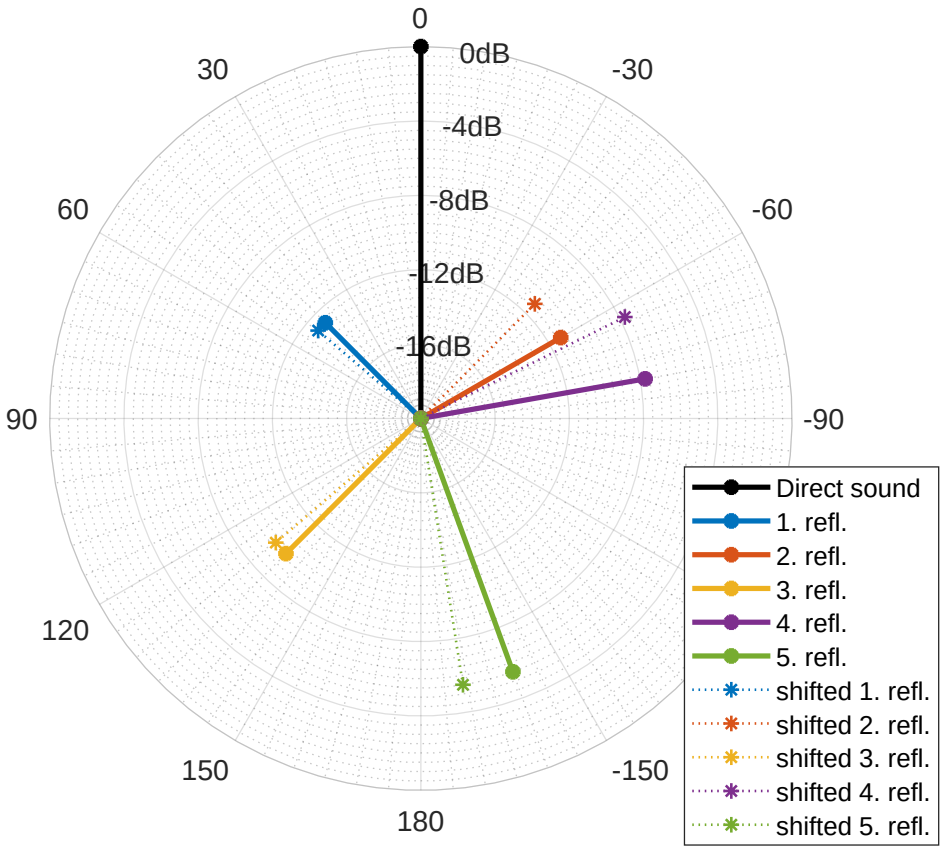

Perceptual Detection Thresholds for Alterations of the Azimuth of Early Room Reflections

Master's thesis in Master Program Sound and Vibration

Felicitas Bederna

DEPARTMENT OF CIVIL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2022
www.chalmers.se

Master's thesis 2022

Perceptual Detection Thresholds for Alterations of the Azimuth of Early Room Reflections

FELICITAS BEDERNA, 2022

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
Gothenburg, Sweden 2022

Perceptual Detection Thresholds for Alterations of the Azimuth of Early Room Reflections
FELICITAS BEDERNA

Supervisor and Examiner: Jens Ahrens, Division of Applied Acoustics

Master's Thesis 2022
Department of Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Incidence angles and relative levels of the direct sound and first reflections or a room impulse response together with shifted versions of the first reflections.

Typeset in LaTeX
Gothenburg, Sweden 2022

Perceptual Detection Thresholds for Alterations of the Azimuth of Early Room Reflections

FELICITAS BEDERNA
Department of Civil Engineering
Chalmers University of Technology

Abstract

Spatial audio, that is, the reproduction of an acoustic event including all spatial information, has become a major research topic in recent years. The computer games industry, for example, is looking for ways to reproduce the created spaces as realistically as possible, not only visually but also acoustically. A particularly large area of interest in spatial audio is the authentic reproduction of a room. The question arises how far the reproduced angles of a reflection can deviate from the actual angle without the room being perceived differently. This has already been investigated for single reflections, but how and when several shifted reflections together cause a perception change is still largely unknown.
Therefore, in this work it is investigated how far the azimuth angles of the first five reflections of a room impulse response have to be shifted, so that the original room impulse response can just be distinguished from the manipulated one. For this purpose, 20 subjects performed a headphone-based adaptive ABX threshold test in which the selected reflections, initially all coming from one side, were shifted closer to or further away from their original positions at each adaptation step until the individual point of just noticeable difference was found. For this, the number of steps for moving each reflection from the adaptation start to the original position was the same, but the step size differed due to the individual angle differences. The recordings of a drum and a male speaker were used in the test. The results show that the five selected reflections for both test signals can be shifted between 4.33° and 17.13°, depending on the step size of the reflection, until a difference is audible. The perceived differences between the rooms were described as changes of the center of gravity in the room or the localization of the sound source. No major difference between the median values of the drum and the speech signal could be found. The inexperienced listening test participants were mainly responsible for all outliers found in the threshold tests while the median thresholds were slightly higher, suggesting that the reproduced angles of the reflections could potentially deviate more for the broad society. An additional ABX test comparing the original room with a room where the first reflections were mirrored showed a clear perceptual difference.

Keywords: Psychoacoustics, Spatial Audio, Room Acoustics, Spatial Impulse Response, Lateral Reflections, Spatial Decomposition Method, Transformed Up-Down Method

Contents

1 Introduction **1**

2 Theory **3**
2.1 Fundamentals of spatial hearing . . . 3
2.1.1 Auditory space . . . 3
2.1.2 Sound source localization . . . 4
2.1.2.1 Localization of one sound source . . . 4
2.1.2.2 Localization and perception of multiple sound sources 5
2.1.3 Perception of reflections . . . 7
2.2 Spatial and binaural room impulse responses . . . 8
2.3 Spatial decomposition method . . . 9
2.3.1 DOA estimation . . . 10
2.3.2 DOA postprocessing . . . 11
2.3.3 Reverberation equalization . . . 11
2.4 Listening tests . . . 12

3 Method **17**
3.1 Creation of the stimuli . . . 18
3.1.1 Used impulse response and DOA estimation . . . 18
3.1.2 Alteration of the first reflections . . . 20
3.1.3 Real-time convolution of BRIRs and audio examples . . . 21
3.2 Experiment setup . . . 23
3.2.1 Experiment procedure . . . 23
3.2.2 User Interface . . . 25
3.2.3 Participants . . . 26
3.3 Analysis of the listening tests . . . 26
3.3.1 Threshold test . . . 27
3.3.2 Discrimination test . . . 27

4 Results **31**
4.1 Threshold test . . . 31
4.2 Discrimination test . . . 34

5 Discussion **35**
5.1 Performance of the participants . . . 35
5.2 Influence of the audio signals . . . 37
5.3 Analysis of the found thresholds . . . 40

5.4 Analysis of the discrimination task 44
5.5 Relevance and limitations of the experiments 44
5.6 Effects of the design 46

6 Conclusion **49**

Bibliography **51**

1
Introduction

In recent years surround sound systems became more and more popular. After such systems were initially rather unusual and perhaps part of a good cinema, they are now even to be found at home both as a speaker system and through headphones. Surround sound systems are particularly popular in the computer games industry, as they advertise the ability to locate opponents at an early stage and to experience the created landscapes not only visually but also acoustically very close to reality. For this growing market it is an ongoing challenge to reproduce different spaces in a plausible way, no matter whether this space is real or virtual. Especially the processing of rooms is challenging, because not only the direct sound and possibly a few reflections are important for the perception, but a multitude of reflections together with reverberation. So in order to fully reproduce a room the direct sound as well as all reflections in addition to the reverberation need to be rendered and thus not only the pressure room impulse response at the receiver position but also all the incidence angles of the reflections need to be known. For virtual spaces this information needs to be simulated while the reproduction of real rooms can be based on measurements or simulation or a combination of both. For a measurement of a room impulse response with spatial information an array of microphones is needed. With all the recordings of the microphones the incidence angles for each sample can be calculated. Generally, the more microphones the array has the higher the spatial resolution of the recording. This is, however, time and cost intensive. In addition to that, the measured room impulse response only represents the acoustic sound field at the measurement position and for one source position. Using the example of a conversation between two people in a room where one is the listener (receiver position) and one the talker (source), this means that the reproduced sound field for the listener is no longer realistic if one of the two people moves even one step to the side. So, theoretically, an infinite number of measurements would have to be made in a room in order to reproduce it authentically for every possible combination of position. Obviously, this is not possible due to time and cost constraints. Therefore, the question arises how precise the spatial information of a spatial impulse response has to be, so that the space is perceived realistically?

To answer this question the auditory system of humans, or rather the sensitivity of the human auditory system to angle changes, needs to be examined. The sensitivity to angle changes of single sound sources, so the amount of angle change of a signal that is needed to hear a difference has been examined for different incident angles [1]. There have also been several examinations on influences of the delay times, incidence angles and levels of single or multiple reflections on the overall

perception of an auditory event, both in anechoic (for example [1], [2]) as well as in real environments (for example [2]–[5]). For the acoustic reproduction of rooms, algorithms have been developed, in the process of which it was investigated how many reflections of a room impulse response need to be rendered in order to plausibly reproduce that room [6]. Another reproduction of a room with synthetic spatial data, so the rendering of the reflections with incidence angles that were randomly generated, even showed no significant perceived difference between the original and the adjusted room [7]. But how much can the incidence angles of early reflections differ from the original ones, without hearing any difference? This question has, to the author's best knowledge, been examined in two works so far, in [4] for a single lateral reflection and in [8] for a single ceiling reflection. However, how perception changes when not only one reflection deviates from the original angle, but several at the same time, has not yet been investigated. This work therefore aims to give insight into the perceptual effect of shifting the azimuth angles of several early reflections. For this insight the maximum azimuth shift of several reflections is searched, which just leads to a change of perception in listening test participants.

This thesis is organized as followed: Firstly, an introduction in the fundamentals of spatial hearing is given as well as an overview over the recording and reproduction of spatial room impulse responses. This is followed by the design and method of analysis of the implemented listening test. Lastly, the obtained thresholds are analyzed, discussed and compared with related work.

2

Theory

2.1 Fundamentals of spatial hearing

2.1.1 Auditory space

A source, or the direction of incidence of sound, is usually described relative to the listener with a so-called head-related coordinate system with the listener in the center (see figure 2.1). A source location is then described with an elevation and azimuth angle and a radius r. The angles for this coordinate system can be assigned in different ways, in this work the frontal direction has the coordinates 0° azimuth and 0° elevation, the back has an azimuth angle of 180° and the same elevation angle of 0°. The position on top of the listener has an elevation angle of 90°, while the position directly below the listener has -90° elevation (both azimuth angles are 0°). The positions left and right to the listener have the azimuth angles of 90° and -90°, respectively.

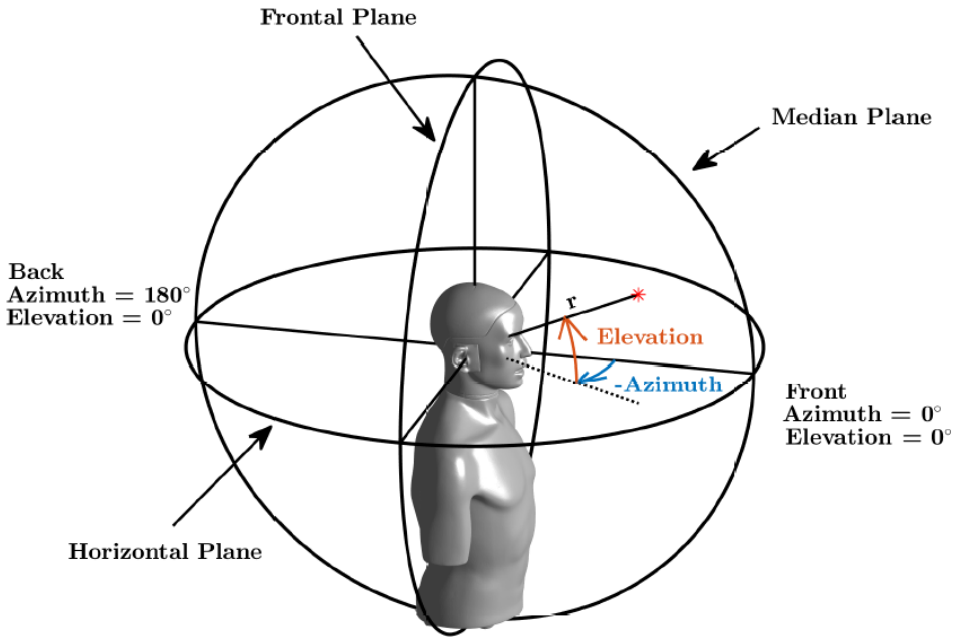

Figure 2.1: Coordinate system to describe a source position by its azimuth and elevation angle and distance r relative to the listener's head, taken from [9].

2.1.2 Sound source localization

Humans localization of sound sources are based on auditory cues which describe characteristics of the sound waves arriving at our ears. These auditory cues can be separated into two subcategories, the monaural cues that are the same for both ears and the binaural cues which describe differences between the two ear signals ([1], p.11). These cues arise from the fact that our body, as any other object as well, has an influence on the sound field which is depending on the frequency range. If a sound comes, for example, from the right side, the sound will arrive at the right ear first and has to bend around the head until it arrives at the second ear. The time difference between the two ears can then be used to determine that the sound came from the right side. Another binaural cue that can be used to localize the sound in this example are the level differences between the ears. Especially higher frequencies with a short wavelength will lose a lot of energy while traveling from the right to the left ear and the resulting level differences indicate the direction of arrival of the corresponding sound event. If the sound source is located on the median plane, however, these binaural cues can not be used since the signal arrives at both ears at the same time and, assuming a symmetrical shape of the head and ears, the signals will be exactly the same for both ears. In this case the monaural cues are of great importance. Here, the specific shapes of our torso, head and ears color the incoming sound spectrally. This spectral coloration is different if the sound comes from the front or the back, so that the spectral coloration can be analyzed to get information about the source location.

2.1.2.1 Localization of one sound source

Lot of studies have been done in the past to study the ability of the human auditory system to localize a single sound source and the ability to detect small local changes of a single source. A word used in this context is the one "Localization blur'' which characterizes the smallest change of an attribute that leads to a change of the auditory event in 50% of the participants ([1], pp.37) . Here it is important to distinguish between the term of the sound source and the auditory event. The sound source is the actual physical sound source while the auditory event is the image that the sound source produces in our auditory system. Localization blur in the azimuth angle, for example, describes how much the angle of a source can be shifted in the horizontal plane so that it evokes a difference of the auditory event. The auditory event and the sound source can have the same or similar attributes, for example a physical and perceived azimuth angle change of 10°, but the auditory system generally has a smaller spatial resolution than it can be achieved by physical measuring techniques ([1], p.38).
The region of the most precise spatial hearing in the auditory system is close the the forward direction in the horizontal plane. The resolution here is around 1° but strongly depends on the test signal. Values measured with speech signal, for example, are smaller than corresponding values measured using broadband noise. The localization blur for sound incidence from the sides is 3 up to 10 times bigger than for the frontal direction ([1], p.40). The spatial resolution in the elevation angle is much coarser than for the azimuth angle. The localization blur for broadband

signals is around 4-17○ for the forward direction, while we're not able to detect any angle changes in the median plane at all when using narrow band signals ([1], p.44). This is due to the missing broadband components in the signal that are needed for the monaural cues since those detect spectral differences between the ear signals, while binaural cues are not present in case of sound sources placed on the median plane. The same spatial resolution is even observed for people with small to moderate symmetrical hearing loss and age-related hearing loss, so that the hearing threshold and the age are not directly related to the spatial resolution of the auditory system ([1], p.49). Asymmetrical hearing loss on the other side has a negative influence on the localization ability and localization blur. It is even possible to have some localization abilities with deafness in one ear. Obviously, no binaural cues can be used in this case to process potential difference between the two ear signals, but monaural cues are used to preserve spatial information.
Generally, it was found that monaural signal attributes help the auditory system to cover the front and the rear sections of the median plane, elevation angles in general as well as a feeling of distance of a sound source. Binaural cues generally help with the perception of angles and angle changes in horizontal direction ([1], p.177). Of course, when localizing sounds, we don't have to stand still and rely on the corresponding cues to give us a rough idea where the source might be. Instead, after the first rough localization we can move our heads and thus bring the source into the region of sharpest local resolution to decrease the localization blur ([1], p.179).

