by using *italic font* and are reviewed one by one.

The first hypothesis was regarding the non-weighted CRM:

1. *The proposed method has higher signal-ratios than DAS.*

The results showed that CRM did have higher signal-ratios than the DAS. This was true under the condition that the reverberation time was above 0.2s. This is the case in almost every room. The analytical calculations showed however that the SNR for CRM is lower than DAS levels when a high number of reflections ($\approx$1000 for the small room test case) were used in the beamformer. The maximum SNR level for CRM were found around 50-100 reflections for the small room test case.

2. *The proposed method will remove the room effects from the emitted source signal. This will be apparent in a smaller deviation from a flat frequency response compared to DAS beamformer.*

The impulse response analysis showed that CRM did remove room-effects. The increased SRR is also a measure for de-reverberation. MVDR-CRM did not have a distinctly flatter frequency response than DAS, but by listening to the recording the de-reverberation effect is detectable.

3. *Some reflections will contribute to higher SRR, some will contribute to a lower SRR.*

4. *Some reflections will contribute to higher SIR, some will contribute to a lower SIR.*

5. *All reflections will contribute to a lower SNR*

The *number of reflections*-test on the simulated data showed that this hypothesis were partly correct. The SRR and SIR contributions of each reflections were both constructive and destructive. This also was the case for the SNR measure, proving hypothesis 5 wrong.

6. *The adaptive algorithm will give better and more stable signal-ratios than the non-adaptive one.*

This was shown to be partially correct. The signal-ratios for MVDR-CRM were more stable than for CRM on increasing the number of reflections. MVDR-CRM performed as good as, or better, on the total signal-ratio (SRINR), while on the individual signal-ratios (SRR, SIR and SNR) the performance was not always better than for CRM. MVDR-CRM would decrease one of them in order to increase another if it increased the total signal-ratio SRINR.

6.2 Reviewing the assumptions

Some of the assumptions presented in chapter 3 have been challenged during the experiments. This section gives an overview over the findings and suggestions to how the new method could become independent of some of the assumptions.

Assumptions that must be valid

- *Spatial position of the source of interest is known.*
- *The spatial positions of the image sources of the source of interest are known.*

The tests show that for CRM it is very important that the exact locations of the source and reflections are known. Only the reflection points used in the beamformer need to be known. MVDR-CRM is less sensitive to this than CRM.

- *The sources are stationary, i.e. they do not move within the timeframe in which they are observed.*
- *The environment is stationary, meaning that reflecting surfaces do not move and the wave speed is constant.*

For the same reason as the first assumption this is a problem for the proposed beamformers. If the trajectory of the moving source is known, and the reflection points for the entire trajectory is known, the beamformer could be dynamically steered towards the source as it moves.

- *The sound emitted from the source of interest is uncorrelated with the sound emitted from interfering sources.*

Correlation between interfering sources caused MVDR-CRM to break down. This can be fixed by using diagonal loading on the $\boldsymbol{R}'$ matrix. CRM has no problem with correlated sources.

Assumptions that do not have to be valid

- *There is no air absorption present.*

The reflection amplitudes needs to be know for the MVDR-CRM method. If the air absorption can be estimated, it can be included in the estimations of the amplitudes.

- *We want to enhance speech which is band limited between 100Hz and 3500Hz.*

Both CRM and MVDR-CRM seems to work for any frequency, as long as it is below the Nyquist frequency.

Assumptions that have not been tested

- *The goal is to listen to a source, not determine DOA.*
- *The source is omnidirectional.*
- *The array has omnidirectional microphone elements.*
- *The array is working and there are no defective elements in the array.*
- *The absorption coefficients for the room is frequency independent.*
- *The array is well sampled ($d \leq \frac{\lambda_{min}}{2}$).*
- *The sampling rate is high enough to get a precise delay, $Fs \geq 10 f_{max}$.*

6.3 Strengths and weaknesses

This section reviews the research questions presented in chapter 1:

- *Will the new method improve the speech intelligibility compared to the DAS beamformer?*
  - *If so, how much will the improvement be compared to DAS beamformer?*
  - *Under which conditions will the new method be a better choice of beamformer than DAS?*
  - *Under which conditions will DAS be the better beamformer?*

CRMs strengths
The main strength for CRM is the higher signal ratios (compared to DAS). This results in better de-reverberation, better spatial filtering (more suppressing of unwanted sources, narrower beam) and an improved array gain.

