Low complexity crosstalk cancellation algorithm for consumer audio systems

Optimizing crosstalk cancellation from a human sound perception perspective

Richard Eveleens

TU Delft

kien

Low complexity crosstalk cancellation algorithm for consumer audio systems

Optimizing crosstalk cancellation from a human sound perception perspective

Thesis report

by

Richard Eveleens

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on June 13, 2023 at 11:00

Thesis committee:	
Chair:	Dr. O.C. Scharenborg
Supervisors:	Dr. J. Martinez
	Dr. ir. A. Noroozi
External examiner:	Dr. R.T. Rajan
Place:	EEMCS, Delft
Project Duration:	September, 2022 - June, 2023
Student number:	4665570

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

EEMCS · Delft University of Technology

Abstract

Over the past decade, spatial audio awareness evolved into an in-demand feature in audio entertainment. The addition of sound source locations to, for instance, movies or music adds a level of auditory envelopment and spatial awareness to the audio experience. Expensive setups present in, for instance, cinema's, are able to create this envelopment by means of a large set of loudspeakers with which the desired sound fields are created. Creating this spatial envelopment in practical consumer living rooms or home cinema setups proves to be a more challenging task due to the impractical amount of loudspeakers required. To create the audio envelopment with a small amount of loudspeakers, crosstalk cancellation can be used. Crosstalk cancellation as posed in literature is, however, not robust enough to be used in practical appliances. The main cause of these bad characteristics is the objective cost function it optimizes which results in an ill-posed problem. In this thesis, the crosstalk cancellation problem is relaxed by aiming for perceptually sufficient results instead of aiming for objectively optimal results. The human auditory system has its limitations in both the perception of audio and localization of audio sources. Exploiting the limitations of the auditory system generates mathematical freedom that can be used to construct a more robust and stable crosstalk cancellation algorithm. This thesis provides the first steps towards the incorporation of audio perception in the domain of crosstalk cancellation and audio envelopment with a limited amount of loudspeakers.

Acknowledgements

The first thoughts and ideas leading to this thesis started to take shape in February 2022. Some preliminary ideas on audio processing with Jorge Martinez eventually lead to a collaboration with Kien and, specifically, Arash Noroozi. We all have an interest in crystal clear and immersive audio experiences and the subject of the thesis represents this. Kien offered a work place in their cozy office in Rotterdam and I worked there four days per week. One day per week I would go to the university to have discussions with Jorge and to eat a pile of cookies he provided to keep us happy. I am using ”us” here since I was not alone throughout this period. All day and every day my good friend and colleague Dimme de Groot was working beside me. Dimme and I had a similar thesis subject and so we worked together most of the time and had numerous discussions to help each other solve our problems. Also, the submitted paper that can be found in Appendix A is primarily written together with Dimme. Discussions with Jorge, Arash and Dimme helped me solve issues, come up with new ideas and eventually finish this thesis with a nice result.

I want to thank the Kien team for the amazing time, the unconditional support and immeasurable amount of laughs (and games of pool or table tennis). Special thanks to Arash for offering this opportunity and always listening to my (sometimes bizarre) ideas. Also special thanks to Nick for helping me set up and use linux and just being an amazing guy.

I want to thank Jorge for being the best supervisor imaginable. Besides being very knowledgeable on the topic, he is a caring and fun guy and I am going to miss the weekly meetings. The only thing I blame him for is the weight I gained passed year due to all the snacks he brought to the meetings.

As already mentioned above, Dimme and I were barely separated last year and I want to thank him for the amazing time. Even though the amount of messages I received on a daily basis at least doubled because of his limitless spamming with equations and figures, the amount of laughter also doubled.

Of course, I want to thank my parents for giving me the opportunity to follow this study and always supporting me when needed. Above all, I want to thank Tal for making my life a dream come true. Because of her, even if I would fail this thesis, I would still be the happiest man alive.

Anyway, without further ado, enjoy reading the thesis and try not to fall asleep.

Contents

Mathematical notation

The following notation is used in the thesis.

$\hat{x}$	frequency domain
$\odot$	pointwise multiplication
$*$	convolution
$\mathbf{x}$	column vector
$\mathbf{X}$	two or higher dim. tensor
$\mathbf{I}$	identity matrix
$\mathbf{F}$	flip operator
$\mathbf{W}$	DFT matrix
$\mathbf{W}^{-1}$	IDFT matrix
$\mathbf{X}^{-1}$	matrix inverse
$\mathbf{X}^{\mathrm{T}}$	transpose
$\mathbf{X}^{\mathrm{H}}$	Hermitian transpose
$\mathbf{X}^{\dagger}$	matrix pseudoinverse
x^{*}	desired / optimal solution
$\|\|\mathbf{x}\|\|_2$	L_2-norm
$\mathbb{R}$	Set of real numbers
$\mathbb{C}$	Set of complex numbers
!	Factorial
!!	Double factorial
$\delta(x)$	Kronecker delta function
$\mathrm{atan2}(x, y)$	2-argument arctangent
$\mathrm{mod}(x, c)$	modulo function
$\mathrm{diag}(\mathbf{x})$	matrix with $\mathbf{x}$ diagonal

The following definitions are used in the thesis:
General definitions:

- f_s: Sample frequency.
- $f_{\mathrm{opt}}()$: Optimization function.
- v_s: The speed of sound.
- w_h: The width of the head.
- $\mathbf{l}_s$: The 3D location of the source.
- $\mathbf{l}_l$: The 3D location of the listener.
- d_L: Distance from source to left ear.
- d_R: Distance from source to right ear.
- p_s: Pressure level of a stimulus.
- p_{SPL}: Pressure level expressed in SPL.
- τ_{60}: The time it takes for the amplitude of the reflections to drop 60 dB.
- $\boldsymbol{\beta}$: Reflection coefficients of the walls in a rectangular room.
- θ_{out} The outgoing (originating from a loudspeaker) azimuth angle.

- ϕ_{out}: The outgoing (originating from a loudspeaker) elevation angle.
- θ_{in}: The incoming (arriving at the head of the listener) azimuth angle.
- ϕ_{in}: The incoming (arriving at the head of the listener) elevation angle.
- $\boldsymbol{\alpha}$: Intermediate or optimization variable.
- e: L_2-norm of the difference between obtained and desired response.
- d_{IACC}: Difference between obtained and desired τ_{IACC}.
- d_θ: Difference between perceived and desired azimuth angle estimation.
- PEAQ(x^*, x): The Perceptual Evaluation of Audio Quality measure given signal x and reference signal x^*

Signal and response definitions:

- $\mathbf{y}_{\text{L}}$: The measured signal deep in the ear canals for the left ear.
- $\mathbf{y}_{\text{R}}$: The measured signal deep in the ear canals for the right ear.
- $\mathbf{c}_{\text{L}}$: The channel composed of all the responses and signal alterations from the input signal fed to the loudspeaker, up until the measurement by the auditory system for the left ear.
- $\mathbf{c}_{\text{R}}$: The channel composed of all the responses and signal alterations from the input signal fed to the loudspeaker, up until the measurement by the auditory system for the right ear.
- $\mathbf{s}$: The input signal fed to a loudspeaker.
- $\mathbf{h}$: The response of a loudspeaker.
- $\mathbf{d}$: The directivity of a loudspeaker.
- $\mathbf{r}$: The response of a single reflection.
- $\mathbf{v}_{\text{L}}$: The HRTF response for the left ear.
- $\mathbf{v}_{\text{R}}$: The HRTF response for the right ear.
- $\mathbf{m}$: The masking signal response.
- $\boldsymbol{\epsilon}$: The distortion response.
- $\mathbf{g}$: The masking curve.
- $\mathbf{u}$: Window response.

Auditory system definition

- $\mathbf{y}_{\text{IACC}}$: The IACC response of vectors $\mathbf{y}_{\text{L}}$ and $\mathbf{y}_{\text{R}}$.
- τ_{IACC}: The time delay corresponding to the peak value of $\mathbf{y}_{\text{IACC}}$.
- T_q: The threshold in quiet.
- γ: Gamma-tone filter.
- p_γ: Power response of gamma-tone filter.
- f_γ: Gamma-tone filter bank centre frequencies.
- η_γ: Gamma-tone filter order.

- κ_γ: Gamma-tone filter constant.
- $\mathbf{h}_{\mathrm{OM}}$: The outer- and middle-ear filter.
- m_p: Masker power.
- ϵ_p: Distortion power.
- ξ: Detectability.
- c_1 & c_2: Calibration constants for detectability.

Optimization variables:

- $\boldsymbol{\mu}_{\mathrm{IACC}}$: The upper limit of the IACC response.
- $\boldsymbol{\alpha}_L$: Optimization response at left ear.
- $\boldsymbol{\alpha}_R$: Optimization response at right ear.
- $\boldsymbol{\alpha}_{\mathrm{IACC}}$: Optimization IACC response.
- $\boldsymbol{\alpha}_{\mathrm{IACC,L}}$: Left optimization IACC response.
- $\boldsymbol{\alpha}_{\mathrm{IACC,R}}$: Right optimization IACC response.

Lengths and iteration variables:

- N_b: Number of bins (samples) in a response, with n_b as iteration variable.
- N_s: Number of loudspeakers, with n_s as iteration variable.
- N_r: Number of reflections, with n_r as iteration variable.
- N_γ: Number of filters in gamma-tone filter bank, with n_γ as iteration variable.
- N_c: Number of channel blocks, with n_c as iteration variable.
- N_i: Number of input signal blocks, with n_i as iteration variable.
- N_y: Number of in-ear response blocks, with n_y as iteration variable.
- N_l: Number of listener locations, with n_l as iteration variable.

Nomenclature

List of Abbreviations

BW_c	Critical Bandwidth
BRIR	Binaural Room Impulse Response
DFT	Discrete Fourier Transform
ERB	Equivalent Rectangular Bandwidth
HRTF	Head Related Transfer Function
IACC	InterAural CrossCorrelation
IDFT	Inverse Discrete Fourier Transform
ILD	Interaural Level Difference
ITD	Interaural Time Difference
JND	Just Noticeable Difference
MDF	Multi-Delay Filter
PEAQ	Perceptual Evaluation of Audio Quality
QP	Quadratic Problem
RIR	Room Impulse Response
SPL	Sound Pressure Level
SRTF	(Loud)Speaker Related Transfer Function
SV	Singular Value
SVD	Singular Value Decomposition
VR	Virtual Reality

Constants

$p_0 = 20\ \mu$Pa	SPL reference value

List of Figures

1

Spatial audio and virtual sources

Recreating audio experiences in regular house holds has been a topic of research for decades. Nowadays, an audio device can be found in nearly every house. Filling a room with sound is usually done with (a set of) loudspeakers and there are a lot of options to chose from. Systems range from one small smart loudspeaker to extensive systems including multiple loudspeakers, subwoofers and sometimes even ceiling mounted loudspeakers. All these systems are meant to give the user a certain experience of audio ranging from movies to (live) music. Especially in movies, but also in music, sound sources have a certain direction or location the sound should come from. For instance, a car that drives at the right side of the television screen, should be heard from the right side as well and when it moves, the location of the sound source should also move accordingly.

Creating and delivering the spatial information of the sound source correctly proves to be a challenging task. A famous solution to this problem is provided by Dolby Digital in the form of the 5.1 audio codec for home cinema setups [1–3]. This codec provides six distinct responses meant for six distinct loudspeakers that are placed around the listener in a prescribed setup. Using this setup, the listener experiences a 360° audio experience and thus receives the spatial information corresponding to, for instance, a driving car.

The major downside to these kind of solutions is that a very specific setup is required to obtain optimal results (see Figure 1.1 for the Dolby Digital 5.1 example). General living rooms or even home cinema rooms do not allow for such a setup because the loudspeakers do not fit in the ideal positions or because of aesthetic reasons. The audio coming from these off-positioned loudspeakers results in a wrong audio perceptual experience. A non-equidistant setup can be corrected using volume adjustment but correcting the difference in angle is more challenging.

1.1. Creating virtual audio sources

The aforementioned audio codec is designed with the purpose of delivering an audio experience to one person. We can redefine this goal into: delivering the desired response to two points in space, the location of both ears. In this case the desired response is the signal measured by our auditory system. This measured signal should be perceptually equivalent to the signal that would have ended up in the ear canals if the simulated source were real. Creating this desired response at both

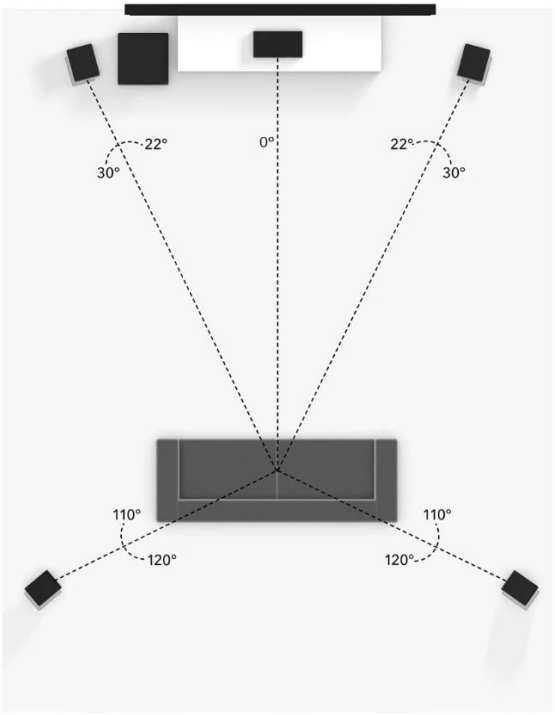


Figure 1.1: Dolby Digital ideal setup for the 5.1 codec. Figure adopted from [4].

ears, given a small amount of loudspeakers placed in a room with no placement limitations, is the general goal of this thesis.

Note that some figures in this thesis make use of an illustrative head sketch, the different views of this head sketch are depicted in Figure 1.2.

An illustrative example of the goal is shown in Figure 1.3. In Figure 1.3a we can see a living room with plants, couches and other furniture and decoration. The shape of the room and the furniture in the room force a non-ideal loudspeaker setup. Given this setup we wish to create a virtual audio source as depicted in Figure 1.3b, where the green arrows indicate that we compute the response at both left and right ear coming from the virtual source.

By defining an audio source and its location, the desired response found at both ears can be calculated. Since the channel from the physical loudspeakers to both ears is known, we can derive the output of these loudspeakers such that the desired response is actually perceived.

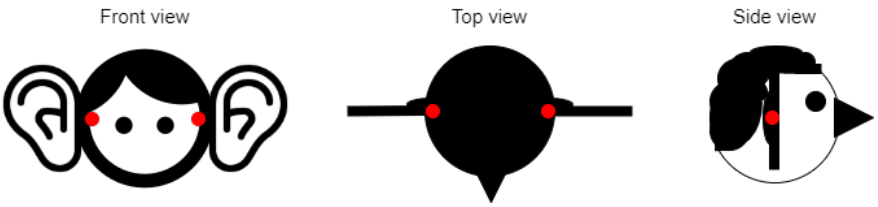
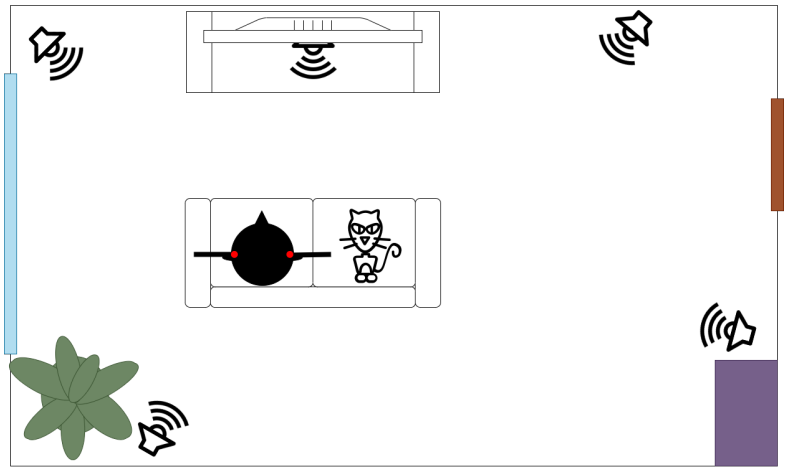
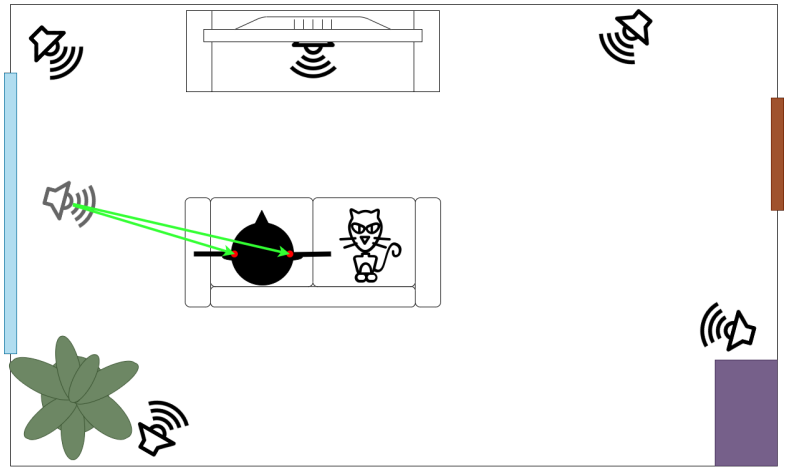


Figure 1.2: Different views of used head sketch.



(a) A room with a non-ideal 5.1 loudspeaker setup.



(b) The created virtual source using the non-ideal setup.

Figure 1.3: Due to the shape of the room and due to the furniture and decoration inside the room, the loudspeakers are placed in a non-ideal setup when compared to the ideal setup in Figure 1.1. Using this setup, we wish to create the virtual source (depicted by the shaded loudspeaker on the left) by creating the response that would be caused by the virtual source at both ears (denoted by the green arrows). The physical loudspeakers can be used to create this response.

1.2. Applications for virtual audio sources

The ability to create virtual audio sources can prove useful in many different applications besides movie and music entertainment, we discuss a few possibilities.

Noise attenuation: In nearly all situations in everyday life there is at least some sort of environmental noise present. Attenuating these noise sources is sometimes desirable for protection of hearing or simply comfort. A situation including a possible noise attenuating system using a virtual source is presented in Figure 1.4

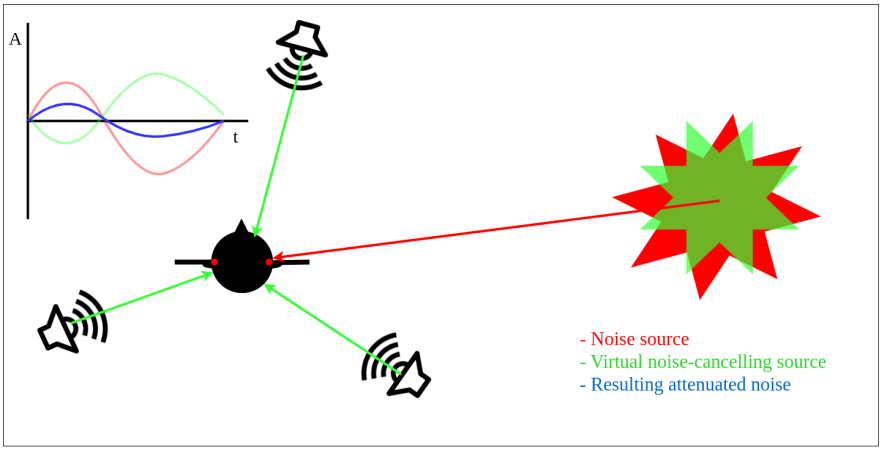


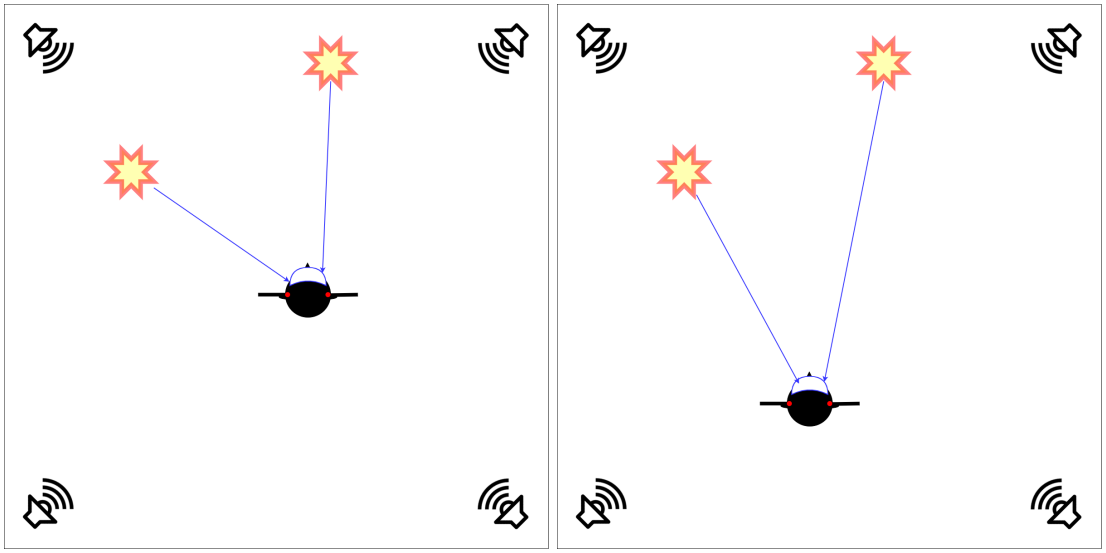
Figure 1.4: An example of virtual sources in noise attenuation. The red noise source should be attenuated. By using the three physical loudspeakers to create a virtual source (the green source), which approximates the inverse of the noise source, the noise can be attenuated. The graph on the top left gives a rough indication of the functioning of noise attenuation, with the blue plot the perceived response.

Virtual Reality experience: When playing a game in Virtual Reality (VR), the audio greatly contributes to the perception. Normally the audio is presented by headphones on the VR-glasses. The audio experience and comfort could be increased if loudspeakers delivered the sound. To be able to do this, the sound and its location corresponding to events that happen in the virtual world should be delivered to the ears correctly. When the player is moving or rotating, the response found at both the ears should move or rotate accordingly. An example of this is given in Figure 1.5 where the player moves but events happening in the virtual world remain at the same location.

1.3. Definitions and variables in the thesis

Before we can continue to the problem definition we must first establish a few Definitions and variables that are used throughout the thesis. Other definitions are discussed throughout the thesis:

- **Room Impulse Response (RIR)**: The Room Impulse Response describes the amplitude, delay and response of all the reflections in a room originating from a sound source being received at a single point in space.
- **Head Related Transfer Function (HRTF)**: The Head Related Transfer Function describes the way our body and ears alter an incoming sound wave before our auditory system measures the sound wave. These alterations allow us to localize sound sources in space. More on this in Chapter 2.



(a) Two virtual events happen in the room. The arrows show how these should be presented to the listener.

(b) The two virtual events still happen at the same location but the perceived response changes as shown by the arrows.

Figure 1.5: Virtual sources in a virtual reality experience.

- **Loudspeaker**: In the thesis the term loudspeaker is mostly used to refer to a loudspeaker but it could also be generalized to any audio source.
- **Listener**: The listener is the human who receives the audio. The listener has two measuring points, the two ears, and the measured audio is altered according to the (listener dependent) Head Related Transfer Function.
- **Physical loudspeakers**: The physical loudspeakers are the set of loudspeakers that are actually present in the room.
- **Virtual loudspeaker**: The virtual loudspeaker is the sound source present at a certain location we want to recreate by tuning the response of the physical loudspeakers. The virtual loudspeaker is thus not actually present but the response that would have come from it is recreated.
- **Sweet spot**: The sweet spot is the region in space around the listeners' head in which the illusion of the virtual loudspeaker is sufficiently recreated. The illusion of the virtual loudspeaker is not sufficiently recreated outside of the sweet spot region.

The definitions are followed by some variables and their notation:

- **Channel**: The channel is composed of all the responses and signal alterations from the input signal fed to the loudspeaker up until the measurement of the audio by the human auditory system. The composition of the channel is further discussed in Chapter 4. The channel is denoted by $\mathbf{c}_{\mathrm{L}}$ and $\mathbf{c}_{\mathrm{R}}$ for left and right ear respectively.
- **Input signal**: The input signal is the signal fed to the loudspeaker and is denoted by $\mathbf{s}$.
- **Received signal**: The received signal is the signal measured by the human auditory system, it is denoted by $\mathbf{y}_{\mathrm{L}}$ and $\mathbf{y}_{\mathrm{R}}$ for the left and right ear respectively.

1.4. Research question and assumptions

With the objective illustrated before in mind, we can construct a proper problem definition. In this thesis we focus mainly on creating one single virtual source given a set of loudspeakers. Here we assume that extending the creation of one virtual source to multiple virtual sources is possible with the required future research. To create the virtual source we use four loudspeakers based on the 5.1 loudspeaker. We omit the centre speaker and subwoofer since they are generally not used for delivering spatial audio cues [5]. The resulting setup consists of a good amount of speakers to create spatial audio without the need to place an over the top amount of loudspeakers in the room. This leads to the following problem definition:

Research Question. *Given a set of four physical loudspeakers in a room, is it possible to create the illusion of a virtual audio source for one listener in the room?*

While solving this problem we make a set of assumptions listed below:

1. The 3D location of the loudspeakers is known.
2. The 3D location of the two ears and the direction of the head is known.
3. The loudspeakers play there content at the same time, so there is no delay between loudspeakers.
4. The loudspeakers' response and directivity is known.
5. The Head Related Transfer Function (HRTF) is known and is applicable to the listener.

Note that some of these assumptions are clarified and described more throughout the thesis. Later on we also consider how the system performs when these assumptions do not exactly hold, since this is where the real challenges lie.

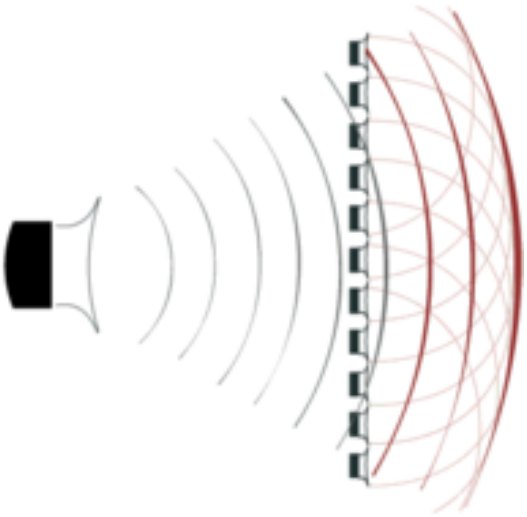
1.5. Possible solutions

In state-of-the-art literature, one can find two possible approaches that try to address to the problem presented in the problem definition: wavefield synthesis and crosstalk cancellation. The performance and possible use of these approaches serves as a base for the newly proposed algorithm in this thesis. The two approaches and their limitations are further discussed in the following sections.

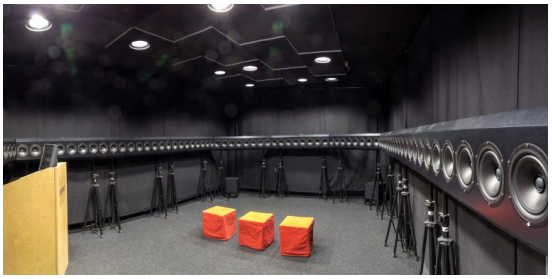
1.5.1. Wavefield synthesis

The principle of wavefield synthesis is creating a sound field in a room corresponding to a sound source which is present somewhere in space using an array of closely spaced loudspeakers [6]. By timing and tuning the output of the loudspeakers in a smart way, the waves coming from these loudspeakers combine into one wave which corresponds to the sound source you want to create. This principle is shown in Figure 1.6a. In the figure we can see that the wavefront that would have been emitted by the large source on the left can be artificially created using a large loudspeaker array. In this case the loudspeaker in the middle plays first and the further we go to the side of the array, the later the loudspeakers starts playing, which, after combining, creates the wavefront corresponding to the large loudspeaker.

A major advantage of wave field synthesis is that we are exactly creating the source we wish to simulate, this means that a large zone in the room is filled with the correct sound waves. Because of this, multiple people in the room can experience this source as it was intended.



(a) The loudspeaker array in the middle of the figure creates the same wave as the large loudspeaker on the left. We can thus create the illusion of the source on the left with the array of loudspeakers. In fact, any arbitrary source can be created with the loudspeaker array (up to some limitations). Figure adopted from [7].



(b) Example of a practical wavefield synthesis setup. This setup is created in the "Casa del Suono", an italian museum dedicated to audio, to show the power of wavefield synthesis. It shows that this involved setup is not practical for endconsumers due to the large amount of precisely placed loudspeakers. Figure Adopted from [8].

Figure 1.6: Wavefield synthesis, the principle (a) and a practical implementation (b).

The downside to this technique is the (large) amount of loudspeakers required to be able to perform wavefield synthesis. A sufficient setup would look like an array of loudspeakers next to each other with very little space between them, an example is shown in Figure 1.6b. Since this thesis centers around a solution achievable with only four loudspeakers and in a normal house-hold situation, wavefield synthesis is not an applicable solution to the problem.

1.5.2. Crosstalk cancellation

Originally, crosstalk cancellation was introduced to mimic the headphone experience on a pair of loudspeakers. The big advantage with headphones is that the audio played to the left ear cannot be heard by the right ear and vice versa. On top of that, there is no influence of the room on the audio quality with headphones. Crosstalk cancellation aims to bring this experience with the comfort of loudspeakers. Normally, with a stereo loudspeakers setup, the audio presented by the left loudspeaker is perceived by the left and right ear. Crosstalk cancellation aims at artificially removing the audio from the left loudspeaker to the right ear and vice versa, in short, cancelling the crosstalk. The crosstalk cancellation framework can be generalized to creating any desired response at the left and right ear using only two loudspeakers [9]. In Figure 1.7, the crosstalk cancellation setup is depicted where the green channels should be preserved and red channels should be cancelled.

In theory, crosstalk cancellation works but in practice only very specific scenarios with near-perfect prior knowledge on the setup can result in satisfying results [10, 11]. Crosstalk cancellation relies on the inversion of the channel response from the loudspeakers to the ears (more on this in Chapter 3). This operation yields a very unstable matrix inversion making a practical implementation nearly impossi-

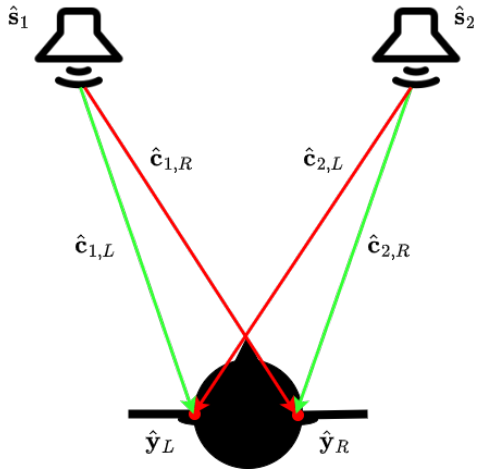


Figure 1.7: Crosstalk Cancellation setup where the green arrows depict the channels that should be preserved and the red arrows depict the crosstalk channels that should be cancelled.

ble. On top of that, crosstalk cancellation requires high accuracy prior knowledge on the room impulse response. As is discussed in Section 3.2, the presence of the room impulse response with high accuracy in a practical, furnitured and decorated room is an unrealistic assumption.

In the thesis we take the principles of crosstalk cancellation but we make an approximation of the optimization to obtain a more stable solution without sacrificing the perceptual target signals. The proposed solution is formulated in the following:

Proposed Solution. *The crosstalk cancellation algorithm described in literature uses an objective cost function (the L_2-norm). By optimizing for a subjective cost function that is related to the crosstalk cancellation problem, we hope to make the solution more stable. The subjective cost function is designed by considering psychoacoutstics. On top of this, a generalized room impulse response is introduced which further increases numerical stability, decreases computational complexity and makes the system less setup dependent without sacrificing on the audio perception.*

1.6. Thesis structure

This thesis is subdivided in several chapters that are connected as shown in Figure 1.8. The literature review on auditory localization is given in Sections 2.1 and the literature review on auditory perception is presented in Section 2.2. To complete the literature review, the crosstalk cancellation problem as posed in literature is discussed in Section 3.1. If the reader is already familiar with these topics, these sections can be skipped. Based on the findings in literature, the crosstalk cancellation problem is analysed in Section 3.2 and the system model is introduced in Chapter 4. Combining the findings results in the proposed algorithm presented in Chapter 5 and the results are analysed in Chapter 6. We finalize with the conclusion and future work in Chapters 7 and 8 respectively. In addition, a submitted paper is presented in Appendix A. This paper presents a stochastical room impulse response model which potentially improves the system model and proposed algorithm when it is incorporated.

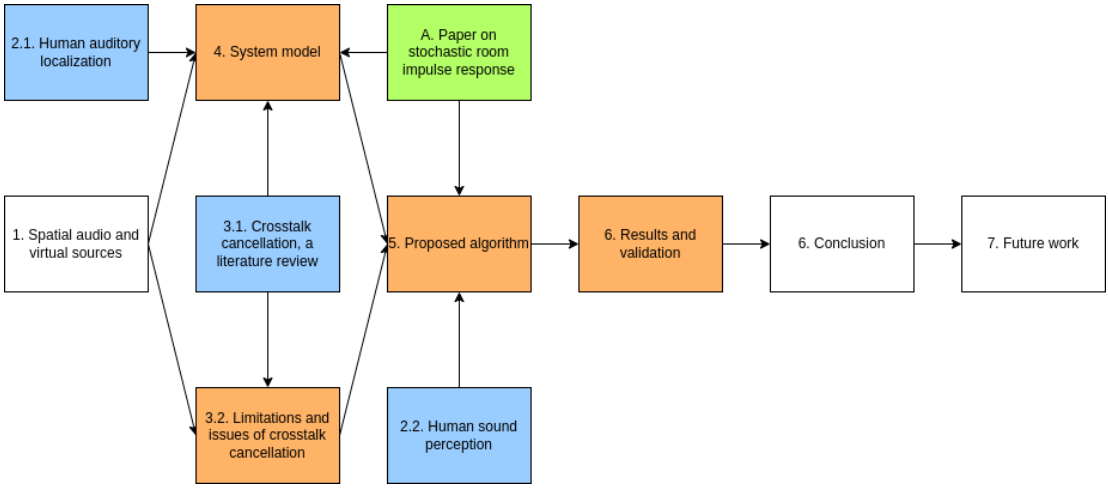


Figure 1.8: Outline of the thesis. The blue coloured sections present the literature reviews and are not required to read if the reader is familiar with the topics. The orange coloured chapters and sections are the ones that contain the main contribution of this thesis. The green coloured appendix presents a submitted paper that presents a stochastic room impulse model that could potentially improve the system model and proposed algorithm.

2

Human auditory localization and audio perception

To improve the crosstalk cancellation problem by means of an approximation of the problem, we aim to apply the perception of sound of human listeners. In this chapter we cover two aspects of the human auditory system that are of interest to us in form of a literature review. First the auditory localization system is discussed, in here we discuss how and how well human listeners can identify the location of a sound source. After this, we cover the threshold of hearing and the masking principles.

2.1. Human auditory localization

Humans are able to localize sound sources with great precision. By means of only two sensors, the two ears, we are able to estimate the horizontal angle (referred to as azimuth angle), the vertical angle (referred to as elevation angle) and also the distance to the sound source. To be able to estimate these parameters, we use a set of characteristics in the perceived sound. These characteristics are usually referred to as auditory cues. The combination of different auditory cues results in our ability to localize sound with surprising precision.

In the following section, the details of human auditory localization are discussed by means of a literature review. First, azimuth estimation is presented which is primarily focused on binaural cues. These are cues that use the response found at both ears and compares them to obtain the required information. Next, the elevation estimation is discussed which primarily uses monaural cues, cues that only use the response found at a single ear. The estimation of the third parameter in a polar coordinate system, namely distance, is discussed in Appendix B.1 since these findings are not directly applied in the thesis. A few interesting but specific finds are presented in Appendix B.2, these are also not directly applied in the thesis.

Please note that most of these findings are based on experiments with human subjects. In my opinion, some sources present conflicting conclusions. This means that they should be treated with caution as is mentioned throughout this chapter.

Before diving into the auditory localization we define the axes system and corresponding angle definition used throughout the thesis, it is shown in Figure 2.1. The axes system is roughly based on the definitions found in [12] and [13]. [12] is

the report corresponding to the Head Related Transfer Function (HRTF) data used in the thesis. In this definition, note that the elevation angle has double defined angles. This is done because only elevation rotations from 0° to 180° are considered for the elevation angle. The elevation angles at the left half of the elevation definition are mirrored to the right side by adding or subtracting 180° to the azimuth angle. This axis system is considered since we do not consider upside-down head orientations which is an uncommon position when listening to music or watching a movie.