2.1.2.2 Localization and perception of multiple sound sources

This chapter is about the localization of perception of multiple sound sources. This can mean additional sources as well as reflections of a single source. The localization of two or more sound sources is significantly different to the one with only one sound source since the sound fields produced by the sources influence each other depending on the delay between them, their levels and their frequency components. The perception of the resulting auditory event also depends on these factors. For this chapter we assume two rather coherent signals, so a first direct signal, followed by a delayed version of more or less the exact same signal. Since this work is about the perception of spatial impulse responses, this is an appropriate assumption, since the additional sources are, due to the reflection slightly changed, versions of the original signal.
In a case where the delay between two signals at the same level is very small below around 1ms, a single auditory event is localized at a position that is dependent on the locations of both sound sources. This situation creates a so-called "phantom source" and the effect behind this is called summing localization ([1], p.204). This effect also occurs when there is no delay but a level difference of up to 30 dB between the signals. As an example, when there is a first sound source at 45° and a second one at 10° degrees playing at the same time, but the second source is 20 dB more silent than the first one. This scenario would create a phantom source probably at around 40°, since the first sound source is much louder than the second one and therefore the "leading" sound source which is, however, slightly influenced by the second source. Thus the auditory event lies between the two sources but closer to

the first one.
In cases of coherent signals with a delay greater than around 1ms or more than 30 dB difference in level, one auditory event is heard whose localization is dependent only by the first wave front or the louder signal. This is called the "law of the first wave front'' ([1], p.204). The upper limit of this effect is impossible to tell for any kind of signal since it strongly depends on the type of signal, the direction of the sound incidence and the level of the two signals. In this scenario, the second wave front can still has an effect on the auditory event, even if it is not perceived as a second auditory event. It can for example influence the spatial extend or the tone color of the auditory event. An explanation for this coloration effect is the comb effect that results from the interference pattern between the direct sound and its reflection. This suggests that tone coloration is a monaural effect since it also occurs with ceiling reflections and is less noticeable when direct and reflected sound are laterally separated [3]. Besides tone coloration and a spatial extension, the reflection can also influence the "center of gravity" which means that the auditory event somewhat shifts to the direction of the second sound source but without a change of the localization of the fist sound source ([1], p.224). This could be interpreted as a part of a change of the spatial extend with a weighting in a specific direction.
Reflections with bigger delay after the direct sound are perceived as two separate auditory events, one for each source ([1], p.204).
The transition between these different perceptions of summing localization, law of the first wave front, and of two separate sources is smooth. As already mentioned, it depends on the delay between the direct component and the reflection, as well as on the corresponding levels. A factor that is often used for the evaluation of a reflection is the spatial impression of the auditory event. In [3] the authors conducted a study with a direct sound and one reflection at 40° with a delay of 40ms. They tested how high the level of the reflection had to be for different delay times to achieve the same spatial impression as the reference reflection with its time and level. It could be seen that the level of the reflection needed to be higher for longer delays in order to produce the same spatial impression. For delays below the approximately 20ms it was seen that the level decreased drastically with shorter delay times, so the reflection could be much more silent than the first reflection in order to produce the same spatial reflection. Below 5ms no data is given, probably the threshold for summing localization was found at this position. The authors also conducted a similar study with a changing angle of the reflection at a constant level. The biggest spatial impression was found at an angle of around 40°. The spatial impression increased again for angles towards 90° and 0°. This experiment, however, had only three participants and the participants gave the feedback that the difference was hard to detect since not only the spatial impression but also the tone color changed. Nevertheless, the results suggest that the perception of a room (here a combination of spatial impression and tone coloration) can significantly change when the incidence angles of the first reflections change and if the relative levels of these reflections are above the threshold detected in their first experiment. To examine the effect of multiple reflections compared to a single reflection, the authors in [3] also conducted an experiment with two side reflections and compared the perception to the reference of only one side reflection. It was found that the side reflections seemed

to add up incoherently because the ratio of the direct to the reflected sound needed for the same spatial impression was the same for both scenarios. This means that the two side reflections had to be more silent than the single side reflection but the total level of the incoherently added side reflections was the same as the one for the single reflection. It did not matter whether one reflection was stronger than the other or both had the same level. The result of this experiment suggest that the spatial impression is dependent on the ratio of level between the direct sound and all the added reflection in the according time frame where the law of the first wave front is valid. However, this statement has only a very limited applicability, since only very limited conditions were investigated. For example, the relationship between the incidence angles of the two reflections and the incidence angle of the reference reflection was not investigated, but according to the findings of the previous experiment, it has a strong influence on the spatial impression.
Even though the spatial impression is often used for the evaluation of rooms and depends on the first reflections, it is not the only perception caused by the first reflections. As already mentioned, the tone color of the auditory event or the "center of gravity" of the event may also change. Thus, it could be that the spatial impression of two or more signals is the same, but they can still be distinguished and there are clear differences in quality. The following chapter is therefore about the more realistic perception of reflections in room impulse responses.

2.1.3 Perception of reflections

If one wants to investigate the influence of the first reflections in real conditions, for example in rooms, it is a completely different situation than in the previous chapters, because there is not only the direct sound and possibly an additional reflection, but there are several reflections in addition to reverberation. In principle all effects covered before are possible to occur in rooms as well, but the perception and importance of one single reflection or a combination of some reflections in such a construct is much harder to predict since it highly dependent on the specific room characteristics. A reflection with the same level, angle and delay time can have a completely different effect in two different rooms. It was also found that one particular reflection is less likely to be audible when there are other reflections between the direct sound and that reflection ([1], p.274). Therefore, it is difficult to impossible to make general statements about the influence of individual and the combined reflections. However, some studies have been conducted that analyze the perceptual differences in changes of individual reflections using one or more RIRs.
The effect of a level and angle change of single reflections in a RIR, for example, was examined in [4]. They used impulse responses from two concert halls and conducted three experiments via loudspeaker representation. All experiments were threshold tests, so the goal was to determine the level (first two experiments) and the angle (third experiment) of the reflections that lead to a change of the auditory event. The first two experiments determined the threshold level of the first (first experiment) and the first and second (second experiment) reflections at 40ms and 60ms in one of the RIRs. It could be seen that the level of the original reflection had to be increased by 3.6 dB in order to be audible which results in a level difference from the direct

sound to that reflection of -2.11 dB. This means that this specific reflection could be removed from the impulse response without being noticed. In the second experiment the level of the two reflections had to be increased even further, resulting in levels slightly above the one of the direct sound, to be audible. These findings are especially interesting regarding the reported perceptual differences. While an image shift was found for the first experiment with a single side reflection at approximately 75°, the perceptual difference for the additional reflection at around -55° was the spatial impression of the auditory event. Here, it is clearly visible how the angles and levels of the reflection can cause different effects when examined alone or on combination with others. That no image shift was found in the second experiment could have been cause by the angles of the two reflections. With 75° and -55°, they are nearly mirrored on the median plane and therefore have an opposite effect on localization and cancel each other out, so to speak. Instead, the increased level of both of them lead to an auditory event that was perceived as more spacious. If the two reflections would have been on one side, they would have been likely to cause an image shift instead and thus resulting in a lower threshold. The goal of the third experiment was to determine the minimum audible angle shift of a single side reflection at 40ms and an original angle of around 70°. It was found that the reflection can be between 12° and 35° closer to the frontal direction while still producing the same impression.

2.2 Spatial and binaural room impulse responses

The behavior of a linear and time-invariant system can be described by an impulse response (IR). If this system is a room, the transfer path from a defined source position to a specific receiver position in that room is called the room impulse response (RIR). This RIR includes the direct path from the source to the receiver as well as all reflections from walls or other objects in that room. A measurement of a RIR can be done in many different ways. If a single omnidirectional microphone is used as a receiver, the reflections from all directional are summed up since the microphone has the same sensitivity in all directions. The resulting signal therefore obtains no spatial information but only the pressure at the receiver position and is thus called pressure room impulse response (PRIR). However, if the PRIR at several non planar receiver positions is measured, it is possible to deviate the direction of arrival (DOA) of different parts of the impulse response. Impulse responses with this additional spatial information are called spatial room impulse responses (SRIR) ([5] pp.231). For the measurement of SRIR, at least four non planar microphones are needed to determine the three-dimensional information of incidence for the sound. Usually, a microphone array with special dimensions is used for this purpose, which facilitates the calculation of the DOAs. If a greater spatial accuracy is needed, microphone arrays with more than four microphone positions can be used. There are several methods of calculating the DOAs of such a set of PRIRs obtained by a microphone array. The method used for this thesis is the so called spatial decomposition method which will be described in section 2.3.

In order to auralize a RIR with headphones, not only the spatial and pressure room impulse response, but also the influence of the listener in the sound field has to be

considered. This effect can be captured by a set of filters that describe the influence of our body on the incoming sound from a specified direction. This transfer function from incoming sound at a precise direction to the ear canal is called Head-related-transfer-function (HRTF) ([1], pp.42). The equivalent in the time domain is the Head-related-impulse-response (HRIR). To obtain a set of these HRTFs, multiple measurements have to be carried out with sound sources from all possible directions, so covering all points in a grid representing the head-related coordinate system. To also include the influence of the shape of the ear, very small microphones have to be positioned inside the ear canal. Since this measurement is very time consuming, HRTFs are usually not person specific but measured with a dummy head that represents mean torso and head dimensions of adults. This means, however, that a subjects actual anatomy differs more or less from that of a dummy head and therefore both, the monaural and the binaural cues can be inaccurate for that person and thus worsen the localization ability with HRTFs. The HRTFs for each DOA in the SRIR can now be used to account for the influence of the human body on the sound field by convolving the HRTF for a specific direction with the according sample. The room impulse response obtained by this method is the binaural room impulse response (BRIR) and describes the influence of the room acoustics at a persons ears.

2.3 Spatial decomposition method

The spatial decomposition method is a method that can be used to parameterize a sound by multichannel measurements to be able to recreate this sound field with these parameters. In our case, a room impulse response measured with a microphone array can be analyzed to obtain the DOA of each sample. With this information the corresponding sample with its energy can be mapped to the corresponding direction. This can be loudspeakers or, as in our case, headphones.
The assumptions made for this method are that the direction of propagation of the sound field at a specific moment in time is the average of all sounds arriving at that time instance and that the sound pressure at that time is defined by the sound pressure in the center of the microphone array [10]. This assumption leads to a limitation of the method. In a case where two reflections arrive at the same time, it is not possible to determine both directions, but the direction with the most energy will lead to the propagation direction of those samples. Other assumptions are that a finite number of image sources can be used to represent the impulse response and that a room impulse response can be split up into an early part and a reverberation part by the so-called mixing time [10]. At this mixing time the energy net flow is zero, so the amplitude is equally distributed and the distribution of the phase and direction are uniform.
In this work the provided SDM implementation [11] of an optimized version of the SDM for binaural reproduction [12] is used. The following sections summarize the main processing steps according to [12]. A schema of this method is displayed in figure 2.2.

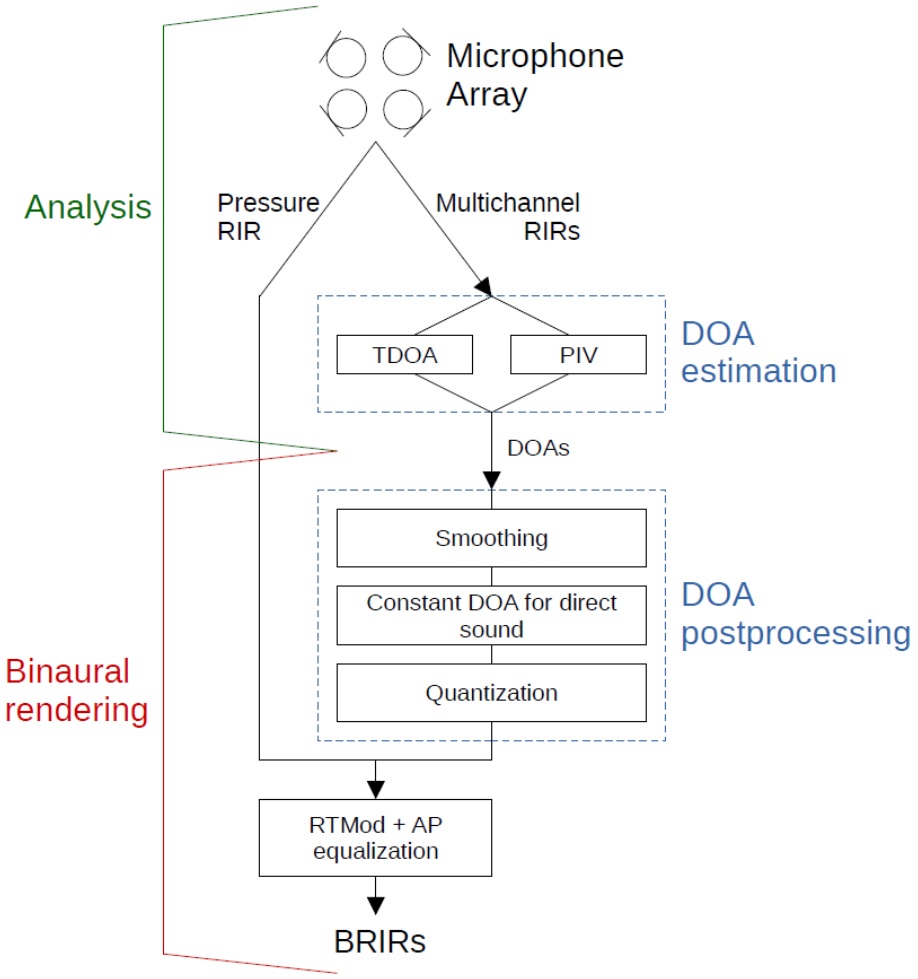

Figure 2.2: Schema of the spatial decomposition method.

2.3.1 DOA estimation

The first step of the SDM is to analyze the multichannel room impulse responses and estimate the direction of arrivals for each sample. There are two different ways to obtain the DOA from a multichannel recording, depending on the kind of signal that is measured and the available microphone array.
The first one is called the Time Difference of Arrival (TDOA) method and is also used in this work. This method requires an open microphone array with at least four microphones arranged in the 3D space and analyzes the time differences between the microphones. To get the direction of each sample, a sliding Hanning window of one step at a time is applied to the microphone signals and the cross-correlation between all signals are calculated. For this the size of this window must be greater or equal to the maximal travel time between the microphones with the greatest distance. The delay that maximizes the cross-correlation is the corresponding time difference of arrival and the direction of arrival can be calculated by using the TDOA between all microphones.
The second method is the so called Pseudo-Intensity Vectors (PIV) method. This method is based on the instantaneous intensity a sound field has. This intensity

is a product of the sound pressure and the particle velocity. Since the particle velocity is a vector and therefore has a directional information, so has the intensity. The direction of arrival of the sound field is then the opposite of the direction of the intensity vector. The intensity for this method can be calculated from a specific microphone array, the B-format microphone. This microphone array has four microphones in a specific setup with one omnidirectional microphone to obtain the sound pressure information and three orthogonal figure-of-eight microphones which output signals are assumed to be proportional to the particle velocity.

