

Room correction for smart speakers

Simon Mårtensson

Master of Science Thesis in Electrical Engineering

Room correction for smart speakers:

Simon Mårtensson

LiTH-ISY-EX-ET--19/5209--SE

Supervisor: **Kamiar Radnosrati**
ISY, Linköpings universitet

Viktor Gunnarsson
Dirac Research AB

Examiner: **Fredrik Gustafsson**
ISY, Linköpings universitet

*Division of Automatic Control
Department of Electrical Engineering
Linköping University
SE-581 83 Linköping, Sweden*

Copyright © 2019 Simon Mårtensson

Abstract

Portable smart speakers with wireless connections have in recent years become more popular. These speakers are often moved to new locations and placed in different positions in different rooms, which affects the sound a listener is hearing from the speaker. These speakers usually have microphones on them, typically used for voice recording. This thesis aims to provide a way to compensate for the speaker position's effect on the sound (so called room correction) using the microphones on the speaker and the speaker itself.

Firstly, the room frequency response is estimated for several different speaker positions in a room. The room frequency response is the frequency response between the speaker and the listener. From these estimates, the relationship between the speaker's position and the room frequency response is modeled. Secondly, an algorithm that estimates the speaker's position is developed. The algorithm estimates the position by detecting reflections from nearby walls using the microphones on the speaker. The acquired position estimates are used as input for the room frequency response model, which makes it possible to automatically apply room correction when placing the speaker in new positions.

The room correction is shown to correct the room frequency response so that the bass has the same power as the mid- and high frequency sounds from the speaker, which is according to the research aim. Also, the room correction is shown to make the room frequency response vary less with respect to the speaker's position.

Acknowledgments

I would like to thank all the people that have helped me with this master's thesis.

I am very grateful for the help, expertise and interest I have gotten from the people at Dirac, especially from my supervisor Viktor Gunnarsson, who's shown great interest and curiosity about my results. Your knowledge about acoustics, audio systems, signal processing and scientific methods has been very valuable.

To my supervisor at Linköping University, Kamiar Radnosrati, a sincere thank you for all the time and effort you have spent helping me. Your knowledge and expertise have been a great resource and pushed the thesis further. Your positive attitude and constructive feedback has been invaluable and one of the best motivators when working with this master's thesis.

Many thanks to my examiner Fredrik Gustafsson, whose experience and feedback have been of great value. Your guidance have pushed the thesis into a direction where the most interesting results have been found.

Lastly, many thanks to the people I have been sharing office with and spent my breaks together with. You have made this spring very enjoyable and given me energy to pursue my work.

Contents

Acknowledgments	v
Notation	ix
1 Introduction	1
1.1 Motivation	1
1.2 Purpose	1
1.3 Research questions	1
1.4 Delimitations	2
1.5 Report structure	3
2 Background and motivation	5
2.1 Related work	5
2.2 Psychoacoustics	5
2.3 Room acoustics	6
2.3.1 Reflections and sound paths	6
2.3.2 Schroeder frequency	8
2.3.3 Output changes due to speaker position	8
2.4 Speaker dynamics	8
2.5 Microphone dynamics	9
2.6 System identification	9
2.6.1 Hammerstein models	9
2.6.2 Identification with log-sine-sweeps	10
2.7 Spectral analysis	13
2.7.1 Normalization	13
2.7.2 Octave smoothing	13
2.8 Regression methods	13
2.8.1 Linear regression	13
2.8.2 L1 regularized linear regression - Lasso	14
2.9 Shelving filter	14
3 Method	17
3.1 Signal and system model	17

3.1.1	Model of whole system	17
3.1.2	Model of speaker	18
3.1.3	Model of room acoustics	19
3.1.4	Model of microphones	19
3.2	Measurements	20
3.3	Finding correct filter parameters	20
3.4	Localization from RIR	21
3.4.1	Correcting impulse responses	22
3.4.2	Lasso for finding reflections	23
4	Results and Discussion	29
4.1	Setup	29
4.1.1	Hardware and software	29
4.1.2	Room description	30
4.2	Estimating filter gain from speaker position	31
4.2.1	Room frequency response measurements	32
4.2.2	Position's impact on bass	34
4.2.3	Model for correction gain from speaker position	35
4.3	Estimation of speaker position	37
4.3.1	Measurements for speaker position estimation	37
4.3.2	Finding reflections in impulse responses	39
4.4	Filter design and implementation	40
4.5	Tests of room correction	40
4.5.1	Tests on positions 1-16	41
4.5.2	Tests on new measurements	44
4.6	Problems and limitations	45
4.6.1	Position to room frequency response mapping	45
4.6.2	Speaker position estimation	45
4.6.3	Correction filter	47
5	Conclusions	51
5.1	Further work	52
A	Additional work	55
A.1	Pairwise DOA	55
A.2	Implementation and results	57
B	Room frequency responses	59
C	Frequency responses for microphones	65
	Bibliography	67

Notation

ABBREVIATIONS

Abbreviation	Meaning
RIR	Room Impulse Response
SNR	Signal-to-Noise Ratio
PSD	Power Spectral Density
RFRM	Room Frequency Response Magnitude
DP	Direct Path
FIR	Finite Impulse Response
DOA	Direction Of Arrival

1

Introduction

1.1 Motivation

Portable smart speakers are often put in different places in a room and moved to new placements by the users. The placement of the speaker affects the sound perceived by the listener, since the acoustical characteristics change depending on the speaker's placement. Especially, the lower hearable frequencies (i.e. the bass) are dependant on the speaker position in the room [12]. E.g., the low frequency components from the speaker increase in power if the speaker is placed close to a corner [12] [16]. This variation in the room frequency response magnitude (RFRM, which is the magnitude of the frequency response for the room and the speaker) makes it difficult to predict how the sound from the speaker will be perceived once in use.

1.2 Purpose

The purpose of this thesis is to investigate methods to make the speaker able to automatically apply correction filters depending on its position. The RFRM between the input of the speaker and what the listener is hearing (hereby called only RFRM) should be corrected to being the same, no matter where the speaker is positioned. This thesis will focus on doing this in a conference room at Linköping University, in the area Visionen, for which the plan can be seen in Figure 1.1.

1.3 Research questions

Three research questions have been formulated to properly define the approach and the aim of the thesis. The research questions which this thesis aims to answer

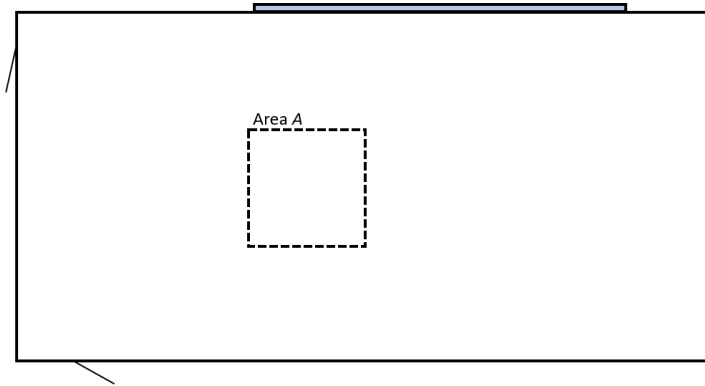


Figure 1.1: Plan of the measurement room. The area A is where the supposed listener is placed.

are

- Can we make a model of how the positioning of a speaker in the room in Figure 1.1 affects the RFRM heard by the listener, who is standing on a position within the area A (Figure 1.1)?
- Can we determine what the RFRM within area A is, by doing measurements with a microphone or a microphone array which is placed on the speaker?
- Can we use simple digital filters so that the RFRM to the listener is identical within area A, no matter which position the speaker is at?

1.4 Delimitations

Some delimitations have been set for this thesis, since otherwise the project would be too complex for a master's thesis.

The properties of the acoustics in a room can be hard to predict for high-frequency sound. Hence, only the room acoustics of the lower hearable frequencies (the bass) will be considered.

For the correction filter, a simple filter design was desired to limit the number of filter parameter estimates needed. A suitable filter for this is a Shelving filter and therefore the thesis is limited to only using Shelving filters [18].

Some speaker positions make it troublesome to identify which reflections come from walls and which come from the ceiling or the roof. Therefore, a limitation is that the distance between the speaker and the ceiling is known beforehand and explicitly put in the developed algorithms. Also, for the same reason, the two closest walls are closer to the speaker than the ceiling and the speaker is not closer than 0.4 meters to the closest wall.

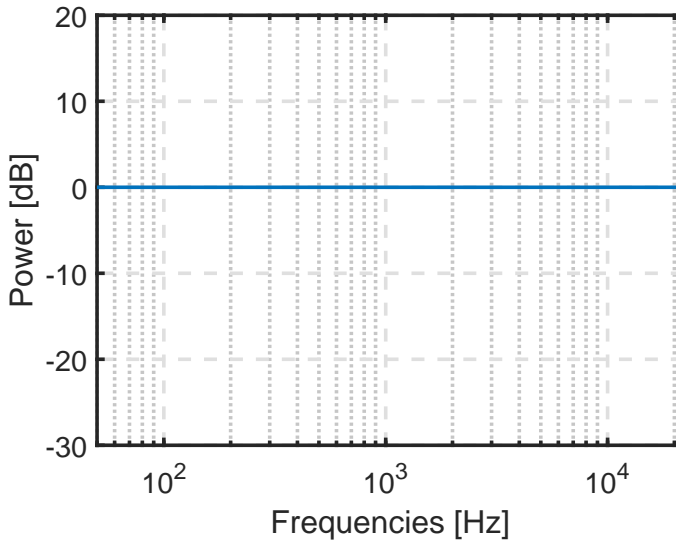


Figure 1.2: The RFRM which is the target in this thesis. Defined between frequencies 50 to 22050 Hz and is compared to other RFRMs normalized to 0 dB.

1.5 Report structure

The thesis consists of five chapters. Chapter 2 discusses related work, presents relevant theory, which is about physical properties of sound, suitable models for acoustical problems, system identification, regression methods and Shelving filters. Chapter 3 presents how the discussed theory is applied to answer the research questions. Chapter 4 presents how measurements have been made and the results with a discussion around it. In Chapter 5, some conclusions about the work are drawn and future work that could be done to improve the results is discussed. Lastly, in the appendices, additional work that could be of use for future work is presented. Also, some plots of room frequency responses and microphone behavior that did not fit in the main parts of the report are put in the appendices.

2

Background and motivation

In this chapter, background and motivation for the thesis are presented. Firstly, some related work is discussed and then some main theory is presented.

2.1 Related work

The authors in [12] provide a room correction method for subwoofers, which includes a movable microphone that does several measurements in different positions. From these measurements, it is then possible to estimate the sound propagation in the room and according to this correct the outputted sound.

For room geometry estimation, the authors of [2] provide a method for which the room geometry can be inferred by a co-located speaker and microphone array. The method builds upon identifying reflections, their direction and the distance to the walls from which the reflection came from. However, to use this method, reference measurements done in an anechoic room are needed. The Lasso linear regression method used in this thesis is largely inspired by this paper.

Another method for room geometry estimation is provided by the authors of [15], who present a method for localizing walls by looking at reflections in room impulse responses (RIR) for several distributed microphones. They use a time-of-arrival (TOA) approach for the wall localization, but do not present a method for automatically identifying reflections in the RIR.

2.2 Psychoacoustics

Psychoacoustics is a term used to describe the study of the physical structure of the ear, the sound pathways, the human perception of sound and their interrelationships. One main area in psychoacoustics is the relationship between

audibility, the frequency and the pressure level of a sound [3]. For applications where the perceived sound is important, psychoacoustic models could be used to evaluate performance of the application.

One notable characteristic of psychoacoustics is how loudness (perceived strength of a sound) differs from the actual sound pressure levels. When studying loudness, Benjamin and Fielder show that a change of ± 1 dB is just audible for low frequencies [5].

2.3 Room acoustics

Room acoustics is what defines the system between outputted sound from the speaker to what will be received by the microphone or a human listener. In this section, some theory about how a room affects the sound in it is presented. Central properties are reflections of the walls, how the sound is spreading in the room and how the sound source position changes the acoustical properties in the room.

2.3.1 Reflections and sound paths

In a room, the sound can take many different paths between the speaker and the microphone, due to the reflections of the walls. Different paths result in different time delays and attenuations. The total attenuation for path i depends on the absorption coefficients of the walls and the total traveled distance of the sound wave, due to the propagation resistance from the air in the room.

Figure 2.1 illustrates three different paths - the direct path between the speaker and the microphone, a first order reflection affected by the absorption coefficient $\rho^{(1)}$ (path 1) and a second order reflection affected by the absorption coefficients $\rho^{(2)}$ and $\rho^{(3)}$. In Figure 2.2, similar examples can be seen, but in this case the speaker and the microphone are co-located. In this case, the distance of the direct path is very small and the first order reflections will be perpendicular to the walls [2].

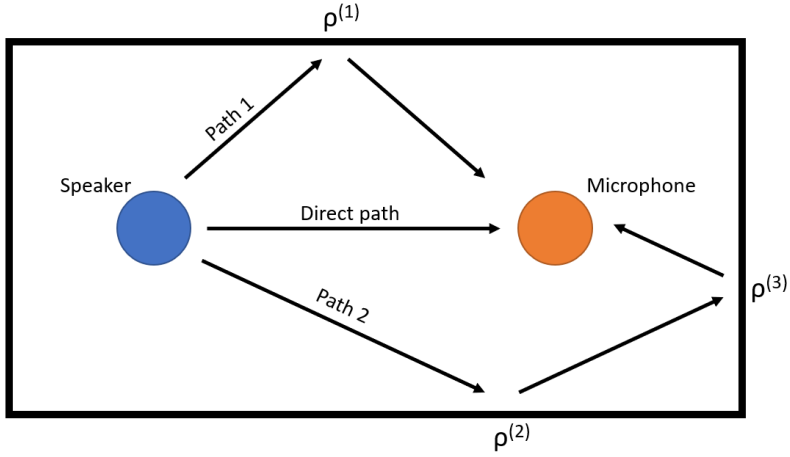


Figure 2.1: Examples of paths the sound can propagate in the room, from the source speaker to the receiver microphone. Path 1 is a first order reflection and path 2 is a second order reflection. $\rho^{(j)}$, $j = 1, 2, 3$ are different attenuation coefficients for the walls.

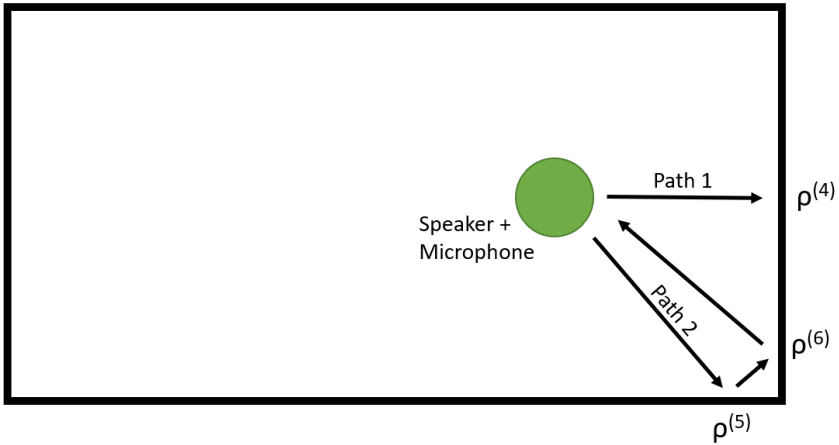


Figure 2.2: Examples of paths the sound can propagate via the room, with the speaker and microphone co-located. Path 1 is a first order reflection and path 2 is a second order reflection. The direct path is not visible, since the speaker and microphone are co-located, and the first order reflections are perpendicular to the walls. $\rho^{(j)}$, $j = 4, 5, 6$ are different attenuation coefficients for the walls.

For a path i , the total attenuation is represented by an attenuation constant, denoted $\alpha^{(i)}$, which includes the absorption from walls and the energy loss due

to air resistance. Since wall reflections do not change the phase of the sound and do not increase the power, $\alpha^{(i)}$ should be positive and below one, i.e. $0 < \alpha^{(i)} < 1$. For some situations, $\alpha^{(i)}$ can vary depending on the frequency of the sound [2].

2.3.2 Schroeder frequency

The Schroeder frequency $f_{\text{Schroeder}}$ is a frequency which approximately separates low frequencies from high- and mid frequencies in a room. The low frequency in this case is defined as frequencies for which standing waves occur and room reverberation dominates. The Schroeder frequency $f_{\text{Schroeder}}$ is defined by

$$f_{\text{Schroeder}} = 2000 \cdot \sqrt{\frac{T_{60}}{V}}, \quad (2.1)$$

where T_{60} is the reverberation time of a room (for when the room impulse response's power has decreased by 60 dB) and V is the volume of the room. [1]

2.3.3 Output changes due to speaker position

In a room, the closer the speaker is placed to a corner, the more power there will be in the output for the lower frequencies (compared to higher frequencies) [12] [16]. As an example, for a specific room and corner, when the authors of [16] moved the speaker away from the corner, they noticed a drop of 20.5 dB in power output for a single frequency with a certain wave length λ .

A common defined transition between low and mid frequencies for a room is the Schroeder frequency.

2.4 Speaker dynamics

Speakers' can sometimes distort the sound in a non-linear way. Hence, an appropriate model for the speaker is a Volterra model of N :th order. The speaker will be modeled as

$$\mathcal{H}_{\text{speaker}}\{x(t)\} = x(t) * k_1(t) + x^2(t) * k_2(t) + \dots + x^N(t) * k_N(t), \quad (2.2)$$

where $x(t)$ is the input, $k_i(t)$, $i = 1, 2, \dots, N$, is a kernel and the operator $*$ denotes a convolution [4] [13] [9].

Note that in Equation 2.2 the first term $x(t) * k_1(t)$ is linear. If $k_2(t), k_3(t), \dots, k_N(t)$ are all close to being all zero-valued, the speaker can be approximated as a linear system. When estimating the impulse response of a total system including a speaker, a room and a microphone, it is often wanted to not let the speaker's non-linear terms $k_i(t) * x^i(t)$, $i = 2, 3, \dots, N$ affect the estimation of the linear part of the total system. When using the method Farina Sweeps, introduced in [4] and expanded by [9], those problems are manageable.