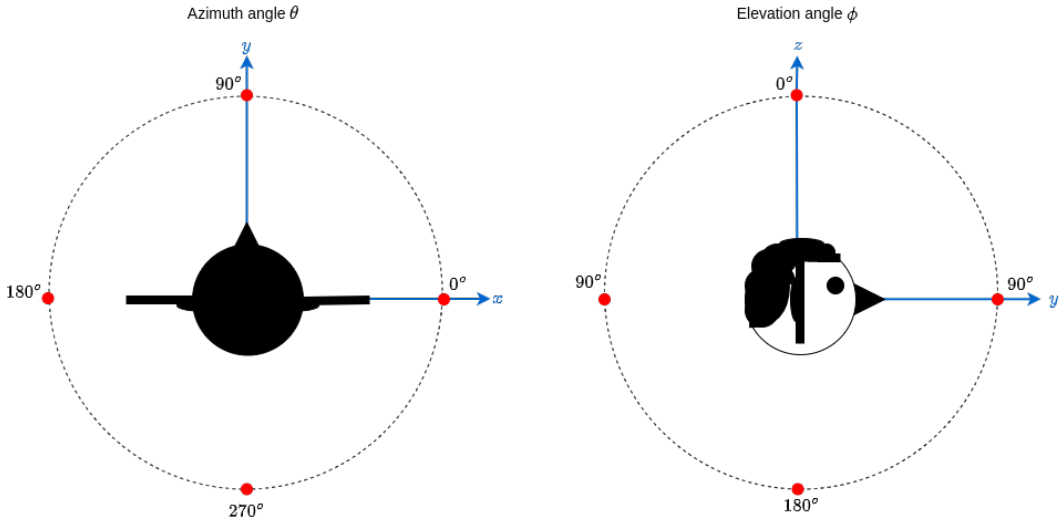


Figure 2.1: Definition of the axis system and corresponding angle definitions. Do note the uncommon definition of the elevation angle. This is done because only elevation rotations from 0° to 180° are considered for the elevation angle. The elevation angles at the left half of the elevation definition are mirrored to the right side by adding or subtracting 180° to the azimuth angle.

2.1.1. Azimuth angle estimation

The estimation of the azimuth angle is performed using the differences between the response found at the left and right ear. Both ears lie in the horizontal plane in which we determine the azimuth angle. It seems intuitive that this placement of the ears allows for great azimuth angle estimation performance and literature shows it does [14–21]. First we discuss the two major cues used for this, Interaural Time Difference (ITD) and Interaural Level Difference (ILD), which is found in the so-called duplex-theory. We follow with a small sidestep to the InterAural CrossCorrelation (IACC) measure, which represents how our brain determines the azimuth angle estimation. After this, we discuss the performance of azimuth estimation based on multiple experiments found in various published literature [18, 22–24] and finalize with a few conflicting sources [25, 26].

Duplex theory

The duplex theory describes how humans are able to localize sound sources in the azimuthal plane using a combination of the ITD and ILD cues [14, 15], first their definitions are given followed by their contribution to the duplex theory.

Interaural time difference

In general, for a given source location and head orientation, the time it takes for the wave to reach the left and right ear is not the same. This difference in arrival

time, referred to as Interaural Time Difference (ITD), can be used to determine the azimuth angle. Figure 2.2 shows an example of this. Throughout our lifetime our brain learns how to interpret certain delays in arrival time and translate them to the corresponding azimuth angle.

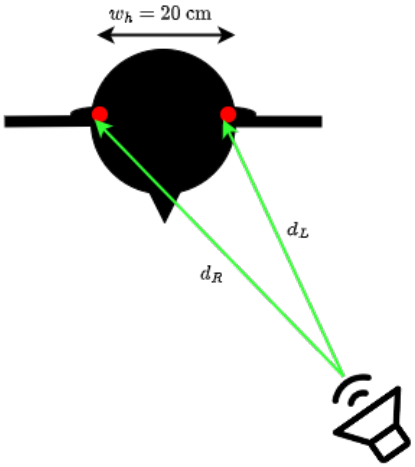


Figure 2.2: Example illustrating the interaural time difference and interaural level difference. Here, the path from the source to the right ear is longer than the path from the source to the left ear. Where d_L and d_R is the distance from source to left and right ear respectively and w_h is the width of the head, estimated around 20 cm.

There is a certain maximum frequency for which ITD can be used, this has to do with the width of the head. When half the wavelength is smaller than the distance between the ears, a certain threshold is reached after which the brain is unsure whether the difference is half a wavelength, one and a half wavelength or more than that. Because of this, we do not use ITD for frequencies higher than $f = 0.86$ kHz, according to Equation (2.1), where v_s is the speed of sound and w_h is the width of the head.

$$f = \frac{1}{2}\frac{v_s}{w_h} = \frac{1}{2}\frac{342}{0.20} = 0.86 \text{ kHz} \tag{2.1}$$

This is roughly in line with the 1.5 kHz generally found in literature [14–18]. Some research has shown that some people have developed the ability to use ITD at a higher frequencies by means of envelope analyses [15].

Interaural Level Difference

The Interaural Level Difference (ILD) is based on the same principle as the ITD, the sound source is measured slightly different by the left and the right ear. In the case of the ILD the small level difference expressed in Sound Pressure Level (SPL) between the two signals is considered. For instance, in Figure 2.2, the SPL at the right ear is less than the SPL at the left ear. This can be used to localize a sound source. On top of the level difference occurring due to the difference in audio path distance, the shadowing effect of the head in between the ears also contributes to the level difference.

The downside of the ILD is that it is less reliable than the ITD. Since the difference in SPL depends on the distance of the listener to the source, this cue is not fully reliable without prior knowledge of the distance to the source. The ITD does not have such limitations.

The ITD and ILD cues together compose the duplex theory. Research shows that the ITD cue is the most dominant cue and when both ITD and ILD are present but opposing, the ITD cue is followed [19]. ITD is only used for sound source with frequencies smaller than about 1.5 kHz after which it becomes less reliable. When localizing higher frequency content, the ILD is used, with limited performance. A sound source can be best localized when both the ITD and ILD can be used. The larger the bandwidth, the better the localization [20, 21].

Interaural crosscorrelation

The InterAural CrossCorrelation (IACC) is an important measure when considering audio perception. In short, as the name suggests, the IACC is the crosscorrelation between the response found at the left and right ear, which are denoted as $\mathbf{y}_{\mathrm{L}} \in \mathbb{R}^{N_b \times 1}$ and $\mathbf{y}_{\mathrm{R}} \in \mathbb{R}^{N_b \times 1}$ respectively where N_b is the length of the response. The IACC is defined as given in Equation (2.2) [27].

$$y_{\mathrm{IACC}}(\bar{n_b}) = \frac{\sum_{n_b=1}^{N_b} y_{\mathrm{L}}(n_b) y_{\mathrm{R}}(n_b - \bar{n_b})}{\sqrt{\mathbf{y}_{\mathrm{L}}{}^{\mathrm{T}} \mathbf{y}_{\mathrm{L}} \mathbf{y}_{\mathrm{R}}{}^{\mathrm{T}} \mathbf{y}_{\mathrm{R}}}} \tag{2.2}$$

Where $y_{\mathrm{IACC}}(\bar{n_b})$ is sample $\bar{n_b}$ of the InterAural Crosscorrelation with $\bar{n_b} = 1, ..., 2N_b - 1$, $y_{\mathrm{L}}(n_b)$ and $y_{\mathrm{R}}(n_b)$ are the n_b^{th} sample of the responses $\mathbf{y}_{\mathrm{L}}$ and $\mathbf{y}_{\mathrm{R}}$ with $n_b = 1, ..., N_b$ and $(\cdot)^{\mathrm{T}}$ denotes the transpose operation. Note that the term in the denominator is a normalization term.

The IACC cue is mainly treated in literature when analysing the quality of audio in terms of the feeling of envelopment and spatial impression in, for instance, concert halls [28–30]. Although these are interesting topics, we are not interested in this.

An aspect of the IACC we are interested in is the relation between the IACC and the azimuth angle estimation. It is shown that it is very likely that the brain calculates (something related to) the IACC and uses the index of the peak, τ_{IACC}, of this response to determine the azimuth angle [31, 32]. τ_{IACC} is determined by means of Equation (2.3), where the limiting set indicates that delays outside this range are too large to correspond to valid auditory cues. The index of the peak value represents the ITD between both ears. A few examples of IACC responses are given in Figure 2.3.

$$\tau_{\mathrm{IACC}} = \arg\max_{\bar{n_b}} \quad y_{\mathrm{IACC}}(\bar{n_b}), \qquad \text{for} \quad \bar{n_b} \in [-1, 1] \text{ ms} \tag{2.3}$$

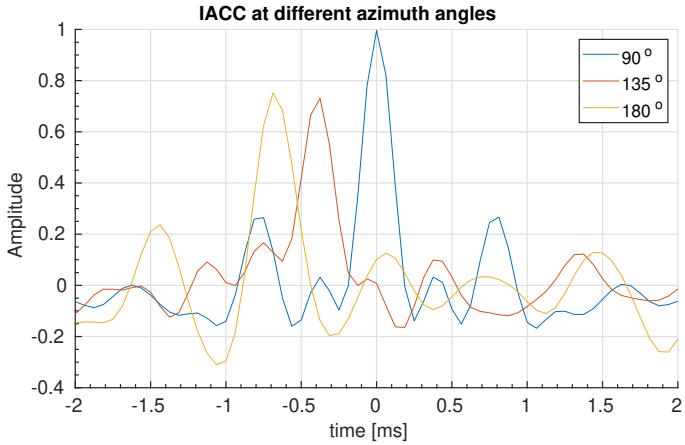


Figure 2.3: Example of IACC responses at different azimuth angles. The time index of the peak values corresponds to the ITD observed.

As can be seen in this figure, when the source is in front of the listener, at 90° azimuth, the IACC peak is at $\tau_{\text{IACC}} = 0$ ms. This makes sense considering the signal arrives at the left and right ear at the same time. Looking at the case of the source being to the left of the listener, at 180° azimuth, the peak of the IACC is at $\tau_{\text{IACC}} \approx -0.75$ ms. Once again, this corresponds to the found ITD. It is interesting to note that this τ_{IACC} corresponds to a distance of $v_s \cdot \tau_{\text{IACC}} = 343 \cdot 750 \cdot 10^{-6} = 0.26$ m, which is roughly the width of the head, w_h. This result is in line with the source being at 180° azimuth since the sound to the right ear has to travel w_h more distance than the sound travelling to the left ear.

Azimuth angle estimation performance

The duplex theory gives us the ability to localize audio sources in the azimuthal plane, but how good is our performance? Our localization performance is generally best in front of us with a maximal error of about 5° and this increases to about 20° when localizing sources to the left or right [18, 22–24].

Conflicting sources

Although the statements above are confirmed by several studies, there are some sources that present different results in my opinion, they are given here for sake of completeness. Note that most of the conflicting information originates from the fact that the data is obtained by means of subjective experiments, there are numerous reasons why this would lead to conflicting findings. [25] states that the most important frequency band to perform localization is $4 \leftrightarrow 16$ kHz which greatly conflicts with the importance of the ITD cue. [26] states a similar find by concluding that content with frequencies under 2 kHz has little contribution to auditory localization.

2.1.2. Elevation angle estimation

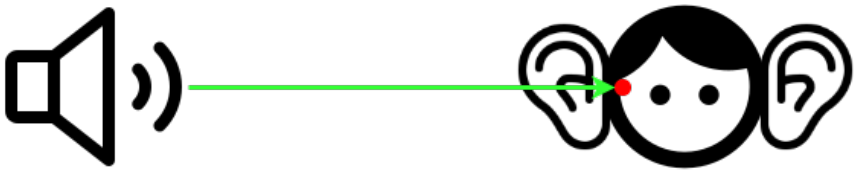
While the estimation in the azimuthal plane is relatively straightforward, this is not true for elevation estimation. In the elevation plane, both ears are at the same location meaning the responses can not be compared to obtain useful information. All the spatial information must be gathered from one sensor in said space. This brings us to the monaural cues, which are the primary source of information for elevation localization. We elaborate on the Head Related Transfer Function (HRTF) and discuss its contribution and key features according to literature.

Head related transfer function

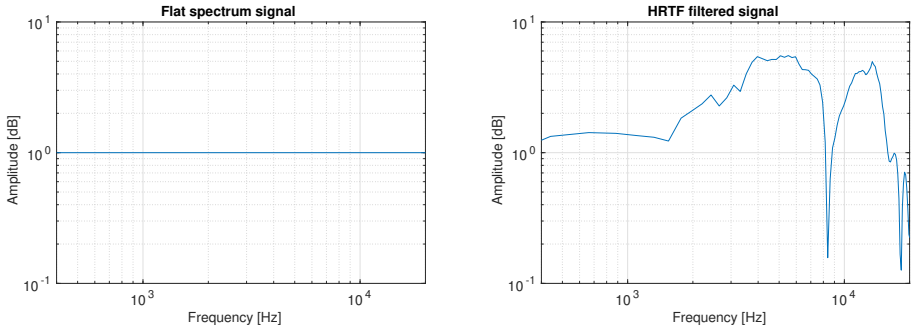
Defining the Head Related Transfer Function (HRTF) is not a straightforward task. In short, the HRTF is how our body and, most importantly, the ears alter the signal that comes in from a source away from the body [14]. The HRTF is generally measured using a microphone placed deep inside the ear canal using a source placed about a meter away from the subject. The HRTF is subject dependent, everyone has a different HRTF. An example of an HRTF response based on measurements on the KEMAR head model [33] is given in Figure 2.4. In the figure, we can see how a flat spectrum signal is altered by the HRTF.

The KEMAR head model is meant to represent the average head across human civilization such that the measurements done on this head should come close to everyone's personal HRTF. Even though this is far from true and there is a serious mismatch between a personal HRTF and the average HRTF, it is the best we can do when trying to generalize the HRTF. The HRTF's are generally measured using a source a meter away from the subject but this HRTF is also applicable to sources placed further away [34].

Throughout our lifetime we learn how to interpret the HRTF and we are also able to adapt to a different HRTF given enough time [35]. Generally, we use the



(a) Filter path described by the HRTF.



(b) An example HRTF corresponding to the filter path indicated in (a).

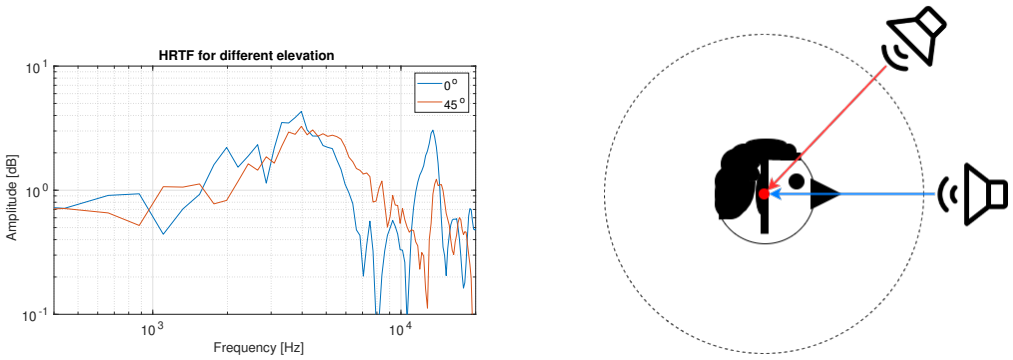
Figure 2.4: Example of an HRTF response given a measurement based on the KEMAR head model.

spectrum of the incoming signal and its features to estimate the elevation angle [36], an example is given in Figure 2.5. In the figure, it is shown that certain peaks and dips are moved when the elevation angle changes and we use the frequencies of these peaks and dips to estimate the elevation angle [37].

The frequency content of the to be localized sound source influences the localization abilities. Peaks and dips in the spectrum of the source can lead to misleading localization cues. Generally it is found that we localize a sound source by assuming the source has a flat spectrum [19].

2.2. Human sound perception

The human auditory system contains an impressive and well performing auditory source localization functionality, but this is not the only impressive characteristic



(a) HRTF responses for different elevation angles.

(b) Filter paths corresponding to the two HRTF responses given in (a).

Figure 2.5: HRTF for different elevation angles, the colors of the arrows and the responses correspond to each other.

of the human auditory apparatus. We are able to detect low power sound sources, even when a louder distortion sound source is present [38]. Although our sound detection is impressive, it still has limited performance. These limitations consist of the minimum detectable sound pressure and the influence of present sound sources on this detectability. In the proposed algorithm, these limitations are exploited by allowing unnoticeable errors which improve the robustness of the algorithm.

Numerous attempts have been done to quantify the ability to perceive sound (e.g. [39]) and in the following, the critically acclaimed measure used in this thesis are discussed. We start with the definition of the threshold in quiet followed by modelling the human auditory apparatus. After this, we combine these two to introduce sound masking principles and models thereof.

2.2.1. Absolute threshold in quiet

The absolute threshold of hearing is a frequency dependent measure, or curve in practise, that shows the just noticeable Sound Pressure Level (SPL) for an average human listener. Sound sources resulting in sound pressure levels underneath the curve are not noticeable by the auditory system and vice versa.

Before we can introduce the definition of the threshold in quiet we must first define the SPL in decibels. The SPL is defined as the detected Sound Pressure Level of a sound stimulus relative to a standardized value [40]. The standardized pressure level is given by $p_0 = 20$ μPa or equivalently $p_0 = 20$ μN/m^2 [41]. The resulting sound pressure level in dB SPL of sound stimulus p_s is given by $p_{\text{SPL}} = 20\log_{10}(p_s/p_0)$ (dB SPL). As an indication, 60 dB SPL corresponds to normal speech at 1 meter distance, 100 dB SPL corresponds to a disco and 150 dB SPL corresponds to a jet engine at 10 meter distance leading to permanent hearing damage [42].

With SPL defined, we can introduce the definition of the absolute threshold of hearing, also known as the threshold in quiet. The threshold in quiet, denoted by $T_q(f)$ with f the frequency in Hz, is given in Equation (2.4) [43].

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 \text{ (dB SPL)} \tag{2.4}$$

The threshold in quiet is also depicted in Figure 2.6. Interpreting the threshold in quiet, as shortly mentioned before, is relatively straightforward. With no other sound sources or noise present, sound stimuli underneath the curve in Figure 2.6 cannot be heard or detected by the human listener. All the stimuli above the curve are heard and detected by the human listener.

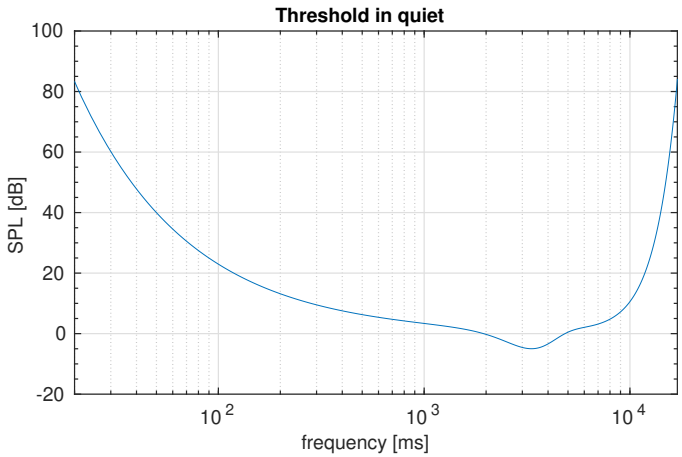


Figure 2.6: The threshold in quiet.

2.2.2. Human auditory apparatus

The threshold in quiet shows the limitations of sound detection in a quiet scenario. If the listener is not listening in quiet but a different sound source is present, the sound detection performance changes. In practise, the sound source raises the threshold in quiet curve at certain frequencies dependent on the characteristics of the sound source. Note that this is beneficial to us, the higher the curve, the more space we have to make unnoticeable errors! When a sound source is present, the threshold curve is generally referred to as the masking curve. Before we can derive the exact shape of the masking curve based on the sound source's properties, we must first discuss the human auditory apparatus and its frequency dependent behaviour.

When a sound wave enters the ear, it eventually reaches the cochlea and the basilar membrane which, roughly speaking, records the sound wave and sends corresponding signals to the brain. See [40, 41] for more details. The functionality of the basilar membrane can be modelled as a filter bank ranging over the entire audible spectrum, 20 to 20000 Hz [44–46]. The filters are divided over the audible spectrum in logarithmic fashion which suits the behaviour of the auditory system [38].

Deriving the filters in the filter bank starts with determining the center frequencies of these filters. For this we use logarithmic Equation (2.5) and find the frequencies f_γ corresponding to the linear spacing of output $\bar{f}$ [47], as shown in Figure 2.8a.

$$\bar{f}(f_\gamma) = 21.4\log_{10}(4.37 f_\gamma/1000 + 1) \tag{2.5}$$

With the centre frequencies defined, the bandwidth of each cochlear filter can be determined. The critical bandwidth measure, as posed in [40], relates centre frequencies and critical bandwidths based on subjective experiments. The critical bandwidth measure, denoted by BW_c, is given by Equation 2.6.

$$\mathrm{BW}_c(f_\gamma) = 25 + 75(1 + 1.4(f_\gamma/1000)^2)^{0.69} \text{ (Hz)} \tag{2.6}$$

Although widely used, other expressions exist. The measure used further on in this thesis is the Equivalent Rectangular Bandwidth (ERB) [47–49] and it is presented in Equation (2.7).

$$\mathrm{ERB}(f_\gamma) = 24.7(4.37(f_\gamma/1000) + 1) \text{ (Hz)} \tag{2.7}$$

Besides the fact that [44–46] use the ERB scale, the main reason for choosing this measure is that it is based on more experiments and it is composed of numerous different more elaborate measures [38].

With the centre frequency and the bandwidth of each filter in the basilar membrane filter bank known, the actual filters can be derived. The filters are modelled by an η_γ^{th} order gamma-tone filter $\gamma(f)$ given by Equation (2.8) [44, 45].

$$\gamma(f) = \frac{1}{\left(1 + \left(\frac{f - f_\gamma}{\kappa_\gamma ERB(f_\gamma)}\right)^2\right)^{\frac{\eta_\gamma}{2}}} \tag{2.8}$$

Here, η_γ is generally taken as $\eta_\gamma = 4$ and $\kappa_\gamma = \frac{2^{(\eta_\gamma - 1)}(\eta_\gamma - 1)!}{\pi(2\eta_\gamma - 3)!!}$ with ! the factorial and !! the double factorial operator.

2.2.3. Auditory maskers

With the model for the human auditory apparatus defined, the sound masking principles can be defined and the so-called masking curve can be constructed. In this

section, a model is discussed that is non-ideal but can be used efficiently in optimization problems, which is an important property for the proposed algorithm. The model and its calibration is primarily based on the findings in [44, 45].

The masking curve model is depicted in Figure 2.7, the model is described in more detail in the following.

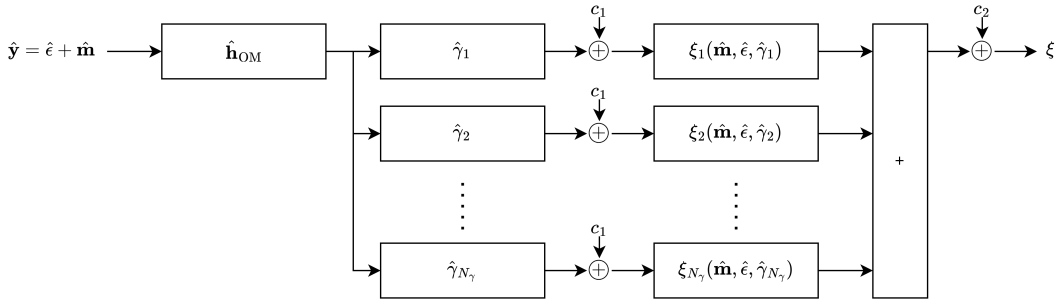


Figure 2.7: Human auditory apparatus model used to construct the masking curve.

We start with the frequency domain signal coming into the model, $\hat{\mathbf{y}} \in \mathbb{C}^{N_b \times 1}$, which represents either $\hat{\mathbf{y}}_{\mathrm{L}}$ or $\hat{\mathbf{y}}_{\mathrm{R}}$, the response found inside the left or right ear. $\hat{\mathbf{y}}$ is divided into the masker signal $\hat{\mathbf{m}} \in \mathbb{C}^{N_b \times 1}$ and the distortion $\hat{\boldsymbol{\epsilon}} \in \mathbb{C}^{N_b \times 1}$.

To simulate the frequency dependent passive filtering done by the shape of the outer and middle ear, $\hat{\mathbf{y}}$ is filtered by outer- and middle-ear filter $\hat{\mathbf{h}}_{\mathrm{OM}} \in \mathbb{R}^{N_b \times 1}$. It is important to note that the system model we use (presented in Chapter 4) considers the Head Related Transfer Function (HRTF) which includes an outer- and middle-ear filter. The reason for this is that the HRTF is measured deep in the ear canals. Because of this, $\hat{\mathbf{h}}_{\mathrm{OM}}$ is not considered in the implemented algorithm presented in Chapter 5 as it is already implied. For the analysis in this chapter the outer- and middle-ear filter presented in [44] is used.

The outer- and middle-ear filter considered in [44] is the inverse of the threshold in quiet as presented in Equation (2.4), $\hat{\mathbf{h}}_{\mathrm{OM}} = -\hat{\mathbf{t}}_q$ (dB).

The gamma-tone filter bank consists of $N_\gamma = 64$ gamma-tone filters. The centre frequencies f_γ are determined by finding the linear spacing of $\bar{f}$ in Equation (2.5) as is depicted in Figure 2.8a. The corresponding gamma-tone filter bank $\hat{\mathbf{\Gamma}} \in \mathbb{R}^{N_\gamma \times N_b}$ derived from Equation (2.8) is given in Figure 2.8b. The black plot in Figure 2.8b shows the sum of the absolute square of all the individual filters $\hat{\mathbf{p}}_\gamma \in \mathbb{R}^{N_b \times 1}$, as also given in Equation (2.9), where $n_\gamma = 1, ..., N_\gamma$ and $n_b = 1, ..., N_b$. We can see that the result is a near flat power response which is desired according to [38, 45], it shows that the filters themselves do not apply a frequency amplification or suppression.

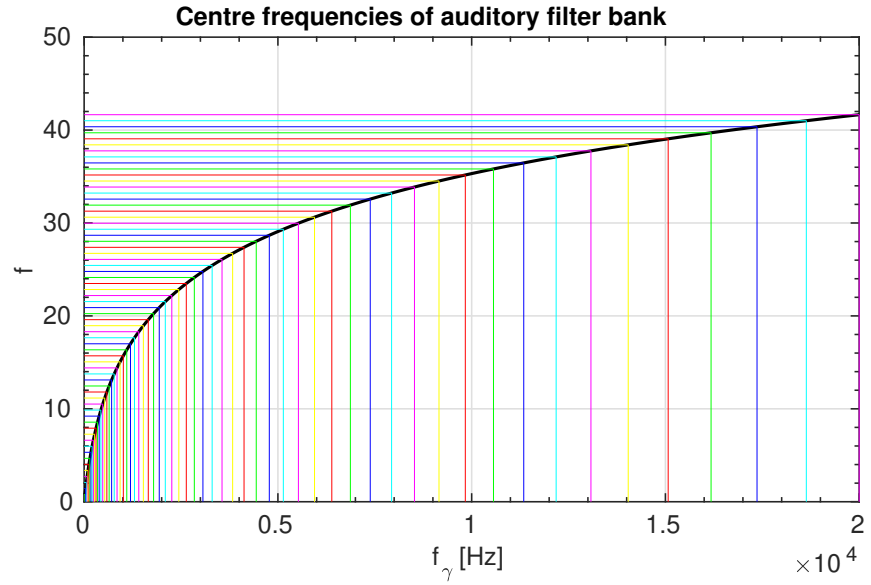
$$\hat{p}_\gamma(n_b) = \sum_{n_\gamma} |\hat{\gamma}(n_\gamma, n_b)|^2 \tag{2.9}$$

As the model in Figure 2.7 shows, a constant c_1 is added to the signal which represents the noise floor of the auditory system. The value of this constant is determined later in the calibration stage of the model.

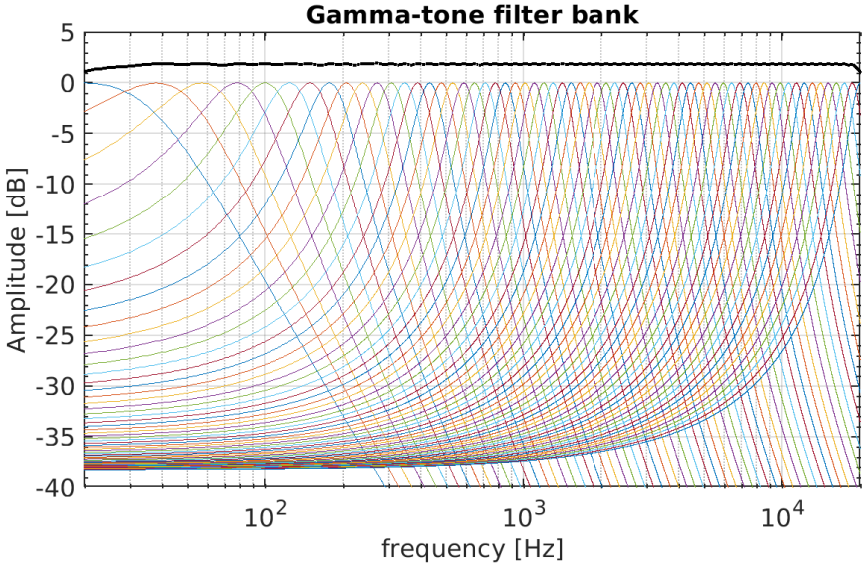
Before we can determine the filter dependent detectability ξ_{n_γ}, we must determine the masker and distortion power for each filter. They are given in Equation (2.10) and Equation (2.11) respectively.

$$m_{p,n_\gamma} = \frac{1}{N_b} \sum_{n_b} |\hat{h}_{\mathrm{OM}}(n_b)|^2 |\hat{\gamma}_{n_\gamma}(n_b)|^2 |\hat{m}(n_b)|^2 \tag{2.10}$$

$$\epsilon_{p,n_\gamma} = \frac{1}{N_b} \sum_{n_b} |\hat{h}_{\mathrm{OM}}(n_b)|^2 |\hat{\gamma}_{n_\gamma}(n_b)|^2 |\hat{\epsilon}(n_b)|^2 \tag{2.11}$$



(a) Centre frequencies of gamma-tone filter. The coloured lines indicate the centre frequencies f_γ corresponding to the linear spacing of $\bar{f}$.



(b) Gamma-tone filter bank modelling the filter bank functionality of the basilar membrane. The black plot at the top shows the sum of the absolute square of the individual filters.

Figure 2.8: Derivation of the gamma-tone filter bank. As shown in (a), the centre frequencies of the filters are determined by finding the frequencies f_γ corresponding to the linear spacing of $\bar{f}$ in Equation (2.5). Given these centre frequencies, the filter bank as shown in (b) is derived.

Given these powers, the detectability for each filter is given by Equation (2.12)

$$\xi_{n_\gamma} = \frac{\epsilon_{p,n_\gamma}}{m_{p,n_\gamma} + c_1} \tag{2.12}$$

The last step represents the ability of humans to combine information over all the filters in the filter-bank to improve detectability. Including this property can be approximated by a summation as given in Equation (2.13). In here, c_2 is a constant used to make sure that $\xi = 1$ represents a just noticeable distortion.

$$\xi = c_2 \sum_{n_\gamma} \xi_{n_\gamma} = c_2 \sum_{n_\gamma} \frac{\sum_{n_b} |\hat{h}_{\text{OM}}(n_b)|^2 |\hat{\gamma}_{n_\gamma}(n_b)|^2 |\hat{\epsilon}(n_b)|^2}{\sum_{n_b} |\hat{h}_{\text{OM}}(n_b)|^2 |\hat{\gamma}_{n_\gamma}(n_b)|^2 |\hat{m}(n_b)|^2 + c_1} \tag{2.13}$$

With the detectability defined, it is possible to define the masking curve. The masking curve represents the frequency dependent, just noticeable distortion given the presence of a masker. To find the masking curve $\hat{\mathbf{g}}$, we solve Equation (2.13) for $\hat{\epsilon}(\bar{n_b})$, where $\bar{n_b} = 1, ..., N_b$, given $\xi = 1$.

$$\frac{1}{\hat{g}^2(\bar{n_b})} = \frac{1}{\hat{\epsilon}^2(\bar{n_b})} = c_2 \sum_{n_\gamma} \frac{|\hat{h}_{\text{OM}}(\bar{n_b})|^2 |\hat{\gamma}_{n_\gamma}(\bar{n_b})|^2}{\sum_{n_b} |\hat{h}_{\text{OM}}(n_b)|^2 |\hat{\gamma}_{n_\gamma}(n_b)|^2 |\hat{m}(n_b)|^2 + c_1} \tag{2.14}$$

Calibrating the model

The calibration of the model is based on two findings in psychoacoustical research [44, 45]. First the threshold in quiet is used, specifically at $f = 1$ kHz [44, 45] (from now on referred to as $n_{b,1\text{kHz}}$). This finding allows us to calibrate constant c_1 by setting $\hat{\mathbf{m}} = 0$, $\hat{\boldsymbol{\epsilon}} = \hat{\mathbf{t}}_q \, \text{diag}(\delta(n_b - n_{b,1\text{kHz}}))$ and $\xi = 1$, with $\delta(\cdot)$ the Kronecker delta and $\text{diag}(\cdot)$ the diagonal matrix with the input vector on the diagonal. Substituting this into Equation (2.13) and rewriting it leads to Equation (2.15)

$$c_1 = c_2 \sum_{n_\gamma} |\hat{h}_{\text{OM}}(n_{b,1\text{kHz}})|^2 |\hat{\gamma}_{n_\gamma}(n_{b,1\text{kHz}})|^2 |\hat{t}_q(n_{b,1\text{kHz}})|^2 \tag{2.15}$$

Psychoacoustics has shown that humans have a 1 dB Just Noticeable Difference (JND) at a volume of 70 dB SPL [44, 45], we use this to calibrate c_2. To obtain the 1 dB level difference, a 1 kHz sinusoid is chosen with amplitude A_{70} and A_{52} for masker and distortion respectfully. Here, A_{70} and A_{52} are the amplitudes corresponding to a 70 and 52 dB SPL signal. Given this setting, and $\xi = 1$, Equation (2.13) can be rewritten to find c_2, it is given in Equation (2.16). Combining Equation (2.15) and (2.16) gives the solution to both constants.

$$\frac{1}{c_2} = \sum_{n_\gamma} \frac{|\hat{h}_{\text{OM}}(n_{b,1\text{kHz}})|^2 |\hat{\gamma}_{n_\gamma}(n_{b,1\text{kHz}})|^2 A_{52}^2}{|\hat{h}_{\text{OM}}(n_{b,1\text{kHz}})|^2 |\hat{\gamma}_{n_\gamma}(n_{b,1\text{kHz}})|^2 A_{70}^2 + c_1} \tag{2.16}$$

To show an example of the functioning of the model, Figure 2.9 shows the masking curve with and without the presence of a 52 dB SPL sinusoid at 1 kHz masker. As can be seen, the masker does not only influence the masking curve at 1 kHz but at a wider frequency range.

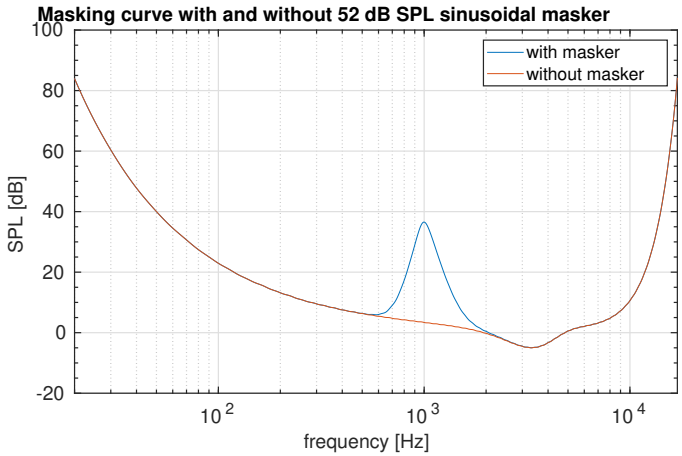


Figure 2.9: Masking curve with and without a 52 dB SPL sinusoid at 1 kHz masker.

3

Crosstalk cancellation

As stated in Section 1.5.2, crosstalk cancellation was first introduced to mimic the headphone experience using a pair of stereo loudspeakers. Although this is an interesting usecase it can, in theory, be extended to create any virtual source using any setup with at least two loudspeakers. As also noted in Section 1.5.2, this approach does not result in a practical stable solution and making crosstalk cancellation applicable in practise is the main goal of this thesis. Before we can improve the crosstalk cancellation algorithm, the original algorithm has to be properly defined and analysed, this is done in this chapter. First, the original crosstalk cancellation algorithm according to literature is discussed and expanded. After this, an in-depth analysis of the original crosstalk cancellation algorithm is performed and the major causes of problems is identified.

3.1. Crosstalk cancellation, a literature review

Given channels $\mathbf{c}_\mathrm{L}$, $\mathbf{c}_\mathrm{R}$, input signal $\mathbf{s}$ and received signals $\mathbf{y}_\mathrm{L}$ and $\mathbf{y}_\mathrm{R}$, crosstalk cancellation is posed in the literature to create any desired response at the two ears with at least two loudspeakers.

The solution to the crosstalk cancellation problem as presented by [10, 50–52] can be described with respect to the setup presented in Figure 3.1. As the figure shows, the goal is to cancel the crosstalk channels and preserve the direct channels.

The equations describing the crosstalk cancellation, as shown in Figure 3.1, are given in Equations (3.1) and (3.2) for the left and right ear response respectively. The equations are presented in the frequency domain. In the equation, $n_s = 1, ...N_s$ denotes the loudspeaker number with $N_s = 2$ the number of loudspeakers in the setup, $\hat{\mathbf{y}}_{\mathrm{L},n_s} \in \mathbb{R}^{N_b \times 1}$ and $\hat{\mathbf{y}}_{\mathrm{R},n_s} \in \mathbb{R}^{N_b \times 1}$ are the responses found at the left and right ear respectively coming from loudspeaker n_s, N_b is the length of the response, $\hat{\mathbf{s}}_{n_s} \in \mathbb{R}^{N_b \times 1}$ is the input signal presented to loudspeaker n_s, $\hat{\mathbf{c}}_{\mathrm{L},n_s} \in \mathbb{R}^{N_b \times 1}$ and $\hat{\mathbf{c}}_{\mathrm{R},n_s} \in \mathbb{R}^{N_b \times 1}$ are the channels from loudspeaker n_s to the left and right ear respectively and $\hat{\mathbf{C}}_{\mathrm{L},n_s} = \mathrm{diag}(\hat{\mathbf{c}}_{\mathrm{L},n_s})$ and $\hat{\mathbf{C}}_{\mathrm{R},n_s} = \mathrm{diag}(\hat{\mathbf{c}}_{\mathrm{R},n_s})$ with $\mathrm{diag}(\cdot)$ denoting a square matrix with the input vector on the diagonal (equivalent to a pointwise multiplication).

