Master thesis

# Multichannel Speech Separation of Dinner Talks at an Antarctic Research Station

conducted at the
Signal Processing and Speech Communications Laboratory
Graz University of Technology, Austria

by
Ing. Martin Finsterer, BSc., 1031364

Supervisor:
Dipl.-Ing. Dr. techn. Martin Hagmüller

Head of Institute:
Univ.-Prof. Dipl.-Ing. Dr.techn. Gernot Kubin

Graz, January 19, 2020

# Abstract

Annually european research teams are dispatched to spend the winter at the Antarctic research station 'Concordia' to perform scientific experiments. Due to the extreme isolation there is an analogy to Mars missions and is therefore of interest for psychological monitoring. For that reason voice recordings of the scientists have been conducted at regular dinner meetings using two microphone arrays.

The recorded speech signals are typically superimposed by interfering speech, environmental noise or also room reverberation. The goal within this thesis lies in implementing an algorithm that extracts each speech signal from the sound mixture and computes one single track for each target speaker, so that linguists can further analyze each person individually in its psychological behaviour. Therefore beamforming, sound source localization and noise-reduction techniques are investigated in theory and in simulations before applying them to the real-world data.

The challenge was that only limited knowledge about the environment was available, as no access was given to the research station and the microphone arrays have been installed on site by the researchers, who were led by instructions remotely. This means the two arrays had to be post-calibrated, but also the sound velocity was unknown. Furthermore also ground truth or clean recordings of the speakers were not given. Despite being unaware of the surrounding, in the end an algorithm has been developed within this thesis that produces useful results.

# Zusammenfassung

Jährlich werden europäische Forschungsteams entsandt, um den Winter in der antarktischen Forschungsstation 'Concordia' zu verbringen und wissenschaftliche Experimente durchzuführen. Aufgrund der extremen Isolation gibt es eine Analogie zu Mars-Missionen und ist daher für die psychologische Überwachung von Interesse. Aus diesem Grund wurden bei regelmäßigen Dinner-Meetings mit zwei Mikrofon-Arrays Sprachaufnahmen der Wissenschaftler durchgeführt.

Die aufgezeichneten Sprachsignale werden typischerweise durch störende Sprachsignale, Umgebungsgeräusche oder auch Raumhall überlagert. Das Ziel dieser Arbeit besteht darin, einen Algorithmus zu implementieren, der jedes Sprachsignal aus dem Klanggemisch extrahiert und für jeden Zielsprecher eine einzelne Spur berechnet, so dass Linguisten jede Person individuell auf ihr psychologisches Verhalten hin analysieren können. Daher werden Techniken wie Beamforming, Schallquellenlokalisierung und Rauschunterdrückung in der Theorie und in Simulationen untersucht, bevor sie auf die realen Daten angewendet werden.

Die Herausforderung bestand darin, dass nur begrenzte Kenntnisse über die Umgebung zur Verfügung standen, da die Forschungsstation nicht zugänglich war und die Mikrofon-Arrays vor Ort von den Forschern installiert wurden, die durch Anweisungen aus der Ferne geleitet wurden. Das bedeutet, dass die beiden Arrays nachträglich kalibriert werden mussten, aber auch die Schallgeschwindigkeit war unbekannt. Weiterhin waren auch Grundwahrheiten oder Nahaufnahmen der Sprecher nicht gegeben. Obwohl die Umgebung nicht bekannt ist, wurde im Rahmen dieser Arbeit ein Algorithmus entwickelt, der nützliche Ergebnisse liefert.

# Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the sources used sources. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

date

(signature)

# Contents

# List of Figures

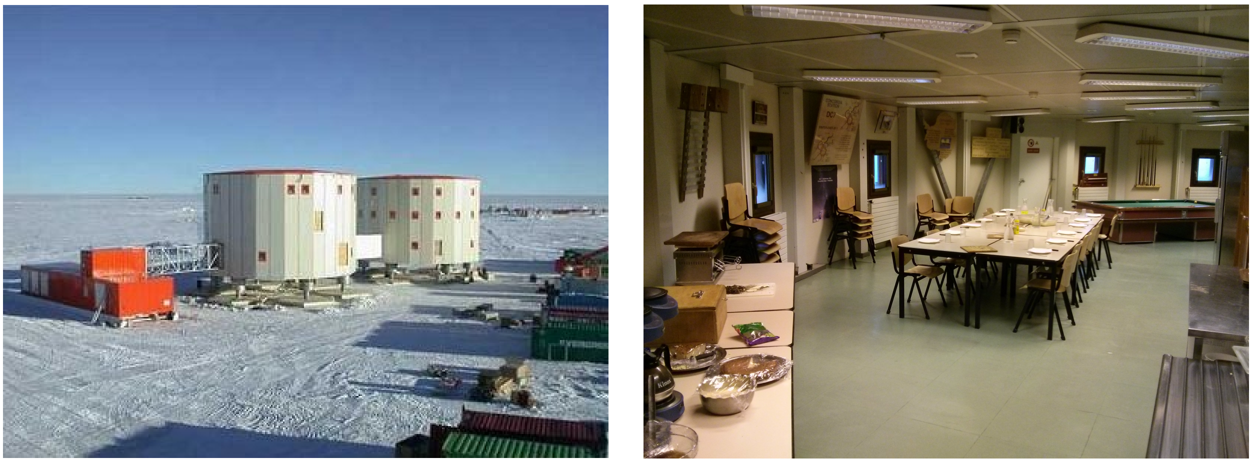
# List of Tables

# 1 Introduction

## 1.1 Background

The main goal of this master's thesis is to perform multichannel source separation and apply speech signal enhancement techniques on real-world data, by using state-of-the-art algorithms. On the way to this achievement, these methods have been investigated in theory and simulations.

The real-world data mentioned above comprises over 30 hours of 32-channel audio data, which has been recorded at the antarctic research station *Station Dome Concordia* [2], illustrated in figure 1.1 (a). The main research fields at the Concordia research station are focused on glaciology and astronomy. Besides the scientists are isolated from the rest of the world for several months, which implies psychological and physiological stress. This makes it also interesting for sociological monitoring because of its analogy to space-missions.



*(a) The building from outside*

*(b) The dinner room*

*Figure 1.1: Pictures of the research station*

In cooperation with the SPSC Lab, the scientists at the research station agreed on having their dinner talks recorded, transcribed and analyzed. The recordings have been conducted for several weeks in 10-minute-long sessions at regular meetings, where most of the researchers gathered to eat in the dinner room, which is shown in figure 1.1 (b).

The microphones have been installed on circular frames in the dinner room on the ceiling above the dinner table, depicted in figure 1.2. The design of the microphone arrays and all of the preparation for the recording was done in advance. This thesis was launched when all of the recorded data was already available at the SPSC laboratory. The main focus was to compute enhanced, separate speech tracks for each speaker at the table, so that the processed audio data can be further transcribed and analyzed by linguists.

The big challenges in this task were that no geometrical measurements could be done on site. This means that the positions of the speakers, the positions of the microphones but also the distance between the microphone arrays and their orientation were unknown. As a further handicap the room temperature also could not be measured, which has an influence on the speed of sound and as a result also on the sound wave propagation.

Furthermore ground truth is not available, which means that it is unknown which persons are talking on the recorded tracks and when they are talking. Also clean speech recordings of the scientists were not available, which would make utilization of machine learning algorithms possible. As a further thought experiment video material would have been nice to have to verify the position changes of the speakers.

But at least little information was available, which included a rough sketch with the top view of the dinner room, the common seat locations of each speaker and four clap signals with the corresponding clap positions. All of those positions were marked on the sketch approximately, but the exact coordinates were not given.

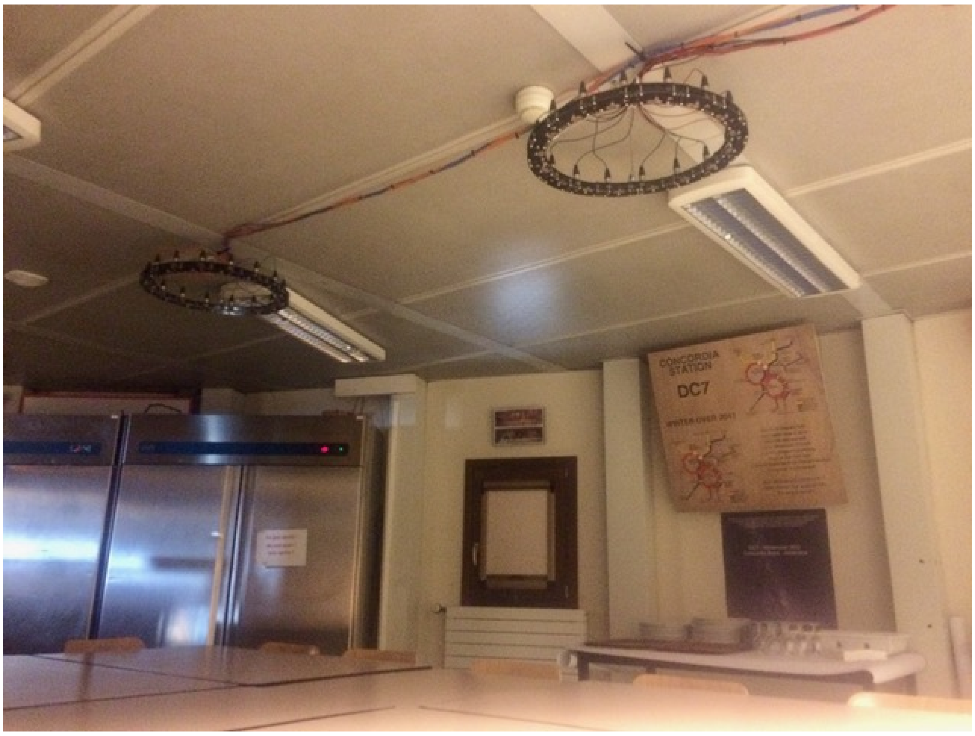


*Figure 1.2: Microphone arrays mounted above the dinner table*

Besides just processing the real-world data, a very important aspect in this thesis will be to show differences or problems that arise when comparing the real-world results with simulated experiments.

## 1.2 Cocktail Party Problem

The problem that is tried to tackle within this thesis is the so-called ”Cocktail party problem”. Usually this scenario is described as the natural ability of the human auditory system to intelligently detect, select and perceive relevant acoustic information at simultaneous presence of several sound sources e.g. human talkers. That’s where the name ”Cocktail party problem” comes from.

The effect arises from binaural mechanisms, which means hearing with both ears is required to localize sound. Interchannel differences in level and phase are substantial to directional hearing. Our auditory system is thus able to extract signals of interest from a mixture of interfering signals or noise.

Figure 1.3 illustrates a typical cocktail party scenario [3]. All sources (target, interferer, noise) possibly radiate acoustic waves directly or reflectively onto a microphone array placed at a specific location inside a reverberated enclosure.

Based on the position information of the array and the desired target, digital filters $h_m$ process the microphone array signals $x_m$ and produce a single output track containing the desired data, where $m$ is the running index of the considered microphones.

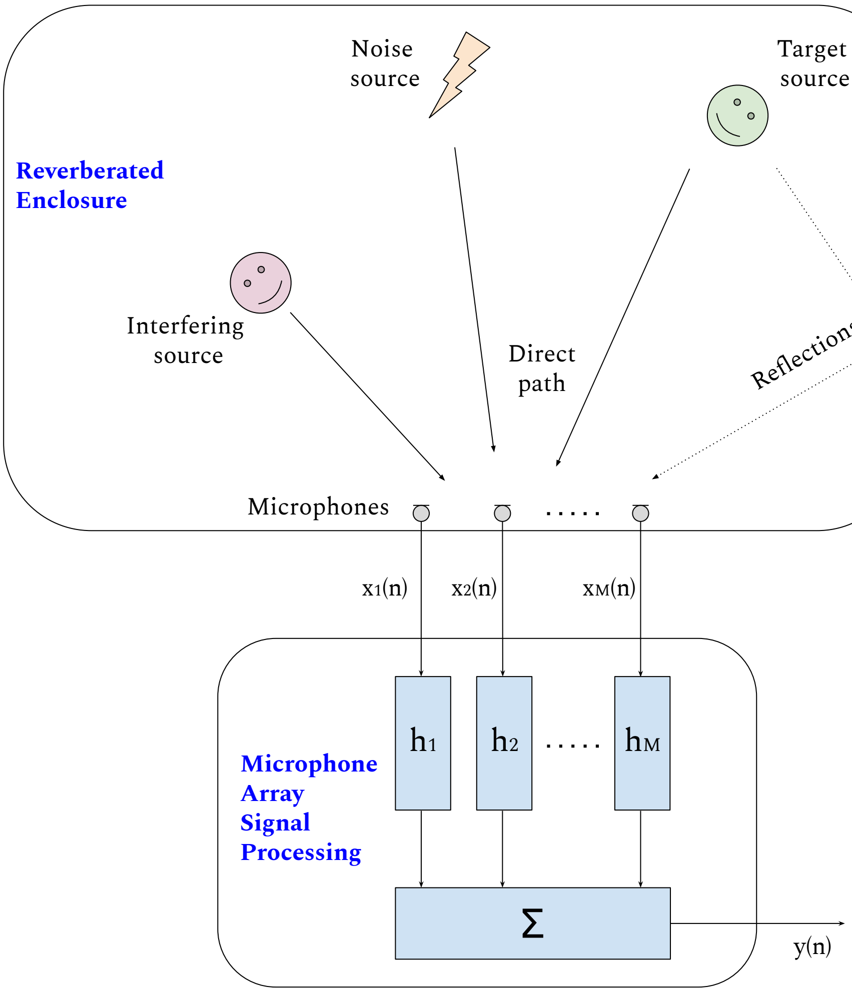


*Figure 1.3: Cocktail party effect*

## 1.3 Array Signal Processing and Source Separation

However, if microphones are used to record such a situation, those recordings would not reproduce a clear, perceptive image of the situation, because they lack of room depth and positioning information. Source separation in general is the task to teach computers on how to process recorded data to achieve signal enhancement with respect to a specific target.

These technologies are used in countless devices to improve speech signal quality such as smartphones, hearing aids, robots, automatic speech recognition (ASR), cars and surveillance. Nevertheless those mechanisms are used not only for audio signals, but also to improve signals in other disciplines like ultrasound, radar or sonar.

It has been observed, that the signal improvement increases with the number of sensors/microphones within the system. In literature it is differentiated between single-channel and multichannel signal enhancement. This thesis focuses on multichannel enhancement for speech signals.

## 1.4 Structure of this Thesis

In the beginning of the thesis the mathematical and physical foundations will be established. Subsequently beamforming (BF) will be introduced in theory as a spatial filtering method by showing different approaches. Afterwards the time-frequency masking (TFM) will be discussed as a post-filtering step to beamforming. The following chapter will be on sound source localization (SSL), which gives us the ability to control the beamformer more precisely. A simplified version of the algorithm, that will be developed within this thesis, is illustrated in figure 1.4.

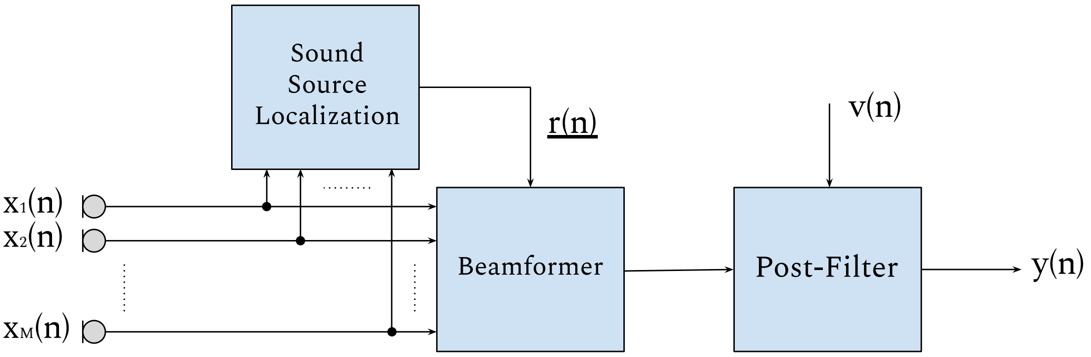


*Figure 1.4: Simplified Algorithm*

where $x_m(n)$ are the multichannel microphone signals, $\underline{r}(n)$ is the position of the detected source, $v(n)$ are estimated noise signals, which allow us to finally produce the enhanced, separated speech signal $y(n)$.

After having discussed all of the algorithms in theory, the beamforming and time-frequency masking algorithms will then be investigated in noise simulations and speech simulations. Afterwards also the real-world data will be processed using the considered algorithms and comparisons to the simulated scenarios will be drawn. At last final conclusions and an outlook will be given.

# 2 Array Processing Fundamentals

This chapter should give an overview of the fundamental theory in the acoustic domain and how the sensor signals can be combined based upon spatial information. These basics are crucial for multichannel array signal processing.

## 2.1 Math Notation Style

The used variables within this thesis are notated in the following manner:

- Scalar symbols are emphasized, e.g.: scalar $s$
- Vector symbols are underlined, e.g.: vector $\underline{v}$
- Matrices are bold, e.g.: Matrix $\mathbf{M}$

## 2.2 Coordinate System Definition

The coordinate system concerning this thesis is defined as follows.

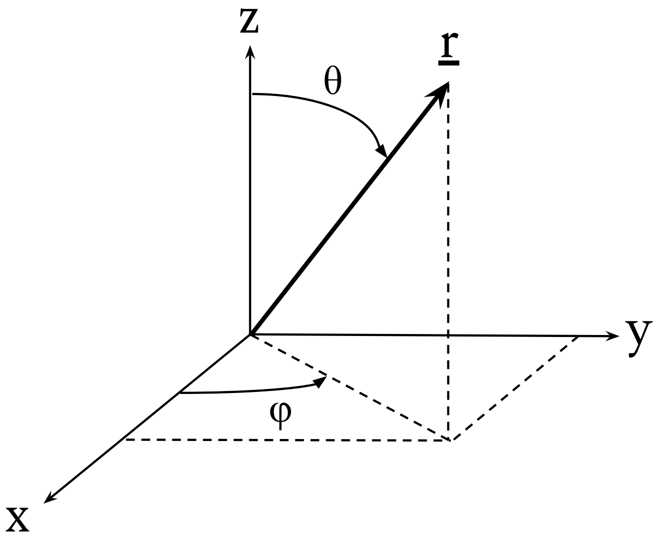


*Figure 2.1: Coordinate system*

The source point $\underline{r}$ in cartesian space is defined as $\underline{r} = \begin{bmatrix} x & y & z \end{bmatrix}^\mathsf{T}$.

The conversion from spherical to cartesian coordinates can be done as

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = |r| \cdot \begin{bmatrix} \sin(\theta)\cos(\varphi) \\ \sin(\theta)\sin(\varphi) \\ \cos(\theta) \end{bmatrix} \tag{2.1}$$

where $|r|$ is the absolute distance from the origin to the reference point, $\varphi$ is the azimuth angle, $\theta$ is the elevation angle. The opposite conversion from cartesian to spherical coordinates is done as

$$|r| = \sqrt{x^2 + y^2 + z^2} \tag{2.2}$$

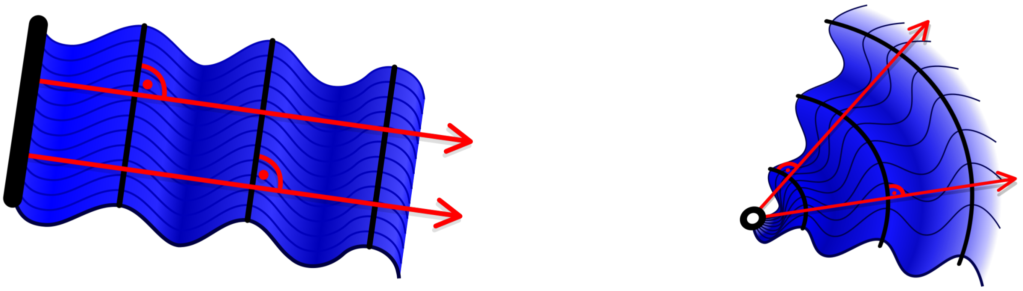
Further the elevation angle $\theta$ and the azimuth angle $\varphi$ can be computed as

$$\theta = \arccos \frac{z}{\sqrt{x^2 + y^2 + z^2}} \tag{2.3}$$

$$\varphi = \mathrm{atan2}(y, x) \tag{2.4}$$

## 2.3 Acoustic Wave Propagation

Microphone arrays pick up acoustic waves emitted by sound sources. Two different models of acoustic wave propagation are considered, namely the *plane wave* and *spherical wave* propagation model. Both models have to be considered because of spatial relationships between the transmitters and receivers within an acoustic field. When a source is at sufficient distance and further the array's aperture is relatively small (so that the curvature of the originating sound wave can be neglected), in the majority of beamforming literature there is talk of *far-field beamforming* and the incoming sound waves are approximated as plane waves. On the contrary if the source is located closely to a relatively large array the beamforming community speaks about *near-field beamforming* (here the curvature of the sound wave can't be neglected). However these attributes insufficiently describe far-field or near-field acoustics, therefore in this thesis these assumptions will be explicitly called *plane waves* and *spherical waves*.



*(a) Planar wave*

*(b) Radial wave*

*Figure 2.2: Comparison of wave propagation models (taken from [1])*

The two mentioned wave propagation models are illustrated in figure 2.2. Planar waves are characterized by parallel wave fronts that move in perpendicular direction. In spherical waves the direction of the wave front's propagation is radial.

The equation for a monochromatic plane wave is given as [4]

$$x(t, \underline{r}) = A \cdot \mathrm{e}^{j(\omega t - \underline{k}\underline{r})} \tag{2.5}$$

where A is the wave amplitude, $\omega$ $(= 2\pi f)$ is the radial frequency, $\underline{r}$ is the distance between source/sink and the wavenumber vector $\underline{k}$:

$$\underline{k} = \frac{2\pi}{\lambda} \begin{bmatrix} \sin(\theta)\cos(\varphi) \\ \sin(\theta)\sin(\varphi) \\ \cos(\theta) \end{bmatrix} \tag{2.6}$$

describes the speed and direction of the propagating wave. The wavelength $\lambda$ is given as $\lambda = \frac{c}{f}$, where $c$ is the speed of sound in air, which can be computed from the ambient temperature $c = (331.5 + 0.6\frac{\vartheta}{{}^\circ C})\frac{m}{s}$, where $\vartheta$ is the ambient temperature in degree Celsius. With the definition of the azimuth angle $\varphi$ and the elevation angle $\theta$ the *direction of arrival (DOA)* is fully determined, which describes the direction of the incoming sound wave.

If the curvature of a spherical sound wave has to be considered, the monochromatic solution is given as

$$x(t, \underline{r}) = \frac{A}{4\pi r}\mathrm{e}^{j(\omega t - \underline{k}\underline{r})} \tag{2.7}$$

## 2.4 Time Difference of Arrival

The *time difference of arrival (TDOA)* is best described as relative time lag between sensor signals originating from a signal source. The phenomenon is best shown by a plane wave arriving at a uniform linear array (ULA) with two microphones in figure 2.3 [5].

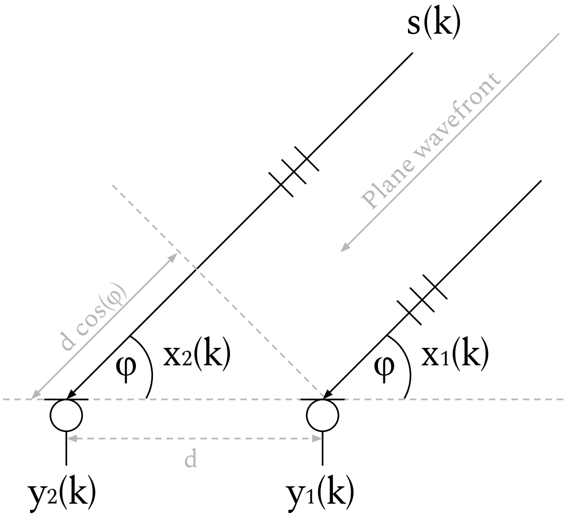


*Figure 2.3: Simple ULA with an impinging plane sound wave*

The single-source signal $s(k)$ convolved with a particular acoustic MIMO impulse response results in the plane wave signal $x_m(k)$. The final microphone signal $y_m(k)$ also includes transfer function and directional characteristics of the microphone.

For the simplified case of a ULA the TDOA between sensors 1 and 2 is then computed as

$$\tau_{12} = \frac{d \cdot \cos(\varphi)}{c} \tag{2.8}$$

In contrast to the TDOA the *time of arrival (TOA)* describes the absolute elapsed time between the onset of a sound wave and the microphone(s) picking up the propagated signal.
The TDOA is especially relevant to estimate the direction of arrival by just looking at the TDOA between several microphone signals. This task is called the time delay estimation problem (TDE). As this aspect is crucial for (automatized) sound source localization (SSL) it will be discussed more in detail in chapter 5.