For the linear part with the kernel $k_1(t)$, important characteristics are the bandwidth, the frequency response's magnitude within the bandwidth and the directionality of the speaker. The characteristics differ depending on which speaker

type is used. Often, speakers for all-around entertainment purposes have a bandwidth within or slightly above the hearing interval, which is 20 Hz to 20 kHz [3]. The bandwidth, together with the speakers frequency response's magnitude, colors the speakers output. The output's directionality of the speaker is mostly to the front, where the speaker elements are pointing at.

2.5 Microphone dynamics

Microphones can generally be modeled as linear systems, since they usually are at least approximately linear systems [9]. Important properties for measurement microphones are that they have a flat frequency response for the magnitude, and preferably a linear phase response. Microphones generally have a polar pattern, which defines how they pick up sound from different directions. Some microphones are (nearly) omni-directional, meaning they pick up sound from every direction with the same strength. As for speakers, the bandwidth of the microphone is also an important property. The bandwidth defines for which frequencies the microphone is suitable for.

In many cases, the microphone dynamics are not of interest in measurements, but only a mean to capture sound with. If a microphone is linear (i.e. can be described with a linear system), it might be possible to find the inverse system for the microphone. The inverse system can be used to inverse filter the recorded output and exclude the effect of the microphone dynamics. Microphones are usually the last part of cascade system such as

$$\begin{aligned} y(t) &= \mathcal{H}_{\text{Microphone}} \{ \mathcal{H}_{\text{Wanted}} \{ x(t) \} \} = \\ &= \mathcal{H}_{\text{Microphone}} \{ y_{\text{Wanted}}(t) \}, \end{aligned} \quad (2.3)$$

where $y(t)$ is the total output, $x(t)$ is the input, $\mathcal{H}_{\text{Microphone}}$ is the system of the microphone, $\mathcal{H}_{\text{Wanted}}$ and $y_{\text{Wanted}}(t)$ are the wanted system and wanted output, respectively. With the inverse linear system, it is possible to find $y_{\text{Wanted}}(t)$ (if the signal is within the systems bandwidth) by

$$\mathcal{H}_{\text{Microphone}}^{-1} \{ \mathcal{H}_{\text{Microphone}} \{ y_{\text{Wanted}}(t) \} \} = y_{\text{Wanted}}(t), \quad (2.4)$$

which is due to the linearity of the microphone.

2.6 System identification

System identification can be done in many ways and is often specific to which type of system that is to be estimated. In this section, an identification method for a type of Hammerstein models is presented.

2.6.1 Hammerstein models

Hammerstein models belong to a class of models which are block-based and consist of a non-linear memory-less block followed by a linear block, where each

block represents a system [11] [17]. In Figure 2.3, a Hammerstein model is shown, where the non-linear block is a Volterra system (which can be read about in Section 2.4).

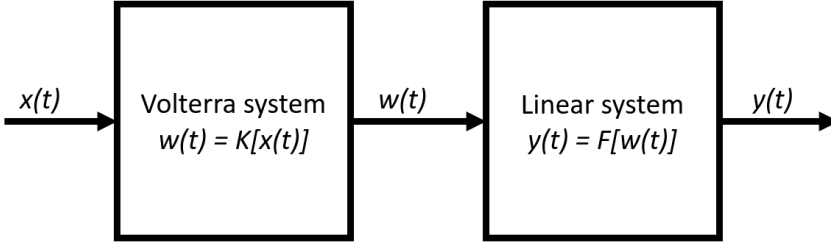


Figure 2.3: System with a non-linear subsystem (a Volterra system) and linear subsystem chained together.

2.6.2 Identification with log-sine-sweeps

When identifying a Hammerstein model with a Volterra system as the non-linear part (as in Figure 2.3) for identifying room acoustics, the linear parts (of both the speaker and room acoustics) are often the interesting parts. Farina has presented a method of doing so [4], which Rébillat et al. has expanded [9]. Advantages of these methods are that

- they do not require tight synchronization between the input and the output, which otherwise can be hard to obtain in digital sound systems including PCs, and that
- the non-linear parts of the system can be easily removed by truncating away the first half of the estimated impulse response

For identification purposes, the input signal $x(t)$ of length T seconds is used by both Farina and Rébillat, defined as

$$\begin{aligned}
 x(t) &= \sin \left(\frac{\omega_1 T}{\ln \left(\frac{\omega_2}{\omega_1} \right)} \cdot \left(e^{\frac{t}{T} \ln \left(\frac{\omega_2}{\omega_1} \right)} - 1 \right) \right) = \\
 &= \sin \left(\frac{2\pi f_1 T}{\ln \left(\frac{f_2}{f_1} \right)} \cdot \left(e^{\frac{t}{T} \ln \left(\frac{f_2}{f_1} \right)} - 1 \right) \right),
 \end{aligned} \tag{2.5}$$

where f_1 and ω_1 are the instantaneous frequencies at $t = 0$ in Hertz and radians per second, respectively and f_2 and ω_2 are the instantaneous frequencies at $t = T$ in Hertz and radians per second, respectively.

The difference between Farina's and Rébillat's input signals is the length T seconds, where Rébillat modifies the length T specified from the user to T_{Reb} , which satisfies $T_{\text{Reb}} > T$, defined as

$$T_{\text{Reb}} = \left(2m\pi - \frac{\pi}{2} \right) \cdot \frac{\ln\left(\frac{f_2}{f_1}\right)}{2\pi f_1 f_s}, \quad (2.6)$$

where

$$m = \left\lceil \frac{2\pi T f_s}{\ln\left(\frac{\omega_2}{\omega_1}\right) \omega_1 + \frac{\pi}{2}} \right\rceil, \quad (2.7)$$

and f_s is the sampling frequency. Using the length T_{Reb} seconds instead of T seconds gives $x(t)$ mathematically the correct phase properties [9].

The inverse signal to $x(t)$ is $x_{\text{inv}}(t)$ is the signal that gives a Dirac's delta function when convoluted with $x(t)$. Although, since $x(t)$ is not infinite in time, a Dirac's delta function is not possible to obtain by convoluting the signal with another signal. In Figure 2.4, 2.5 and 2.6 examples of an input signal, its inverse and the convolution between them is shown. The inverse signal is calculated using the Hammerstein toolbox in Matlab [9].

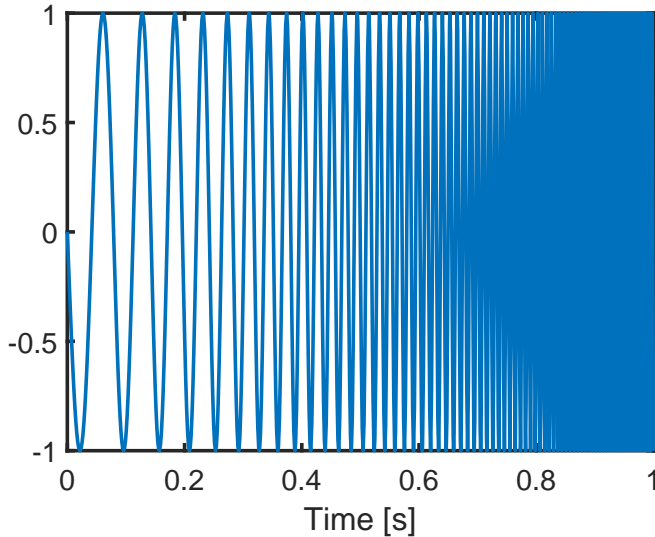


Figure 2.4: Log-sine-sweep (Rébillat's method), with $f_1 = 10$ Hz, $f_2 = 200$ Hz and a $T = 1$ second.

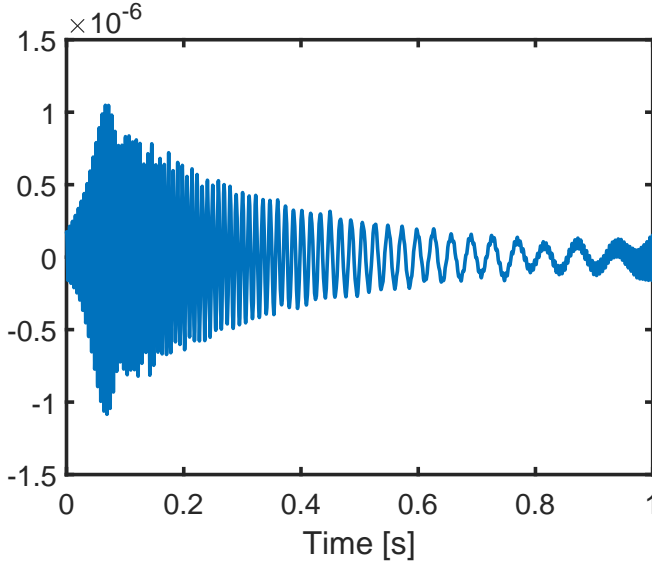


Figure 2.5: Inverse filter for the log-sine-sweep shown in Figure 2.4, with $f_1 = 10$ Hz, $f_2 = 200$ Hz and a $T = 1$ second.

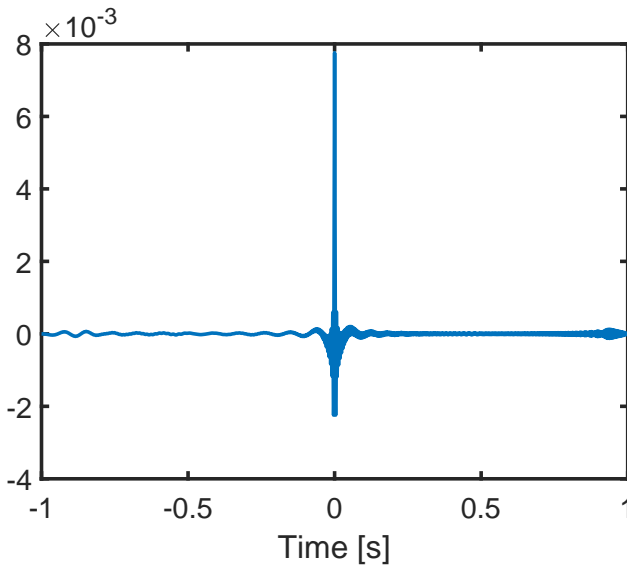


Figure 2.6: Convolution of Log sine sweep and its inverse filter from Figures 2.4 and 2.5, with $f_1 = 10$ Hz, $f_2 = 200$ Hz and a $T = 1$ second. Result is a sinc-like function which is approximately a Dirac's delta function.

2.7 Spectral analysis

2.7.1 Normalization

Normalization of measurements might be of interest when comparing different measurements to each other, where the absolute gain is not of interest. If signal $y(t)$ is received, it can be normalized to 0 dB for the frequency interval $[f_{\text{lower}}, f_{\text{upper}}]$ by

$$y_{\text{Norm}} = \frac{1}{\frac{1}{f_{\text{upper}} - f_{\text{lower}}} \cdot \sum_{f=f_{\text{lower}}}^{f_{\text{upper}}} \Phi(f)} y(t), \quad (2.8)$$

where y_{Norm} is the normalized signal and $\Phi(f)$ is the power spectral density (PSD) of $y(t)$, where $\Phi(f)$ can be replaced with an estimate of the PSD $\hat{\Phi}(f)$.

2.7.2 Octave smoothing

Octave smoothing is a smoothing method where the size of the smoothing window increases for larger frequencies. When smoothing with $1/N$ -octave smoothing, the window size is $1/N$ octave large. E.g., for the frequency interval 50-100 Hz, the window size is 50 Hz, and for the frequency interval 1000-2000 Hz, the window size is 1000 Hz.

This method lets important details be left in the lower frequencies (the bass, for acoustical problems), and make a smooth spectra for high frequencies.

2.8 Regression methods

In this section, different regression methods and their properties are presented.

2.8.1 Linear regression

For prediction or estimation problems, linear regression can be used to create models which maps a set of features $X \in \mathbb{R}^{M \times N}$ to a set of target variables $y \in \mathbb{R}^N$. This is done by finding coefficients $a \in \mathbb{R}^M$ and creating a model

$$y = a^T X + \epsilon, \quad (2.9)$$

where N is the amount of data points gathered, M the dimension of each data point and $\epsilon \in \mathbb{R}^N$ is the error of each prediction. A common way to find a suitable coefficient vector a is to minimize the MSE $\|\epsilon\|_2^2$. The optimization problem is then given by

$$\min_{a \in \mathbb{R}^M} \frac{1}{N} \|y - a^T X\|_2^2, \quad (2.10)$$

for which

$$a = (X^T X)^{-1} X^T y \quad (2.11)$$

is the optimal solution. [7]

2.8.2 L1 regularized linear regression - Lasso

Lasso optimization is an regression analysis method aimed to only let important features get non-zero coefficients. The optimization setup is

$$\min_{a \in \mathbb{R}^N} \frac{1}{N} \|y - a^T X\|_2^2 + \lambda \|a\|_1, \quad (2.12)$$

where $\lambda \in \mathbb{R}^+$ is design parameter which regulates size of the coefficients a . Equation 2.12 is the Lagrangian form of

$$\begin{aligned} \min_{a \in \mathbb{R}^N} \quad & \frac{1}{N} \|y - a^T X\|_2^2 \\ \text{s.t.} \quad & \|a\|_1 < t, \end{aligned} \quad (2.13)$$

where the t is a constant dependent on λ and the relationship between t and λ that makes the forms equivalent is data dependent. In this form, it is clear that the choice of the design parameter t restricts the size of the coefficients a . [14]

If λ is increased in size, less coefficients in a will have non-zero values. Therefore, to find a solution which gives a given amount of non-zero coefficients D_{\max} , a grid search of lambdas can be made to find a satisfying solution.

2.9 Shelving filter

A Shelving filter is a filter that increases the magnitude either above or below a certain cut-off frequency, while keeping all the other frequencies magnitudes the same [18]. If the cut-off frequency is low enough and the Shelving filter is constructed so that it increases the gain below the cut-off frequency, the filter is a suitable filter for increasing the bass in audio applications.

A second order filter as

$$y(t) = -a_1 y(t-1) - a_2 y(t-2) + b_0 x(t) + b_1 x(t-1) + b_2 x(t-2) \quad (2.14)$$

is a bass boosting Shelving filter with cut-off frequency f_c , sample frequency f_s and gain G (in dB) if the coefficients are defined as

$$\begin{aligned} b_0 &= \frac{1 + \sqrt{2V_0}K + V_0K^2}{1 + \sqrt{2}K + K^2}, & a_1 &= \frac{2(K^2 - 1)}{1 + \sqrt{2}K + K^2} \\ b_1 &= \frac{2(V_0K^2 - 1)}{1 + \sqrt{2}K + K^2}, & a_2 &= \frac{1 - \sqrt{2}K + K^2}{1 + \sqrt{2}K + K^2} \\ b_2 &= \frac{1 - \sqrt{2V_0}K + V_0K^2}{1 + \sqrt{2}K + K^2} \end{aligned} \quad (2.15)$$

as described in [18]. The parameters K and V_0 are defined by

$$\begin{aligned} K &= \tan\left(\frac{\pi f_c}{f_s}\right) \\ V_0 &= 10^{\frac{G}{20}}. \end{aligned} \quad (2.16)$$

If the cut-off frequency is set to $f_c = 3000$ Hz and the gain to $G = 5$ dB the magnitude response will be as in Figure 2.7

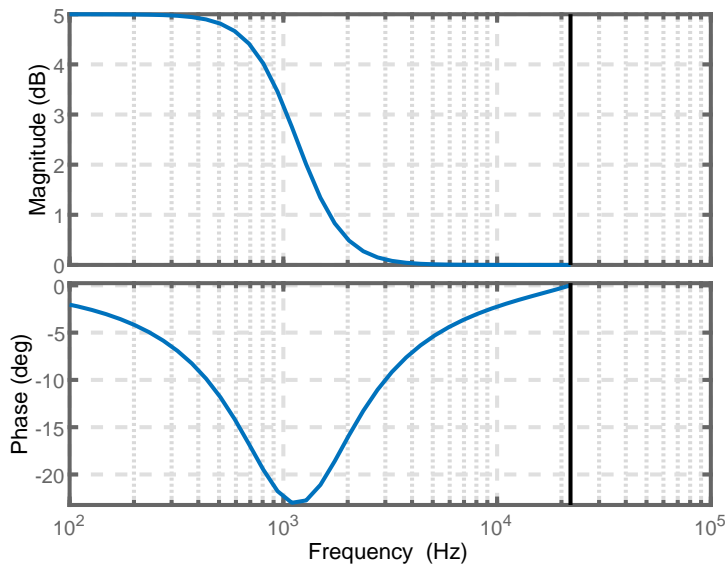


Figure 2.7: Magnitude response for a Shelving filter with boost for low frequencies, where $f_c = 3000$ Hz and $G = 5$ dB.

3

Method

3.1 Signal and system model

In this section, models for systems and subsystems are discussed, such as for the speaker, the microphones and the room acoustics.

3.1.1 Model of whole system

In Figure 3.1 the whole system, from input $x(t)$ to final output $y_{\text{tot},s}(t)$, is presented. The system that generates output $y_{\text{tot},s}(t)$ is specific for a set of variables \mathbf{s} , called the setup \mathbf{s} . The setup \mathbf{s} defines the system and is defined as

$$\mathbf{s} = (\mathbf{p}_{\text{speaker}}, \mathbf{p}_{\text{mic}}, m), \quad (3.1)$$

where

$$\mathbf{p}_{\text{speaker}} = (p_{\text{speaker}}^{(x)}, p_{\text{speaker}}^{(y)}) \quad (3.2)$$

defines the x- and y-coordinates of the speaker,

$$\mathbf{p}_{\text{mic}} = (p_{\text{mic}}^{(x)}, p_{\text{mic}}^{(y)}) \quad (3.3)$$

defines the x- and y-coordinates of the microphone and m defines what microphone is used.