2.3.2 DOA postprocessing

The obtained DOA from with the TDOA or the PIV could directly be used for the binaural rendering, so the creation of the BRIR as the weighted and delayed sums of the HRTFs for each DOA. This, however, can lead to distortions and tone coloration of the output signal. This is due to unstable DOA estimations. Sound events can spam over several samples and should be mapped to the same direction, which is not guaranteed with the methods described in section 2.3.1. This can have different reasons. There could be multiple reflections interfering and only the direction with the most energy is used to determine the direction, so that one sample could be assigned to the direction of the first reflection, while the next sample could be assigned to the direction of the other reflection. It could also be that the wavelength of a signal is longer than the analysis window for the DOA estimation and thus the corresponding event is split up into different time analysis windows and mapped to different directions due to strong signal components in other frequency ranges and the analysis being based on a broadband DOA estimation. The common problem of the mentioned situations is that a sound event is mapped onto different locations which means that one event is split up into several impulses from different direction. Due to the nature of impulses they are broadband signals, which can thus lead to a spatial spread and distortions of the original sound event.
To avoid this a moving average filter is applied on the DOA estimations and the direct sound is forced to have a stable DOA by taking the sample with the largest amplitude of the direct sound to obtain the corresponding DOA for it. An additional quantization step is included after the smoothing of the DOA in order to have a fixed number of positions for the early reflections and thus no spatial and timbrel degradation. There are different ways to quantize the DOAs, for this work a Lebedev grid with a size of 50 points is used. The DOAs are assigned to these points by determining the closest direction of a DOA to any point on the grid.

2.3.3 Reverberation equalization

The final step of the rendering part is the equalization of the reverberation, so everything after the mixing time. As explained earlier, after this point in the SRIR the amplitude is assumed to be equally distributed and the distribution of the phase and direction are assumed to be uniform. So from this point the DOA estimation becomes unreliable since it will result in random directional information which will lead to broadband impulses and thus a spectral whitening and an increase in the

reverberation time at high frequencies. The solution for this problem is to split the BRIR into octave bands and to modify the energy envelope of each band and additionally process the reconstructed BRIR with allpass filters in order to increase the late reverberation and the echo density. This way the quality of the late reverberation is less dependent on the signal.
The first part of adjusting the energy envelope is called RTMod and is based on the information of the reverberation times of the original pressure RIR. Correction constants are calculated using the sub-band reverberation times of the original and the synthesized RIRs and used for adjusting exponential functions that are multiplied with the subband BRIRs.
The aim of the allpass filter (AP) equalization is to break up possible strong reflections that might have build up due to constructive interferences of late reflections. Without having an influence on the spectral characteristics, the AP filters can increase the echo density which results in a smoother time envelope.

2.4 Listening tests

Generally there are many different psychoacoustic tasks. They can yield different results and small changes can have an influence on the measurement time, the complexity of the experimental setup and the need to fulfill certain requirements. Psychoacoustic tests can be divided into two groups, the ones that yield a result directly after one trial, so only one answer from the test subject is needed and those that need several trials until a result can be found [13].
Example for the first group are for example adjustment methods, where the test subject has full control over the stimulus and can for example set the frequency or the level to a specific value that for them matches some asked perception of that stimuli. A similar method, where the subject can only control the direction of the change of the stimuli, for example higher of lower frequency, until the required perception is achieved, is called a tracking method. A last example method for the first group is the magnitude estimation method. Here, for example the loudness of a group of stimuli needs to be assigned using a reference stimuli with a number that corresponds to its loudness.
Examples for the second category are yes-no procedures, interval forced-choice tests and adaptive procedures [13]. In the yes-no procedure, the test subject needs to state whether a signal was present in the current interval of not. The stimuli changes during the experiment. Similar to this is the interval forced-choice method, where the participant needs to decide whether the signal to be detected was in the first of the second interval. The last method mentioned here is also the one that is used for this work, it is called an adaptive method. Here, not the examiner chooses the stimuli but they are dependent on the subjects answers. These methods are also called up-down or staircase procedures because the intensity if a stimulus is decreased if the subject is able to identify the stimulus and increased if not. There are many ways on how to implement an up-down test. The number of trials can be adjusted as well as the possible answers for a DOWN or an UP decision and the step size in which the stimulus is changing (and whether this step size is constant or changing during the experiment). But generally this test is used to determine some kind of

threshold which is achieved at the point where the stimulus intensity will half of the time be increased and half of the time decreased. This results in a specific point on a so-called psychometric function which describes the relationship between a physical stimulus and the ability of a human to detect that stimulus. An example how this psychometric function can look like is displayed in figure 2.3. It can be seen that

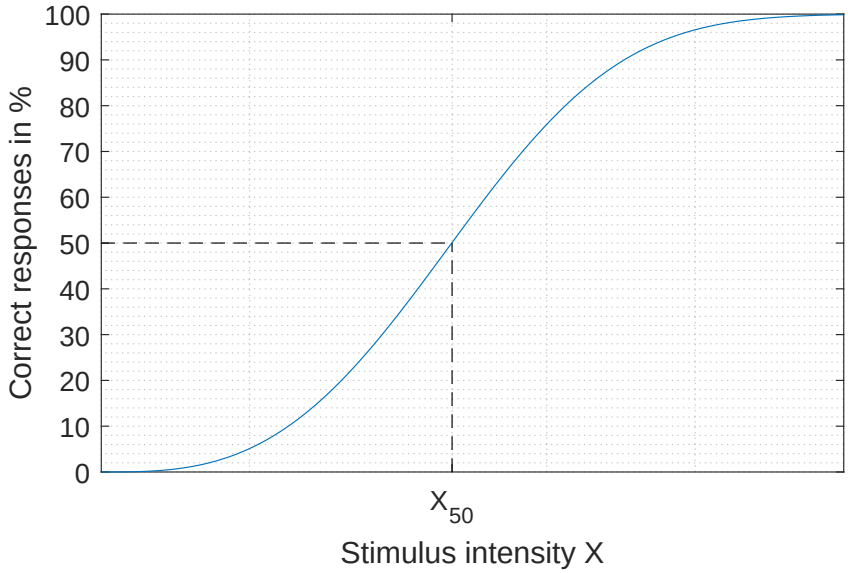

Figure 2.3: Example of a psychometric function with the guessing point of 50% correct answers marked.

with a low stimulus intensity the percentage of correct responses is zero, so the stimulus is not identifiable. With increasing intensity the subject is able to detect more stimuli until, from a specific intensity, the participant is able to correctly identify the stimulus in 100% of the cases. It is important to note here, that the intensity of the stimulus can relate to many different characteristics of a physical stimulus. A possible test could for example be to examine the hearing threshold of a pure tone. In this case the intensity of the stimulus would be its sound pressure level. If the tone is very silent the subject is not able to hear it. With increasing level the tone will be more and more audible and from one specific level on remain completely audible for all higher sound pressure levels. With this example it is becomes clear, that every psychometric function looks different from person to person and also different for one person but for different tasks. Some psychometric curves can for example be steeper than others if the turning point from hearing to not hearing some change of the stimulus is very sudden, for example if the resolution of the auditory system for that sensation is sharp. On the other side it can also be that some some psychometric curves are rather flat if some characteristic of a stimuli is hard to process for the human auditory system. It can also happen, that the curve does not reach 100% because a person is just not able to clearly identify a characteristic. Of course a psychometric function is not only limited to psychoacoustic but can also be used to test other abilities of humans, for example to smell or to see.
Going back to the up-down methods, it was mentioned already that there are many

ways to conduct that kind of test. Regarding the psychometric function, it is very important how the rules of increasing and decreasing the intensity are since they will determine the convergence point of the test and thus the resulting point on the individual psychometric function. Let's consider the case of a simple one-up-one-down method first. As the name suggests, there is one condition, a correct answer, that leads to a decrease of the intensity and one, a wrong answer, that leads to an increase. Since the guessing probability of both, the UP decision and the DOWN decision are the same with 50%, the convergence of this method is the 50% correct responses point of the psychometric function for a stimulus intensity of X_{50} as marked in figure 2.3. However, also the other factors such as the number of trials and the step size can have an influence on the convergence point. It was shown in [14] that the landing point, so the resulting percentage point on the psychometric function differed for a fixed x-up-y-down method with varying relations between the step size for the up and the down steps as well as the with varying relation between the relative size of the step up to the spread of the psychometric function. These factors can lead to drastic differences from the targeted point on the psychometric function in ranges of up to 10% or more. This will further be dealt with in the discussion.
In order to initially target a different point than 50%, for example to be able to estimate the behavior of the psychometric function, it is possible to use transformed up-down methods. These have more than one condition for the UP and/or DOWN decision, for example the one-up-two-down method, which is used in this thesis. Here, the stimulus intensity is increased only if two correct answers have been given in succession. Thus there is still only one DOWN condition. But the UP group now consists two conditions, one false response or one correct followed by one false response. The convergence of the listening test is always the point where the probability of going up is the same as the one for going down, so the UP group has a probability of 50% as well as the DOWN group. This is visible in figure 2.4, where both, the one-up-one-down and the one-up-two-down method are shown. For the simple one-up-one-down method there is only one condition in both the UP and the DOWN group thus the function displayed here corresponds to the psychometric function in figure 2.3. However, in the new method, there are two conditions in the UP group, just one in the DOWN group, but both groups have the same probability at the convergence point. This means the probability of choosing a correct answer is higher than choosing a wrong answer and thus the point on the psychometric function must lay above the 50% point. Thus the resulting convergence point for this method is at 70.7% of the psychometric function [15].
This way a lot of different points on the psychometric function can be targeted with the according step size ratios and a sufficient number of trials in a test to ensure convergence.

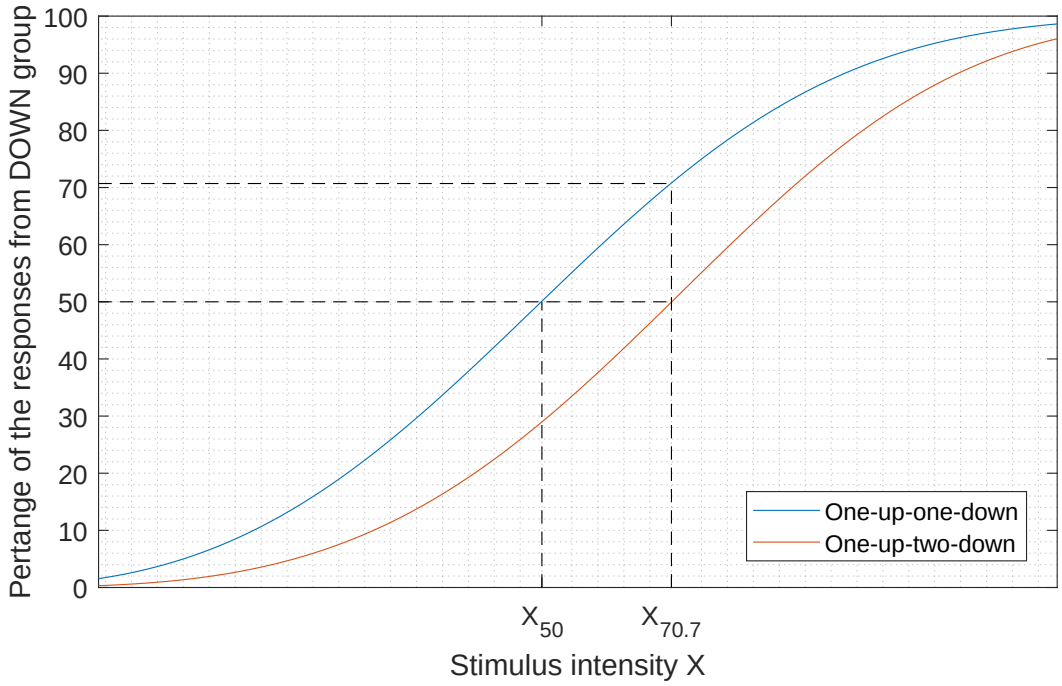

Figure 2.4: Percentage of responses from the DOWN group over the stimulus intensity for the simple one-up-one-down and the one-up-two-down method.

3

Method

The purpose of this work is to examine how important the directional information of the first reflections of a room impulse response is for the subjective perception of that room. It is, however, very difficult to give a general answer to that question since every room is different and so are the reflections. As discussed in the theory section, the perceptual importance of a first reflection depends on may factors, including the time it appears in the RIR, the level of that reflection and the directional information. Additionally, the perception also strongly depends on all other reflections in that RIR. This could result in a situation where a person might rate a RIR with altered azimuth reflections as plausible when positioned in the middle of the room, while rating the same alteration as implausible when that person is close to a wall. This is due to permanent changes of all the including factors mentioned above when changing the receiver or source position in a room. Therefore, many studies of different scenarios would be needed to make a general statement about the importance of the first reflections, if this is possible at all.
In this work, therefore, only a small aspect of perception is investigated using one RIR with only one source and receiver position. The aim of this work is to find out, how far the first reflections in this specific scenario can be shifted in the horizontal plane until the subjects hear a difference. This can give an indication of how sensitive human hearing is to changes of the first reflections in a RIR. For this purpose, a ABX listening test is performed to determine the smallest possible angle shift of the first reflections that the subject can perceive. However, this hearing threshold, or just noticeable difference, must be distinguished from plausibility tests. Just because a difference between the signals can be detected does not mean that both RIRs cannot plausibly reproduce the corresponding space.

This chapter describes the method used for the ABX threshold tests and the statistical analysis of them. First, section 3.1 gives information on the used room impulse response and how the directional information is obtained and altered for the experiment. It also explains the creation of the test stimuli in real time during the listening experiment. The hardware setup, experiment procedure, user interface and participating test subjects are described in section 3.2 followed by the statistical analysis in section 3.3.

3.1 Creation of the stimuli

3.1.1 Used impulse response and DOA estimation

The room impulse response used for this thesis is from a measurement of the Audio Lab at the Division of Applied Acoustics in Chalmers University of Technology. The room has a volume of approximately 46m^2 with the dimensions of 4.75m width, 3.7m length and 2.6m height. The room impulse response was measured with a 32-channel Eigenmike em32 with the source being a Genelec 8030A loudspeaker. The room impulse response can be seen in figure 3.1. It has a short reverberation time of approximately 0.27 s. The directions of arrival for all samples of this RIR can

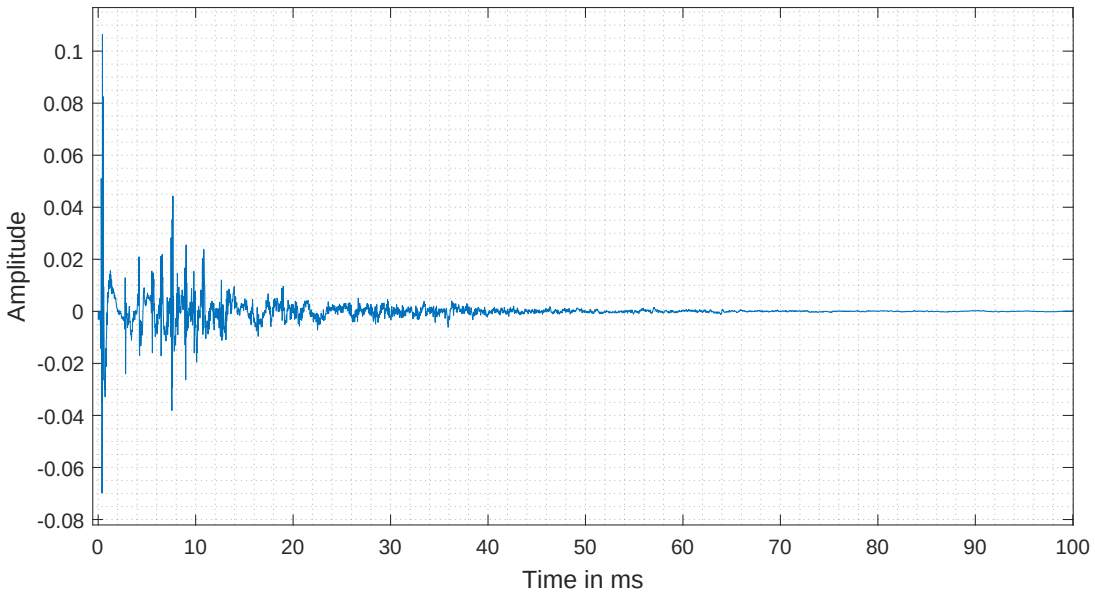

Figure 3.1: Room impulse response of the audio lab.