CRMs weaknesses
Given the assumptions from chapter 3, the most fundamental prerequisite for CRM is that there are strong reflections. It will only work in a reverberating environment. It is likely that the method would not work as well in a large room because it would mean weaker reflections.

The source and image source positions needs to be very precisely known to maintain the good signal ratios. (The speech intelligibility, measured through the early-ratios, is less sensitive to this.)

Also, the array should have an open construction to be able to include the reflections behind the array. This also poses a question about how the algorithm would work in a furnished room where the reflections may not be uniformly positioned.

MVDR-CRMs strengths
The greatest strength for MVDR-CRM is the same as for CRM, the signal-ratios increases. In addition it is more robust against reflection beams with much unwanted sound.

MVDR-CRMs weaknesses
The simulations showed that there are some challenges for MVDR-CRM, but they where less dramatic than assumed. When an interfering source is emitting a correlated signal with the source of interest, the performance of MVDR-CRM degrades. Although it was robust on a small error in $\boldsymbol{A}$-estimate the performance dropped when the error became large. This also represents a weakness for MVDR.

6.4 Further work

6.4.1 Blind estimation of reflection points

This method would be very attractive if it could be combined with a blind estimation of the reflection points and their time of arrivals relative to the direct sound. Then the CRM beamformer could be used in a room without any a priori knowledge about the room. In "Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing." [40] the authors investigate methods for localizing the reflection points using different array processing algorithms. On a real measurement, using a 32 element spherical array, they achieve localization of 5 reflections with a precision of 4°, when there are no interfering sources present. The *number of reflections*-test showed than even a small number of reflections in the CRM beamformer could give a substantial increase in the signal-ratios.

Another possibility is to investigate how the image sources can be extracted from an RIR measurement.

6.4.2 Extending the MVDR-CRM

The measurements showed that MVDR-CRM used reflections to remove low frequency interference/noise. This observation leads to ideas for other variants of the beamformer:

MVDR for frequency bands The first idea is to formulate an MVDR algorithm that works on narrow frequency bands. The output from each reflection beam is split into frequency bands, and there is individual MVDR weighting for each band. MVDR could be formulated in the frequency domain giving the possibility to have complex weights.

MVDR-CRM on sensor output before beamforming The MVDR-CRM for a single microphone is interesting. It does not require information about where the reflections are positioned; only their relative time delay to the direct sound. This could be seen in a plot of the RIR. The algorithm could be used on a single microphone or on each element of an array in advance of the beamforming. Each element must not use the same reflection points.

6.4.3 More testing of the methods

The simulations and measurements in this thesis are limited. To verify that the algorithm works under other conditions different scenarios should be explored.

Different room configurations The simulations were done for only one type of room. The experiments should be done for larger rooms, to see how this affects the beamformer performance. Also it would be interesting to conduct more real-life experiments, for example in a furnished room with real people as talking sources.

Use CRM with other methods As discussed in section 3.2.4 there are multiple other methods that could be used to create the reflection beams for CRM and for comparison. Since the reflection points are known, especially the positioning of zeros is interesting. This would remove the most prominent reverberation in both the direct beam and in the reflection beams.

List of Symbols

i	- the imaginary unit
t, T	- time
Fs	- the temporal sampling frequency
$\vec{}$	- denotes a three dimensional vector (i.e $\vec{x} = (x, y, z)$ or $\vec{k} = (k_x, k_y, k_z)$)
^{H}	- the hermitian operator

Coordinate system

$\vec{x} = (x, y, z)$- the three dimensional spatial position vector (cartesian coordinates)

$\vec{\Omega} = (r, \theta, \phi)$- the spatial position vector (spherical coordinates), where

r	- radius (spherical coordinates)
θ	- elevation angle (z) (spherical coordinates)
ϕ	- angle in the horizontal plane (x, y) (spherical coordinates)
	(') denotes the array centered coordinate system, else the symbol refers to the room centered coordinate system.