Subsystem 1, representing the properties for the speaker, is denoted $\mathcal{H}_{\text{speaker}}$. Subsystem 2, representing the room's acoustical properties, is denoted $\mathcal{H}_{\text{room},s}$ and is parameterized by \mathbf{s} , as described above. Subsystem 3, representing a microphone's properties, is denoted $\mathcal{H}_{\text{mic},s}$ and is also parameterized by \mathbf{s} . Each subsystem is explained in detail in the Sections 3.1.2, 3.1.3 and 3.1.4. The total system is defined as $\mathcal{H}_{\text{tot},s}$. The total system, with input $x(t)$, then becomes

$$y_{\text{tot},s}(t) = \mathcal{H}_{\text{tot},s}\{x(t)\} = \mathcal{H}_{\text{mic},s}\left\{\mathcal{H}_{\text{room},s}\left\{\mathcal{H}_{\text{speaker}}\{x(t)\}\right\}\right\}, \quad (3.4)$$

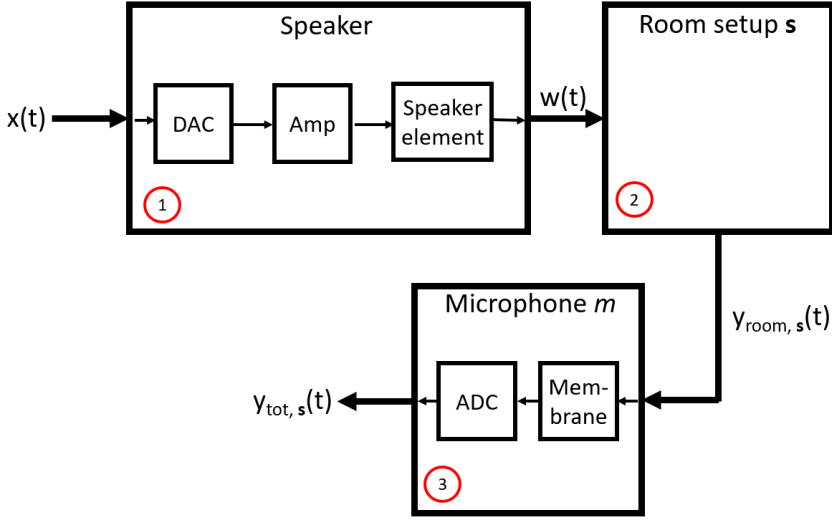


Figure 3.1: Flow chart of the whole system, from the input $x(t)$ generated from the computer to the DAC, to output $y_{\text{tot},s}(t)$ that is what microphone m outputs. Note that the system characteristics differ depending on which microphone m is observed and on the location of the speaker and the microphone. The different subsystem are numbered 1-3, seen in the red circle for each subsystem box.

for some setup s .

The whole system is built out of blocks, where (as later described) the first block have some non-linear properties and the second and third block can be considered linear. In following sections, the blocks will be described in more detail. Although, this makes the whole system a type of Hammerstein model, which are described in Section 2.6.1.

3.1.2 Model of speaker

The model used for modeling the speaker is a Volterra model of degree N , that is

$$\mathcal{H}_{\text{speaker}}\{x(t)\} = x(t) * k_1(t) + x^2(t) * k_2(t) + \dots + x^N(t) * k_N(t). \quad (3.5)$$

The method for estimating the impulse response for the whole system $\mathcal{H}_{\text{tot},s}\{x(t)\}$, which is described in Section 2.6.2, have the properties of being able to extract only the linear parts of the system. Therefore, it is possible to only consider the linear part of the system and do the approximation

$$\mathcal{H}_{\text{speaker}}\{x(t)\} \approx x(t) * k_1(t), \quad (3.6)$$

where the non-linear parts of the speaker are ignored.

3.1.3 Model of room acoustics

The RIR of the rectangular room associated with subsystem $\mathcal{H}_{\text{room},s}$, for sound outputted from a speaker and received at a microphone m , for a setup s , can be viewed as a sum of the all paths the sound impulse can take. Since the RIR only depends on reflections, the acoustics of the room form a linear system, i.e. $\mathcal{H}_{\text{room},s}$ is linear. The RIR for setup s can be modeled with the finite impulse response (FIR) model

$$h_{\text{room},s}(t) = \alpha_s^{(dp)} \delta(t - \tau_s^{(dp)}) + \sum_{i=1}^R \alpha_s^{(i)} \delta(t - \tau_s^{(i)}) + v_s(t), \quad (3.7)$$

where $\alpha_s^{(dp)}$ and $\alpha_s^{(i)}$ are attenuation coefficients, $\tau_s^{(dp)}$ and $\tau_s^{(i)}$ are the lags of a path (in samples). The term $\alpha_s^{(dp)} \delta(t - \tau_s^{(dp)})$ corresponds to the direct path between the speaker and the microphone. The term $\alpha_s^{(i)} \delta(t - \tau_s^{(i)})$ resembles a path i , for which the path includes at least one reflection on a wall, the ceiling or the floor (as [1] mentions in Section 4.3). Attenuation constant $\alpha_s^{(dp)}$ do not include any energy absorption of reflections and the energy loss comes only from the distance traveled by the sound wave. R is the number of paths of interest (excluding the direct path) and $v_s(t)$ holds all information about paths not of interest. The paths not of interest are path with very small $|\alpha_s^{(i)}|$. This is similar to how [2] have modeled a similar system.

For this thesis, the attenuation coefficients $\alpha_s^{(i)}$ for all possible i will be assumed frequency independent, as the authors in [2] have done. With this assumption, the system can be interpreted as

$$\begin{aligned} \mathcal{H}_{\text{room},s}\{w(t)\} &= h_{\text{room},s} * w(t) = \\ &= \alpha_s^{(DP)} w(t - \tau_s^{(DP)}) + \sum_{i=1}^R \alpha_s^{(i)} w(t - \tau_s^{(i)}) + v'_s(t) \end{aligned} \quad (3.8)$$

for an input $w(t)$ and output $y_{\text{room},s}$, where $v'_s(t)$ includes all paths not of interest.

3.1.4 Model of microphones

The system for the microphones, $\mathcal{H}_{\text{mic},s}$, is assumed to be linear. Hence, if wanted and if the system of the microphone is known, the system frequency response's magnitude can be inverted to find the input (within the systems bandwidth). The microphone frequency response magnitude is also nearly flat within its bandwidth and all microphones are omni-directional [8] [10]. Therefore, the microphones are not considered affecting the signal in a significant way.

The bandwidth of the Umik-1 microphone is 20 - 20000 Hz [10] and for UMA-8:s microphones it is 100-10000 Hz [8].

3.2 Measurements

In this section the approach and methodology for the measurements are described. There are two types of measurements made, which are:

1. Room frequency response measurements
2. Measurements for speaker position estimation

where the first type, room frequency response measurements, have the aim to identify how the speaker position affects what the listener (within area A , as described in Chapter 1) hears and then make a model of how to correct for the room acoustics according to the speaker position. The second type of measurements, measurements for speaker position estimation, have the aim to identify the speaker position as well as possible. This could either be estimating the room coordinates of the speaker or the distance to the two closest walls to the speaker.

For the room frequency response measurements, a reference microphone with nearly flat frequency response magnitude is used and is placed in 12 different positions within area A . The average frequency response over the 12 positions is calculated. In this case, the speaker and microphones are not co-located.

For the measurements for speaker position estimation, a circular microphone array is used and is placed on top of the speaker. In this case, the microphone array and the speaker are co-located, i.e. the microphone array and the speaker are approximately in the same position.

For the main part of the thesis, 16 speaker positions are used for measurements (labeled 1-16). These positions form a 4x4-grid and are chosen due to being suitable for algorithm development. In the later part of the thesis, 4 new measurements with new speaker positions will be done (labeled 17-20), in order to evaluate the results. These 4 new speaker positions are randomly chosen.

3.3 Finding correct filter parameters

To be able to correct the room frequency response, the speaker position's effect on the room frequency response is studied. Specifically, the goal is to model how the Shelving filter parameters G and f_c should be set in order to correct the bass for the listener. As stated in Section 2.3.3, the bass increases in power (compared to the other frequencies) when the speaker is placed near a corner. From this statement, the interesting frequencies to study are frequencies where the power of the output decreases if the speaker is placed further away from a corner. From this it is possible to find a suitable cut-off frequency f_c . Then, to find a suitable gain parameter G , a model is made of how the speaker position affects the output's bass power.

Several models are tested and evaluated for predicting the desired magnitude correction G . The most suitable set of features will then be chosen to predict G . Features that have been used in these models are:

- Position: x, y

- Distance to the two closest walls: $d_{\min} = \min\{x, y\}$ and $d_{\max} = \max\{x, y\}$.
- Distance to closest corner: d_{corner}
- Distance to listener: d_{listener}
- Position squared: x^2, y^2
- Distance to closest corner, squared: d_{corner}^2
- Distance to listener, squared: d_{listener}^2

Note that using the features d_{\min} and d_{\max} is basically the same as using the position x and y , with the difference that it is not possible to tell which distance to the wall belongs to which axis.

In following Table 3.1, features used for each model tested are shown, where the bass gain G is predicted with a linear regression model (for which the coefficients minimize the MSE):

Model label	Features used
1	x, y
2	d_{\min}
3	d_{\min}, d_{\max}
4	d_{corner}
5	d_{listener}
6	$x, y, d_{\text{listener}}$
7	x
8	x, y, x^2, y^2
9	$d_{\text{listener}}, d_{\text{listener}}^2$
10	$x, y, d_{\text{listener}}, d_{\text{listener}}^2$
11	x^2, y^2
12	$d_{\min}, d_{\max}, d_{\text{corner}}$

Table 3.1: Features that have been used for different models.

E.g., for Model 1 the model will be

$$G = \beta_0 + \beta_1 x + \beta_2 y, \quad (3.9)$$

where G is the magnitude for the correction filter (in dB), x and y are the features and $\beta_i, i = 1, 2, 3$, are the model coefficients which minimize the MSE.

3.4 Localization from RIR

Estimates of the speaker position are of interest, so that a correction filter can be calculated using the model from Section 3.3. To do this, the goal is to search for

reflections from the walls, the floor and the ceiling that hold information about the room and the speaker position.

In order to estimate the speaker position with the microphone array co-located with the speaker, Rébillat's method (described in Section 2.6.2) is used to find the impulse response $h_{\text{tot},s}(t)$ for each microphone m on the microphone array. The log-sine-sweep is played through the speaker and recorded with the microphone array, for which the microphones are synchronized with each other.

3.4.1 Correcting impulse responses

The deconvolution using Rébillat's method gives a raw impulse response $h_{\text{tot},s}(t)$ and in order to use this impulse response it has to be corrected in several ways. First, the beginning of the DP (direct path) reflection, t_{start} , is gotten by finding the lowest time that satisfies

$$|h_{\text{tot},s}(t_{\text{start}})| > 0.05 \cdot \max_{\tau} \{|h_{\text{tot},s}(\tau)|\}, \quad (3.10)$$

which should be the same for all microphones on the array. Removing the part of the impulse response before t_{start} removes the non-linear properties in the impulse response [4] [9].

For example, the impulse responses for a certain position of the measurements explained in Chapter 4 can be seen in Figure 3.2, which shows the impulse response after removing the first unnecessary part. Henceforth, the clipped impulse response (without the first part) is called only 'impulse response'.

In the first milliseconds of the impulse response for each microphone, the direct path impact of the sound from the speaker can be seen. This part has significantly more energy than the rest of the impulse response, which is due to that the sound has traveled only a very small distance and have also not lost any energy from absorption of wall reflections.

From the measurements in a certain speaker position for which the distance to the walls is large, the part in the time interval $0-t_{\text{ref}}$ ms can be extracted to use as an estimate of the direct path part of each microphone (from now on denoted $h_m^{(\text{DP,ref})}(t)$), for which t_{ref} is a (preferably large) time t where no wall or ceiling reflections are included in the impulse response for time interval $0-t_{\text{ref}}$. Hence, this part has no wall or ceiling reflections in it, although the reflections from the floor are included. Then, this direct path estimation has been subtracted from the other measurements, so that only the impulse response without the direct path is left. The corrected impulse response is defined as

$$h_{\text{tot},s}^{(\text{corr})}(t) = \begin{cases} h_{\text{tot},s}(t) - h_m^{(\text{DP,ref})}(t - \tau_{\text{lag}}), & 0 < t < t_{\text{ref}} \\ h_{\text{tot},s}(t), & t \geq t_{\text{ref}} \end{cases} \quad (3.11)$$

for each measurement setup s , where $h_{\text{tot},s}^{(\text{corr})}(t)$ is the corrected impulse response of $h_{\text{tot},s}(t)$ and τ_{lag} is a lag constant due to not being able to synchronize the speaker and the microphone array. To synchronize the speaker and the microphone array's synchronization differences, the lag τ_{lag} between them has been

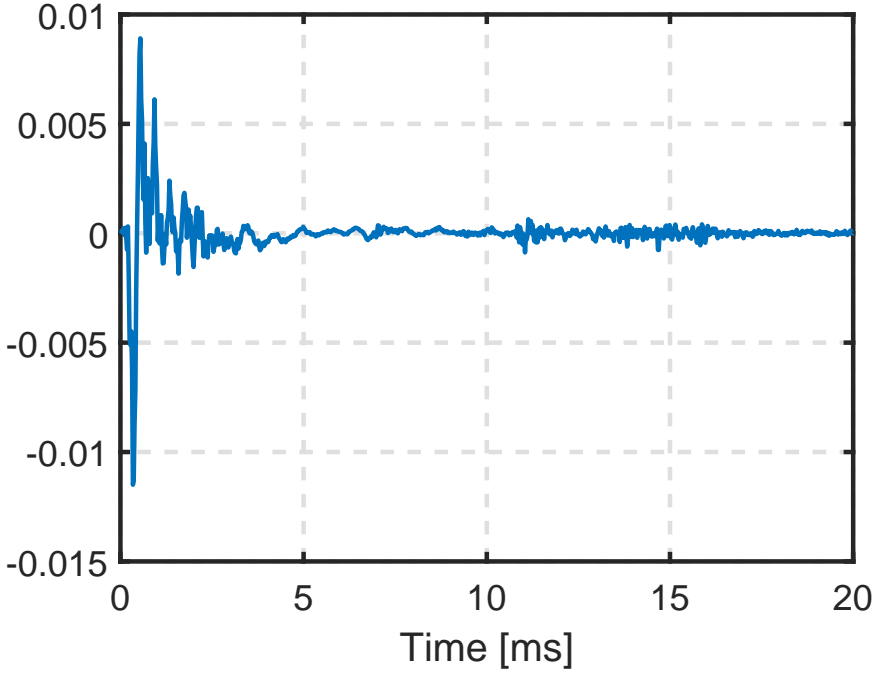


Figure 3.2: Example of relevant impulse response for setup \mathbf{s} , for a certain microphone on the microphone array.

found by using cross-correlations maximum point for $h_{\text{tot},\mathbf{s}}(t)$ and $h_m^{(\text{DP},\text{ref})}$. An example of the result of subtracting the direct path can be seen in Figure 3.3. There are still some artifacts from the direct path left, but has considerably less energy than before.

In the corrected impulse response $h_{\text{tot},\mathbf{s}}^{(\text{corr})}(t)$ in Figure 3.3, it is possible to see the ceiling reflection (at about 11 ms) and a wall reflection at about 7 ms. Around 4 ms after the ceiling reflection, there is a lot of energy in the impulse response, which supposedly mostly comes from second order reflection from the ceiling and the walls.

3.4.2 Lasso for finding reflections

To find out where the reflections are in the corrected impulse response $h_{\text{tot},\mathbf{s}}^{(\text{corr})}(t)$ for a setup \mathbf{s} , a Lasso linear regression is used to estimate the attenuation constants $\alpha_s^{(i)}$, notated $\hat{\alpha}_s^{(i)}$. Since the microphone has some gain, it is not the true $\alpha_s^{(i)}$ that are estimated, but a scaled attenuation constant. Although, $\alpha_s^{(i)}$ for all i are scaled the same for each microphone, and the relevant information is the proportion between $\alpha_s^{(i)}$ for different i . Hence, the scaling factor is simply ignored.

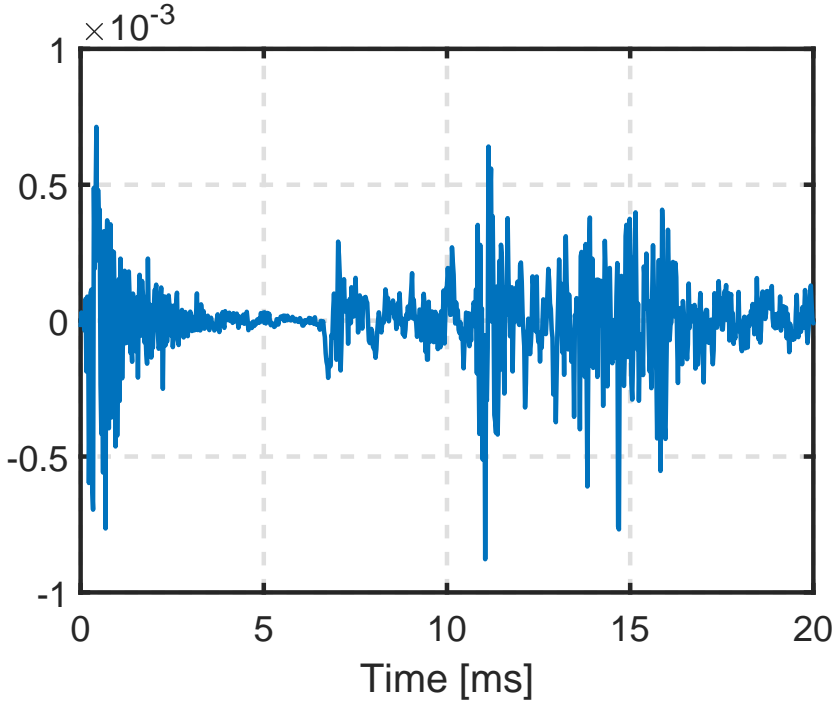


Figure 3.3: Impulse response for the same setup s as in Figure 3.2, but with the direct path removed. Time constant t_{ref} is set to 9.1 ms.