## 2.5 Spatial Aliasing

When designing a sensor array the so-called spatial aliasing has to be considered. Similarly to the sampling theorem in signal theory, design errors in spatial sampling can lead to spatial images (or grating lobes) and therefore unwanted directional characteristics. The spatial sampling theorem can be determined as

$$d < \frac{\lambda}{2} \tag{2.9}$$

where $d$ is the distance between the sensors in meters and $\lambda$ is the observed wavelength ($\lambda = \frac{c}{f}$). To avoid spatial aliasing for a specific frequency $f$ the microphone distance $d$ should be kept below $\frac{\lambda}{2}$ ($= \frac{c}{2\,f}$). In other words for optimal operation at higher frequencies the array should have a smaller sensor spacing. On the other hand to guarantee spatial selectivity also for lower frequencies large sensor spacing is required, which always results in a tradeoff between both requirements.

## 2.6 Uniform Circular Array

The geometry of the sound field that is picked up by a microphone array is depending on the geometry of the used microphone array itself and also the directional characteristics of the microphones. A radial pattern with equal sensitivity in all horizontal directions is given by a circular array as it is proposed in this work. Although also other geometrical forms of arrays are possible, the circular array shape will be used for this thesis.

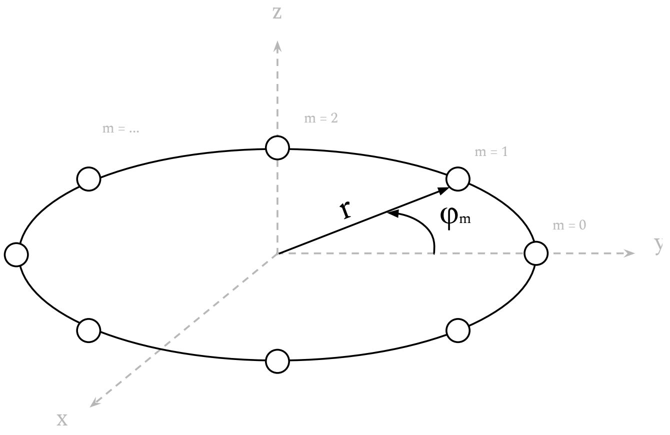


*Figure 2.4: Uniform circular array*

The steering vector determines the direction of the incoming wave and can be generally expressed, without assuming a specific wave propagation model, via the according time-shift constants $\tau_m$ where $m = 0, 1, 2, \ ..., \ M-1$ as depicted in figure 2.4.

$$\underline{d}_m(f) = \mathrm{e}^{-j2\pi f \tau_m} = \begin{bmatrix} \mathrm{e}^{-j2\pi f \tau_0} \\ \mathrm{e}^{-j2\pi f \tau_1} \\ ... \\ \mathrm{e}^{-j2\pi f \tau_{M-1}} \end{bmatrix} \tag{2.10}$$

The steering vector gives us the ability to compute the so-called beamforming weights, which further allows us to focus a sensor array to a desired direction. This will be on topic in chapter 3. The computation of the time-shift constants will be specified more in detail in the following sections.

## 2.7 Plane and Spherical Waves

In array processing a rule-of-thumb has been established [4], which determines if a source is at sufficient distance, so that the arriving sound wave can be approximated as plane wave. We estimate that distance $r$ as shown below.

$$r > \frac{2L^2}{\lambda} \tag{2.11}$$

where $L$ is the maximum dimension of the array that picks up the incoming sound wave and $r$ is the absolute distance from the emitting sound source to the sound receiver.

In speech signal processing we are dealing with a limited bandwidth of interest up to around 12kHz and therefore those presumptions can already be included into the array design considerations, beside the relevant source distance. But this also means that not only the distance between source and sensor has to be considered for the wave propagation model but also the size of the sensor aperture is relevant. This will be considered later when the specific array geometry is discussed in section 6.1.

### 2.7.1 Uniform Circular Array Receiving Plane Waves

According to [6] the steering vector for an uniform circular array under plane wave assumption is defined as

$$\underline{d}_m(f,\varphi_s,\theta_s) \;=\; \mathrm{e}^{-j2\pi f\frac{r}{c}\sin(\theta_s)\cos(\varphi_s-\varphi_m)} \;=\; \begin{bmatrix} \mathrm{e}^{-j2\pi f\frac{r}{c}\sin(\theta_s)\cos(\varphi_s-\varphi_0)} \\ \mathrm{e}^{-j2\pi f\frac{r}{c}\sin(\theta_s)\cos(\varphi_s-\varphi_1)} \\ \ldots \\ \mathrm{e}^{-j2\pi f\frac{r}{c}\sin(\theta_s)\cos(\varphi_s-\varphi_{M-1})} \end{bmatrix} \tag{2.12}$$

where $m$ is the microphone index, $\varphi_s$ is the azimuth angle of the incoming plane wave source, $\theta_s$ is the elevation angle of the incoming plane wave source and $\varphi_m$ is the azimuth angle of that particular microphone.

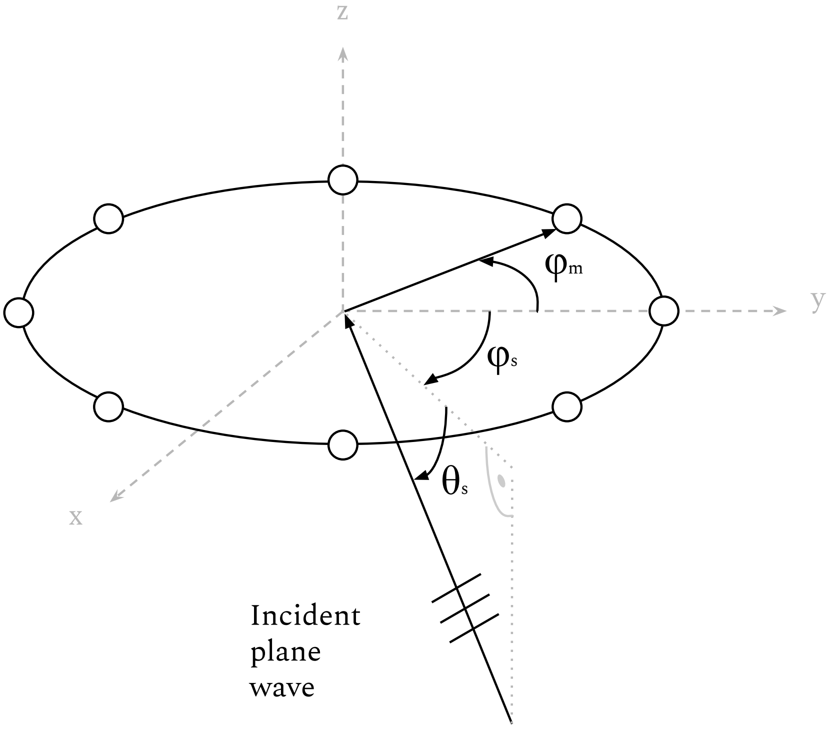


*Figure 2.5: Circular array receiving plane waves*

If a sensor array is steered towards a specific source assuming plane waves, each sensor signal can be interpreted as virtually shifted onto a straight line in the array's origin, perpendicular to the direction of arrival.

### 2.7.2 Uniform Circular Array Receiving Spherical Waves

If the sound source is assumed to propagate spherical waves, the steering vector is specified as

$$\underline{d}_m(f,\ l_m,\ \alpha_m) \ = \ \alpha_m \ \mathrm{e}^{-j2\pi f \frac{l_m}{c}} \ = \begin{bmatrix} \alpha_0 \ \mathrm{e}^{-j2\pi f \frac{l_0}{c}} \\ \alpha_1 \ \mathrm{e}^{-j2\pi f \frac{l_1}{c}} \\ \dots \\ \alpha_{M-1} \ \mathrm{e}^{-j2\pi f \frac{l_{M-1}}{c}} \end{bmatrix} \tag{2.13}$$

where $l_m$ is the euclidean distance from the source position vector $\underline{r}_s$ to the $m$-th microphone position vector $\underline{r}_m$ and $\alpha_m$ is the according attenuation factor.

$$l_m = ||\underline{r}_s - \underline{r}_m|| = \sqrt{(x_s - x_m)^2 + (y_s - y_m)^2 + (z_s - z_m)^2} \tag{2.14}$$

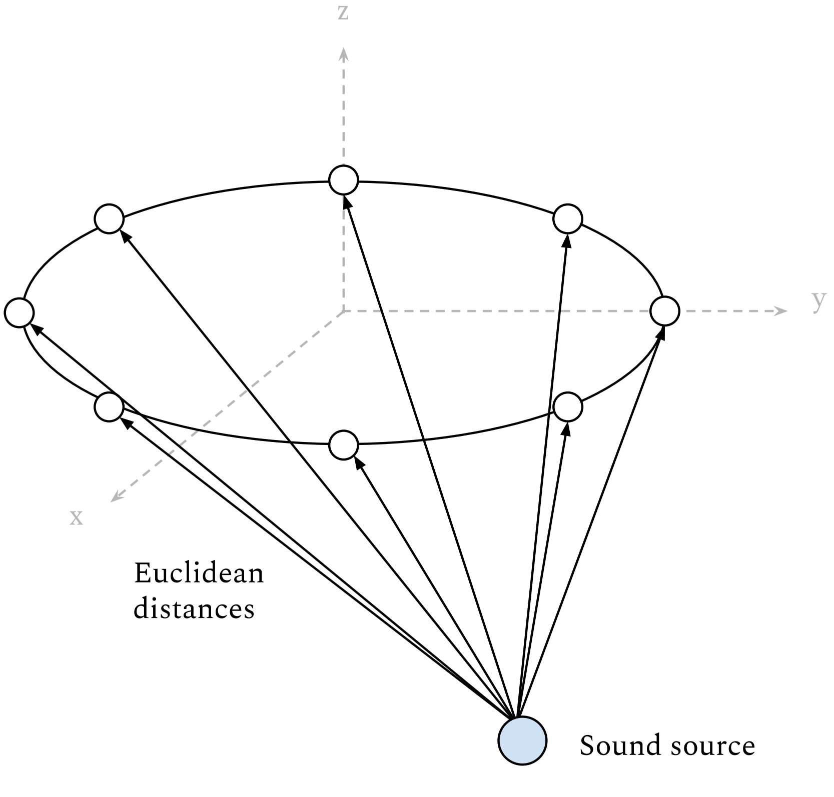


*Figure 2.6: Circular array receiving spherical waves*

If a sensor array is steered towards a specific source assuming spherical waves, each sensor signal can be imagined as shifted into the spot right where the sound source is sitting. Of course the position of the sound source has to be known. At a later point in this thesis it has been shown, that also a rough estimate of the source position produces meaningful results.

### 2.7.3 Combined Array

So far the steering vector models for a single array have been described under the assumption of a plane waves and spherical waves. On the other hand the sensor setup has been introduced as dual array configuration in chapter 1. There is definitely a benefit in improving the signal quality by using a bigger amount of microphones. Still the question is, how the models previously introduced could be applied to the dual array configuration.

Imagine both arrays receiving *plane waves* from an arbitrarily located sound source. The two steering vectors $\underline{d}_{1s}(f, \varphi_{1s}, \theta_{1s})$ and $\underline{d}_{2s}(f, \varphi_{2s}, \theta_{2s})$ for both arrays can be computed once both DOAs are known (assuming all other variables in the formula are given) and the beamforming result will be two separate signals, one for each array, which need further processing. Neither cross-correlation based time-shifts nor estimated-distance-based time-shifts have been proven as a suitable method to combine the two array outputs under plane wave assumption into one single output.

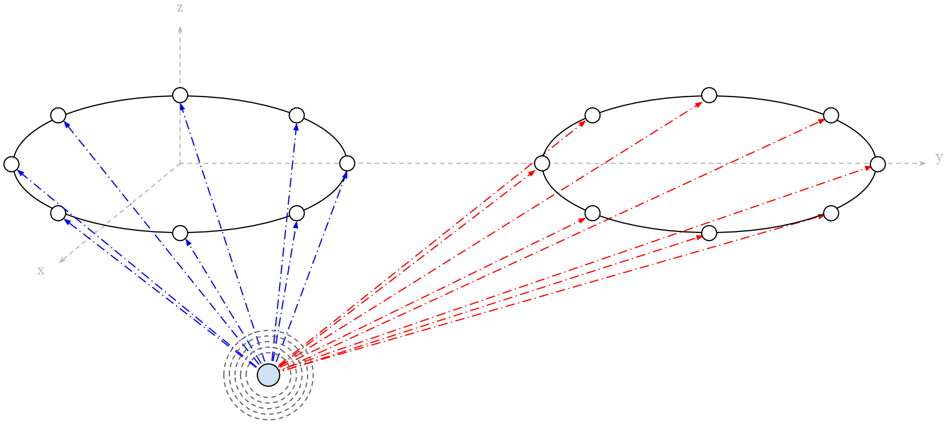


*Figure 2.7: Spherical waves striking the dual array*

The second method shown in figure 2.7 assumes the *spherical wave propagation model*, which requires a defined source position to compute the euclidean distances and the attenuation gains. This sounds like an obstacle at first, but in this thesis it has been experimentally found in subjective listening tests, that an estimated position given by a *sound source localization algorithm* (see chapter 5) in combination with the spherical wave propagation model delivers superior results compared to the plane wave approach as the interfering noise level and reverberation is reduced and therefore the signal-to-noise-ratio (SNR) is improved.

By knowing the source location estimate the euclidean distances $l_m$ to all microphones and thus the steering vector for the dual array $\underline{d_s}(f, l_m, \alpha_m)$ can be computed. The attenuation gains are taken care of by using spatial weights (according to the inverse distance law) which is described in section 3.3. Using the estimated source position also the steering vector estimate can be formulated.

$$\underline{\hat{d}}_m(f,\ \hat{l}_m,\ \hat{\alpha}_m) \;=\; \hat{\alpha}_m\ \mathrm{e}^{-j2\pi f \frac{\hat{l}_m}{c}} \;=\; \begin{bmatrix} \hat{\alpha}_0\ \mathrm{e}^{-j2\pi f \frac{\hat{l}_0}{c}} \\ \hat{\alpha}_1\ \mathrm{e}^{-j2\pi f \frac{\hat{l}_1}{c}} \\ \cdots \\ \hat{\alpha}_{M-1}\ \mathrm{e}^{-j2\pi f \frac{\hat{l}_{M-1}}{c}} \end{bmatrix} \tag{2.15}$$

## 2.8 Noise Field Statistics

Before statements about the performance of beamforming algorithms are made, the statistics of the noise field surrounding the sensor array must be taken into account. The noise field can be described generally by looking at the coherence of the noise signals $\Gamma_{V_nV_m}$.

$$\Gamma_{V_nV_m}(e^{j\omega}) = \frac{\Phi_{V_nV_m}(e^{j\omega})}{\sqrt{\Phi_{V_nV_n}(e^{j\omega})\Phi_{V_mV_m}(e^{j\omega})}} \tag{2.16}$$

where $\Phi_{V_nV_m}$ is the cross power spectral density (CPSD) of the noise signals $v_n$ and $v_m$ and $n$, $m$ are the associated microphone indices. The noise field across all sensor signals can be further described with the coherence matrix $\mathbf{\Gamma_{VV}}$.

$$\mathbf{\Gamma_{VV}} = \begin{bmatrix} 1 & \Gamma_{V_0V_1} & \Gamma_{V_0V_2} & \cdots & \Gamma_{V_0V_{N-1}} \\ \Gamma_{V_1V_0} & 1 & \Gamma_{V_1V_2} & \cdots & \Gamma_{V_1V_{N-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Gamma_{V_{N-1}V_0} & \Gamma_{V_{N-1}V_1} & \Gamma_{V_{N-1}V_2} & \cdots & 1 \end{bmatrix} \tag{2.17}$$

However the noise coherence model can be viewed as theoretical as in practice it is complicated to estimate noise signals. Starting from this observation, other defined noise fields with more practical meaning have been established (which are not signal-dependent anymore, but position-dependent instead). Those will be discussed at a later point, when arriving at the *Superdirective MVDR beamformer* in section 3.5.

# 3 Beamforming

This chapter introduces widely used beamforming algorithms which combine sensor data and positioning information to produce enhanced output. For our task we are interested in separating speech signals, hence beamforming helps us to focus the array on a desired target location, which also means suppressing interfering signals from other directions and therefore improving the signal-to-noise-ratio.

Beamforming algorithms are divided in data-independent and data-dependent beamformers. The delay-and-sum beamformer, a representative of the signal-independent algorithm class, uses spatial information to steer the sensor array towards a specified target.

The second class of algorithms further applies information based on sensor positions (Superdirective MVDR) or signal statistics (Classic MVDR). Further the Generalized Sidelobe Canceller uses an adaptive filtering approach and therefore also belongs to the class of data-dependent beamformers.

## 3.1 Beampattern

The directional power characteristics of a sensor array can be computed by evaluating $H$ for all possible angles, which results in the beampattern [7].

$$H(e^{j\omega}, \varphi_0, \theta_0, \varphi, \theta) = 10 \cdot log_{10}\Big\{|\underline{w}^H(e^{j\omega}, \varphi_0, \theta_0)\ \underline{d}(e^{j\omega}, \varphi, \theta)|^2\Big\} \tag{3.1}$$

where $\underline{w}$ is the beamformer filter-weight vector that is computed with respect to the target direction at azimuth angle $\varphi_0$ and elevation angle $\theta_0$. The steering vector $\underline{d}$ will be iterated over all desired azimuth angles $\varphi$ and elevation angles $\theta$. Examples for beampatterns will be shown in figure 3.4.

A balloon plot can be achieved if all possible azimuth angles $\varphi$ and all elevation angles $\theta$ will be evaluated for one single frequency. Examples will be shown in section 3.2, where the delay-and-sum beamformer will be discussed.

## 3.2 Delay-and-sum Beamformer

The *delay-and-sum beamformer (D&S-BF)* is the most fundamental beamforming concept which is also known as classic beamformer or conventional beamformer [4]. The idea is to time-align all given sensor signals according to the desired steering direction and to sum the time-aligend signals and compute the mean afterwards. The output of the D&S-BF is computed as follows

$$y_i(n) = \frac{1}{M} \sum_{m=1}^{M} x_m(n - \tau_{mi}) \tag{3.2}$$

where $y_i(n)$ is the delay-and-sum beamformer output steered to the desired source $i$ with cartesian coordinates $\underline{r}$, $M$ is the total number of used microphone channels, $m$ is the running microphone index, $x_m$ is the considered microphone signal, $n$ is the sample index and $\tau_{mi}$ is the TDOA of the considered microphone (in samples) with respect to steering direction belonging to the desired target $i$.

The block diagram of the D&S-BF is illustrated in figure 3.1

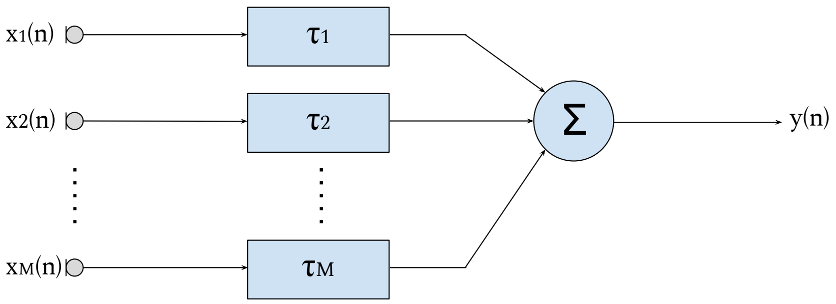


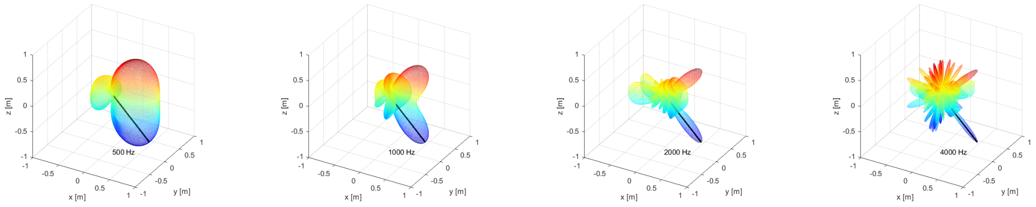
*Figure 3.1: Delay-and-sum beamformer*

In practice the delay-and-sum beamforming can be done in time-domain by performing discrete shifts on the data-vectors according to the TDOA. For higher accuracy oversampling or fractional delays can be applied. The other method is done by taking the input signals via Fourier transform (e.g. FFT) into frequency-domain and by applying a complex signal multiplication with the steering vector to achieve the time-shift operation.

$$Y_i(k,f) = \frac{1}{M}\sum_{m=1}^{M} X_m(k,f)\cdot W_m^*(k,f) \qquad \text{where } W_m(k,f) = \mathrm{e}^{-j\cdot 2\pi f\cdot\tau_{mi}} \tag{3.3}$$

where $Y_i(k,f)$ is the output in frequency-domain at frame $k$ and discrete frequency bin $f$. The time-domain D&S-BF output $y_i$ can be computed by taking the inverse Fourier transform of $Y_i$.

Using equation 3.1 the wideband-beampattern will be evaluated and plotted in figure 3.4 (a). The circular array (specified in section 6.1) is steered towards 0° azimuth and -45° elevation. The beampattern can also be shown in a 3-dimensional balloon plot by stepwise evaluating the monochromatic beampattern at specific azimuth and elevation angles. The black line in figure 3.2 marks the steering direction. The lengths of the lobes denote the directional sensitivity, whereas the z-axis is coded in color for a better visualization.



*(a) 500 Hz* *(b) 1000 Hz* *(c) 2000 Hz* *(d) 4000 Hz*

*Figure 3.2: Balloon plots for delay-and-sum beamformer*

## 3.3 Spatial windowing

The plain delay-and-sum beamformer performs spatial filtering by shifting the signals according to the direction the array is steered to. Before going deeper into more advanced spectral weighting techniques, there is also another method called spatial windowing, which allows putting emphasis on specific microphone gains [8].

Usually this is done by multiplying classic windowing functions (such as Hanning window, Kaiser window, Dolph-Chebyshev window,etc.) directly onto the microphone tracks, so that microphones close to the desired target remain having higher gain values and the microphones at greater distance are attenuated. One has to keep in mind, that this technique also reduces the beamformer output gain in general.

Spatial windowing is interesting, especially because it affects the microphone array's beampattern by a tradeoff between main lobe width and side lobe suppression. The reason the parametrizable Kaiser window is very popularly used for this method is, because that tradeoff can be easily controlled.

## 3.4 Classic MVDR Beamformer

The target of the *minimum variance distortionless response (MVDR)* beamformer is to minimize the variance of the output while keeping the signal for a desired direction distortionless [9]. The solution can be found in optimizing the criterion:

$$\underset{\underline{w}}{min}\ \underline{w}^H\ \mathbf{R_{vv}}\ \underline{w} \qquad \text{subject to: } \underline{d}^H\ \underline{w} = 1 \tag{3.4}$$

where $\underline{w}$ is the beamformer weights vector, $\mathbf{R_{vv}}$ is the spatial correlation matrix of the noise signals and $\underline{d}$ is the steering vector. By using Lagrange multipliers the solution is given without proof as:

$$\underline{w}_{mvdr} = \frac{\mathbf{R_{vv}}^{-1}\ \underline{d}}{\underline{d}^H\ \mathbf{R_{vv}}^{-1}\ \underline{d}} \tag{3.5}$$

In literature the terms MVDR and MPDR (minimum power distortionless response) are often mixed up. The key difference is, that the MPDR beamformer uses the spatial correlation matrix $\mathbf{R_{xx}}$ which is computed directly from the input signals $x(n)$. However the MVDR beamformer is preferred over the MPDR beamformer as the MPDR beamformer is sensitive to steering vector errors much more.

To compute the spatial correlation matrix $\mathbf{R_{vv}}$ additional processing is needed to compute the noise signals $v(n)$ first, e.g. by invoking the spectral subtraction [10]. In the upcoming section 3.4.1 the details, how the spatial correlation matrix is composed, will be discussed. Generally the noise correlation matrix $\hat{\mathbf{R}}_{\mathbf{vv}}$ is estimated as

$$\hat{\mathbf{R}}_{\mathbf{vv}} = \mathbb{E}\Big[\ \underline{v}\ \underline{v}^{\intercal}\ \Big] \tag{3.6}$$

The correlation matrix of highly correlated signals often require regularization to put out meaningful computations, which is why techniques such as Diagonal Loading are used. In practice also the theoretical MVDR beamformer suffers from sensor mismatch and array imperfections.

### 3.4.1 Estimating the Spatial Correlation Matrix

One of the challenging tasks concerning the MVDR beamformer is to estimate the noise signal. In simulated environments the noise signal is mostly accessible, which does not apply to real environments. Therefore in this thesis the delay-and-sum beamformer is invoked to acquire the noise signal via spectral subtraction. The algorithm is depicted in figure 3.3.
For MPDR beamformers the PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{XX}}$ is initially computed as follows.

$$\hat{\mathbf{\Phi}}_{\mathbf{XX}}(f) = \mathbb{E}\Big[\underline{X}(f)\ \underline{X}^{\intercal}(f)\Big] \tag{3.7}$$

where $\underline{X}$ is the multichannel signal input block transformed into a frequency domain vector. It is very common to use Welch's method [11] to compute smoothed PSD matrices for each timeframe introducing the forgetting factor $\lambda$.

$$\hat{\mathbf{\Phi}}_{\mathbf{XX}}(f) = \lambda\ \hat{\mathbf{\Phi}}_{\mathbf{XX-1}}(f) + (1-\lambda)\ \underline{X}(f)\ \underline{X}^{\intercal}(f) \tag{3.8}$$

After the PSD matrix has been computed, $\hat{\mathbf{\Phi}}$ is then substituted with $\mathbf{R}$ from equation 3.5 to compute the beamformer weights $\underline{w}_{mpdr}$.