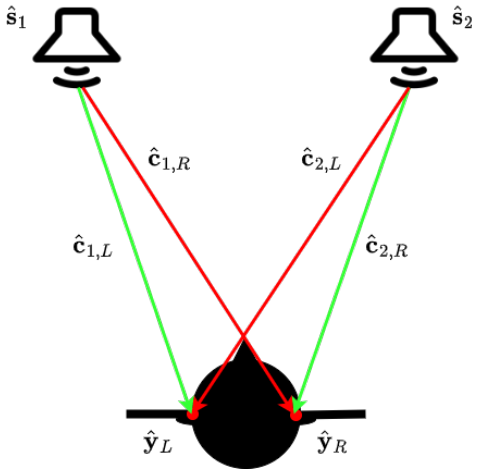


Figure 3.1: Crosstalk Cancellation setup where the green arrows depict the channels that should be preserved and the red arrows depict the crosstalk channels that should be cancelled.

$$\hat{\mathbf{y}}_{\mathrm{L}} = \sum_{n_s} \hat{\mathbf{y}}_{\mathrm{L},n_s} = \sum_{n_s} \hat{\mathbf{C}}_{\mathrm{L},n_s} \hat{\mathbf{s}}_{n_s} \tag{3.1}$$

$$\hat{\mathbf{y}}_{\mathrm{R}} = \sum_{n_s} \hat{\mathbf{y}}_{\mathrm{R},n_s} = \sum_{n_s} \hat{\mathbf{C}}_{\mathrm{R},n_s} \hat{\mathbf{s}}_{n_s} \tag{3.2}$$

The two equations can be rewritten into a matrix form given by Equation (3.3).

$$\begin{bmatrix} \hat{y}_{\mathrm{L}}(1) \\ \hat{y}_{\mathrm{R}}(1) \\ \hat{y}_{\mathrm{L}}(2) \\ \hat{y}_{\mathrm{R}}(2) \\ \vdots \\ \hat{y}_{\mathrm{L}}(N_b) \\ \hat{y}_{\mathrm{R}}(N_b) \end{bmatrix} = \begin{bmatrix} \hat{c}_{\mathrm{L},1}(1) & \hat{c}_{\mathrm{L},2}(1) & 0 & 0 & \dots & 0 & 0 \\ \hat{c}_{\mathrm{R},1}(1) & \hat{c}_{\mathrm{R},2}(1) & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \hat{c}_{\mathrm{L},1}(2) & \hat{c}_{\mathrm{L},2}(2) & \dots & 0 & 0 \\ 0 & 0 & \hat{c}_{\mathrm{R},1}(2) & \hat{c}_{\mathrm{R},2}(2) & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \hat{c}_{\mathrm{L},1}(N_b) & \hat{c}_{\mathrm{L},2}(N_b) \\ 0 & 0 & 0 & 0 & \dots & \hat{c}_{\mathrm{R},1}(N_b) & \hat{c}_{\mathrm{R},2}(N_b) \end{bmatrix} \begin{bmatrix} \hat{s}_1(1) \\ \hat{s}_2(1) \\ \hat{s}_1(2) \\ \hat{s}_2(2) \\ \vdots \\ \hat{s}_1(N_b) \\ \hat{s}_2(N_b) \end{bmatrix} \rightarrow \hat{\mathbf{y}} = \hat{\mathbf{C}}\hat{\mathbf{s}} \tag{3.3}$$

With received signal vector $\hat{\mathbf{y}} \in \mathbb{C}^{2N_b \times 1}$, channel matrix $\hat{\mathbf{C}} \in \mathbb{C}^{2N_b \times 2N_b}$ and input vector $\hat{\mathbf{s}} \in \mathbb{C}^{2N_b \times 1}$. The crosstalk cancellation problem as treated in literature aims to achieve $\hat{\mathbf{y}}^* = \hat{\mathbf{s}}$, where $(\cdot)^*$ denotes the desired or optimal solution. The solution posed in literature is a simple matrix inversion presented in Equation (3.4).

$$\hat{\mathbf{y}}^* = \hat{\mathbf{C}}\hat{\mathbf{C}}^{-1}\hat{\mathbf{s}} = \mathbf{I}\hat{\mathbf{s}} = \hat{\mathbf{C}}\hat{\mathbf{s}}^* \tag{3.4}$$

Here $(\cdot)^{-1}$ denotes the matrix inverse, $\mathbf{I}$ is the identity matrix and $\hat{\mathbf{s}}^* = \hat{\mathbf{C}}^{-1}\hat{\mathbf{s}}$ is the optimal solution for $\hat{\mathbf{s}}$ to achieve $\hat{\mathbf{y}}^* = \hat{\mathbf{s}}$. Do note that the inverse of the channel matrix may not exist leading to a non-existing solution.

The given setup can be extended to contain more than two loudspeakers, $N_s \geq 2$, resulting in the matrix form given in Equation (3.5) (Note that Equation (3.3) is a special case, $N_s = 2$, of Equation (3.5)).

$$\begin{bmatrix} \hat{y}_{\mathrm{L}}(1) \\ \hat{y}_{\mathrm{R}}(1) \\ \hat{y}_{\mathrm{L}}(2) \\ \hat{y}_{\mathrm{R}}(2) \\ \vdots \\ \hat{y}_{\mathrm{L}}(N_b) \\ \hat{y}_{\mathrm{R}}(N_b) \end{bmatrix} = \begin{bmatrix} \hat{c}_{\mathrm{L},1}(1) & \dots & \hat{c}_{\mathrm{L},N_s}(1) & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ \hat{c}_{\mathrm{R},1}(1) & \dots & \hat{c}_{\mathrm{R},N_s}(1) & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & \hat{c}_{\mathrm{L},1}(2) & \dots & \hat{c}_{\mathrm{L},N_s}(2) & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & \hat{c}_{\mathrm{R},1}(2) & \dots & \hat{c}_{\mathrm{R},N_s}(2) & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & \hat{c}_{\mathrm{L},1}(N_b) & \dots & \hat{c}_{\mathrm{L},N_s}(N_b) \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & \hat{c}_{\mathrm{R},1}(N_b) & \dots & \hat{c}_{\mathrm{R},N_s}(N_b) \end{bmatrix} \begin{bmatrix} \hat{s}_1(1) \\ \vdots \\ \hat{s}_{N_s}(1) \\ \hat{s}_1(2) \\ \vdots \\ \hat{s}_{N_s}(2) \\ \vdots \\ \hat{s}_1(N_b) \\ \vdots \\ \hat{s}_{N_s}(N_b) \end{bmatrix}$$

$$\rightarrow \hat{\mathbf{y}} = \hat{\mathbf{C}}\hat{\mathbf{s}} \tag{3.5}$$

In this equation, the sizes of the variables are given by $\hat{\mathbf{y}} \in \mathbb{C}^{2N_b \times 1}$, $\hat{\mathbf{C}} \in \mathbb{C}^{2N_b \times N_s N_b}$ and $\hat{\mathbf{s}} \in \mathbb{C}^{N_s N_b \times 1}$. With this renewed setup, the goal $\hat{\mathbf{y}}^* = \hat{\mathbf{s}}$ is not viable anymore since we now have N_s loudspeaker outputs to obtain a desired response for only two points in space. The additional loudspeakers give us more freedom for optimization, possibly leading to more stable and robust solutions.

Now lets say we wish to obtain any $\hat{\mathbf{y}}^*$ given the multiple loudspeaker setup, the solution can be found using Least Squares minimisation. This solution is the original crosstalk cancellation solution as posed in literature. The solution is found in Equation (3.6) [53].

$$\hat{\mathbf{s}}^* = \arg\min_{\hat{\mathbf{y}}} ||\hat{\mathbf{y}}^* - \hat{\mathbf{y}}||_2^2 = \arg\min_{\hat{\mathbf{s}}} ||\hat{\mathbf{y}}^* - \hat{\mathbf{C}}\hat{\mathbf{s}}||_2^2 \tag{3.6}$$

Where $||(\cdot)||_2$ is the L_2-norm. The well known solution to this is given in Equation (3.7) and substitution of the solution for $\hat{\mathbf{s}}^*$ is shown in Equation (3.8).

$$\hat{\mathbf{s}}^* = (\hat{\mathbf{C}}^{\mathrm{H}}\hat{\mathbf{C}})^{-1}\hat{\mathbf{C}}^{\mathrm{H}}\hat{\mathbf{y}}^* = \hat{\mathbf{C}}^{\dagger}\hat{\mathbf{y}}^* \tag{3.7}$$

$$\hat{\mathbf{C}}\hat{\mathbf{s}}^* = \hat{\mathbf{C}}\hat{\mathbf{C}}^{\dagger}\hat{\mathbf{y}}^* = \mathbf{I}\hat{\mathbf{y}}^* = \hat{\mathbf{y}}^* \tag{3.8}$$

Here, $(\cdot)^{\mathrm{H}}$ denotes the Hermitian (conjugate) transpose and $(\cdot)^{\dagger}$ denotes the Moore-Penrose (pseudo) inverse. Note that the two loudspeaker solution presented in Equation 3.4 is a special case of the solution presented in Equations 3.7 and 3.8.

3.2. Limitations and issues of crosstalk cancellation

The CrossTalk Cancellation (CTC) approach presented thus far theoretically gives good results. In practise, the issues of CTC result in a poorly functioning algorithm. The reason CTC is not functioning is generally described by a solution that is not stable, not robust and vulnerable to errors and noise. The major cause of the inaccuracies can be found in the (pseudo) inverse of the channel found in the solutions (see Equation (3.4) and Equation (3.7)). The channel matrix is a badly conditioned matrix not suited for inversion and its inverse thus leads to the enhancement of noise and errors.

To further understand the problems with the original implementations of crosstalk cancellation, we discuss the underlying reasons for the bad inversion properties of the channel matrix. Understanding the factors that contribute to these properties hopefully allows us to improve the crosstalk cancellation algorithm. The found factors are described in the following sections.

3.2.1. Condition number and inversion quality

To show that the inversion of the channel matrix could potentially lead to implementation problems, we analyse the condition number of the matrix [54]. A similar analyses is presented in [55]. The condition number is a measure of how good the matrix is conditioned and thus how well the result of the inversion is defined. A high condition number means that the matrix is ill-conditioned and vice versa.

The condition number of the channel matrix is calculated using the singular values (SV) of the matrix [56]. The channel matrix $\hat{\mathbf{C}}$ as given in Equation (3.5) describes the full frequency range of interest of the channel. For better insight, we wish to evaluate the condition number per frequency bin and so $\hat{\mathbf{C}}$ is divided into one channel matrix for each frequency bin, given by $\hat{\mathbf{C}} \in \mathbb{C}^{2\times N_s}$. Due to the size of the channel matrices, the Singular Value Decomposition (SVD) provides only two SV's. Dividing the largest SV over the smallest SV gives the condition number of the channel matrix.

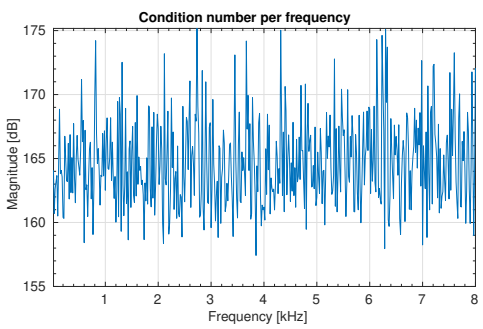
The stability of the matrix inversions is greatly dependent on the parts that compose the channels. These parts are the Room Impulse Response (RIR), the Head Related Transfer Function (HRTF), the (loud)Speaker Related Transfer Function (SRTF) and the number of and the position of the loudspeakers. Later in this chapter, these parts are discussed in more detail. Different situations and the corresponding condition numbers are depicted in Figure 3.2. The figures show an indication of the influence of different channel parts on the quality of the inversion.

As can be seen in Figure 3.2a, when we only consider the Room reflections, the condition number is very high. The obvious reason is that, in this case, the left and right response are the same leading to an inversion of a rank 1 matrix, the smallest singular value is practically zero.

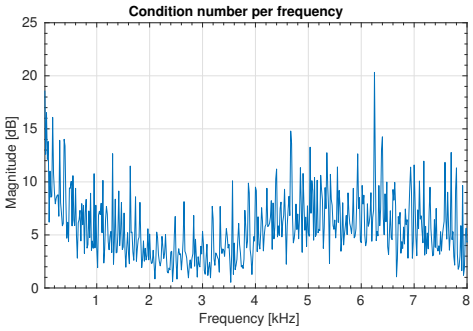
When adding the HRTF (described and evaluated in Section 2.1.2) to the model, the condition number drastically drops since the left and right response now clearly differ, this can be seen in Figure 3.2b. The fact that the condition number drops so drastically, is a very convenient property. It indicates that a more elaborate and realistic model actually improves crosstalk cancellation performance.

Next we include the SRTF to the analyses, which, in short, is a combination of the directivity and the response of the loudspeakers. Including this SRTF slightly decreases the conditionality of the matrix, as can be seen when comparing Figure 3.2b and 3.2c. The general behaviour of an SRTF is that the loudspeaker emits more energy from the front and less from the side, top, bottom and the back. Since we assume that the loudspeaker is always front-facing the listener, the most energy emitting part of the loudspeaker directly faces the listener. This direct path from loudspeaker to listener, without sound reflections from the wall, is referred to as the direct path and is the first soundwave coming in. An example of this can be found in Figure 3.3a, where the first non-zero peak is the direct path. Before the SRTF was added, the loudspeakers are modelled as omnidirectonal emitters, meaning that all outgoing directions receive the same energy from the sound source. Including the SRTF thus resulted in a relative energy decrease for the reflections compared to the direct path, resulting in a channel with a more impulse like shape. In the example of Figure 3.3a, this would result in a decrease of the amplitude of the signal after the direct path. The impulse like shape results in a flat frequency response, which is beneficial for the inversion. The condition number shows this by means of the improvement after the addition of the SRTF.

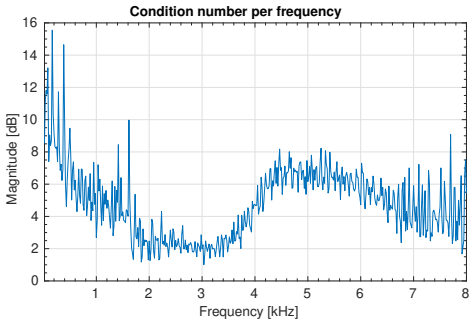
Finally, adding more loudspeakers to the setup makes a clear difference as is shown in Figure 3.2d. Adding more loudspeakers to the setup gives more options and freedom to obtain the desired response which is reflected in a smaller condition number, corresponding to more stable inversion.



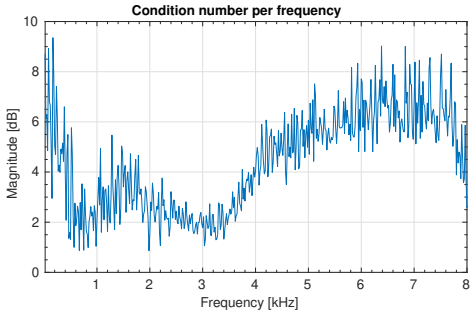
(a) Condition number per frequency bin with no HRTF, no SRTF and only 2 loudspeakers.



(b) Condition number per frequency bin with HRTF, no SRTF and only 2 loudspeakers.

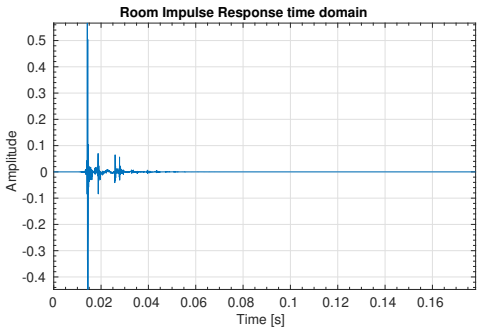


(c) Condition number per frequency bin with HRTF, SRTF and only 2 loudspeakers.

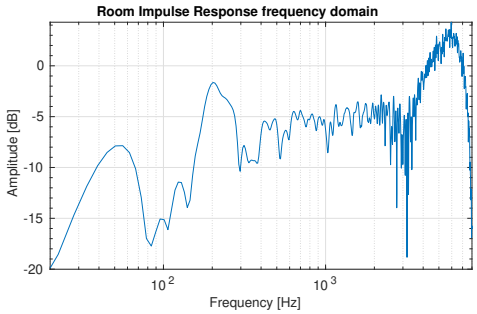


(d) Condition number per frequency bin with HRTF, SRTF and 4 loudspeakers.

Figure 3.2: Condition number per frequency bin in different situations. In these figures we subsequently add the Head Related Transfer Function (HRTF), the Speaker Related Transfer Function (SRTF) and more speakers to the model that at first consists of the Room Impulse Response (RIR) and two speakers. Take note of the different amplitude axis definitions in these figures.



(a) Example of a RIR in time domain.



(b) Example of a RIR in frequency domain.

Figure 3.3: Example of a Room Impulse Response (RIR) computed using the image source method. Code to generate these figures is provided by [57].

3.2.2. Room impulse response

The Room Impulse Response (RIR) describes at what time delay, with which amplitude and with what response each sound reflection path reaches the listener. This RIR is causing the most trouble when trying to invert the channel. Looking at the frequency domain signal of a generic RIR, as the one given in Figure 3.3b, it is clear that it contains a lot of low valued samples. These low valued samples are amplified by the inversion causing noise and out of the ordinary sound behaviour. To properly invert the channel, some specific frequencies have to be amplified substantially to properly cancel the channel. In practise this results in strange sounding audio causing some frequencies to explode and others to vanish.

In theory, the RIR is an infinitely long response in time but in practise the RIR is considered truncated at the τ_{60} time [58]. In figure 3.3a an example RIR is shown cut off at the τ_{60} time. The τ_{60} time is defined as the time at which the amplitude of the RIR is decayed 60 dB with respect to the maximum magnitude. The τ_{60} time in a regular living room we consider is about 200 ms, which translates to $0.2 * 16000 = 3200$ samples given our sampling frequency of 16 kHz. This size results in a computational complex inversion.

3.2.3. Non-personal head related transfer function

As also mentioned in Section 2.1.2, the Head Related Transfer Function (HRTF) is the response describing how the human body and, primarily, the ears alter an incoming sound wave. This response depends on the direction the sound wave comes from and this direction dependency is used for the auditory localization. The HRTF is considered personal, any two people have clearly different HRTF's and listening to someone else's HRTF generally leads to wrong auditory cues, presenting the listener with a wrong sound localization experience. As mentioned before, a generalized HRTF is used in the model based on the general KEMAR head model. Even though this is a general head model, it is far from ideal. The perceptual cues that the response is optimized for might be invalid for the listener, generally leading to a wrong experience.

3.2.4. Errors in location and orientation of loudspeakers and listener

When optimizing the loudspeaker output to generate the desired response at both ears, the position and orientation of all the loudspeakers and the listener are assumed to be known. These positions are used to determine all the reflections and

calculate all the outgoing and incoming angles. Small differences between the known position and the actual physical position lead to substantial errors in the responses [10, 59]. The spikes that appear in the RIR, as shown in Figure 3.3, correspond to a soundwave that has been reflected from one or multiple walls. When the loudspeaker or receiver position slightly changes, the amplitude and arriving time of these reflections generally change considerably with random behaviour [60]. This change of the reflections results in a wrong response being used in the optimization procedure. The unpredictability and randomness of the result of these small errors could have detrimental effects on the optimization.

3.2.5. General errors in models

Apart from the above mentioned specific problems, model simplifications with respect to the practical scenario also contribute to errors in the found crosstalk cancellation solution. One of these simplifications is the assumption that a room is perfectly rectangular and also empty, which in practise is not correct and also contradicts the purpose of this thesis. Another assumption is that a loudspeaker is a point source even though they generally consist of multiple drivers and have a non-negligible physical appearance.

4

System model

Since the goal of the thesis is to make a system that works in practise, the system model we define and use throughout the thesis is involved and a lot of factors are considered to approach reality. The advantage of an involved system, apart from being closely related to the physical situation, is that the numerical properties of the model improve with a more involved model, as presented in Section 3.2.1. In our model we consider the listener being located in a (rectangular) room and the reflections of the wall are considered, this is done by means of the Room Impulse Response (RIR). The model consists of a Head Related Transfer Function (HRTF) model that includes all the auditory localization cues (ILD, ITD and HRTF) that are mentioned in Section 2.1. The directivity of the speakers and their sound response is considered by means of the (loud)Speaker Related Transfer Function (SRTF). The model is introduced after which the individual building blocks are further elaborated on and, if applicable, the implementation is discussed. A portion of the system model is already covered in Chapter 3 but this is repeated for sake of completeness.

4.1. Structure and signals of the channel

Here the system model structure is described and it is presented in frequency domain, denoted by $\hat{(\cdot)}$. The response from loudspeaker n_s, where $n_s = 1, ..., N_s$ with N_s the number of loudspeakers, to each ear is described in Equations (4.1) and (4.2) for left and right ear respectively.

$$\hat{\mathbf{y}}_{\mathrm{L},n_s} = \hat{\mathbf{H}}_{n_s} \left(\sum_{n_r} \hat{\mathbf{D}}_{n_s}(\theta_{\mathrm{out},n_r}, \phi_{\mathrm{out},n_r}) \hat{\mathbf{R}}_{n_r}(\mathbf{l}_s, \mathbf{l}_l) \hat{\mathbf{V}}_{\mathrm{L}}(\theta_{\mathrm{in},n_r}, \phi_{\mathrm{in},n_r}) \right) \hat{\mathbf{s}}_{n_s} \tag{4.1}$$

$$\hat{\mathbf{y}}_{\mathrm{R},n_s} = \hat{\mathbf{H}}_{n_s} \left(\sum_{n_r} \hat{\mathbf{D}}_{n_s}(\theta_{\mathrm{out},n_r}, \phi_{\mathrm{out},n_r}) \hat{\mathbf{R}}_{n_r}(\mathbf{l}_s, \mathbf{l}_l) \hat{\mathbf{V}}_{\mathrm{R}}(\theta_{\mathrm{in},n_r}, \phi_{\mathrm{in},n_r}) \right) \hat{\mathbf{s}}_{n_s} \tag{4.2}$$

Where for all vectors $\mathbf{A} = \mathrm{diag}(\boldsymbol{\alpha})$ with $\mathrm{diag}(\cdot)$ denoting a square matrix with the input vector on the diagonal, $\hat{\mathbf{y}}_{\mathrm{L},n_s}$ and $\hat{\mathbf{y}}_{\mathrm{R},n_s}$ describe how the signal originating from the n_s^{th} loudspeaker is received inside the left and right ears, $\hat{\mathbf{s}}_{n_s}$ is the signal fed to the n_s^{th} loudspeaker, $\hat{\mathbf{h}}_{n_s}$ is the response of the n_s^{th} loudspeaker, $n_r = 1, ..., N_r$ with N_r the number of reflections, $\hat{\mathbf{d}}_{n_s}(\theta_{\mathrm{out},n_r}, \phi_{\mathrm{out},n_r})$ is the directivity of the n_s^{th} loudspeaker

for outgoing angles θ_{out,n_r} and ϕ_{out,n_r}, $\hat{\mathbf{r}}_{n_r}(\mathbf{l}_s, \mathbf{l}_l)$ is the response of the n_r^{th} reflection given 3D loudspeaker location $\mathbf{l}_s$ and listener location $\mathbf{l}_l$ and $\hat{\mathbf{v}}_{\text{L}}(\theta_{\text{in},n_r}, \phi_{\text{in},n_r})$ and $\hat{\mathbf{v}}_{\text{R}}(\theta_{\text{in},n_r}, \phi_{\text{in},n_r})$ are the HRTF's given the angles of incidence θ_{in,n_r} and ϕ_{in,n_r} for the left and right ear.

The vectors applied in Equation (4.1) and (4.2) are all of size $\mathbb{C}^{N_b \times 1}$ and are also zero-padded prior to calculations if necessary. The full size of these variables are given in Equation (4.3).

$$\begin{aligned} &\hat{\mathbf{Y}}_{\text{L}},\ \hat{\mathbf{Y}}_{\text{R}},\ \hat{\mathbf{S}},\ \hat{\mathbf{H}} \in \mathbb{C}^{N_b \times N_s} \\ &\hat{\mathbf{D}}(\theta_{\text{out},n_r}, \phi_{\text{out},n_r}),\ \hat{\mathbf{R}}(\mathbf{l}_s, \mathbf{l}_l),\ \hat{\mathbf{V}}_{\text{L}}(\theta_{\text{in},n_r}, \phi_{\text{in},n_r}),\ \hat{\mathbf{V}}_{\text{R}}(\theta_{\text{in},n_r}, \phi_{\text{in},n_r}) \in \mathbb{C}^{N_b \times N_s \times N_r} \end{aligned} \tag{4.3}$$

Figure 4.1 shows what the variables correspond to given a few reflection examples.

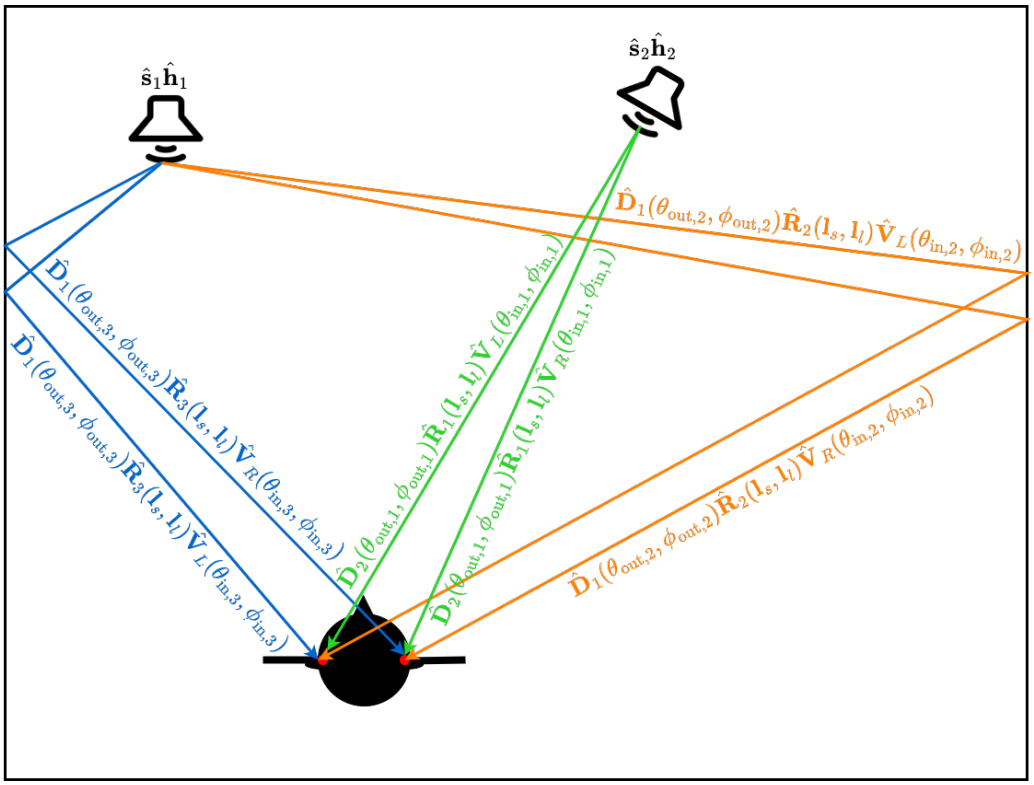


Figure 4.1: Example of a few reflections that happen in a room. The direct path and two single reflections are shown in the figure, denoted by reflections $n_r = 1, 2, 3$. The indices in the responses show how the contributions of every path are determined.

For convenience, we summarize all the terms contributing to the perceived channel in the variables $\hat{\mathbf{c}}_{\text{L},n_s} \in \mathbb{C}^{N_b \times 1}$ and $\hat{\mathbf{c}}_{\text{R},n_s} \in \mathbb{C}^{N_b \times 1}$ (or in general, $\hat{\mathbf{C}}_{\text{L}} \in \mathbb{C}^{N_b \times N_s}$ and $\hat{\mathbf{C}}_{\text{R}} \in \mathbb{C}^{N_b \times N_s}$), which represent the channel responses from the n_s^{th} loudspeaker to the left and right ear respectfully. This results in Equations (4.4) and (4.5). Note that this notation comes in useful since the goal of the thesis is to optimize the speaker output of all the loudspeakers.

$$\hat{\mathbf{y}}_{\text{L},n_s} = \hat{\mathbf{C}}_{\text{L},n_s}\hat{\mathbf{s}}_{n_s} \tag{4.4}$$

$$\hat{\mathbf{y}}_{\text{R},n_s} = \hat{\mathbf{C}}_{\text{R},n_s}\hat{\mathbf{s}}_{n_s} \tag{4.5}$$

The total response found at both ears given all the loudspeakers in the system is given by Equations (4.6) and (4.7) with total responses $\hat{\mathbf{y}}_{\mathrm{L}} \in \mathbb{C}^{N_b \times 1}$ and $\hat{\mathbf{y}}_{\mathrm{R}} \in \mathbb{C}^{N_b \times 1}$ for left and right ear respectfully.

$$\hat{\mathbf{y}}_{\mathrm{L}} = \sum_{n_s} \hat{\mathbf{y}}_{\mathrm{L},n_s} = \sum_{n_s} \hat{\mathbf{C}}_{\mathrm{L},n_s} \hat{\mathbf{s}}_{n_s} \tag{4.6}$$

$$\hat{\mathbf{y}}_{\mathrm{R}} = \sum_{n_s} \hat{\mathbf{y}}_{\mathrm{R},n_s} = \sum_{n_s} \hat{\mathbf{C}}_{\mathrm{R},n_s} \hat{\mathbf{s}}_{n_s} \tag{4.7}$$

In the following sections, the parts of the model are further elaborated.

4.2. Room impulse response and reflections

The Room Impulse Response (RIR) is the transfer function from a single point source (loudspeaker) in the room to a listener point in space caused by the room and its sound characteristics. This response is composed of the direct path response (shortest path from source to listener) and all the reflections from one or multiple walls or objects. The longer the sound travels in the room the more the amplitude decreases with $A \propto \frac{1}{\alpha^2}$ with A the amplitude and α the distance travelled by the sound wave. When the sound travels longer, the response is also perceived later in time. This information is included in the RIR and also the sound energy decay when the wave reflects from a wall is included. In the end, the RIR thus shows all the weighted, time-delayed direct-path and reflections happening in a room given a single loudspeaker and listener location pair. An example is given in Figure 4.2. Please note that currently the RIR is only representing the amplitude of the incoming response, no direction information can be found in the response thus far. Later in this chapter we add this information and thus create a so-called Binaural Room Impulse Response (BRIR) [61, 62].

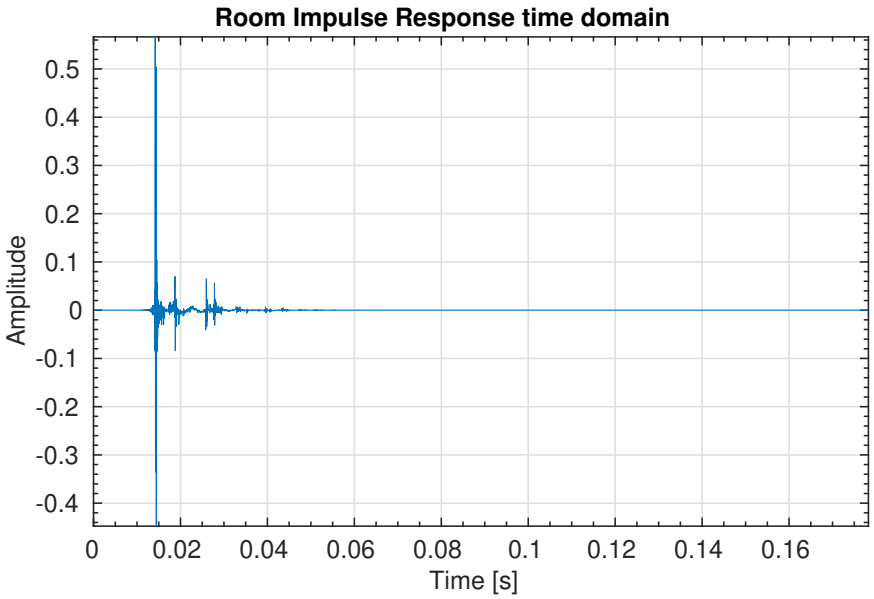


Figure 4.2: Example of a Room Impulse Response computed using the image source method. All the peaks in the response represent one reflection path. Code to generate this figure is provided by [57].

For physical rooms, it is best to measure the RIR [63–65]. In the special case of an empty rectangular room, the RIR can be modelled and simulated efficiently and

with good precision using the so-called image-source method [60], as is also used to obtain Figure 4.2.

In the model, the assumption is made that we have an empty rectangular room to obtain the RIR required and thus we use the image source method. For now, the assumption is made that this simple case is a good enough approximation of a real physical room but a new model is proposed in the to be published paper found in Appendix A.

Figure 4.3 shows a small scale example of the image source method in 2D. In the figure, the white rectangle in the middle represents the physical room including the receiver and the physical loudspeaker. The black arrow indicates the sound path of the direct path response from physical loudspeaker to physical receiver. The other gray shaded rooms represent the functioning of the image-source method, they are folded versions of the original room and loudspeaker to make it easier to determine the reflection paths. Because of these virtual rooms, an actual reflection path with properly placed wall reflections and proper angles does not have to be determined. A straight arrow from the loudspeaker in the virtual room to the physical receiver gives all the information required. The reflection sound paths are represented by the coloured arrows and as can be seen from the figure, the distance the wave travels and the amount of walls it passes can be easily determined.

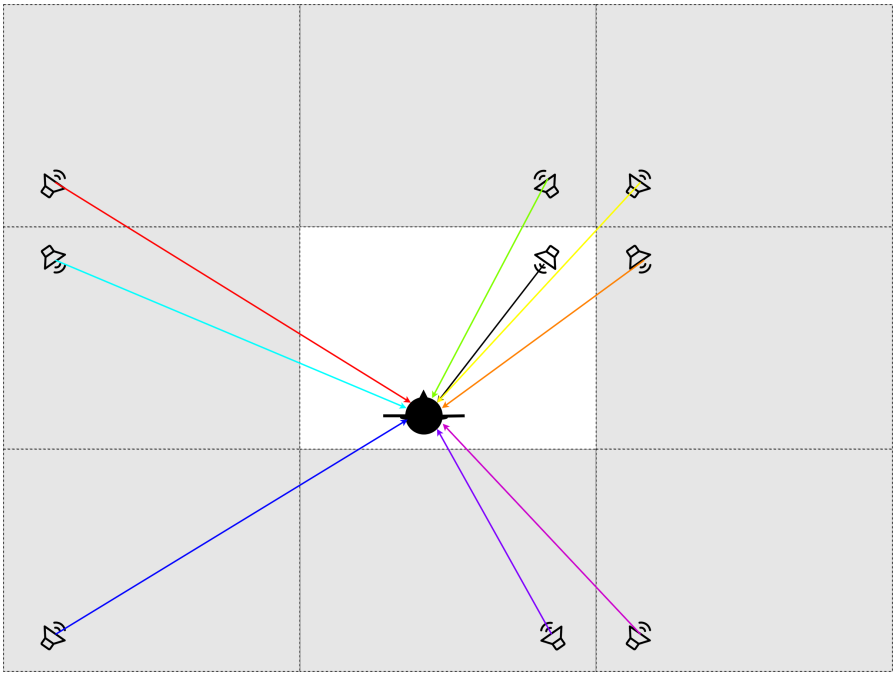


Figure 4.3: Representation of the image-source method. The white rectangle in the middle of the picture represents the physical room including the receiver and loudspeaker. The gray rectangles around this room are folded versions of the original room that make the computation of reflections paths straight-forward. The coloured arrows indicate the sound paths of reflected sound waves. As can be seen, due to the folded rooms, it is easy to determine the traveled distance of the sound wave and also the number of walls it reflected from.

To compute the RIR's using the image-source method, the Room Impulse Response Generator given in [57] is used. This code can be computed using mex to be used in MATLAB and uses the image source method to efficiently compute the responses. The original source code is modified to give more outputs that are mentioned throughout this chapter. Other modifications are applied to be able to do multi-thread computation for increased processing speed and to be able to handle

multiple speaker and listener locations as input.

To allow for all the factors in the model to be represented and implemented as desired, the most important output, the response, is also changed. Originally, the entire response would be returned including all reflections. Since we wish to modify each reflection path (all the arrows in Figure 4.3) and their response individually, all the reflections, their start time and the response itself are returned individually. The individual reflection responses, as denoted by $\hat{\mathbf{R}}_{n_r}(\mathbf{l}_s, \mathbf{l}_l)$, are constructed by means of a Hanning windowed sinc function with proper delay and amplitude. The delay and amplitude of the responses are determined by the distance between loudspeaker location $\mathbf{l}_s$ and listener location $\mathbf{l}_l$ and also the reflection coefficients of each wall the sound wave reflects from. [57] provides more details on the exact implementation. Each reflection can now be processed individually after which they are added together to construct the full response.