For the estimation, a matrix H_m consisting of delayed and zero-padded direct path is made for each microphone m . Let τ_{max} be the maximum delay considered (in seconds), which corresponds to the k_{max} th sample in the impulse responses. Then let $h_m^{(\text{DP,short})}(t) = h_m^{(\text{DP,ref})}(t)$, $t=[0 \ 1.1]$ ms, i.e. truncate $h_m^{(\text{DP,ref})}(t)$ to the first 1.1 ms. With this, the matrix H_m is constructed as

$$H_m = \begin{bmatrix} f_{\text{delay}} \left\{ h_m^{(\text{DP,short})}(t), 0 \right\} \\ f_{\text{delay}} \left\{ h_m^{(\text{DP,short})}(t), 1 \right\} \\ \vdots \\ f_{\text{delay}} \left\{ h_m^{(\text{DP,short})}(t), k_{\text{max}} \right\} \end{bmatrix}, \quad (3.12)$$

where the dimension of the matrix is $H_m \in \mathbb{R}^{N \times k_{\text{max}}}$, N is the amount of samples in $h_{\text{tot},s}^{(\text{corr})}(t)$ and the function f_{delay} is defined as

$$\begin{aligned} f_{\text{delay}} \{x(t), k\} &= \\ &= \underbrace{\begin{bmatrix} 0 & \dots & 0 \end{bmatrix}}_{k \text{ elements}} \quad x(0 \cdot T_s) \quad x(1 \cdot T_s) \quad \dots \quad x(k_{\text{max}} \cdot T_s) \quad 0 \quad \dots \quad 0 \end{aligned} \quad (3.13)$$

for a vector $x(t) \in \mathcal{R}^{k_{\max}}$ and gives $f_{\text{delay}} \{x(t), k\} \in \mathcal{R}^N$, for which $T_s = 1/f_s$ is the sampling period time. Using the Lasso, the optimization problem is defined as

$$\min_{\hat{\alpha}_s \in \mathbb{R}^N} \frac{1}{N} \left\| h_{\text{tot},s}^{(\text{corr})} - \hat{\alpha}_s^T H_m^T \right\|_2^2 + \lambda \|\hat{\alpha}_s\|_1. \quad (3.14)$$

where

$$h_{\text{tot},s}^{(\text{corr})} = [h_{\text{tot},s}^{(\text{corr})}(0) \quad h_{\text{tot},s}^{(\text{corr})}(1 \cdot T_s) \quad \dots \quad h_{\text{tot},s}^{(\text{corr})}((N-1) \cdot T_s)]. \quad (3.15)$$

When solving (3.14), λ is set such that the maximum number of non-zero values in $\hat{\alpha}_s$ is as close to design parameter D_{\max} as possible.

From this, the vector $\hat{\alpha}_s$ non-zero values represent reflections, and the time delay can be found by finding the index for each non-zero value. For each estimation, a maximum delay τ_{\max} is set to limit the amount of non-zero $\hat{\alpha}_s^{(i)}$ and to exclude reflections coming from the ceiling. Also, all reflections that come from objects and walls closer than 0.3 meters are ignored, since there are no walls at this distance. This is done by setting $h_{\text{tot},s}^{(\text{corr})}(t) = 0$ for $t = 0, \dots, t_{0.3m}$, where time $t_{0.3m}$ corresponds to distance 0.3 meters (see Equation 3.18).

Each microphone's $\hat{\alpha}_s$ are then summed up, for each time sample and creates $\hat{\alpha}_s^{(\text{sum})}$ as

$$\hat{\alpha}_s^{(\text{sum})} = \sum_{m \in \mathbb{M}} \hat{\alpha}_s, \quad (3.16)$$

where \mathbb{M} is the set of microphones on the microphone array. Then, the summed vector $\hat{\alpha}_s^{(\text{sum})}$ is smoothed by convoluting a Hanning window

$$w_{\text{Hanning}}(k) = 0.5 \left(1 - \cos \left(2\pi \frac{k}{20} \right) \right), \quad 0 < k < 20 \quad (3.17)$$

for which the result can be seen in the blue solid line Figure 3.5. All negative values are set to 0 for the convoluted $\hat{\alpha}_s^{(\text{sum})}$, since negative values do not affect the reflection estimation and are of no interest. From this, the two maximum points' indices are used for the estimated wall distances \hat{d}_{\min} and \hat{d}_{\max} (after recalculating index numbers to distances, see Equation 3.18).

Although, some clutter is present (e.g. at 0.4m in the top plot). The distance d_{meter} in the x -axis is calculated with the function

$$d_{\text{meter}} = \frac{t_{ms}}{1000} \cdot \frac{c}{2}, \quad (3.18)$$

where d_{meter} is the distance to the speaker (in meters), t_{ms} is the time as in the x -axis for Figure 3.3 (in milliseconds) and $c = 343$ m/s is the speed of sound.

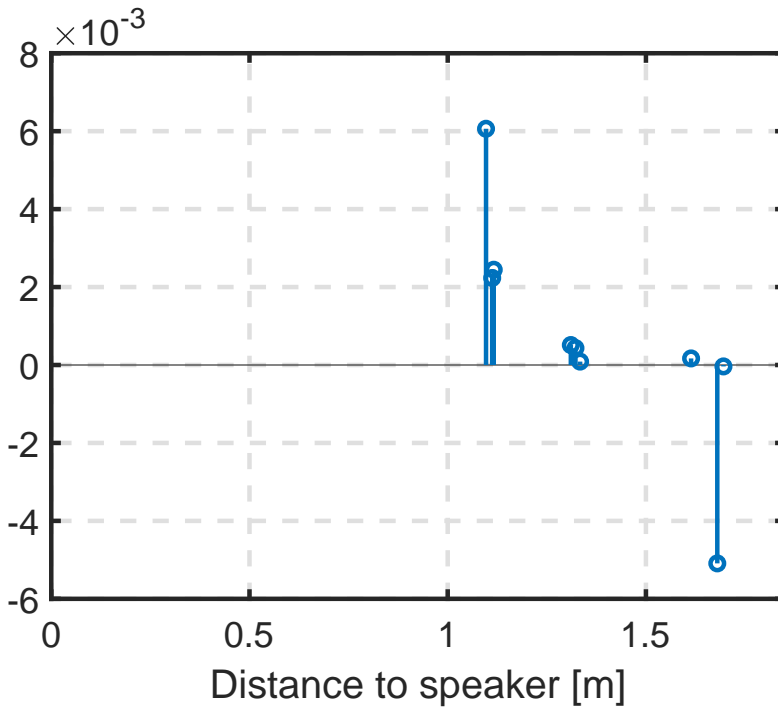


Figure 3.4: Example of estimated attenuation constants \hat{a}_s for the corrected impulse response in 3.3, with optimization parameter $D_{\max} = 10$. Only parts of impulse response corresponding to 0.3-1.85m have been considered.

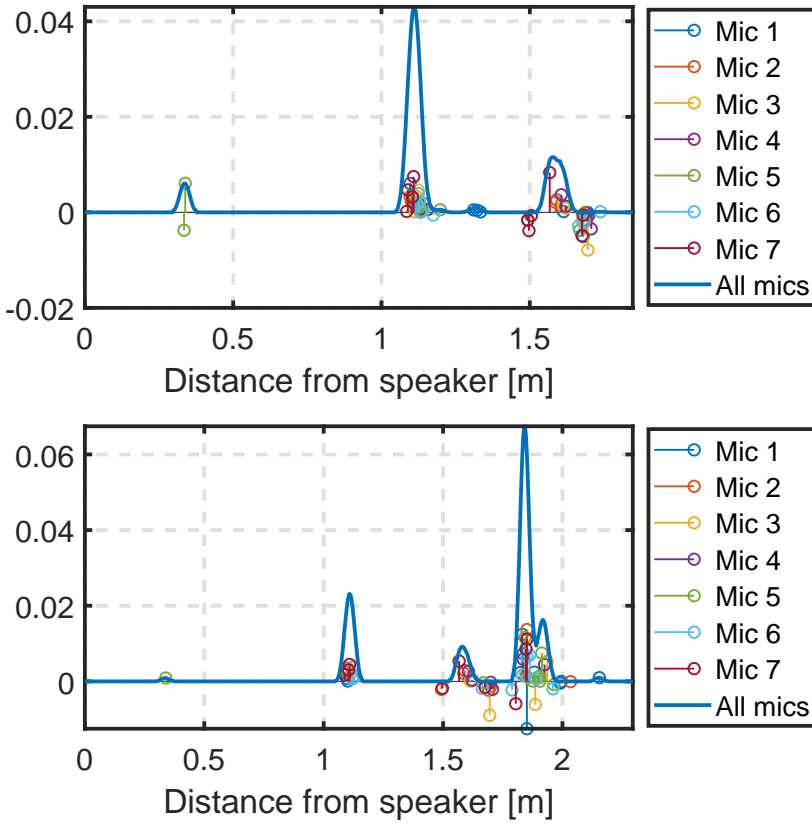


Figure 3.5: Example of estimated attenuation constants $\hat{\alpha}_s$, with optimization parameter $D_{max} = 10$. Only parts of impulse response corresponding to 0.3-1.85m have been considered for the top plot, and 0.3-2.3m for the bottom plot. Sum of the attenuation constant estimation over all the microphones, for each time sample, has been smoothed with a Hanning window of size 20 samples, represented by the blue line.

4

Results and Discussion

4.1 Setup

In this section, the hardware, software and measurement room used in the thesis are described.

4.1.1 Hardware and software

The software used in this thesis is

- **Dirac's HDSound** - for measuring the room frequency response for different speaker positions.
- **Matlab 2018a** - for analyzing measurement data and implementing algorithms that outputs a correction filter from the speaker position estimates. Matlab packages used includes:
 - **Hammerstein toolbox**, based on [9], available at Matlab's File Exchange.
- **Dirac Studio** - for real-time implementation of correction filters (with functionality to turn on/off in real time).

The hardware used is

- **Behringer's 1C-BK** as a speaker
- **t.amp's TA50** as an amplifier to the speaker
- **miniDSP's Umik-1** as reference microphone (representing the listener in the room) for which the magnitude of the frequency response can be seen in Figure C.2, Appendix C.

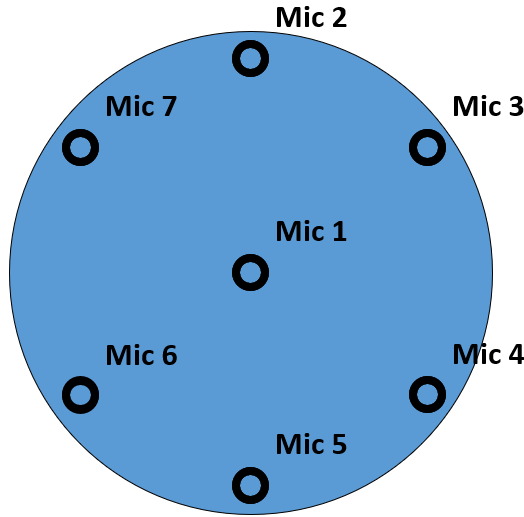


Figure 4.1: Layout of UMA-8 microphone array, seen from above.

- **miniDSP's UMA-8 Microphone array** as the on-board microphone array on the speaker with a microphone setup seen in Figure 4.1 and the magnitude of the frequency response in Figure C.2, Appendix C. The distance between Mic 1 and the other microphones is about 4.6 cm.
- **Focusrite's Scarlett 2i4 2nd Generation** as an audio interface for the PC
- **ASUS UX305CA Zenbook with Windows 10** as a PC that everything is connected to and where the software runs

The setup for the speaker with the UMA-8 mounted can be seen in Figure 4.2. The speaker was placed on a stool to get some height above the floor. All hardware and software are set to sample frequency $f_s = 44100$ Hz.

4.1.2 Room description

The room used for measurements is a conference room in the area Visionen at Linköping University. The plan for the room can be seen in Figure 4.3. The room is rectangular with the sides being 5.85 and 3.40 meters long, and the ceiling height is about 2.40 meters. One of the walls mainly consists of a glass wall, which starts 5 cm into the wall. For the wall to the right in Figure 4.3, there hangs a white board of a size which is common in conference rooms. In the left bottom corner in Figure 4.3, there were a table and some chairs pushed into the corner when the measurements were done.

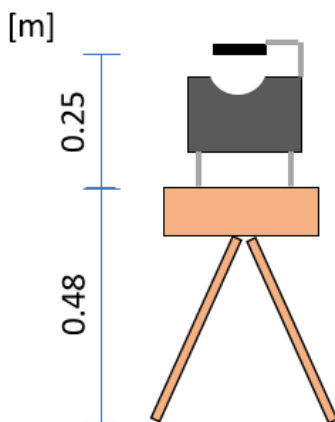


Figure 4.2: Speaker setup for doing measurements. The setup shows a speaker placed on a stool, with the microphone array UMA-8 (the black box) on top of the speaker.

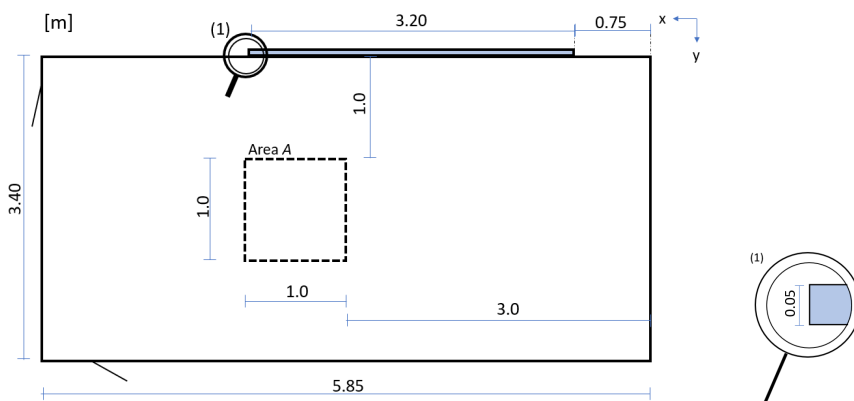


Figure 4.3: Drawing of the measurement room. The area A is where the supposed listener is placed.

4.2 Estimating filter gain from speaker position

In this section, results for finding a model that predicts filter gain G from a speaker position searched for. To find this model, the room frequency response measurements are studied in order to find reliable features. Then, linear regression is used to find model the coefficients that will be used.

4.2.1 Room frequency response measurements

Measurements were done to estimate the room frequency response, i.e. the frequency response derived from $\mathcal{H}_{\text{tot},s}\{x(t)\}$. For these measurements, the reference microphone Umik-1 is used (denoted $m = 0$), and therefore the setup is $\mathbf{s} = (\mathbf{p}_{\text{speaker}}, \mathbf{p}_{\text{mic}}, 0)$ for each speaker and microphone position $\mathbf{p}_{\text{speaker}}$ and \mathbf{p}_{mic} .

Since Umik-1's magnitude of frequency response do not differ more than ± 1 dB for different frequencies in the interval 20-20000 Hz, the approximation and assumption that the Umik-1 does not colorize the sound is used. Therefore, for some $\mathbf{s} = (\mathbf{p}_{\text{speaker}}, \mathbf{p}_{\text{mic}}, 0)$, the approximation

$$\begin{aligned} \mathcal{H}_{\text{tot},s}\{x(t)\} &= \mathcal{H}_{\text{mic},s}\left\{\mathcal{H}_{\text{room},s}\left\{\mathcal{H}_{\text{speaker}}\{x(t)\}\right\}\right\} \approx \\ &\approx \mathcal{H}_{\text{room},s}\left\{\mathcal{H}_{\text{speaker}}\{x(t)\}\right\} \end{aligned} \quad (4.1)$$

is done, as mentioned in Section 3.1.4 about microphone modeling.

For each speaker position $\mathbf{p}_{\text{speaker}}$ which was considered (seen in Table 4.2), twelve estimates of the frequency response were made. Each of those twelve measurements were done for a different reference microphone position \mathbf{p}_{mic} (seen in Table 4.1), all within the area A in the room. In Figure 4.4 the reference microphone positions are labeled with the letters **a-i**. Nine measurements, one for each reference microphone position **a-i**, were done for when the microphone was 158 cm. In addition to those nine measurements, three measurements for position **d-f** were made with the microphone being on the height 120 cm above the floor. The microphone was always pointing upwards, towards the ceiling.

In Table 4.1 the position of the microphone can be seen, if the coordinate system is set as described in Figure 4.4.

Position label	Position x-direction [m]	Position y-direction [m]	Height above floor [m]
a	3.0	1.0	1.58
b	3.0	1.5	1.58
c	3.0	2.0	1.58
d-high	3.5	1.0	1.58
e-high	3.5	1.5	1.58
f-high	3.5	2.0	1.58
g	4.0	1.0	1.58
h	4.0	1.5	1.58
i	4.0	2.0	1.58
d-low	3.5	1.0	1.20
e-low	3.5	1.5	1.20
f-low	3.5	2.0	1.20

Table 4.1: Reference microphone positions. Postfix -low and -high is to separate measurements where the microphone was 120 cm or 158 cm above the ground.

The settings for volume were set so that the volume knobs on both the audio interface and the amp were set to the middle position. The PC's internal volume was set to a value of 80, out of a 100.

In Table 4.2 the position of the speaker for the 16 different speaker positions on which measurement were made. The position makes up a 4x4 grid of positions, where the aim was to identify how the distance to walls and corners affected the frequency response for the lower frequencies. For all positions the speaker's front was pointing upwards.

For position 16, an additional measurement was done with the exact same speaker position, to see if the frequency response estimates were stationary within a time interval of a minute.

Then, for each speaker and microphone position $\mathbf{p}_{\text{speaker}}$ and \mathbf{p}_{mic} , Dirac's software HDSound was used to estimate the frequency response magnitude. For each speaker position $\mathbf{p}_{\text{speaker}}$, the estimated frequency response magnitudes were averaged over the twelve microphone positions \mathbf{p}_{mic} . After that, the frequency response magnitude was smoothed using an 1/8-octave filter and normalized. For the normalization, the lower frequency bound f_{lower} was set to 120 Hz and the higher bound f_{upper} to 3000 Hz. The resulting frequency response magnitude estimate is used as the RFRM estimate.