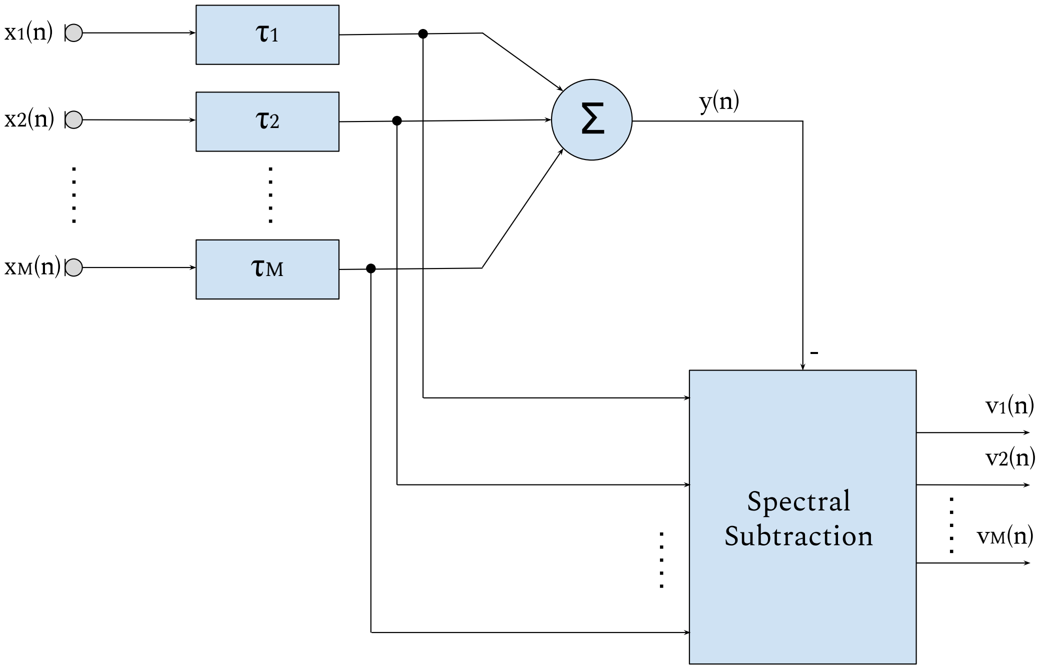


*Figure 3.3: Estimating the noise signal*

If MVDR beamformers are used instead of MPDR beamformers, the noise PSD can be achieved in a similar way, by first pre-processing the input signals e.g. with the spectral subtraction algorithm shown in figure 3.3 to obtain the noise signals $v(n)$.

$$\hat{\mathbf{\Phi}}_{\mathbf{VV}}(f) = \mathbb{E}\Big[\underline{V}(f)\ \underline{V}^{\intercal}(f)\Big] \tag{3.9}$$

$$\hat{\mathbf{\Phi}}_{\mathbf{VV}}(f) = \lambda\ \hat{\mathbf{\Phi}}_{\mathbf{VV-1}}(f) + (1-\lambda)\ \underline{V}(f)\ \underline{V}^{\intercal}(f) \tag{3.10}$$

By analogy to the MPDR beamformer, to achieve the MVDR beamformer weights $\underline{w}_{mvdr}$ the PSD matrix $\hat{\mathbf{\Phi}}_{\mathbf{VV}}$ is substituted with $\mathbf{R_{vv}}$ from equation 3.5.

However the correlation matrices that are usually dealt with in such tasks are prone to ill-conditioning due to highly correlating signals. This further leads to numerical inaccuracy when computing the inverse of ill-conditioned matrices. To work around such phenomenons several strategies have been introduced of which two are mentioned next.

### 3.4.2 Diagonal Loading

The diagonal loading method [12] is the most simple approach, where a small proportion of the identity matrix $\mathbf{I}$ is added to the correlation matrix $\hat{\mathbf{R}}$ that needs to be processed. The scaling factor $\gamma$ controls the amount of the identity matrix $\mathbf{I}$ that is added.

$$\underline{w}_{DL} = \frac{\hat{\mathbf{R}}_{\mathbf{DL}}^{-1}\ \underline{d}}{\underline{d}^H\ \hat{\mathbf{R}}_{\mathbf{DL}}^{-1}\ \underline{d}} \qquad \text{with } \hat{\mathbf{R}}_{\mathbf{DL}} = \hat{\mathbf{R}} + \gamma\mathbf{I} \tag{3.11}$$

By increasing $\gamma$ also the robustness of the algorithm increases and the solution approaches the delay-and-sum beamformer.

### 3.4.3 Variable Loading

Another variant is the variable loading technique [13], where the regularization is scaled dynamically with the input signal by adding a small proportion of the correlation matrix inverse onto the correlation matrix itself.

$$\underline{w}_{VL} = \frac{\hat{\mathbf{R}}_{\mathbf{VL}}^{-1}\ \underline{d}}{\underline{d}^H\ \hat{\mathbf{R}}_{\mathbf{VL}}^{-1}\ \underline{d}} \qquad \text{with } \hat{\mathbf{R}}_{\mathbf{VL}} = \hat{\mathbf{R}} + \delta\ \hat{\mathbf{R}}^{-1} \tag{3.12}$$

## 3.5 Superdirective MVDR beamformer

The first approach to make the MVDR beamformer more applicable to real-time environments is called the *superdirective MVDR beamformer (SD-BF)* [7]. Using a correlation matrix, that is built from noise field statistics instead of instantaneous signal statistics, makes the beamformer signal-independent but position-dependent instead. Within this workaround the estimated, spatial correlation matrix $\hat{\mathbf{R}}_{\mathbf{vv}}$ is substituted by a theoretically defined noise field coherence $\mathbf{\Gamma}_{\mathbf{vv}}$.

$$\underline{w}_{sd} = \frac{\mathbf{\Gamma}_{\mathbf{vv}}{}^{-1}\underline{d}}{\underline{d}^H\mathbf{\Gamma}_{\mathbf{vv}}{}^{-1}\underline{d}} \tag{3.13}$$

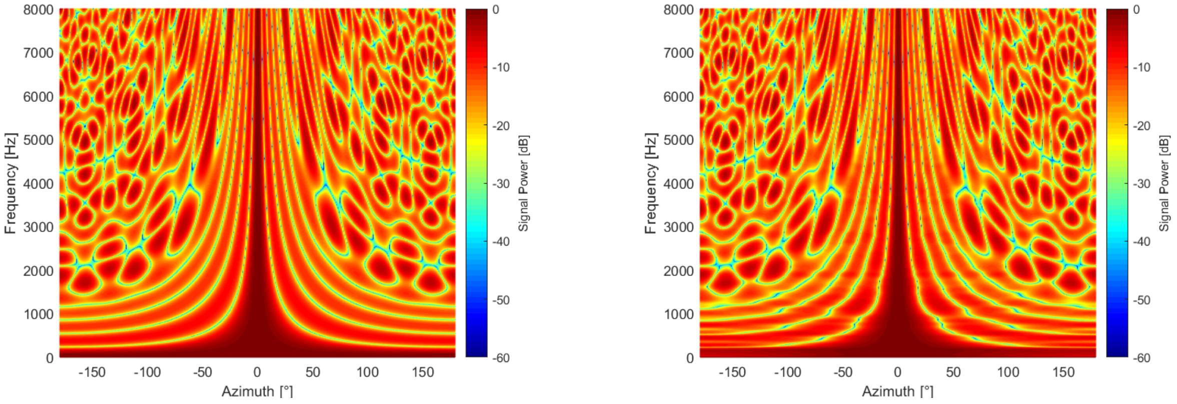
The noise field coherence $\mathbf{\Gamma}_{\mathbf{vv}}$ has already been introduced in section 2.8. Whereas the simple D&S-BF is unable to attenuate low frequency signals from non-target directions, the SD-BF performs better to do so.

#### Diffuse Noise Field

The coherence function $\Gamma_{V_nV_m}$ of a diffuse noise field can be computed as

$$\Gamma_{V_nV_m}\Big|_{diffuse} = \text{sinc}\left\{\frac{2\pi f\ l_{nm}}{c}\right\} \tag{3.14}$$

where $l_{nm}$ is the distance between the according sensor positions.
In figure 3.4 the spectral beampatterns of different noise fields are compared, where especially at low frequencies the differences become significant. As before the array is steered towards 0° azimuth and -45° elevation.



*(a) D&S-BF*

*(b) SD-BF: Diffuse noise*

*Figure 3.4: Theoretical wideband beampatterns*

## 3.6 Generalized Sidelobe Canceller

The *generalized sidelobe canceller (GSC)* [14] is an adaptive beamforming method, which separates the signal enhancement task in two processing paths. The GSC has been invented by Lloyd J. Griffiths and Charles W. Jim and is therefore also known as Griffiths Jim beamformer.

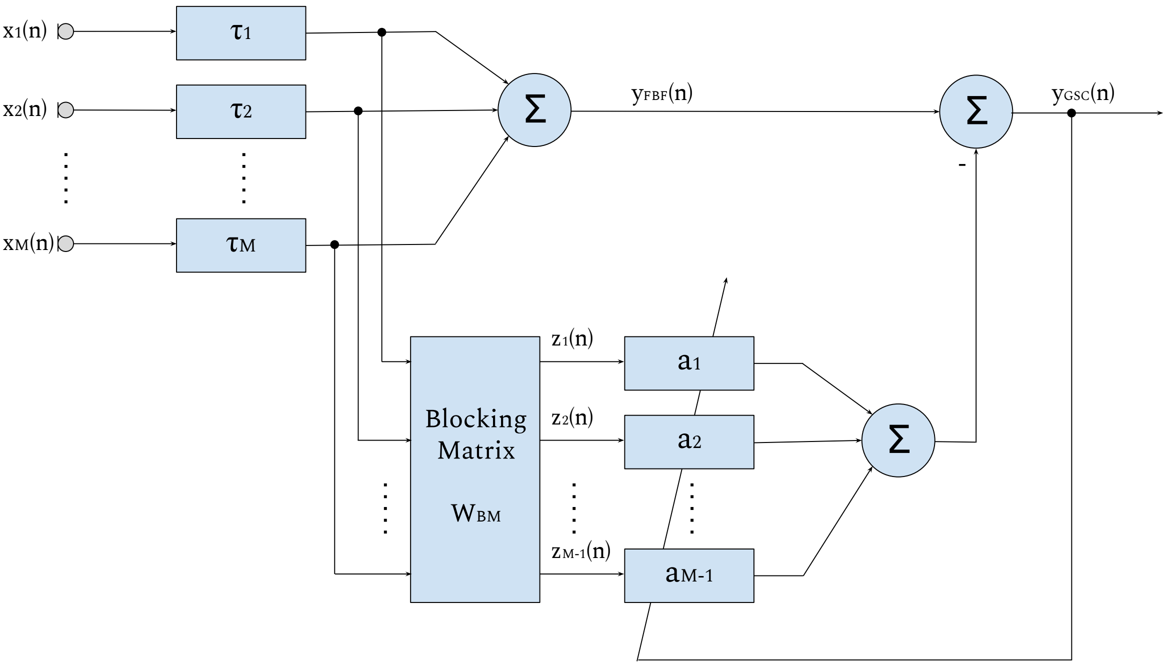


*Figure 3.5: The Generalized Sidelobe Canceller*

The upper branch of the GSC (in figure 3.5) is a fixed beamformer (FBF), that enhances the signal in one single, desired target direction. A very popular choice for the FBF is the simple delay-and-sum beamformer. The adaptive beamformer is implemented in the lower path, which consists of a blocking matrix (BM) and the adaptive input canceller (AIC). The blocking matrix takes the pre-steered signals and ensures that the desired signal is eliminated at the blocking matrix output. The adaptive input canceller contains a set of filters, that adaptively optimizes the GSC output power in a minimum mean-square error sense.

The adaptive filters are often implemented as (normalized) least-mean-squares (NLMS/LMS) filters. The GSC in general can be implemented in time-domain or in frequency-domain, while having convergence advantages in frequency-domain.

Several extensions have been introduced to make the GSC more robust to steering direction errors, such as adaptive filters in the blocking matrix. The signal matrix $\mathbf{Z}(n)$ at the blocking matrix output can be computed as

$$\mathbf{Z}(n) = \mathbf{W_{BM}}\ \mathbf{X}(n - \tau_m) \tag{3.15}$$

where $\mathbf{X}(n-\tau_m)$ are the pre-steered signals and $\mathbf{W_{BM}}$ are the fixed weights of the $[M-1 \times M]$ blocking matrix. Several techniques exist for the blocking matrix design, where the classic blocking matrix of the inventors Griffiths and Jim is composed as shown below.

$$\mathbf{W_{BM}} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix} \tag{3.16}$$

The multiplication of the input signal with the blocking matrix above can be interpreted as subtracting adjacent channels. The resulting M-1 audio channels mostly contain noise signals. With the help of the adaptive filters in the lower path the output of the GSC is computed as

$$y_{GSC}(n) = y_{FBF}(n) - \sum_{m=1}^{M-1} \mathbf{a_m}^{\intercal}(n)\ \mathbf{z_m}(n) \tag{3.17}$$

where $y_{GSC}(n)$ is the output of the generalized sidelobe canceller, $y_{FBF}(n)$ is the output of the fixed beamformer, $m$ is the running microphone index, $M$ is the total number of microphones, $\mathbf{w_m}(n)$ are the adaptive filter weight vectors, $\mathbf{z_m}(n)$ are the blocking matrix output vectors. According to the NLMS update rule, the filter weights are updated each cycle.

$$\mathbf{a_m}(n+1) = \beta\ \mathbf{a_m}(n) + \mu\ y_{GSC}(n)\ \frac{\mathbf{z_m}(n)}{||\mathbf{z_m}(n)||^2} \tag{3.18}$$

where $\beta$ is the forgetting factor of the old filter weights, $\mu$ is the step-size of the adaptive filter [15]. Since the outputs of the blocking matrix optimally just contain noise and interference signals, the finding of the unconstrained adaptive filter weights ultimately also minimizes the output power in $y_{GSC}$.

# 4 Time-Frequency Masking

With the renaissance of deep neural networks (DNN) time-frequency masking (TFM) has become a crucial part of speech separation. While being used as training targets in DNNs, in acoustic beamforming time-frequency masks are used as a post-processing step to further improve a beamformed signal with the help of simultaneous interfering signals. In the next sections several methods will be shown [16].

## 4.1 Theoretical Ideal Mask

The *ideal mask* shows the general formulation of the problem. By assuming additive noise, the signal model is considered to be:

$$Y(k,f) = X(k,f) + V(k,f) \tag{4.1}$$

where $X$ is the source signal, $V$ the additive noise signal, $Y$ the resulting output signal, $k$ the block index and $f$ the frequency bin index. Further the Mask $M$ is introduced to filter the noise-inflicted signal $Y$, which results in the estimate $\hat{X}$ of the source signal $X$ as follows

$$\hat{X}(k,f) = M(k,f) \cdot |Y(k,f)| e^{j\theta_{Y(k,f)}} \tag{4.2}$$

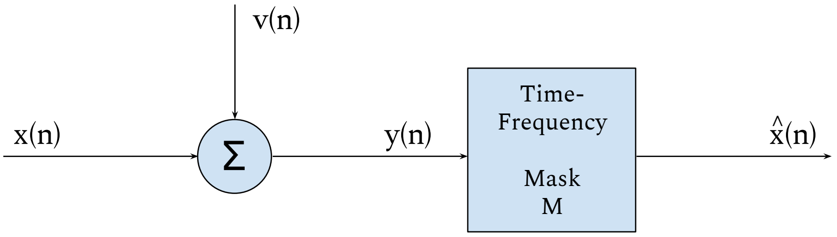


*Figure 4.1: Ideal mask application*

The block diagram in figure 4.1 illustrates how the time-frequency masking is used to enhance a signal. The time-frequency mask block also includes the forth- and back-transformation between time- and frequency-domain. However it is still not clear how the filter mask $M$ can be determined or how the inaccessible noise signal $v(n)$ can be estimated, which will be discussed in the next sections.

## 4.2 Ideal Binary Mask

The most basic idea that is presented is the theoretical *ideal binary mask (IBM)*, where the spectral power ratio $SPR$ is introduced [17]. Frequency bins of the target signal that are more dominant than the interfering frequency bins shall be preserved, while weaker bins are blocked:

$$SPR(k,f) = \frac{|Y(k,f)|^2}{|V(k,f)|^2} \tag{4.3}$$

After computing the spectral power ratio the decision criterion $\gamma$ is introduced, which determines a hard boundary where frequency bins are kept or tossed away.

$$M_{IBM}(k,f) = \begin{cases} 1 & \text{if } SPR(k,f) > \gamma \\ 0 & \text{otherwise} \end{cases}$$

In conclusion this means the IBM frequency bin is assigned to one, if the target power is relatively larger than the noise power. Due to its data-dependent nature, the threshold parameter $\gamma$ has to be found experimentally. By choosing $\gamma$ a hard threshold region around the $\gamma$ parameter is produced, which can be softened by applying e.g. a sigmoid function. The softening of the threshold area reduces the output signal suffering from musical noise.

It has been found that the ideal binary mask is able to produce separated signals, that improve the speech intelligibility dramatically and therefore also improve the overall performance of automatic-speech-recognition units, while having weak performance in perceived quality [18], due to the introduction of musical artifacts (”burbling”).

## 4.3 Ideal Ratio Mask

It has been shown that the *ideal ratio mask (IRM)* [18] produces a better trade-off between separation and intelligibility results than the ideal binary mask, while also keeping the musical noise low.

$$M_{IRM}(k,f) = \left( \frac{|Y(k,f)|^2}{|Y(k,f)|^2 + |V(k,f)|^2} \right)^{\frac{1}{\beta}} \tag{4.4}$$

where $\beta$ can be interpreted as a tunable separation level parameter. Small $\beta$ values will improve the separation, while worsening the quality. High $\beta$ values decrease the separation, while keeping the quality. For $\beta = 2$ the formula is equivalent to the square-root Wiener filter in frequency-domain, which is considered to be the optimal estimator of the power spectrum. The common Wiener filter is defined as

$$H_{Wiener}(k,f) = \frac{|Y(k,f)|^2}{|Y(k,f)|^2 + |V(k,f)|^2} \tag{4.5}$$

## 4.4 Extension To The Multichannel Case

In the multichannel case [17] it is considered that several targets are going to be separated in a parallel fashion, as figure 4.2 suggests. To do so the partial time-frequency masks $M_{ij}$ have to be applied right after beamforming.

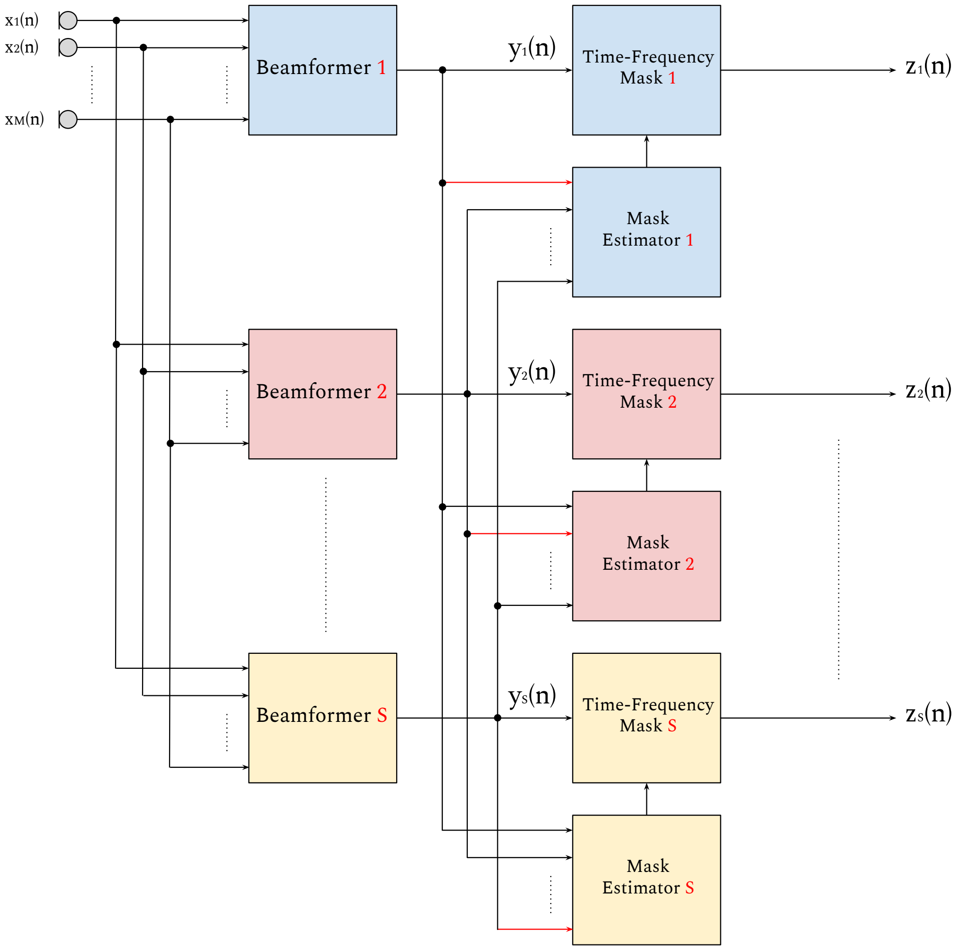


*Figure 4.2: Using time-frequency masking within a beamforming application*

The approach will be sketched in detail on the IRM example, based on equation 4.4. The target signal $Y$ and the noisy signal $V$ from the original formula will be substituted with the beamformed signals $Y_i$ and $Y_j$, where $i$ is the index for the current target source, while $j$ being the indices for all the other interfering sources.

$$M_{ij}^{IRM}(k,f) = \left( \frac{|Y_i(k,f)|^2}{|Y_i(k,f)|^2 + |Y_j(k,f)|^2} \right)^{\frac{1}{\beta}} \tag{4.6}$$

Once all the partial masks $M_{ij}$ for one desired target $i$ have been computed, the full mask $M_i$ can be calculated as a product of all the partial masks.

$$M_i(k,f) = \prod_{j=1, j\neq i}^{J} M_{ij}(k,f) \tag{4.7}$$

As a final step the full mask $M_i$ is applied to the associated beamformed output signal $Y_i$ to receive the time-frequency masked output $Y_i$, while keeping the unaltered phase of the original beamformed signal.

$$Z_i(k,f) = M_i(k,f) \cdot |Y_i(k,f)| e^{j\theta_{Y_i(k,f)}} \tag{4.8}$$

# 5 Sound Source Localization

Sound source localization (SSL) is the task to detect and find a single or even multiple sound sources in a defined search space, by utilizing multiple, simultaneous sensor signals originating from a primarily known, spatially distributed array geometry. SSL crucially improves the performance of beamforming algorithms.

## 5.1 Overview

As presented in [19] three main categories of sound source localization techniques have arisen, each with their individual benefits or drawbacks. Those techniques use ...
$(a)$ ... the maximization of a beamformer's steered response power (SRP)
$(b)$ ... high-resolution spectral estimation methods such as MUSIC, ESPRIT
$(c)$ ... time-difference-of-arrival (TDOA) methods.
In practice each method also comprises the other methods partially, so the boundaries between those three methods are loose.

In method $(a)$ the source location is estimated by filtering, weighting and finally summing the signals from multiple microphones. The result can be considered as the maximum beamformer output with respect to the considered TDOA. On the contrary in method $(b)$ signal statistics are considered by using spatio-spectral correlation matrices. Within method $(c)$ the source locations are computed from a set of delay estimates regarding considered microphone combinations. In the first step pairwise TDOAs between sets of microphones are computed. In the second step, from the time delay estimates (TDE) hyperbolic curves and their intersections are computed, where the sound source may be present.