Transformation from room centered to array centered

α	- rotation about x-axis
β	- rotation about y-axis
γ	- rotation about z-axis
$\boldsymbol{R_x}$	- transformation matrix with rotation about x-axis
$\boldsymbol{R_y}$	- transformation matrix with rotation about y-axis
$\boldsymbol{R_z}$	- transformation matrix with rotation about z-axis
$\boldsymbol{T}$	- transformation matrix with translation

Wave properties

R - the fraction of the amplitude of a wave being reflected at a surface

Z - the wave impedance

ρ - the density of the medium the wave is traveling in

P - pressure

A - amplitude of the wave

ω - the angular frequency of a wave

$\vec{k}$ - the wave number

$c/$ - the wave speed

λ - the wavelength of a wave

θ - angle

Acoustic

α - absorbtion coefficient

L_p - Sound pressure level

p_0 - atmospheric pressure

p_{rms} - root mean square of the wave pressure amplitude

$T60$ - The reverberation time

$C50$ - Clarity

STI - Speech transmission index

Indexing

N - number of samples (in time)

n - counting variable from 0 to $N-1$

M - number of elements in an array

m - counting variable from 0 to $M-1$

J - number of image sources

j - counting variable from 0 to $J-1$

L - number of image sources included in the SODAS beamformer

l - counting variable from 0 to $L-1$

Q - number of sound sources in room

q - counting variable from 0 to $Q-1$

Array signal processing

D - Array size

d - spacing between elements in a regular array

$\vec{x}_m$ - spatial position vector of the m-th element

$\vec{x}_source$ - spatial position vector of the source

$\vec{x}_j$ - spatial position vector of the j-th image source

$\vec{x}_{phase\,center}$ - the spatial position vector of the phase center of an array

$y_m(t)$ - recorded signal at the m-th element

w_m - the weighting of the m-th element

$z_j(t)/z_l(t)$ - DAS output for image source j/l

$z(t)$ - $(= z_0(t))$ DAS output when steered to original source $(j/l = 0)$

$z'(t)$ - SODAS beamformer output

P - power of beamformer output

E - Expectation value

σ^2 - Variance

δ - dirac delta function

G - array gain (SNR)

Array signal processing: vector notation

$\boldsymbol{y}(t)$ - a vector containing the delayed output for all (including the virtual) the elements

$\boldsymbol{z}(t)$ - a vector containing the delayed first order DAS outputs

$\boldsymbol{w}$ - a vector containing the weights for DAS

$\boldsymbol{w'}$ - a vector containing the weights for SODAS

$\boldsymbol{R}$ - the covariance matrix for $\boldsymbol{y}$

$\boldsymbol{R'}$ - the covariance matrix for $\boldsymbol{p \cdot z}$

Δ - diagonal loading constant

ϵ - diagonal loading factor

$\boldsymbol{I}$ - identity matrix

Array signal processing: time delays

τ_m - time delay from the source to the m-th sensor (anechoic model)

$\tau_{m,j}/\tau_{m,l}$ - time delay from image source j/l to the m-th sensor (reverberant model)

Image source method

$\hat{x}_j/\hat{x}_l$ - the spatial position vector of the j/l-th image source

p_j/p_l - the phase shift of the j/l-th image source

$\boldsymbol{p}$ - a vector containing all p_j/p_l

A_j/A_l - the amplitude of the j/l-th image source

$\boldsymbol{A}$ - a vector containing all A_j/A_l

R_j - the attenuation factor of an image source due to reflection losses

Signal model

$s_q(t)$ - the signal emitted from a source q

$y_m^s(t)$ - signal part of output on element m

$y_m^r(t)$ - reflection part of output on element m

$y_m^i(t)$ - interference part of output on element m

$y_m^n(t)$ - noise (spatially uncorrelated) part of output of element m

Signal model: vector notation

$\boldsymbol{y}^s(t)$ - a vector containing the signal part on all elements

$\boldsymbol{y}^r(t)$ - a vector containing the reflection part on all elements

$\boldsymbol{y}^i(t)$ - a vector containing the interference part on all elements

$\boldsymbol{y}^n(t)$ - a vector containing the noise (spatially uncorrelated) part of all elements

$\boldsymbol{z}^s(t)$ - a vector containing the signal part on all DAS outputs

$\boldsymbol{z}^r(t)$ - a vector containing the reflection part on all DAS outputs

$\boldsymbol{z}^i(t)$ - a vector containing the interference part on all DAS outputs

$\boldsymbol{z}^n(t)$ - a vector containing the noise part of all DAS outputs

$\boldsymbol{z}'^s(t)$ - a vector containing the signal part on the CRM output

$\boldsymbol{z}'^r(t)$ - a vector containing the reflection part on the CRM output