For each speaker position $\mathbf{p}_{\text{speaker}}$ 12 measurements for different reference microphone positions (position a-f) were made. This was done to so that the correction filter do not overfit to just one position, but will make a satisfying correction for the listener for an area. Otherwise, if only one or a very few reference

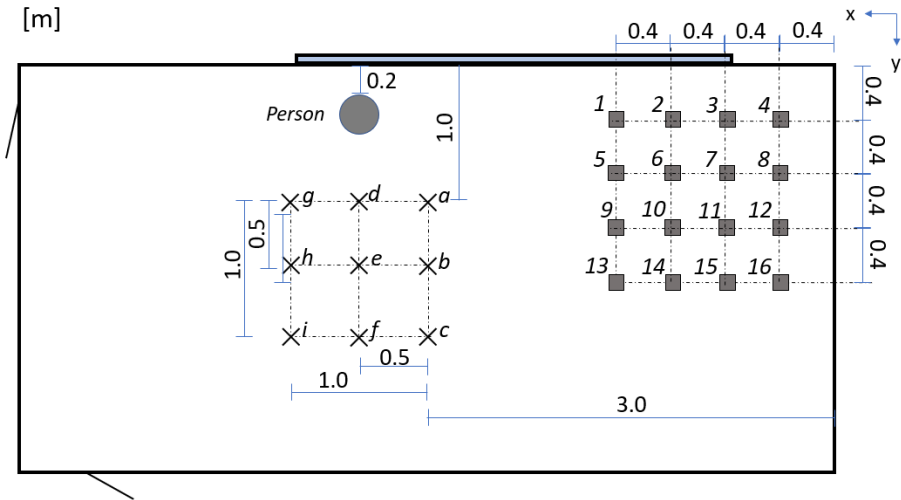


Figure 4.4: Positions for reference microphone (crosses) and speaker (squares). The point $(x, y) = (0, 0)$ is in the upper right corner of the room.

Speaker position label	Position x-direction [m]	Position y-direction [m]
1	0.4	0.4
2	0.8	0.4
3	1.2	0.4
4	1.6	0.4
5	0.4	0.8
6	0.8	0.8
7	1.2	0.8
8	1.6	0.8
9	0.4	1.2
10	0.8	1.2
11	1.2	1.2
12	1.6	1.2
13	0.4	1.6
14	0.8	1.6
15	1.2	1.6
16	1.6	1.6

Table 4.2: Speaker positions for measurements.

microphone positions had been used, the risk of making a correction for a specific frequency peak is higher. In Figure 4.5 the standard deviation for the 12 measurements can be seen, for speaker position 14. The standard deviation is quite high. Although, in Figure 4.6 it is possible to see that if the speaker is in the same position, the RFRM estimate does not change much between different measurements. Therefore, it is reasonable to assume that the noise in these measurements is low, that the signal-to-noise ratio (SNR) is high and that the measurements of the RIR are valid.

4.2.2 Position's impact on bass

To find for which frequency interval this bass-position-relation holds, nine patterns are looked at, where their frequency gains for the interval 40-250 Hz can be seen in Appendix B. The interval 40-250 Hz is suitable since the speaker's output power is very low below 40 Hz and 250 Hz is approximately the Schroeder frequency. In every legend in the figures, the positions at the top should have the highest bass gain and then the bass gain should decrease when going down in the list.

In all spectra, a peak at about 60 Hz can be seen. The height of the peak seems to decrease when placing the speaker further away from a corner or a wall, especially when placing the speaker further away from the wall with a whiteboard on it. This can partly be due to the whiteboards reflective properties, since a whiteboard does absorb very little of the sound in comparison to most walls [2], or could also partly be due to that the speaker is closer to the corner, it is also

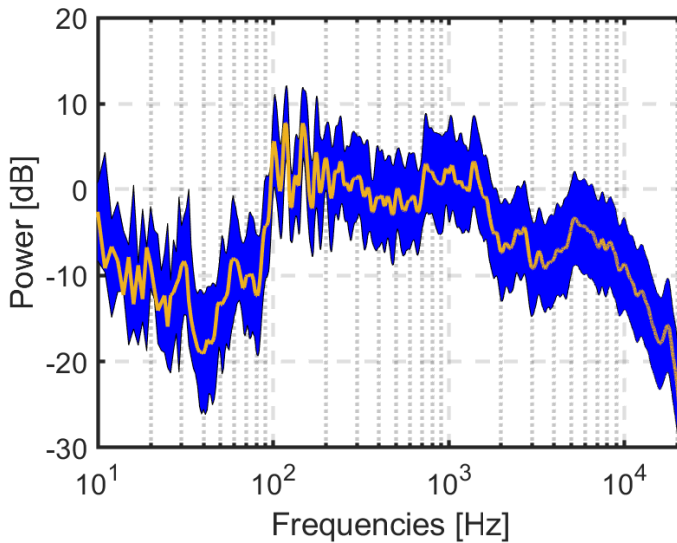


Figure 4.5: The estimated RFRM for the measured RIR (yellow solid line) for the speaker, with \pm one standard deviation added (blue area). All for speaker position 14.

further away from the listener (area A).

The tendency that bass has a higher gain if the speaker is closer to a wall can also be seen if looking at an average of frequency gain in the interval 50-80 Hz. For the frequency interval right above 80 Hz (about 80-110 Hz) there are attenuations for position 6 and 7, which do not follow the above mentioned pattern looked for. Therefore, the interval 50-80 Hz seems to be a reasonable interval to look at for room correction in this thesis and for this room.

Hence, a suitable value for the cut-off frequency is $f_c = 80$ Hz, since this is the highest frequency for which a good prediction can be made with the methodology in this thesis.

4.2.3 Model for correction gain from speaker position

To be able to make correction filters for the speaker, a model for the speaker position's effect on the RFRM has to be made. The aim is to make the average frequency gain in the interval 50-80 Hz having an average magnitude of 0 dB. For each position, that would need a filter that increases the magnitude by the amount seen in Table 4.3 and Figure 4.7.

In this section, the linear regression models presented in Section 3.3 are evaluated and the most suitable model is then used for predicting bass gain correction (in dB) from the speaker position. In Tables 4.3a and 4.3b (distinguished by feature types in the models) the RMSE and the maxima of absolute errors are shown. The RMSE and the maxima of absolute errors were calculated on the same

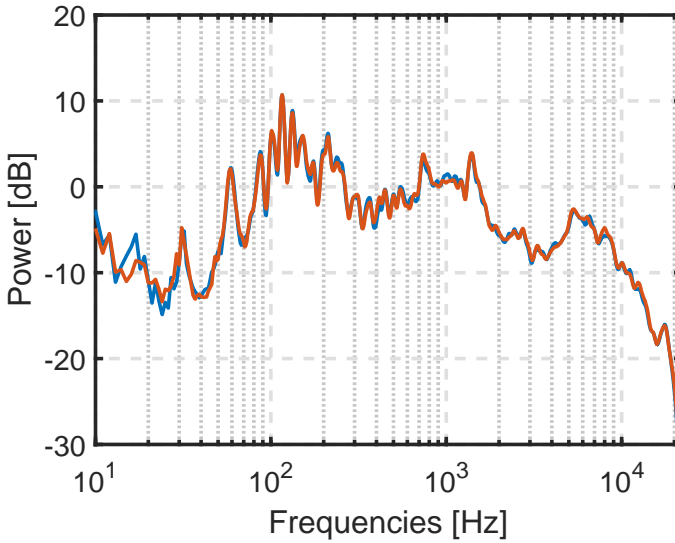


Figure 4.6: Two separate estimates of RFRM for speaker position 16. Both are very similar to each other, which shows that there seems to be consistency between different room measurements.

measurements as was used for finding model coefficients, i.e. measurements for speaker position 1-16. From these tables, it is possible to see that the best predictions are gained when the features x and y are given to the model. Adding more features do not result in an significantly lower RMSE, so to avoid using unnecessarily many features, Model 1 (with only x and y as features) seems to be the best model.

Features that do not need the information of coordinates, but only the distances to the walls, are d_{\min} , d_{\max} and d_{corner} , are shown in Table 4.4b. There seems to be a benefit in being able to identify from which wall the distance to the wall corresponds to, since the models in Table 4.4a seem to perform better than the models in Table 4.4b. This is reasonable, since the bass gain has a higher derivate along the x -axis than the y -axis, as seen in Tables 4.5 and 4.6. Therefore, a model with information about the coordinates, so that the model can differentiate between x - and y -values, should in general result in a better RMSE.

In the current stage of the speaker positions estimation algorithm, it is not possible to acquire estimates of the coordinates, but only of the distances to the walls. Therefore, a model using only the features d_{\min} , d_{\max} , d_{corner} is needed.

If the coordinates are known in the estimates from the speaker position estimation part, the best model that is not overfitted seems to be Model 1. It uses few features, has one of the best RMSEs and the max error is in the magnitude of 1 dB (which is within acceptable limits, as stated in Section 2.2). Adding features to model 1 does not improve the RMSE significantly.

If only information about distance to the two closest walls is given (Model 2,

Speaker position label	Magnitude correction for interval 50-80 Hz, G [dB]
1	8.44
2	7.16
3	4.17
4	1.98
5	9.15
6	7.85
7	4.89
8	2.43
9	10.42
10	8.77
11	5.51
12	3.23
13	11.57
14	10.55
15	6.29
16	3.98

Table 4.3: The correction needed for the frequency response within interval 50-80 Hz, for each speaker position.

3, 4, 12) in the room modeling part, Model 3 seems to perform the best. Adding features to this model, as in Model 12, does not improve it in any noticeable sense. Since only estimates of wall distances can be gotten, Model 3 is the model that will be used.

The model, with its coefficients, is then

$$G = -1.39 + 4.67d_{\min} + 3.63d_{\max}, \quad (4.2)$$

where G is the predicted bass gain for the correction filter (in dB).

4.3 Estimation of speaker position

In this section, the results for speaker position estimation is presented and evaluated.

4.3.1 Measurements for speaker position estimation

Measurements with an onboard microphone array (the UMA-8) was done to estimate the impulse response of the whole system for each speaker position. With this impulse response, the speaker's position can then be estimated, which is described in Section 3.4. The microphone array UMA-8 has 7 microphones on it, denoted $m = 1, \dots, 7$.

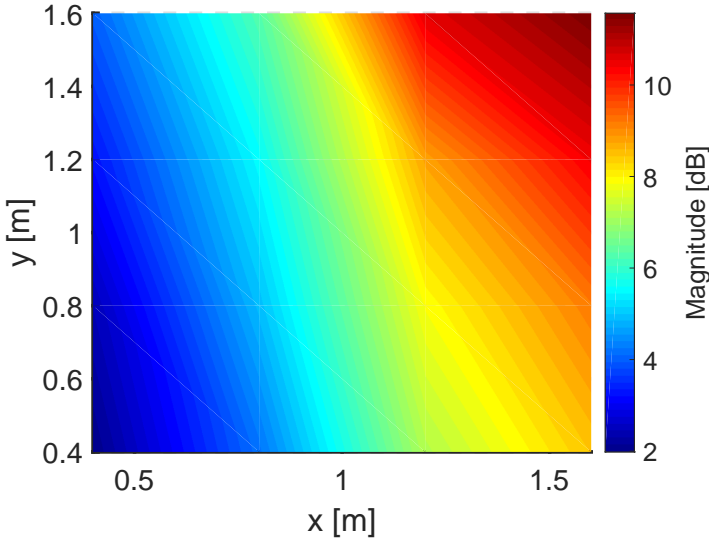


Figure 4.7: Colormap of magnitude correction for interval 50-80 Hz with x and y coordinates of the room on the horizontal and vertical axis. Values on the speaker positions 1-16 are the same as in Table 4.3. Values between speaker position 1-16 are calculated through linear interpolation between the data points.

In order to do this, the speaker was placed on the speaker positions 1-16 (as in Table 4.2). For each position a log-sine-sweep (using to Rébillat's method, as described in Equation 2.5), starting at 50 Hz ($f_1 = 50$) and up to 20 kHz ($f_2 = 20000$) and slightly longer than 2 seconds ($T = 2$), was played through the speaker. The volume was set to 70 out of 100 on the computer, and the knobs set to their middle position at the sound card and the amplifier. The log-sine-sweep was recorded simultaneously by each of the seven microphones on the microphone array. The microphones on the array are synced with each other, but the speaker is not synced with the microphone array (except that the speaker plays only approximately at the same time as the microphone start recording). The log-sine-sweep recorded by the microphones was then deconvoluted (using Rébillat's method described in Section 2.6.2) to find the impulse response of each system from the speaker to each microphone. The received impulse responses are denoted $h_{\text{tot},s}(t)$, and there are 7 for each speaker position, since $\mathbf{s} = (\mathbf{p}_{\text{speaker}}, \mathbf{p}_{\text{mic}}, m)$, $m = 1, \dots, 7$ for some speaker position $\mathbf{p}_{\text{speaker}}$. Note that a speaker position $\mathbf{p}_{\text{speaker}}$ also defines \mathbf{p}_{mic} , since the speaker and microphone array are approximately co-located and with known geometry. In Figure 4.8 a recorded signal for a single microphone on the UMA-8 can be seen. After 1.5 seconds, the signal is considerably lower in amplitude, which is due to that the microphones on the UMA-8 have a bandwidth of 100 to 10000 Hz.

(a)

Model label	Features used	RMSE [dB]	max{absolute error} [dB]
1	x, y	0.51	1.36
5	d_{listener}	0.71	2.05
6	$x, y, d_{\text{listener}}$	0.51	1.35
7	x	1.12	2.69
8	x, y, x^2, y^2	0.43	0.99
9	$d_{\text{listener}}, d_{\text{listener}}^2$	0.69	1.95
10	$x, y, d_{\text{listener}}, d_{\text{listener}}^2$	0.49	1.26
11	x^2, y^2	0.84	1.71

(b)

Model label	Features used	RMSE [dB]	max{absolute error} [dB]
2	d_{min}	1.77	4.00
3	$d_{\text{min}}, d_{\text{max}}$	1.31	2.33
4	d_{corner}	1.46	3.61
12	$d_{\text{min}}, d_{\text{max}}, d_{\text{corner}}$	1.31	2.33

Table 4.4: The result for different models and which features each model includes. The RMSE and maximum absolute error of these models is shown to the right. In Table 4.4a the models with features that need information about the coordinates x and y are shown. In Table 4.4b the models with features that do not need information about the coordinates x and y are shown, but only about the wall distances to the closest walls.

4.3.2 Finding reflections in impulse responses

Estimation with the help of the Lasso linear regression has been made as described in Section 3.4.2. The resulting peaks after the summing of coefficients and smoothing with Hanning window can be seen in Figure 4.9, with maximum distance from speaker set to $\tau_{\text{max}} = 1.85$ meter, optimization parameter $D_{\text{max}} = 10$ and impulse response correction parameters $t_{\text{start}} = 1.1$ ms and $t_{\text{ref}} = 9.1$ ms. The speaker position used for extracting $h_m^{(\text{DP}, \text{ref})}$, $m = 1, \dots, 7$, was speaker position 13. Parameter τ_{max} is set to 1.85 meter since that is the distance between the speaker and the ceiling.

In Table 4.8 the estimates for speaker position 1-16 are shown. The error is usually within ± 10 cm, if disregarding very big errors such as for speaker position 1, 2, 5, 9, 10 and 15. For those speaker positions, it seems that the highest peaks do not correspond to the correct wall reflection, resulting in errors which are very large.

The error tends to be positive for most speaker positions (as seen in Figure 4.10), which is supposedly due to the fact that $\tau_s^{(\text{DP})} > 0$ seconds. The UMA-8 microphone array is approximately 8 cm away from the closest speaker element

$\begin{matrix} \text{x} \\ \text{y} \end{matrix}$	0.4	0.8	1.2	1.6
0.4	-2.0	-4.2	-7.2	-8.4
0.8	-2.4	-4.9	-7.9	-9.2
1.2	-3.2	-5.5	-8.8	-10.4
1.6	-4.0	-6.3	-10.5	-11.6

Table 4.5: Coordinates (**bold**), in meters, and which mean gain they yield (**not bold**), in dB.

$\begin{matrix} d_{\min} \\ d_{\max} \end{matrix}$	0.4	0.8	1.2	1.6
0.4	-2.0	////////	////////	////
0.8	-2.4 and -4.2	-4.9	////////	////
1.2	-3.2 and -7.2	-5.5 and -7.9	-8.8	////
1.6	-4.0 and -8.4	-6.3 and -9.2	-10.5 and -10.4	-11.6

Table 4.6: Distances to two closest walls (**bold**), in meters, and which mean gain they yield (**not bold**), in dB.

(the bass element), which means that the true value of $\tau_s^{(DP)}$ should be $\tau_s^{(DP)} \approx 0.23$ ms. The resulting wall distances with bias correction can be seen in Table 4.9. The bias correction is calculated as the mean of errors that are not larger than ± 20 cm, which results in a bias correction of +5.5 cm.

4.4 Filter design and implementation

The Shelving filter is set to have a cut-off frequency $f_c = 80$ Hz, as discussed in Section 4.2.2, and the sampling frequency is set to $f_s = 44100$ Hz. The filter parameters are set to

$$K = \tan\left(\frac{80\pi}{44100}\right) \quad (4.3)$$

$$V_0 = 10^{\frac{G}{20}}$$

for which G is estimated with model in Equation 4.2. The filter coefficients can then be calculated as in Equation 2.15. The filter is implemented using Dirac's software Dirac Studio.

4.5 Tests of room correction

The algorithms in Sections 4.2 and 4.3 were combined and the result of is evaluated in this section. In Section 4.5.1, tests are done on the measurement data which was used to create the room correction algorithm and find suitable parameter values. In Section 4.5.2, the room correction is evaluated using new mea-

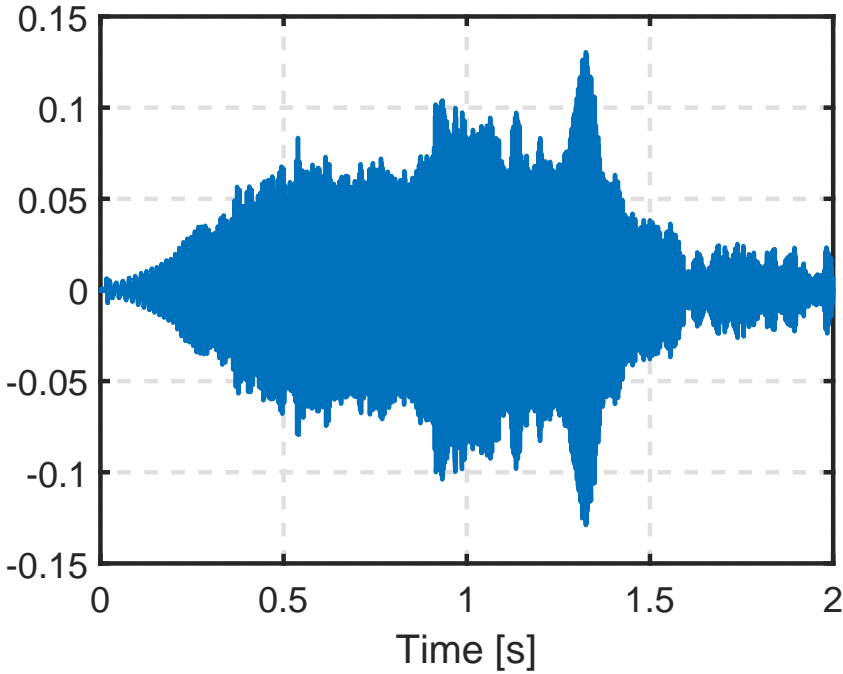


Figure 4.8: Recorded signal for microphone 1 for speaker position 14.

surement data, which was created after the room correction algorithm had been created and the parameter values set.