## 5.2 Cross-Correlation

The correlation function is a signal-statistics-based measure to determine the similarity between two signals $x_1(k)$ and $x_2(k)$. The auto-correlation function is a special case of the correlation function, which involves just one signal $x(k)$ and can be formulated as

$$r_{xx}(p) = \mathbb{E}\Big[x(k)\ x(k+p)\Big] \tag{5.1}$$

where $k$ is the time index of the signal and $p$ is the time shift index. For $p = 0$ the auto-correlation function $r_{xx}(p)$ reaches its maximum value, which is also equivalent to the signal power of signal $x(k)$. Similarly for two sensor signals $x_1(k)$ and $x_2(k)$ the cross-correlation is computed:

$$r_{12}(p) = \mathbb{E}\Big[x_1(k)\ x_2(k+p)\Big] \tag{5.2}$$

## 5.3 Time Delay Estimation Using Cross-Correlation

By finding the maximum of the cross-correlation function at sample offset $p$, the time-shift $\hat{\tau}_{12}$ between the two sensor signals $x_1(k)$ and $x_2(k)$ can be estimated:

$$\hat{\tau}_{12} = \max_p \, r_{12}(p) \tag{5.3}$$

Conveniently the cross-correlation can be computed via the frequency domain, starting by slicing the signal into signal blocks:

$$x_n(t+k) \;=\; \{x_n(t), x_n(t+1), ..., x_n(t+K-1)\} \quad \text{for } t = 0, 1, ... \text{ and } n = 1, 2 \tag{5.4}$$

where $t$ is the absolute time index, $k$ is the running block index and $K$ is the total size of the signal block. The frequency spectrum $X_n(\omega)$ of the signal block $x_n(t+k)$ is computed by invoking the Fourier transform.

$$X_n(\omega) \;=\; \sum_{k=0}^{K-1} x_n(t+k)\ \mathrm{e}^{-j\omega k} \;=\; FT\Big\{x_n(t+k)\Big\} \tag{5.5}$$

The cross-correlation is then obtained by multiplying the frequency-domain signals and taking the inverse Fourier transform:

$$\hat{r}_{12}(p) \;=\; \frac{1}{K}\sum_{k=0}^{K-1} X_1(\omega)\ X_2^*(\omega)\ \mathrm{e}^{j\omega k} \;=\; FT^{-1}\Big\{X_1(\omega)\ X_2^*(\omega)\Big\} \tag{5.6}$$

Finally the time delay estimate between the two sensor signals can be found as before in searching the maximum of the cross-correlation function:

$$\hat{\tau}_{12} = \max_p \, r_{12}(p) \tag{5.7}$$

## 5.4 Generalized Cross-Correlation

The *generalized cross-correlation (GCC)* is a modification to the framework presented in section 5.3 by applying frequency-dependent weights, which potentially improves the TDE performance. The GCC method is popular and was presented in [20]. For the sake of completeness the steps from section 5.2 are repeated and the modifications are mentioned.

$$x_n(t+k) \;=\; \{x_n(t), x_n(t+1), ..., x_n(t+K-1)\} \quad \text{for } t = 0, 1, ... \text{ and } n = 1, 2 \tag{5.8}$$

$$X_n(\omega) \;=\; \sum_{k=0}^{K-1} x_n(t+k)\ \mathrm{e}^{-j\omega k} \;=\; FT\Big\{x_n(t+k)\Big\} \tag{5.9}$$

The desired weighting function $\Psi(\omega)$ will be multiplied with the cross-spectrum.

$$\hat{r}_{12}^{GCC} \;=\; \frac{1}{K}\sum_{k=0}^{K-1} \Psi(\omega)\ X_1(\omega)\ X_2^*(\omega)\ \mathrm{e}^{j\omega k} \;=\; FT^{-1}\Big\{\Psi(\omega)\ X_1(\omega)\ X_2^*(\omega)\Big\} \tag{5.10}$$

As before the time-lag will be found in maximizing the cross-correlation function.

$$\hat{\tau}_{12}^{GCC} = \max_{p} r_{12}^{GCC}(p) \tag{5.11}$$

## 5.5 GCC-PHAT

Once room reverberation rises up to a certain level the cross-correlation TDE starts to degrade significantly in their performance. Therefore several robust weighting methods have been proposed in [19, 21] such as the *Phase transform (PHAT)*, the Smoothed coherence transform (SCOT), the 'Eckart' weighting function and the Maximum likelihood (ML) weighting method. Each of them features individual properties e.g. improving the TDE performance against additive noise (uncorrelated) or multipath effects (highly correlated signals).

$$\Psi_{12}(\omega) = \frac{1}{|X_1(\omega)\ X_2^*(\omega)|} \tag{5.12}$$

The PHAT weighting equalizes the emphasis of each cross-spectrum component, by normalizing the signal spectral density with the spectrum magnitude. In case of the GCC-PHAT this causes a spectral whitening on the input signals, which leads to a sharpening effect on the peak in the cross correlation function. Subsequently the maximum detection and thus finding the correct time-lag will be easier.

## 5.6 Steered Response Power

The *steered response power (SRP)* is described as the output power of a beamformer. For simplicity the steered response power $P(\underline{r})$ will be introduced in time-domain using the simple delay-and-sum beamformer.

$$P(\underline{r}) = \Big| \sum_{m=1}^{M} x_m(n - \tau_m(\underline{r})) \Big|^2 \tag{5.13}$$

where $n$ is the time index, $m$ is the microphone index, $x_m$ are the according input signals, $\tau_m$ is the associated TDOA and $\underline{r}$ is the defined location, where the SRP value $P(\underline{r})$ is going to be determined.

## 5.7 SRP-PHAT

The SRP-PHAT algorithm combines the steered response power concept and the PHAT weighting and is therefore applied in frequency-domain. Previously only the two-channel case has been discussed, whereas in the multichannel case the PHAT weighting is computed in the following way.

$$\Psi_{kl}(\omega) = \frac{1}{|X_k(\omega)\ X_l^*(\omega)|} \tag{5.14}$$

To compute the steered response power, a particular set of TDOAs associated with the microphone indices $k$ and $l$ is needed.

$$\tau_{kl} = \Delta_k - \Delta_l \tag{5.15}$$

where $\tau_{kl}$ is the TDOA between the microphones $k$ and $l$ and $\Delta_k$ respectively $\Delta_l$ are the absolute TOAs. As an example the discrete time TOA for the $k$-th sensor is computed from the euclidean norm as

$$\Delta_k = \frac{fs}{c} \left\| \underline{r} - \underline{r_k} \right\| \tag{5.16}$$

where $\underline{r}$ is the candidate position of the source, $\underline{r_k}$ is the position of the $k$-th sensor, $fs$ is the sampling rate and $c$ is the speed of sound. $\Delta_k$ is computed by analogy.

$$P(\tau_{kl}) = \sum_{k=1}^{M} \sum_{l=1}^{M} \int_{-\infty}^{+\infty} \Psi_{kl}(\omega)\ X_k(\omega)\ X_l^*(\omega)\ \mathrm{e}^{j\omega(\tau_{kl})}\ d\omega \tag{5.17}$$

The desired source position $\hat{\mathbf{r}}_s$ can be estimated by evaluating the SRP-PHAT values $P$ for the considered TDOAs $\tau_{kl}$ and finding the maximum.

$$\hat{\mathbf{r}}_s = \max_{\tau_{kl}} P(\tau_{kl}) \tag{5.18}$$

The algorithm is illustrated as a block diagram in figure 5.1. From the blocks it can be seen that the number of computations depend on the size of the search space and the number of candidate points we want to evaluate.

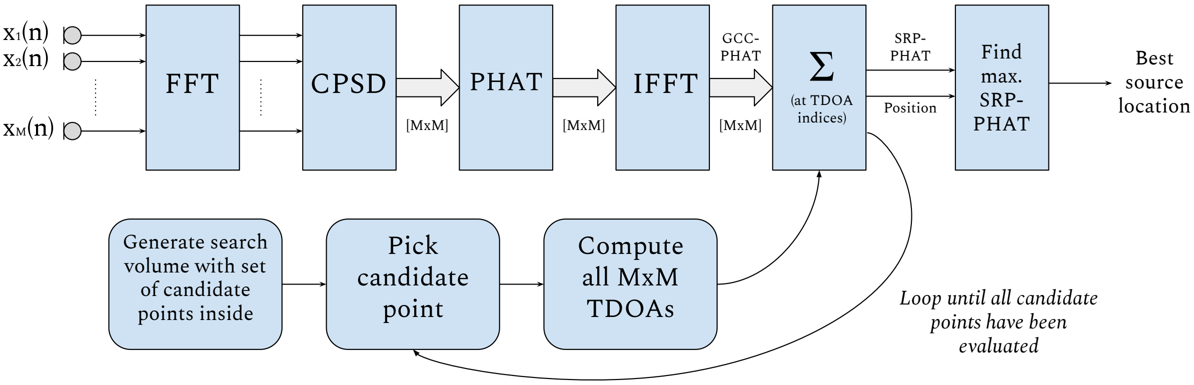


*Figure 5.1: SRP-PHAT algorithm*

To find the maximum SRP-PHAT within the defined search space efficiently, further processing is needed, which will be discussed in the following sections.

## 5.8 Maximum Optimization Techniques

Once the SRP-PHAT values for all considered positions have been calculated the maximum needs to be found as equation 5.7 suggests. As already mentioned the source location can also be estimated in a reverberated environment. However the search space consists of many local maxima which subsequently possibly require optimization techniques to accelerate the search process. Three methods are presented to generate appropriate candidate positions for the SRP-PHAT computation.

### 5.8.1 Full-Grid Search

As the name implies, within the *full-grid search* the candidate points are positioned in an equidistant manner. Of course the distance in between candidate points should match the dimension of the search space at some degree, e.g. $1cm$ segments inside a $1m^3$ search cube. Clearly the full-grid search is computationally intensive which leads to another method that is presented hereinafter.

### 5.8.2 Coarse-to-Fine Region Contraction

A very intuitive and effective optimization to the full-grid search and is called *Coarse-to-Fine Region Contraction (CFRC)*, which was presented in [22]. The idea is to initially divide the search volume $V_0$ in search regions with coarser segments compared to the full-grid search. After evaluating the generated candidate positions $J_0$, a subset of the best candidates $N_0$ (with the highest SRP-PHAT values) is picked, which spans a new, smaller search sub-volume $V_{i+1}$ inside the initial search volume. Hence the name *contraction.* This is repeated until an exit criterion is met, e.g. if the number of evaluations $\phi$ is exceeded or if the new sub-volume is smaller than the initially specified minimum volume $V_{min}$. The algorithm overview is illustrated in table 5.1.

```
1. Initialize $i = 0$, $\phi$, $V_{min}$, $V_0$, $J_0$, $N_0$
2. Compute SRP-PHAT $P(\tau_{kl})$ for all $J_i$
3. Find best points $N_i$ in all $J_i$
4. Generate new sub-volume $V_{i+1}$ that encloses all $N_i$
5. Loop 2. to 4. until $i = \phi$ or $V_{i+1} < V_{min}$
6. Return best candidate in $N_i$ with highest SRP-PHAT
```

*Table 5.1: CFRC algorithm pseudo code*

As this algorithm is not self-adapting, all parameters have to be tested and adjusted as required to the considered problem.

### 5.8.3 Performance

As experiments in [22] have shown, both CFRC and *Stochastic Region Contraction (SRC)* (which uses randomized candidate positions) potentially decrease the computational complexity of finding the maximum in SRP-PHAT to only a few percent compared to a full-grid search by keeping the accuracy. This is especially useful for real-time applications.

CFRC is less costly in noisy cases, whereas SRC performs faster in cases with higher SNR. The explanation is that higher noise also causes more local maxima. The deterministic approach from CFRC is then potentially more accurate.

The stochastic method from SRC is said to have higher probability to hit the true location when fewer local maxima are inside the search volume, which is the case for high SNR samples. The algorithm in general is designed for localization of single sound sources. In this thesis this concept has been extended to a multi-target case, by simply splitting the whole environment into separate search volumes for each target source.

# 6 Experiments on Simulated Data

## 6.1 Investigated Array

After the preceding discussions concerning sensor arrays and propagation models in chapter 2, a defined array geometry has been considered throughout this thesis. The device is depicted in figure 6.1.

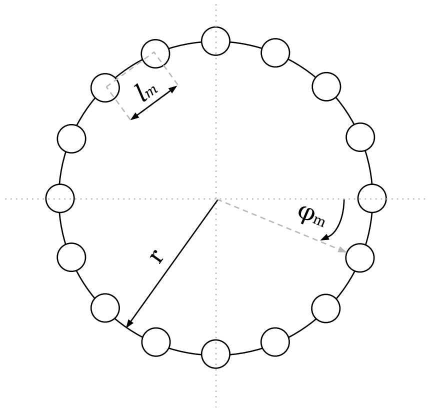


*Figure 6.1: Device under test*

The tested array consists of 16 equidistantly distributed microphones, which means the azimuth spacing is $\varphi_m = \frac{360°}{16} = 22.5°$. The radius of the mounting ring is $r = 0.275\ m$ and the euclidean minimum distance between two adjacent microphones is $l_m = 0.1073\ m$. So far the single array (16 channels) has been investigated theoretically in chapter 3, e.g. in beampatterns or balloon plots. The dual array configuration (32 channels) has been used in simulation experiments (in chapter 6, e.g. simulated beampatterns, simulated speech evaluation) and also for the real recordings (see chapter 7) where both exemplars are mounted on the ceiling at a horizontal (estimated) distance of 2 meters (measured from the array centers). As the array has been designed before this thesis' kickoff, there is still need for verification concerning the wave propagation model. The rule-of-thumb mentioned in section 2.7 can also be interpreted as a transition area between the spherical wave propagation model and the plane wave propagation model.

$$r_{spherical} < \frac{2L^2}{\lambda} < r_{plane} \tag{6.1}$$

Both variants single array (16 microphones) and dual array (32 microphones) will be considered to determine the appropriate wave propagation models according to the rule-of-thumb estimate. The maximum sensor distance for single array usage is $L_{single} = 0.55\ m$. On the other hand the maximum sensor distance for dual array usage is $L_{dual} = 2.55\ m$ *($L_{dual}$ will be estimated initially in section 7.2, but is here already used for comparative reasons)*. In table 6.1 the formula $\frac{2L^2}{\lambda} = \frac{2L^2 \cdot f}{c}$ will be evaluated for both array configurations single/dual at particular frequencies.

$$r_{single} = \frac{2 \cdot L_{single}^2 \cdot f}{c} \quad \text{where } L_{single} = 0.55m \tag{6.2}$$

$$r_{dual} = \frac{2 \cdot L_{dual}^2 \cdot f}{c} \quad \text{where } L_{dual} = 2.55m \tag{6.3}$$

| Frequency | $r_{single}$ | $r_{dual}$ |
|---|---|---|
| 125 Hz | 0.22 m | 4.74 m |
| 250 Hz | 0.44 m | 9.48 m |
| 500 Hz | 0.88 m | 14.22 m |
| 1000 Hz | 1.76 m | 18.96 m |
| 2000 Hz | 3.52 m | 23.70 m |
| 4000 Hz | 7.10 m | 28.44 m |

*Table 6.1: Transition boundaries between spherical and plane wave propagation*

The values for $r_{single}$ and $r_{dual}$ in table 6.1 illustrate the suggested transition boundaries between the spherical and the plane wave propagation model for both array configurations at several, interesting frequencies. Below any specific $r_{single}$ or $r_{dual}$ from table 6.1 spherical waves shall be considered, above any specific $r_{single}$ or $r_{dual}$ plane wave assumption is valid according to the rule-of-thumb.

The transition boundaries for the single array are located mostly at rather small distances, which means the single array should be most likely receiving plane waves. For the dual array configuration the opposite can be assumed, so that the transition boundaries are at sufficient distance and the array most likely receives spherical waves.

It should also become clear, that the considered propagation model is of course determined by the application and its frequency range (e.g. speech). However from table 6.1 it can be also seen that no clear decision criterion can be concluded, whether if the single array data shall be computed considering the spherical wave or the plane wave model, because most of the distances are relevant for the real experiments. With the dual array configuration for the lower frequencies both propagation models could be taken into consideration, for the higher frequencies spherical waves can be assumed.

In listening tests with single array configuration (16 channels) and the real recordings it has been found, that the computed outputs, using both propagation models, subjectively do not differ. In single array simulations (beampatterns, simulated speech experiments) the differences between both propagation models achieved just marginal differences.
In the dual array case (32 channels) the spherical wave assumption has been found to be the superior method (described more in detail in section 2.7.3).

## 6.2 Room Simulation

To determine the characteristics of the aforementioned signal enhancement strategies, several experiments have been conducted. Therefore the *Roomsimove Toolbox* (embedded within the *Multichannel BSS Locate Toolbox*) [23, 24] has been used to generate simulated microphone signals in a shoebox-like room environment. The toolbox is supplied with the room geometry, each wall's absorption coefficients, the positions of the microphones and the position(s) of the sound source(s). The sources are assigned to the desired input signals (e.g. WAV files). From the geometry data the multipath impulse responses are computed and used as filters for the input signals.

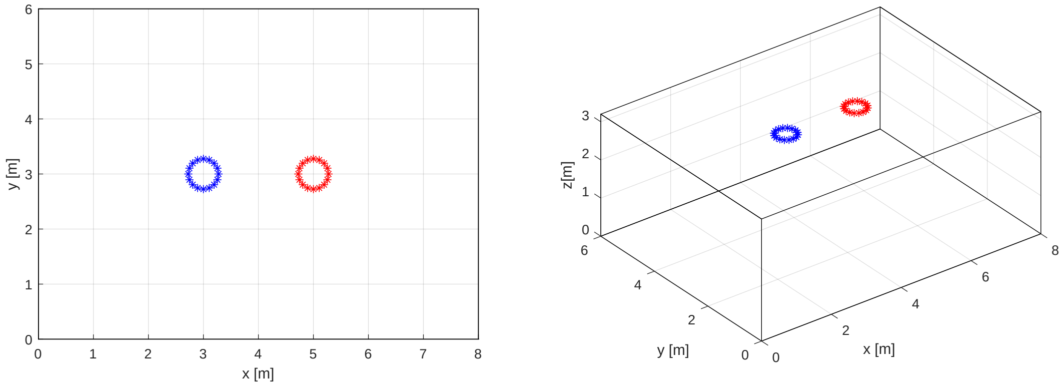


*Figure 6.2: Room simulation*

The simulated room geometry (illustrated in figure 6.2) has been adapted to match the reverberation time ($RT_{60}$) from the real recordings that are discussed in the upcoming chapter 7. The $RT_{60}$ is the time constant where 60 dB of the supplied energy has been decayed. Usually short impulses such as claps, pistol shots, sine sweeps or pseudo-random noise signals are used to measure the reverberation time because of their dense signal energy. For this application the four calibration claps (from the real recordings mentioned before) have been determined to have a $RT_{60}$ between 200 and 250 milliseconds. Therefore the simulated room volume $V$ has been specified as 8m x 6m x 3.2m (length x width x height) and the absorption coefficients for all surfaces $S$ have been set to $\alpha = 0.65$. Using the Sabine formula this results in a reverberation time of 200 milliseconds.

$$RT_{60} = \frac{24 \cdot \ln(10)}{c} \frac{V}{S \cdot \alpha} \tag{6.4}$$

The toolbox finally generates the simulated audio tracks for the specified microphone array geometry. For the current chapter two circular arrays with 16 omnidirectional microphones each and radius $r = 0.275m$ have been simulated. The distance between both array centers will be estimated in chapter 7, but is already used for the simulations in this chapter and specified as $d = 2.0m$.

The upcoming experiments can be separated into two parts:
○ In section 6.3 the dual array is simulated to be exposed to a circulating noise sound source and the directional sensitivity (beampattern) of the investigated algorithms is analyzed.
○ In section 6.4 the dual array is steered towards a varying number of simulated speech sources and the perceived quality will be observed.

The implementation details for the investigated algorithms can be found in section 7.5. Except for experiments 6.3.3 and 6.4.6 the separation parameter $\beta$ is fixed to 3.0, as it has shown to produce a suitable enhancement of separation, while keeping the quality degradation of the target signal moderate. Changes to that parameter consequently result in a tradeoff between interference reduction and quality loss.

## 6.3 Beampatterns - Moving Noise Source in a Simulated Room

For this experiment one sound source (S1) has been specified to emit a zero-mean, Gaussian-distributed signal with variance $\sigma^2 = 0.05$. The noise source (S1) starts at an elevation angle of -45° (measured from the dual array's center horizontal plane) and at a constant euclidean distance of 2.775 meters to circulate. The source moves virtually at a speed of 1° of azimuth angle per second, rotating in a circular shape around both arrays resulting in a 360 second long audio signal. The arrays are steered towards 180° azimuth and -45° elevation.

The 32 microphone signal tracks are then processed by a particular signal beamforming method (and optional post-filter) and the normalized output will be plotted in two variants:
◦ Via Short Time Fourier Transform the signal is transformed frame-by-frame into frequency domain to display the spectral content over time (resp. azimuth) in a spectrogram.
◦ When the noise source is located exactly in steering direction, the respective STFT snapshots are time-averaged and plotted in logarithmic power scale [25].

The beampattern experiments within this chapter are divided into three parts:

◦ Section (6.3.1) Beamformers only, Steering direction 180°.
◦ Section (6.3.2) Beamformers only, Steering direction 270°.
◦ Section (6.3.3) Beamformers + TFM and variable $\beta$, Steering direction 180°.

In the experiments in sections 6.3.1 and 6.3.2 the beampatterns for plain beamforming techniques (that have been discussed in chapter 3) will be computed. This will allow us to make statements about the directional characteristics of each method.

In section 6.3.3 the simple delay-and-sum beamformer will be revisited and extended by the ideal ratio mask presented in 4.3 and its application in 4.4. The steering direction will be used as the target signal and a fixed number of interfering directions will be specified as noise signals. This means that the array is steered towards each specified direction sequentially and all of those resulting beamformed tracks will be combined within the time-frequency masking algorithm to produce a single output. After beamforming the ideal ratio mask serves to remove noisy content from the target signal. The impact of using several interfering signals and a variable separation parameter $\beta$ will be shown in this experiment. This allows us to control the intensity of separation from the target signal versus the interferer signals.

### 6.3.1 Beampattern, Steering Direction at 180°

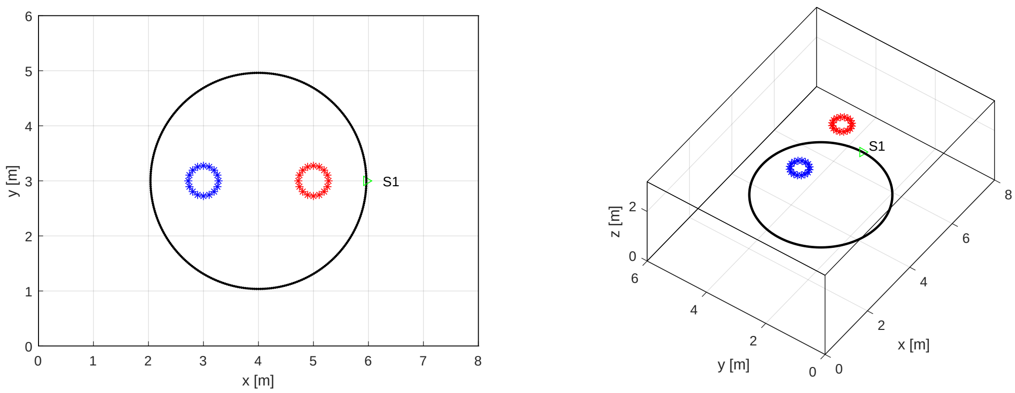
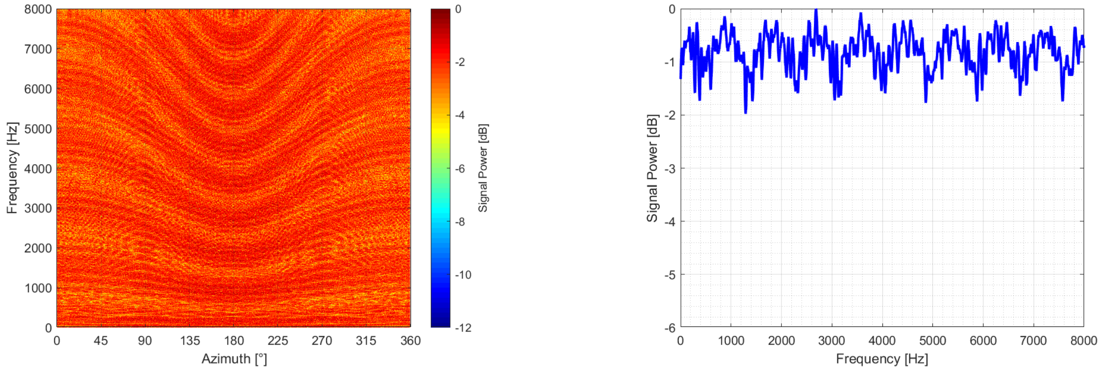


*Figure 6.3: Room simulation using a rotating noise source, Steering direction at* 180°

In the first experiment we investigate the presented beamformer characteristics, when the array is steered to the front-end at 180° (in analogy to *endfire* operation for ULAs). The green triangle (S1) in figure 6.3 marks the steering direction and the black circle shows the movement route of the noise source around the dual array.