4.3. Head related transfer function model

To make the model for human listeners instead of, for instance, microphones, the influence of the human body and its ears are included. This is done by adding the Head Related Transfer Function (HRTF) to the system model which includes all the auditory cues discussed in Section 2.1.

The HRTF response transforms a single reflection path into two paths to the left and right ear. Since the measurements of the HRTF are made at 1 meter distance [12], applying the HRTF to the response is depicted as given in Figure 4.4. Do note that this is not a depiction of the path the sound waves travel but merely a depiction of the way in which the HRTF influences the response. This response is not equivalent to the actual response path to the left and right ear, as is depicted in Figure 4.1. Literature shows however that the response depicted in Figure 4.4 is a sufficient approximation of the actual reflection paths [34] and this response is thus considered in the model.

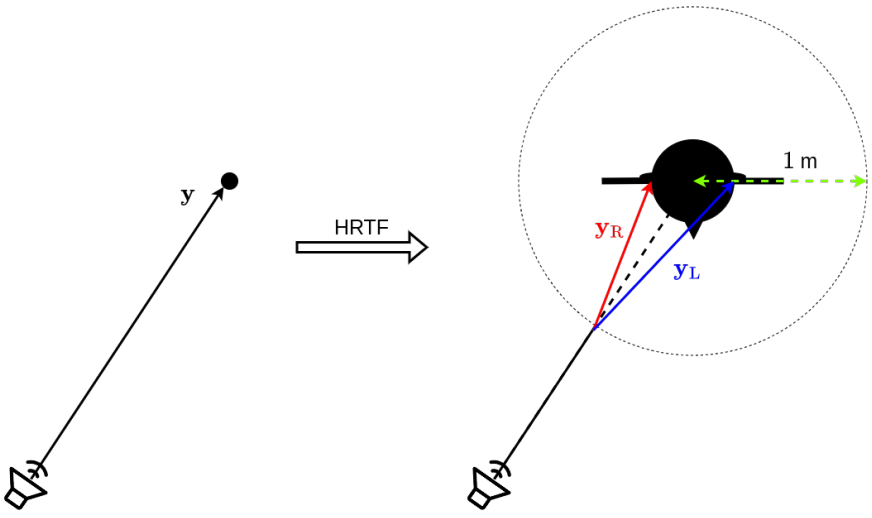


Figure 4.4: The influence of the HRTF on the response. Note that the paths depicted in the right picture do not correspond to the sound wave path but to the precise way the HRTF influences the response. The reason for this is that the HRTF is measured at 1 meter distance.

To apply the HRTF responses $\hat{\mathbf{V}}_\mathrm{L}(\theta_{\mathrm{in},n_r}, \phi_{\mathrm{in},n_r})$ and $\hat{\mathbf{V}}_\mathrm{R}(\theta_{\mathrm{in},n_r}, \phi_{\mathrm{in},n_r})$ to the channel responses $\hat{\mathbf{C}}_{\mathrm{L},n_s}$ and $\hat{\mathbf{C}}_{\mathrm{R},n_s}$ respectively, (as done in Equations (4.1) and (4.2)) the incoming angle of the sound wave must be determined. Due to the image-source

method as depicted in Figure 4.3, it is straightforward to calculate the azimuth angle θ_{in,n_r} and elevation angle ϕ_{in,n_r} for each n_r^{th} reflection. The 3D position of the sources is denoted by $\mathbf{l}_{s,n_r}$ and $\mathbf{l}_l$ denotes the position of the listener. The corresponding azimuth and elevation angle of the incoming sound wave are given in Equation (4.8), where atan2$(\cdot)$ denotes the 2-argument arctangent function, mod$(\cdot)$ is the modulo function and $\mathbf{d}_{n_r} = \mathbf{l}_{s,n_r} - \mathbf{l}_l$. Given these angles, the correct HRTF response can be chosen and applied to the response.

$$\begin{aligned}
\theta_{\text{in},n_r} &= \text{mod}\left(\text{atan2}\left(d_{n_r}(y), d_{n_r}(x)\right) + 360°, 360°\right) \\
\phi_{\text{in},n_r} &= \text{atan2}\left(\sqrt{d_{n_r}^2(x) + d_{n_r}^2(y)}, d_{n_r}(z)\right) \\
\text{atan2}(y,x) &= \begin{cases} \text{atan}\left(\frac{y}{x}\right), & x > 0 \\ \text{atan}\left(\frac{y}{x}\right) + 180, & x < 0 \text{ and } y \geq 0 \\ \text{atan}\left(\frac{y}{x}\right) - 180, & x < 0 \text{ and } y < 0 \\ 90, & x = 0 \text{ and } y > 0 \\ -90, & x = 0 \text{ and } y < 0 \\ \text{undefined}, & \text{else} \end{cases}
\end{aligned} \tag{4.8}$$

The used HRTF's are based on a database of measurements given by [12]. This database provides measurements $360°$ in azimuth angle and $120°$ in elevation angle, both have a step size of $1°$. The measurements are performed on the KEMAR head model. Even though this makes a very well defined HRTF description with 43560 measurement points, it is still not a continuous set. To obtain the required continuous set, the measurement points are linearly interpolated and extrapolated.

4.4. Speaker directivity and transfer function

Recent studies have shown that the directivity of the loudspeakers has a substantial impact on the auditory localization of that source [66–68]. Because of this, we wish to add the directivity of the speakers to the model. Directivity data is available for a set of commercial loudspeakers in the database provided by [69]. For the simulations, the directivity data on the KEF LS50 bookshelf loudspeakers is used. Due to the KEF's affordable price, high quality audio and modern and compact design, the KEF's fit the goal of the thesis well. The KEF LS50 loudspeaker, that is depicted in Figure 4.5, consists of a coaxial driver meaning that the sound emitted by the tweeter and the woofer have the exact same origin.

The measurements done for this speaker include a set of measurements along the azimuth circle and along the elevation circle with steps of $5°$. This is little data especially when we require a continuous spherical data set. Interpolating the data is prone to errors but a method to give reasonable results is presented in [71]. This method is used to obtain a continuous data set on the directivity of the KEF LS50 loudspeakers.

To apply the correct directivity response $\hat{\mathbf{D}}_{n_s}(\theta_{\text{out},n_r}, \phi_{\text{out},n_r})$ to the channel responses $\hat{\mathbf{C}}_{\text{L},n_s}$ and $\hat{\mathbf{C}}_{\text{R},n_s}$, (as done in Equations (4.1) and (4.2) respectively) the outgoing angles of the sound wave from the loudspeaker must be determined. Calculating the outgoing azimuth angle θ_{out,n_r} and elevation angle ϕ_{out,n_r} is, unlike determining θ_{in,n_r} and ϕ_{in,n_r}, not a straightforward task. As can be seen in Figure 4.3, with every folded room, the loudspeaker direction also folds which proves to be difficult to track. To obtain a computationally efficient and elegant way to correctly compute the angles, quaternions are used. Quaternions are an extension to imaginary numbers by adding two additional imaginary parts, they are applied mainly in 3D graphics and computations [72]. Basic theory on quaternions and its use in

Figure 4.5: KEF LS50 loudspeakers. These loudspeakers and their directivity measurements are considered in the system model. The loudspeakers only consist of one coaxial driver which consists of one tweeter and one woofer. Figure adopted from [70].

the implementation of the system model is presented in Appendix C.1. The theory in the appendix provides a method to compute the angles θ_{out,n_r} and ϕ_{out,n_r} for all reflection paths n_r.

Given the outgoing sound wave angle from the loudspeakers and the continuous loudspeaker directivity data set, the loudspeaker and its characteristics can be added to the model and the channel response. The response corresponding to the directivity of the loudspeaker is referred to as the (loud)Speaker Related Transfer Function) (SRTF).

5

Proposed algorithm

The CrossTalk Cancellation (CTC) problem as discussed in chapter 3 gives theoretically correct results but a practical use for the current implementation is very unlikely. To solve the practical issues, we wish to optimize for perceptual measures instead of the objective measure found in the original CTC problem. In this chapter, the steps towards the final proposed algorithm which includes the perceptual measures are discussed. First, the auditory localization validation metric is discussed. This metric is used to support the choices made in the derivation of the proposed algorithm. The metric definition is followed by the evaluation of the original CTC algorithm using this metric. After this, two relatively simple algorithms are discussed based on the findings in Chapters 2 and 3. These algorithms provide slightly better performance than the original CTC solution but do not offer enough improvements to be applicable in a practical scenario. These simple algorithms serve as an introduction to the building blocks required for the final algorithm. This more involved algorithm gives slightly worse results compared to the simple algorithms but also comes with some advantages that are discussed throughout this and the following chapters.

5.1. Simulation setup and interaural cross-correlation response as validation metric

The method of characterizing the performance of the algorithms is the size of the sweetspot. As mentioned in Section 1.3, the sweet spot is defined as the region in space in which the illusion we wish to create is sufficiently present. Before diving into the algorithms, we first specify the definition of the sweet spot and show how it is determined.

As stated in Section 2.1.1, the time index of the peak in the InterAural Cross-Correlation (IACC) is a measure that reliably shows the Interaural Time Difference (ITD) interpretation of the auditory system. This is the reason the IACC is used to determine the size of the sweet-spot. The sweet spot is defined as the largest sphere in space, in which for all points in the sphere, the peak of the IACC is found at the time index corresponding to the correct ITD cue. In here, a point in space

represents the centre of the head. This metric is used throughout this chapter to determine the performance of the algorithms, the bigger the sweet spot, the better the performance. We assume that a better performance indicates that the system is more stable and more robust against the errors and inaccuracies described in Chapter 3. The practical implementation of the validation metric becomes more clear later on.

The simulation setup is the same for all the proposed algorithms to obtain a clear and valid comparison between them. The setup is placed in a room with size $4.5 \times 4.4 \times 3$ m with the centre of the listeners' head positioned at $[2.2, 2.1, 1.3]$ m. The reflection coefficients for the walls are chosen as $\boldsymbol{\beta} = [\beta_{x_1}, \beta_{x_2}, \beta_{y_1}, \beta_{y_2}, \beta_{z_1}, \beta_{z_2}] = [0.4, -0.15, 0.15, -0.6, 0.6, -0.4]$, where the average over all reflection coefficients is zero. This results in an average of the RIR that is zero which simulates practical scenarios best [60]. For computational purposes, the τ_{60} is chosen as a constant instead of a variable and is assumed to be 0.17 s. The setup consists of four loudspeakers roughly placed in the corners of a rectangular room, with locations $[0.8, 1.1, 1.5]$ m, $[3.8, 0.8, 1.4]$ m, $[4.2, 4.0, 1.5]$ m and $[1.1, 4.1, 1.2]$ m. The goal is to create a virtual source to the left of the listener at location $[0.7, 2.3, 1.4]$ m. The exact setup is depicted in Figure 5.1.

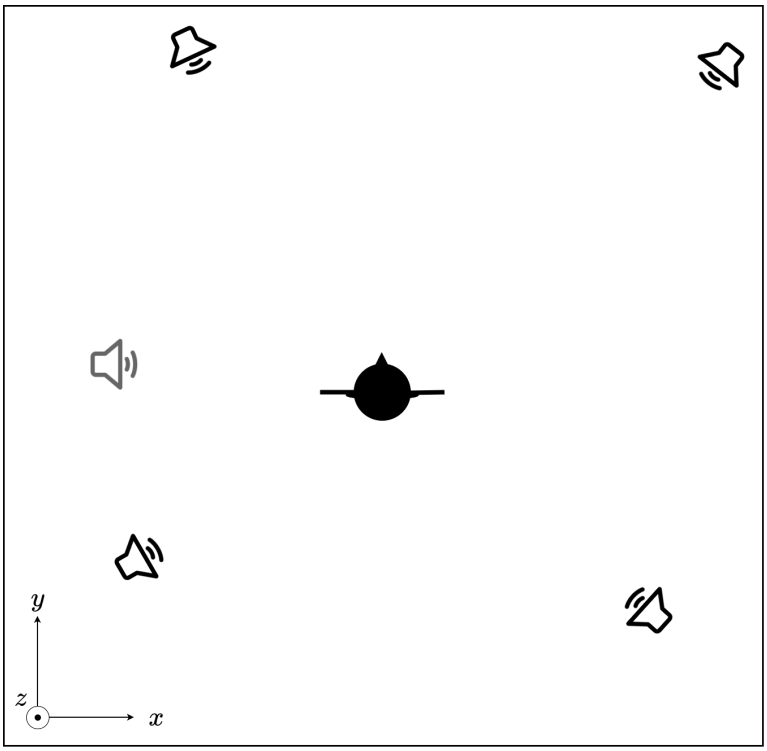
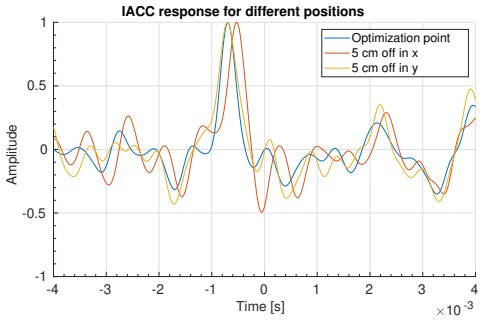
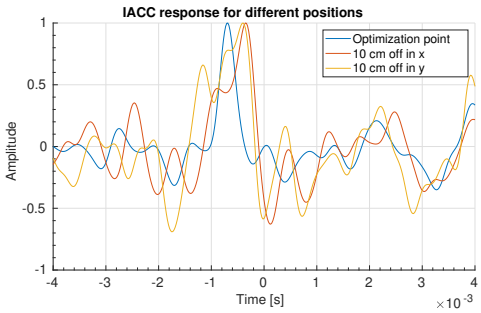


Figure 5.1: Simulation setup used to validate the algorithms. The loudspeakers roughly placed in the corners of the room are the physical loudspeakers and the to be created virtual source is the shaded loudspeaker to the left of the listener. For details see text.

In the simulation, the "gong©" audio from MATLAB (R2022b) is used as audio stimulus. A sampling frequency of $f_s = 16$ kHz is used for the simulation instead of a higher sampling frequency, for example $f_s = 44.1$ kHz, as is usual the case for audio processing. The reason for this is that it greatly reduces computational expenses without loss of validity since nearly all the auditory localization is performed with frequency content under 8 kHz as stated in Section 2.1. The speed of sound is assumed to be $v_s = 342$ m/s.



(a) The IACC response for the optimization point and two points on a 5 cm radius sphere around this point. The IACC response shows that the the illusion is sufficiently present at this point.



(b) The IACC response for the optimization point and two points on a 10 cm radius sphere around this point. The IACC response shows that the the illusion is not present at this point.

Figure 5.2: IACC responses found for original CTC optimization results. The figures show that the sweet spot size is small for the original CTC problem. Do note that the optimization is not performed for the off-positions, only the IACC response is evaluated here.

5.2. Crosstalk cancellation performance

To be able to compare the performance of the new algorithms with the original CrossTalk Cancellation (CTC) problem, its performance in terms of the validation metric is determined. The original CTC optimizes for two points in space, the ears, or, equivalently, one point in space representing the centre of the head. This simple solution, as already presented in Section 3.1, is given by Equation (5.1) where $\hat{\mathbf{y}}^*$ denotes the desired solution for left and right ear, $\hat{\mathbf{C}}$ is the channel matrix and $\hat{\mathbf{s}}$ the signal fed to the loudspeaker with $\hat{\mathbf{s}}^*$ the optimal solution (For variable definitions and explanation see Section 3.1).

$$\hat{\mathbf{s}}^* = \arg\min_{\hat{\mathbf{s}}} ||\hat{\mathbf{y}}^* - \hat{\mathbf{C}}\hat{\mathbf{s}}||_2^2 \tag{5.1}$$

This optimization gives a sweet spot size of a small circle around the optimization point. To illustrate the size of the sweet spot, we show the IACC response found at the point of optimization and also two points slightly off the optimization point. The results for two different sets of off optimization points are shown in Figure 5.2. Note that we do not optimize for these off points but only show the found IACC responses.

Figure 5.2a indicates that the sweet spot given a single point CTC solution is valid in a 5 cm radius sphere surrounding this point. Figure 5.2b on the other hand shows that the illusion is not valid anymore 10 cm away from the optimization point. These results led to believe that there exists a sweet spot sphere with a radius between 5 and 10 cm.

Although these results sound promising, they are not. These results would only be valid if there was perfect knowledge on all the attributes of the setup. In a practical scenario we can not assume this and these results are thus an overestimation of the practical performance.

Increasing the robustness and practical applicability of the system can be interpreted as increasing the sweet spot size so that hopefully the practical sweet spot is large enough for the illusion to hold. The CTC algorithm is modified in the following to increase this sweet spot size.

5.3. Multi-point crosstalk cancellation

One possible way of increasing the sweet spot size is optimizing for multiple sweet spots in close proximity. A similar implementation is done in [73] for multiple sound zones. The hypothesis here is that multiple smaller sweet spots close to each other combine to one bigger sweet spot. The implementation is done by placing one optimization point at the centre of the head (as was the case in the original CTC problem) and surrounding it by 6 optimization points on a sphere with a certain radius. The scenario is illustrated in Figure 5.3 and Equation (5.2) shows the new optimization problem.

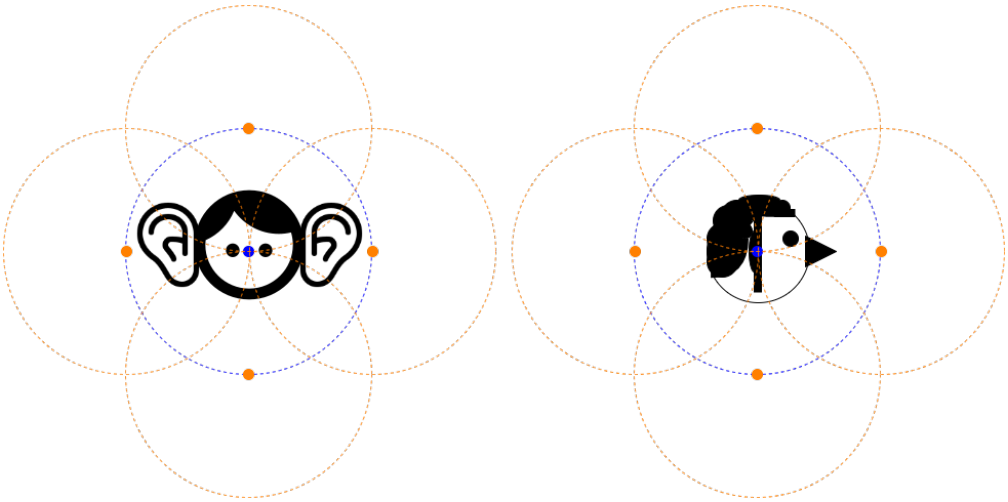


Figure 5.3: Illustration of the 7 optimization points in multi-point crosstalk cancellation. The colored dots indicate the position of the optimization points and the dotted circles indicate the corresponding sweet spot sizes. The colours in the figure correspond to the ones given in Equation (5.2).

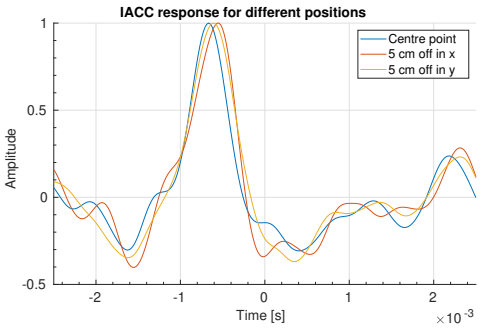
$$\hat{\mathbf{s}}^* = \arg\min_{\hat{\mathbf{s}}} ||\hat{\mathbf{y}}_1^* - \hat{\mathbf{C}}_1\hat{\mathbf{s}}||_2^2 + ||\hat{\mathbf{y}}_2^* - \hat{\mathbf{C}}_2\hat{\mathbf{s}}||_2^2 + \cdots + ||\hat{\mathbf{y}}_7^* - \hat{\mathbf{C}}_7\hat{\mathbf{s}}||_2^2 \tag{5.2}$$

In the figure and equation, the blue parts represent the original CTC setup and the orange parts are the added optimization points. As is shown in the figure, the extra optimization points are placed such that they are on the border of the original CTC sweet spot placed on the crossings with the positive and negative x-, y- and z-axis.

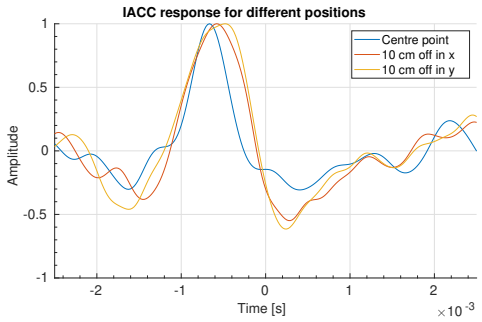
To test the performance of the algorithm, we assume an original CTC sweet spot sphere with radius 6 cm and place the extra points accordingly. Similar results as presented in Figure 5.2 for the new optimization are presented in Figure 5.4.

Comparing the results found in Figures 5.2a and 5.4a, we can see that the multi-point optimization makes the peaks in the IACC more profound and the response more similar to the desired result found at the centre optimization point. The best improvement can be found when comparing Figures 5.2b and 5.4b. Here we see that the IACC peaks for the latter are found closer to the correct time delay compated to the original one-point CTC solution. Based on this and similar proof the conclusion is drawn that the multi-point algorithm increases the sweet spot size compared to the original one-point CTC solution.

This method can be extended by including more optimization points to further increase the sweet spot size, but limitations are found relatively quickly. Adding more optimization points leads to less pronounced peaks in the IACC response wich



(a) The IACC response for the centre optimization point and two points on a 5 cm radius sphere around this point. The IACC responses show that the illusion is clearly present at these points.



(b) The IACC response for the centre optimization point and two points on a 10 cm radius sphere around this point. The IACC responses show that the illusion is still present at these further away points.

Figure 5.4: IACC responses found for multi-point CTC optimization results. The figures show that the sweet spot size is substantially larger compared to the original CTC sweet spot size. Do note that the optimization is not performed for the off-positions, only the IACC response is evaluated here.

eventually results in a vanishing illusion. On top of that, adding more points drastically increases the computational complexity of the problem with bad scalability in the 3D field.

All in all, the multi-point optimization leads to increased performance at the cost of computational complexity. Due to limited scalability, only a small increase in sweet spot size is realistic and practical. More improvements to the CTC solution are required to achieve a substantial increase in sweet spot size resulting in valid illusions in practical scenarios.

5.4. Optimizing for the interaural crosscorrelation

In the previous approach, the entire signal $\hat{\mathbf{s}}$ is optimized such that it matches the desired response $\hat{\mathbf{y}}$ as closely as possible. The primary goal of the algorithms is, however, not to obtain the best signal reproduction (including the spatial cues) but to present the desired spatial audio cues to the listener while not noticeably deterring the audio quality. By focusing more on the primary goal of the algorithm and treating the audio quality as a secondary constraint, a more robust and efficient solution might be found.

An implementation of this can be found in the validation measure, the InterAural CrossCorrelation (IACC). Optimizing directly for the IACC response instead of the entire response might improve the robustness and effectiveness of the solution and thus increase the sweet spot size.

As mentioned in Section 2.1.1, the IACC is defined as the cross correlation of the response found at the left and right ear, as given in Equation (2.2). First, we are going to write the cross correlation in a more applicable form, as presented in Equation 5.3.

$$y_{\mathrm{IACC}}(\bar{n_b}) = \frac{\sum_{n_b=1}^{N_b} y_{\mathrm{L}}(n_b) y_{\mathrm{R}}(n_b - \bar{n_b})}{\sqrt{\mathbf{y}_{\mathrm{L}}{}^{\mathrm{T}}\mathbf{y}_{\mathrm{L}}\mathbf{y}_{\mathrm{R}}{}^{\mathrm{T}}\mathbf{y}_{\mathrm{R}}}} \rightarrow \mathbf{y}_{\mathrm{IACC}} = \frac{\mathbf{Y}_{\mathrm{L}}\mathbf{F}\mathbf{y}_{\mathrm{R}}}{\sqrt{\mathbf{y}_{\mathrm{L}}{}^{\mathrm{T}}\mathbf{y}_{\mathrm{L}}\mathbf{y}_{\mathrm{R}}{}^{\mathrm{T}}\mathbf{y}_{\mathrm{R}}}} = \frac{\mathbf{Y}_{\mathrm{R}}\mathbf{F}\mathbf{y}_{\mathrm{L}}}{\sqrt{\mathbf{y}_{\mathrm{L}}{}^{\mathrm{T}}\mathbf{y}_{\mathrm{L}}\mathbf{y}_{\mathrm{R}}{}^{\mathrm{T}}\mathbf{y}_{\mathrm{R}}}} \tag{5.3}$$

$$\mathbf{F} = \begin{bmatrix} 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 1 & 0 \\ \vdots & \vdots & \iddots & \vdots & \vdots \\ 0 & 1 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \end{bmatrix} \tag{5.4}$$

In here, $\mathbf{y}_{\text{IACC}} \in \mathbb{R}^{2N_b-1\times 1}$ is the IACC with $\bar{n_b} = 1, ..., 2N_b-1$ and $n_b = 1, ..., N_b$ with N_b the length of $\mathbf{y}_{\text{L}}$ and $\mathbf{y}_{\text{R}}$ and $\mathbf{F} \in \mathbb{R}^{N_b\times N_b}$, as defined in Equation (5.4), represents the exchange matrix that flips the entries of a vector. $\mathbf{Y}_{\text{L}} \in \mathbb{R}^{2N_b-1\times N_b}$, and equivalently $\mathbf{Y}_{\text{R}}$, represents the toeplitz matrix of the left ear response defined in Equation (5.5) that is used to implement the convolution.

$$\mathbf{Y}_{\text{L}} = \begin{bmatrix} y_L(1) & 0 & 0 & \cdots & 0 \\ y_L(2) & y_L(1) & 0 & \cdots & 0 \\ y_L(3) & y_L(2) & y_L(1) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_L(N_b-1) & y_L(N_b-2) & y_L(N_b-3) & \cdots & 0 \\ y_L(N_b) & y_L(N_b-1) & y_L(N_b-2) & \cdots & y_L(1) \\ 0 & y_L(N_b) & y_L(N_b-1) & \cdots & y_L(2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & y_L(N_b) \end{bmatrix} \tag{5.5}$$

The interesting property of the IACC we wish to optimize for is the time-delay corresponding to the peak of the IACC response. As mentioned in Section 2.1.1, this time-delay, denoted by τ_{IACC}, is derived by means of Equation (2.3) and repeated in Equation (5.6).

$$\tau_{\text{IACC}} = \arg\max_{\bar{n_b}} \quad y_{\text{IACC}}(\bar{n_b}), \qquad \text{for} \quad \bar{n_b} \in [-1, 1] \text{ ms} \tag{5.6}$$

With these (revisited) definitions given, its application to the usecase can be discussed.

Since the location of the virtual loudspeaker is known, the desired τ_{IACC} is known or can be easily calculated prior to running the algorithm. For now, we assume it to be known and denoted by τ^*_{IACC}. This gives us the optimization problem given in Equation (5.7), where we omitted the normalization term since we are not interested in the amplitude of the IACC peak. In here, $\hat{\mathbf{C}}_{\text{L},n_s} = \text{diag}(\hat{\mathbf{c}}_{\text{L},n_s})$ and $\hat{\mathbf{C}}_{\text{R},n_s} = \text{diag}(\hat{\mathbf{c}}_{\text{R},n_s})$ with $\text{diag}(\cdot)$ the square matrix with the input vector on the diagonal, $\boldsymbol{\alpha} \in \mathbb{R}^{N_b\times 1}$ and $\mathbf{A} \in \mathbb{R}^{2N_b-1\times N_b}$ are intermediate optimization variables and $\mathbf{W} \in \mathbb{C}^{N_b\times N_b}$ denotes the DFT matrix with $\mathbf{W}^{-1} \in \mathbb{C}^{N_b\times N_b}$ its inverse, the IDFT. Note that the size of the optimization problem is mainly caused by the dependence of the IACC on the loudspeaker signals $\mathbf{s}_{n_s}$ and the corresponding required calculations. The problem can be summarized by: minimizing the IACC peak index error for the IACC calculated by the sum of the signals measured by the ears originating from all loudspeakers.

$$\begin{aligned} \arg\min_{\mathbf{s_1},\mathbf{s_2},\dots,\mathbf{s_{N_S}}} \quad & ||\tau^*_{\text{IACC}} - \tau_{\text{IACC}}||_2 \\ \text{s.t.} \quad & \tau_{\text{IACC}} = \arg\max_{\bar{n_b}\in[-1,1] \text{ ms}} \quad \alpha_{\text{IACC}}(\bar{n_b}) \\ & \boldsymbol{\alpha}_{\text{IACC}} = \mathbf{A}_{\text{L}}\mathbf{F}\boldsymbol{\alpha}_{\text{R}} \\ & \boldsymbol{\alpha}_{\text{L}} = \sum_{n_s} \mathbf{W}^{-1}\hat{\mathbf{C}}_{\text{L},n_s}\hat{\mathbf{s}}_{n_s} \\ & \boldsymbol{\alpha}_{\text{R}} = \sum_{n_s} \mathbf{W}^{-1}\hat{\mathbf{C}}_{\text{R},n_s}\hat{\mathbf{s}}_{n_s} \end{aligned} \tag{5.7}$$

The sole purpose and functioning of this problem is to achieve a desired τ_{IACC}. To make sure the solution results in an audio experience similar to the desired response, we add a constraint to the optimization problem as given in Equation (5.8). In here, a_1 defines the allowed error in terms of the L_2-norm.

$$\begin{aligned}
\arg \min_{\mathbf{s_1},\mathbf{s_2},\ldots,\mathbf{s_{N_s}}} \quad & ||\tau^*_{\text{IACC}} - \tau_{\text{IACC}}||_2 \\
\text{s.t.} \quad & ||\mathbf{y}^*_{\text{L}} - \boldsymbol{\alpha}_{\text{L}}||_2^2 + ||\mathbf{y}^*_{\text{R}} - \boldsymbol{\alpha}_{\text{R}}||_2^2 \leq a_1 \\
& \tau_{\text{IACC}} = \arg \max_{\bar{n_b} \in [-1,1] \text{ ms}} \quad \alpha_{\text{IACC}}(\bar{n_b}) \\
& \boldsymbol{\alpha}_{\text{IACC}} = \mathbf{A}_{\text{L}} \mathbf{F} \boldsymbol{\alpha}_{\text{R}} \\
& \boldsymbol{\alpha}_{\text{L}} = \sum_{n_s} \mathbf{W}^{-1} \hat{\mathbf{C}}_{\text{L},n_s} \hat{\mathbf{s}}_{n_s} \\
& \boldsymbol{\alpha}_{\text{R}} = \sum_{n_s} \mathbf{W}^{-1} \hat{\mathbf{C}}_{\text{R},n_s} \hat{\mathbf{s}}_{n_s}
\end{aligned} \tag{5.8}$$

This problem can also be easily rewritten to have a close resemblance with the original CTC problem as posed in Equation (5.1), it is given in Equation (5.9). In this formulation, the optimization problem is essentially the same as the original CTC problem with the addition of a hard constraint on the desired IACC peak time. Do note here that in case of a near perfect CTC solution in the original problem, the constraints do not have any influence on the optimization outcome, apart from a substantial increase in computational expenses.

$$\begin{aligned}
\arg \min_{\mathbf{s_1},\mathbf{s_2},\ldots,\mathbf{s_{N_s}}} \quad & ||\mathbf{y}^*_{\text{L}} - \boldsymbol{\alpha}_{\text{L}}||_2^2 + ||\mathbf{y}^*_{\text{R}} - \boldsymbol{\alpha}_{\text{R}}||_2^2 \\
\text{s.t.} \quad & \tau^*_{\text{IACC}} = \tau_{\text{IACC}} \\
& \tau_{\text{IACC}} = \arg \max_{\bar{n_b} \in [-1,1] \text{ ms}} \quad \alpha_{\text{IACC}}(\bar{n_b}) \\
& \boldsymbol{\alpha}_{\text{IACC}} = \mathbf{A}_{\text{L}} \mathbf{F} \boldsymbol{\alpha}_{\text{R}} \\
& \boldsymbol{\alpha}_{\text{L}} = \sum_{n_s} \mathbf{W}^{-1} \hat{\mathbf{C}}_{\text{L},n_s} \hat{\mathbf{s}}_{n_s} \\
& \boldsymbol{\alpha}_{\text{R}} = \sum_{n_s} \mathbf{W}^{-1} \hat{\mathbf{C}}_{\text{R},n_s} \hat{\mathbf{s}}_{n_s}
\end{aligned} \tag{5.9}$$

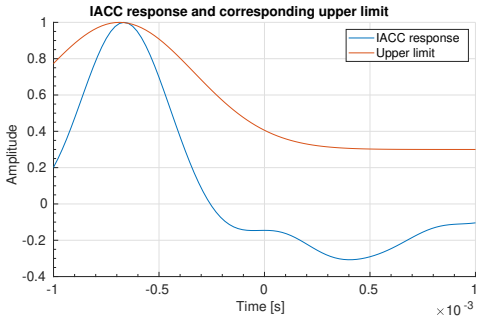
The benefit of the problem found in Equation (5.9) over the one found in Equation (5.8) is that the value of τ_{IACC} is harshly constraint. This indirectly causes the optimization to be forced to have a correct IACC peak before the audio experience is optimized, which is in line with purpose of the algorithm.

To further improve the optimization problem, and, essentially, provide an actual benefit of this optimization problem over the original CTC problem, we also consider the profoundness of the IACC peak. To do this, we analyse the shape of the IACC found for a near perfect CTC solution in the search interval for τ_{IACC}. It is given in Figure 5.5a, which is a zoomed in version of the optimization point response found in Figure 5.2b. In Figure 5.5b an IACC originating from a loudspeaker to the right of the listener is shown.

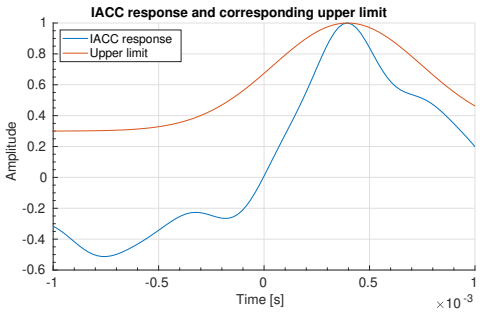
As seen in the figures, a so-called upper limit is plotted above the IACC response. This limit is an empirically determined overestimate of IACC responses. In the optimization, the IACC response should be entirely underneath this curve to force a profound IACC peak at the desired τ^*_{IACC}. The curve is defined in Equation (5.10).

$$\mu_{\text{IACC}}(t) = 0.3 + 0.7e^{-4 \cdot 10^6 (t - \tau^*_{\text{IACC}})^2} \tag{5.10}$$

Adding the upper limit to the optimization problem results in the problem found in Equation (5.11), where c_3 is an empirically determined scaling factor used to tune the impact of this limit.



(a) The IACC response originating from loudspeakers to the left of the listener and the corresponding upper limit.



(b) The IACC response originating from loudspeakers to the right of the listener and the corresponding upper limit.

Figure 5.5: IACC responses and corresponding upper limit.

$$
\begin{aligned}
\arg \min_{\mathbf{s_1},\mathbf{s_2},\ldots,\mathbf{s_{N_s}}} \quad & ||\mathbf{y}_{\mathrm{L}}^{*} - \boldsymbol{\alpha}_{\mathrm{L}}||_2^2 + ||\mathbf{y}_{\mathrm{R}}^{*} - \boldsymbol{\alpha}_{\mathrm{R}}||_2^2 \\
\text{s.t.} \quad & \tau_{\mathrm{IACC}}^{*} = \tau_{\mathrm{IACC}} \\
& \tau_{\mathrm{IACC}} = \arg \max_{\bar{n_b} \in [-1,1] \text{ ms}} \quad \alpha_{\mathrm{IACC}}(\bar{n_b}) \\
& \boldsymbol{\alpha}_{\mathrm{IACC}} \leq c_3 \boldsymbol{\mu}_{\mathrm{IACC}} \\
& \boldsymbol{\alpha}_{\mathrm{IACC}} = \mathbf{A}_{\mathrm{L}} \mathbf{F} \boldsymbol{\alpha}_{\mathrm{R}} \\
& \boldsymbol{\alpha}_{\mathrm{L}} = \sum_{n_s} \mathbf{W}^{-1} \hat{\mathbf{C}}_{\mathrm{L},n_s} \hat{\mathbf{s}}_{n_s} \\
& \boldsymbol{\alpha}_{\mathrm{R}} = \sum_{n_s} \mathbf{W}^{-1} \hat{\mathbf{C}}_{\mathrm{R},n_s} \hat{\mathbf{s}}_{n_s}
\end{aligned}
\tag{5.11}
$$

Solving this problem results in a desired IACC response while maximizing the resemblance of the experienced audio with the desired audio in terms of the L_2-norm. The major downside to this problem is its optimization properties.

In the current form, the problem is non-convex and it is thus difficult to find an optimal solution while also being computationally expensive. This is far from ideal for the real-time audio appliances this algorithm is aimed for. In the following sections, some relaxation methods (approximation methods) are proposed to decrease computational complexity and improve optimization properties.