$\boldsymbol{z}'^i(t)$ - a vector containing the interference part on the CRM output

$\boldsymbol{z}'^n(t)$ - a vector containing the noise part of the CRM output

Speech enhancement

$h_{0,m}(t)$ - room impulse response from source 0 to element m

$\hat{h}_{0,m}(t)$ - estimate of $h_{0,m}(t)$

$h(t)$ - beamformer room impulse response from source 0

$\hat{h}(t)$ - estimate of $h(t)$

$h^n_m(t)$ - noise cancelation filter for the noise on element m

$\boldsymbol{h}^n(t)$ - a vector containing all M $h^n_m(t)$

$h^n(t)$ - noise cancelation filter for the noise on beamformer output

$h^i_m(t)$ - interference cancelation filter for the noise on element m

List of Acronyms

CRM constructive reflections method

DAS delay and sum

DOA directions of arrival

ELR early to late ratio

EIR early to interference ratio

ENR early to noise ratio

FIR finite impulse response

GSC generalized sidelobe canceller

HVAC heat, ventilation, air conditioning

IR impulse response

ISM image source method

MVDR minimum power/distortionless response

MVDR-CRM minimum power/distortionless response CRM

MP/MV minimum power/minimum variance

RIR room impulse response

SIR signal to interference ratio

SNR signal to noise ratio

SRR signal to reverberation ratio

SRINR signal to reverberation, interference and noise ratio

STIPA speech transmission index for public address systems

TDOA time differences of arrival

ULA uniform linear array

URA uniform rectangular array

Appendix A: MATLAB source code

The MVDR-CRM beamformer

```
function [MVDR_CRM_out] = MVDR_CRM(data, array, img_srcs, phases, ...
    signal_amplitudes, diag_loading_d, DAS_weights)
% MVDR-CRM beamformer
%
% data : Each row is data on a single channel (M x N)
% array : Positions of the mics (M x 3)
% img_srcs : Positions for the image sources (L x 3)
% phases : Phase hifts, p_l  (L x 1)
% signal_amplitudes : Estimate of the signal strengths (L x 1)
% diag_loading_d : Diagonal loading (0 = no loading)
% DAS_weights : Weights for DAS beams, w'_{m,l} (M x L)
A = signal_amplitudes;

%Get the reflection beams (multiplied by the phase shift
CRM_weights = ones(size(img_srcs,1), 1);
[¬,DAS_beams] = CRM(data, array, img_srcs, phases, DAS_weights, ...
    CRM_weights);

%Calculate R
R = DAS_beams * DAS_beams' / size(DAS_beams,2);

%Diagonal loading
I = eye(size(R,1));
d = diag_loading_d*trace(R);
R = d*I+R;

%Inverse of R
R_inv = inv(R);

%The MVDR weights
w = (R_inv * A) / (A' * R_inv * A)

%Normalize the weigths
w = w./sum(w);

%Apply the weights and sum the output
MVDR_CRM_out = w'*DAS_beams;
end
```

The CRM beamformer

```
function [CRM_out, DAS_beams] = CRM(data, array, img_srcs, phases, ...
    DAS_weights, CRM_weights)
% CRM
%
% data : Each row is data on a single channel (M x N)
% array : Positions of the mics (M x 3)
% img_srcs : Positions for the image sources (L x 3)
% phases : Phase shifts, p_l  (L x 1)
% DAS_weights : Weights for DAS beams, w'_{m,l} (M x L)
% CRM_weights : Weights for the DAS beams, w'_l (L x 1)


global c fs

% Extract important sizes
M = size(array,1); %array size
L = size(img_srcs,1); %Number of reflections

%Splitting matrix for better readabillity
xm = array(:,1);      ym = array(:,2);      zm = array(:,3);

%Beamform array with DAS to each image source
DAS_beams = {};

%Loop through image sources and beamform with DAS
for l = 1:L;
    src = img_srcs(l,:);
    xl=src(1);  yl=src(2); zl=src(3);

    %The DAS weights (w_{m,l})
    w_m = DAS_weights(l,:)'./sum(DAS_weights(l,:));

    %the delay from source to receiver element in samples
    df = round(sqrt((xm-xl).^2 + (ym-yl).^2 + (zm-zl).^2 )/c*fs);

    %Delay and sum
    out = frameshift_signals(data, df);
    out = w_m'*out;

    %Phaceshift factor:
    DAS_beams{l,1} = out.*phases(l);
end

%Convert cell-array to matrix
DAS_beams = cell2mat2D(DAS_beams);

%Sum the DAS beams (they are already delayed)
CRM_out = CRM_weights'*DAS_beams;

end
```