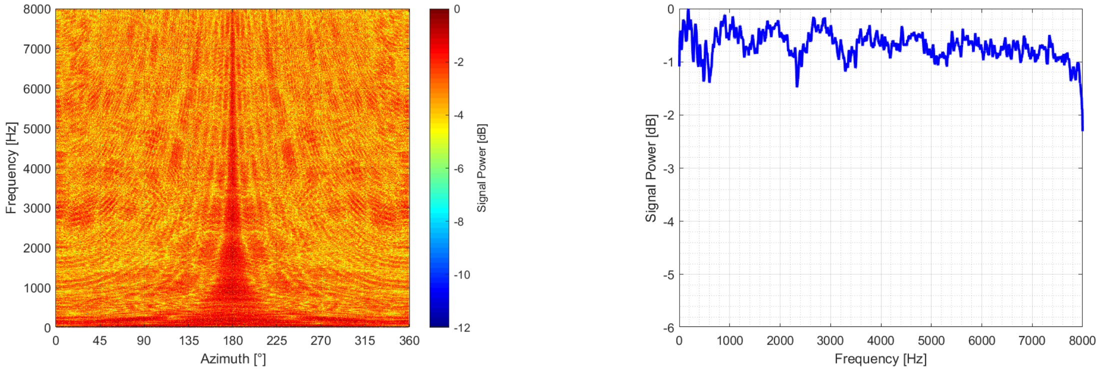
Parameters:
- Steering direction: $\varphi = 180°$, $\theta = -45°$
- Source position: $\varphi = [0° \ ... \ 360°]$, $\theta = -45°$



*(a) Beampattern*

*(b) PSD esimate at* 180°

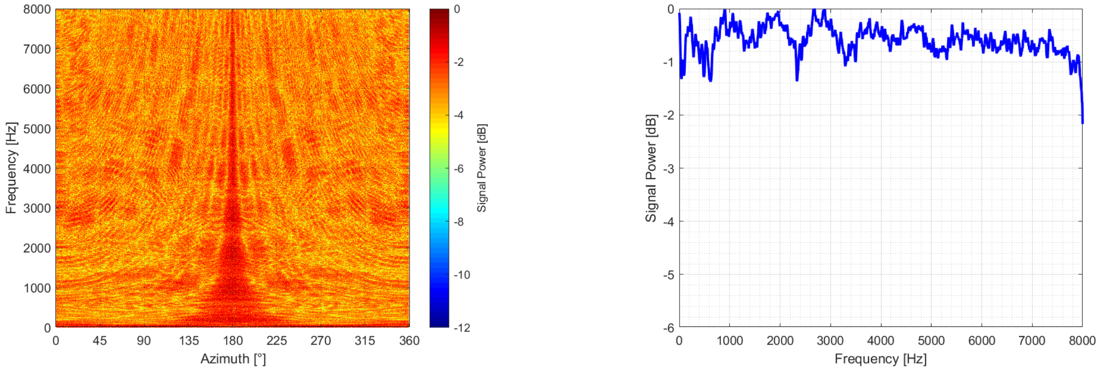
*Figure 6.4:* ***Closest microphone signal****, Steering direction at* 180°



*(a) Beampattern*

*(b) PSD esimate at* 180°

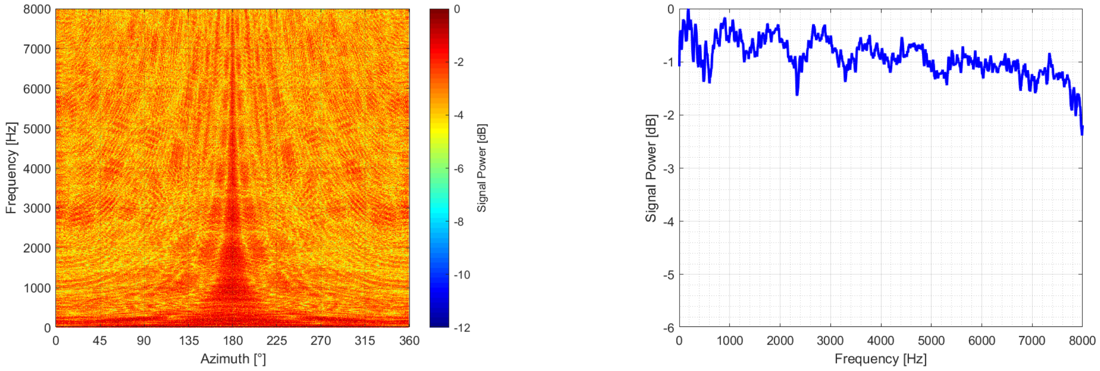
*Figure 6.5:* ***D&S-BF*** *output, Steering direction at* 180°



*(a) Beampattern*

*(b) PSD esimate at* 180°

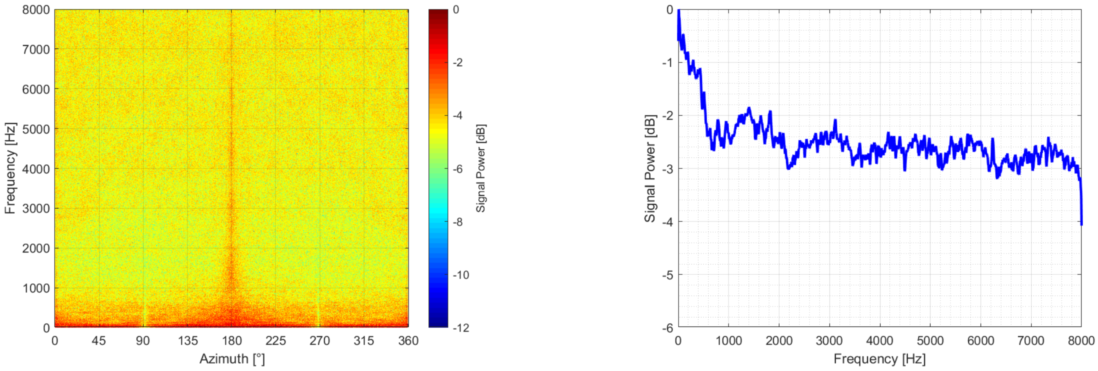
*Figure 6.6:* ***SD-BF*** *output, Steering direction at* 180°



*(a) Beampattern*

*(b) PSD esimate at* 180°

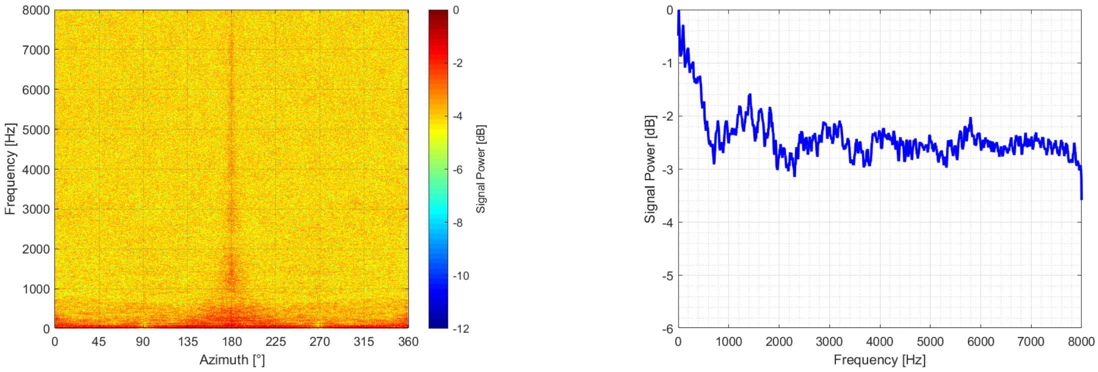
*Figure 6.7:* ***GSC*** *output, Steering direction at* 180°



*(a) Beampattern*

*(b) PSD esimate at 180°*

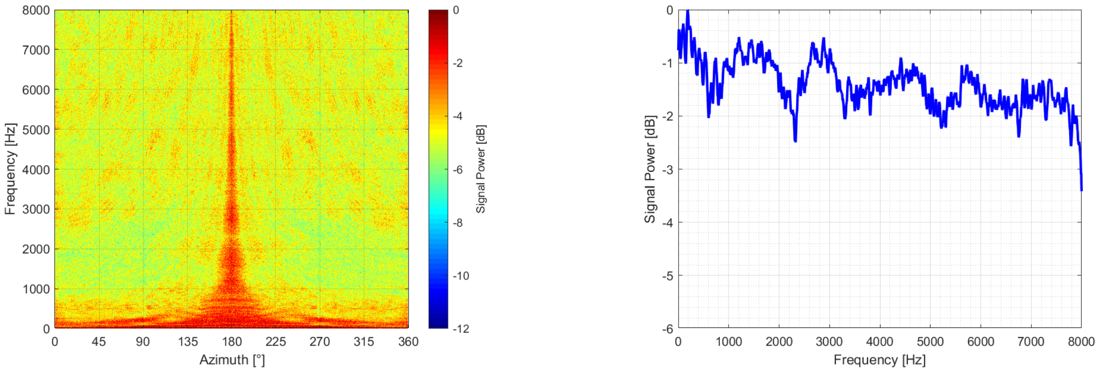
*Figure 6.8: **MPDR-DL** output, Steering direction at 180°*



*(a) Beampattern*

*(b) PSD esimate at 180°*

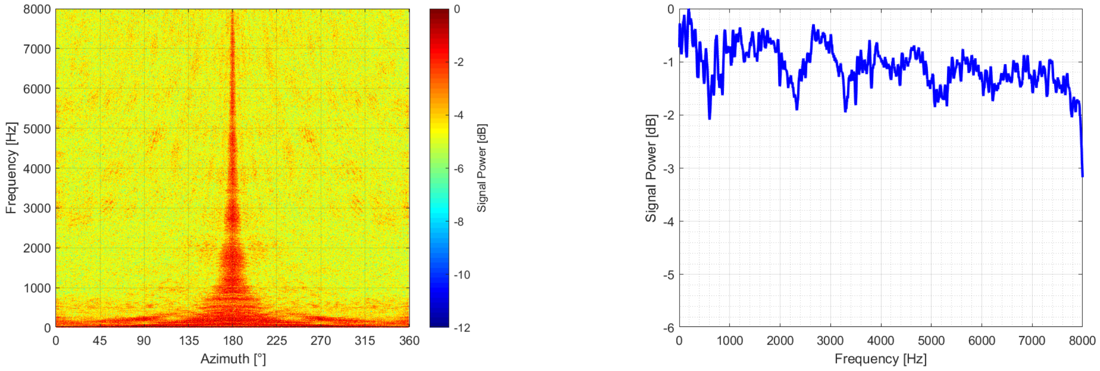
*Figure 6.9: **MPDR-VL** output, Steering direction at 180°*



*(a) Beampattern*

*(b) PSD esimate at 180°*

*Figure 6.10: **MVDR-DL** output, Steering direction at 180°*



*(a) Beampattern*

*(b) PSD esimate at* 180°

*Figure 6.11:* ***MVDR-VL*** *output, Steering direction at* 180°

∘ The SD-BF is very capable of suppressing low-frequent side lobes, but has worse performance in suppressing higher-frequent side lobes (compared to e.g. D&S-BF).

∘ At lower frequencies the GSC performs similar to the D&S-BF, but maintains a better side lobe suppression up to higher frequencies. Interestingly the GSC has a slight high frequency roll-off at steering direction, which is contradictory in comparison to listening tests of processed speech signals.

∘ The MPDR beamformers have an extremely narrow main lobe, which is very difficult to use in real applications. However in practice the MPDR signal output is afflicted with heavy distortion and therefore does not produce a very meaningful output. Hence both MPDR variants will be skipped in further experiments.

∘ The MVDR beamformers in general succeed much more in suppressing side lobes compared to all the other techniques. The signal in steering direction also shows distinct resonances, where the D&S-BF, SD-BF and the GSC signals are rather flat in comparison. The MVDR-VL produces a more pronounced main lobe compared to the MVDR-DL method.

∘ The beampatterns illustrate significant differences between the presented beamforming techniques in terms of main lobe width and side lobe suppression. These insights may already affect the choice for a specific beamforming method, but as we will see in other experiments at a later point, there are other qualities that will have to be considered.

### 6.3.2 Beampattern, Steering Direction at $270°$

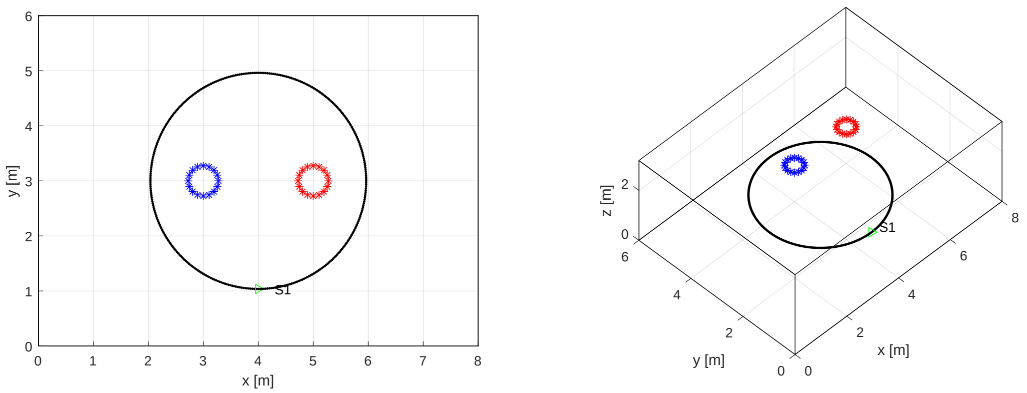
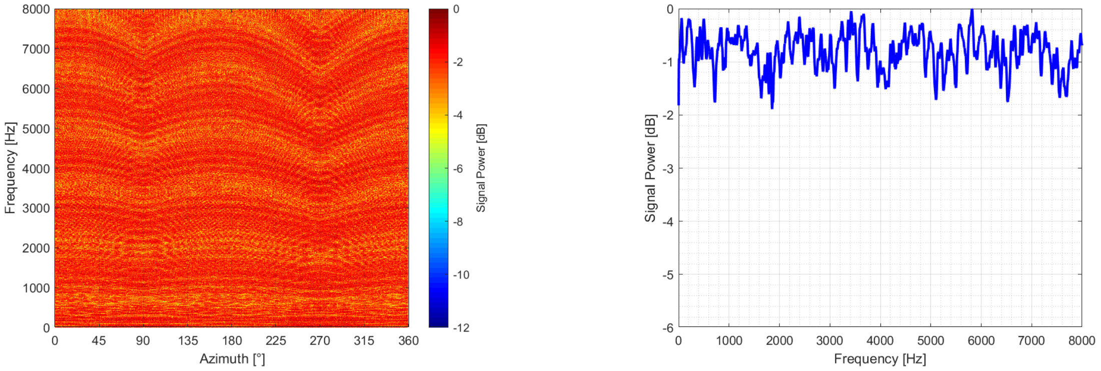


*Figure 6.12: Room simulation using a rotating noise source, Steering direction at* $270°$

In the second experiment the beamformer characteristics are investigated, when the array is steered to the broad-end at $270°$ (in analogy to *broadside* operation for ULAs). Again the green triangle (S1) in figure 6.3 marks the steering direction and the black circle shows the movement route of the noise source around the dual array.

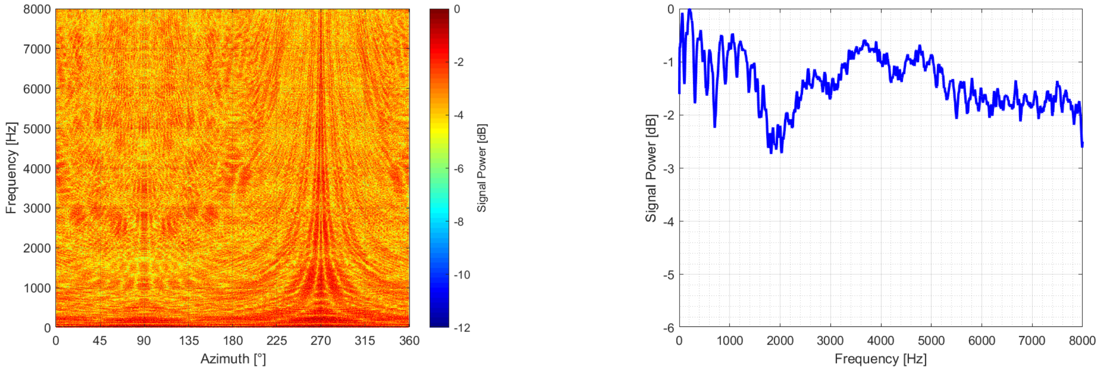
Parameters:
- Steering direction: $\varphi = 270°$, $\theta = -45°$
- Source position: $\varphi = [0° \ldots 360°]$, $\theta = -45°$



*(a) Beampattern*

*(b) PSD esimate at* $270°$

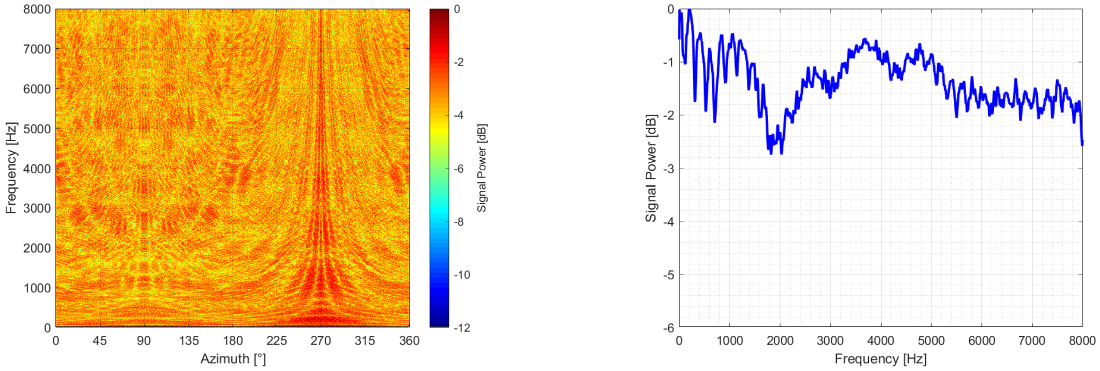
*Figure 6.13:* ***Closest microphone signal****, Steering direction at* $270°$



*(a) Beampattern*

*(b) PSD esimate at 270°*

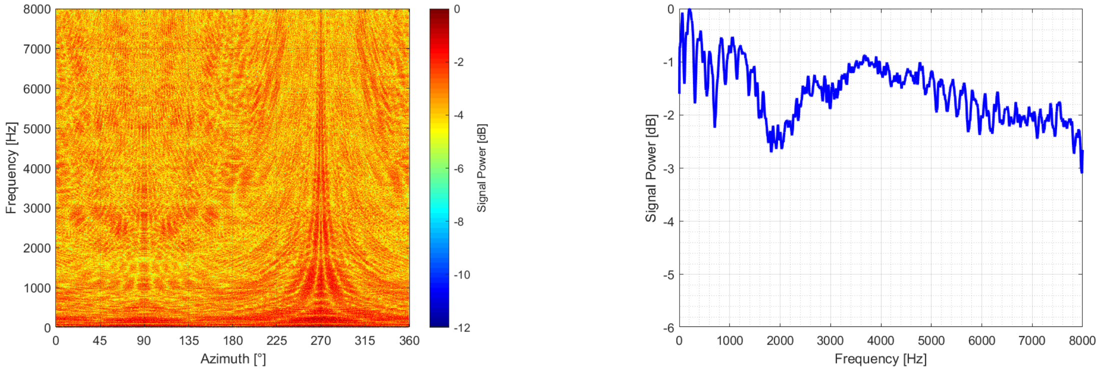
*Figure 6.14: **D&S-BF** output, Steering direction at 270°*



*(a) Beampattern*

*(b) PSD esimate at 270°*

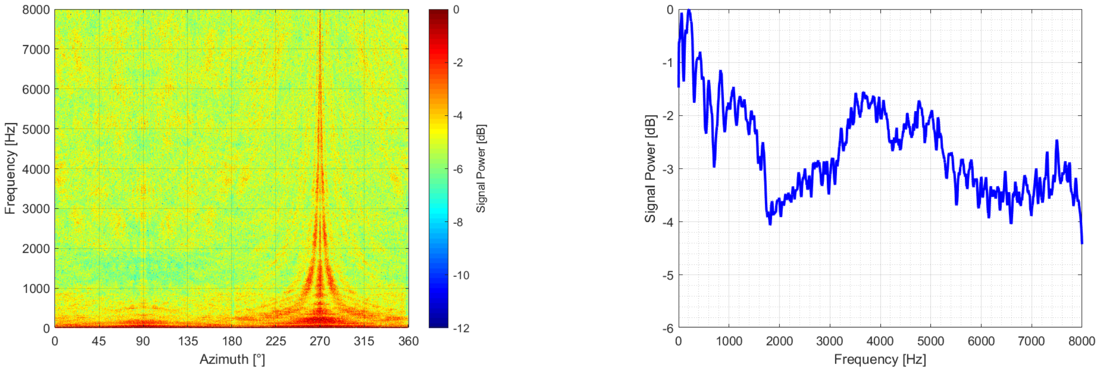
*Figure 6.15: **SD-BF** output, Steering direction at 270°*



*(a) Beampattern*

*(b) PSD esimate at 270°*

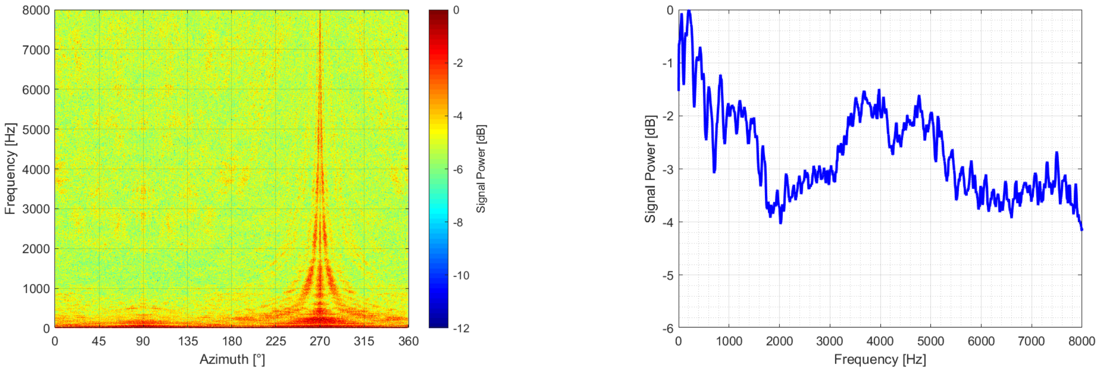
*Figure 6.16: **GSC** output, Steering direction at 270°*



*(a) Beampattern*

*(b) PSD esimate at* 270°

*Figure 6.17: **MVDR-DL** output, Steering direction at* 270°



*(a) Beampattern*

*(b) PSD esimate at* 270°

*Figure 6.18: **MVDR-VL** output, Steering direction at* 270°

○ The SD-BF has similar qualities as in the former experiment (low frequency suppression) but here also has smaller side lobe levels in higher frequency regions than the D&S-BF.

○ The GSC again performs similar to the D&S-BF at low frequencies, has a better side lobe suppression at higher frequencies and again a subtle high frequency roll-off in steering direction.

○ The MVDR beamformers again have a much higher side lobe suppression. The 1st side lobe is still recognizable, but the 2nd side lobe is almost gone. MVDR-VL and MVDR-VL produce a rather similar output here.

○ In general the main lobe in this experiment is more narrow compared to the first experiment, but also the 1st and the 2nd side lobe may be considered as contributing to the target direction. The signals in the steering direction show more resonances than in the first experiment.

### 6.3.3 Beampatterns - D&S-BF + TFM, Varying Separation Parameter $\beta$

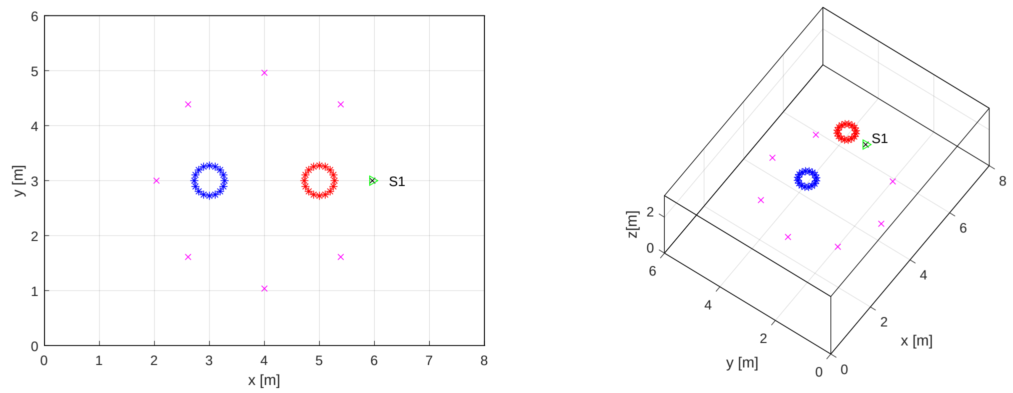
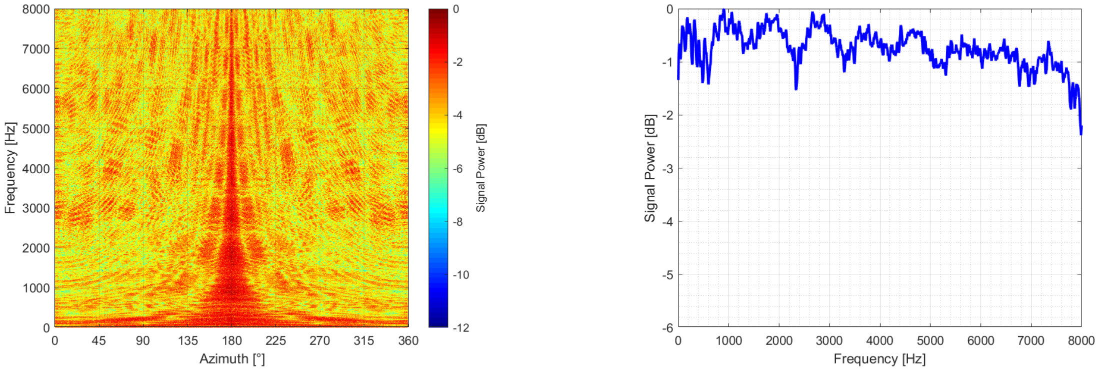


*Figure 6.19: Room simulation using a rotating noise source, Steering direction at* 180°

To make the impact of the separation parameter visible, in the third experiment a delay-and-sum beamformer has been used in combination with time-frequency masking (ideal ratio mask) with eight, equiangular target directions (one desired target + seven interfering targets) and a varying separation parameter $\beta$. In figure 6.19 the interfering directions are marked as pink crosses, the desired target is shown as black cross (S1). In section 6.4.6 a similar setup with speech signals will be considered in terms of subjective quality evaluation.