be extracted using the SDM method discussed in the theory part. This information together with a squared version of the RIR, which corresponds to the energy, can be used to extract the first reflections. Figure 3.2 shows the corresponding graphs. The marked first five reflections can clearly be seen due to the combination of high energy and a stable azimuth and elevation estimation. With the DOA information it can also be seen where the reflections come from in the room. The azimuth angle is stable at 0° for the first two reflections while the elevation is first positive at around 40° and then negative at about -48°. This means the first reflections in this source-receiver setup com from the ceiling and the floor. The following three reflections have a stable elevation angle at 0°, so they come from the same height as the source. The azimuth angle for the third reflection is at -54° which means it comes from the wall of the right side. With an azimuth angle of 55°, the fourth reflection comes from the left side wall. The last reflection here as an azimuth value of 178° so it comes from the back wall of the room. These reflections are first order reflections since they are reflected by one surface only. All of the following reflections

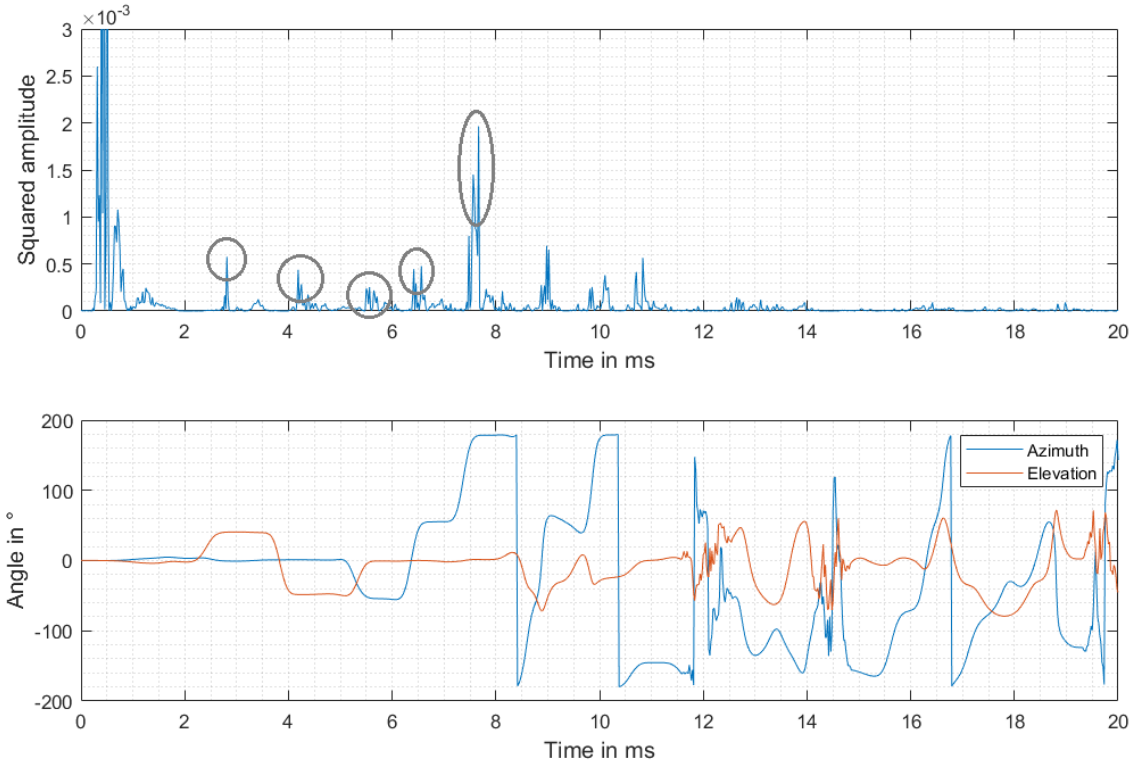

Figure 3.2: Top: Squared impulse response of the audio lab. The markings show the first five reflections. **Bottom:** Azimuth and elevation calculated with the SDM method.

are combinations of several reflections, so second and higher order reflections. Since the room is quite small, the reflections are very close to each other in time so that the following reflections could arrive at the microphone at the same time. This can lead to unstable direction of arrival estimations. Additionally, the energy of these reflections decreases which also leads to a more difficult analysis. This is why only the clearly visible first order reflections are used in this work.

The DOAs of the whole RIR are in the next step quantized to a Lebedev sphere grid of 50 points to avoid spreading the reflections onto multiple directions. Since the purpose of this work is to investigate the perceptual changes when the azimuth of the first reflections is changed without any influence of the elevation angle, the elevation angle is set to 0° over the whole signal. In addition, the reflections should not lie in the median plane and should be spatially distributed as well as possible. The chosen reflections therefore have azimuth angles of 45°, -60°, 135°, -80° and -160° (see table 3.1 and figure 3.6). Since the azimuth angle of the reflections change with small steps in the next part, they need to be at exactly that position they are assigned to and are therefore excluded from the quantization process. The area around the reflections that has been changed is based on the quantized DOAs of the original RIR. The resulting DOAs are used for the experiment to render the reference signal. They can be seen in figure 3.3.

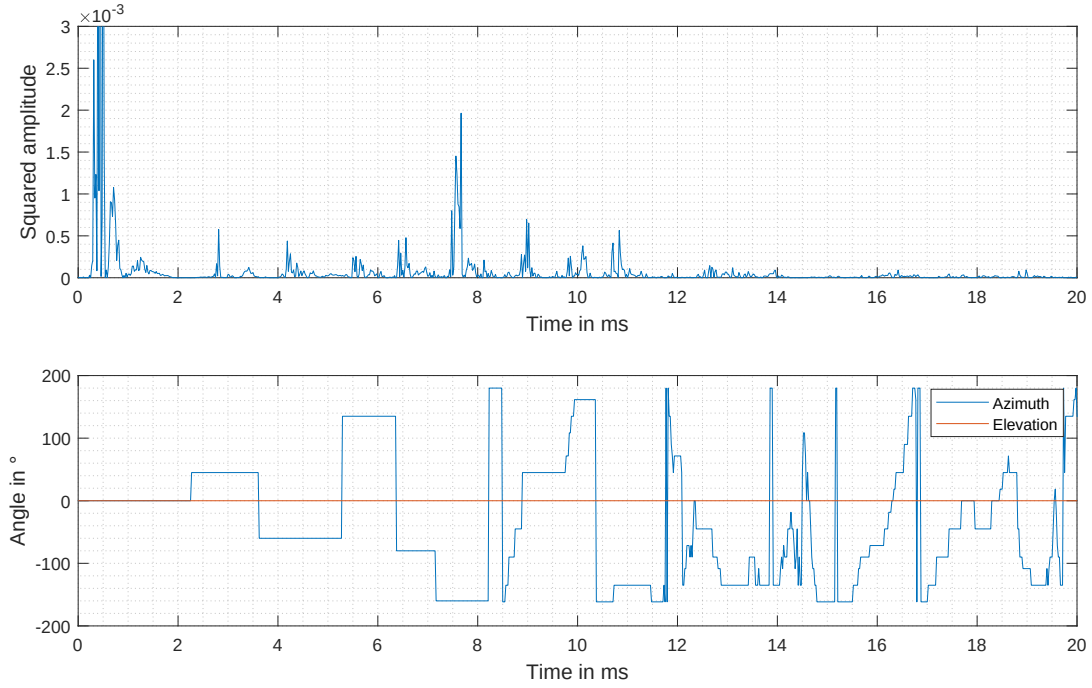

Figure 3.3: Top: Squared impulse response of the audio lab. **Bottom:** Quantized and adjusted azimuth and elevation angles.

3.1.2 Alteration of the first reflections

In order to be able to examine how much the early reflections can be shifted so that a test subject can just hear a difference between the signal with the shifted angles and the reference signal, the chosen five reflections have to change gradually in azimuth angle. The beginning conditions should also be clearly distinguishable from the reference. In the self-test it was found that the greatest difference between the signals was found when all reflections start at a common point and this point has at the same time the greatest possible binaural differences. Thus this signal lacks some spaciousness compared to the reference and the whole signal has some kind of weighting to the corresponding side of the reflections. It was therefore decided to have a gradual change in the azimuth angles for the chosen reflections all starting at 90° to their corresponding reference positions. This way the signal gradually gains spaciousness and loses the weighting towards one side.

The step size of this change is oriented on the biggest angle change for the reflections, which is the fourth reflection with a reference azimuth angle of -80°, which results in a maximum angle shift of 170°. This reflection changes in angle with a step size of 1°. To have a gradual change for all reflections all other reflections also need 170 steps from the maximum to the reference position but with an accordingly smaller step size. The step sizes for all reflections can be seen in table 3.1.

Table 3.1: Reference azimuth angles and step sizes for the angle shifting for all five reflections.

Reflection number	Reference azimuth angle	Step size
1.	45°	0.2647°
2.	-60°	0.8824°
3.	135°	0.2647°
4.	-80°	1°
5.	-160°	0.6471°

Again, the quantization of the impulse response is done for the whole signal except the first reflections, so the small changes can be rendered correctly. After the quantization and changing of the angles for the first reflections, the BRIRs are rendered using the HRTF set up to the mixing time, which is 40ms for this RIR. For this a resolution of 2° in azimuth and 0° in elevation is used, since no changes in elevation angle are of interest and thus not to be tracked. The HRTF set is used from [16] and was measured with Kowles FG-23329 miniature microphones in the ear canel of a KEMAR dummy head. The data set is in the spherical harmonics domain, which allows to interpolate the HRTFs for an arbitrary direction. This way no other quantization or clustering of the data is needed to fit the HRTF set, since the HRTFs for all the quantized and defined directions of the DOAs can directly be calculated. After this the late reverberation after the mixing time of 40ms is rendered for one direction since it is independent of the direction. The resulting BRIRs for all angle shifts of the first reflections and rendered in steps of 2° azimuth are saved for the main experiment.

3.1.3 Real-time convolution of BRIRs and audio examples

The convolution of the BRIRs with the audio examples is a real-time convolution during the experiment. This way the different parts of the convolution can simply be exchanged when the audio example of the head orientation changes. For the real-time convolution a overlap-add fast convolution was used. A fast convolution means that the convolution is not performed in the time domain but in the frequency domain instead. For this the two input signals have to be zero-padded to have the same length that both signals have added together. Then the signals are transformed into the frequency domain and multiplied with each other. The output signal then is transformed back to the time domain with a iFFT and the output is the result of the convolution. The fast convolution results in a much shorter execution time if the signals to be convoluted are long. Figure 3.4 shows the procedure of the overlap-add convolution. This is performed in a black processing, so the audio signal is separated into blocks of 512 samples. In each processing step, a signal block $x_i(n)$ of the audio signal is convolved with a BRIR, $h(n)$, using fast convolution. The result of the convolution is called $y_i'(n)$ and has the same length as $x_i(n) + h(n)$. Since the BRIR is much longer than the audio frame, the convolution result is several blocks long. Therefore only the first block of the current convolution is used for the actual output signal of this block. For the very first block (orange), there is no convolution rest so the output block is just the first 512 samples of the convolution result. For all

other blocks, however, the convolution rest of the previous blocks has to be taken into account. for example in the third case in figure 3.4 (blue), there are still the remainders of the convolution of the two previous blocks. Thus all three blocks have to be added to obtain the final output block.

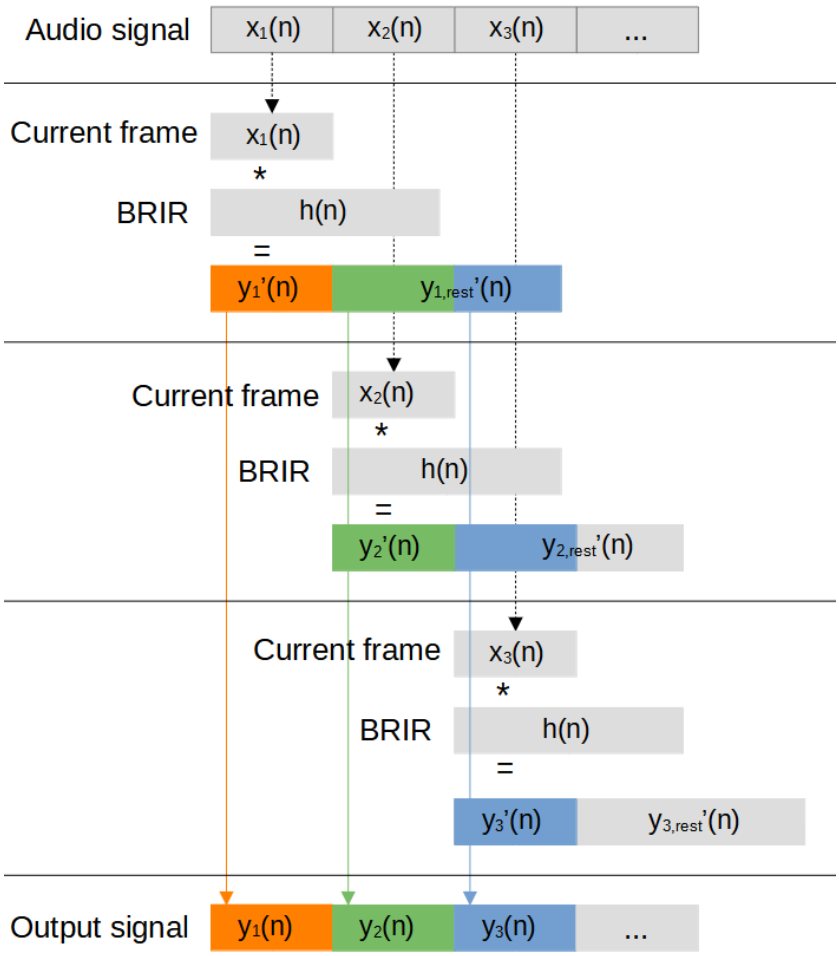

Figure 3.4: Schema of the overlap-add convolution.

The advantage of the real-time convolution is that both, the input signal and the BRIRs can easily be changed during the convolution. If there is a different audio signal, the new block can simply be taken from the new audio signal. Correspondingly, if the BRIR changes, this new BRIR will be taken for the current and the future convolutions. However, it is important to still add the old convolution remainders so that there are no artifacts between the blocks.

This convolution is implemented in MATLAB and starts running as soon as the test subject selects the first example with the user interface (see chapter 3.2.2). The audio file will repeat when it reaches the end so the signal is played until the participant has reached a decision. During a trial only the BRIRs change, either due to head movement or because a different stimulus is selected. Both changes are represented in the BRIRs, either with different HRTFs in case of head movements,

or with different SRIRs in case of a different stimuli (different stimuli is equal to different angle shift).

3.2 Experiment setup

The general hardware setup for the experiment can be seen in figure 3.5. Besides the computer, AKG K-702 headphones, a Focusrite Scarlett 2i2 audio interface and a Polhemus Patriot magnetic tracking system. The participant sits in front of the computer which shows the starting screen of the experiment which can be controlled with the mouse and the keyboard. The headphones are connected with the head-tracker which is located behind the participants to track the head movements. The stimuli are played at a level that roughly corresponds to the normal conversational level of about 65 dB. During the training phase, this can be slightly adjusted according to personal perception.