5.4.1. Channel interaural crosscorrelation optimization

The size of the data vectors has a significant impact on the computational complexity of the optimization problem. Generally speaking, the size of $\mathbf{s}$, is much bigger than the (non zero-padded) size of $\mathbf{c}$, which is roughly τ_{60} long. The channel response contains all the spatial information while the audio fed to the loudspeaker generally does not contain any spatial information. A possible relaxation (or approximation) would be to optimize for the channel response and the corresponding IACC only, after which the found channel response can be convolved with the audio. This drastically decreases the computational complexity of the problem. The corresponding problem is presented in Equation (5.12), where $\mathbf{s}_{n_s}$ now corresponds to the channel optimization variable.

$$
\begin{aligned}
\arg \min_{\mathbf{s_1},\mathbf{s_2},\ldots,\mathbf{s_{N_s}}} \quad & ||\mathbf{c}_{\mathrm{L}}^{*} - \boldsymbol{\alpha}_{\mathrm{L}}||_2^2 + ||\mathbf{c}_{\mathrm{R}}^{*} - \boldsymbol{\alpha}_{\mathrm{R}}||_2^2 \\
\text{s.t.} \quad & \tau_{\mathrm{IACC}}^{*} = \tau_{\mathrm{IACC}} \\
& \tau_{\mathrm{IACC}} = \arg \max_{\bar{n_b} \in [-1,1]\ \mathrm{ms}} \quad \alpha_{\mathrm{IACC}}(\bar{n_b}) \\
& \boldsymbol{\alpha}_{\mathrm{IACC}} \leq c_3 \boldsymbol{\mu}_{\mathrm{IACC}} \\
& \boldsymbol{\alpha}_{\mathrm{IACC}} = \mathbf{A}_{\mathrm{L}} \mathbf{F} \boldsymbol{\alpha}_{\mathrm{R}} \\
& \boldsymbol{\alpha}_{\mathrm{L}} = \sum_{n_s} \mathbf{W}^{-1} \hat{\mathbf{C}}_{\mathrm{L},n_s} \hat{\mathbf{s}}_{n_s} \\
& \boldsymbol{\alpha}_{\mathrm{R}} = \sum_{n_s} \mathbf{W}^{-1} \hat{\mathbf{C}}_{\mathrm{R},n_s} \hat{\mathbf{s}}_{n_s}
\end{aligned}
\tag{5.12}
$$

5.4.2. From non-convex to quadratic program

The optimization problem contains two constraints that are not convex, namely: $\tau_{\mathrm{IACC}} = \arg\max_{\bar{n_b} \in [-1,1]\ \mathrm{ms}} \alpha_{\mathrm{IACC}}(\bar{n_b})$ and $\boldsymbol{\alpha}_{\mathrm{IACC}} = \mathbf{A}_{\mathrm{L}} \mathbf{F} \boldsymbol{\alpha}_{\mathrm{R}}$.

First we rewrite $\tau_{\mathrm{IACC}} = \max_{\bar{n_b} \in [-1,1]\ \mathrm{ms}} \alpha_{\mathrm{IACC}}(\bar{n_b})$ into a similar but linear form. The desired time-delay τ_{IACC}^{*} is known prior to the start of the optimization, we thus know that the maximum value of $\boldsymbol{\alpha}_{\mathrm{IACC}}$ should be at τ_{IACC}^{*} and the other values should be lower than that. This can easily be added to the problem as presented in Equation (5.13). Note that the $\boldsymbol{\alpha}_{\mathrm{IACC}} \leq c_3 \boldsymbol{\mu}_{\mathrm{IACC}}$ constraint is also merged with this. As can be seen, the constraint is now linear.

$$
\begin{aligned}
\arg \min_{\mathbf{s_1},\mathbf{s_2},\ldots,\mathbf{s_{N_s}}} \quad & ||\mathbf{c}_{\mathrm{L}}^{*} - \boldsymbol{\alpha}_{\mathrm{L}}||_2^2 + ||\mathbf{c}_{\mathrm{R}}^{*} - \boldsymbol{\alpha}_{\mathrm{R}}||_2^2 \\
\text{s.t.} \quad & \alpha_{\mathrm{IACC}}(\tau_{\mathrm{IACC}}^{*}) > \alpha_{\mathrm{IACC}}(\bar{n_b}) + c_3(1 - \mu_{\mathrm{IACC}}(\bar{n_b})), \quad \text{for} \quad \bar{n_b} \neq \tau_{\mathrm{IACC}}^{*} \\
& \boldsymbol{\alpha}_{\mathrm{IACC}} = \mathbf{A}_{\mathrm{L}} \mathbf{F} \boldsymbol{\alpha}_{\mathrm{R}} \\
& \boldsymbol{\alpha}_{\mathrm{L}} = \sum_{n_s} \mathbf{W}^{-1} \hat{\mathbf{C}}_{\mathrm{L},n_s} \hat{\mathbf{s}}_{n_s} \\
& \boldsymbol{\alpha}_{\mathrm{R}} = \sum_{n_s} \mathbf{W}^{-1} \hat{\mathbf{C}}_{\mathrm{R},n_s} \hat{\mathbf{s}}_{n_s}
\end{aligned}
\tag{5.13}
$$

The convolution constraint $\boldsymbol{\alpha}_{\mathrm{IACC}} = \mathbf{A}_{\mathrm{L}} \mathbf{F} \boldsymbol{\alpha}_{\mathrm{R}}$ is a more involved to relax and it is not possible to make it convex without loss of accuracy. An approximation of the convolution can be made using the prior knowledge of the desired (channel) response. Currently, we optimize the IACC determined by the optimized response of the left and right ear. We can change this to the IACC determined using the optimized response of left ear and desired right ear and also the optimized response of the right ear and desired response of the left ear. This new problem is formulated in Equation (5.14), where c_4 is an additional scaling constant and $\boldsymbol{\alpha}_{\mathrm{IACC,L}}$ and $\boldsymbol{\alpha}_{\mathrm{IACC,R}}$ denote the so defined left and right IACC. The flip operators in the calculations for the left and right IACC's are applied such that the convolution matrix is made with the prior known desired response and the IACC peaks end up at the the same τ_{IACC} for efficiency and simplicity.

$$
\begin{aligned}
\arg \min_{\mathbf{s_1},\mathbf{s_2},\ldots,\mathbf{s_{N_s}}} \quad & ||\mathbf{c}_{\mathrm{L}}^{*} - \boldsymbol{\alpha}_{\mathrm{L}}||_2^2 + ||\mathbf{c}_{\mathrm{R}}^{*} - \boldsymbol{\alpha}_{\mathrm{R}}||_2^2 \\
\text{s.t.} \quad & \alpha_{\mathrm{IACC,L}}(\tau_{\mathrm{IACC}}^{*}) > \alpha_{\mathrm{IACC,L}}(\bar{n_b}) + c_4(1 - \mu_{\mathrm{IACC}}(\bar{n_b})), \quad \text{for} \quad \bar{n_b} \neq \tau_{\mathrm{IACC}}^{*} \\
& \alpha_{\mathrm{IACC,R}}(\tau_{\mathrm{IACC}}^{*}) > \alpha_{\mathrm{IACC,R}}(\bar{n_b}) + c_3(1 - \mu_{\mathrm{IACC}}(\bar{n_b})), \quad \text{for} \quad \bar{n_b} \neq \tau_{\mathrm{IACC}}^{*} \\
& \boldsymbol{\alpha}_{\mathrm{IACC,L}} = \mathbf{F}\mathbf{C}_{\mathrm{R}}^{*}\mathbf{F}\boldsymbol{\alpha}_{\mathrm{L}} \\
& \boldsymbol{\alpha}_{\mathrm{IACC,R}} = \mathbf{C}_{\mathrm{L}}^{*}\mathbf{F}\boldsymbol{\alpha}_{\mathrm{R}} \\
& \boldsymbol{\alpha}_{\mathrm{L}} = \sum_{n_s} \mathbf{W}^{-1}\hat{\mathbf{C}}_{\mathrm{L},n_s}\hat{\mathbf{s}}_{n_s} \\
& \boldsymbol{\alpha}_{\mathrm{R}} = \sum_{n_s} \mathbf{W}^{-1}\hat{\mathbf{C}}_{\mathrm{R},n_s}\hat{\mathbf{s}}_{n_s}
\end{aligned}
\tag{5.14}
$$

The applied change results in different, non-ideal, results that hopefully come close to the original optimization problem, which holds when the optimization result comes close to the target signal. The great benefit of this change is that the non-convex constraint is now split into two linear constraints. The optimization problem is now not only convex but also a Quadratic Problem (QP) which has beneficial optimization properties.

The aforementioned improvements did not prove to be sufficient to give desirable results. The simplification of optimizing for the channel instead of the entire received audio response does not hold and is not reliable enough. This is shown in Appendix C.2. Not doing this simplification however results in computational expenses too great to handle.

A solution to this can be found when subdividing the audio signal and also the channel response into smaller time blocks. This not only allows for improved computation times but also allows near real time processing of the audio. The theory and implementation of this Multi-Delay Filter (MDF) inspired setup is discussed in the next section.

5.5. Near real-time optimization framework

Real-time computation and processing is important for audio appliances. Especially when watching a movie, the audio stream coming in should be processed and outputted with an unnoticeable delay. To achieve this, we subdivide the incoming audio and also the channel into small time blocks and do the optimization block by block. As mentioned previously, this method also greatly decreases the computational expenses required.

Even though the Multi-Delay Filter [74] is not implemented and used directly, its setup is used. Instead of the conventional overlap-add method, an odd-even overlap-add method is used for smoother results [75]. In short, $\mathbf{s}$ and $\mathbf{c}$ are subdivided in time blocks of a power 2 length with 50% overlap and proper windowing.

Given a desired block length $\bar{L_b}$ (s), we define the actual block length L_b (s) as defined in Equation (5.15), with f_s the sample frequency. This block length is the next power of 2 of the desired block length in samples.

$$
L_b = 2^{\lceil \log_2(\bar{L_b} f_s) \rceil} \quad \text{(s)} \tag{5.15}
$$

The input audio signal $\mathbf{s_{n_s}}$, the channel signals $\mathbf{c_{L,n_s}}$ and $\mathbf{c_{R,n_s}}$ and the response signals $\mathbf{y_L}$ and $\mathbf{y_L}$ are divided into N_i, N_c and N_y time blocks, respectively. These time blocks do not yet include the 50% overlap for notation purposes which become clear later on. The response blocks are defined in Equation (5.16) with $n_i = 1, ..., N_i$, $n_c = 1, ..., N_c$ and $n_y = 1, ..., N_y$. Where the set of all the blocks is denoted by $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{N_s \times L_b \times N_i}$, $\boldsymbol{\mathcal{C}}_{\mathrm{L}} \in \mathbb{R}^{N_s \times L_b \times N_c}$, $\boldsymbol{\mathcal{C}}_{\mathrm{R}} \in \mathbb{R}^{N_s \times L_b \times N_c}$, $\boldsymbol{\mathcal{Y}}_{\mathrm{L}} \in \mathbb{R}^{1 \times L_b \times N_y}$ and $\boldsymbol{\mathcal{Y}}_{\mathrm{R}} \in \mathbb{R}^{1 \times L_b \times N_y}$.

$$\begin{aligned}
\mathcal{S}_{n_s,n_i} &= [s_{n_s}(n_i L_b - L_b + 1), ..., s_{n_s}(n_i L_b)] \\
\mathcal{C}_{\mathrm{L},n_s,n_c} &= [c_{\mathrm{L},n_s}(n_c L_b - L_b + 1), ..., c_{\mathrm{L},n_s}(n_c L_b)] \\
\mathcal{C}_{\mathrm{R},n_s,n_c} &= [c_{\mathrm{R},n_s}(n_c L_b - L_b + 1), ..., c_{\mathrm{R},n_s}(n_c L_b)] \\
\mathcal{Y}_{\mathrm{L},n_y} &= [y_{\mathrm{L}}(n_y L_b - L_b + 1), ..., y_{\mathrm{L}}(n_y L_b)] \\
\mathcal{Y}_{\mathrm{R},n_y} &= [y_{\mathrm{R}}(n_y L_b - L_b + 1), ..., c_{\mathrm{R}}(n_y L_b)]
\end{aligned} \tag{5.16}$$

With the time blocks defined, the optimization setup can be defined. First, one time block is added at the beginning and end of the time block set containing only zeros to allow for a smooth beginning and end of the algorithm. The new set sizes are now given by $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{N_s \times L_b \times N_i + 2}$, $\boldsymbol{\mathcal{C}}_{\mathrm{L}} \in \mathbb{R}^{N_s \times L_b \times N_c + 2}$ and $\boldsymbol{\mathcal{C}}_{\mathrm{R}} \in \mathbb{R}^{N_s \times L_b \times N_c + 2}$.

Since the algorithm is intended for real-time appliances, the computation of the desired response is also done real-time. Since all responses are prior known, the desired response can be computed by convolution in parts using zero padding as shown in Equation (5.17), where $\mathbf{0} \in \mathbb{R}^{1 \times L_b}$ is the zero vector and the stars indicate that the signals correspond to the desired response.

$$\begin{aligned}
\mathcal{Y}^*_{\mathrm{L},n_y} = \text{first } L_b \text{ of } & \sum_{n_c} \mathbf{W}^{-1}(\mathbf{W}\ \mathrm{diag}([\mathcal{S}^*_{n_y - n_c + 1}, \mathbf{0}])\, \mathbf{W}\ \mathrm{diag}([\mathcal{C}^*_{\mathrm{L},n_y + n_c + 1}, \mathbf{0}])) \quad + \\
\text{last } L_b \text{ of } & \sum_{n_c} \mathbf{W}^{-1}(\mathbf{W}\ \mathrm{diag}([\mathcal{S}^*_{n_y - n_c}, \mathbf{0}])\, \mathbf{W}\ \mathrm{diag}([\mathcal{C}^*_{\mathrm{L},n_y + n_c + 1}, \mathbf{0}])) \\
\mathcal{Y}^*_{\mathrm{R},n_y} = \text{first } L_b \text{ of } & \sum_{n_c} \mathbf{W}^{-1}(\mathbf{W}\ \mathrm{diag}([\mathcal{S}^*_{n_y - n_c + 1}, \mathbf{0}])\, \mathbf{W}\ \mathrm{diag}([\mathcal{C}^*_{\mathrm{R},n_y + n_c + 1}, \mathbf{0}])) \quad + \\
\text{last } L_b \text{ of } & \sum_{n_c} \mathbf{W}^{-1}(\mathbf{W}\ \mathrm{diag}([\mathcal{S}^*_{n_y - n_c}, \mathbf{0}])\, \mathbf{W}\ \mathrm{diag}([\mathcal{C}^*_{\mathrm{R},n_y + n_c + 1}, \mathbf{0}]))
\end{aligned} \tag{5.17}$$

With the desired signal defined, the optimization setup can be constructed. To keep the optimization computationally cheap and also introduce as little delay as possible, the optimization is only done for the first two channel blocks. Due to the shape of the RIR (see Figure 4.2), the most power is in these first two channel blocks which is beneficial for this optimization setup. The signal is also filtered by the other channel blocks and added to the entire response, but optimization does not control this. The hypothesis is that these uncontrolled channel blocks do not cause much trouble since the channel power is small and the next iteration can compensate for this.

To further improve performance and general smoothness of the algorithm, an odd and even set is created for optimization. These two sets are optimized independently but the signals to be send to the loudspeaker are added together. The separation of these two sets prevents sharp edges in the optimization signals resulting in smoother results [76].

The optimization process can be found in Equation 5.18, where notation $\mathcal{A}_{[\alpha,\alpha+1]}$ is introduced as more compact replacement for $[\mathcal{A}_\alpha, \mathcal{A}_{\alpha+1}]$. In the equation, $\mathcal{Y}_{[n_i,n_i+1]}$ denotes the current illusion at these time samples given previously computed responses and f_{opt} represents the used optimization function as, for instance, the one given in Equation (5.14).

$$\begin{aligned}
\mathcal{S}_{\text{odd},[n_i,n_i+1]} &= f_{\text{opt}}(\mathcal{Y}_{\text{odd},[n_i,n_i+1]}, \mathcal{Y}^*_{[n_i,n_i+1]}, \mathcal{C}_{[n_i,n_i+1]}), \quad \text{for } n_i = 1, 3, ..., N_i \\
\mathcal{S}_{\text{even},[n_i,n_i+1]} &= f_{\text{opt}}(\mathcal{Y}_{\text{even},[n_i,n_i+1]}, \mathcal{Y}^*_{[n_i,n_i+1]}, \mathcal{C}_{[n_i,n_i+1]}), \quad \text{for } n_i = 2, 4, ..., N_i
\end{aligned} \tag{5.18}$$

As can be seen, the optimization is now done over two time blocks with 50% overlap. Eventually, the overlapping time blocks are added together to obtain the desired audio signal to be presented to the loudspeakers.

The odd and even obtained loudspeaker signals $\mathcal{S}_{\text{odd}}$ and $\mathcal{S}_{\text{even}}$ are added together at the required time index but the sets always remain separated for optimization. For clarity, this is illustrated with proper windowing in Figure 5.6.

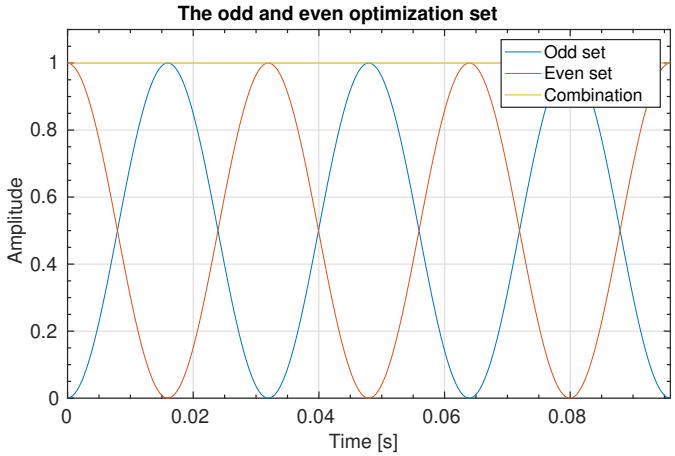


Figure 5.6: Illustration of the odd and even optimization sets. The two sets add op to one due too proper windowing.

Given a newly optimized loudspeaker signal, the odd or even illusion becomes as denoted in Equation (5.19), where a new incoming solution $\mathcal{S}_{[n_i,n_i+1]}$ gives an update to the existing illusion.

$$\begin{aligned}
\mathcal{Y}_{\mathrm{L},n_y} = \text{first } L_b \text{ of } & \sum_{n_c} \mathbf{W}^{-1}(\mathbf{W}(\mathbf{U}\ \mathrm{diag}(\mathcal{C}_{\mathrm{L},[n_y+n_c,n_y+n_c+1]}))\mathbf{W}(\mathbf{U}\mathcal{S}_{[n_y-n_c+1,n_y-n_c+2]})) \quad + \\
\text{last } L_b \text{ of } & \sum_{n_c} \mathbf{W}^{-1}(\mathbf{W}(\mathbf{U}\ \mathrm{diag}(\mathcal{C}_{\mathrm{L},[n_y+n_c,n_y+n_c+1]}))\mathbf{W}(\mathbf{U}\mathcal{S}_{[n_y-n_c,n_y-n_c+1]})) \\
\mathcal{Y}_{\mathrm{R},n_y} = \text{first } L_b \text{ of } & \sum_{n_c} \mathbf{W}^{-1}(\mathbf{W}(\mathbf{U}\ \mathrm{diag}(\mathcal{C}_{\mathrm{R},[n_y+n_c,n_y+n_c+1]}))\mathbf{W}(\mathbf{U}\mathcal{S}_{[n_y-n_c+1,n_y-n_c+2]})) \quad + \\
\text{last } L_b \text{ of } & \sum_{n_c} \mathbf{W}^{-1}(\mathbf{W}(\mathbf{U}\ \mathrm{diag}(\mathcal{C}_{\mathrm{R},[n_y+n_c,n_y+n_c+1]}))\mathbf{W}(\mathbf{U}\mathcal{S}_{[n_y-n_c,n_y-n_c+1]}))
\end{aligned} \tag{5.19}$$

Here, $\mathbf{U} = \mathrm{diag}(\mathbf{u})$, with $\mathbf{u} \in \mathbb{R}^{2L_b}$ denoting a properly sized Hanning window. With the optimization setup defined, the optimization function f_{opt} can be defined, which is done in the next section.

5.6. The proposed optimization function

The optimization function optimizes a small time block of audio being played by the loudspeakers to obtain the desired response at the ears. In general, given the channel and the desired response, a suitable loudspeaker output signal is derived. To do so efficiently and as good as possible, a few alterations are applied to the input signals that are further emphasized in Appendix C.3. In this chapter we do not take these alterations into account to obtain a general description of the proposed algorithm. For the implementation details of the proposed algorithm as presented in Equation (5.22), see Appendix C.4.

First the required response, $\bar{\mathbf{y}}$, at time block n_i is defined given the desired response $\mathcal{Y}^*_{[n_i,n_i+1]}$ and the current present illusion originating from responses generated in previous time blocks. It is defined as given in Equation (5.20).

$$\bar{\mathbf{y}} = \mathcal{Y}^*_{[n_i,n_i+1]} - \mathcal{Y}_{[n_i,n_i+1]} \tag{5.20}$$

With this required response and the MDF inspired framework, the derived optimization functions, for instance the one found in Equation (5.14), can be implemented. A new and improved optimization function is defined instead. Before doing

so, a new optimization feature is introduced that greatly reduces optimization complexity, the masking curve.

As stated and explained in Section 2.2, the masking curve gives the frequency dependent audio power that cannot be heard in presence of a masker. In this case, the masker is the required response and the audio power that should be masked is the difference (error) between the required response and the obtained response found by the optimization.

The masking curve is calculated as presented in Section 2.2 and is implemented by means of the masking matrices $\hat{\mathbf{G}}_{\mathrm{L}} \in \mathbb{R}^{2L_b \times 2L_b} = \mathrm{diag}(\hat{\mathbf{g}}_{\mathrm{L}})$ and $\hat{\mathbf{G}}_{\mathrm{R}} \in \mathbb{R}^{2L_b \times 2L_b} = \mathrm{diag}(\hat{\mathbf{g}}_{\mathrm{R}})$ for left and right ear respectively [77], where $\mathrm{diag}(\cdot)$ converts the vector into a matrix with the vector on the diagonal . This matrix serves as a weighting matrix which indicates the frequency dependent error that can be made.

Due to the masking curve, the minimization of the 2-norm as presented in Equation (5.14) can be replaced by a constrained as shown in Equation (5.21). Note that in contrast to Equation (5.14) we do not optimize for the channel but for the response found at the ears.

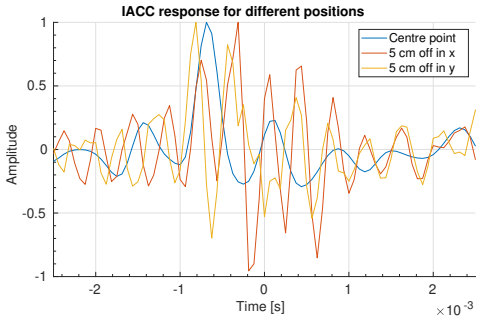
$$
\begin{aligned}
\arg \min_{\mathbf{s_1},\mathbf{s_2},\ldots,\mathbf{s_{N_s}}} \quad & 0 \\
\text{s.t.} \quad & ||\hat{\mathbf{G}}_{\mathrm{L}}(\hat{\bar{\mathbf{y}}}_{\mathrm{L}} - \hat{\boldsymbol{\alpha}}_{\mathrm{L}})||_2 \leq c_1 \\
& ||\hat{\mathbf{G}}_{\mathrm{R}}(\hat{\bar{\mathbf{y}}}_{\mathrm{R}} - \hat{\boldsymbol{\alpha}}_{\mathrm{R}})||_2 \leq c_2 \\
& \alpha_{\mathrm{IACC,L}}(\tau^*_{\mathrm{IACC}}) > \alpha_{\mathrm{IACC,L}}(\bar{n_b}) + c_3(1 - \mu_{\mathrm{IACC}}(\bar{n_b})), \quad \text{for} \quad \bar{n_b} \neq \tau^*_{\mathrm{IACC}} \\
& \alpha_{\mathrm{IACC,R}}(\tau^*_{\mathrm{IACC}}) > \alpha_{\mathrm{IACC,R}}(\bar{n_b}) + c_4(1 - \mu_{\mathrm{IACC}}(\bar{n_b})), \quad \text{for} \quad \bar{n_b} \neq \tau^*_{\mathrm{IACC}} \\
& \boldsymbol{\alpha}_{\mathrm{IACC,L}} = \mathbf{F}\bar{\mathbf{Y}}_{\mathrm{R}}\mathbf{F}\boldsymbol{\alpha}_{\mathrm{L}} \\
& \boldsymbol{\alpha}_{\mathrm{IACC,R}} = \bar{\mathbf{Y}}_{\mathrm{L}}\mathbf{F}\boldsymbol{\alpha}_{\mathrm{R}} \\
& \boldsymbol{\alpha}_{\mathrm{L}} = \sum_{n_s} \mathbf{W}^{-1}\hat{\mathbf{C}}_{\mathrm{L},n_s}\hat{\mathbf{s}}_{n_s} \\
& \boldsymbol{\alpha}_{\mathrm{R}} = \sum_{n_s} \mathbf{W}^{-1}\hat{\mathbf{C}}_{\mathrm{R},n_s}\hat{\mathbf{s}}_{n_s}
\end{aligned}
\tag{5.21}
$$

This special problem is referred to as a constraint satisfaction problem. Since all the requirements can be defined in constraints and there is no necessity for any minimization, this optimization form can be used. The lack of a minimization greatly decreases the computational complexity since all the optimizer has to do is find a solution within the feasible set.

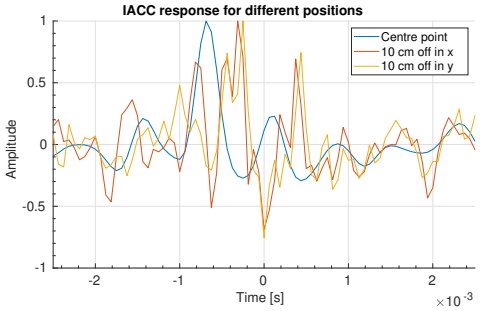
The introduced constraint satisfaction problem finds a solution for one single point in space. The goal of the thesis is to create an as big as possible sweet spot to increase the chances of a satisfying illusion. Just like the multi-point CTC algorithm, we include multiple points in space surrounding the centre point in the optimization.

The problem found in Equation (5.21) allows for a computationally cheap extension to a so defined semi multi-point optimization. As discussed in section 2.1.1, the only interesting information of the IACC is in the rough range of $-1 \leftrightarrow 1$ ms wherein the peak index corresponding to the ITD is present. Due to this property, the computational expense of the IACC constraint can be greatly reduced up to a point where it is nearly negligible compared to the computational expense of the masking constraint.

Because of this, it is relatively cheap to add multiple IACC constraints for N_l different points in space, as is done in the problem posed in Equation (5.22). In this problem, $n_l = 1$ corresponds to the centre optimization point. Do note that in the implementation of the problem zero padding is required, this is further emphasized in Appendix C.4. With this problem, the assumption is made that the masking curve constraint holds for all the different points in space for which the IACC constraint is



(a) The IACC response for the optimization point and two points on a 5 cm radius sphere around this point. The IACC responses show that the illusion is not entirely valid at these points.



(b) The IACC response for the optimization point and two points on a 10 cm radius sphere around this point. The IACC responses show that the illusion is not entirely valid at these points.

Figure 5.7: IACC responses found for the proposed algorithm with 7 optimization points, one centered point surrounded by 6 equally spaced points on a 6 cm radius sphere surrounding the centre point.

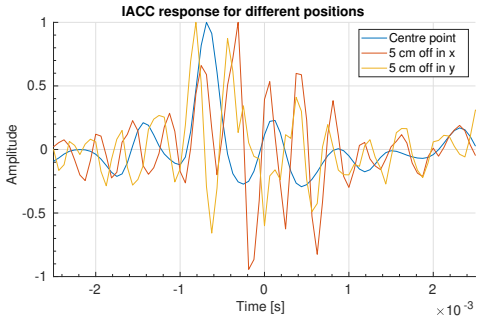
introduced. Also, do note that all $\boldsymbol{\alpha}_{\text{IACC,L}}$ and $\boldsymbol{\alpha}_{\text{IACC,R}}$ are calculated with the desired responses at the centre optimization point, which also introduces errors.

$$
\begin{aligned}
\arg \min_{\mathbf{s_1},\mathbf{s_2},\ldots,\mathbf{s_{N_s}}} \quad & 0 \\
\text{s.t.} \quad & ||\hat{\mathbf{G}}_{\text{L}}(\hat{\bar{\mathbf{y}}}_{\text{L}} - \hat{\boldsymbol{\alpha}}_{\text{L},1})||_2 \leq c_1 \\
& ||\hat{\mathbf{G}}_{\text{R}}(\hat{\bar{\mathbf{y}}}_{\text{R}} - \hat{\boldsymbol{\alpha}}_{\text{R},1})||_2 \leq c_2 \\
& \alpha_{\text{IACC,L},n_l}(\tau^*_{\text{IACC}}) > \alpha_{\text{IACC,L},n_l}(\bar{n_b}) + c_3(1 - \mu_{\text{IACC}}(\bar{n_b})), \quad \text{for} \quad \bar{n_b} \neq \tau^*_{\text{IACC}} \\
& \alpha_{\text{IACC,R},n_l}(\tau^*_{\text{IACC}}) > \alpha_{\text{IACC,R},n_l}(\bar{n_b}) + c_4(1 - \mu_{\text{IACC}}(\bar{n_b})), \quad \text{for} \quad \bar{n_b} \neq \tau^*_{\text{IACC}} \\
& \boldsymbol{\alpha}_{\text{IACC,L},n_l} = \mathbf{F}\bar{\mathbf{Y}}_{\text{R}}\mathbf{F}\boldsymbol{\alpha}_{\text{L},n_l} \\
& \boldsymbol{\alpha}_{\text{IACC,R},n_l} = \bar{\mathbf{Y}}_{\text{L}}\mathbf{F}\boldsymbol{\alpha}_{\text{R},n_l} \\
& \boldsymbol{\alpha}_{\text{L},n_l} = \sum_{n_s} \mathbf{W}^{-1}\hat{\mathbf{C}}_{\text{L},n_s,n_l}\hat{\mathbf{s}}_{n_s} \\
& \boldsymbol{\alpha}_{\text{R},n_l} = \sum_{n_s} \mathbf{W}^{-1}\hat{\mathbf{C}}_{\text{R},n_s,n_l}\hat{\mathbf{s}}_{n_s} \\
& \text{for} \quad n_l = 1, \ldots, N_l \quad \& \quad \bar{n_b} \in [-1, 1] \text{ ms}
\end{aligned}
\tag{5.22}
$$

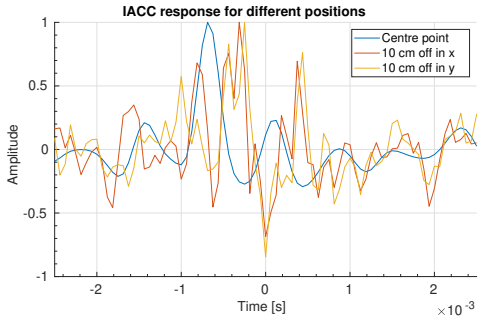
The result of the algorithm is, just like the other algorithms, evaluated by the size of the sweet spot. In addition to the simulation setup mentioned above, we use time blocks of length $L_b = 16$ ms for the simulation and just like the multi-point CTC algorithm, the 6 extra optimization points are placed on the positive and negative axis on a 6 cm radius around the centre point. The results of the simulation for the problem defined in Equation (5.22) are given in Figure 5.7.

The results show a worse IACC response when compared to the results found for the multi-point CTC as presented in Figure 5.4. Although this seems disappointing, the amount of optimization points can easily be extended without adding too much computational expenses. Instead of taking 6 optimization points on one sphere with a radius of 6 cm, we now take 6 optimization points on three spheres with radii 2, 4 and 6 cm. The results are presented in Figure 5.8.

As can be seen, the addition of the spheres hardly makes a difference on the result. This indicates that the optimization problem and its structure does not lead to the desired results. Further expanding the set of optimization points is possible and leads to little added computational expenses. Optimization points on spheres with



(a) The IACC response for the optimization point and two points on a 5 cm radius sphere around this point. The IACC responses show that the illusion is not entirely valid at these points.



(b) The IACC response for the optimization point and two points on a 10 cm radius sphere around this point. The IACC responses show that the illusion is not entirely valid at these points.

Figure 5.8: IACC responses found for the proposed algorithm with 19 optimization points, one centered point surrounded by 6 equally spaced points on a 2, 4 and 6 cm radius sphere surrounding the centre point.

bigger radii lead to infeasible sets however, meaning it is not possible to greatly expand the optimization problem.

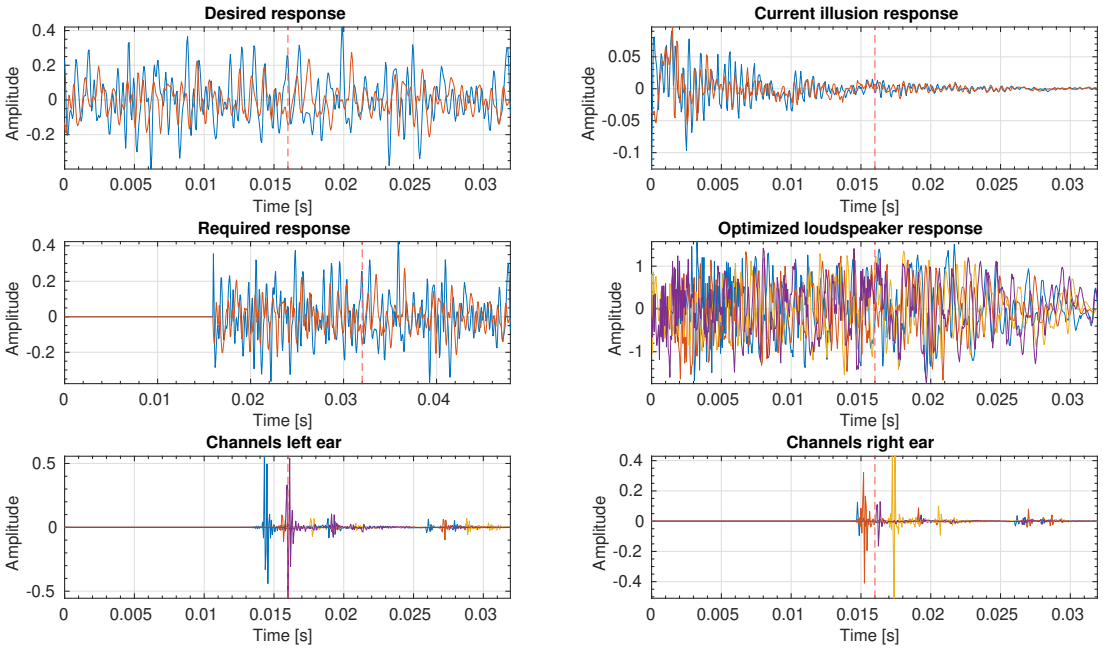
5.7. Insight in time block optimization

Before discussing the results of the algorithm, the functioning of the optimization in a time block is shown. At a certain time block, the optimization consists of the signals given in Figure 5.9, with the time domain signals given in Figure 5.9a and the constraints related signals given in Figure 5.9b. The different signals are treated to get a better idea of the functioning of the algorithm.

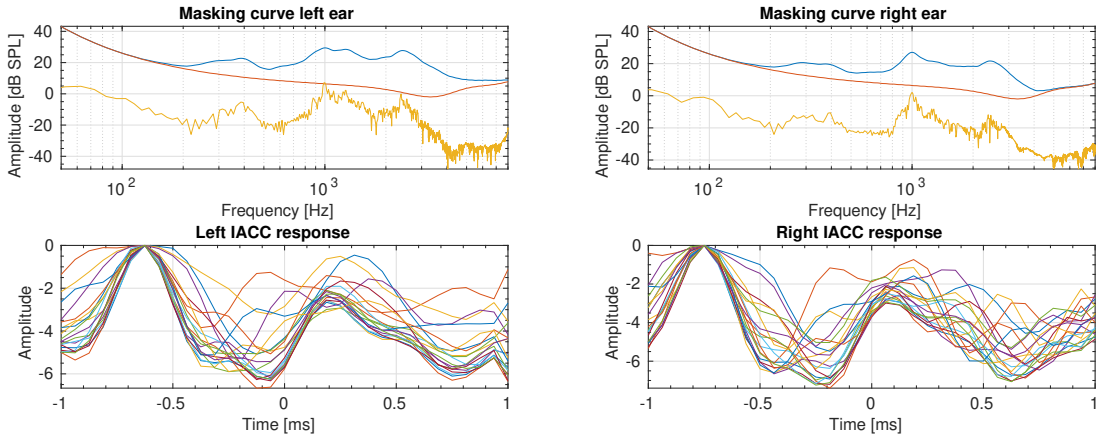
First of all, The required response is defined by means of the desired response and the current response by means of Equation (5.20). This can be deduced from Figure 5.9a with close inspection. Do note that the required response has an additional zero block at the left, this has practical reasons that are discussed in Appendix C.3. The channels from the loudspeakers to the left and right ear are plotted in the bottom two plots in Figure 5.9a. The clear difference between the two show the influence of the HRTF. Especially the time difference but also the level differences shows the effect of the head shadowing and of course the difference in position of the ears. The loudspeaker signals convolved with channels for the left channel should create the blue plot in the required response plot while those same loudspeaker signals should create the red plot with the right channels. A solution for the loudspeaker signals that does this sufficiently according to the optimization constraints is given in the optimized loudspeaker response plot.