Parameter values were set to $D_{\max} = 10$, $\tau_{\max} = 1.85$ meter, $t_{\text{start}} = 1.1$ ms and $t_{\text{ref}} = 9.1$ ms. The measurements for RIR correction were taken from the measurements made at speaker position 13.

4.5.1 Tests on positions 1-16

To calculate which correction filter to use, model in Equation 4.2 is used for the part where the filter magnitude (parameter G) is estimated from the position. When estimating the wall distances \hat{d}_{\min} and \hat{d}_{\max} , a bias correction of +5.5 cm has been done.

For the frequency interval 50-80 Hz, the correction seem to be satisfying. Although, due to the filter slope for frequencies right above 100 Hz being not being steep enough, some peaks gets an extra unnecessary boost. E.g., for position 14 in Figure 4.11, the spectrum has very high peaks at 100 Hz, 118 Hz and 147 Hz, which gets boosted even more.

To evaluate the performance of the correction algorithm, the power mean and standard deviation over speaker positions are studied. The metrics are calculated for each frequency, as seen in Figure 4.12. The first moment and second moment

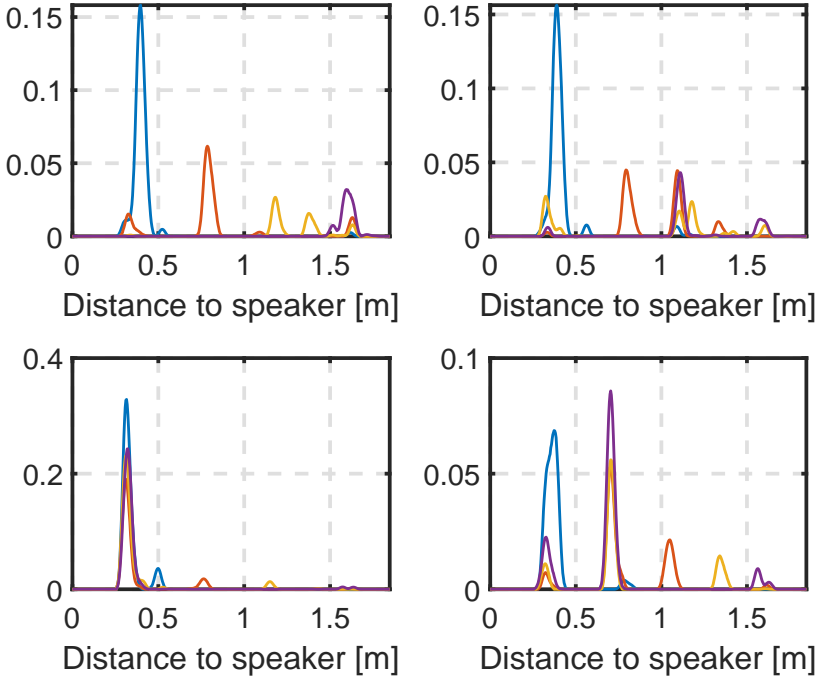


Figure 4.9: Estimates $\hat{\alpha}_s^{(\text{sum})}$ (convolved with a Hanning window) for the different speaker positions 1-16. The plots are divided into four parts for clarity and readability. Only parts of impulse response corresponding to 0.3-1.85m have been considered. Color encoding for the lines can be seen in Table 4.7

are calculated in dB, as

$$\begin{aligned}\mu_{\text{dB}}(f) &= \frac{1}{\#\mathcal{N}} \sum_{n \in \mathcal{N}} 20 \log(|H_n(f)|) \\ \sigma_{\text{dB}}^2(f) &= \frac{1}{\#\mathcal{N}} \sum_{n \in \mathcal{N}} (20 \log(|H_n(f)|) - \mu_{\text{dB}}(f))^2,\end{aligned}\tag{4.4}$$

where μ_{dB} is the mean in dB, $\sigma_{\text{dB}}(f)$ the standard deviation in dB, $H_n(f)$ the room frequency response for speaker position n and \mathcal{N} the set of speaker positions included.

As seen in Table 4.9, some position estimates for speaker positions 1-16 have an absolute error above 20 cm for the wall distance estimations. These positions are disregarded when evaluating the performance in Figure 4.12. To improve the results further, the main focus should be on lowering the occurrence of these large error estimates. Because of this, only positions that have less

Color Plot	Blue	Orange	Yellow	Purple
Top left	Pos. 1	Pos. 5	Pos. 9	Pos. 13
Top right	Pos. 2	Pos. 6	Pos. 10	Pos. 14
Bottom left	Pos. 4	Pos. 8	Pos. 12	Pos. 16
Bottom right	Pos. 3	Pos. 7	Pos. 11	Pos. 15

Table 4.7: Description of color encoding for lines in the plots in Figure 4.9

Speaker position	True ($d_{min}; d_{max}$) [m]	Est. ($\hat{d}_{min}; \hat{d}_{max}$) [m]	Error ($d_{min} - \hat{d}_{min}; d_{max} - \hat{d}_{max}$)
1	(0.45; 1.6)	(0.32; 0.40)	(0.14; 1.2)
2	(0.45; 1.2)	(0.39; 0.57)	(0.06; 0.63)
3	(0.45; 0.8)	(0.38; 0.79)	(0.07; 0.01)
4	(0.4; 0.4)	(0.32; 0.50)	(0.08; -0.10)
5	(0.85; 1.6)	(0.33; 0.79)	(0.52; 0.81)
6	(0.85; 1.2)	(0.80; 1.10)	(0.05; 0.10)
7	(0.8; 0.85)	(0.71; 1.05)	(0.09; -0.20)
8	(0.4; 0.8)	(0.32; 0.77)	(0.08; 0.03)
9	(1.25; 1.6)	(1.19; 1.38)	(0.06; 0.22)
10	(1.2; 1.25)	(0.32; 1.18)	(0.87; 0.07)
11	(0.8; 1.25)	(0.71; 1.35)	(0.09; -0.10)
12	(0.4; 1.2)	(0.32; 1.16)	(0.09; 0.05)
13	(1.6; 1.65)	(1.52; 1.60)	(0.08; 0.05)
14	(1.2; 1.65)	(1.12; 1.58)	(0.08; 0.07)
15	(0.8; 1.65)	(0.33; 1.71)	(0.47; 0.94)
16	(0.4; 1.6)	(0.33; 1.58)	(0.07; 0.02)

Table 4.8: Wall distance estimations for speaker positions 1-16. Rounded to closest whole (integer) centimeter

than 20 cm in absolute error for wall distance estimations have been included in the solid lines in the plots in Figure 4.9. That is, for the solid lines, $\mathcal{N} = \{3, 4, 6, 7, 8, 9, 11, 12, 13, 14, 16\}$. For the dashed line, all speaker positions are included, which means that $\mathcal{N} = \{1, 2, \dots, 16\}$.

It is clear that the standard deviation over different speaker positions for the frequency interval [50 80] Hz is lower for the corrected RFRM. For frequencies in the interval 100-200 Hz, the standard deviation for the power (in dB) is higher for the increased for the corrected RFRM. This is due to that the tail of the slope for the Shelving filter affects the power for frequencies above the cut off frequency of 80 Hz. For frequencies above 200 Hz the RFRM:s of the corrected and uncorrected RFRM:s are almost the same. The reason for this is that the Shelving filter's frequency response's magnitude tend to 0 dB for frequencies above the cut off frequency.

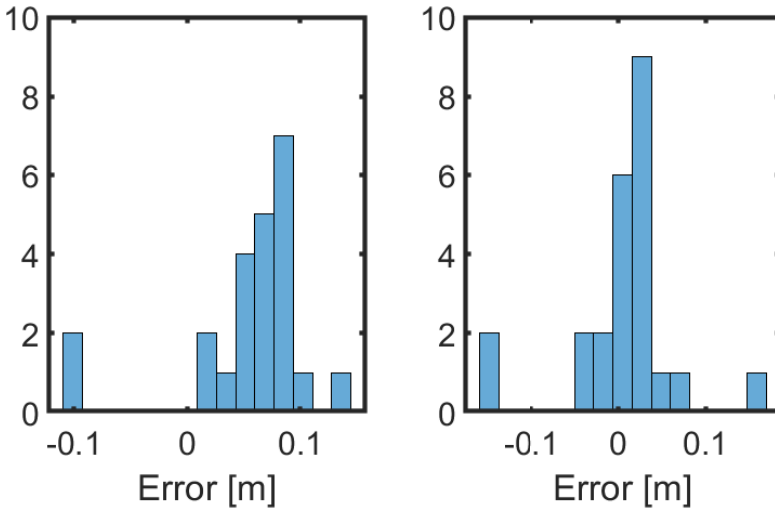


Figure 4.10: Histograms of error which have an absolute value below 20 cm, for speaker position 1-16 (values from Table 4.8). The left one is without bias correction, and the right one is with a bias correction of 5.5 cm.

When considering all the speaker positions 1-16 (dashed lines in Figure 4.12), the same effect as for the speaker positions without large errors (solid lines), which is as expected when removing badly estimated speaker positions.

4.5.2 Tests on new measurements

For evaluation, 4 measurements were done a few months after the first measurements. The speaker positions' coordinates were randomly drawn from a uniform distribution $U(0.4, 1.6)$ (in meters). These measurements were labeled speaker position 17-20. In Table 4.11, the speaker positions and the estimated wall distances are shown.

For speaker position 20, one wall is estimated to be at the distance 0.41 meters away instead of 1.53 meters. This is a very big error and also causes that \hat{d}_{\min} is wrongly an estimate of d_{\max} instead of d_{\min} . Because of the big error, the filter for the correction is not very good, for which the effect can be seen in Figure 4.13. There, the standard deviation including speaker position 20 (dashed lines) are worse (i.e. greater than) the standard deviation excluding speaker position 20 (solid lines), for some frequencies within the interval 50-80 Hz.

The estimated filter has the bass gain of G dB and the error for the estimates for speaker position 17-19 where almost within ± 1 dB, as seen in the Difference column in Table 4.12. For speaker position 20, the error is large, which is due to the poorly estimated wall distances.

The performance on the measurements for speaker positions 17-20 are similar to the measurements for speaker positions 1-16. Although, due to the low

Speaker position	True ($d_{min}; d_{max}$) [m]	Bias corrected est. ($\hat{d}_{min}; \hat{d}_{max}$) [m]	Error ($d_{min} - \hat{d}_{min}; d_{max} - \hat{d}_{max}$)
1	(0.45; 1.6)	(0.37; 0.46)	(0.06; 1.12)
2	(0.45; 1.2)	(0.45; 0.62)	(-0.02; 0.55)
3	(0.45; 0.8)	(0.43; 0.84)	(-0.01; -0.07)
4	(0.4; 0.4)	(0.37; 0.56)	(0.00; -0.18)
5	(0.85; 1.6)	(0.39; 0.84)	(0.43; 0.73)
6	(0.85; 1.2)	(0.85; 1.15)	(-0.03; 0.02)
7	(0.8; 0.85)	(0.76; 1.11)	(0.01; -0.28)
8	(0.4; 0.8)	(0.37; 0.82)	(0.00; -0.05)
9	(1.25; 1.6)	(1.24; 1.44)	(-0.02; 0.14)
10	(1.2; 1.25)	(0.38; 1.24)	(0.79; -0.01)
11	(0.8; 1.25)	(0.76; 1.40)	(0.01; -0.18)
12	(0.4; 1.2)	(0.37; 1.21)	(0.00; -0.04)
13	(1.6; 1.65)	(1.58; 1.65)	(0.00; -0.03)
14	(1.2; 1.65)	(1.17; 1.64)	(0.00; -0.01)
15	(0.8; 1.65)	(0.39; 0.76)	(0.39; 0.86)
16	(0.4; 1.6)	(0.38; 1.63)	(-0.01; -0.06)

Table 4.9: Wall distance estimations for speaker positions 1-16, with a bias correction of +5.5 cm for the estimations \hat{d}_{min} and \hat{d}_{max} . Rounded to closest whole (integer) centimeter,

amount of measurements in total, it is hard to draw any statistically supported conclusions.

4.6 Problems and limitations

In this section the problems and limitation of the methods used and results gained are discussed.

4.6.1 Position to room frequency response mapping

The frequency interval 50-80 Hz which has been focused at is rather small and is barely within the speaker's bandwidth. This most certainly affects the satisfaction of the listener. The target frequency response (Figure 1.2) might not be the best for the speaker model used (which is Behringer's 1C-BK) in respect to listener satisfaction. A better model for predicting G for a wider frequency interval would be beneficial for the results, but was not found in this thesis.

4.6.2 Speaker position estimation

For all speaker positions tested there are negative peaks values for the estimated attenuation constants, $\hat{a}^{(i)}$. Two plausible explanations for these negative values

Speaker position	Uncorrected mean bass power P_{Uncorr} [dB]	Estimated bass gain for filter G [dB]	Difference $G - P_{Uncorr}$ [dB]
1	-8.43	1.53	-6.90
2	-7.16	2.50	-4.65
3	-4.16	3.23	-0.93
4	-1.98	1.91	-0.06
5	-9.14	3.01	-6.12
6	-7.84	6.31	-1.53
7	-4.88	5.74	0.85
8	-2.43	2.87	0.44
9	-10.41	9.16	-1.25
10	-8.76	4.42	-4.34
11	-5.51	6.79	1.28
12	-3.22	4.27	1.04
13	-11.56	11.51	-0.05
14	-10.55	9.56	-0.98
15	-6.29	2.72	-3.56
16	-3.98	5.86	1.88

Table 4.10: Mean bass power (for the frequency interval 50-80 Hz), the estimated bass gain for the Shelving filter G and the difference between them, which should ideally be 0 dB.

are that they come from:

1. Reflections that correspond to negative $\hat{\alpha}^{(i)}$. This could be from either that the floors frequency dependent absorption coefficients give this effect, or that the microphones properties give this effect.
2. The DP part used in the Lasso linear regression (Equation 3.12) is too short and the negative $\hat{\alpha}^{(i)}$ has the property of minimizing the later part of the reflections.

The negative estimated constants are not generally a problem, since the reflec-

Speaker position	Coordinates (x, y) [m]	Wall est. \hat{d}_{\min} [m]	Wall est. \hat{d}_{\max} [m]
17	(0.75; 0.84)	0.72	1.12
18	(0.82; 1.16)	0.81	1.43
19	(1.20; 1.59)	1.20	1.61
20	(1.53; 0.82)	0.41	0.89

Table 4.11: The coordinates for the measurements for speaker position 17-20, together with the estimated wall distances \hat{d}_{\min} and \hat{d}_{\max} . For the shown wall distance estimates, bias correction of +5.5cm has been done.

Speaker position	Uncorrected mean bass power P_{Uncorr} [dB]	Estimated bass gain for filter G [dB]	Difference $G - P_{\text{Uncorr}}$ [dB]
17	-5.08	5.56	0.48
18	-6.77	7.13	0.36
19	-11.43	9.61	-1.82
20	-9.51	3.28	-6.23

Table 4.12: Mean bass power (for the frequency interval 50-80 Hz), the estimated bass gain for the Shelving filter G and the difference between them, which should ideally be 0 dB.

tions are found using the maximum peak (which excludes all negative peaks since they are lesser than 0). In some cases there could be problems with negative coefficients. If the negative coefficients coincidence in time with clear positive peaks and therefore decreases the height of the reflection peak. If this becomes a problem, it is possible to only consider positive $\hat{\alpha}^{(i)}$.

A problem with estimating the two closest walls is when they are at the same distance. For speaker position 10, the wall distance difference of 5 cm is enough to identify two clear peaks, and the algorithm can estimate the walls distances with a satisfying precision. For the case of speaker position 4 and 7, the reflection classified as the second strongest seems to come from the corner (the distance to the corner is 57 cm and 113 cm respectively). This makes for a greater error. Although, the error can in an application eventually be satisfying enough for a listener to perceive the room correction as an improvement of the sound quality.

For the measurements with the co-located microphone array (for speaker position estimation), log-sine-sweeps from the frequency 50 to 20000 Hz were used, even though the bandwidth of the UMA-8:s microphones only is 100-10000 Hz. This have probably resulted in poorer estimates of the RIRs. Knowledge about the bandwidth of the microphones was obtained after the measurements were done, which is the reason for the poorly motivated choice of start- and end-frequencies of the log-sine-sweeps.

4.6.3 Correction filter

Due to the slope of the Shelving filter, the correction make the sound worse for frequencies right above 80 Hz. To fix these, the Shelving filter could be modified to have a faster decreasing slope around the cut-off frequency f_c . Another fix could be to use another filter type for correction filter. If a more advanced filter was used, it could also be possible to correct standing waves in the room.