Parameters:

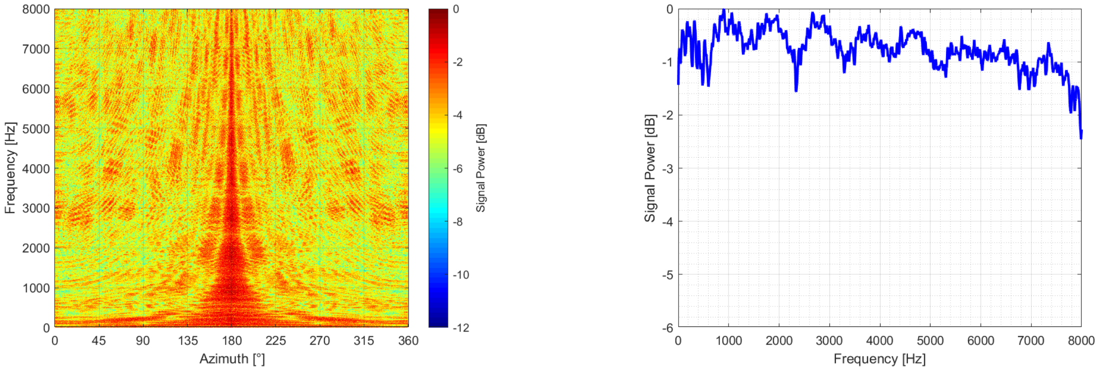
- Steering direction: $\varphi = 180°$, $\theta = -45°$
- Source position: $\varphi = [0° \; ... \; 360°]$, $\theta = -45°$
- Interfering direction #1: $\varphi = 0°$, $\theta = -45°$
- Interfering direction #2: $\varphi = 45°$, $\theta = -45°$
- Interfering direction #3: $\varphi = 90°$, $\theta = -45°$
- Interfering direction #4: $\varphi = 135°$, $\theta = -45°$
- Interfering direction #5: $\varphi = 225°$, $\theta = -45°$
- Interfering direction #6: $\varphi = 270°$, $\theta = -45°$
- Interfering direction #7: $\varphi = 315°$, $\theta = -45°$
- Used TF-Mask: Ideal Ratio Mask
- Varying separation parameter $\beta = \{\ 4.0,\ 2.0,\ 3.0,\ 1.0,\ 0.5\ \}$



*(a) Beampattern*

*(b) PSD esimate at 180°*

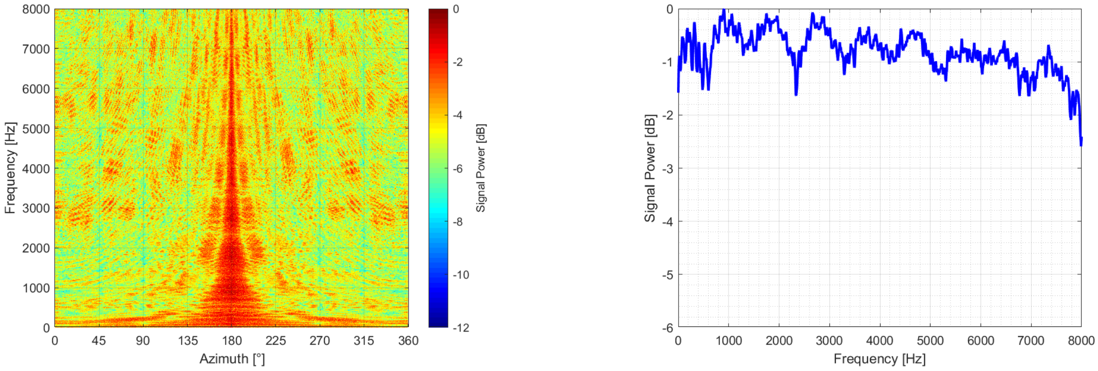
*Figure 6.20:* ***D&S-BF + TFM****, $\beta = 4.0$, Steering direction at 180°*



*(a) Beampattern*

*(b) PSD esimate at 180°*

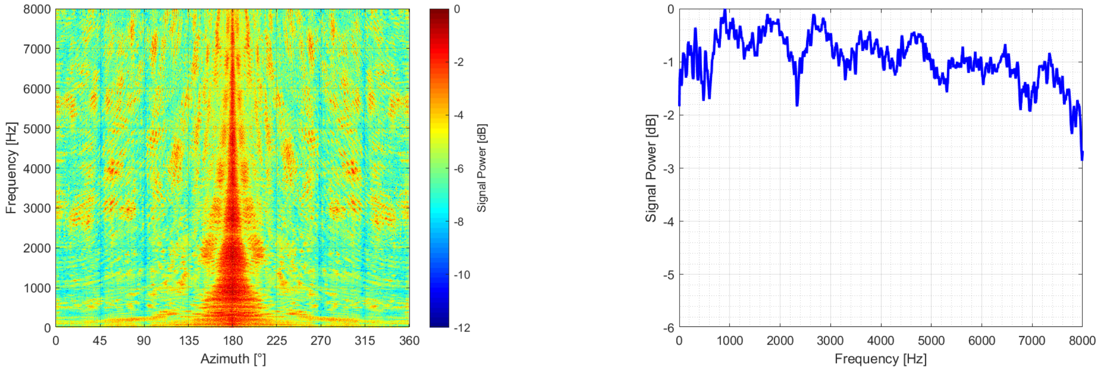
*Figure 6.21:* ***D&S-BF + TFM****, $\beta = 3.0$, Steering direction at 180°*



*(a) Beampattern*

*(b) PSD esimate at 180°*

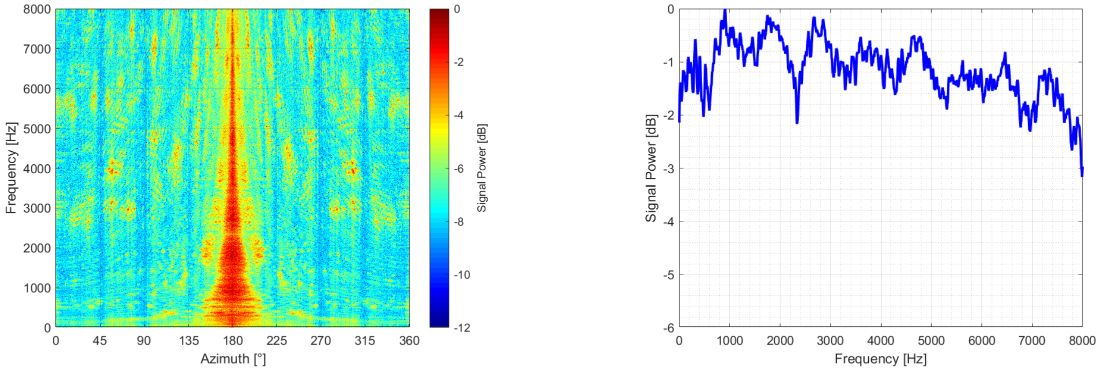
*Figure 6.22:* ***D&S-BF + TFM****, $\beta = 2.0$, Steering direction at 180°*



*(a) Beampattern*

*(b) PSD esimate at* 180°

*Figure 6.23:* ***D&S-BF + TFM****,* $\beta = 1.0$*, Steering direction at* 180°



*(a) Beampattern*

*(b) PSD esimate at* 180°

*Figure 6.24:* ***D&S-BF + TFM****,* $\beta = 0.5$*, Steering direction at* 180°

∘ For the 8-target case the notches can be seen in the beampattern at 0°, 45°, 90°, 135°, 225°, 270° and 315°. By choosing smaller $\beta$, the notches are getting deeper. The side lobe level in general is also drastically reduced by decreasing $\beta$, which subsequently leads to an SNR improvement.

∘ We can conclude that by increasing the amount of interfering targets, the main lobe in steering direction becomes more and more prominent compared to the side lobes, which also leads to SNR improvement. This approach could be relevant to shape the beam in applications, where only the desired target location is given. By specifying several arbitrary interfering directions (e.g. at equidistant angles as in the experiment) all of the directions other than the desired direction are forced to be suppressed.

∘ The signal in steering direction starts to dissolve for low separation parameters, which results in degradation of the desired signal. One has to keep in mind that despite the enormous potential of time-frequency masking, the desired signal also suffers from coloration and introduction of artifacts for lower separation parameters, which will be also shown in subjective experiments in section 6.4.6.

## 6.4 Signal Enhancement - Static Speech Sources in a Simulated Room

The second part of the simulation experiments focuses on the subjective quality features produced by the signal enhancement algorithms that have been discussed so far. As described in section 6.3 the same room simulation environment has been used, but the used source(s) have been changed. For the following experiments no random-generated signals have been used, but instead speech recordings, that are provided free of charge by the *Centre for Speech Technology Research (University of Edinburgh)*. The speaker recordings consist of native english speech with various accents and is named *CSTR VCTK Corpus* [26].

Further the sound sources are not moving (as in the former beampattern experiments) but at fixed positions instead. The number of present talkers starts with one talker and will be raised up to three talkers. The speakers S1 to S3 are alternating gender: male, female, male. The composition of the simultaneous talkers is not intended to represent real conversations with normal speech pauses. Instead each speaker is talking constantly. Random utterances for each speaker are appended to form 40 seconds of contextless sentences. The signal amplitudes of all speakers are normalized to have the same RMS value over the whole time period.

Similarly to section 6.3.3 the impact of the separation parameter will be investigated more in detail for the 3-speaker case in section 6.4.6.

Time-alignment of the compared signals is crucial to the evaluation outcome, therefore all signals are time-aligned based on cross-correlation in advance.

### 6.4.1 Evaluation Framework

Figure 6.25 illustrates the block diagram of the evaluation algorithm. According to the geometrical information of each speaker and the microphone positions the MIMO impulse responses are computed to simulate room-like reverberation. The clean speech signals $s_j(n)$ (where $j$ is the speaker index) are then filtered with the room impulses to generate the simulated microphone tracks $x_m(n)$ (where $m$ is the microphone index). The microphone that is nearest to the desired target is labelled as *Closest microphone signal* and the associated audio track will be considered as the reference input signal.

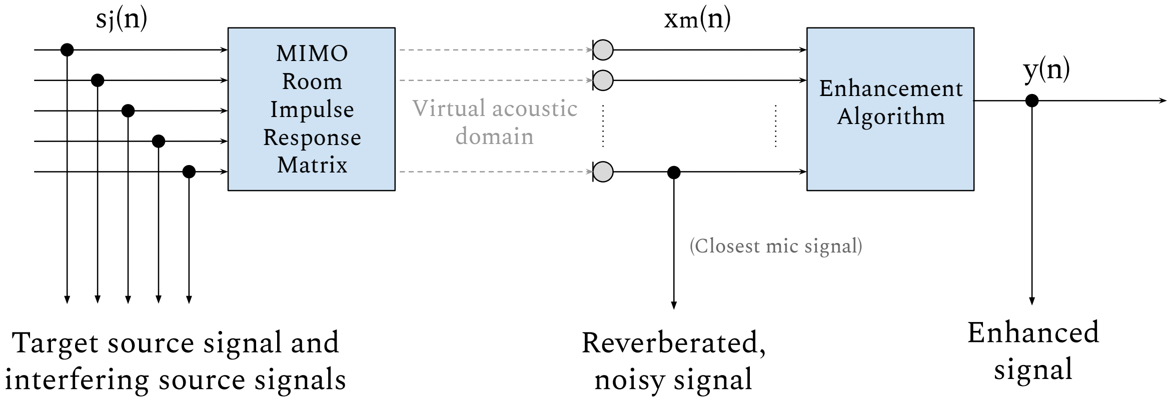

sj(n)
MIMO Room Impulse Response Matrix
Virtual acoustic domain
xm(n)
Enhancement Algorithm
y(n)
(Closest mic signal)
Target source signal and interfering source signals
Reverberated, noisy signal
Enhanced signal

*Figure 6.25: Evaluation Block Diagram*

All microphone signals $x_m(n)$ will be processed by the enhancement algorithm under investigation to produce the enhanced output signal $y(n)$, considered as reference output signal.

### 6.4.2 Perceptual Evaluation Methods for Audio Source Separation

In this section the metrics of the *Perceptual Evaluation Methods for Audio Source Separation (PEASS)* project [27, 28] will be presented very briefly. Many other popular evaluation algorithms have been considered for this thesis at first, but in the end the PEASS method was the only one to remain due to reasons of clarity, comprehensibility and because of its compact way to represent the most important aspects of the processed speech signal.

In contrast to all other assessment methods the PEASS toolkit does not only expect the clean source signal and the enhanced signal but further also the interfering signals are needed as inputs, which allows to give more elaborate prediction about the perceived quality. The required inputs can be recognized in the tapped signal paths in figure 6.25.

The PEASS algorithm is separated into three stages:

○ In the first step the estimation error is decomposed and expressed as energy ratios: Source to distortion ratio ($SDR$), Source image to spatial distortion ratio ($ISR$), Source to interference ratio ($SIR$) and Source to artifacts ratio ($SAR$).

○ In the second stage an auditory model (PEMO-Q [29]) will be applied on the decomposed signals to compute the PEMO-Q quality features: $q_{overall}$, $q_{target}$, $q_{interf}$, and $q_{artif}$.

○ In the final step the PEMO-Q features are mapped non-linearly by a feedfoward neural network, which is trained on subjective opinion scores to finally produce the PEASS metrics: overall perceptual score ($OPS$), target-related perceptual score ($TPS$), interference-related perceptual score ($IPS$) and artifacts-related perceptual score ($APS$).

In this thesis only the last four scores will be considered. High $TPS$ values indicate how well the desired signal has been preserved, whereas high $IPS$ values imply good suppression of interfering signals and high $APS$ values denote a small amount of artifacts getting introduced. The $OPS$ value illustrates the total signal quality, but is actually not the mean value of the preceding scores. For all of the 4 values it is valid to say: the higher the number, the better the quality.

### 6.4.3 Speech Simulation for 1 Speaker, Steering direction at 180°

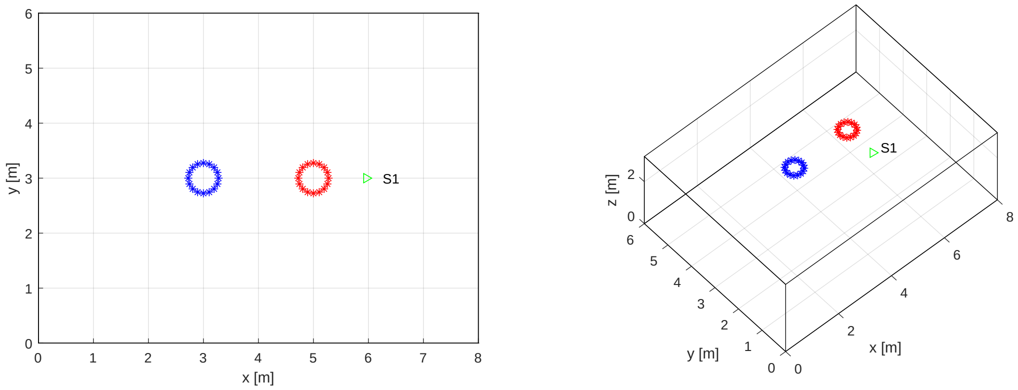


*Figure 6.26: Room simulation using 1 Speaker, Steering direction at* 180°

```
Parameters:
- Steering direction:  φ = 180°, θ = −45°
- Source position (S1):  φ = 180°, θ = −45°
```

| | *OPS* | *TPS* | *IPS* | *APS* |
|---|---|---|---|---|
| D&S-BF | 68.82 | **96.22** | 71.30 | 73.56 |
| GSC | 74.64 | 94.66 | 68.89 | 65.25 |
| SD-BF | **79.67** | 95.06 | **78.99** | **76.10** |
| MVDR-DL | 66.41 | 90.88 | 61.26 | 60.33 |
| MVDR-VL | 59.59 | 84.46 | 60.68 | 53.58 |

*Table 6.2: PEASS metrics for 1 Speaker, Steering direction at* 180°

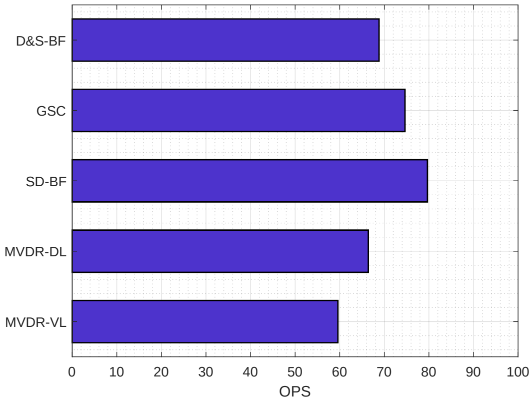


*Figure 6.27: Overall perceptual score for 1 Speaker, Steering direction at* 180°

The most significant feature of the close microphone signal is the room reverberation of the speech signal, which will be processed by the beamformers at different qualities. The D&S-BF sounds a bit dull, whereas the GSC adds more brightness, but both sounding the most neutral and having similar performance at reverberation suppression. The SD-BF is the only method to suppress the room reverberation at lower frequencies, but already produces a totally different tone color compared to the clean speech signal. Both MVDR beamformers have worse performance in reverberation suppression and very extreme tonal, slightly distorted characteristics. Beyond that the MVDR-DL sounds very boomy and the MVDR-VL sounds piercingly bright.

### 6.4.4 Speech Simulation for 1 Speaker, Steering direction at 270°

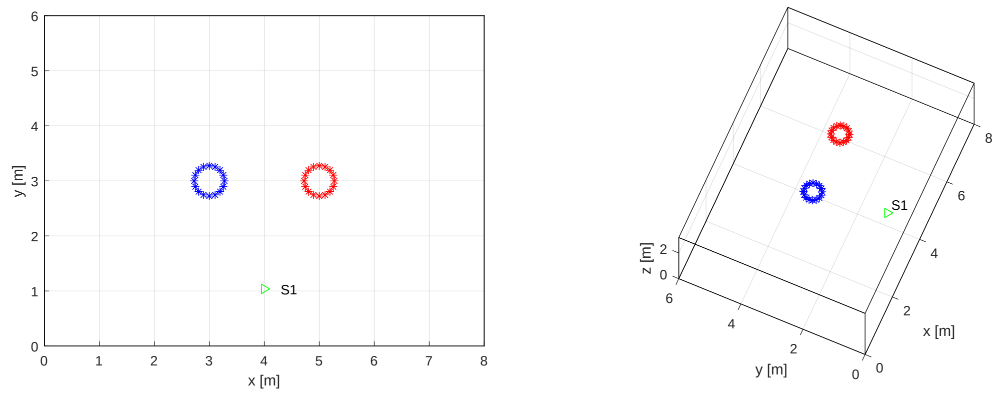


*Figure 6.28: Room simulation using 1 Speaker, Steering direction at* 270°

Parameters:
- Steering direction: $\varphi = 270°$, $\theta = -45°$
- Source position (S1): $\varphi = 270°$, $\theta = -45°$

| | *OPS* | *TPS* | *IPS* | *APS* |
|---|---|---|---|---|
| D&S-BF | 58.31 | **97.87** | 64.19 | 67.77 |
| GSC | **67.98** | 94.21 | 62.61 | 62.32 |
| SD-BF | 62.79 | **97.87** | **69.77** | **70.95** |
| MVDR-DL | 65.19 | 93.68 | 59.99 | 60.46 |
| MVDR-VL | 64.16 | 90.88 | 59.78 | 58.49 |

*Table 6.3: PEASS metrics for 1 Speaker, Steering direction at* 270°

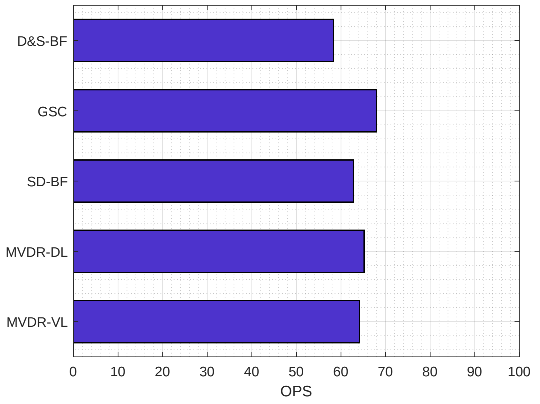


*Figure 6.29: Overall perceptual score for 1 Speaker, Steering direction at* 270°

It is noticeable that all beamformers produce a much more balanced sound in the 180°-task, whereas in the 270°-task they all sound much thinner. For this case the D&S-BF and the SD-BF sound very similar and the GSC is the only method to suppress low-frequent interference. Both MVDR beamformers sound similar to the former task which is interestingly also captured by the PEASS measurements.

### 6.4.5 Speech Simulation for 3 Speakers, Steering direction at $180°$

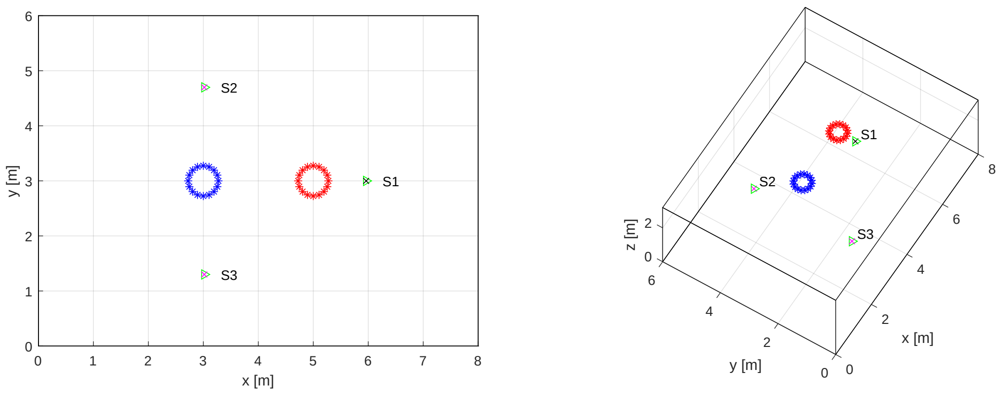


*Figure 6.30: Room simulation using 3 Speakers, Steering direction at* $180°$

```
Parameters:
- Steering direction:  $\varphi = 180°$, $\theta = -45°$
- Interfering direction #1:  $\varphi = 60°$, $\theta = -45°$
- Interfering direction #2:  $\varphi = 300°$, $\theta = -45°$
- Source position (S1):  $\varphi = 180°$, $\theta = -45°$
- Source position (S2):  $\varphi = 60°$, $\theta = -45°$
- Source position (S3):  $\varphi = 300°$, $\theta = -45°$
- Used TF-Mask:  Ideal Ratio Mask, Separation parameter $\beta = 3.0$
```

| | *OPS* | *TPS* | *IPS* | *APS* |
|---|---|---|---|---|
| D&S-BF | 24.36 | 55.74 | 28.90 | 68.75 |
| D&S-BF + TFM | 35.12 | 58.38 | 46.02 | 62.67 |
| GSC | 32.74 | 59.24 | 39.12 | 62.56 |
| GSC + TFM | 41.74 | 62.56 | 54.36 | 57.53 |
| SD-BF | 26.99 | 59.69 | 39.57 | **70.19** |
| SD-BF + TFM | 39.69 | 63.48 | 54.82 | 63.83 |
| MVDR-DL | **46.54** | **66.61** | 66.40 | 51.19 |
| MVDR-DL + TFM | 41.95 | 53.29 | **71.60** | 45.92 |
| MVDR-VL | 40.44 | 61.54 | 63.22 | 49.21 |
| MVDR-VL + TFM | 41.21 | 60.27 | 70.93 | 44.24 |

*Table 6.4: PEASS metrics for 3 Speakers, Steering direction at* $180°$

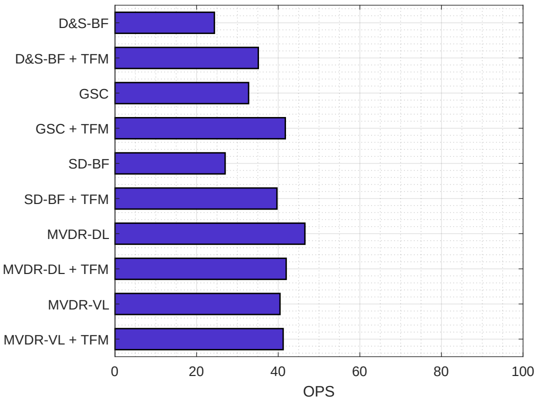


*Figure 6.31: Overall perceptual score for 3 Speakers, Steering direction at* $180°$

By listening at the close mic recording it is hard to tell which of the talkers actually is the target speaker, because of the heavy crosstalk. This 3 speaker task is already technically very challenging, but still each improvement step presented here enhances the target speaker substantially.