Utility functions

```
function [ a ] = cell2mat2D( c )
%Takes a 2d cellarray as an argument and convert it to a array. ...
    Pads with
%zeros at the end of the rows if nessecary.
if size(c,2) == 1
    [max_size, max_index] = max(cellfun('size', c, 1));
    RIRs = zeros(length(c),max_size);
    for i = 1 : length(c)
        a(i,1:length(cell2mat(c(i)))) = cell2mat(c(i));
    end
else
    max_size = max(cellfun('size', c, 1));
    a = zeros(length(c),max_size);
    for i = 1 : length(a)
        el = cell2mat(c(i));
        a(i,1:length(el)) = el;
    end
end
```

```
function data = frameshift_signals(data, Δ)
% Edited version from code handed out in INF5410-project work
% -------
% data : Data matrix. each row is a single signal
% Δ : samples to shift the signals
%
% Outdata
% -------
% data : the shifted matrix of signals.

[M, N] = size(data);


for (ii = 1:M)
    if n(ii) > 0
        data(ii,:) = [data(ii,(n(ii)+1):end)   zeros(1, n(ii))];
    elseif n(ii) < 0
        data(ii,:) = [zeros(1,-n(ii))    data(ii,1:(end+n(ii)))];
    else
        data(ii,:) = data(ii,:);
    end
end
```

Appendix B: Position of sources

The source positions are generated with the MATLAB function:

```
1 bsxfun( @plus, bsxfun(@times, rand(500,3), [3.5,4,1.5]), [1,0.5,1]);
```

This creates random source positions within a rectangular box of the room. The spaces close to the wall and the array are not populated with sources, since it is unlikely that the beamformers would be steered here. The random positions for the interference and source are available as a MATLAB *.mat*-file at: *https://db.tt/oRiBDLiA*

References

[1] *The American Heritage Dictionary of the English Language,* 4th edition. Houghton Mifflin Company, 2010.

[2] Norsonic. (2014). Acoustic camera, nor848a, [Online]. Available: `http://www.norsonic.no/en/products/acoustic_camera`.

[3] Squarehead Technology AS. (2014), [Online]. Available: `http://www.sqhead.com`.

[4] "Personopplysningsloven" - from norwegian law. (2000), [Online]. Available: `http://lovdata.no/dokument/NL/lov/2000-04-14-31#KAPITTEL_1`.

[5] D. Johnson and D. E. Dudgeon, *Array Signal Processing, concepts and techniques*, 1th edition. PRT Prentice Hall, Upper Saddle River, 1993, ISBN: 0-13-048513-6.

[6] J. Billingsley and R. Kinns, "The acoustic telescope," *Journal of Sound and Vibration*, vol. 48, no. 4, pp. 485–510, 1976.

[7] J. Billingsley, *An acoustic telescope*, Aeronautical Research Council ARC 35/364, 1974, 1974.

[8] W. K. III and D. Bechert, "On the sources of wayside noise generated by high-speed trains," *Journal of Sound and Vibration*, vol. 66, no. 3, pp. 311 –332, 1979.

[9] G. Howell, A. Bradley, M. McCormick, and J. Brown, "De-dopplerization and acoustic imaging of aircraft flyover noise measurements," *Journal of Sound and Vibration*, vol. 105, no. 1, pp. 151 –167, 1986.

[10] J. Piet, U. Michel, and P. Böhning, "Localization of the acoustic sources of the a340 with a large phased microphone array during flight tests," in *8th AIAA/CEAS Aeroacoustics Conference & Exhibit*, ser. Aeroacoustics Conferences, American Institute of Aeronautics and Astronautics, 2002.

[11] U. Michel, "History of acoustic beamforming," *Proceedings of the Berlin Beamforming Conference*, pp. 1–17, 2006.

[12] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[13] O. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[14] R. Masak, "Auxiliary array processing for improved adaptive sidelobe canceller performance," in *1976 Antennas Propag. Soc. Int. Symp.*, vol. 14, Institute of Electrical and Electronics Engineers, 1976, pp. 443–446.