In Figure 5.9b the status of the constraints is given. The above two plots show the masking curve, threshold in quiet and the current error between obtained response and desired response. As can be seen from these plots, the error takes the shape of the masking curve but stays far below its limits. Hard conclusions cannot be drawn on this find but it seems to indicate that a solution with the desired IACC requires a solution closely related to the exact desired response. The bottom two plots in Figure 5.9b show the IACC constraints. As can be seen, the algorithm struggles most with this constraint since the peak IACC value is not that much higher than the other points. With constants c_3 and c_4 we are able to force these side lobes to be smaller but this generally results in an infeasable problem.

This indicates that the freedom we give the algorithm by means of the masking



(a) Optimization function time domain signals. In the desired, current illusion and required response plot, the blue plots represent the response at the left ear and the red plots represent the response at the right ear. In the other plots, the optimized loudspeaker response and the channels are depicted. Each individual loudspeaker is represented by the same colour in these three plots.



(b) Optimization function constraints. The top two plots show the masking curve of the left and right ear respectively. The red plot represents the threshold in quiet, the blue plot represents the masking curve and the yellow curve represents the error. The bottom two plots show the IACC response used in the IACC related constraint, where $\tau^*_{\text{IACC}} \approx -0.7$ ms.

Figure 5.9: An example of signals present in one optimization iteration . Both the available and derived time signals are presented in (a). The constraints and their status is depicted in (b).

curve can not be used by the algorithm to create a more reliable IACC response. Consequently, this freedom can not be used to increase the size of the sweet spot.

6

Results and validation

Throughout Chapter 5, the performance of the introduced algorithms' is presented by validating the IACC response of the algorithms. This is primarily done to motivate choices and improvements that are made to the algorithms to eventually end up with the proposed algorithm. In this chapter, a proper comparison and validation of the algorithms' performance is done. The considered algorithms are the original crosstalk cancellation algorithm, the multi-point crosstalk cancellation algorithm and the proposed algorithm. The validation is done by considering three aspects: auditory localization performance, error between desired and obtained result, and the perceptual difference between desired and obtained result. The auditory localization performance is tested by means of the IACC response, the error between desired and obtained result is evaluated by means of the L_2-norm and the perceptual difference between desired and obtained result is validated by means of the Perceptual Evaluation of Audio Quality (PEAQ) measure. The results are obtained using the setup presented in Section 5.1. First, the validation metrics are discussed after which a comparison and analyses of the results is performed.

6.1. Validation metrics

In this section, the validation metrics are introduced and discussed. To do this we use the obtained response at left and right ear, $\hat{\mathbf{y}}_\mathrm{L}$ and $\hat{\mathbf{y}}_\mathrm{R}$ respectively, and the desired response at left and right ear, $\hat{\mathbf{y}}^*_\mathrm{L}$ and $\hat{\mathbf{y}}^*_\mathrm{R}$ respectively.

6.1.1. Interaural crosscorrelation

The InterAural CrossCorrelation (IACC) is discussed in Section 2.1.1 but is shortly repeated here for sake of completeness. We are interested in the time-index corresponding to the peak value of the IACC response since this indicates the observed Interaural Time Difference (ITD). The ITD is the primary auditory cue used by the auditory system to estimate the azimuth angle when localizing a sound source.

Deriving the IACC response, $\mathbf{y}_\mathrm{IACC}$, is done by calculating the cross correlation of $\hat{\mathbf{y}}_\mathrm{L}$ and $\hat{\mathbf{y}}_\mathrm{R}$ as shown in Equation 5.3. Given the IACC response, the time-index corresponding to the maximum value of the response is denoted by τ_IACC as shown in Equation (5.6). Here, τ^*_IACC denotes the time delay corresponding to the desired response and τ_IACC denotes the time delay corresponding to the obtained result.

In the validation, we rank the IACC performance based on the difference d_{IACC} between τ^*_{IACC} and τ_{IACC}, as presented in Equation (6.1).

$$d_{\text{IACC}} = |\tau^*_{\text{IACC}} - \tau_{\text{IACC}}| \quad \text{(s)} \tag{6.1}$$

To be able to relate d_{IACC} to the localization performance, d_{IACC} must be translated to an angle difference. In the simulation, the to be recreated source is placed at the left of the listener and will give the reference time delay τ^*_{IACC}. The angle of the perceived source relative to the reference source, denoted by d_θ, in terms of τ_{IACC} and τ^*_{IACC}, is done by means of Equation (6.2). This equation originates from the goniometric evaluation of the situation shown in Figure 6.1. Do note that we assume a plane wave coming from the perceived loudspeaker in this situation. This is a valid estimation since we are only interested in rough estimates of the angles for the performance assessment as shown next.

$$d_\theta = 90^\circ - \text{asin}\left(\frac{\tau_{\text{IACC}}}{\tau^*_{\text{IACC}}}\right) \tag{6.2}$$

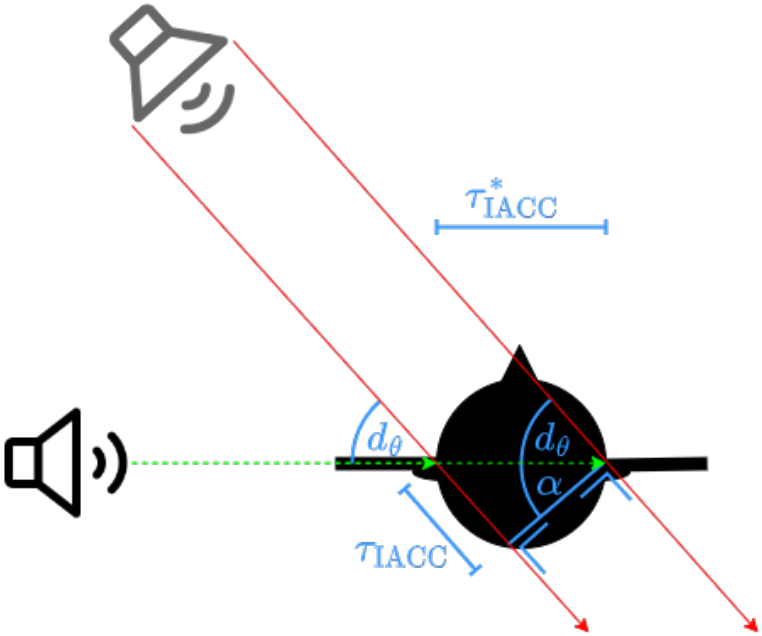


Figure 6.1: From time index to azimuth angle of incidence. In the figure, the loudspeaker on the left corresponds to the reference loudspeaker with delay τ^*_{IACC} and the shaded loudspeaker on the topleft corresponds to the perceived loudspeaker corresponding to time-delay τ_{IACC}. Given these two time delays, the angle α can be calculated by means of the asin$(\cdot)$ function in Equation (6.2). Calculating the angle d_θ by using this value is done as given in Equation (6.2).

The auditory localization performance of the algorithms is evaluated by the ability to correctly estimate the azimuth angle of the to be localized source. This ability is evaluated by categorizing the results for each algorithm and scenario in four different groups. These groups are defined by means of the d_θ that corresponds to the τ_{IACC} found in the IACC of the perceived response.

The four different groups are presented in Figure 6.2. The first group is the one corresponding to the green region. As presented in Section 2.1.1, the accuracy of the localization of sources to the left or right of the listener is roughly 20°. Because

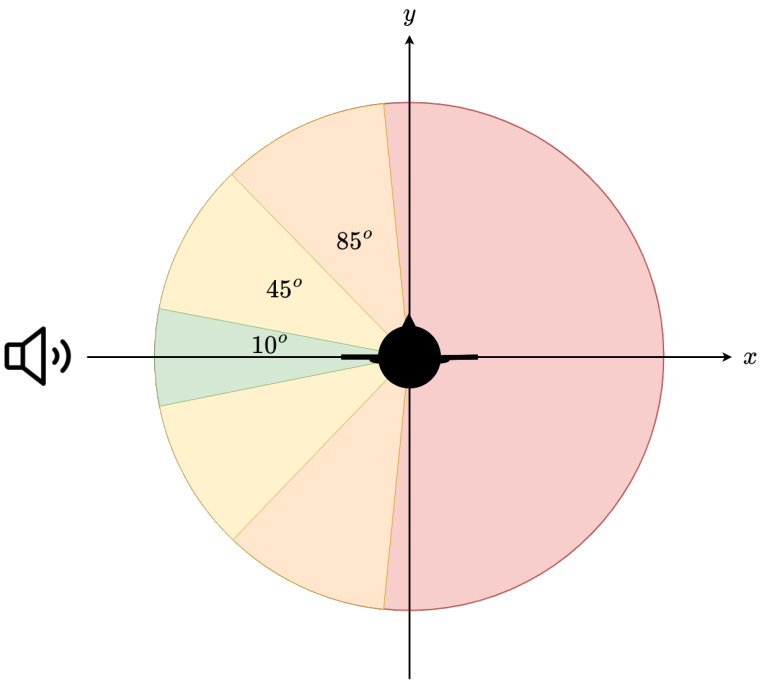


Figure 6.2: Angle groups defining the different score groups. The green 10° region indicates ”good” localization performance. The yellow 45° region indicates ”sufficient” localization performance. The orange 85° region indicates ”indicating” localization performance. The red region indicates ”poor” localization performance.

of this, we consider the green region, defined by a 10° span in both the positive and negative y-plane, to be the region with ”good” localization properties.

The yellow region indicates the correct classification between a source originating from either the left, right, front or back. This region does not give correct localization cues to localize the source at the correct location but at least gives the correct direction. This region is referred to as ”sufficient”.

The orange region is a broader definition of the yellow region. The orange region indicates the correct distinction between left and right. According to Section 2.1.1, the accuracy of localization of sources in front of the listener is roughly 5°. Because of this, the threshold for this region is not 90° but 85° to ensure that the correct distinction between left and right is made (85° is chosen instead of 87.5° as an additional safety margin). This region is referred to as ”indicating”.

The red region means that the listener receives auditory cues that do not at all indicate the desired location of the virtual source. The localization of the source is clearly wrong in this case and thus this region is referred to as ”poor”.

In some cases, the IACC response consists of two peaks of similar amplitude at different time-indices. Two source locations will be perceived in this situation and the worst time delay will be chosen when grading the localization.

To convert the regions in terms of d_θ to d_{IACC}, we use Equation (6.3), which is based on Equation (6.2). The summary of all the scores and their definitions is given in Table 6.1.

$$\frac{\tau_{\text{IACC}}}{\tau^*_{\text{IACC}}} = \sin(90^\circ - d_\theta) \rightarrow d_{\text{IACC}} = |\tau^*_{\text{IACC}} - \tau_{\text{IACC}}| = |\tau^*_{\text{IACC}} - \sin(90^\circ - d_\theta)\tau^*_{\text{IACC}}| \tag{6.3}$$

Score	d_θ	d_{IACC}
good	10°	0.02 (ms)
sufficient	45°	0.25 (ms)
indicating	85°	0.79 (ms)
poor	else	else

Table 6.1: Localization scores. The d_{IACC} is determined using Equation 6.3 and the result is round up. The values d_{IACC} serve as an upper limit difference to be categorized in the corresponding group.

Difference grade	Subjective Description of difference
0	Imperceptable
-1	Perceptable but not annoying
-2	Slightly annoying
-3	Annoying
-4	Very annoying

Table 6.2: Difference grade corresponding to the PEAQ measure. These scores relate the output grades to the perceived audio quality.

6.1.2. L_2-norm of the error

The original CrossTalk Cancellation (CTC) is an objective optimization problem and we thus wish to validate all the algorithms by means of an objective measure. On top of this, the validation by an objective metric gives us the means to make a comparison between objective and perceptual performance. This is especially interesting since the goal of the thesis is to change the objective measure of CTC into a perceptual measure. The objective measure chosen is the L_2-norm of the error between the desired and obtained response, denoted by e, as given in Equation (6.4).

$$e = ||\hat{\mathbf{y}}_{\text{L}}^* - \hat{\mathbf{y}}_{\text{L}}||_2 + ||\hat{\mathbf{y}}_{\text{R}}^* - \hat{\mathbf{y}}_{\text{R}}||_2 \tag{6.4}$$

Validating the error based on this value is done by comparing the error values of all the different algorithms and scenarios.

6.1.3. Perceptual evaluation of audio quality

The Perceptual Evaluation of Audio Quality (PEAQ) measure is used to determine the audio quality of the obtained result in comparison with the desired result. The PEAQ is a standardized measure to evaluate audio quality and it characterizes how the audio quality would be perceived by human test subjects, as proposed in [78]. The used PEAQ implementation is based on [79] and the implementation can be found in [80].

The PEAQ measure gives a subjective score that indicates the audibility of the differences between the desired signal and the obtained signal. It does so in terms of the difference grade which ranges from 0 to -4 as given in Table 6.2.

Since there is a response for the left and right ear, the average of the PEAQ grades for both ears is taken, as presented in Equation 6.5. Here, PEAQ$(\cdot)$ denotes the PEAQ measure.

$$\text{PEAQ grade} = \frac{\text{PEAQ}(\hat{\mathbf{y}}_{\text{L}}^*, \hat{\mathbf{y}}_{\text{L}})}{2} + \frac{\text{PEAQ}(\hat{\mathbf{y}}_{\text{R}}^*, \hat{\mathbf{y}}_{\text{R}})}{2} \tag{6.5}$$

Centre point	CTC	M-point CTC	Prop. alg. 7	Prop. alg. 19
IACC	**good**	**good**	**good**	**good**
L_2-norm	$\mathbf{8.28 \cdot 10^{-6}}$	3.76	66.8	66.5
PEAQ	**0.136**	−2.11	−2.51	−2.65

Table 6.3: Results for centre optimization point. The bold-faced results correspond to the best performing algorithm in that particular metric.

6.2. Results and validation of the algorithms

The validation methods provide a way to compare the discussed algorithms and draw conclusions on the performance of the proposed algorithm. The compared algorithms are the original CrossTalk Cancellation (CTC), the multi-point CTC, the proposed algorithm with 7 optimization points and the proposed algorithm with 19 optimization points. All the algorithms have a centre optimization point that represents the centre of the head. The mutli-point CTC and the proposed algorithm with 7 optimization points have 6 additional optimization points evenly placed on a sphere with radius 6 cm centered around the centre optimization point. The proposed algorithm with 19 optimization points has 6 additional optimization points evenly placed on three spheres with radii 2, 4 and 6 cm centered around the centre optimization point. In the presented result tables, CTC refers to the original crosstalk cancellation algorithm, M-point CTC refers to multi-point crosstalk cancellation, Prop. alg. 7 refers to proposed algorithm with 7 optimization points and Prop. alg. 19 refers to proposed algorithm with 19 optimization points.

The algorithms are evaluated on the results obtained from a simulation with the setup as presented in Section 5.1. The results are analysed at the centre optimization point and two points 5, 10, 20 and 30 cm off the centre optimization points. Apart from the centre optimization point, these validation points are not located at an optimization point for a proper validation. Based on these findings, conclusions are drawn.

An important note to the obtained result is the alteration of the results of the proposed algorithms. Since the small time frames combined with the time independence of the masking curve results in a form of musical noise, the audio quality is degraded significantly. The PEAQ results are thus bad. A solution to this is not researched nor implemented in this thesis. Since this musical noise is mainly found in frequencies higher than the maximum frequency of the audio content, a low pass filter is applied to the obtained results to mimic the effect of a solution to the musical noise problem. We only apply this filter when determining the PEAQ results. This filter is a minimum-order low-pass filter with a stopband attenuation of 60 dB and a cut-off frequency of 950 Hz.

6.2.1. Results centre optimization point

The results for the centre optimization point are presented in Table 6.3.

These results show that the original CTC algorithm gives a near perfect reconstruction since the L_2-norm error is close to zero and the PEAQ shows an imperceptable difference. Comparing the L_2-norm and the PEAQ results for the multi-point CTC and the original CTC shows that the multi-point CTC optimizes for multiple points and is thus unable to perfectly recreate the de desired response. The results for the multi-point CTC and the proposed algorithms show that we do indeed optimize for the audio perception and not an objective measure. Tis because the IACC is good for both algorithms and the PEAQ score is also similar. The L_2-norm of the proposed algorithm solutions is however significantly higher compared to the multi-point CTC. This clearly indicates that the optimization is done for the perception of audio and not the objective signal.

5 cm off in x	CTC	M-point CTC	Prop. alg. 7	Prop. alg. 19
IACC	**good**	sufficient	indicating	indicating
L_2-norm	8.62	**8.19**	196.5	192.8
PEAQ	**−2.45**	−2.59	−3.30	−3.31
5 cm off in y	CTC	M-point CTC	Prop. alg. 7	Prop. alg. 19
IACC	sufficient	**good**	sufficient	sufficient
L_2-norm	8.07	**6.58**	210	206.6
PEAQ	−2.97	**−2.82**	−3.88	−3.88

Table 6.4: Results for 5 cm off centre. The bold-faced results correspond to the best performing algorithm in that particular metric.

10 cm off in x	CTC	M-point CTC	Prop. alg. 7	Prop. alg. 19
IACC	**good**	indicating	indicating	indicating
L_2-norm	13.72	**12.49**	218.1	218.2
PEAQ	**−2.73**	−2.85	−3.31	−3.38
10 cm off in y	CTC	M-point CTC	Prop. alg. 7	Prop. alg. 19
IACC	sufficient	**good**	indicating	indicating
L_2-norm	13.22	**11.15**	217.8	220.5
PEAQ	−3.23	**−2.70**	−3.36	−3.39

Table 6.5: Results for 10 cm off centre. The bold-faced results correspond to the best performing algorithm in that particular metric.

6.2.2. Results 5 cm off optimization point

The results for the 5 cm off optimization points are presented in Table 6.4.

These results show that the CTC performance degrades greatly in the L_2-norm and PEAQ score compared to the results for the centre optimization point as found in Table 6.3. The multi-point CTC performance suffered a small degration in the PEAQ score but the IACC and L_2-norm remained similar. The proposed algorithms show a great decrease in performance. The PEAQ is far worse compared to the centre optimization point and the IACC score is degraded. These results indicate that the methods used to increase the sweet spot size do not give the desirable result. Also do note that the results of the 7 and 19 point algorithm are very similar, indicating that the additional optimization points do not improve the performance.

6.2.3. Results 10 cm off optimization point

The results for the 10 cm off optimization points are presented in Table 6.5.

These results are similar to the ones found in Table 6.4. The performance of the proposed algorithms are slightly worse in terms of the IACC. Comparing the CTC and the multi-point CTC shows that the multi-point CTC algorithm slightly outperforms the CTC algorithm at 10 cm off the centre point, this was not the case for 5 cm off. The localization performance is still roughly the same but the PEAQ score and the L_2-norm are better for the multi-point CTC algorithm.

6.2.4. Results 20 cm off optimization point

The results for the 20 cm off optimization points are presented in Table 6.6.

These results show the effect of the multiple optimization points. The IACC results of the CTC algorithm are considered poor and do not contain the correct auditory cues to recreate the virtual source. The other three algorithms do not show perfect results but on average the performance is not poor. The other metrics have similar scores compared to the ones given in Table 6.5.

20 cm off in x	CTC	M-point CTC	Prop. alg. 7	Prop. alg. 19
IACC	**poor**	**poor**	**poor**	**poor**
L_2-norm	17.93	**17.3**	216.1	214.0
PEAQ	−2.87	**−2.77**	−3.29	−3.34
20 cm off in y	CTC	M-point CTC	Prop. alg. 7	Prop. alg. 19
IACC	poor	**good**	indicating	indicating
L_2-norm	16.90	**15.1**	240.3	240.5
PEAQ	**−2.72**	−2.73	−3.44	−3.50

Table 6.6: Results for 20 cm off centre. The bold-faced results correspond to the best performing algorithm in that particular metric.

30 cm off in x	CTC	M-point CTC	Prop. alg. 7	Prop. alg. 19
IACC	poor	poor	poor	**indicating**
L_2-norm	**16.88**	**16.88**	232.4	233.4
PEAQ	**−3.29**	−3.33	−3.81	−3.84
30 cm off in y	CTC	M-point CTC	Prop. alg. 7	Prop. alg. 19
IACC	**indicating**	poor	**indicating**	**indicating**
L_2-norm	15.86	**15.40**	244.6	243.6
PEAQ	−3.49	**−2.91**	−3.89	−3.88

Table 6.7: Results for 30 cm off centre. The bold-faced results correspond to the best performing algorithm in that particular metric.

6.2.5. Results 30 cm off optimization point

The results for the 30 cm off optimization points are presented in Table 6.7.

These results show that for both the original CTC and the multi-point CTC algorithm the illusion of the recreated virtual source is nearly lost. The multi-point CTC solution has a slight advantage in the PEAQ score. The proposed algorithm gives a better results in the IACC metric, an indication of the illusion of the virtual source is recreated at this distance. This performance in terms of the localization performance comes at the cost of audio quality. The PEAQ score indicates a score equivalent to very annoying meaning that the quality is deterred significantly.

6.2.6. Summary of the results

From the above results we can draw some general conclusions. The CTC and the multi-point algorithms generate results that are closely related to the required response as can be seen in the L_2-norm results. The IACC and the PEAQ results for the multi-point CTC are slightly better than the ones for the CTC algorithm, meaning that multi-point CTC is an improvement over CTC. The proposed algorithm with 7 and 19 optimization points show similar results meaning that the addition of the optimization points does not have a significant impact. The L_2-norm error of the proposed algorithm is significantly higher than the other algorithms while the PEAQ score is not significantly worse. This indicates that the proposed algorithms do not optimize for an objective but a perceptual measure. Comparing the proposed algorithm results to the multi-point CTC results shows that the IACC performances are better at long distance and the PEAQ results are worse. This indicates that we were successful in increasing the size of the sweat spot but this came with the cost of degraded audio quality.

7 Conclusion

The main focus of the thesis is the recreation of a virtual audio source in a practical consumer living room situation. The rooms resembling a living room consist of a small set of two to five loudspeakers and partly reflective walls. In this room, an illusion resembling the virtual source must be delivered to a single listener in the room. The resulting research question is:

> *Given a set of four physical loudspeakers in a room, is it possible to create the illusion of a virtual audio source for one listener in the room?*

A solution to this problem can be found in crosstalk cancellation. Crosstalk cancellation gives theoretically viable results but suffers from lack of robustness and stability. Increasing this robustness and stability is attempted in this thesis by considering the audio perception of the human listener.

The bad characteristics of crosstalk cancellation originate from the objective nature of the problem. The crosstalk cancellation problem implies the inversion of the room impulse response resulting in an ill-posed problem. The stability and robustness of the solution is defined in terms of the size of the sweet spot. The sweet spot is defined as the largest sphere in space surrounding the head in which the illusion is sufficiently present. A larger sweet spot indicates a more stable and robust solution. To increase the sweet spot size, the problem is relaxed by finding solutions that are perceptually sufficient instead of objectively optimal. One of the perceptual measures used is the masking curve, which describes the frequency dependent detectability of audio with a masker sound source present. The masking curve is used to determine which errors are inaudible and this freedom is used to optimize for the audio spatial perception and increase the stability and robustness of the problem. The second used measure is the interaural crosscorrelation. This measure is closely related to the way the human brain identifies the interaural time difference which is primarily used when localizing an audio source in azimuth direction. These two measures are added to the original crosstalk cancellation problem to form a constraint satisfaction problem. This constraint satisfaction problem finds a solution in the feasible set defined by the perceptual constraints. As a result, the sweet spot size is increased compared to the original crosstalk cancellation algorithm. This comes however at the cost of deterred audio quality. The resulting conclusion is that the proposed solution is more robust and stable than the original crosstalk cancellation but lacks the desired audio quality.

8

Discussion and future work

The presented work shows a step towards the inclusion of perceptual measures in the field of audio envelopment and localization with a limited amount of speakers. The proposed algorithm shows promising results but the validity, applicability and expandability of the solution is questionable and requires further research.

First of all, the validation of the algorithms' performance by means of the IACC is questionable. To properly validate the performance of the algorithms, an involved subjective experiment with a large number of listeners is required. Since such an experiment is outside the scope of this thesis, the IACC measure is used as replacement. A subjective validation experiment is desirable as expansion to this work.

Another questionable aspect of the validation method is the assumption that a larger sweet spot translates to a more robust and stable system. A larger sweet spot means that there is more displacement of the head possible before the illusion is lost, but that does not necessarily mean that the system is more robust against other possible inaccuracies. The validation of this relation is required to show that the proposed method of increasing robustness and stability holds.

The Room Impulse Response model used in this thesis results in a specific response for a specific room. Since the goal is to construct an algorithm that functions for practical and more general situations, this RIR model does not fit. To improve this, a stochastical Room Impulse Response has been derived by means of simulations. The work is presented in the submitted paper that can be found in the Appendix A. This RIR model can improve the stability of the found solutions and also the practicality of the system. A next step would be to use this stochastical RIR to derive, for instance, a minimum variance solution to the problem instead of the current deterministic algorithm. Due to limited time, the implementation of the RIR model or a stochastical analyses could not be performed.

The "gong" audio stimulus used to verify the algorithms is a simple, slowly varying and low frequency signal. A more varying audio stimulus could be used to analyze the proposed algorithm more in depth. Similarly, the influence of higher frequency content in the audio stimuli should be discovered. The length of the channel, about 200 ms, could become a problem when using a faster varying and higher frequency sound stimulus.

In the proposed algorithm we only consider the interaural time difference auditory cue. In practice auditory localization uses more cues as, for instance, the interaural level difference. Taking these auditory cues into account in the opti-

mization can potentially increase the performance and also enable the possibility to include elevation and distance estimation.

The masking curves used in the optimization give some problems caused by the time dependency of the sound stimulus. Since the masking curve is determined in the frequency domain, time dependency information is lost. This leads to proper masking in the frequency domain but clear errors can be found in the time domain signal. Improved implementation of the masking curve is required for audibly pleasing signals.

The HRTF model used to determine the channel from the loudspeaker to the signal found inside the ears is based on a large set of measurements performed on the KEMAR head model. Instead of this advanced model, a more simplified model can be used to lower the computational expenses and potentially improve the stability and robustness. The proposed model consists of the outer- and middle-ear filter used in the masking curve definition in combination with a delay corresponding to the ITD cue.

Finally, when all the aforementioned issues are taken care of, the creation of only one virtual source can be expanded to multiple virtual sources to be able to create even more impressive audio experiences. Whether the creation of virtual sources creates new problems should be investigated.

Bibliography

[1] D. L. Engineers, "Dolby digital 5.1." Available at https://professional.dolby.com/tv/dolby-digital/#gref (2022/12/20).

[2] L. Gaston and R. Sanders, "Acoustic control by wave field synthesis," *Evaluation of he-acc, ac-3, and e-ac-3 codecs*, vol. 56, no. 3, pp. 140–155, 2008.

[3] C. Todd, "Ac-3 audio coder," *Proc. SPIE 10282, Standards and Common Interfaces for Video Information Systems: A Critical Review*, pp. 24–35, 1995.

[4] D. L. Engineers, "5.1 virtual speaker setup." Available at https://www.dolby.com/about/support/guide/speaker-setup-guides/5.1-virtual-speakers-setup-guide (2023/01/11).

[5] R. S. PLC, "An introduction to 5.1 speaker systems." Available at https://blog.richersounds.com/an-introduction-to-5-1-speaker-systems/ (2023/05/30).

[6] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2778, 1993.

[7] C. Raffel, "Wave field synthesis vs. higher order ambisonics; audio codec improvement." Available at https://ccrma.stanford.edu/events/wave-field-synthesis-vs-higher-order-ambisonics-audio-codec-improvement (2022/12/22).

[8] N. Failla, "Wave field synthesis." Available at https://proyectoidis.org/wave-field-synthesis/(2022/12/20).

[9] B. S. Atal and M. R. Schroeder, "Apparent sound source translator," *U.S. patent*, vol. 3,236,949, pp. 1–10, 1966.

[10] T. Takeuchi and P. A. Nelson, "Optimal source distribution for binaural synthesis over loudspeakers," *The Journal of the Acoustical Society of America*, vol. 112, no. 6, pp. 2786–2797, 2002.

[11] O. Kirkeby, P. A. Nelson, and H. Hamada, "Local sound field reproduction using two closely spaced loudspeakers," *The Journal of the Acoustical Society of America*, vol. 104, pp. 1973–1981, 1998.

[12] H. Braren and J. Fels, "A high-resolution head-related transfer function data set and 3d-scan of kemar," tech. rep., Institute and Chair for Hearing Technology and Acoustics, Aachen, 2020.

[13] J. Stewart, D. Clegg, and S. Watson, *Calculus Early Transcendentals*. Cengage Learning, 9 ed., 2019.

[14] T. R. Letowski and S. T. Letowski, *Auditory Spatial Perception: Auditory Localization*. Aberdeen Proving Ground, MD 21005-5425: U.S. Army Research Laboratory, 1 ed., 2012.

[15] J. C. Middlebrooks and D. M. Green, "sound localization by human listeners," *Annual review of psychology*, vol. 42, pp. 135–159, 1991.

[16] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *The Journal of the Acoustical Society of America*, vol. 111, pp. 2219–2236, 2002.

[17] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *The Journal of the Acoustical Society of America*, vol. 91, pp. 1648–1661, 1992.

[18] G. H. Recanzone, S. D. D. R. Makhamra, and D. C. Guard, "Comparison of relative and absolute sound localization ability in humans," *The Journal of the Acoustical Society of America*, vol. 103, pp. 1085–1097, 1998.

[19] F. L. Wightman and D. J. Kistler, "Monaural sound localization revisited," *The Journal of the Acoustical Society of America*, vol. 101, pp. 1050–1063, 1997.

[20] K. C. Wood and J. K. Bizley, "Relative sound localisation abilities in human listeners," *The Journal of the Acoustical Society of America*, vol. 138, pp. 674–686, 2015.

[21] C. V. Parise, K. Knorre, and M. O. Ernst, "Natural auditory scene statistics shapes humanspatial hearing," *PNAS*, vol. 111, no. 16, pp. 6104–6108, 2014.

[22] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners," *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2188–2200, 1990.

[23] S. Carlile, P. Leong, and S. Hyams, "The nature and distribution of errors in sound localization by human listeners," *Hearing Research*, vol. 114, no. 1-2, pp. 179–196, 1997.

[24] C. Giguère and S. M. Abel, "Sound localization: Effects of reverberation time, speaker array, stimulus frequency, and stimulus rise/decay," *The Journal of the Acoustical Society of America*, vol. 94, pp. 769–776, 1993.

[25] E. H. A. Langendijk and A. W. Bronkhorst, "Contribution of spectral cues to human sound localization," *The Journal of the Acoustical Society of America*, vol. 112, pp. 1583–1596, 2002.

[26] S. Carlile, S. Delaney, and A. Corderoy, "The localisation of spectrally restricted sounds by human listeners," *Hearing Research*, vol. 128, no. 1-2, pp. 175–189, 1999.

[27] B. F. G. Katz and M. Noisternig, "A comparative study of interaural time delay estimation methods," *The Journal of the Acoustical Society of America*, vol. 135, pp. 3530–3540, 2014.

[28] T. Hidaka, L. L. Beranek, and T. Okano, "Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls," *The Journal of the Acoustical Society of America*, vol. 98, pp. 988–1007, 1995.

[29] T. Okano, L. L. Beranek, and T. Hidaka, "Relations among interaural cross-correlation coefficient lateral fraction and apparent source width in concert halls," *The Journal of the Acoustical Society of America*, vol. 104, pp. 255–265, 1998.

[30] R. Mason, T. Brookes, and F. Rumsey, "Frequency dependency of the relationship between perceived auditory source width and the interaural cross-correlation coefficient for time-invariant stimuli," *The Journal of the Acoustical Society of America*, vol. 117, pp. 1337–1350, 2005.

[31] Y. Soeta, "Autocorrelation function mechanism for pitch salience and cross-correlation function mechanism for sound localization revealed by magnetoencephalography," *Proceedings of Meetings on Acoustics*, vol. 19, pp. 1–8, 2013.

[32] H. SAKAI, T. HOTEHAMA, Y. ANDO, N. PRODI, and R. POMPOLI, "Diagnostic system based on the human auditory–brain model for measuring environmental noise—an application to railway noise," *Journal of Sound and Vibration*, vol. 250, no. 1, pp. 9–21, 2002.

[33] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 99–102, 2001.

[34] K. Iida, *Head-Related Transfer Function and Acoustic Virtual Reality*. 152 Beach Road, 21-01/04 Gateway East, Singapore 189721: Springer, 1 ed., 2017.

[35] P. Zahorik, P. Bangayan, and V. Sundareswaran, "Perceptual recalibration in human sound localization: Learning to remediate front-back reversals," *The Journal of the Acoustical Society of America*, vol. 120, pp. 343–359, 2006.

[36] C. L. Searle, L. D. Braida, D. R. Cuddy, and M. F. Davis, "Binaural pinna disparity: another auditory localization cue," *The Journal of the Acoustical Society of America*, vol. 57, pp. 448–455, 1975.

[37] P. X. Zhang and W. M. Hartmann, "On the ability of human listeners to distinguish between front and back," *Hearing Research*, vol. 260, no. 1-2, pp. 30–46, 2010.

[38] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[39] H. Fletcher, "Auditory patterns," *Reviews Of Modern Physics*, vol. 12, pp. 47–66, 1940.

[40] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*. Munchen, Germany: Springer, 3 ed., 2006.

[41] B. Moore, *An Introduction to the Psychology of Hearing*. Cambridge, United Kingdom: Brill, 5 ed., 2008.

[42] F. E. Toole, *Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers and Rooms*. New York, United States of America: Routledge, 3 ed., 2016.

[43] E. Terhardt, "Calculating virtual pitch," *Hearing Research*, vol. 1, pp. 155–182, 1979.

[44] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psychoacoustical masking model for audio coding applications," *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1805–1808, 2002.

[45] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP Journal on Advances in Signal Processing volume*, pp. 1292–1304, 2005.

[46] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A low-complexity spectro-temporal distortion measure for audio processing applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1553–1564, 2012.

[47] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.

[48] B. C. J. Moore, "Masking in the human auditory system," *AES: Collected Papers on Digital Audio Bit-Rate Reduction*, pp. 9–19, 1996.

[49] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The Journal of the Acoustical Society of America*, vol. 74, pp. 750–753, 1983.

[50] K.-S. Lee and S.-P. Lee, "A real-time audio system for adjusting the sweet spot to the listener's position," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 2, pp. 835–843, 2010.

[51] J. Rose, P. Nelson, and B. Rafaely, "Sweet spot size of virtual acoustic imaging systems at asymmetric listener locations," *The Journal of the Acoustical Society of America*, vol. 112, pp. 1992–2002, 2002.

[52] D. B. Ward and G. W. Elko, "Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation," *IEEE Signal Processing Letters*, vol. 6, no. 5, pp. 106–108, 1999.

[53] T. Kariya and H. Kurata, *Generalized Least Squares*. Chichester, West Sussex: John Wiley and Sons Ltd, 1 ed., 2004.

[54] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Hoboken, New Jersey: John Wiley and Sons Ltd, 1 ed., 2004.

[55] E. C. Hamdan and F. M. Fazi, "A modal analysis of multichannel crosstalk cancellation systems and their relationship to amplitude panning," *Journal of sound and vibration*, vol. 490, pp. 1–21, 2021.

[56] O. Rojo, "Further bounds for the smallest singular value and the spectral condition number," *Computers and Mathematics with Applications*, vol. 38, no. 7, pp. 215–228, 1999.

[57] E. Habets, "Room impulse response generator," *Internal Report*, pp. 1–17, 2006.

[58] H. Kuttruff, *Room acoustics*. London, New York: Spon Press, 5 ed., 2009.

[59] T. Takeuchi and P. A. Nelson, "Robustness of the performance of the "stereo dipole" to head misalignment," *ISVR Technical Report*, vol. 285, pp. 1–30, 1999.

[60] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[61] P. Mackensen, U. Felderhoff, and G. Theile, "Obtaining binaural room impulse responses from b-format impulse responses using frequency-dependent coherence matching," *The Journal of the Acoustical Society of America*, vol. 105, no. 2, pp. 1343–1344, 1999.

[62] F. Menzer, C. Faller, and H. Lissek, "Binaural room scanning-a new tool for acoustic and psychoacoustic research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 396–405, 2011.

[63] Stan, Guy-Bart, Embrechts, Jean-Jacques, Archambeau, and Dominique, "Comparison of different impulse response measurement techniques," *JAES*, vol. 50, no. 4, pp. 249–262, 2002.

[64] W. T. Chu, "Impulse-response and reverberation-decay measurements made by using a periodic pseudorandom sequence," *Applied Acoustics*, vol. 29, pp. 193–205, 1990.

[65] A. J. Berkhout, D. de Vries, and M. M. Boone, "A new method to acquire impulse responses in concert halls," *The Journal of the Acoustical Society of America*, vol. 68, pp. 179–183, 1980.

[66] F. Wendt, F. Zotter, M. Frank, and R. Höldrich, "Auditory distance control using a variable-directivity loudspeaker," *Applied Sciences*, vol. 7, pp. 666–681, 2017.

[67] F. Zotter and M. Frank, "Investigation of auditory objects caused by directional sound sources in rooms," tech. rep., Institute of Electronic Music and Acoustics, Graz, 2014.

[68] H. Steffens, S. van der Par, and S. D. Ewert, "Perceptual relevance of speaker directivity modelling in virtual rooms," *Proceedings of the 23rd International Congress on Acoustics*, vol. 23, pp. 2651–2658, 2019.

[69] J. G. Tylka, R. Sridhar, and E. Choueiri, "a database of loudspeaker polar radiation measurements," *journal of the audio engineering society*, pp. 1–4, 2015.

[70] KEF, "Kef ls50 meta." Available at `https://nl.kef.com/products/ls50-meta` (2023/05/16).