The GSC in this case shows again a very similar quality as the D&S-BF, but keeps a better intelligibility in the target direction due to more pronounced higher frequencies and is therefore more advantageous to time-frequency masking. Again the SD-BF shows its quality at suppressing low-frequent interference. All of the afore-mentioned beamformers profit from time-frequency masking. The MVDR beamformers produce quite an amount of artifacts and are therefore not very suitable for time-frequency masking.

### 6.4.6 Speech Simulation for 3 Speakers and variable $\beta$, Steering direction at 180°

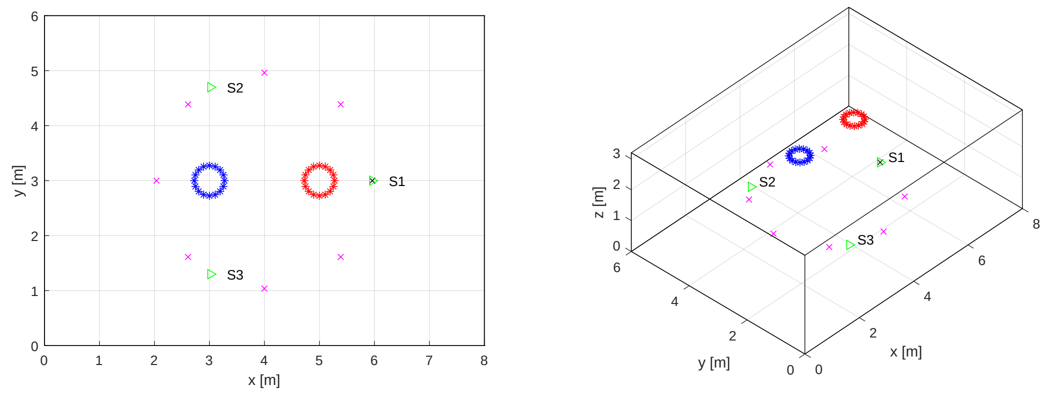


*Figure 6.32: Room simulation for 3 Speakers and variable $\beta$, Steering direction at 180°*

Parameters:
- Steering direction: $\varphi = 180°$, $\theta = -45°$
- Interfering direction #1: $\varphi = 0°$, $\theta = -45°$
- Interfering direction #2: $\varphi = 45°$, $\theta = -45°$
- Interfering direction #3: $\varphi = 90°$, $\theta = -45°$
- Interfering direction #4: $\varphi = 135°$, $\theta = -45°$
- Interfering direction #5: $\varphi = 225°$, $\theta = -45°$
- Interfering direction #6: $\varphi = 270°$, $\theta = -45°$
- Interfering direction #7: $\varphi = 315°$, $\theta = -45°$
- Source position (S1): $\varphi = 180°$, $\theta = -45°$
- Source position (S2): $\varphi = 60°$, $\theta = -45°$
- Source position (S3): $\varphi = 300°$, $\theta = -45°$
- Used TF-Mask: Ideal Ratio Mask
- Varying separation parameter $\beta = \{\ 4.0,\ 3.0,\ 2.0,\ 1.0,\ 0.5\ \}$

| | *OPS* | *TPS* | *IPS* | *APS* |
|---|---|---|---|---|
| D&S-BF | 24.36 | 55.74 | 28.90 | **68.75** |
| D&S-BF + TFM, $\beta = 4.0$ | 36.77 | 60.45 | 46.56 | 60.65 |
| D&S-BF + TFM, $\beta = 3.0$ | 38.63 | 61.37 | 49.68 | 59.52 |
| D&S-BF + TFM, $\beta = 2.0$ | 41.85 | **62.07** | 55.89 | 57.14 |
| D&S-BF + TFM, $\beta = 1.0$ | **44.08** | 62.06 | 62.29 | 53.47 |
| D&S-BF + TFM, $\beta = 0.5$ | 41.84 | 58.68 | **66.63** | 48.44 |

*Table 6.5: PEASS metrics for 3 Speakers and variable $\beta$, Steering direction at 180°*

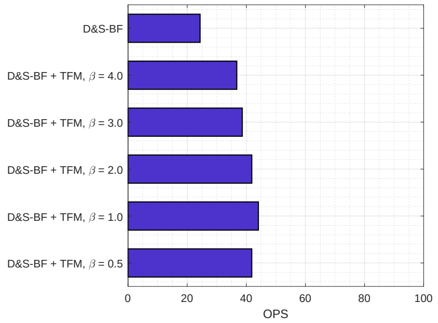


*Figure 6.33: Overall perceptual score for 3 Speakers and variable $\beta$, Steering direction at 180°*

By decreasing the separation parameter the target signal starts to stick out of the mixture more and more. Beginning from $\beta = 4.0$ up to $\beta = 1.0$ the separation and also the intelligibility improves drastically. Around $\beta = 2.0$ the target signal is starting to get narrow-band. At $\beta = 1.0$ and below the separation still increases, but the degradation of the target signal quality is also significantly perceivable.

The evaluation experiments from the last sections have shown that all measurements deteriorate significantly if the number of interfering speakers is increased, which of course was expected but is now captured in terms of numbers.

Characteristics that really stick out when listening at all the samples computed by the presented algorithms are: the quality and intelligibility of the target source and the perturbation by interfering signals and/or artifacts.

Concerning the subjective sound quality MVDR beamformers have shown to produce very colored output, e.g. the MVDR-DL sounds damped and inarticulate, whereas the MVDR-VL sounds exorbitant bright and thus more intelligible. D&S beamformers have a more neutral character and both SD-BF and GSC lie somewhere in between neutral and bright. Generally the sound quality of D&S-BF, GSC and SD-BF are comparable to each other, where as the MVDR beamformers play in their own class due to their extremely colored tone.

Interference suppression is superiorly executed by MVDR beamformers. The side effect is, that the remaining interfering components are distorted and afflicted with artifacts. The other beamformers are not capable of performing such impressive interference suppression, but at the other hand maintaining only a small amount of interference distortion. If time-frequency-masking is considered, small separation parameter values also cause the artifact perturbation to increase.

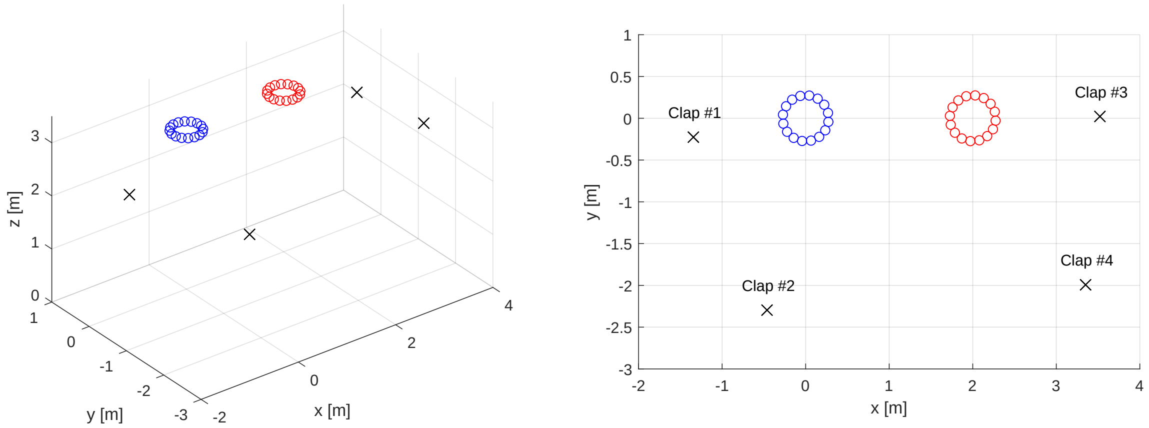
After all this insights the algorithms have been considered for the usage with real-world data instead of simulations, which will be discussed in chapter 7. The conditions for that case will be slightly different compared to the simulated case, but nevertheless the past experiments deliver a good basis to score the given speech enhancement algorithms.

# 7 Experiments on Real-World Data

## 7.1 Description of the Scenario at Concordia Station

One of the biggest challenges in this thesis was to investigate the surrounding environment of the given speaker scenario to perform the best possible speech separation. Prior knowledge was limited to the array geometry of one array (two identical exemplars were used), a rough sketch of the situation without any distance measurements and recordings of claps to post-calibrate the microphones (including rough clap positions on the sketch). Accurate knowledge of the microphone positions is crucial for the signal improvement. From that point the best fitting algorithms to localize and separate the talkers had to be found, to process approximately 30 hours of audio recordings. Further it also had to be considered, that in the future the separated speech tracks will be manually transcribed instead of running them directly into an automatic-speech-recognition engine.

The plots in figure 7.1 illustrate the directions of the claps, which were identified by the SRP-PHAT algorithm after calibrating the arrays. These are not the true positions but instead just the estimates for the direction of arrival in cartesian coordinates, which is sufficient.



*(a) Isometric view*

*(b) Top view*

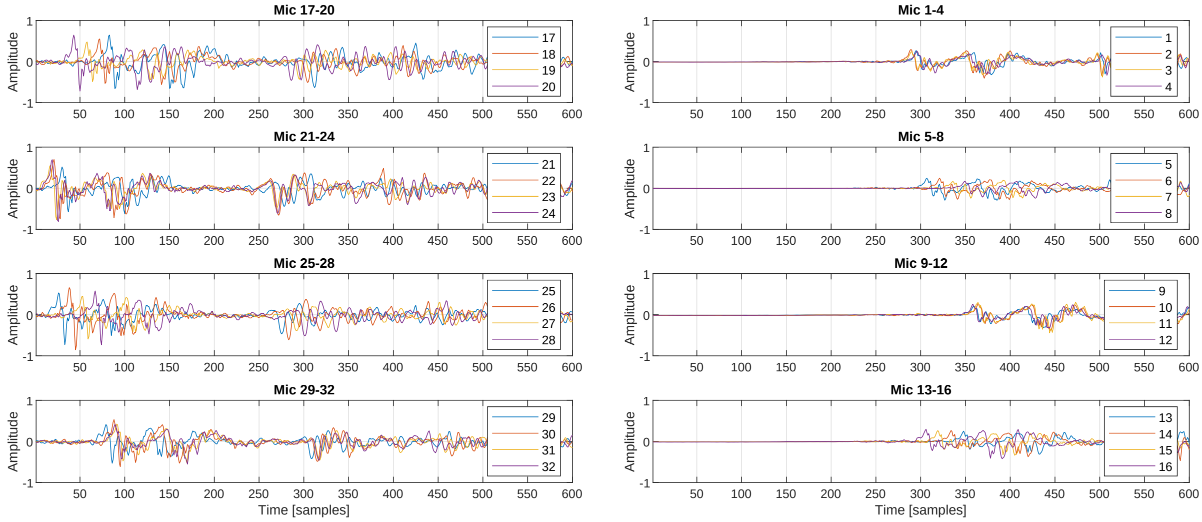
*Figure 7.1: Estimated clap directions*

To illustrate the situation, the pictures in figure 7.1 already show the microphone position configuration that was used in the end, but still it has not been discussed in detail how they were obtained: The array geometry of each specific array is known, but neither the absolute mounting positions nor relative distances between both arrays or other positions were known.
Beyond that the microphone signals were provided as 32 tracks sampled at 48kHz. It was not clear at first how these 32 tracks belonged to the microphone positions, if they were enumer-

ated after their geometrical positions at all, clockwise or counter-clockwise, which of the tracks belonged to the blue array (left), which of them belonged to the red array (right), etc.

So the first idea was to look at the TDOAs of the calibration claps to see how the signals fit to the microphone positions. Of course it was assumed, that the clap positions were sketched correctly and the timings of the microphone signals are synchronized.

The plots in figure 7.2 show the microphone signals of clap #1. By looking at the TDOA the orientation of the microphone signals can be roughly assumed and also the signal amplitudes give information about the distance from the microphone array to the clap position. The enumeration of the microphone tracks can be clearly assigned to both arrays: Tracks 1-16 belong to the red array, whereas tracks 17-32 belong to the blue array. For a better overview the signals in figure 7.2 have been grouped into quartets.



(a) Blue array

(b) Red array

*Figure 7.2: Clap #1 arriving at both arrays*

This procedure has been repeated for the remaining 3 recorded claps. After looking at all of the 4 claps' TDOAs, the orientation of both arrays showed a much clearer but still incomplete picture. The following image 7.3 illustrates the state of knowledge at this point.

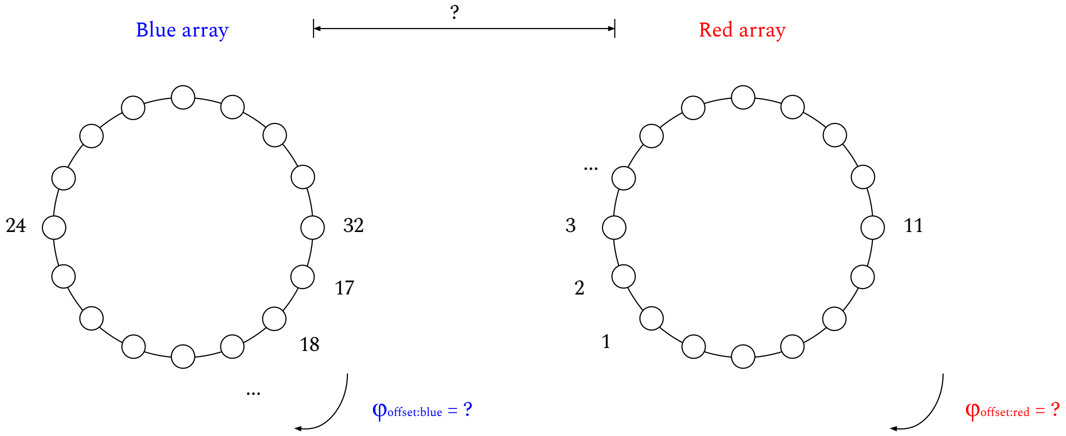


*Figure 7.3: Array without calibration*

Some of the microphone positions are already indicated in figure 7.3 but at this point it is still unknown if there is an azimuth offset in the mounting of the arrays and the distance between both arrays is also unknown.

## 7.2 Microphone Calibration

Existing strategies for microphone calibration often require the knowledge of accurate calibration source positions or the sound velocity to minimize distance errors based on triangulation (such as [30]). As all of that prior knowledge was not given a different approach has been chosen: The distance between the arrays and each azimuth offset was found by doing a brute force parameter search, which means evaluating the SRP-PHAT scores among meaningful parameters. Height differences or angular elevation differences between both arrays have been neglected.

Further also the room temperature was unknown and thus the actual speed of sound could not be computed. But still an assumption had to be made, so the room temperature was supposed to be fixed at 20°C. The brute force method's pseudo code, to estimate the best parameters for one clap, consists of nested for-loops (one loop for each parameter) and looks as follows:

```
for $\varphi_{offset:blue} = -45° : 1° : +45°$ do
{
    for $\varphi_{offset:red} = -45° : 1° : +45°$ do
    {
        for $distance = 1.00m : 0.05m : 1.80m$ do
        {
            Compute the position at SRP-PHAT maximum
        }
        end for
    }
    end for
}
end for
```

The algorithm above has been performed for all of the 4 clap samples. The search volumes required for the SRP-PHAT computations have been roughly fitted to the guessed clap positions. For each clap the position estimates have been ranked by their SRP-PHAT value. The best compromise parameters among all 4 claps have been chosen as the final calibrated configuration and have been used for all further computations.

The optimal parameters have been found as: azimuth offset of the blue array $\varphi_{offset:blue} = -9°$, azimuth offset of the red array $\varphi_{offset:red} = -6°$, distance between both arrays (inner spacing) $d = 1.45m$, which means $2.00m$ distance between both array centers.

The sketch from before has been updated with the acquired parameters and the expected microphone locations in figure 7.4. The negative sign of the azimuth offset in the figure has been dropped, by turning the array into the direction opposite to the default convention. The dotted lines in figure 7.4 should indicate the slanted rotational alignment of both arrays.

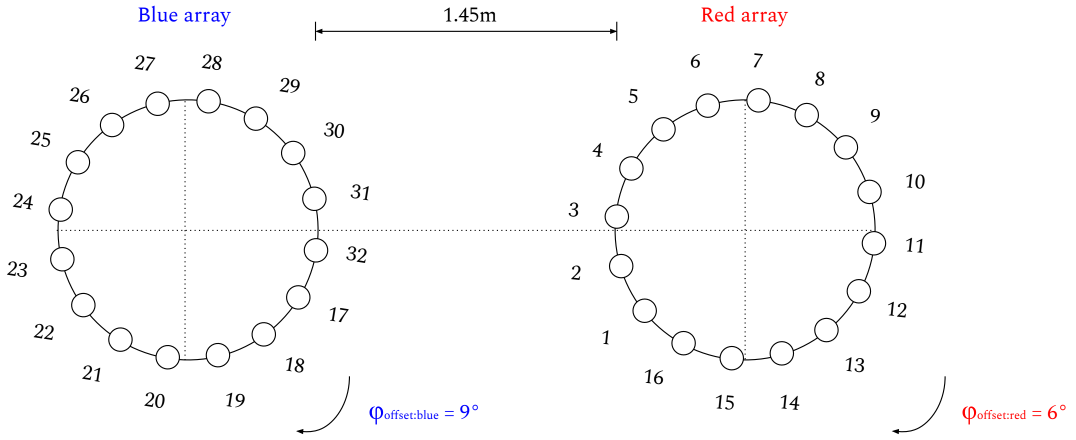
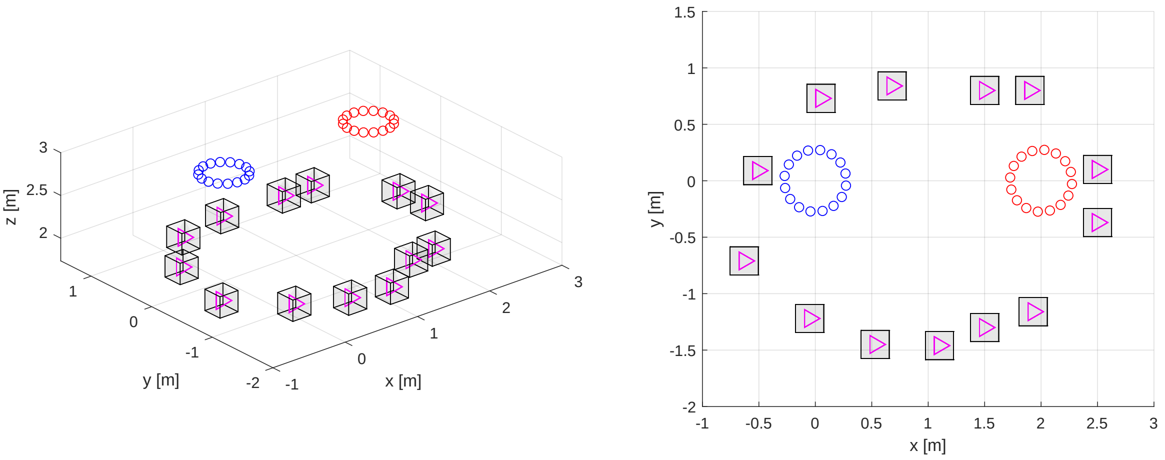


*Figure 7.4: Array with calibration*

## 7.3 Array Steering Improvement

The SRP-PHAT algorithm has not only been used to perform the post-calibration for the microphone array positions. It also turned out that the steering vector robustness and therefore also the separated signal quality can be drastically improved, if the speakers are tracked permanently using the SRP-PHAT algorithm. Figure 7.5 illustrates the estimated directions of the target speakers.



*(a) Isometric view*

*(b) Top view*

*Figure 7.5: Estimated speaker directions*

The pink triangles in figure 7.5 show the estimated, average direction of each target speaker. Those position estimates have been semi-manually found by observing the SRP-PHAT positions for a few minutes of audio material. Some of the speakers were not able to be located at first, because they were not present at the dinner table at the observation time. But the picture could be completed by looking at the other positions, that already have been found in advance. The grey constraint boxes around the estimated positions mark the SRP-PHAT search space, where the candidate points for each target are generated. In other words this means that the arrays are always steered within the boundaries of the grey boxes, independent if that specific speaker's mouth is actually located within the box or the target speaker is talking at all. The algorithm has not been extended by a voice detection mechanism and therefore does not make a difference if a speaker is talking inside a grey box or not. Keep in mind, that the grey boxes do not represent the actual positions of each speaker, but instead the constraint for the target directions. It also has been noticed, that the the direction estimates on the lower side of the table are arranged in a more dispersed way compared to other speakers. A possible explanation for causing a different localization behaviour could be that the persons are located at a bigger distance or also interfering signals might be a factor. Nevertheless in listening tests it has been concluded, that the given SRP-PHAT direction estimates provided a good target signal quality. Further the algorithm is not trained on specific talkers, which means the algorithm also does not differentiate if a speaker is possibly located at another seat or somewhere else. Simply put, each grey box represents one audio-track that is computed for each speaker in the end.

## 7.4 Separation Algorithm Choice Discussion

In section 2.7.3 it has already been discussed how the spherical wave propagation model will be combined for the dual array configuration. Based on the *'Combined array'* approach the steering vectors will be computed for the system illustrated in figure 7.6.

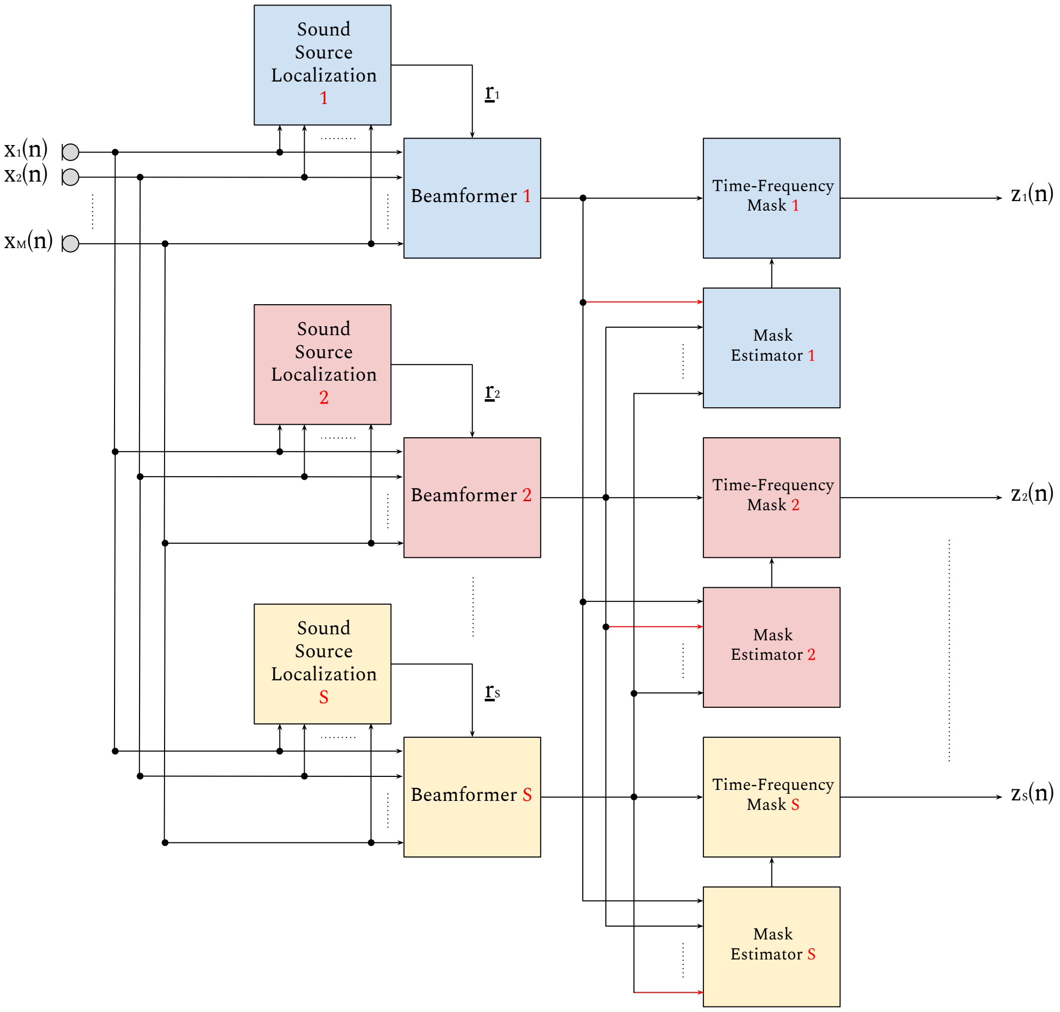


*Figure 7.6: The final system consists of SRP-PHAT, GSC and IRM algorithms*

The knowledge achieved from the evaluation measurements in chapter 6 can be roughly transferred to the real-world data case with one big exception: The MVDR and the SD-BF do not perform as good as the simulation results may imply. The sound quality of those beamformers combined with time-frequency masking is inferior compared to the GSC and even the D&S-BF (both combined with TFM of course) for the real-world data case.