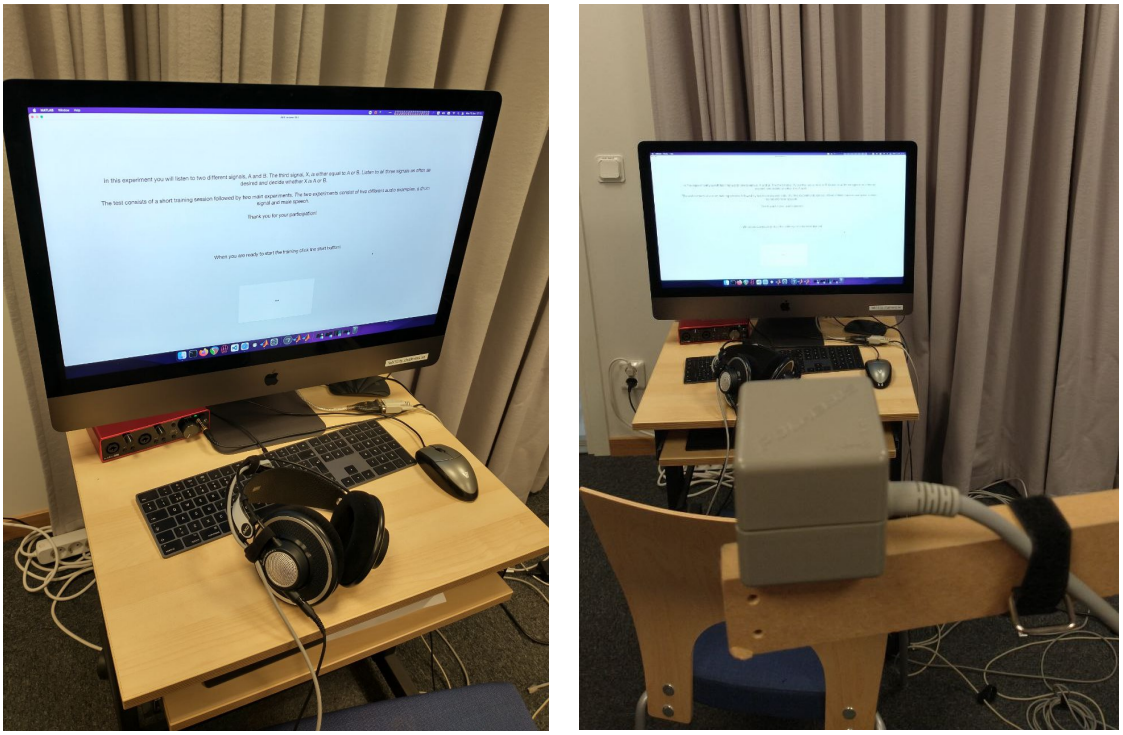

Figure 3.5: Left: Experiment setup with the starting screen and the used headphones. **Right:** Head tracker behind the seat for the participants to track the current head position.

3.2.1 Experiment procedure

The whole test is divided into a short training session followed by the two main experiments. Before the start of the experiment the subject gets written instructions about the procedure of the test. This is followed by an initialization of the head tracking system during which the participants are instructed to look straight ahead at the screen. After this the test subjects are allowed to move their heads freely

again. The last part before the beginning of the training session is an information screen on the computer showing the essential task one more time.

After clicking on the start button, the test begins with a short training session. The audio used for this depended on the randomized starting audio for the actual test. The training uses the same audio as the first experiment. Six conditions are tested in the training, which are the angle shifts of 171°, 121°, 141°, 101°, 11° and 71°. With this the training included mainly audible differences between the signals so that the participants are able to get used to signals and the kind of difference between the conditions. Besides that, also harder conditions are used in the later part of the training. These were chosen to get the subjects used to the fact that it will also be difficult to detect differences in the test. With the end of the training session the participants again have to click a start button to precede to the main experiments to make sure they concentrate properly.

The following two main experiments are threshold tests, so the goal is to find the individual just noticeable difference between the signal with the original and the shifted azimuth angles of the first reflections. To obtain the threshold an adaptive one-up-two-down method is used in combination with an ABX test. The ABX test is used for the comparisons between two conditions A and B, which are the reference signal and the signal with shifted reflections. The answer of these trials are used in the adaptive one-up-two-down method. The two experiment starts at an angle shift of 101° with a step size of 20°. Using table 3.1 it can be seen that the initial angle shift of 101° means that all five reflections start the adaptation process at different points between 90° and their individual reference position. Due to the different step sizes of the reflections this also means that the initial angle shift is not 101° for all reflections but only for the fourth. All other reflections have accordingly smaller initial angle shifts. Again, this is because the biggest step size of 1° of the fourth reflection is used as a representative for the angles and angle shifts for all other reflections as well even though the actual values are smaller. To calculate the initial angle shifts for each of the five reflections the representative value of 101° has to be multiplied with with their individual step sizes given in table 3.1. With the initial angle shifts for all reflections the starting points for the adaptation process can be calculated and are displayed in figure 3.6 together with their reference incidence angles. The plot also shows the relative level of the reflections compared to the direct sound. The initial angle shift was set to 101° and not to the maximum of 171° (in that case all reflections would have started at 90°) because informal tests showed that an angle shift of 101° was clearly audible and a higher value would unnecessarily lengthen the adaptation process. Additionally, a starting point at 90° for all reflections would mean that the maximum angle shift is already reached and thus no upwards step in the adaptation process would be possible in the beginning of the experiment.

Starting the experiment, the initial angle shift is increases if a wrong answer or a correct answer followed by a wrong answer is given and increased if two correct answers are given. When a reversal from increasing to decreasing of the angle shift is found the step size is halved. The test ends after 8 reversals for each threshold test. More reversals could give more accurate results, but were not possible in terms of the time involved. Since the signals used here can not exceed an angle shift of a

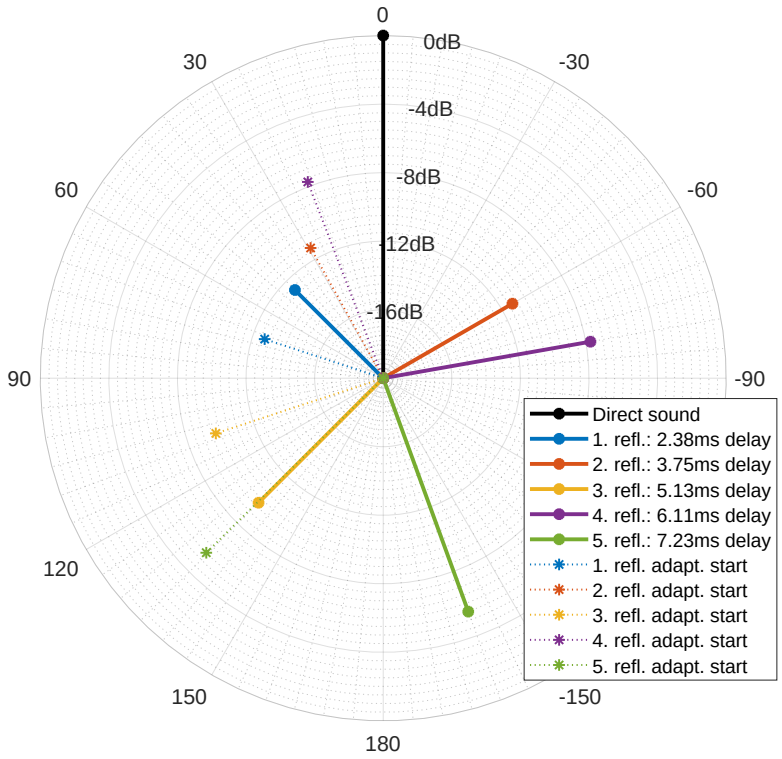

Figure 3.6: Reference angle of incidence of the direct sound and the five first reflections with their relative level to the direct sound. The dotted lines show the angle of incidence of each reflection at the beginning of the adaptation process.

maximum of 171° and a minimum of 0°, the algorithm reaches limits at these points. If the maximum shift of 171° is reached the adaptation will stop and it is concluded that the test subject is not able to hear any difference between the signals or did not understand the task of the experiment correctly. At the other end the limit is set to a minimal angle shift of 1° but the normal procedure of the one-up-two-down method continues to find the threshold.
When the threshold tests are finished after 8 reversals, one additional condition is tested and this sub test is in the following referred to as the discrimination test. The test signal of this condition is a version where the reference first reflections are mirrored on the frontal axis, so a reflection from the right side turns into a reflections from the left side and vice versa. This single condition is tested twice before continuing with the second threshold test or finishing the experiment, respectively.

3.2.2 User Interface

The user interface for this experiment is implemented as a MATLAB GUI. It is connected to an additional MATLAB instance that controls the playback of the stimuli when the user chooses a signal or gives an answer. The pick of the stimulus can be controlled by the mouse as well as with keyboard shortcuts. This enables the participants to quickly change between the three signals and therefore an easier comparison of them. The GUI for the ABX experiment used for the training and

the main tests can be seen in figure 3.7. The main parts of the GUI are the three

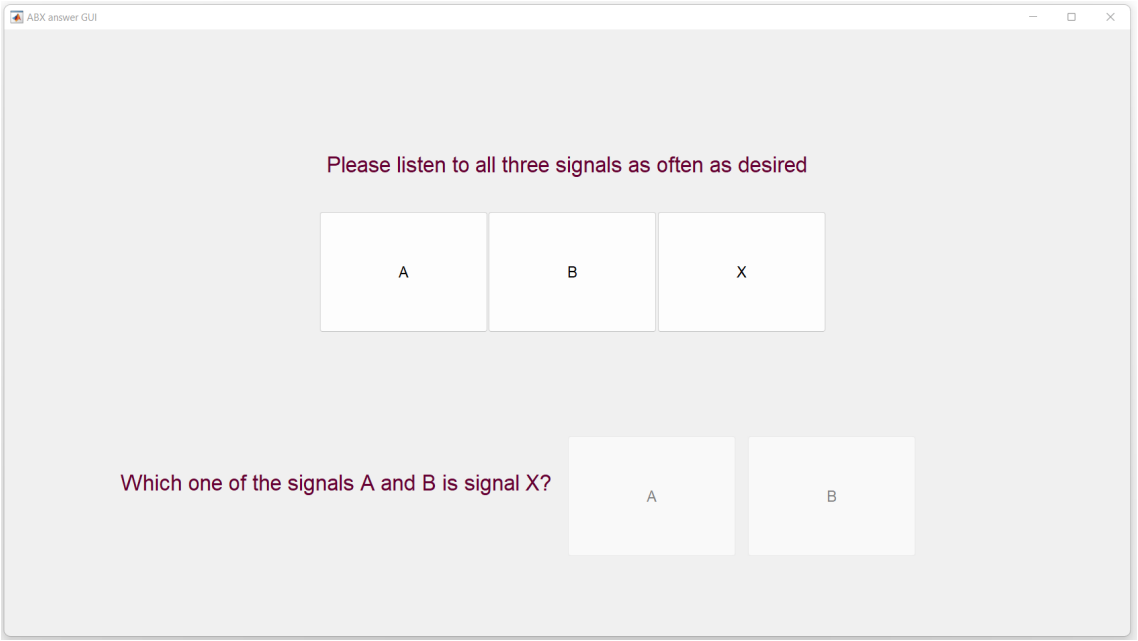

Figure 3.7: Listening experiment user interface.

signals A, B and X. The test subjects were able to listen to the signals as often as desired and the audio examples were running in a loop, so that they never stopped playing. The answer to the ABX task could only be given after the test subject listened to all three signals at least once, until then the buttons were disabled.

3.2.3 Participants

The experiment was performed with 20 participants, including PhD students at the Division of Applied Acoustics, students in the Master program "Sound and Vibration" and people without a background in acoustics. The age of the subjects ranged from 23 to 46 years with a mean age of 28.95 years and a median of 29 years. 8 participants were female, 12 male. 13 of the 20 participants stated to have some experience with listening test, while the 7 people without a connection to acoustics have never taken part in a hearing test. All test subjects stated to have a normal hearing.

3.3 Analysis of the listening tests

The listening test in this work consists of two parts. The first and main one is the threshold test to determine the angle shift of the first five reflections that give a perceivable difference to the reference signal. The other test is the simple discrimination task where two additional signals with mirrored reflection angles are given at the end of the threshold tests.
The basis of both tests is composed of the individual ABX tasks. The ABX test is a Bernoulli trial, so the outcome of it is either success of failure and the probability

of answering correct is always the same for a fixed condition. If A and B are indistinguishable, so if the subject is guessing, the probability of picking X=A or X=B is the same with 50%. Or in other words, when the outcome of the test is that 50% of the participants answered correctly, it means that there is no difference between the two signals. If the outcome would be 0% it means that all subjects constantly picked the wrong one which could indicate an error or basic misunderstanding of the task. However, since every person picked the wrong one it also means that there is a clear difference between the signals. The ABX task is also only applicable to tell whether there is a difference between A and B and not to proof that there is none. With this as a basis for both experiments the following two sections explain how to analyze both the threshold and the discrimination test.

3.3.1 Threshold test

The outcome of the threshold test will be a sequence of angle shifts that describe the convergence process of the participant and thus converges to the 70.7% point on the psychometric function. However, the last point of this process is not used as the resulting threshold. This point could be distorted, for example, by correct guessing of the last two trails. The usual method for staircase methods is thus to average the stimulus values at the reversal points and use this average as a threshold estimate. For this averaging, the first reversals are usually omitted, since they could bias the results due to initial uncertainties [14]. A common number of reversals for tests of approximately the length used here is to average over the last 4-5 reversals[17], [18]. Since the step size was halved at upper reversals the averaging should contain an even number of reversal points to avoid any bias due to the influence of a different step size in only one direction. Therefore the stimulus level at the last four reversal points were used to calculate the resulting threshold for each participant and test signal.
These individual threshold are then used to calculate the median of the group of 20 participants was calculated. This was done to minimize the influence of outliers since it was found later that the thresholds of some participants were very high compared to the majority of the group. In addition to that, the 25th and 75th percentiles are given as well as the region of 1.5 times the interquartile range and the outliers in order to gain an understanding of the spread of the measurement.

3.3.2 Discrimination test

According to [19], on which this section is based on, there are two different ways of statistically analyzing discrimination tasks. One is based on a binomial distribution while the other is the signal detection theory. As the name suggests, the first method assumes a binomial distribution. Two requirements need to be fulfilled to fulfill this assumption. First of all, the correct answers A or B need to be randomly assigned to X, while also the two signals to be compared have to randomly vary between being A and B. Secondly, the guessed response has to be uncorrelated to the audio signals and random, so the probability of answering X=A is the same as answering X=B. Uncorrelated to the audio signals means that there must not be any other

difference between A and B other than the one under test, so that A and B are indistinguishable if this difference to be studied does not exist. This sounds trivial, but it can happen that e.g. small artifacts, interruptions between the stimuli or level differences occur due to signal processing, so that X can correctly be assigned to A or B despite A and B being (almost) identical signals. If these requirements can not be fulfilled, the distribution is not binomial but has some kind of bias. This can for example mean, that the probability of answering X=A while the true answer is X=B is more likely than answering X=B while truly X=A. In a binomial distribution these probabilities are the same, but if they are not, a more complex analysis needs to be done to make up for those biases. Here, however, we can proof due to the method and measurement setup that the requirements for having a binomial distribution are fulfilled and can thus continue with the more basic statistical analysis for this discrimination test.
With the assumed binomial distribution the binomial probability mass function can be calculated with

$$f(s) = \binom{n}{s} \cdot p^s \cdot (1-p)^{n-s}, \tag{3.1}$$

where n is the number of number of trials, s is the number of correct answers and p is the probability of a correct answer. In the case of this experiment, $n = 40$ (2 each for 40 participants) and $p = 0.5$ for the guessing case. With this the probability mass function can be calculated and is displayed in figure 3.8 on the left side. It shows the probability if having s correct answers when the participants are just guessing in all 40 trials. It is clear, that the most likely number of correct answers in this case with a 50% probability of guessing the correct answer is exactly 20, so at half of the cases. More interesting is thus the right graph in figure 3.8 which shows the inverse cumulative binomial probability function and is calculated by summing all probabilities of the current s up to a total of all n correct answers:

$$g(s) = \sum_{m=s}^{n} \binom{n}{m} \cdot p^m \cdot (1-p)^{n-m}. \tag{3.2}$$

So the graph shows the probability of achieving a minimum of s correct answers purely by guessing. Thus the probability of having at least 0 correct answers is 100% and this probability decreases with more correct answers. The chance of guessing all 40 answers correctly is only approximately $9{\cdot}10^{-11}$%. For psychoacoustic experiments the 95% confidence interval is sufficient in order for the result to be statistically significant, which means that less than 5% can achieve a specific number of correct answers by guessing. This means for example that at least eight out of ten people need to correctly identify X in order to find a significant difference between the signals A and B. In this work with 40 trials that results in a number of at least 25 to 26 correct answers in order to be able to say that a clear difference between A and B could be heard.