[71] J. Ahrens and S. Bilbao, "Computation of spherical harmonics based sound source directivity models from sparse measurement data," tech. rep., Forum Acusticum, Lyon, 2020.

[72] J. Voight, *Quaternion Algebras*. Gewerbestrasse 11, 6330 Cham, Switzerland: Springer, 1 ed., 2021.

[73] T. Betlehem, W. Zhang, M. A. Poletti, and T. D. Abhayapala, "Personal sound zones," *IEEE signal processing magazine*, vol. 81, pp. 1–11, 2015.

[74] J. S. Soo and K. K. Pang, "Multidelay block frequency domain adaptive filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 2, pp. 373–376, 1990.

[75] D. de Groot, "A heuristic approach to spatial audio using consumer viable loudspeaker systems." Unpublished thesis, 2023.

[76] J. O. Smith, *Spectral audio signal processing*. "`http://ccrma.stanford.edu/~jos/sasp/`", accessed 31-05-2023. online book, 2023 edition.

[77] A. Jeannerot, N. de Koeijer, P. Martínez-Nuevo, M. B. Møller, J. Dyreby, and P. Prandoni, "Increasing loudness in audio signals: A perceptually motivated approach to preserve audio quality," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1001–1005, 2022.

[78] I. T. Union, "Recommendation bs.1387-1, method for objective measurements of perceived audio quality." Available at `https://www.itu.int/rec/R-REC-BS.1387-1-200111-I/en` (2023/05/21).

[79] P. Kabel, *An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality*. McGill university: Telecommunications and signal processing laboratory, 1 ed., 2003.

[80] N. Andersson, "Peaq." Available at `https://github.com/NikolajAndersson/PEAQ` (2023/05/21).

[81] A. W. Bronkhorst and T. Houtgast, “Auditory distance perception in rooms,” *Nature*, vol. 397, pp. 517–520, 1999.

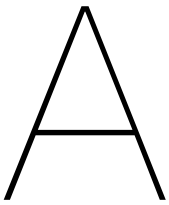
[82] D. H. Mershon and L. E. King, “Intensity and reverberation as factors in the auditory perception of egocentric distance,” *Perception and Psychophysics*, vol. 18, pp. 409–415, 1975.

[83] A. T. Sabin, E. A. Macpherson, and J. C. Middlebrooks, “Human sound localization at near-threshold levels,” *Hearing Research*, vol. 199, no. 1-2, pp. 124–134, 2005.

[84] G. Andéol, E. A. Macpherson, and A. T. Sabin, “Sound localization in noise and sensitivity to spectral shape,” *Hearing Research*, vol. 304, pp. 20–27, 2013.

[85] E. J. Golob, J. Lewald, J. Jungilligens, and S. Getzmann, “Interaction of number magnitude and auditory localization,” *Sage journal, Perception*, vol. 45, pp. 165–179, 2015.

[86] E. J. Golob, J. Lewald, S. Getzmann, and J. R. Mock, “Numerical value biases sound localization,” *Scientific Reports*, vol. 7, pp. 1–14, 2017.



paper ARIR

In the next pages, a draft version of the to be submitted paper can be found on the stochastic angular room impulse response model.

Perceptually Accurate Stochastic Angular Room Impulse Response

Dimme de Groot, Richard Eveleens, Arash Noroozi, and Jorge Martinez

***Abstract*—The room and its contents have a great impact on the behaviour of the sound field in a room. The reverberation time in a small and densely furnitured room can be roughly 100 ms, while the reverberation time in concert halls can be multiple seconds. Modelling the sound behaviour of a room is generally done using the Room Impulse Response (RIR). The RIR has many appliances in the field of audio processing. Think of, for example, concert hall design, commercial entertainment and localization. The estimation and modelling of the RIR is generally performed in a deterministic manner with techniques ranging from the idealistic image source method to highly complex room models. In this paper, a zeroth order stochastic model of the RIR is proposed for rooms that can be considered average living rooms. By means of randomly generated rooms, a large set of simulations is performed on which time-delay dependent distributions are derived. Additionally, it is shown that the directivity and transfer function of a loudspeaker has a great impact on the RIR and can thus not be ignored when modeling the RIR in a practical scenario.**

***Index Terms*—Room Impulse Response (RIR), Speaker directivity, Probability Density Function (PDF)**

I. Introduction

THE behaviour of a sound field in a room greatly depends on the shape and contents of the room. One way to characterize this influence of the room is the acoustic channel from a source to a receiver inside a room. This response is known as the Room Impulse Response (RIR), and it is of interest in many algorithms [1], [2], [3]. Examples include spatial sound [4], [5], [6], acoustic echo cancellation [3], [7], [8], [9], blind source separation [3], [9], speech dereverberation [3], [10], [11], [12] and beamforming [13].

Modelling the RIR has proven to be a computationally expensive task. Approaches include numerically solving the wave- or Helmholtz-equation with proper boundary conditions [14], [15] and geometrical acoustics [1], [16]. The former provides accurate RIRs, but is computationally too expensive to be used in real-time algorithms [2], [16]. Geometrical acoustic based approaches can be used in some real-time algorithms, but lack accuracy most pronounced in the low-frequency range [2], [16]. Additionally, properly modelling a (furnitured) room is difficult. All reflections' behaviour depends on the type of material, angle of incidence, and signal frequency [1]. On top of these problems, additional challenges are introduced by the loudspeaker and receiver directivity pattern that need to be taken into account [17] [18].

Due to the above mentioned challenges, algorithms may profit from a stochastic characterisation of the RIR. The RIR can be decomposed in three parts, the direct path, the early reflections and the (late) reverberation [1], [2]. The stochastic characterisation of the reverberation is well known and may be modelled using plane-waves which arrive from all directions [13], [19]. The resulting distribution can be approximated using a a Gaussian or logistic probability density function (pdf) [20]. Literature on the stochastic characterisation of the early reflections is limited (TODO: check ook dat de gaussian en logistic pdf wordt genoemd in literature terwijl wij laplace distribution vinden).

In this paper, we aim to stochastically characterise the impulse response for an isotropic receiver located in a small room. The stochastic characterisation is limited to a zeroth order Markov process. So, subsequent samples are considered to be independent. The RIR's are simulated using the mirror-image source method [21], [22] and a number of different sources of variation are considered.

In the following, we first describe the simulation setup in Section II. This gives rise to three different scenarios with increasing source of variation. The results of these simulations are presented in Section III-A and further analysed in Section III-B. We finalize with the conclusion in Section IV.

II. Simulation setup

The simulation setup described in the following is designed to limit any biases in the obtained data. This is done by identifying parameters of interest and randomising these in some range of interest. We consider box-shaped rooms with length L_x, width L_y and height L_z. The room has origin $(0, 0, 0)$ and its corners are given by non-negative coordinates. The six walls have reflection coefficients specified by $\beta_i \in \mathbb{R}$, $i \in \{1, \ldots, 6\}$. For each of the three simulations described below, eight isotropic receivers are considered. These receivers have a fixed location $\mathbf{x}_l = (x_l, y_l, z_l)$ given by $\mathbf{x}_l = (\{2, 3.5\}, \{1, 2.5\}, \{1, 1.8\})$ (m).

A. Simulation 1

In the first simulation, only the reflection coefficients are randomised. The room dimensions are $(L_x, L_y, L_z) = (5, 5, 2.5)$ (m) and the loudspeakers have fixed coordinates which are specified by a spherical coordinate system centered

Paper submitted on:... This work was performed in collaboration with the research team Kien

Dimme de Groot is with the Delft University of Technology, Delft 2628 CD The Netherlands (e-mail: dccjdegroot@student.tudelft.nl)

Richard Eveleens is with the Delft University of Technology, Delft 2628 CD The Netherlands (e-mail: reveleens@student.tudelft.nl)

Arash Noroozi is with Kien, Rotterdam 3013 AK The Netherlands (e-mail: arash@kien.io)

Jorge Martinez is with the Delft University of Technology, Delft 2628 CD The Netherlands (e-mail: J.A.MartinezCastaneda@tudelft.nl)

around $\mathbf{x}_l$. Using the coordinate convention of [23], the speaker locations are given by all possible combinations of

$$\begin{aligned} \theta_i &\in \{0, 20, \ldots, 340\}\,(^\circ), \\ \phi_j &\in \{60, 80, \ldots, 100\}\,^\circ), \\ r_k &\in \{1, 1.4, 2, 2.8, 4\}\,(\mathrm{m}). \end{aligned} \tag{1}$$

Depending on $\mathbf{x}_l$, some of the pairs (i, j, k) are discarded in accordance with the procedure described in Section II-B. The reflection coefficients are drawn from a uniform distribution $U(\cdot)$ according to

$$\begin{aligned} \beta_c &\sim U[0.5, 0.7] \\ \beta_{w_1}, \beta_{w_2} &\sim U[0.05, 0.5] \\ \boldsymbol{\beta} &= \text{shuffle}(\{\beta_c, \beta_{w_1}, \beta_{w_2}, -\beta_c, -\beta_{w_1}, -\beta_{w_2}\}), \end{aligned} \tag{2}$$

where β_c is a ceiling reflection coefficient, β_{w_1} and β_{w_2} are wall reflection coefficients. The range of the distribution is chosen to representing practical living rooms (bron). This definition of the reflection coefficients results into an average reflection amplitude of zero (waarom willen we dit).

B. Simulation 2

Simulation 2 adds a few randomizing factors to simulation 1, with first varying speaker locations. Per listener location, a set of speaker locations is drawn uniformly in each of the regions defined by

$$\begin{aligned} \theta_i &= [i, i+2)\,(^\circ), && i \in \{0, 20, \ldots, 340\}, \\ \phi_j &= [j, j+5)\,(^\circ), && j \in \{60, 80, \ldots, 100\}, \\ r_k &= [f(k), f(k+1))\,(\mathrm{m}), && k \in \{1, 2, \ldots, 5\}, \end{aligned} \tag{3}$$

with $f = \{1, 1.4, 2, 2.8, 4, 5.6\}$. Regions of which at least one of the corner points is located within 20 cm from at least one of the walls are discarded. The sizes of regions are based on the findings in psychoacoustic literature. Humans are best at localizing the azimuth direction θ ($\approx 2^\circ$ accuracy), worse at elevation direction ϕ ($\approx 5^\circ$ accuracy) and worst in distance r detection. The exact values depend on, among others, the type of signal [24], [25]. Uniformly sampling the regions on a Cartesian grid is achieved through the inversion method [26].

The rooms will also be varied in simulation 2. The size of the rooms is chosen to reflect typical listening rooms. The room-dimensions are drawn according to

$$\begin{aligned} L_x &\sim U[5.0, 7.0] \quad (\mathrm{m}) \\ L_y &\sim U[5.0, 7.0] \quad (\mathrm{m}) \\ L_z &\sim U[2.5, 3.0] \quad (\mathrm{m}). \end{aligned} \tag{4}$$

To randomize the room placement, the origin of the room is shifted from $(0, 0, 0)$ to $(x_0, y_0, 0)$. The value (x_0, y_0) is drawn according to

$$(x_0, y_0) \sim (U[-L_x + 5, 0],\, U[-L_y + 5, 0]) \quad (\mathrm{m}). \tag{5}$$

The origin in z-direction is not varied since a listener standing on the floor is considered in this setup.

C. Simulation 3

Simulation 3 is equal to simulation 2 with the addition of a directive loudspeaker. The normal (point of maximum gain) of the loudspeaker is always set such that it points towards the listener location. Note that this implies that it varies per speaker region and even per drawn speaker. The loudspeaker considered in our simulation is the KEF LS50. The directivity patterns are obtained through the implementation provided by [27], where a spherical harmonics representation is fitted on sparse measurement data provided by [28] to form a complete directivity pattern.

D. Implementation Details

The simulations were performed on MATLAB R2021b with default settings. For each of the simulations and for each loudspeaker-listener pair, at least 60000 runs are considered. Since the simulations were collected over multiple runs, the random number generator was set to "shuffle". The RIRs were simulated through a modified version of [22]. The modified code returns individual reflections and the incidence and outgoing angles which are combined to form a RIR. No high pass filter is applied. As further explained in [22], the sampling occurs through a Hanning-windowed ideal low-pass filter with a length of 8 ms and a cutoff-frequency $f_s/2$ with $f_s = 16$ kHz the sampling frequency. The speed of sound was set to $c = 342$ m/s. For the KEF LS50, a linear interpolation object with 5° resolution was created based on the directivity pattern. Downsampling was done using the resample function. The length of the considered impulse responses is 170 ms, which captures the majority of the possible rooms their T_{60}. The calculated RIR is normalized by shifting and scaling the response such that the reflection corresponding to the direct-path starts at the same time-sample and has its magnitude multiplied by $4\pi r_{l,s}$, with $r_{l,s}$ the distance between the transmitter and receiver.

For each simulation and speaker-receiver pair, each time-sample n of the computed RIRs corresponds to one histogram. The histogram limits are given by a time-sample dependent range which is derived based on the RIR energy curves e presented in [29]. The resulting curve is shown in Fig. 1. The centers of the 301 histogram bins are linearly spaced between $[-e(n), e(n)]$, so that bin 151 serves as the zero amplitude bin.

III. Simulation results

The results of the three simulations are shown in Fig. 2 and 3. In both figures, the number of the simulation is given by the number of the column. In Fig. 3, the row number corresponds to a time index. In Fig. 2 the simulation results are presented by time-dependent histograms. Do note that the y-axis does not denote the value of the bin-center. Instead, it denotes the bin number. At any time instant, as given by the x-axis, the amplitude corresponding to a given bin number may vary. In Fig. 3 we zoom in on a few selected histograms (denoted by the red dotted lines in Fig. 2). In here, the amplitude corresponding to the histogram bin is given.

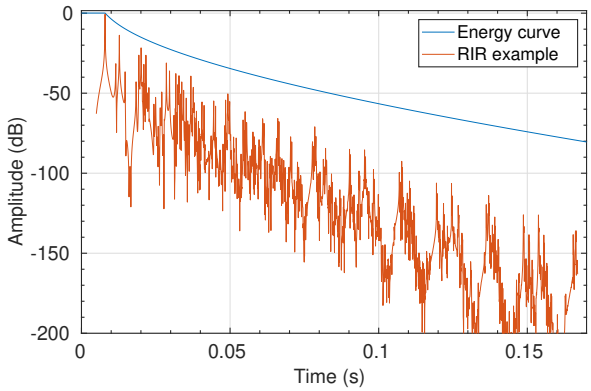


Fig. 1. The energy curve e on which the range of the histograms is based. An example RIR is added for reference

A. Interpreting simulation results

The results from simulation 1 clearly show the almost absence of randomness. As seen in Fig. 3, at $t = 0.02$ s, an early reflection arrives at the receiver, hence it is not possible for the amplitude to be zero. The two step shape of this plot also shows the two reflection coefficients sizes, where a lower coefficient (β_ω) occurs more often than a higher coefficient (β_e). At the late reverberation, $t = 0.07$ s and $t = 0.12$ s, the distribution becomes more random as is predicted in (bron).

Both the histogram and the highlighted distributions of simulation 2 show a great increase in randomness. Apart from the direct path, all time samples show a zero-centered random behaviour. Especially at $t = 0.02$, the increased randomness from simulation 2 compared to simulation 1 can be found. The hypothesis is that the time sample dependent distributions obtained from simulation 2 serve as the fundamental distributions for the RIR of rooms fitting the room type considered.

Simulation 3 serves as a small but interesting sidestep to simulation 2. In simulation 3, the directivity and the directive transfer functions from a speaker, the KEF LS50's, are added to the equation. The histogram in Fig. 2 of simulation 3 shows the great impact of the characteristics of a speaker on the RIR. Although concise conclussions can not be drawn on the exact influence of speakers on the RIR, it is clear that the speakers' directivity and transfer function can not be ignored when modelling the RIR in a practical scenario.

B. Estimating RIR distributions

The time-sample dependent histograms, as depicted in Fig. 3, can be used to derive Probability Density Functions (PDF) for each time sample. The PDF's will be derived for the data from simulation 2. Three possible distributions are selected to be fit on the data: the Laplace distribution, the normal distribution and the logistic distribution. The three distributions are fitted on the histograms and the best fitting distribution is selected by comparing the L_2-norms of the difference between PDF and the data. Doing this for all the obtained data shows that the direct path response (response up until $t \approx 0.01$ s) is best described by a normal distribution and the reverberations are best described by a Laplace distribution. (figure required to show this?)

An example of the mean and standard deviation of a speaker-receiver pair is given in Fig. 4. The figure shows that, apart from the direct path, the mean of the distributions is zero. This is expected and validates that the definition of the reflection coefficients results in a zero mean RIR on average. The standard deviation shows a clear pattern in the early reflections (response up until $t \approx 0.025$ s) after which a logarithmic decay is observed.

An example of a generated RIR based on the proposed model is shown in Fig. 5. The figure shows that the major difference between the proposed model and the deterministic model is that the deterministic response consists of a few distinct peaks while the proposed model shows a more smoothed out response. The behaviour of the standard deviation before $t \approx 0.025$ s as presented in Fig. 4 can be found in Fig. 5.

IV. Conclusion

In this paper, a stochastic Room Impulse Response (RIR) is introduced that is applicable for any furnitured and decorated room that can be characterised as standard shoebox living room. Classic approaches to derive a RIR are based on deterministic simulations or models that require precise prior knowledge on the room properties. In practical room scenarios, this prior knowledge is generally not available and expensive to obtain. A more general and widely applicable model of the RIR is desirable in this case. Deriving the stochastic RIR is performed by simulating a large amount of deterministic RIR's and fitting a probability density function on the set of simulated RIR's. The simulated RIR's show a direct path response best described by a normal distribution and the response caused by the reflections are best described by a Laplace distribution. Additionaly, it is shown that the influence of the directivity and response of the speaker on the RIR can not be ignored due to its significant impact.

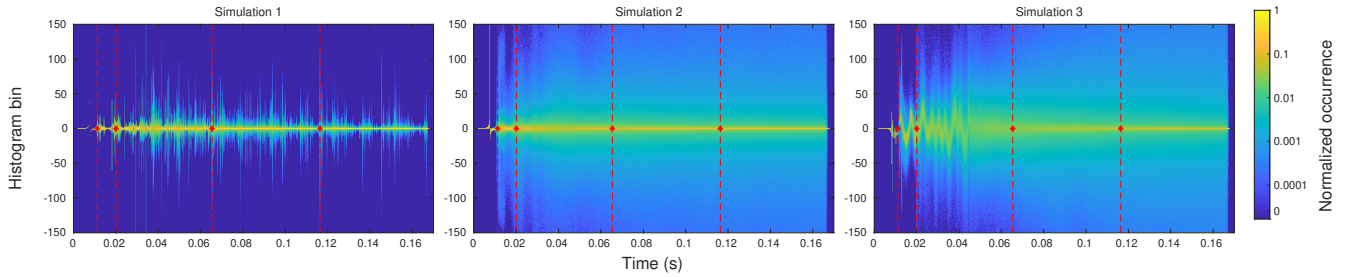


Fig. 2. The histograms resulting from the simulation. The colorbar shows a logarithmically scaled normalized occurrence of a certain histogram bin. Note that the y-axis represents the histogram bin number, this bin number should be translated to a time dependent amplitude by means of the energy curve found in Fig. 1. This is done for a few time samples in Fig. 3.

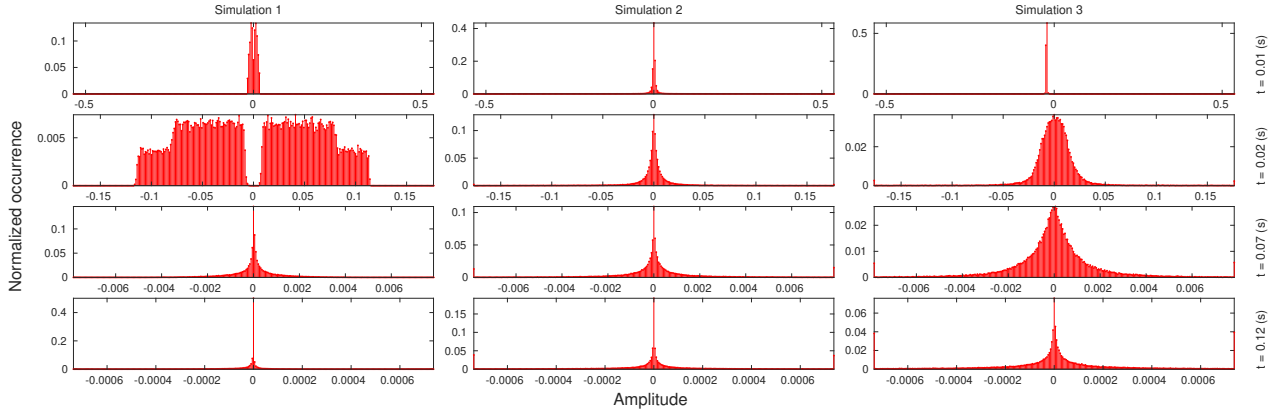


Fig. 3. The amplitude - normalized occurrence plots of a few time samples from Fig. 2 (indicated by the red-dotted line). The amplitude distribution of each time sample of the RIR for each simulation is estimated using these amplitude occurrences.

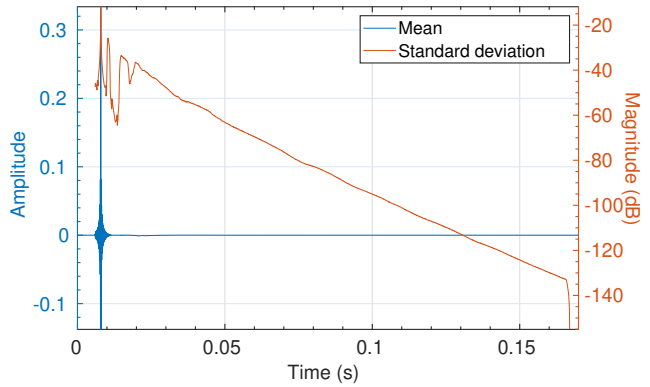


Fig. 4. An example of the time-dependent mean and standard deviation of one speaker-receiver pair. As expected, the mean is zero apart from the direct path. The standard deviation shows a clear pattern in the early reflections and shows a logarithmic delay afterwards.

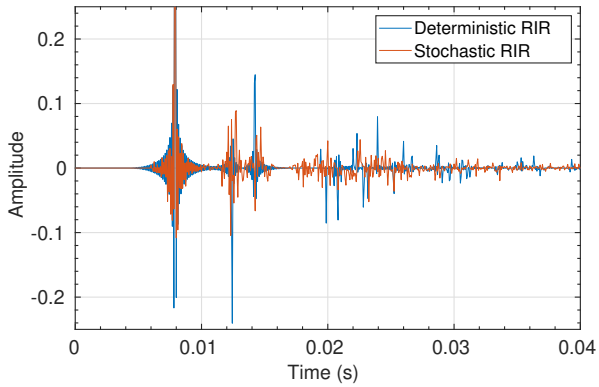


Fig. 5. A zoomed in example of a generated RIR with the proposed stochastic model and a single deterministic RIR from the same region. The major difference between the proposed solution and the deterministic model is that the deterministic model results in a few distinct peaks in the response while the proposed stochastic model has a more smoothed out response which should represent the generalized RIR for different but similar speaker-receiver pairs.

REFERENCES

[1] H. Kuttruff, *Room Acoustics*, 5th ed, Taylor & Francis, 2009.
[2] J. Martinez, "Low-complexity computer simulation of multichannel room impulse responses," Ph.D. dissertation, Delft Univ. Technol., Delft, The Netherlands, 2013.
[3] Y. Huang, J. Benesty and J. Chen, *Acoustic MIMO Signal Processing*, 1st ed, New York, NY, USA: Springer, 2006.
[4] F. Melchior, "Investigations on spatial sound design based on measured room impulse responses," Ph.D. dissertation, Delft Univ. Technol., Delft, The Netherlands, 2011.
[5] M. Kolundzija, C. Faller, M. Vetterli. (2009, May). Sound field reconstruction: an improved approach for wave field synthesis. presented at AES Conv. 126.
[6] E. C. Hamdan, F. M. Fazi, "A modal analysis of multichannel crosstalk cancellation systems and their relationship to amplitude panning," *Journal of Sound and Vibr.*, vol. 490, pp. 115473, 2021, DOI. https://doi.org/10.1016/j.jsv.2020.115743.
[7] H. Buchner, J. Benesty, W. Kellermann, "Generalized multichannel

frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication," *Signal Processing*, vol. 85, pp. 549-570, 2005. DOI. https://doi.org/10.1016/j.sigpro.2004.07.029.

[8] K. Nathwani, "Joint acoustic echo and noise cancellation using spectral domain Kalman filtering in double-talk scenario" in IWAENC2018, Tokyo, Japan, 2018, pp. 326–330.

[9] M. Souden, Z. Liu, "Optimal joint linear acoustic echo cancelation and blind source separation in the presence of loudspeaker nonlinearity" in ICME, New York City, NY, USA, 2009, pp. 117–120.

[10] E.A.P. Habets, S. Gannot, "Dual-microphone speech dereverberation using a reference signal" in ICASSP '07, Honolulu, HI, USA, 2007, pp. 901–904.

[11] J. Zhang, M. D. Plumbley, W. Wang, "Weighted magnitude-phase loss for speech dereverberation" in ICASSP '21, Toronto, ON, Canada, 2021, pp. 5794–5798.

[12] P. A. Naylor, N. D. Gaubitch, *Speech Dereverberation*, 1st ed, Springer, 2010.

[13] J. Martinez, N. Gaubitch, W. B. Kleijn, "A robust region-based near-field beamformer" in ICASSP '15, South Brisbane, QLD, Australia, 2015, pp. 2494–2498.

[14] L. Thompson, "A review of finite-element methods for time-harmonics acoustics" *The Journal of the Acoustical Society of America*, vol. 119, pp. 1315-1330, 2006. DOI. https://doi.org/10.1121/1.2164987.

[15] R. Mehra, N. Raghuvanshi, L. Savioja, M. C. Lin, D. Manocha, "An efficient GPU-based time domain solver for the acoustic wave equation" *Applied Acoustics*, vol. 73, pp. 83-94, 2012. DOI. https://doi.org/10.1016/j.apacoust.2011.05.012.

[16] L. Savioja, U. P. Svensson, "Overview of geometrical room acoustic modeling techniques" *The Journal of the Acoustical Society of America*, vol. 138, pp. 708-730, 2015. DOI. https://doi.org/10.1121/1.4926438.

[17] F. Zotter, M. Frank, "Investigation of auditory objects caused by directional sound sources in rooms" *Acoustical Engineering*, vol. 128, pp. A5-A10, 2015. DOI. https://doi.org/10.12693/APhysPolA.128.A-5.

[18] H. Steffens, S. van der Par, S. D. Ewert "Perceptual relevance of speaker directivity modelling in virtual rooms" in ICA2019, Aachen, Germany, 2019, pp. 2651–2658.

[19] T. D. Abhayapala, R. A. Kennedy, R. C. Williamson "Isotropic noise modelling for nearfield array processing" in WASPAA'99, New Paltz, NY, USA, 1999, pp. 11-14.

[20] E. A. Lehmann, A. M. Johansson, "Diffuse Reverberation Model for Efficient Image-Source Simulation of Room Impulse Responses" *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1429-1439, 2010. DOI. https://doi.org/10.1109/TASL.2009.2035038.

[21] J. B. Allen, D. A. Berkley, "Image method for efficiently simulating small-room acoustics" *The Journal of the Acoustical Society of America*, vol. 65, pp. 943-950, 1979. DOI. https://doi.org/10.1121/1.382599.

[22] E. A. P. Habets, "Room Impulse Response Generator", Internal Report, pp. 1-17, 2006.

[23] J. Stewart, D. Clegg, S. Watson, *Calculus: Early Transcendentals*, 9th ed. Metric Version, Cengage, 2021

[24] T. R. Letowski, S. T. Letowski, "Auditory Spatial Perception: Auditory Localization", ARL, Aberdeen P.G., MD, Scotland, Tech. Rep. ARL-TR-6016, May. 2012

[25] J. C. Middlebrooks, D. M. Green, "sound localization by human listeners" *Annual review of psychology*, vol. 42, pp. 135-159, 1991. DOI. https://doi.org/10.1146/annurev.ps.42.020191.001031.

[26] L. Devroye, *Non-Uniform Random Variate Generation*, New York, NY, USA: Springer, 1986.

[27] J. Ahrens, S. Bilbao, "Computation of Spherical Harmonics Based Sound Source Directivity Models from Sparse Measurement Data" *Forum Acusticum*, pp. 2019-2026, 2020. DOI. https://doi.org/10.48465/fa.2020.0042.

[28] J. G. Tylka, R. Sridhar, E. Y. Choueiri "A database of loudspeaker polar radiation measurements" in AES convention 139, New York, NY, USA, 2015, pp. 1-4.

[29] E. A. Lehmann, A. M. Johansson, "Prediction of energy decay in room impulse response simulated with an image-source model" *The Journal of the Acoustical Society of America*, vol. 124, pp. 269-277, 2008. DOI. https://doi.org/10.1121/1.2936367.

B

Auditory localization extra findings in literature

B.1. Distance estimation

Distance estimation is not our strongest trait as performance is generally considered worst when compared to azimuth and elevation estimation. The main reason for this is that we do not have the sensors required to measure distance based on sound only. To be able to do some distance estimation we apply some simple tricks.

First of all the measured SPL can be used to estimate the distance with a lower SPL corresponding to a further away distance. The major difficulty here is that prior knowledge is required on the volume of the sound source itself, which is generally not the case of course [15].

In the special case of a source being present to the left or right of us, the ITD can also be used to perform some distance estimation [15].

The above mentioned solutions work given that the subject is present in free space. When the subject is in a room with reflective walls it becomes easier to perform distance estimation. It is found that the ratio between direct path sound energy and reflective path energy is used to perform distance estimation [81, 82]. Generally, if the direct path energy is much higher than the reflective path energy, the source is close by and when the reflective path energy increases relative to the direct path, the source is further away.

B.2. Specific finds on auditory localization

In addition to the findings discussed thus far, there are a few more interesting properties in human auditory localization that are worthwhile to discuss even though they will have no significant impact on the rest of the thesis.

The sound detection threshold, the SPL after which a sound source is audible, is dependent of the azimuth angle to the sound source with a maximal difference of 8 dB SPL [83].

A sound source placed in an environment containing a broad band noise stimulus can be detected with an SNR of about -8 dB or higher. Auditory localization

performance is poor at this point but becomes nominal after increasing the SNR with about 6 dB [20, 84].

The sense of motion and our perception of it is still up for discussion. Two popular theories exist, the snapshot and continuous motion theory. The snapshot theory states that we take frames of audio and than unconsciously perceive it as continuous, just like our vision works and why televisions work. The continuous motion theory states that we continuously measure and estimate the position of the sound source. Which of these theories holds true was not found [14, 15].

Finally, an interesting out of the box find. Our localization performance can be biased by saying numbers prior to a localization task. If we hear a ”one” being pronounced before we localize a sound source, our localization is biased to the left since we expect a ”one” to be present at the left. The opposite holds true for a pronounced ”nine” [85, 86]. This indicates that our localization is most likely influenced by many different unexpected factors inducing bias in our localization performance.

C

Proposed algorithm implementation details

In this appendix, a few implementation problems are emphasized and the implemented solutions are shown.

C.1. Outgoing loudspeaker angles and quaternions

Calculating the 3D orientation of a speaker for a reflection path is a challenging task in the image-source method framework. To indicate the challenging aspect, the functioning of the image-source method is shown in 2D in Figure C.1, which is partly a repetition of Section 4.2. In the figure, the white rectangle in the middle represents the physical room including the receiver and the physical loudspeaker. The black arrow indicates the sound path of the direct path response from physical speaker to physical receiver. The other gray shaded rooms represent the functioning of the image-source method, they are folded versions of the original room and speaker to make it easier to determine the reflection paths. Because of these virtual rooms, an actual reflection path with properly placed wall reflections and proper angles does not have to be determined. A straight arrow from the speaker in the virtual room to the physical receiver will give all the information required. The reflection sound paths are represented by the coloured arrows and as can be seen from the figure, the distance the wave travels and the amount of walls it passes can be easily determined. This method does, however, not provide a straight-forward way to determine the outgoing direction of sound from the virtual loudspeakers.

As shown in Figure C.1, the outgoing direction of the sound from the loudspeaker is different for every reflection path. The position of the reflection loudspeaker and the receiver is known but due to all the reflections occurring, the angle can not be calculated efficiently given this information. This problem gets even more challenging when expanding to the 3D field as is required for the implementation.

An elegant and efficient way of calculating the outgoing angles of reflection paths is found in quaternions. Quaternions are an extension to imaginary numbers by the addition of two extra imaginary numbers. These extra imaginary numbers allow for an efficient computation of 3D angles. Applying the quaternion theory

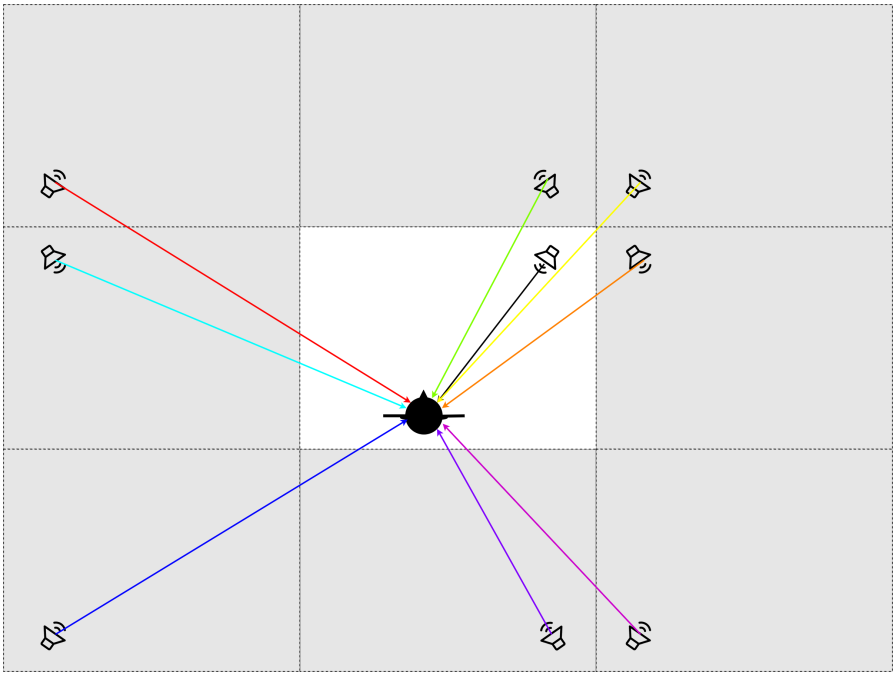


Figure C.1: Representation of the image-source method. The white rectangle in the middle of the picture represents the physical room including the receiver and loudspeaker. The gray rectangles around this room are folded versions of the original room that make the computation of reflections paths straight-forward. The coloured arrows indicate the sound paths of reflected sound waves. As can be seen, due to the folded rooms, it is easy to determine the traveled distance of the sound wave and also the number of walls it reflected from.

to the outgoing loudspeaker angle is possible since we have knowledge on the reflection history of the sound wave. We know exactly how often a sound wave has reflected from every wall in the room. This information is conveniently found in the code used for the image-source code method [57]. In this appendix, the knowledge of quaternions required for the implementation are discussed after which the implementation to obtain the outgoing loudspeaker angles is given.

C.1.1. Quaternions and basic operations

Quaternions and the corresponding algebra provided in this section are primarily based on [72]. Quaternions are an extension to the complex numbers, they are denoted as given in Equation (C.1).

$$a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} \tag{C.1}$$

In the definition, a, b, c and d are real numbers and $\mathbf{1}$, $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are the basis vectors spanning the 4D quaternion space. Important multiplication properties of the basis elements are given in Equation (C.2).

$$\begin{aligned} &\mathbf{i1} = \mathbf{1i} = \mathbf{i}, \qquad \mathbf{j1} = \mathbf{1j} = \mathbf{j}, \qquad \mathbf{k1} = \mathbf{1k} = \mathbf{k} \\ &\mathbf{ij} = -\mathbf{ji} = \mathbf{k}, \qquad \mathbf{jk} = -\mathbf{kj} = \mathbf{i}, \qquad \mathbf{ki} = -\mathbf{ik} = \mathbf{j} \\ &\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -\mathbf{1} \end{aligned} \tag{C.2}$$

These multiplication rules expand to the quaternion multiplication given in Equation (C.3).

$$
\begin{aligned}
a_3 + b_3\mathbf{i} + c_3\mathbf{j} + d_3\mathbf{k} &= (a_1 + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k})(a_2 + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k}), \quad \text{with} \\
a_3 &= a_1a_2 - b_1b_2 - c_1c_2 - d_1d_2 \\
b_3 &= a_1b_2 + b_1a_2 + c_1d_2 - d_1c_2 \\
c_3 &= a_1c_2 - b_1d_2 + c_1a_2 + d_1b_2 \\
d_3 &= a_1d_2 + b_1c_2 - c_1b_2 + d_1a_2
\end{aligned} \tag{C.3}
$$

Another important operation is the inverse of a quaternion, it is given by Equation (C.4).

$$
(a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k})^{-1} = \frac{1}{a^2 + b^2 + c^2 + d^2}(a - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}) \tag{C.4}
$$

Determining the outgoing loudspeaker angles can be done with the above described operations as is explained in the next section.