The MVDR beamformers have shown to successfully reject interfering signals, but besides also introduce artifacts, which makes them perform worse in combination with time-frequency masking. The SD-BF shows nice suppression of low frequency interference signals, but besides also produces inferior output combined with time-frequency masking.

One possible explanation is, as those beamformers depend on signal statistics (MVDR noise correlation matrix) or position data (SD-BF noise field coherence matrix), they are more vulnerable to errors for those parameters. Although the calibration and also the array steering has been optimized with the SRP-PHAT algorithm, the used parameters possibly do not really match the true positions.

Another conclusion is, that the time-frequency masking based on beamformed signals with intense tone coloration or affliction with artifacts produces inferior TFM-outputs, whereas time-frequency masking based on beamformed signals with neutral tone characteristics and moderate introduction of artifacts produces superior TFM-outputs.

However the adaptive approach of the GSC clearly outperforms the sound quality given by the D&S-BF, SD-BF and the MVDR beamformers in the real-world data task at 24kHz and therefore has been finally chosen to process the full dataset of 30 hours audio material. The complete system that was used in the end is depicted in figure 7.6.

## 7.5 Implementation Details

In this section the implementation details of the algorithm are described, that have been finally used for the separation task at the Concordia research station. The descriptions refer to the block diagram in figure 7.6. The parameters have been found empirically by means of objective measures or subjective listening tests.

The 32-channel audio has been provided at 48kHz sampling rate. Due to memory overruns of the local workstations at the SPSC laboratory, the sampling rate of all audio data has been decreased. 16kHz was considered at first, which also implicated signal degradation at higher frequencies. Therefore 24kHz was the preferred sampling rate. The microphone signals are segmented in an overlap-and-add block processing framework with a window size of 4096 samples, Hann window function and an overlap factor of 50%.

Due to the block processing technique the SRP-PHAT computes position updates for the targets approximately every 85 milliseconds. The coordinate estimates obtained by the SRP-PHAT are typically very jumpy during non-speech sequences and stabilize when the target speaker is talking. As the beamformed signals interact with each other in the time-frequency masking, the jumpiness of the acoustic tracking had to be reduced. Therefore the coordinate estimates have been smoothed by a small moving average buffer with only 4 stored, previous positions. This buffer size reduced the jumpiness of the estimated positions sufficiently during sequences, where no target-speech is active. Bigger buffers cause the tracking speed to slow down for sequences where the target is active and therefore also causes the audio quality of the target signal to suffer.

In figure 7.5 the search cubes for each speaker already have been illustrated. Each of these cubes has been specified with a side length of 25cm, which ensures that neighbouring cubes do not overlap.

The coarse-to-fine-region-contraction algorithm then computes 125 equally distributed candidate points inside the initial search cube and evaluates the SRP-PHAT for those positions. The 5 best candidates will be kept and enclosed by a new search cube for the next iteration. The iterative shrinking will be quit, if the new search cube is smaller than $1 * 10^{-4} m^3$ (e.g. a cube with 5 cm side length) and the position with the maximum SRP-PHAT will be returned.

The parameters of the beamformers evaluated in chapters 6 and 7 are listed in this paragraph. Both the MPDR and the MVDR beamformer have been implemented with diagonal loading (MPDR-DL and MVDR-DL) at a diagonal loading level $\gamma = 1 \cdot 10^{-3}$. The second loading technique that has been implemented with the MPDR and MVDR beamformers was the variable loading (MPDR-VL and MVDR-VL) with variable loading level $\delta = 1 \cdot 10^{-3}$. The PSD smoothing factor was chosen as $\lambda = 0.95$. Beyond that also the FFT block size has to be fine-tuned for the MVDR beamformers in dependence of the TDOAs, e.g. 25 milliseconds has been found as a good value. The superdirective MVDR beamformers (SD-BF) also have been implemented with diagonal loading at a diagonal loading level $\gamma = 1 \cdot 10^{-3}$. As the diffuse noise field coherence matrix of the SD-BF does not vary (the microphone positions do not change over time), the variable loading approach would not make sense for the SD-BF. The parameters of the Generalized Sidelobe Canceller (GSC) were forgetting factor $\beta = 0.9$, step-size $\mu = 0.01$ and filter length $N = 80$. These parameters provide stability according to [15] and also cause subtle intelligibility benefits compared to the D&S-BF (especially at 24kHz sampling rate or higher) and is therefore preferable to time-frequency masking. From the experiments it has been found that bigger step-sizes cause the speech signals to distort and pump, whereas smaller forgetting factors tend do produce boomy sounds.

The time-relations (TDOA, reverberation) between the beamformed signals have a big influence on the time-frequency masking, therefore also the parameters have to be fitted to the considered problem. It has been found that an increased overlapping factor of 75% or above (for the block processing of the time frequency masking) produces smoother output and is therefore preferable. Further smoothing can be already achieved if zero-padding is used at the FFT size is doubled. The separation parameter that has been finally used with the ideal ratio mask (IRM) is $\beta = 3.0$, if not declared otherwise for some experiments in section 6.3.3 and 6.4.6.

## 7.6 Visualizations of the Separated Real-World Data

Besides subjective listening tests, objective assessment of the processed, real-world data is difficult, which justifies the experiments in chapter 6. In this section audio snippets of the computed samples will be plotted in time-domain and time-frequency-domain. In figure 7.7 the enumeration associated with the targets around the dinner table is listed.

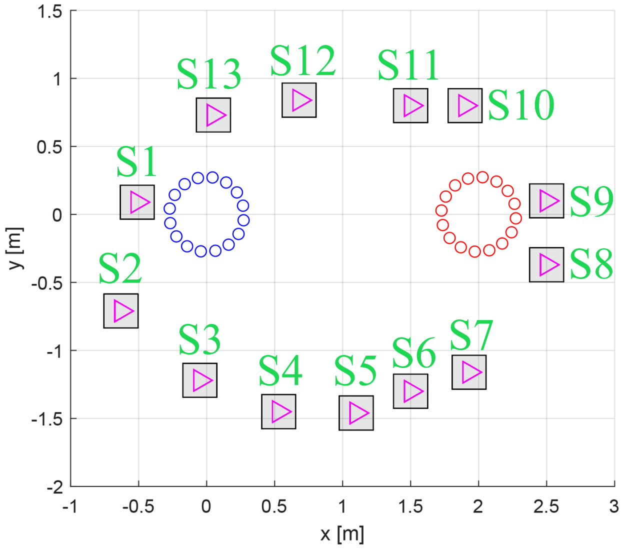
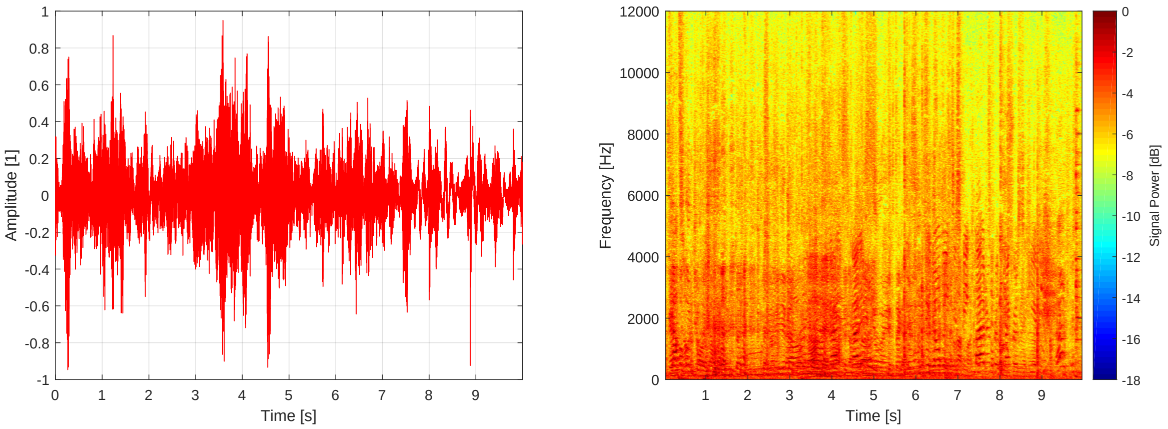


*Figure 7.7: Speaker enumeration*

From figure 7.8 to 7.11 the speech separation is shown step by step in a 10-second-long recording. At that moment all speakers except for S8, S9, S10 and S11 are present at the dinner table. Around the timestamp at 4 seconds to 9 seconds the desired target in this example (S2) is making a statement, while the interfering speech signals are at very high amplitudes.



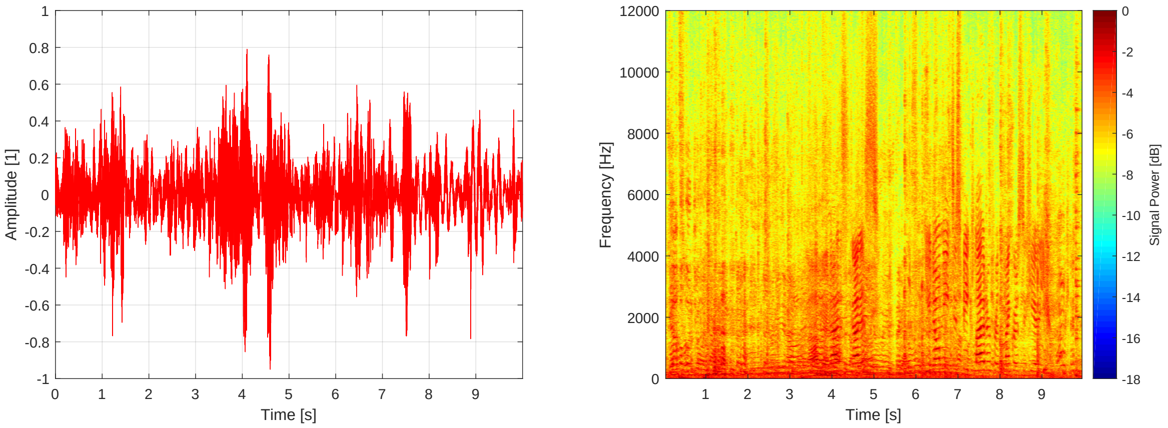
*(a) Time-domain waveform*

*(b) Spectrogram*

*Figure 7.8: Closest microphone signal*

By just listening at the closest microphone signal (depicted in figure 7.8) it is hard to tell which one of the speakers is the target source, because of the high interfering speech signal levels (especially for the nearby seatmates of the target S2).

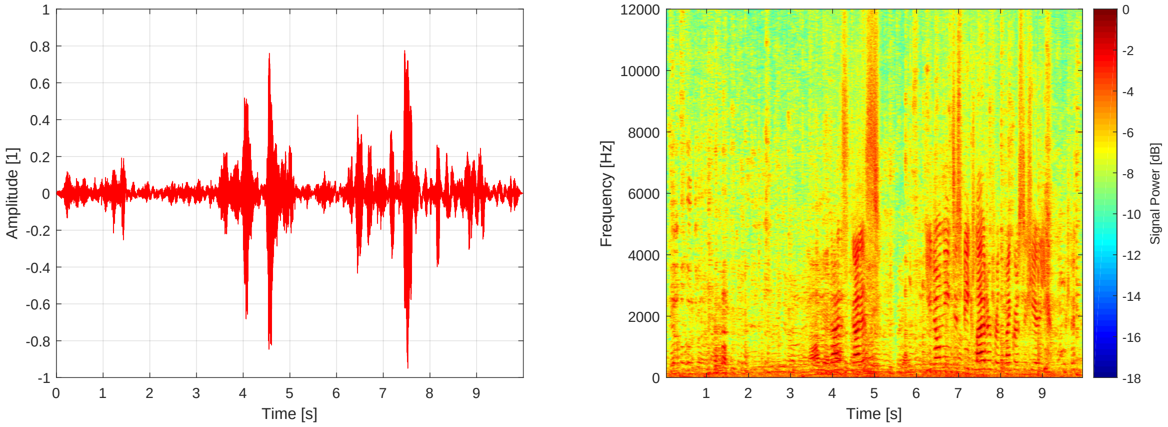
The output of the GSC (depicted in figure 7.9) shows the capability of using several microphones. The signal of the target S2 starts to stick out of the mixture.



*(a) Time-domain waveform*

*(b) Spectrogram*

*Figure 7.9: GSC output*



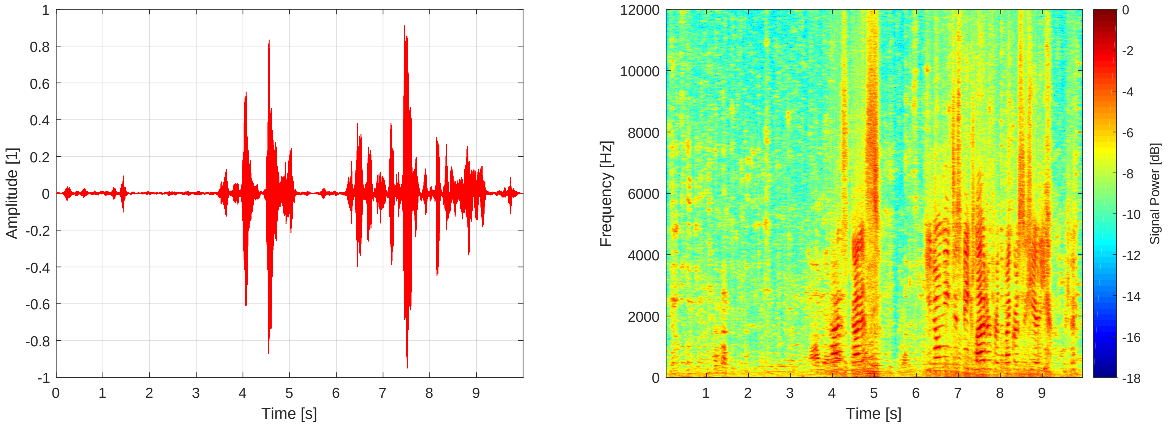
*(a) Time-domain waveform*

*(b) Spectrogram*

*Figure 7.10: GSC output and IRM with* $\beta = 3.0$

Figure 7.10 shows the output of the setup, that was finally used. All of the 13 GSC outputs have been combined with the IRM to produce a signal, where the target is still intelligibly recognizable and clearly separated from the interfering speech.

In figure 7.11 we go a step further to simply show, how the output is influenced, if a smaller separation parameter $\beta$ would be used. The interfering speech disappears in an indistinctive noise floor so that the target speech is emphasized even more. But we also can see the negative aspect of time-frequency masking, when we look at the high frequency content of the target speech, where the target signal suffers from intelligibility loss. As the computed speech data will be listened by linguists in the end, we have decided for $\beta = 3.0$ (depicted in figure 7.10), accepting possible higher interference in favour of a better target intelligibility.



*(a) Time-domain waveform*

*(b) Spectrogram*

*Figure 7.11: GSC output and IRM with* $\beta = 1.0$

# 8 Conclusions and Outlook

In this thesis several beamforming methods have been investigated and their advantages and drawbacks have been discussed. Further also the benefit of using a sound source localization algorithm has been experienced when the beamformers are used on a real-world dataset, not only for microphone post-calibration but also for making the steering vector estimation more robust. As a completing step the time-frequency masking has been pointed out as a powerful method to reduce interfering signals.

At the bottom line the separated speech audio, that has been computed with the system illustrated in figure 7.6, produces useful results and the improvement compared to using one single microphone has been shown. Each intermediate step to enhance the result is worthwhile.

In practice also the computational complexity of the considered algorithms comes into play. The computation period for the whole dataset (approximately 30 hours of audio) took around two and a half months, where six PC workstations simultaneously processed the whole dataset. Depending on the algorithm complexity the choice of the used algorithms can drastically influence the computation period. Of course this also has an impact on the obtained signal quality and should be therefore well-matched with the intended application (e.g. real-time implementation vs. offline processing).

The performance of beamforming algorithms improves significantly with prior knowledge. Concerning this project that desired prior knowledge was strictly limited in advance, because the access to the research station was not given and the scientists on site did not have the expertise concerning that topic to supply more information. Nevertheless a more accurate sketch with geometrical dimensions would have already helped a lot (e.g. to limit the parameter set for calibration in advance). As all those array processing algorithms depend on spatial positions, any kind of prior knowledge included is helpful to improve the results.

However array processing is an intensively researched field and there are a number of possibilities to extend the presented algorithms. There will always be room left for improvement, that is mostly limited by the available data, processing power and/or prior knowledge.

Some possible improvements that came up while conducting this thesis are specified below:

○ Putting more focus on array calibration and considering a bigger set of possible parameters could eliminate gain, phase and positioning errors and therefore improve the overall beamforming performance, especially for the SD-BF and MVDR-based beamformers.

○ Even though no true geometrical dimensions were given, the direction estimates of the SRP-PHAT algorithm have been trusted, because it has turned out to improve the separated signals. The SRP-PHAT could be reinvestigated in combination with room simulations to make a statement about its estimation errors. In such a simulation setup also other localization algorithms could be considered for comparison.

# Bibliography

[1] Wikipedia, "Wavefronts," https://de.wikipedia.org/wiki/Wellenfront, Last edit 23-09-2016, last access 15-12-2019.

[2] ——, "Concordia station," https://en.wikipedia.org/wiki/Concordia_Station, Last edit 31-10-2019, last access 15-12-2019.

[3] J. Benesty et al., *Microphone Array Signal Processing.* Springer, 2008, ch. Introduction.

[4] I. McCowan, "Microphone arrays: A tutorial," *Queensland University, Australia*, 2001.

[5] J. Benesty et al., *Microphone Array Signal Processing.* Springer, 2008, ch. Direction-of-Arrival and Time-Difference-of-Arrival Estimation.

[6] S. Stergiopoulos, *Advanced Signal Processing Handbook: Theory and Implementation for Radar, Sonar and Medical Imaging Real-Time Systems.* CRC press, 2001.

[7] J. Bitzer and K. U. Simmer, *Microphone Arrays.* Springer Science & Business, 2001, ch. Superdirective Microphone Arrays.

[8] J. Benesty et al., *Adaptive Signal Processing.* Springer, 2003, ch. Adaptive Beamforming for Audio Signal Acquisition.

[9] H. L. V. Trees, *Optimum Array Processing.* John Wiley & Sons, Inc., 2002, ch. Optimum Beamformers.

[10] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing 27(2)*, 1979.

[11] P. Welch, "The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms." *IEEE Transactions on audio and electroacoustics 15(2)*, 1967.

[12] H. L. V. Trees, *Optimum Array Processing.* John Wiley & Sons, Inc., 2002, ch. Mismatched MVDR and MPDR Beamformers.

[13] J. Gu and P. J. Wolfe, "Robust adaptive beamforming using variable loading," *Fourth IEEE Workshop on Sensor Array and Multichannel Processing*, 2006.

[14] L. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on antennas and propagation 30(1)*, 1982.

[15] J. P. Townsend and K. D. Donohue, "Stability analysis for the generalized sidelobe canceller," *IEEE Signal Processing Letters 17(6)*, 2010.

[16] Z. Wang et. al, "Oracle performance investigation of the ideal masks," *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016.

[17] J. P. Morgan, "Time-frequency masking performance for improved intelligibility with microphone arrays." Master's thesis, 2017.

[18] P. Mowlaee et. al., *Single Channel Phase-Aware Signal Processing in Speech Communication.* Wiley, 2017, ch. Phase Processing for Single-Channel Source Separation.

[19] J. H. DiBiase et al., *Microphone Arrays.* Springer, 2001, ch. Robust localization in reverberant rooms.

[20] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing 24(4)*, 1976.

[21] J. Benesty et al., *Acoustic MIMO Signal Processing.* Springer, 2006, ch. Time Delay Estimation and Acoustic Source Localization.

[22] H. T. H. Do, "Real-time SRP-PHAT source location implementations on a large-aperture microphone array," Ph.D. dissertation, Brown University, 2007.

[23] R. Lebarbenchon and E. Camberlein, "Multi-channel BSS locate, a toolbox for source localization in multi-channel convolutive audio mixtures," http://bass-db.gforge.inria.fr/bss_locate/, Last edit September 2019, last access 15-12-2019.

[24] C. Blandin et al., "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing 92(8)*, 2012.

[25] D. G. Manolakis et al., *Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing.* ARTECH HOUSE, INC., 2005, ch. 5.3.3 Power Spectrum Estimation by Averaging Multiple Periodograms - The Welch-Bartlett Method.

[26] J. Yamagishi, "English multi-speaker corpus for CSTR voice cloning toolkit (CSTR VCTK corpus)," https://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html, 2010, last access 15-12-2019.

[27] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," *International Conference on Latent Variable Analysis and Signal Separation. Springer, Berlin, Heidelberg*, 2012.

[28] V. Emiya et al, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing 19(7)*, 2011.

[29] R. Huber and B. Kollmeier, "PEMO-Q—a new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on audio, speech, and language processing 14(6)*, 2006.

[30] A. Plinge and G. A. Fink, "Geometry calibration of multiple microphone arrays in highly reverberant environments," *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

# A

# List of Abbreviations

| | | |
|---|---|---|
| ABF | ... | Adaptive Beamformer |
| ASR | ... | Automatic Speech Recognition |
| BM | ... | Blocking Matrix |
| CFRC | ... | Coarse-to-Fine Region Contraction |
| DOA | ... | Direction of Arrival |
| D&S-BF | ... | Delay-and-Sum Beamformer |
| FBF | ... | Fixed Beamformer |
| FFT | ... | Fast Fourier Transform |
| GCC | ... | Generalized Cross-Correlation |
| GJBF | ... | Griffiths Jim Beamformer |
| GSC | ... | Generalized Sidelobe Canceller |
| LMS | ... | Least Mean Squares |
| MPDR | ... | Minimum Power Distortionless Response |
| MVDR | ... | Minimum Variance Distortionless Response |
| NLMS | ... | Normalized Least Mean Squares |
| PEASS | ... | Perceptual Evaluation Methods for Audio Source Separation |
| PHAT | ... | Phase Transform |
| PSD | ... | Power Spectral Density |
| SD-BF | ... | Superdirective Beamformer |
| SNR | ... | Signal-to-Noise Ratio |
| SRC | ... | Stochastic Region Contraction |
| SRP | ... | Steered Response Power |
| SSL | ... | Sound Source Localization |
| STFT | ... | Short Time Fourier Transform |
| TDE | ... | Time Delay Estimation |
| TDOA | ... | Time Difference of Arrival |
| TOA | ... | Time of Arrival |

# B
# List of Symbols

Below the symbols are specified that are valid throughout the whole thesis. In cases where definition conflicts would occur (e.g. cartesian coordinates $x, y, z$ vs. block diagram signals $x, y, z$), the symbols are defined locally at each section.

| | | |
|---|---|---|
| $APS$ | ... | Artifacts-related perceptual score (PEASS) |
| $c$ | ... | Speed of sound |
| $\underline{d}$ | ... | Steering vector |
| $fs$ | ... | Sampling rate |
| $\mathbf{I}$ | ... | Identity matrix |
| $IPS$ | ... | Interference-related perceptual score (PEASS) |
| $\underline{k}$ | ... | Wavenumber vector |
| $OPS$ | ... | Overall perceptual score (PEASS) |
| $r_{xx}$ | ... | Auto-correlation function |
| $r_{xy}$ | ... | Cross-correlation function |
| $\mathbf{R_{vv}}$ | ... | Noise signal correlation matrix |
| $\mathbf{R_{xx}}$ | ... | Input signal correlation matrix |
| $SNR$ | ... | Signal-to-noise ratio |
| $TPS$ | ... | Target-related perceptual score (PEASS) |
| $\underline{w}$ | ... | Beamforming weights |
| $\mathbf{\Gamma_{VV}}$ | ... | Noise field coherence matrix |
| $\mathbf{\Phi_{VV}}$ | ... | Noise signal PSD correlation matrix |
| $\mathbf{\Phi_{XX}}$ | ... | Input signal PSD correlation matrix |
| $\varphi$ | ... | Azimuth angle |
| $\Psi$ | ... | PHAT weighting function |
| $\theta$ | ... | Elevation angle |
| $\tau$ | ... | Time-shift variable |
| $\omega$ | ... | Radial frequency |