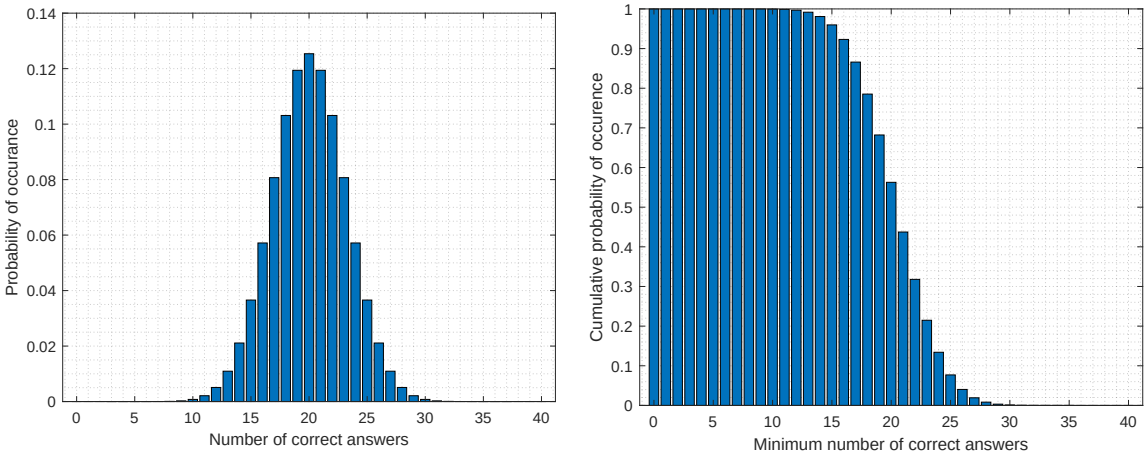

Figure 3.8: Left: Binomial probability mass function for n=40. **Right:** Inverse binomial cumulative distribution function for n=40.

4

Results

4.1 Threshold test

All measurement curves of the threshold test for the 20 participants using the speech signal can be seen in figure 4.1. The measurement curves are distributed over a wide

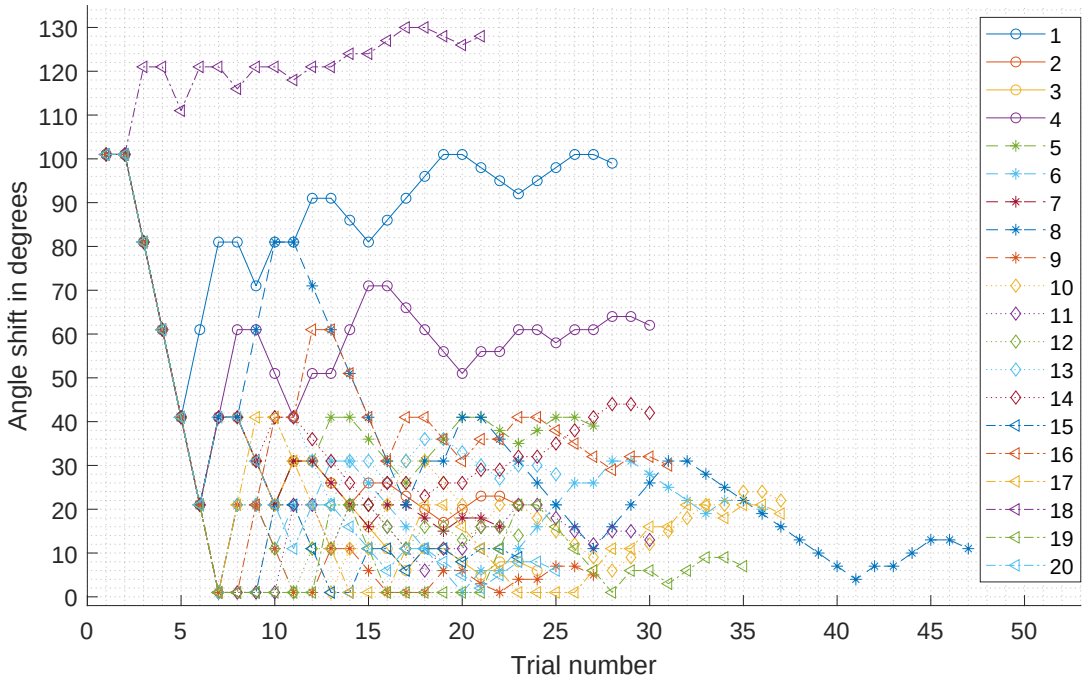

Figure 4.1: Listening test curves for all participants for the speech signal.

range of angle shifts with the focus on the range between approximately 0-40° (17 out of 20 participants). One curve of subject number 18 even exceeds the initial angle shift of 101° which means the experiment started below the threshold of this person. The interpretation of the measurement curves can be explained using the curve of test person 2 in graph 4.2 since it represents the ideal shape of a one-up-two-down threshold test. The participant starts with correct answers and the angle shift decreases with the initial step size of 20°. At around 20° the participant answers wrongly and thus the angle shift increases directly. At this angle shift, the difference seems to be audible again for the participant since two correct answers are given, leading to a decrease in the step size and decreasing angle shift. This continues until

the eights reversal can be seen at trial number 23, where the tests stops. The final threshold for this test person is defined as the average over the last four reversal points, which are marked in the picture. The general shape of this measurement curve is the desired shape since it shows a fast approximation of the area of the threshold for the angle shift and then slowly approaches a more accurate estimation with the decreasing step sizes after an upper reversal. It is obvious from 4.1 that not all measurements correspond to this ideal curve.

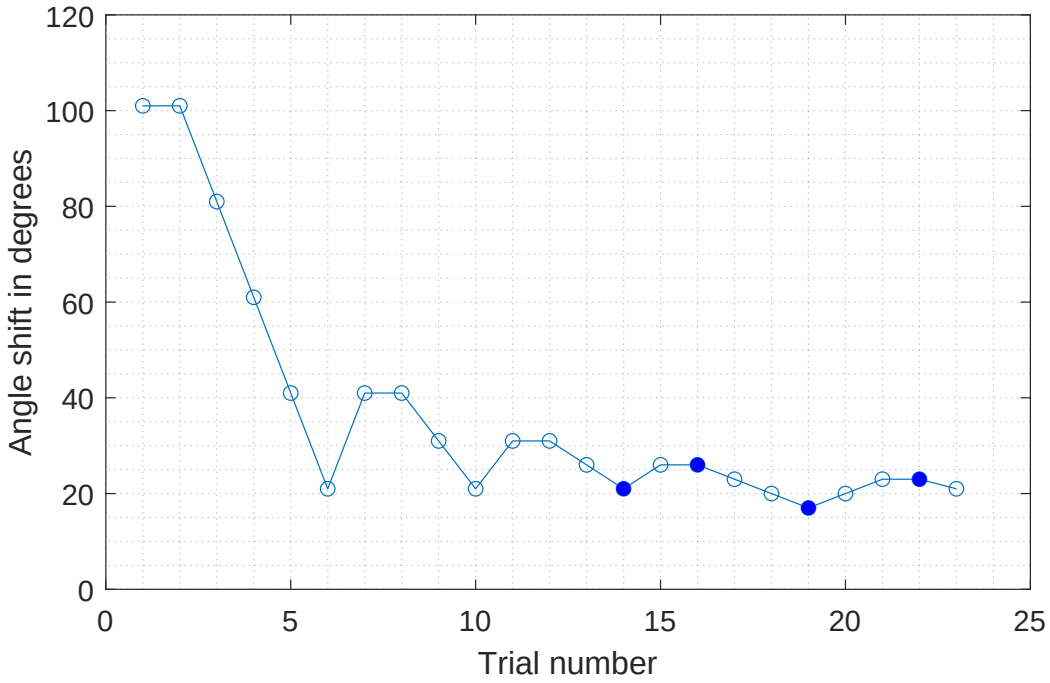

Figure 4.2: Listening test curve for participant 2 for the speech signal.

With 23 trials it is a comparatively short test, so the adaptive method was fast in approximating the threshold and the participants showed very little uncertainties in the answers. More fluctuations and longer tests suggest stronger difficulties of the subjects to manage the task thus making the results less reliable. More about the reasons behind this and the general reliability and reproducibility of the measurements can be found in the discussion.
The range of trials needed to finish the test varies in the range between 23 to 47 trials. The shortest time to complete the test was 4 min and the longest 20 min with a mean time of 11.65 min.
The measurement curves for the drum signal for all participants can be seen in figure 4.3. Again, 17 out of 20 curves are mainly in the range of approximately 0-40° with three exceptions above that range. These three curves also exceed the initial value of 101°, but only in the beginning. This indicates some early uncertainty while the resulting thresholds are below 80° and thus much lower than the initial value. The shortest time to complete the measurement was 6 min, the longest was

30 min and the mean time was 13.7 min. The number of trials needed varied between 22 and 44. The final result of the obtained thresholds for the two signals can be seen

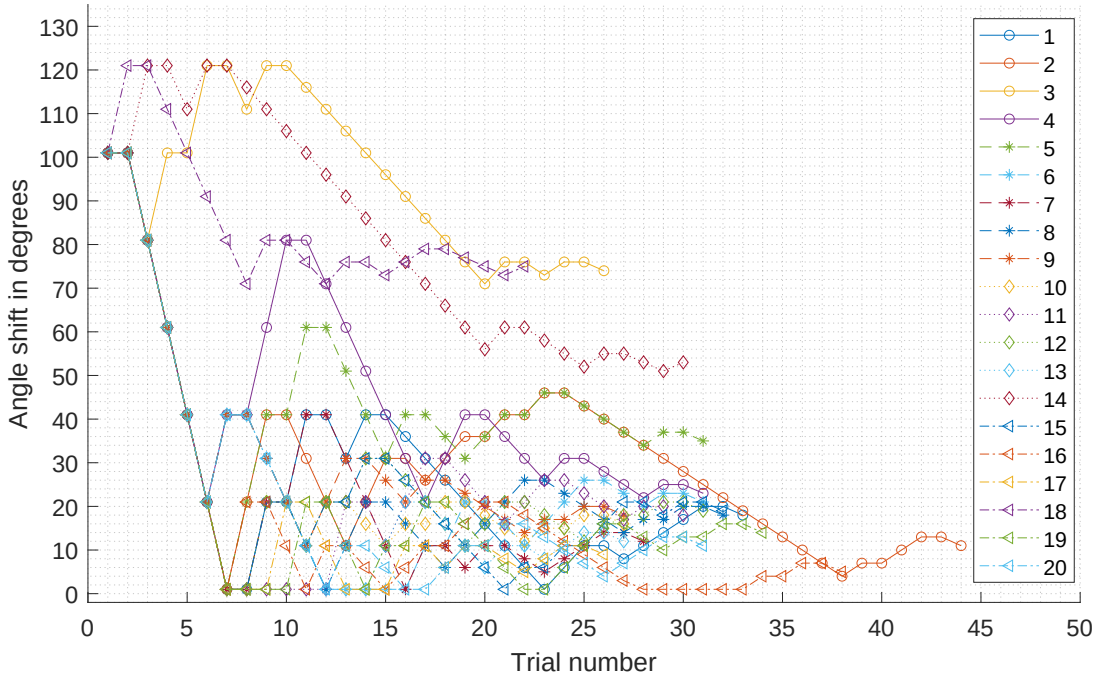

Figure 4.3: Listening test curves for all participants for the drum signal.

in figure 4.4. Here, the median is marked in the box showing the 25th and the 75th percentiles. The whiskers show a range up and down to the extreme values within the range of 1.5 times the interquartile range. Outliers are marked with a red cross. It can be seen that the median values for both tests are very close to each other with 17.125 ° for the speech signal and 16.375 ° for the drum signal. This indicates that the just noticeable difference of angle shift is independent of the test signals used here. However, it must be noted that the angle shifts given here are not the actual angle shifts of all five reflections but it is only the biggest angle shift of the fourth reflection. The other four reflections have, according to their individual step sizes, smaller angle shifts at the threshold. The median angle shifts of all five reflections at the measured threshold are given in table 4.1. When looking at the distribution of the mean thresholds for the 20 participants it can be seen that the percentiles and whiskers for the speech signal are further apart than the corresponding values for the drum signal. This indicates a higher uncertainty during the test with the speech signal. The outliers for the speech signal are also higher than the ones for the drum signal. This fits with the subjective perception of the participants. Several of them reported after the test that they perceived the test with the drum signal as easier.

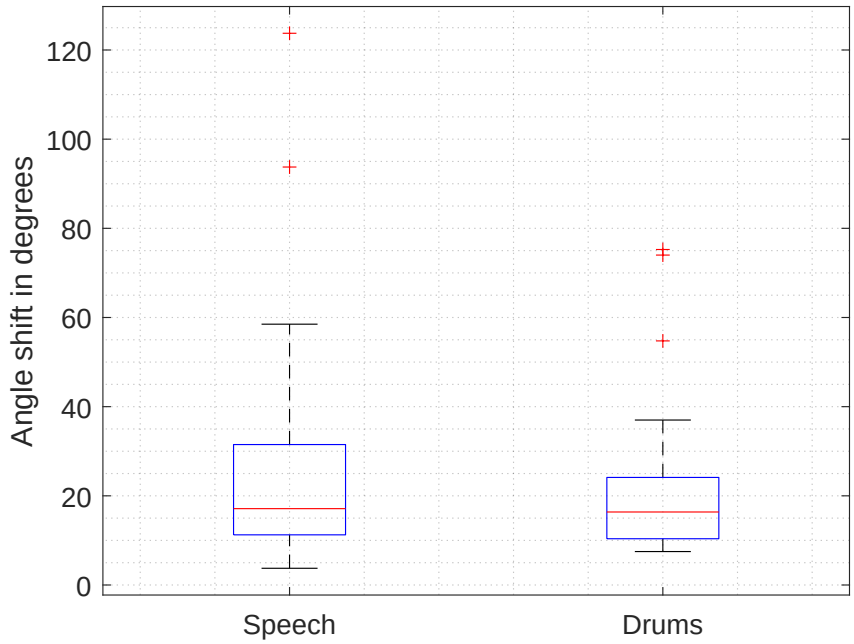

Figure 4.4: Median and the 25th and 75th percentiles of the thresholds for both the speech and the drum signal. The whiskers extend to the extreme values up to 1.5 times the interquartile range and outliers are marked with a cross.

Table 4.1: Median angle shifts of all five reflections for the speech and the drum signal at the measured threshold.

Reflection number	Median shift speech	Median shift drums
1.	4.53°	4.33°
2.	15.11°	14.45°
3.	4.53°	4.34°
4.	17.13°	16.38°
5.	11.08°	10.6°

4.2 Discrimination test

The results of the discrimination test for both the speech and the drum signal showed that 37 out of 40 answers in total have been answered correctly which was the same for both signal types. In order for this result to be statistically significant, at least 25-26 participants were needed to indicate a perceptual difference between the reference and the test signal (see section 3.3.2). Thus, this result is highly statistically significant and shows a clear perceptual difference between the room with the reference first reflection angles and the mirrored reflection angles. Since both signals show the exact same number of correct answers, the kind of test audio signal does not seem to have an influence on the ability to distinguish the version with the reference and the mirrored reflections.