C.1.2. Outgoing loudspeaker angles

For each reflection path, the virtual loudspeaker location, $\mathbf{l}_s \in \mathbb{R}^{3\times1}$, and the physical receiver location, $\mathbf{l}_l \in \mathbb{R}^{3\times1}$, is known. Given these locations we can derive the vector corresponding to the traveled path of the sound wave, this vector corresponds to the outgoing sound direction from the (virtual) loudspeaker. It is given by $\mathbf{d} \in \mathbb{R}^{3\times1}$ and calculated as $\mathbf{d} = \mathbf{l}_l - \mathbf{l}_s$. The quaternion required to perform quaternion algebra based on $\mathbf{d}$ is given in Equation (C.5).

$$
d = 0 + d_x\mathbf{i} + d_y\mathbf{j} + d_z\mathbf{k} \tag{C.5}
$$

As shown in Figure C.1, instead of drawing the actual sound path including reflections, a straight direction is drawn through a set of folded rooms. When folding a room, the loudspeaker and its direction is also folded. We can interpret this as a folding of loudspeaker direction for each wall passing. For instance, the top left virtual loudspeaker in Figure C.1 has been folded once to the left and once to the top. Implementing a fold and applying this fold to the outgoing direction vector is done by means of quaternion algebra.

A reflection crosses a wall in either x-, y- or z-direction. For each of these three cases, a rotation matrix is defined. When a reflection crosses a wall in x-direction, meaning that it crosses a wall with a normal in x-direction, the direction of the outgoing vector is rotated around the y-axis. The corresponding angle at which it should rotate is determined by d_x and d_z. Since we are treating a reflection of an x-direction wall, the z component must remain unchanged and the x component should be mirrored over the z-axis. To do this, the angle between the x and z component is calculated by $\phi = \tan(\frac{d_x}{d_z})$. The direction vector is rotated twice around the y-axis with this angle. An example of this is given in Figure C.2.

The quaternions used to implement this operation is given in Equation (C.6). With similar derivation, the rotation quaternion for y- and z-direction can be derived. They are given in Equation (C.7) and (C.8) respectively. In here $\text{atan2}(\cdot)$ is the 2-argument arctangent function as defined in Equation (C.9).

$$
r_x = \cos(\phi) + \sin(\phi)\mathbf{j}, \quad \text{with} \quad \phi = \text{atan2}(\frac{d_x}{d_z}) \tag{C.6}
$$

$$
r_y = \cos(\phi) + \sin(\phi)\mathbf{k}, \quad \text{with} \quad \phi = \text{atan2}(\frac{d_y}{d_x}) \tag{C.7}
$$

$$
r_z = \cos(\phi) + \sin(\phi)\mathbf{i}, \quad \text{with} \quad \phi = \text{atan2}(\frac{d_z}{d_y}) \tag{C.8}
$$

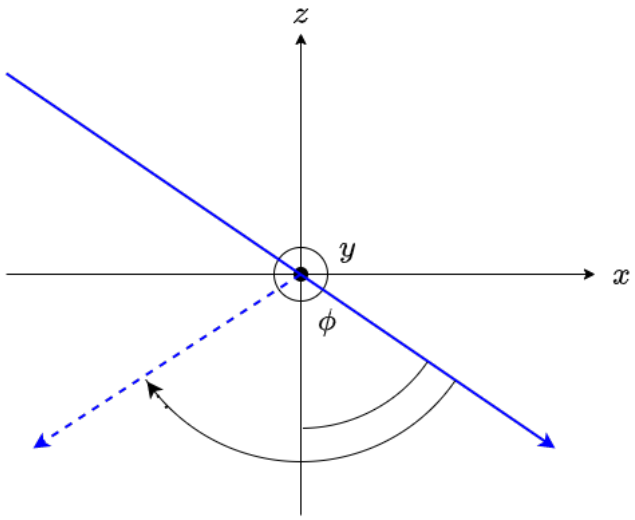


Figure C.2: Example of a quaternion rotation.

$$\text{atan2}(y,x) = \begin{cases} \text{atan}\left(\frac{y}{x}\right), & x > 0 \\ \text{atan}\left(\frac{y}{x}\right) + 180, & x < 0 \text{ and } y \geq 0 \\ \text{atan}\left(\frac{y}{x}\right) - 180, & x < 0 \text{ and } y < 0 \\ 90, & x = 0 \text{ and } y > 0 \\ -90, & x = 0 \text{ and } y < 0 \\ \text{undefined}, & \text{else} \end{cases} \tag{C.9}$$

Given these rotation quaternions, we can apply the rotation on the direction quaternion by means of Equation (C.10).

$$\bar{d} = rdr^{-1} \tag{C.10}$$

We know how many times an x-, y- and z-direction reflection occurs for each reflection path. Applying the corresponding reflections to the direction quaternion gives the correct outgoing sound wave direction from the loudspeaker, denoted by $\bar{d}$.

Calculating the azimuth and elevation angles corresponding to this direction quaternion is not a trivial task. The angle definition of the loudspeakers is the same as the one used for the listener as shown in Figure 2.1 and repeated in Figure C.3. The double defined elevation angle causes problems when calculating the angles. Another challenge is that all the angles have to be calculated with respect to the direction of the loudspeaker in the physical room. The direction of the physical loudspeaker is set such that the front of the loudspeaker faces the listener, this direction is defined by quaternion d_l.

To calculate the correct angles relative to d_l, rotations are applied such that d_l is placed exactly like the system shown in Figure C.3. Reflection loudspeaker directivity $\bar{d}$ will be rotated with the same rotations.

In the implementation, d_l and $\bar{d}$ are rotated among the z- and x-axes with proper angles such that d_l has the same orientation as the one given in Figure C.3. This orientation is denoted by the quaternion $d_l = 0+\mathbf{j}$. With the obtained orientation, we can calculate the angles corresponding to $\bar{d}$. It is shown in Equation (C.11), where the two cases for the azimuth angle θ are caused by the double defined elevation angle ϕ.

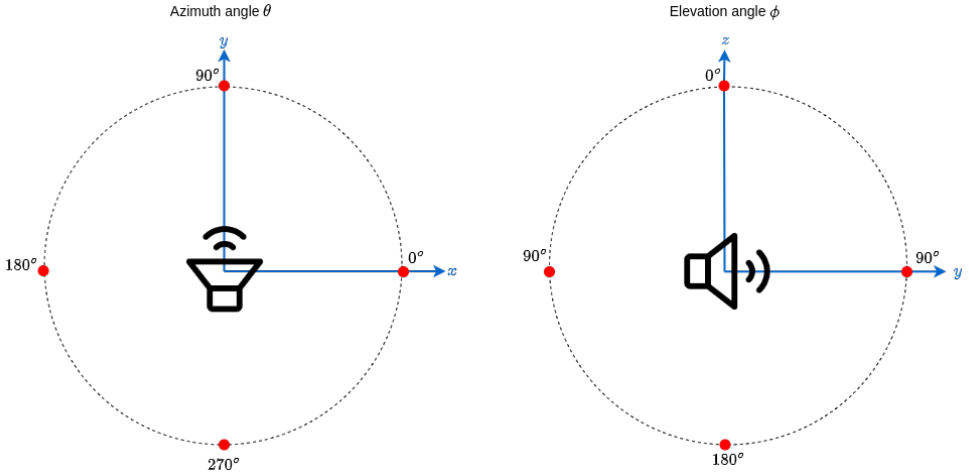


Figure C.3: Angle definitions loudspeaker.

$$\theta = \begin{cases} \text{atan2}\left(\frac{\bar{d}_y}{\bar{d}_x}\right), & \bar{d}_z \geq 0 \\ \text{atan2}\left(\frac{\bar{d}_y}{\bar{d}_x}\right) + 180°, & \bar{d}_z < 0 \end{cases}$$
$$\phi = \text{atan2}\left(\frac{\sqrt{\bar{d}_x^2 + \bar{d}_y^2}}{\bar{d}_z}\right) \tag{C.11}$$

C.2. Difference between channel and received response interaural crosscorrelation

In the proposed algorithm, an attempt was made to relax the optimization of the InterAural CrossCorrelation (IACC) of the received responses $y_{\text{L}}(n)$ and $y_{\text{R}}(n)$ for left and right ear respectively. This IACC would be replaced by the IACC of the received channel at the left and right ear given by $c_{\text{L}}(n)$ and $c_{\text{R}}(n)$ respectively. Here we show why this method did not succeed.

The received response for the left and the right ear is described by Equations (C.12) and (C.13) respectively. Here, $s(n)$ denotes the audio stimulus and n is the discrete time index. Note that the vector notation is replaced with a discrete signal notation in this section, which simplifies the derivation.

$$y_{\text{L}}(n) = c_{\text{L}}(n) * s(n) \tag{C.12}$$

$$y_{\text{R}}(n) = c_{\text{R}}(n) * s(n) \tag{C.13}$$

Given these definitions, the IACC responses for the received response and the channel response are given in Equation (C.14) and (C.15) respectively.

$$\begin{aligned} y_{\text{IACC}}(n) &= y_{\text{L}}(n) * y_{\text{R}}(-n) \\ &= c_{\text{L}}(n) * s(n) * c_{\text{R}}(-n) * s(-n) \\ &= c_{\text{L}}(n) * c_{\text{R}}(-n) * s(n) * s(-n) \end{aligned} \tag{C.14}$$

$$c_{\text{IACC}}(n) = c_{\text{L}}(n) * c_{\text{R}}(-n) \tag{C.15}$$

The term $s(n) * s(-n)$ found in Equation (C.14) represents the autocorrelation of $s(n)$. The autocorrelation is a symmetric signal around $n = 0$ and the peak of the

response is located at $n = 0$. The term $c_{\mathrm{L}}(n) * c_{\mathrm{R}}(-n)$ is present in both IACC equations with the only difference being the addition of the autocorrelation of $s(n)$ in Equation (C.14). Due to the autocorrelation properties, the assumption was made that $s(n) * s(-n)$ would not have impact on the peak location of the IACC peak. Given this assumption holds, $c_{\mathrm{IACC}}(n)$ could be considered for IACC peak index optimization instead of $y_{\mathrm{IACC}}(n)$.

The abovementioned assumption proved to be too unreliable to be used in the optimization. An example to show this is given in Figure C.4 and the inequality is presented in Equation (C.16). In the figure, the inequality is shown with a simple example. The received IACC response is determined by convolving the audio stimulus autocorrelation with the channel IACC respone, as per Equation (C.14). The figure shows that the assumption does not hold in general and throughout the development of the proposed algorithm it became clear that this assumption is not reliable enough to be used in the optimization.

$$\arg\max_n \quad y_{\mathrm{IACC}}(n) \neq \arg\max_n \quad c_{\mathrm{IACC}}(n) \tag{C.16}$$

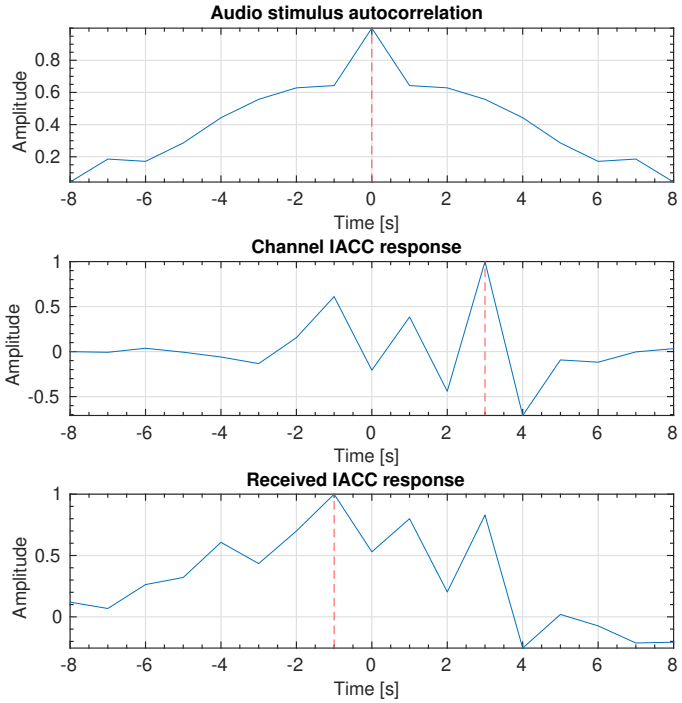
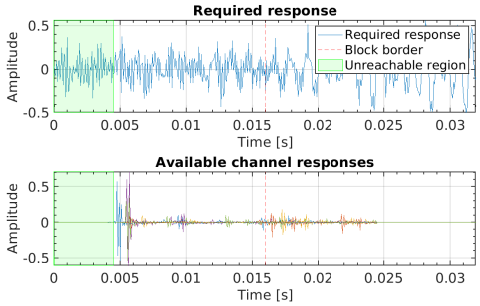


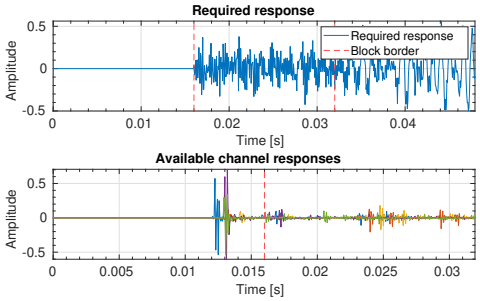
Figure C.4: Interaural crosscorrelation difference channel and received response. The received IACC response is determined by convolving the audio stimulus autocorrelation with the channel IACC respone, as per Equation (C.14). The figure shows that the peak index of the IACC response can be different for $c_{\mathrm{IACC}}(n)$ and $y_{\mathrm{IACC}}(n)$.

C.3. ”Unreachable” target signal

One problem when optimizing small time frames is the general shape of the channel response. The channel response only contains a few significant peaks that will mainly be used by the optimizer to achieve a satisfying result. A problem with this is that these peaks only start relatively late, because of this, the desired output



(a) Due to the near-zero valued samples in the green unreachable region, the required response cannot be obtained in this area



(b) With the addition of a zero block and slight shift of the channel response, all parts of the required response are reachable

Figure C.5: Illustration of unreachable response in MDF framework.

samples before this first significant peak can not be ”reached” by the optimized signal. This is shown by means of an example in Figure C.5a.

To fix this, the required signal is zero padded with a full block length at the beginning and the first significant channel peak is shifted such that it lays somewhere around one block length. The downside to this is that the optimization is now done over three block lengths instead of two. By extra zero padding we still make sure that the optimized x only has non zero samples in the first two time blocks. The adjustment of the example in Figure C.5a is shown in Figure C.5b. The definition of the the required response $\bar{\mathbf{y}} \in \mathbb{R}^{3L_b \times 1}$ is now as given in Equation (C.17) with $\mathbf{0} \in \mathbb{R}^{L_b \times 1}$.

$$\bar{\mathbf{y}} = [\mathbf{0} \quad \mathcal{Y}^*_{[n_i, n_i+1]} - \mathcal{Y}_{[n_i, n_i+1]}] \tag{C.17}$$

C.4. Implementation optimization function

Implementing the proposed algorithm as given in Equation (5.22) requires zero padding and different definitions for efficient implementation. The optimization function in Equation (5.22) is repeated with the original variable sizes as described in Chapter 5. The optimization problem given in Equation (C.18) can be evaluated given the required responses $\bar{\mathbf{y}}_{\mathrm{L}} \in \mathbb{R}^{2L_b \times 1}$ and $\bar{\mathbf{y}}_{\mathrm{R}} \in \mathbb{R}^{2L_b \times 1}$ where L_b is the block length, the desired InterAural CrossCorrelation (IACC) time delay index τ^*_{IACC} and the channel responses $\mathbf{c}_{\mathrm{L},n_s,n_l} \in \mathbb{R}^{2L_b \times 1}$ and $\mathbf{c}_{\mathrm{R},n_s,n_l} \in \mathbb{R}^{2L_b \times 1}$ for $n_s = 1, ..., N_s$ and $n_l = 1, ..., N_l$ with N_s the number of loudspeakers and N_l the number of optimization points with $n_l = 1$ the main centre optimization point.

$$
\begin{aligned}
\arg \min_{\mathbf{s_1},\mathbf{s_2},...,\mathbf{s_{N_s}}} \quad & 0 \\
\text{s.t.} \quad & ||\hat{\mathbf{G}}_{\mathrm{L}}(\hat{\bar{\mathbf{y}}}_{\mathrm{L}} - \hat{\boldsymbol{\alpha}}_{\mathrm{L},1})||_2 \leq c_1 \\
& ||\hat{\mathbf{G}}_{\mathrm{R}}(\hat{\bar{\mathbf{y}}}_{\mathrm{R}} - \hat{\boldsymbol{\alpha}}_{\mathrm{R},1})||_2 \leq c_2 \\
& \alpha_{\mathrm{IACC,L},n_l}(\tau^*_{\mathrm{IACC}}) > \alpha_{\mathrm{IACC,L},n_l}(\bar{n_b}) + c_3(1 - \mu_{\mathrm{IACC}}(\bar{n_b})), \quad \text{for} \quad \bar{n_b} \neq \tau^*_{\mathrm{IACC}} \\
& \alpha_{\mathrm{IACC,R},n_l}(\tau^*_{\mathrm{IACC}}) > \alpha_{\mathrm{IACC,R},n_l}(\bar{n_b}) + c_4(1 - \mu_{\mathrm{IACC}}(\bar{n_b})), \quad \text{for} \quad \bar{n_b} \neq \tau^*_{\mathrm{IACC}} \\
& \boldsymbol{\alpha}_{\mathrm{IACC,L},n_l} = \mathbf{F}\bar{\mathbf{Y}}_{\mathrm{R}}\mathbf{F}\boldsymbol{\alpha}_{\mathrm{L},n_l} \\
& \boldsymbol{\alpha}_{\mathrm{IACC,R},n_l} = \bar{\mathbf{Y}}_{\mathrm{L}}\mathbf{F}\boldsymbol{\alpha}_{\mathrm{R},n_l} \\
& \boldsymbol{\alpha}_{\mathrm{L},n_l} = \sum_{n_s} \mathbf{W}^{-1}\hat{\mathbf{C}}_{\mathrm{L},n_s,n_l}\hat{\mathbf{s}}_{n_s} \\
& \boldsymbol{\alpha}_{\mathrm{R},n_l} = \sum_{n_s} \mathbf{W}^{-1}\hat{\mathbf{C}}_{\mathrm{R},n_s,n_l}\hat{\mathbf{s}}_{n_s} \\
& \text{for} \quad n_l = 1,...,N_l \quad \& \quad \bar{n_b} \in [-1,1] \text{ ms}
\end{aligned}
\tag{C.18}
$$

In the optimization problem, $\bar{n}_b = 1,...,4L_b - 1$ denotes the IACC sample indexes, $\mathbf{W}^{-1} \in \mathbb{C}^{2L_b \times 2L_b}$ denotes the Inverse Fast Fourier Transform (IFFT) matrix, $\mathbf{s}_{n_s} \in \mathbb{R}^{2L_b \times 1}$ for $n_s = 1,...,N_s$ are the loudspeaker signal optimization variables, $\mathbf{F} \in \mathbb{R}^{2L_b \times 2L_b}$ is the exchange (flip) matrix, $\bar{\mathbf{Y}}_{\mathrm{L}} \in \mathbb{R}^{4L_b-1 \times 2L_b}$ and $\bar{\mathbf{Y}}_{\mathrm{R}} \in \mathbb{R}^{4L_b-1 \times 2L_b}$ denote the toeplitz matrix of $\bar{\mathbf{y}}_{\mathrm{L}}$ and $\bar{\mathbf{y}}_{\mathrm{R}}$ respectively, $\boldsymbol{\mu}_{\mathrm{IACC}} \in \mathbb{R}^{4L_b-1 \times 1}$ denotes the upper limit of the IACC, $\bar{\mathbf{G}}_{\mathrm{L}} \in \mathbb{R}^{2L_b \times 2L_b}$ and $\bar{\mathbf{G}}_{\mathrm{R}} \in \mathbb{R}^{2L_b \times 2L_b}$ are the masking curve matrices with the masking curve on the diagonal for left and right ear respectively, $||(\cdot)||$ denotes the L_2-norm, $\boldsymbol{\alpha}_{\mathrm{L},n_l} \in \mathbb{R}^{2L_b \times 1}$, $\boldsymbol{\alpha}_{\mathrm{R},n_l} \in \mathbb{R}^{2L_b \times 1}$, $\boldsymbol{\alpha}_{\mathrm{IACC,L},n_l} \in \mathbb{R}^{4L_b-1 \times 1}$ and $\boldsymbol{\alpha}_{\mathrm{IACC,R},n_l} \in \mathbb{R}^{4L_b-1 \times 1}$ are optimization variables and $c_1,...,c_4$ are optimization constants.

Implementing this problem in a CVX problem in MATLAB requires a series of computations that are described here. The main challenges that arise in the implementation is the addition of proper zero padding. The zero padding is required to make sure the sizes of the variables in the calculations match and also to make sure that a convolution performed using the FFT does not result in a circular convolution. In the following, the size of some variables is redefined and new variables are introduced that correspond to the implementation used to compute the results. First a set of general signal definitions is given after which the masking constraint and the IACC constraint are described.

The optimization variables $\mathbf{s_1}, \mathbf{s_2}, ..., \mathbf{s_{N_s}}$ are implemented by placing the column vectors underneath each other to form $\mathbf{s} \in \mathbb{R}^{2N_sL_b \times 1}$. This means that all the constraints are rewritten to be compatible with this shape. To make sure the speaker signals do not take a too large amplitude and the results remain stable over multiple time blocks, the maximum absolute value of the optimized speaker signals is limited by means of an infinity norm. This simple constraint forces the algorithm to produce solutions that are closer to the desired signal which is desirable when considering future time blocks.

C.4.1. Masking constraint

The new definition of the required response as described in Equation (C.17) leads to a redefinition of the variables $\hat{\bar{\mathbf{y}}}_{\mathrm{L}}$ and $\hat{\bar{\mathbf{y}}}_{\mathrm{R}}$ in the masking constraints in Equation (C.18). On top of this, some additional zero padding is required before converting from time to frequency domain to have a compatible size with $\hat{\boldsymbol{\alpha}}_{\mathrm{L},1}$ and $\hat{\boldsymbol{\alpha}}_{\mathrm{R},1}$.

The renewed and combined definition of the zero padded $\hat{\bar{\mathbf{y}}}_{\mathrm{L}}$ and $\hat{\bar{\mathbf{y}}}_{\mathrm{R}}$ is given in Equation (C.19), where $\mathbf{0} \in \mathbb{R}^{3L_b \times 1}$ is a zeros matrix and $\mathbf{W} \in \mathbb{C}^{6L_b \times 6L_b}$ is the FFT matrix.

$$\hat{\bar{\mathbf{y}}} \in \mathbb{R}^{12L_b \times 1} = \begin{bmatrix} \mathbf{W} \begin{bmatrix} \bar{\mathbf{y}}_{\mathrm{L}} \\ \mathbf{0} \end{bmatrix} \\ \mathbf{W} \begin{bmatrix} \bar{\mathbf{y}}_{\mathrm{R}} \\ \mathbf{0} \end{bmatrix} \end{bmatrix} \tag{C.19}$$

Computing a compatible combination of $\hat{\boldsymbol{\alpha}}_{\mathrm{L},1}$ and $\hat{\boldsymbol{\alpha}}_{\mathrm{R},1}$ starting from $\mathbf{s}$ is a more challenging task. First the speaker signals must be zero-padded, which is done by means of the zero-padding matrix defined in Equation (C.20), where $\otimes$ denotes the Kronecker product and $\mathbf{I}$ denotes the identity matrix.

$$\begin{aligned} \mathbf{Z}_{n_s} \in \mathbb{R}^{6L_b \times 2L_b} &= \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix}, & \mathbf{I} \in \mathbb{R}^{2L_b \times 2L_b}, \mathbf{0} \in \mathbb{R}^{4L_b \times 2L_b} \\ \bar{\mathbf{Z}} \in \mathbb{R}^{6N_s L_b \times 2N_s L_b} &= \mathbf{I} \otimes \mathbf{Z}_{n_s}, & \mathbf{I} \in \mathbb{R}^{N_s \times N_s} \end{aligned} \tag{C.20}$$

Similarly, the matrix implementing the FFT on all the speaker signals is defined. It is given in Equation (C.21).

$$\bar{\mathbf{W}} \in \mathbb{C}^{6N_s L_b \times 6N_s L_b} = \mathbf{I} \otimes \mathbf{W}, \qquad \mathbf{I} \in \mathbb{R}^{N_s \times N_s}, \mathbf{W} \in \mathbb{C}^{6L_b \times 6L_b} \tag{C.21}$$

To complete the response, the channel corresponding to each ear-loudspeaker pair must be multiplied with the speaker signal. This is done by means of the matrix given in Equation (C.22), where $\mathrm{diag}(\cdot)$ converts the vector into a matrix with the vector on the diagonal.

$$\begin{aligned} \hat{\mathbf{C}}_{\mathrm{L},n_s} \in \mathbb{C}^{6L_b \times 6L_b} &= \mathrm{diag}\left(\mathbf{W} \begin{bmatrix} \mathbf{c}_{\mathrm{L},n_s,1} \\ \mathbf{0} \end{bmatrix}\right), & \mathbf{0} \in \mathbb{R}^{4L_b \times 1}, \mathbf{W} \in \mathbb{C}^{6L_b \times 6L_b} \\ \hat{\mathbf{C}}_{\mathrm{R},n_s} \in \mathbb{C}^{6L_b \times 6L_b} &= \mathrm{diag}\left(\mathbf{W} \begin{bmatrix} \mathbf{c}_{\mathrm{R},n_s,1} \\ \mathbf{0} \end{bmatrix}\right), & \mathbf{0} \in \mathbb{R}^{4L_b \times 1}, \mathbf{W} \in \mathbb{C}^{6L_b \times 6L_b} \\ \hat{\mathbf{C}} \in \mathbb{C}^{12L_b \times 6N_s L_b} &= \begin{bmatrix} \hat{\mathbf{C}}_{\mathrm{L},1} & \dots & \hat{\mathbf{C}}_{\mathrm{L},N_s} \\ \hat{\mathbf{C}}_{\mathrm{R},1} & \dots & \hat{\mathbf{C}}_{\mathrm{R},N_s} \end{bmatrix} \end{aligned} \tag{C.22}$$

To complete the masking curve constraints, the final part is the masking curves themselves. The masking curve matrices are calculated as $\hat{\mathbf{G}}_{\mathrm{L}} \in \mathbb{R}^{6L_b \times 6L_b} = \mathrm{diag}(\hat{\mathbf{g}}_{\mathrm{L}})$ and $\hat{\mathbf{G}}_{\mathrm{R}} \in \mathbb{R}^{6L_b \times 6L_b} = \mathrm{diag}(\hat{\mathbf{g}}_{\mathrm{R}})$. Calculating $\hat{\mathbf{g}}_{\mathrm{L}}$ and $\hat{\mathbf{g}}_{\mathrm{R}}$ is performed according to Section 2.2 and does not require any further implementation details.

To efficiently implement the masking curve, the masking curve matrices are combined to form the matrix as defined in Equation (C.23).

$$\hat{\mathbf{G}} \in \mathbb{R}^{12L_b \times 12L_b} = \begin{bmatrix} \hat{\mathbf{G}}_{\mathrm{L}} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{G}}_{\mathrm{R}} \end{bmatrix}, \quad \mathbf{0} \in \mathbb{R}^{6L_b \times 6L_b} \tag{C.23}$$

Given the above defined matrices, the masking curve constraints can be rewritten to Equation C.24, which is the constraint implemented in the MATLAB implementation.

$$||\hat{\mathbf{G}}(\hat{\bar{\mathbf{y}}} - \hat{\mathbf{C}}\bar{\mathbf{W}}\bar{\mathbf{Z}}\mathbf{s})||_2 \leq c_1 \tag{C.24}$$

C.4.2. IACC constraint

Implementing the IACC constraints is more involved than the masking constraint and entails some extra steps. First the audio input signal $\mathbf{s}$ needs to be converted to the perceived response. The difference with the definition of the perceived response in the masking curve constraint is that we remain in the time domain.

First, the loudspeaker signals must be zero padded with length L_b, this is done using the zero padding matrix as given in Equation (C.25).

$$\begin{aligned}
\mathbf{Z}_{n_s} \in \mathbb{R}^{3L_b \times 2L_b} &= \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix}, & \mathbf{I} \in \mathbb{R}^{2L_b \times 2L_b},\, \mathbf{0} \in \mathbb{R}^{L_b \times 2L_b} \\
\mathbf{Z} \in \mathbb{R}^{3N_s L_b \times 2N_s L_b} &= \mathbf{I} \otimes \mathbf{Z}_{n_s}, & \mathbf{I} \in \mathbb{R}^{N_s \times N_s}
\end{aligned} \tag{C.25}$$

Convolving all the loudspeaker signals with the ear to loudspeaker channels is done by the matrix described in Equation (C.26), where the function toep$(\cdot)$ is the toeplitz matrix defined as given in Equation (C.27) (Note the slight different definition than the one given in Equation (5.5), a row of zeros is added). As shown, we do this for all optimization points n_l.

$$\begin{aligned}
\mathbf{C}_{\mathrm{L},n_s,n_l} \in \mathbb{R}^{6L_b \times 3L_b} &= \text{toep}\left(\begin{bmatrix} \mathbf{c}_{\mathrm{L},n_s,n_l} \\ \mathbf{0} \end{bmatrix}\right), & \mathbf{0} \in \mathbb{R}^{L_b \times 1} \\
\mathbf{C}_{\mathrm{R},n_s,n_l} \in \mathbb{R}^{6L_b \times 3L_b} &= \text{toep}\left(\begin{bmatrix} \mathbf{c}_{\mathrm{R},n_s,n_l} \\ \mathbf{0} \end{bmatrix}\right), & \mathbf{0} \in \mathbb{R}^{L_b \times 1} \\
\mathbf{C}_{n_l} \in \mathbb{R}^{12L_b \times 3N_s L_b} &= \begin{bmatrix} \mathbf{C}_{\mathrm{L},1,n_l} & \dots & \mathbf{C}_{\mathrm{L},N_s,n_l} \\ \mathbf{C}_{\mathrm{R},1,n_l} & \dots & \mathbf{C}_{\mathrm{R},N_s,n_l} \end{bmatrix}
\end{aligned} \tag{C.26}$$

$$\mathbf{A} = \text{toep}(\mathbf{a}) = \begin{bmatrix} a(1) & 0 & 0 & \cdots & 0 \\ a(2) & a(1) & 0 & \cdots & 0 \\ a(3) & a(2) & a(1) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a(N_b-1) & a(N_b-2) & a(N_b-3) & \cdots & 0 \\ a(N_b) & a(N_b-1) & a(N_b-2) & \cdots & a(1) \\ 0 & a(N_b) & a(N_b-1) & \cdots & a(2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a(N_b) \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \quad \mathbf{a} \in \mathbb{R}^{N_b \times 1},\, \mathbf{A} \in \mathbb{R}^{2N_b \times N_b} \tag{C.27}$$

To calculate the IACC, the found responses at the ears must be flipped which is done by means of the exchange matrix as presented in Equation (C.28).

$$\begin{aligned}
\bar{\mathbf{E}} \in \mathbb{R}^{6L_b \times 6L_b} &= \begin{bmatrix} 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 1 & 0 \\ \vdots & \vdots & \iddots & \vdots & \vdots \\ 0 & 1 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \end{bmatrix}, \\
\mathbf{E} \in \mathbb{R}^{12L_b \times 12L_b} &= \mathbf{I} \otimes \bar{\mathbf{E}}, & \mathbf{I} \in \mathbb{R}^{2 \times 2}
\end{aligned} \tag{C.28}$$

To convolve the responses at left and right ear with the required response to obtain the so defined left and right IACC's is done by means of the required response convolution matrix as presented in Equation (C.29)

$$\begin{aligned}
\bar{\mathbf{Y}}_{\mathrm{L}} \in \mathbb{R}^{12L_b \times 6L_b} &= \text{toep}\left(\bar{\mathbf{y}_{\mathrm{L}}}\right), \\
\bar{\mathbf{Y}}_{\mathrm{R}} \in \mathbb{R}^{12L_b \times 6L_b} &= \text{toep}\left(\bar{\mathbf{y}_{\mathrm{R}}}\right), \\
\bar{\mathbf{Y}} \in \mathbb{R}^{24L_b \times 12L_b} &= \begin{bmatrix} \bar{\mathbf{Y}}_{\mathrm{L}} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{Y}}_{\mathrm{R}} \end{bmatrix}, & \mathbf{0} \in \mathbb{R}^{12L_b \times 6L_b}
\end{aligned} \tag{C.29}$$

For clarity, we flip one of the IACC responses to make sure that both IACC responses have the same axes system. This is done using the flip matrix as presented in Equation (C.30). After this operation, both IACC responses are given such that a time delay index to the left of the zero delay index corresponds to a source perceived from the left.

$$\mathbf{F} = \begin{bmatrix} \bar{\mathbf{E}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \quad \mathbf{0}, \mathbf{I}, \bar{\mathbf{E}} \in \mathbb{R}^{12L_b \times 12L_b} \tag{C.30}$$

With all the above mentioned matrices, we are able to calculate the left and right IACC. To convert this IACC into a proper constraint, first the desired time index of the IACC peak must be determined. This is done by finding the time indexes of the peak value of the IACC computed with solely the required response. For convenience we define a left and right peak time index $\tau^*_{\text{L,IACC}}$ and $\tau^*_{\text{R,IACC}}$ but in practise they are the same.

The range of interest in the IACC is given by $[-1, 1]$ ms. Translating this to discrete samples results in a small range of $2L_I$ samples. For $f_s = 16000$, $L_I = 16$ (we force $2L_I$ to be even for convenience although this is not correct).

In the constraints we wish the IACC at $\tau^*_{\text{L,IACC}}$ and $\tau^*_{\text{R,IACC}}$ to be larger than the other samples in the response in the range of interest($\boldsymbol{\mu}_{\text{IACC}}$ is added later in the implementation). To do this, we use the difference matrix as presented in Equation (C.31), where $\delta[\cdot]$ is the discrete impulse response function, $n_I = -L_I, ..., L_I$ and $\mathbf{1}$ is a ones vector. Do note the drastic decrease in vector size after the multiplication with this matrix.

$$\begin{aligned}
\mathbf{D}_{\text{L}} \in \mathbb{C}^{2L_I \times 2L_I} &= -\mathbf{I} + \boldsymbol{\delta}[n_l - \tau^*_{\text{L,IACC}}]^{\text{T}} \otimes \mathbf{1}, & \mathbf{I} \in \mathbb{R}^{2L_i \times 2L_i},\ \mathbf{1} \in \mathbb{R}^{2L_i \times 1} \\
\mathbf{D}_{\text{R}} \in \mathbb{C}^{2L_I \times 2L_I} &= -\mathbf{I} + \boldsymbol{\delta}[n_l - \tau^*_{\text{R,IACC}}]^{\text{T}} \otimes \mathbf{1}, & \mathbf{I} \in \mathbb{R}^{2L_i \times 2L_i},\ \mathbf{1} \in \mathbb{R}^{2L_i \times 1} \\
\mathbf{D} \in \mathbb{R}^{4L_i \times 24L_b} &= \begin{bmatrix} \mathbf{0} & \mathbf{D}_{\text{L}} & \mathbf{0} & \bar{\mathbf{0}} & \\ & \bar{\mathbf{0}} & & \mathbf{0} & \mathbf{D}_{\text{R}} & \mathbf{0} \end{bmatrix}, & \mathbf{0} \in \mathbb{R}^{2L_i \times 6L_b - L_i},\ \bar{\mathbf{0}} \in \mathbb{R}^{2L_i \times 12L_b}
\end{aligned} \tag{C.31}$$

In order for the IACC constraint to hold, the resulting vector must only contain non-negative values. Including the IACC upper limit $\boldsymbol{\mu}_{\text{IACC}} \in \mathbb{R}^{2L_i \times 1}$, as defined and determined by means of Equation (5.10), is done by stating that the resulting vector should be larger than the one given in Equation (C.32).

$$\boldsymbol{\mu} \in \mathbb{R}^{4L_i \times 1} = \begin{bmatrix} \mathbf{1} - \boldsymbol{\mu}_{\text{IACC}} \\ \mathbf{1} - \boldsymbol{\mu}_{\text{IACC}} \end{bmatrix}, \quad \mathbf{1} \in \mathbb{R}^{2L_i \times 1} \tag{C.32}$$

Combining all, the IACC constraint can be implemented for each optimization point n_l by means of Equation (C.33) and expanded to include all the optimization points n_l as presented in Equation (C.34).

$$\mathbf{DF\bar{Y}EC}_{n_l}\mathbf{Zs} \geq \boldsymbol{\mu} \tag{C.33}$$

$$\begin{bmatrix} \mathbf{DF\bar{Y}EC}_1\mathbf{Zs} \\ \vdots \\ \mathbf{DF\bar{Y}EC}_{N_l}\mathbf{Zs} \end{bmatrix} \geq \begin{bmatrix} \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{\mu} \end{bmatrix} \tag{C.34}$$

To finalize, the optimzation problem used in the implementation defined in terms of the constraints described above is given in Equation (C.35).

$$\begin{aligned}
\arg\min_{\mathbf{s}} \quad & 0 \\
\text{s.t.} \quad & ||\hat{\mathbf{G}}(\hat{\bar{\mathbf{y}}} - \hat{\mathbf{C}}\bar{\mathbf{W}}\bar{\mathbf{Z}}\mathbf{s})||_2 \leq c_1 \\
& ||\mathbf{s}||_\infty \leq c_2 \\
& \begin{bmatrix} \mathbf{DF\bar{Y}EC}_1\mathbf{Zs} \\ \vdots \\ \mathbf{DF\bar{Y}EC}_{N_l}\mathbf{Zs} \end{bmatrix} \geq \begin{bmatrix} \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{\mu} \end{bmatrix}
\end{aligned} \tag{C.35}$$