[15] P. Annibale, F. Antonacci, P. Bestagini, A. Brutti, A. Canclini, L. Cristoforetti, J. Filos, E. Habets, W. Kellerman, K. Kowalczyk, A. Lombard, E. Mabande, D. Markovic, P. Naylor, and M. Omologo, "The SCENIC Project: Space-Time Audio Processing for Environment-Aware Acoustic Sensing and Rendering," *131st Audio Eng. Soc. Conv.*, 2012.

[16] Y. Peled and B. Rafaely, "Method for dereverberation and noise reduction using spherical microphone arrays," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 113–116, 2010.

[17] ——, "Linearly constrained minimum variance method for spherical microphone arrays in a coherent environment," in *2011 Jt. Work. Hands-free Speech Commun. Microphone Arrays*, IEEE, 2011, pp. 86–91, ISBN: 978-1-4577-0997-5.

[18] *ISO 80000-3:2006 Quantities and units Part 3: Space and time*. International standardization organization, 2006.

[19] J. R. Wright, "Fundamentals of diffraction," *Journal of Audio Engineering Society*, vol. 45, no. 5, 1997.

[20] *ISO 2533:1975 Standard Atmosphere*. International standardization organization, 1976.

[21] Miller-Keane, *Encyclopedia and Dictionary of Medicine, Nursing, and Allied Health*, 7th edition. Saunders, 2005.

[22] H. Bass, L. Sutherland, A. Zuckerwar, D. Blackstock, and D.M.Hester; "Atmospheric absorption of sound: further developments," *Journ. Acoust. Soc. Am., vol. 97, No. 1*, Jan. 1995.

[23] L. Kinsler, A. Frey, A. Coppens, and J. Sanders, *Fundamentals of Acoustics*, 4th edition. John Wiley and Sons, New York, 2000.

[24] JBL, *Speech Intelligibility - A JBL professional Technial Note*. [Online]. Available: `https://www.jblpro.com/pub/technote/spch\_intl\_1.pdf`.

[25] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America, vol. 65(4), pp. 943-950*, April 1979.

[26] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *Journal of Acoustical Society of America*, vol. 62, p. 912, 1977.

[27] N. D. Gaubitch and P. A. Naylor, "Analysis of the dereverberation performance of microphone arrays," *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC-05)*, 2005.

[28] X. Hu, A.-Q. Hu, Q. Luo, and T.-Y. Cai, "A novel adaptive acoustic echo cancellation for teleconferencing systems," in *Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on*, vol. 2, 2002, 1005–1009 vol.2.

[29] S. Holm and K. Kristoffersen, "Analysis of worst-case phase quantization sidelobes in focused beamforming," *IEEE Transactions on ultrasonics,ferroelectrics and frequency control*, vol. 39, no. 5, September 1992.

[30] D. A. Gray, "Effect of time-delay errors on the beam pattern of a linear array," *IEEE Journal of oceanic engineering*, vol. OE. 10, no. 3, JULY 1985.

[31] S. Müller and P.Massarani, "Transfer-function measurement with sweeps," *J. Audio Eng. Soc*, vol. 49, no. 6, pp. 443–471, 2001.

[32] Oygo-Sound, *Matlab code: swept-sine analysis*. [Online]. Available: `http://www.mathworks.com/matlabcentral/fileexchange/29187-swept-sine-analysis/content/extractIR.m`.

[33] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. 2005, ISBN: 9781849960557.

[34] Meyer-Sound. (2014). Machine measures of speech intelligibility, [Online]. Available: `http://www.meyersound.com/support/papers/speech/section4.htm`.

[35] H Kuttruff, *Room acoustics*. 2000, ISBN: 9780415480215.

[36] International Telecommunication Union. (2014). Itu test signals, [Online]. Available: `http://www.itu.int/net/itu-t/sigdb/genaudio/Pseries.htm`.

[37] MathWorks. (2014). Matlab programming language and environment, [Online]. Available: `http://www.mathworks.se/products/matlab/`.

[38] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *Journal of the Acoustical Society of America, vol. 124(1), pp. 269-277*, July 2008.

[39] J.-F. Synnevåg, "Adaptive beamforming for medical ultrasound imaging," PhD thesis, University of Oslo, Department of Informatics, 2008.

[40] H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, "Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing.," *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 2828–40, 2012.