5

Discussion

5.1 Performance of the participants

In the figures 4.1 and 4.3 it can be seen that the the results of the participants vary in a big range. While the angle shift for most of the participants lies in a range of 0-40° during the main part of the experiment, some participants show much higher values. The question is whether these differences are due to actual differences in the auditory system so that they are less sensitive to angle changes or whether another reason can be found to explain this behavior. Firstly it could be that the participants experienced a training effect during the two experiences and therefore got better in the second one. Since the order of the two experiments was random this could maybe explain why there are different people with high angle shifts in the two experiments (except number 18, which has high values in both tests). For this purpose figure 5.1 compares the results of the two experiments divided into the subcategories of the individual first and second experiment to see whether the thresholds of the second experiments (regardless of the audio example) are lower than the one of the first experiment. As mentioned before the order of the experiments was randomized and it turned out to be an exact split of the group, so that half of the group had the speech signal first and the other half the drum signal. It can be seen that the the medians are again very close to each other, the one of the second experiment is even slightly higher. The variance of the second experiment is also bigger as indicated by the larger difference of the percentiles. This implies that the training in the beginning of the experiment was sufficient and thus no training effect influenced the found results.

Related to the training effect the degree of experience with listening tests in general could have influenced the results. 7 of the 20 subjects have never participated in a listening experiment before and thus maybe had a harder time completing the task in this new environment. The participants were only asked whether they have experience with listening tests in general regardless of the kind of listening test and how extensive this experience is. Thus only the factor listening test experience in general can be examined here. Figure 5.2 shows the thresholds for the two audio signals for the two groups of participants with and without experience with listening tests.

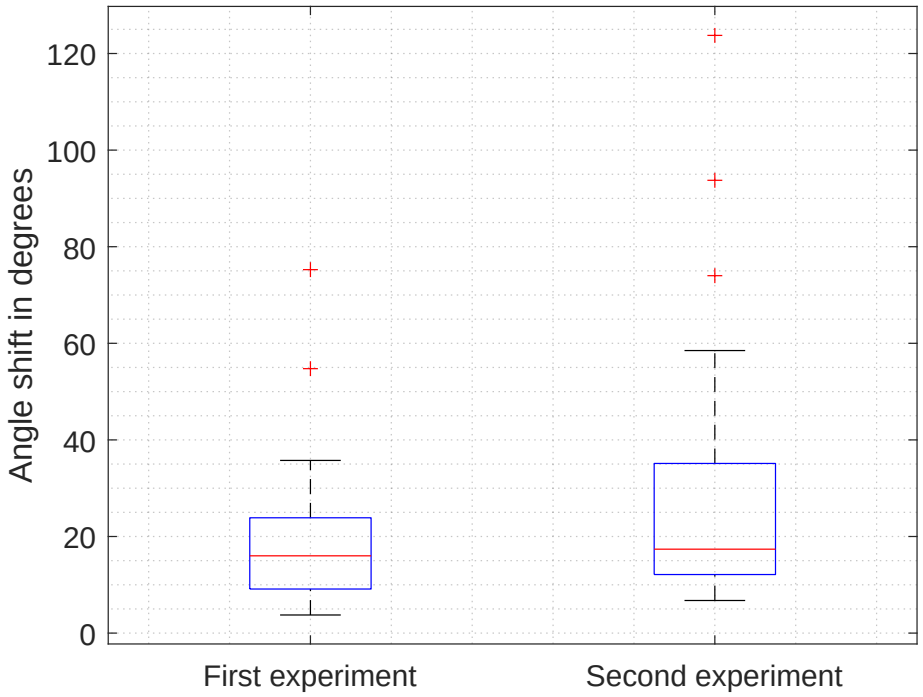

Figure 5.1: Median and the 25th and 75th percentiles of the thresholds for the individual first and the second experiment. The whiskers extend to the extreme values up to 1.5 times the interquartile range and outliers are marked with a cross.

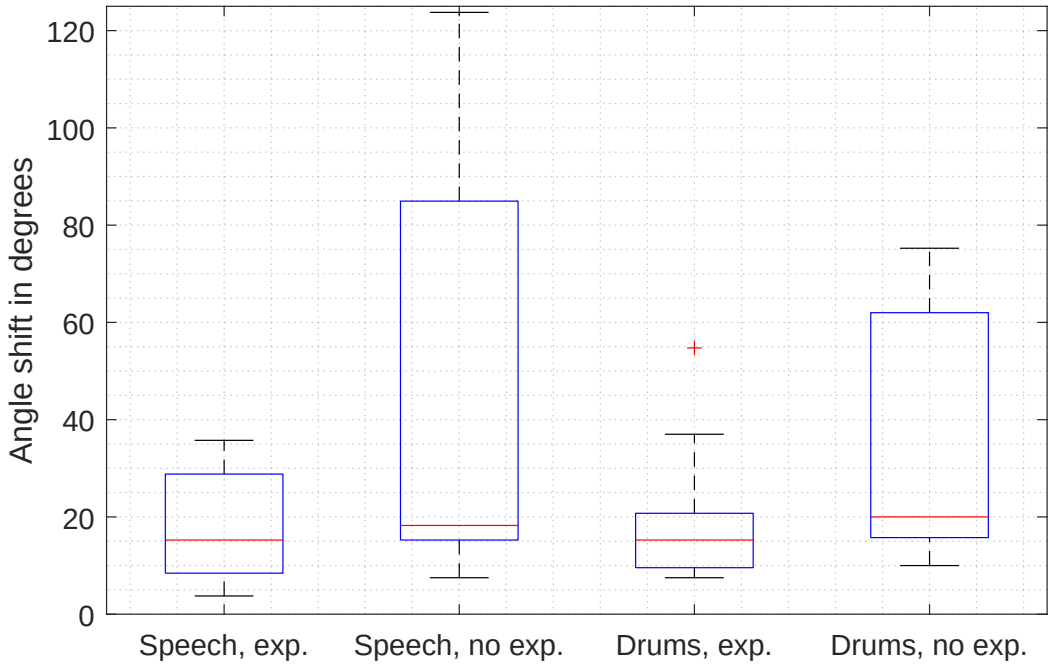

Figure 5.2: Median and the 25th and 75th percentiles of the thresholds for the speech and drum signal for the group of experienced (exp.) and inexperienced (no exp.) listening test participants. The whiskers extend to the extreme values up to 1.5 times the interquartile range and outliers are marked with a cross.

Here a clear difference between the experienced and inexperienced group can be seen. Even though the median is still very similar between the groups it can be seen that the majority of the outliers originate from test subjects in the inexperienced group. On the one side this suggests that experienced listeners generally could hear more subtle differences between the signals than the more inexperienced listeners. The fact that the huge majority of the experienced listeners were students or PhD students from the department of technical acoustics probably also contribute to this picture. Some of the students knew what the test was about before participating in it and thus had a better understanding of what they were supposed to look for in the signals. On the other side, the median and the lower whisker of the inexperienced group is at a very similar angle shift to that of the experienced group which shows that no experience in listening test is required in order to achieve the same result as experienced listeners.

5.2 Influence of the audio signals

It could be seen in figure 4.4 that the difference between the speech and the drum signal was minor even though many participants experienced the test with the drum signal as easier. This could have been caused by the different spectrum and time structure of the two signals. Spectrograms of both signals can be seen in figures 5.3 and 5.4. The speech signal has the most energy in the lower frequency range of below 5 kHz and has some rather stationary parts where there is constant energy in some frequency ranges over time. Generally the speech signal has energy up to 20 kHz, so it covers the whole audible frequency range and thus enables the listener to experience the influence of the room in the whole frequency range while some ranges have more energy than others. Additionally, speech is a very well known signal of the daily life and thus needs no time to getting used to. The drum signal in figure 5.4 shows a more even picture than the spectrogram of the speech signal because the frequency range does not change that much over time. The signal has energy up to approximately 18 kHz with the most energy in the lower frequency range. However, the energy is more evenly distributed over the whole frequency range and the impulse structure of the signal can clearly be seen. Between seconds 5 to 7 is a transition between two different drum parts in the signal which is concentrated in the lower frequency range. From the 7th second on the second section of the drum signal started which has more energy in the middle and higher frequency range than the previous part.

From looking at the spectrograms of the used audio examples it is clear that the main difference between the speech and the drum signals is the more impulsive structure of the drum signal while having a more evenly distributed energy all over the frequency range.

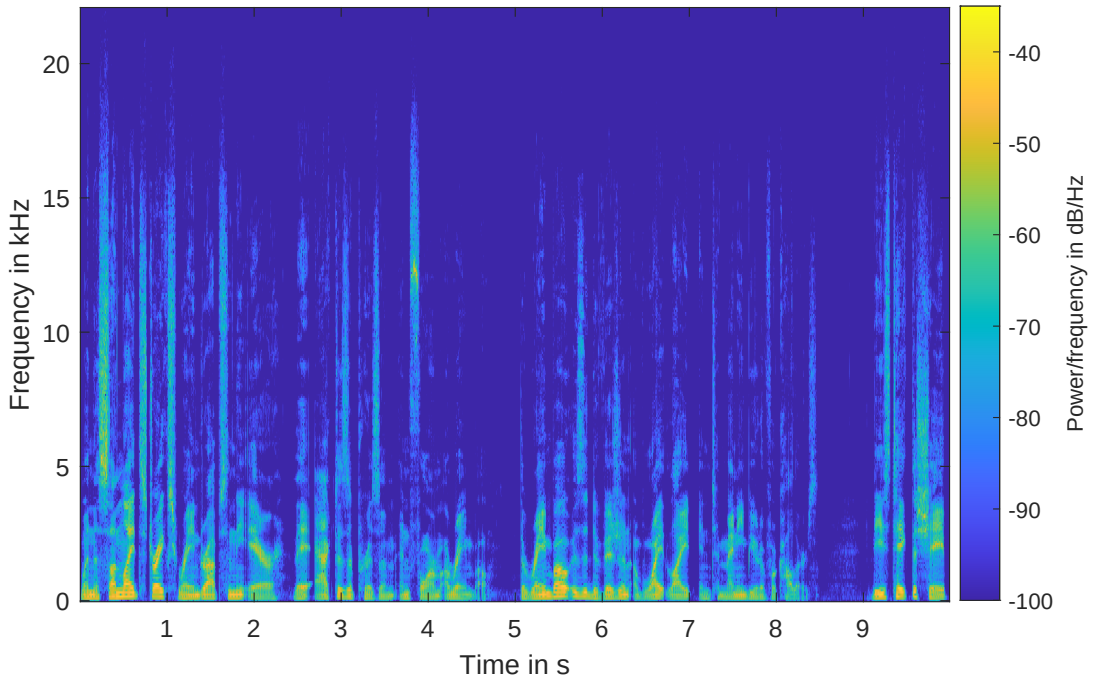

Figure 5.3: Spectrogram of 10 s of the speech signal.

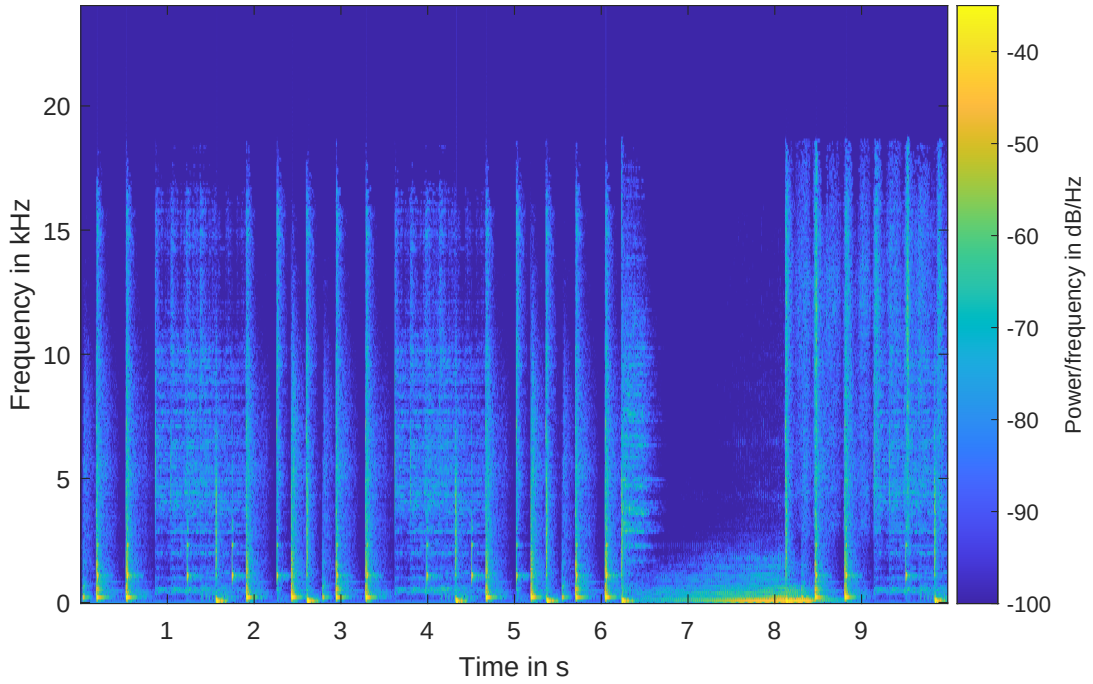

Figure 5.4: Spectrogram of 10 s of the drum signal.

Just like the speech and drum signal, the room impulse response and in that the reflections have a spectrum as well. It makes sense that every single reflection is only audible if the exciting signal has energy in that frequency range. If, for example, a

wall only reflects frequencies above 10 kHz and absorbs everything below that, no reflection will be audible if the exciting signal is a sinusoidal signal with a frequency of 200 Hz. Thus it is interesting to see in what magnitude response the reflections have in order to see whether some of them are less excited than others or not at all. Thus the importance of every single reflection in combination with the used excitation signals can be examined. In order to do so a method described in [16] to estimate the reflection filters will be used. Here, only a short description of the method will be given, for further information see [16].

The difficulty in describing the frequency components of reflections is the short time of the signal. In order to analyze frequencies down to 100 Hz a window length of 10 ms around the reflection is required, while the actual reflection is usually not longer than a few ms maximum. This means that other signal parts are included in the 10 ms window and thus results in an inaccurate reflection filter. To avoid that, this method uses three asymmetric windows to include pre-ringing artifacts before the main peak as well as at least 1 ms after the main peak to include the region of summing localization. The first one has a size of 1.5 ms. The size of the second one is between the first and the third und thus depends on the third and longest window, which itself is determined by the lower frequency limit that should be observed. If a lower frequency of 100 Hz is required, the window sizes of the second and third window could be approximately 2.6 ms and 10 ms. With these three windows, the magnitude response can be calculated by using the information obtained by all three windows so that most reflections can be described correctly in the frequency range of above 667 Hz (limit for the shortest window of 1.5 ms). Since the spectrum of the reflections depends on the exciting source, the resulting spectra are filtered with the inverse spectrum of the direct sound (obtained in the same way as the reflection spectra). In a final step the filters are octave-smoothed. Even though the method can satisfactory described some reflections, the inaccuracies increase for reflection at later points in the RIR since the reflection density increases and thus the signal parts in the windows that are not a part of the reflection under examination.

The normalized reflection filters for the first five reflections of the RIR used in this work can be seen in figure 5.5. It can be seen that all five reflections have a drastically decreasing magnitude spectrum above approximately 8 kHz. The fifth and the fourth reflection seem to be quite strong since they have a high magnitude spectrum for frequencies below 8 kHz with fluctuations between 0-5 dB only. The shape of the second, third and fourth reflection are quite similar since they have more reflective in the range of 400-1500 kHz while being more absorbing in below that at around 200-300 Hz and above that range at approximately 2-3 kHz. However, the second and third reflection have the highest reflection at low frequencies the middle and high frequency range is shifted to lower magnitudes by around 5-10 dB. The first reflection seems to be very reflective in the low frequency range and less so in the remaining frequency range. However, as mentioned before, the results are less reliable below around 677 Hz, so one should be careful with analyzing the results in the low frequency range.