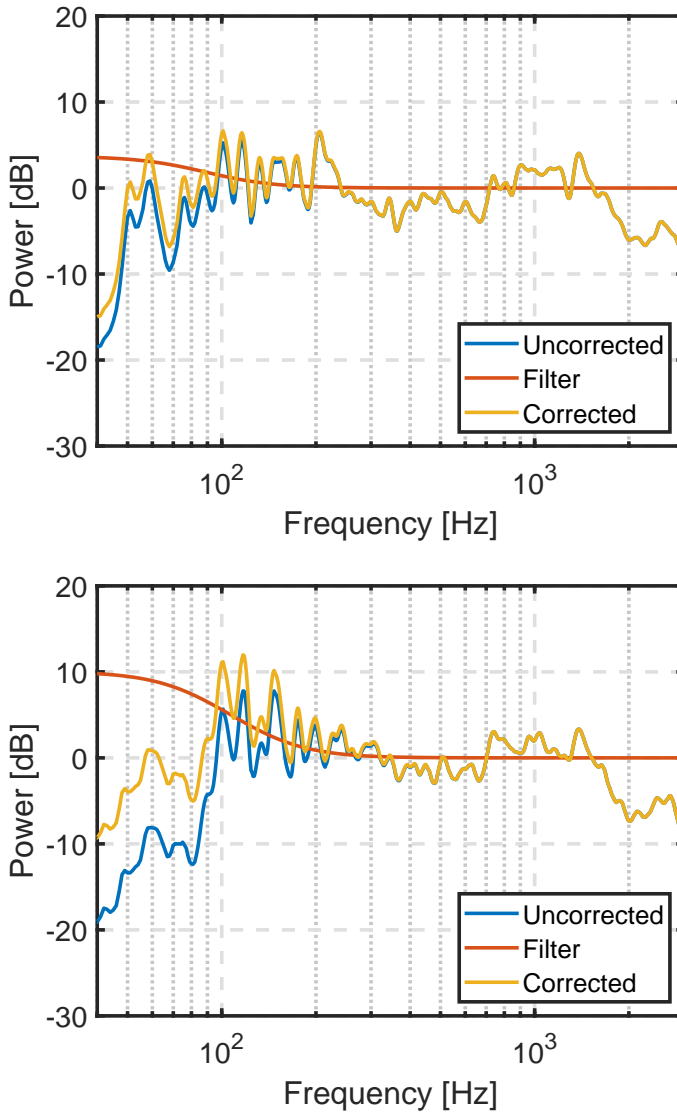


Figure 4.11: Correction of spectrum. The top plot is for speaker position 3 and the bottom is for speaker position 14. For frequencies 1000 Hz and higher, the difference between the corrected and uncorrected spectrum is not visible.

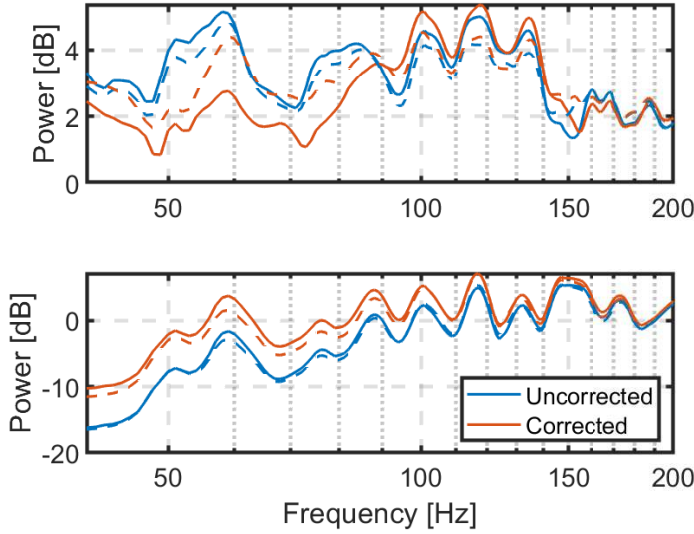


Figure 4.12: Above plot shows the standard deviation per frequency $\sigma_{\text{dB}}^2(f)$ and the bottom plot shows mean power per frequency $\mu_{\text{dB}}(f)$. Speaker positions included in the solid lines are $\mathcal{N} = \{3, 4, 6, 7, 8, 9, 11, 12, 13, 14, 16\}$ and for the dashed lines $\mathcal{N} = \{1, 2, \dots, 16\}$.

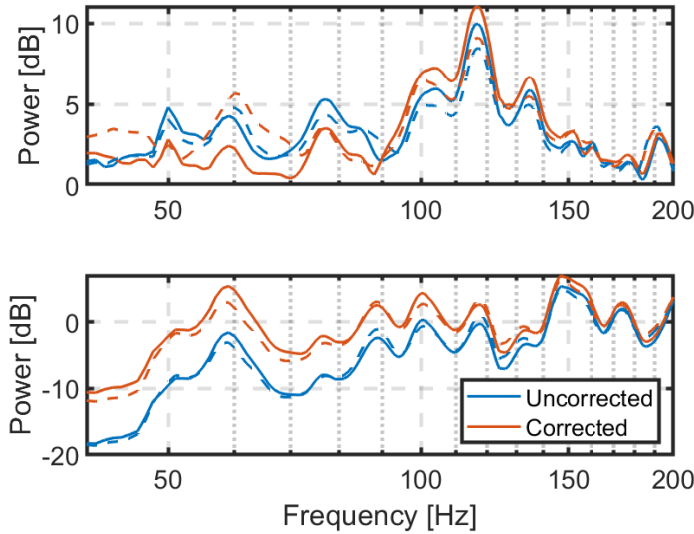


Figure 4.13: Above plot shows the standard deviation per frequency $\sigma_{\text{dB}}^2(f)$ and the bottom plot shows mean power per frequency $\mu_{\text{dB}}(f)$. Speaker positions included in the solid lines are $\mathcal{N} = \{17, 18, 19\}$ and for the dashed lines $\mathcal{N} = \{17, 18, 19, 20\}$.

5

Conclusions

In this thesis, a method to make room correction for a speaker using a co-located microphone array was searched for. The room correction should correct for the speaker's position in the room, so that the speaker sounds the same no matter of its position. The basis for the thesis is that the speaker's outputted bass increases when placed near corners. This was done and evaluated for a small conference room at Linköping University.

The approach was to firstly

1. construct a model for which correction filter should be used for each speaker position in the room, and secondly
2. develop an algorithm that estimates the speaker's position using the co-located microphones

and then combine 1 and 2 (in list above) to be able to automatically create a filter using the speaker and the co-located microphones.

For the correction filter, a bass boosting Shelving filter was used. A model for the gain parameter G for the bass constructed, which made it possible to estimate G by looking at the distance to the two closest walls. The correction filter has a cut-off frequency of 80 Hz and do focus on the output for frequencies below the cut-off frequency. The usage of a Shelving filter had some consequences. One was that the slope around the cut-off frequency f_c (where $f_c = 80$ Hz) was not very steep, and therefore the power for the frequencies right above f_c varied more with the room correction, than without. Another that there was no way to adjust certain frequency peaks coming from the standing waves in the room.

The speaker position estimation was done by looking at the wall reflections in the impulse response for each microphone in the co-located microphone array. Only a way of finding the distance to the two closest walls was found, and the algorithm is not able to tell which reflection comes from which wall. For some

speaker positions, the wall distance estimates were very poor, which considerably worsened the performance of the room correction. These very poor estimates were partly due to that some peaks in the estimated attenuation constants did not correspond to first order wall reflections, but instead to second order wall reflections.

If G is estimated by inputting the estimated wall distances from 2 (in list above), to the model in 1 (in list above), and the resulting correction filter is applied, the result is that the speaker's position is corrected for in the frequency interval 50-80 Hz. This means, if the room correction is applied, the sound from the speaker varies less with the speaker's position than if the room correction was not applied, for most speaker positions. For the speaker positions where the wall distance estimates were very poor, the result could be that the room correction increased the variety in the bass instead of decreasing it.

5.1 Further work

Improvements of the room correction can be done. Listed below are areas possible to do further studies within, for which the results could improve the room correction.

- In the current stage of the speaker position estimation, only the wall distances d_{\min} and d_{\max} can be found, and there is no further information about the room orientation and coordinates (x, y) in the estimates. Finding ways to estimate (x, y) instead of d_{\min} and d_{\max} would improve the room correction, since a better model for estimating G can then be used. The work in Appendix A could potentially be of help for developing a method to estimate (x, y) .
- The effect of the parameters f_1 and f_2 for the log-sine-sweeps in the speaker position estimation were not evaluated fully. A better SNR for the RIR estimates could possibly be achieved if having a smaller bandwidth for the log-sine-sweep, and especially making sure that the interval $[f_1 f_2]$ is within the bandwidth of the microphones, which was not the case for the UMA-8 in this thesis.
- The sampling frequency f_s used in this thesis was set to 44100 Hz. If the speed of sound is 343 m/s, a sound wave can travel about 8 mm in between two samples are recorded. This gives the wall distance estimation no better precision than about 4 mm. Increasing the sampling frequency f_s or upsampling the RIRs would allow for better precision in the distance estimates.

Appendix

A

Additional work

This appendices presents a method for estimating the direction of arrival (DOA) for a wall reflection. This method does not provide any contributions to the thesis's research aims on its own, but can be an important part of future improvements of the room correction in this thesis.

A.1 Pairwise DOA

DOA estimation is done by pairwise comparing wall distance estimations from the microphones on the microphone array UMA-8. The microphones on the UMA-8 are labeled $m = 1, \dots, 7$. Denote a pair of microphone i and j as (i, j) . Let \mathcal{P} be the set of all microphone pairs and be defined by

$$\mathcal{P} = \{(i, j) \in \mathcal{R}^2 : j > i\}. \quad (\text{A.1})$$

For wall k and microphone i , denote a wall distance estimation as $\hat{d}_{k,i}$. The wall distance estimation $\hat{d}_{k,i}$ is defined as

$$\hat{d}_{k,i} = \operatorname{argmax}_{d \in \mathcal{W}_k} \{\hat{\alpha}^{(i)}(d)\}, \quad (\text{A.2})$$

where $\hat{\alpha}^{(i)}(d)$ is the element in $\hat{\alpha}^{(i)}$ corresponding to distance from the speaker d , the set \mathcal{W}_k is $\mathcal{W}_k = \{d \in \mathcal{R} : \hat{d}_{(\text{wall } k)} - 0.15 < d < \hat{d}_{(\text{wall } k)} + 0.15\}$ for which $\hat{d}_{(\text{wall } k)}$ is the estimated wall distance for wall k as seen in Tables 4.8 and 4.11 (Note that in the tables have subindices min and max, instead of $k = 1, 2$). In other words, \mathcal{W}_k restricts that only values within ± 15 cm from the estimated wall distance are considered.

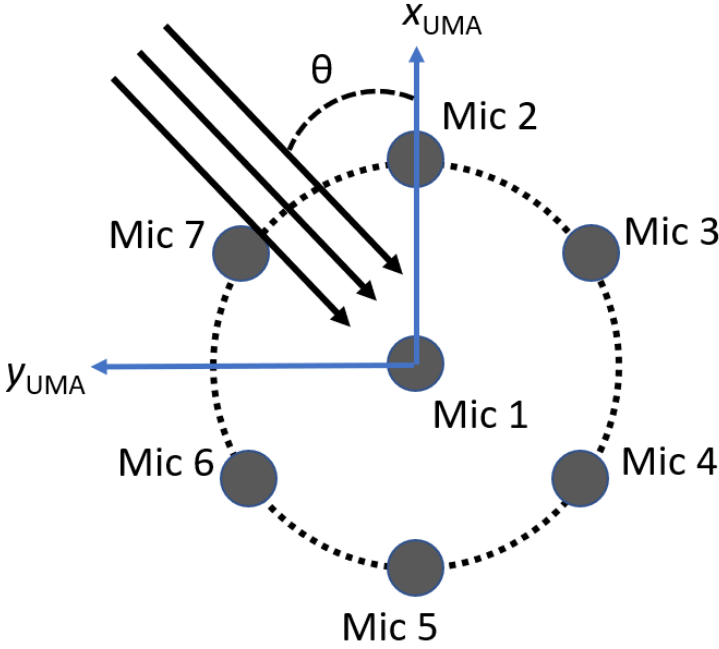


Figure A.1: The circular microphone array UMA-8 seen from above, with sound from a wall reflection coming from the angle θ , in the relative coordinate system $(x_{\text{UMA}}, y_{\text{UMA}})$. Microphone 2-7 are evenly spread around the dashed circle, and therefore the angle between two adjacent microphones is 60 degrees.

Let the angle θ be defined as positive if going from the x-axis to the y-axis (as in Figure A.1) and $r = 0.046$ meters be the radius of the UMA-8. If the wall reflection arrives from the angle θ , the geometry of the microphone array gives

$$\hat{d}_{k,i} - \hat{d}_{k,j} = r(\cos(\theta_j) - \cos(\theta_i)) + \epsilon_{k,i,j} \quad (\text{A.3})$$

for $(i, j) \in \mathcal{P}$ with $i > 1$, and

$$\hat{d}_{k,1} - \hat{d}_{k,j} = r(\cos(\theta_j)) + \epsilon_{k,1,j} \quad (\text{A.4})$$

for $(i, j) \in \mathcal{P}$ with $i = 1$. In these equations, we define angle $\theta_2 = 300^\circ - \theta$, angle $\theta_3 = 240^\circ - \theta$, angle $\theta_4 = 180^\circ - \theta$, angle $\theta_5 = 120^\circ - \theta$, angle $\theta_6 = 60^\circ - \theta$, each corresponding to the microphones on the UMA, angle θ as the DOA and $\epsilon_{k,i,j}$ as the error.

Let all possible equations from (A.4) and (A.3) for $(i, j) \in \mathcal{R}$ form an equation system, with top-to-bottom order as $(i, j) = (1, 2), (1, 3), \dots, (1, 7), (2, 3), (2, 4), \dots, (6, 7)$. This equation system will be equivalent to

$$Ly_{\text{meas}} = F(\theta) + \epsilon_k \quad (\text{A.5})$$

where ϵ_k is all $\epsilon_{k,i,j}$ on a vector for walls $k = 1, 2$, matrix L is defined by

$$L = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ & & & \vdots & & & \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ & & & \vdots & & & \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \quad (\text{A.6})$$

measurement vector y_{meas} is defined by

$$y_{\text{meas}} = [\hat{d}_{k,1} \quad \hat{d}_{k,2} \quad \hat{d}_{k,3} \quad \hat{d}_{k,4} \quad \hat{d}_{k,5} \quad \hat{d}_{k,6} \quad \hat{d}_{k,7}]^T, \quad (\text{A.7})$$

and function $F(\theta)$ defined by

$$F(\theta) = \begin{bmatrix} r \cos(\theta_2) \\ r \cos(\theta_3) \\ \vdots \\ r \cos(\theta_7) \\ r(\cos(\theta_3) - \cos(\theta_2)) \\ r(\cos(\theta_4) - \cos(\theta_2)) \\ \vdots \\ r(\cos(\theta_7) - \cos(\theta_6)) \end{bmatrix} \quad (\text{A.8})$$

Then a grid search is performed for θ with a resolution of 1 degree, to find θ that gives the lowest cost $V^{\text{NLS}} = \|\epsilon_k\|_2^2$, which is the DOA estimate. [6]

A.2 Implementation and results

The DOA estimation was tested for speaker positions 1-20. The design parameter value D_{max} was increased to $D_{\text{max}} = 20$, in comparison to earlier parts of the thesis. This was due to $D_{\text{max}} = 10$ gave too few non-zero values for $\hat{\alpha}^{(i)}$ to be able to do DOA for two walls.

The resulting DOA estimations can be seen in Table A.1. The true value for the DOAs should be approximately 30° or 120° . Although, during the measurements, the accuracy of placing the speaker correctly rotated was not very high. That means that some difference from 30° or 120° may occur due to how the measurements were done.

The histogram in Figure A.2 shows that most DOA estimations are around 30° and 120° . This shows that the DOA approach might be a part future work for better room geometry estimation, but probably needs improved accuracy to be useful.

Speaker position	Wall est. \hat{d}_1 [m]	DOA $\hat{\theta}_1$ [°]	Wall est. \hat{d}_2 [m]	DOA $\hat{\theta}_2$ [°]
1	0.46	129	0.37	129
2	0.44	140	1.15	0
3	0.44	127	0.84	61
4	0.37	150	0.56	83
5	0.85	128	0.38	56
6	0.86	133	1.16	30
7	0.76	36	1.10	51
8	0.37	46	0.83	134
9	1.24	298	1.44	124
10	1.24	14	0.39	123
11	0.77	37	1.40	72
12	0.37	34	1.21	85
13	1.67	275	1.58	275
14	1.17	15	1.67	342
15	0.77	34	0.39	58
16	0.38	19	1.69	0
17	0.72	35	1.12	64
18	0.81	35	1.42	89
19	0.37	195	1.61	100
20	0.89	122	0.40	300

Table A.1: Estimated DOA and wall distances for speaker position 1-20.

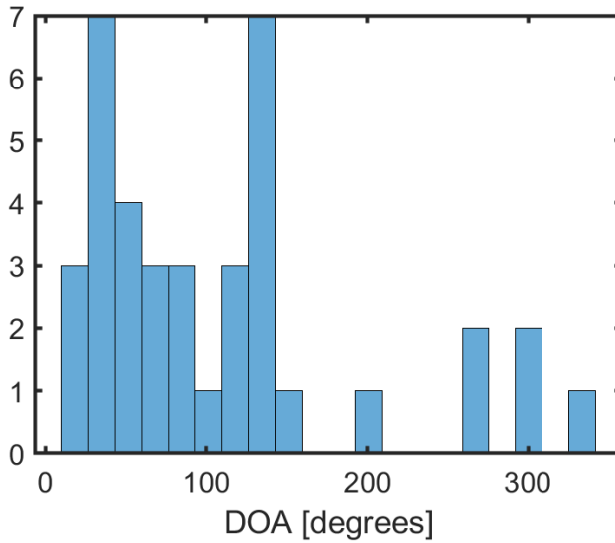
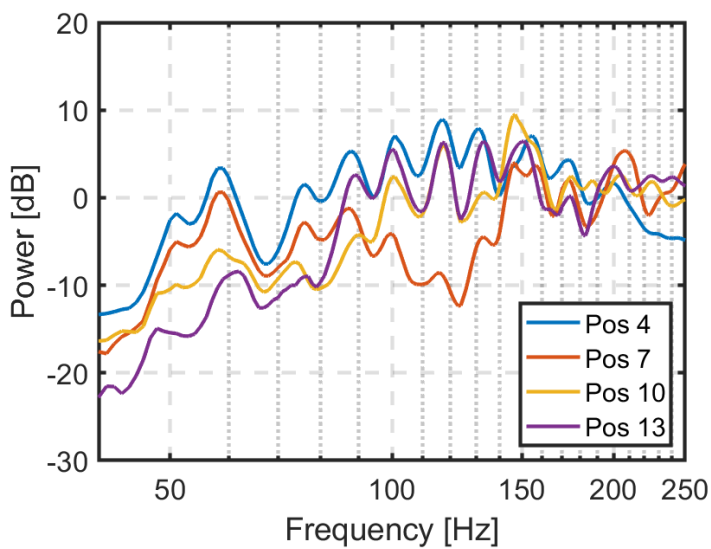


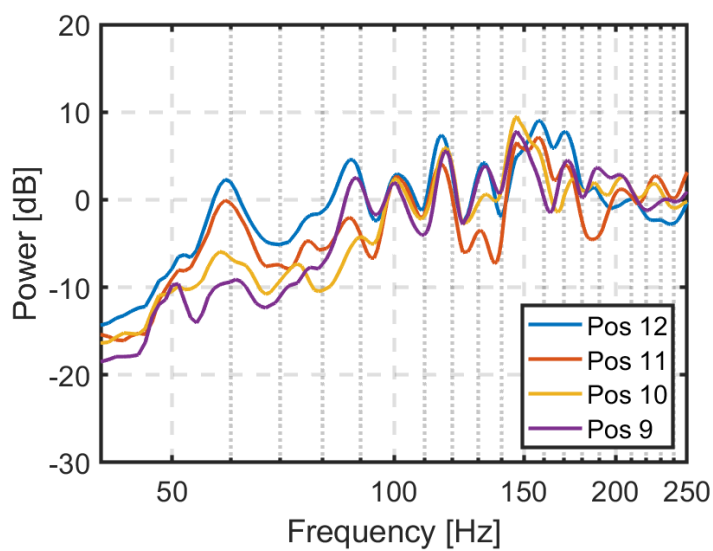
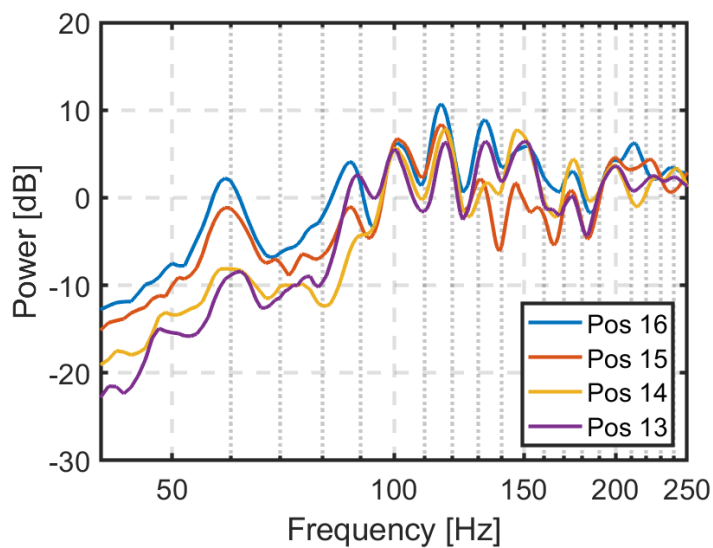
Figure A.2: DOA estimations for each wall for each speaker position 1-12.

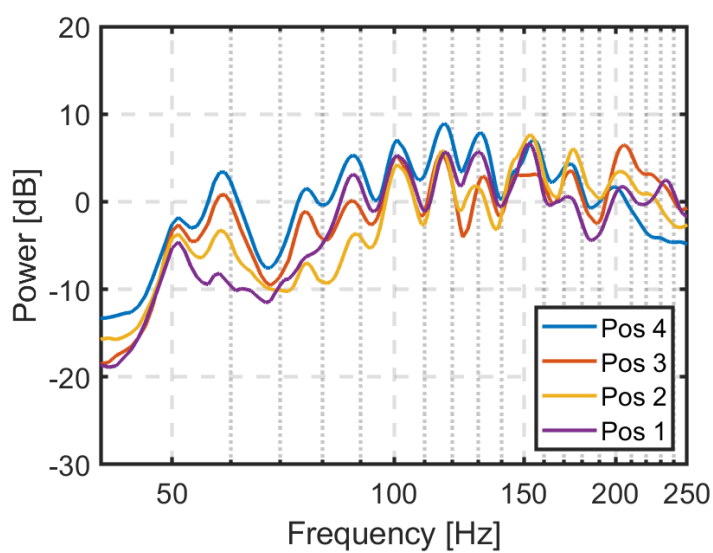
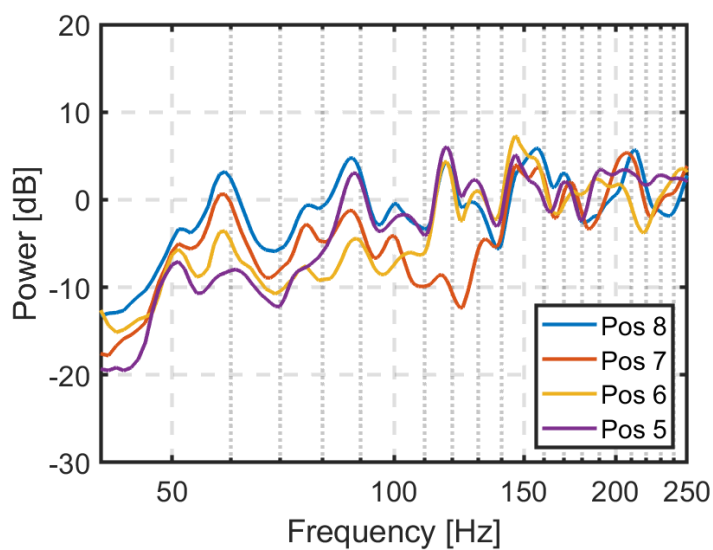
B

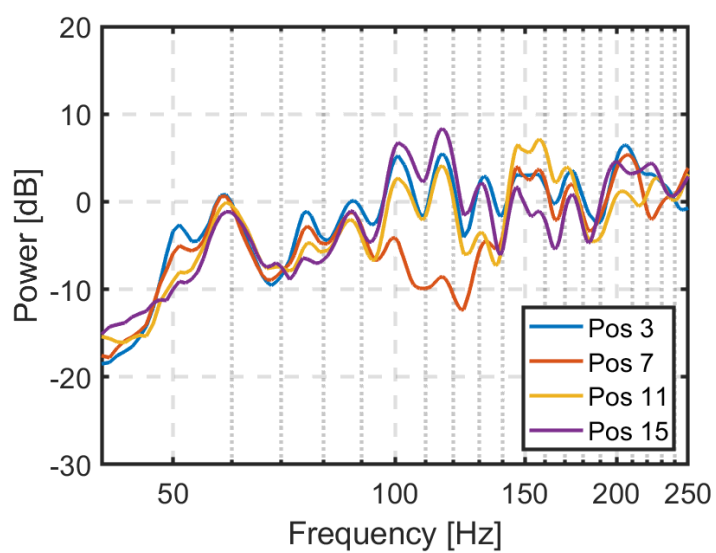
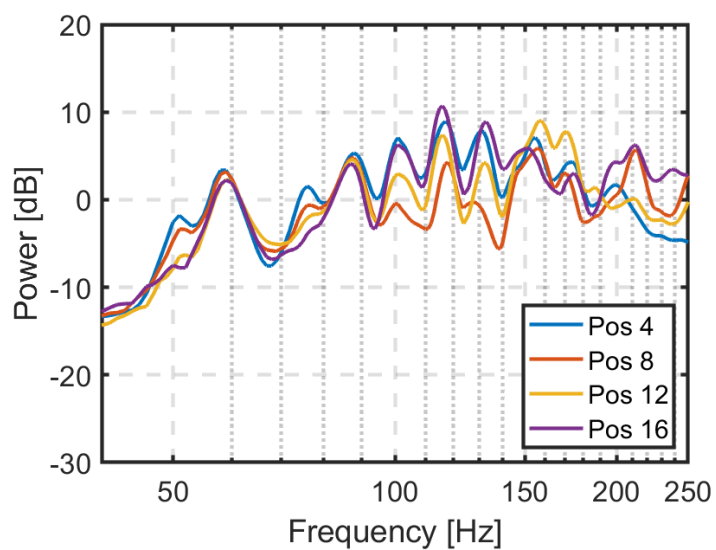
Room frequency responses

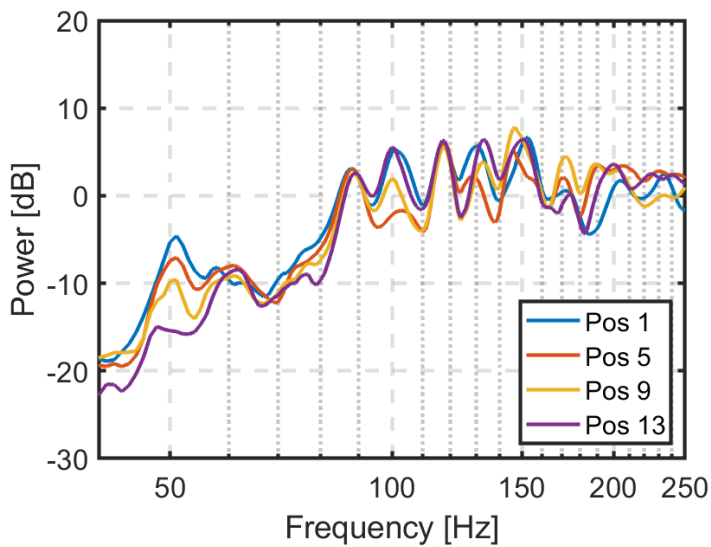
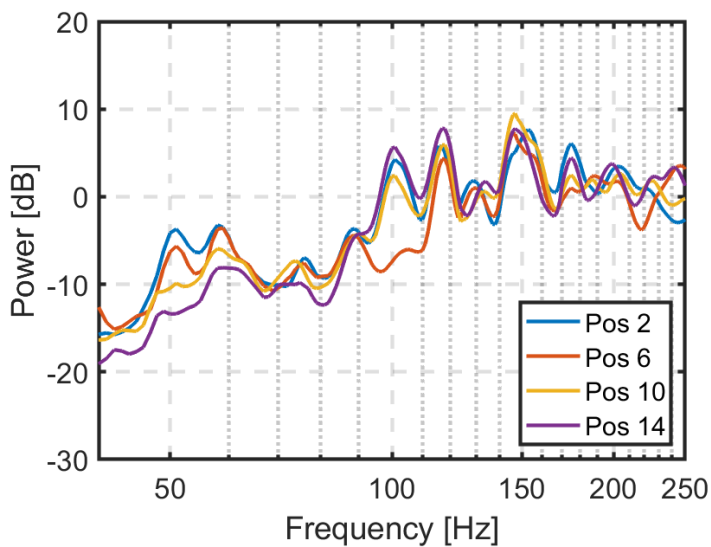
Room frequency responses from Visionen.











C

Frequency responses for microphones



UMIK-1

Template calibration file loaded into Room EQ Wizard

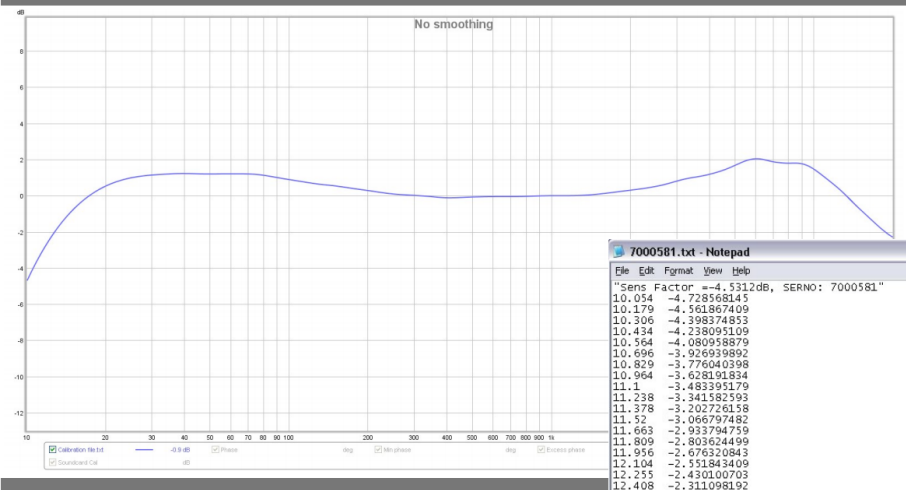


Figure C.1: Frequency responses magnitude for UMIK-1. [10]

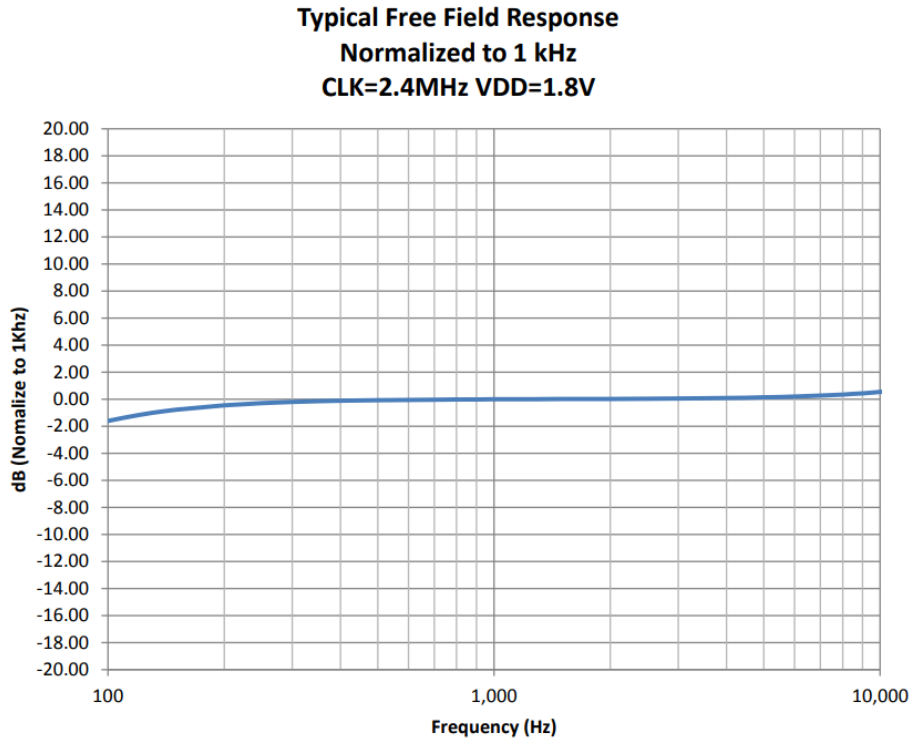


Figure C.2: Frequency responses magnitude for each microphone on the UMA-8. [8]

Bibliography

- [1] Malcolm J. Crocker. *Handbook of noise and vibration control*. J. Wiley, 2007. ISBN 9780471395997.
- [2] Cha Zhang Dinei Florêncio Demba Ba, Flávio Ribeiro. Geometrically constrained room modeling with compact microphone arrays.
- [3] F. Alton Everest. *Master Handbook of Acoustics*. McGraw-Hill, fourth edition, 2001.
- [4] Angelo Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. *108th AES Convention*, 2000. URL <http://pcfarina.eng.unipr.it/Public/Papers/134-AES00.PDF>.
- [5] Louis D. Fielder and Eric M. Benjamin. Subwoofer performance for accurate reproduction of music. In *Audio Engineering Society Convention 83*, Oct 1987. URL <http://www.aes.org/e-lib/browse.cfm?elib=4865>.
- [6] Fredrik Gustafsson. *Statistical sensor fusion*. Studentlitteratur, 2018. ISBN 9789144127248.
- [7] Steven M. Kay. *Fundamentals of statistical signal processing*. Prentice-Hall signal processing series. Prentice Hall PTR, 1993. ISBN 0133457117.
- [8] *Digital Zero-Height SiSonic Microphone*. Knowles Electronics, 4 2015. Rev. A.
- [9] Etienne Corteel Brian Katz Marc Rebillat, Romain Hennequin. Identification of cascade of Hammerstein models for the description of nonlinearities in vibrating device. *Journal of Sound and Vibration*, pages 1018–1038, 2010.
- [10] *Umik-1 Product brief*. miniDSP Ltd, 2019.
- [11] Saban Ozer, Hasan Zorlu, and Selcuk Mete. System identification application using Hammerstein model. *Sadhan*, 41(6):597–605, Jun 2016. ISSN 0973-7677. doi: 10.1007/s12046-016-0505-8. URL <https://doi.org/10.1007/s12046-016-0505-8>.

- [12] Jan Abildgaard Pedersen. Adaptive Bass Control - the ABC room adaption system. *AES 23rd International Conference*, 2003.
- [13] Chuang Shi and Yoshinobu Kajikawa. Volterra model of the parametric array loudspeaker operating at ultrasonic frequencies. *The Journal of the Acoustical Society of America*, 140(5):3643–3650, 2016. doi: 10.1121/1.4966962. URL <https://doi.org/10.1121/1.4966962>.
- [14] Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL <http://www.jstor.org/stable/2346178>.
- [15] T. Wang, F. Peng, and B. Chen. First order echo based room shape recovery using a single mobile device. pages 21–25, March 2016. ISSN 2379-190X. doi: 10.1109/ICASSP.2016.7471629.
- [16] Richard V. Waterhouse. Output of a sound source in a reverberation chamber and other reflecting environments. *The Journal of the Acoustical Society of America*, 30(1):4–13, 1958. doi: 10.1121/1.1909380. URL <https://doi.org/10.1121/1.1909380>.
- [17] Adrian Wills, Thomas B. Schön, Lennart Ljung, and Brett Ninness. Identification of Hammerstein–Wiener models. *Automatica*, 49(1):70 – 81, 2013. ISSN 0005-1098. doi: <https://doi.org/10.1016/j.automatica.2012.09.018>. URL <http://www.sciencedirect.com/science/article/pii/S0005109812004815>.
- [18] Udo Zölzer. *DAFX : digital audio effects, second edition*. John Wiley and Sons Ltd, 2011. ISBN 9780470665992.