Generally, it can be seen that all reflections should be sufficiently excited with both the speech and the drum signal to be be audible (assuming the level is high enough). The more even distribution in energy and especially the impulsive structure of the

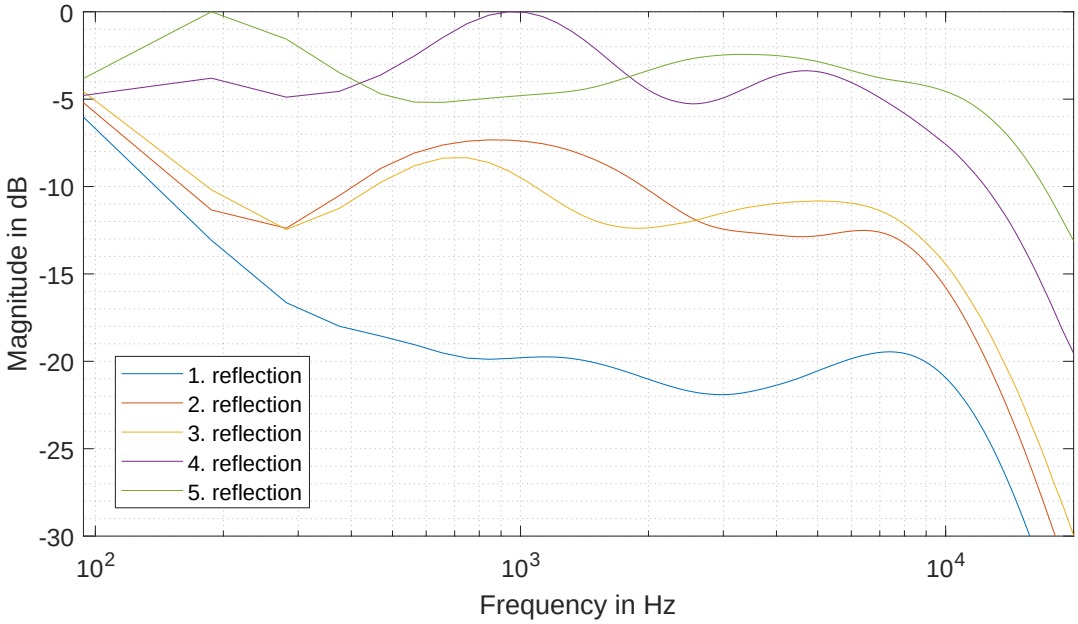

Figure 5.5: Estimated reflection filters of the first five reflections.

drum signal could have helped the participants in perceiving the task are slightly easier. But overall the spectrograms of the two audio example as well as the five reflection filters are quite similar so that no significant difference in the excitation of the individual reflections with the two audio examples can be suspected.

5.3 Analysis of the found thresholds

In order to understand which effects are caused by the first reflections that could lead to the thresholds found, it is useful to first look at the individual reflections in relation to the direct sound and then try to use this information and similar measurements to explain the thresholds and establish connections to other work. This is especially useful because there is already a lot of research on the perception of individual reflections (for example [1], [3], [18]), which can give first hints. To evaluate all five reflections directly in context is very difficult. Although there are studies with several reflections, which were also partly examined in the context of real impulse responses ([2], [4]), one can read off by these studies as good as no general rules. This is due to the already described influences of many factors like delays and levels of the reflections among each other, which make such measurements very specific and generalizations therefore extremely difficult.

Starting with the analysis of the individual reflections, a good way to begin is to look at the level of the five reflections and compare these levels to the one of the direct sound to figure out whether the reflections are audible or masked by the direct sound or any of the other first reflections. The relative level of the reflections can be seen in figure 5.6. It is interesting to see that the relative level of the first reflection

is the smallest with a value of -12.7 dB. The level of the other reflections increases monotonously with increasing number of the reflection with values of -11.3 dB for the second reflection, -9.7dB for the third, -7.7 dB for the fourth and -5.5 dB for the fifth reflection. These relative levels can now be used to evaluate the audibility of the reflections. In [18] the authors examined the audibility threshold of single

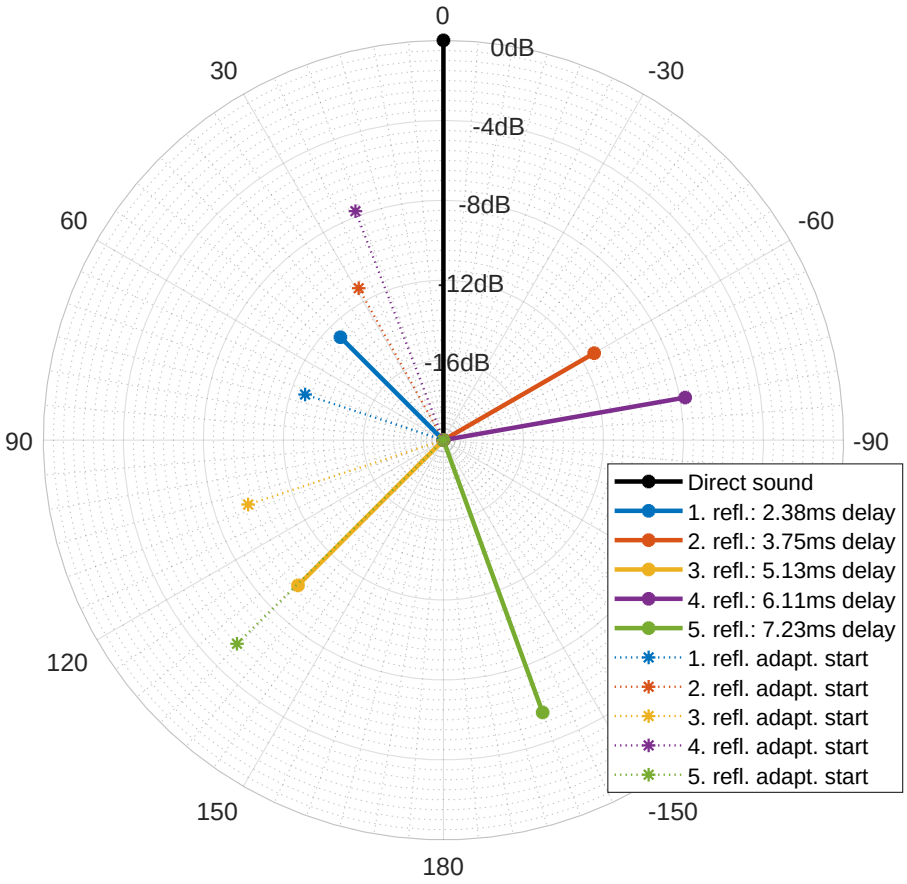

Figure 5.6: Reference angle of incidence of the direct sound and the five first reflections with their relative level to the direct sound. The dotted lines show the angle of incidence of each reflection at the beginning of the adaptation process.

side reflections with different delays and incidence angles. They observed that the thresholds for all incidence angles decreased monotonously with increasing delay between the direct sound and that reflection due to masking effects of the direct sound. Regarding the angle of incidence they found that the highest threshold is found when the reflection comes from the front, so the same direction as the direct sound. The thresholds are then decreasing with increasing angle towards 90°. They also found a front and back symmetry which means the thresholds are again increasing in the range from 90° to 180° in approximately the same manner as they did in the frontal half plane. In order to estimate the audibility of the reflections in this work, values from the measurements at 10 ms are used which are the shortest delay times that have been examined in [18]. With the respective angles of incidence, the approximate thresholds for the five reflections can be derived and are about -

17 dB, -18 dB, -17 dB, -19 dB and -12 dB. However, it must be noted here that especially the early reflections have a much bigger time delay than the 10 ms that we used for these values. Thus the threshold values are probably between 1-3 dB higher than given here. Also, the calculated values here were measured for only the direct sound and one reflection. Additional early and late reflections result in an increase of the threshold, again due to masking effects [18]. The authors in [2] observed an increase in the audibility threshold when measuring in rooms with some reverberation and an even bigger increase in rooms with strong early reflections and reverberation. However, the increase was mainly found at time delays greater than 10ms. Thus the derived thresholds from [18] are assumed to be a sufficient approximation of the true thresholds. Comparing the relative values of the five reflections with their approximated audibility thresholds it can be seen that all five reflections should lay above their thresholds and thus be audible.

Besides the audibility threshold other mechanisms for example the temporal structure can have in influence on how important a single reflection is. In [5] the authors examined the perceived weighting of individual clicks in a click train with different delay times. When the clicks had gaps of more than 5ms the clicks were equally weighted. For smaller gaps the first click was more strongly weighted than the following ones. Transferred to this case it means, that a reflection after a longer gap could be perceived more intensely than a reflection of the same level and incidence angle that follows another reflection with a shorter gap between. In our case, the first reflection after the direct sound has a delay of approximately 2.4ms. All the other reflections follow with gaps shorter than 2ms. This means that none of the first reflections appears after a bigger gap and thus all reflections should have the same intensity weighting with regard to the temporal structure.

With the assumption that all five reflections are above the individual audibility threshold and they have an equal temporal weighting, the individual effects of every single reflection only in combination with the direct sound can be examined. These are likely to have some influence on the overall perception of the room. All reflections appear in the frame between 1ms and 50ms, so they are in the time frame of the effect of the law of the first wave front. So the localization of the sound source should be determined only by the direct sound direction for all direct sound and reflection combinations. Thus it is very likely that also the room with the shifted reflections should always result in the same localization of the sound source. I also means that the reflections are not perceived as echos either, since they appear very early and thus far before the approximate time of the echo threshold is reached. This results in the assumption, that the reflections can for example influence the spatial extend, tone color or the center of gravity [1]. This is somewhat in accordance with the subjective impression of the listening test participants. They mainly perceived differences in the center of gravity and the localization of the auditory event. First of all it is important to understand that the participants were not asked to rate the impression at different points throughout the experiment and that not all participants have been asked at all. This means the subjective impression may have been different for some of the participants and also that this perception could have changed during the experiment. Maybe the perceived difference between the rooms in the

beginning of the adaption process was a different one than in the end. However, the perceived difference in the center of gravity seemed to be an impression that many subjects had and which fits to the theory. The change in localization, on the other hand, is an effect that was not anticipated. There are two possible explanations for this. Firstly, it could be that the very early side reflections (much closer to the region to summing localization than to the echo threshold) could have influenced the localization of some subjects. Secondly, the actual perceived difference was the center of gravity and the subjects verbally expressed this perception as a change in localization possibly due to missing knowledge about the terminology of different effects. This second one is the more likely theory in the author opinion since the subjects that perceived a localization difference were all inexperienced listeners and thus had no knowledge about acoustic terminology and were generally insecure how to express their perception correctly. Generally it can be concluded, that the main perceptual change between the rooms with the original and the shifted reflections was a shift of the center of gravity, while a change of the localization can not be ruled out.

The thresholds that were found in the experiment lay in a range between 4.3° and 17.1°. The 25th and 75th percentiles of the biggest angle shift lay in the range of approximately 12°-32° for the speech signal and 10°-24° for the drum signal. To the authors knowledge there is only one paper that did a comparable examination. In [4] the authors obtained the maximum angle shift of a single side reflection in a RIR that was not distinguishable from the original reflection in that room. This single reflection was at an angle of approximately 70°, so at a region with a bigger localization blur. It was found, that the reflection could be shifted by around 12° to 35°. Even though only one reflection was examined here, it can be compared with this work since the angle is (even though mirrored) very close to the fourth one in this work with an angle of -80°. This is also the reflection with the biggest step size and thus has the biggest angle shift remaining at the threshold. It is therefore assumed that this reflection has a direct influence on the found threshold since the change that this angle shift causes the perceptual difference is greater than for example the first reflection with a median angle shift of around 4.3-4.5° (for simplicity, the following ranges include the results with both test signals, since they are so similar). Also, the found angle shifts between the percentiles of 10°-32° in this work is very similar to the range of 12° and 35° for one lateral reflection in [4]. The median value is with 16.4°-17.1° a bit smaller than the ones in [4] with -22°. This could be due to the combined effect of multiple shifted reflections here that add to the perceived difference and thus can indicate that several reflection can be shifted less than a single one to perceive a difference. This would be interesting to test in new experiments with different angles. However, the comparably lower results found here could also have been caused by a difference in the test design of the two studies and thus in a difference of the convergence point on the psychometric function. This will be discussed further in section 5.6. In this case the lower thresholds here could merely be the reason of a lower convergence point on the psychometric function in this experiment and the results for the same could could be identical. This would indicate that the fourth reflection with the biggest angle shift is the one that

determined the threshold without an influence of the other four reflections. Further tests are important to examine this.

5.4 Analysis of the discrimination task

The discrimination task was especially interesting because of the constant total lateral energy. Even though the reflection angles were mirrored the total spatial information in the signal was still the same. In [3] the authors concluded from their experiments that the spatial impression seems to be determined by the ratio of lateral to direct sound. In their experiments they changed the level of one and multiple reflections and found that the level of a single side reflection needed to be higher than the ones of two side reflections in order to produce the same spatial impression. Thus, if the spatial impression was the only way or was at least a high contributing factor in these experiments, the participants should have had much more trouble to distinguish the rooms with mirrored and original angles. But since the signals were clearly distinguishable as well as none of the participants mentioned perceived changes of the spatial impression, it does not seem to be a significant factor in these experiments. As mentioned before, this can also be due to the design of the experiment, because the reflections start the adaptation process from one side rather than from the front, so that some degree of spatial information is given throughout the experiment. However, this does not mean that the spatial impression did not change in the experiments at all, but that other perceptional changes were more important in the distinction process.

5.5 Relevance and limitations of the experiments

As already discussed, no general statements about the perception of angle shifts of reflections in an impulse response can be made on the basis of this work. Only a very specific case was examined in this work, one impulse response in which the first five reflections have been picket and assigned to random azimuth angles while their levels have been maintained. The real impulse response with its energy distribution and temporal structure can create a realistic situation, while the change of azimuth and elevation allows to study the influence of lateral reflections in interaction. Thus, this experiment can provide an interesting insight into what perceptual effects lateral reflections can have in context. But this does not mean that the found thresholds are directly transferable to real world scenarios. A threshold means that a person is just able to perceive a difference between two signals while the focus is lying on detecting that difference. It does not mean that, if we were to reproduce a room and changed the angles with the here found threshold, a person would notice a difference when they are not trying to. This also means that a person who is in a room where reflections are suddenly changed with the threshold found here would notice these changes when the attention is focused on something else. It is also different from so-called plausibility tests, which check whether a representation of a space seems plausible to a person. Here, the angles could presumably be changed significantly more until the room is considered an implausible reproduction. So it

is important to understand that this experiment is just a small insight into a huge area of psychoacoustics with a lot of different effects interacting. This area is just in the beginning of being researched. To give an idea of how future work in this area could look like, the main limiting aspects of this work will be discussed in the following.

First of all, only one RIR was used in this work which has a very short reverberation time. Studies of for example the audibility threshold of single reflections showed a strong influence of the thresholds with the reverberation time [2]. Thus experiments with different rooms and reverberation times are absolutely necessary. A second important factor are the levels of the reflections and as well their time structure. These were kept original in this work but compared to other authors who examined the effect of reflections the delay time of the first reflections are very short with delays below 10ms for all five reflections. This temporal structure as well as the levels of the reflections determine the occurring masking effects and thus the audibility of the single reflections. And the audibility of the reflections is of course crucial to examine the effect of the incidence angle of the reflections on the way we perceive a space. The results of the authors in [4] illustrate this. They found that the original level of the first reflection with a delay of around 20ms had to be increased by several decibels in order for the test subject to hear a difference of the RIR with and without the reflection. This means the first reflection in this RIR is without modifications not loud enough to be heard at all, which means it has no influence in the way the space is perceived. Also the significant difference between the found audibility thresholds found for anechoic ([18]) and real environments ([2]) show that it is important to include reflections of different level in future examinations. Besides just making sure that the reflections are above the threshold, the levels also determine masking effects as well as the strength and the way the angles of incidence associated with the reflection affect the spatial perception.

Of course, an extremely important point of the experiment is the choice of the angles of incidence of the reflections. Here, they were chosen largely at random, with care taken to ensure that the reflections were distributed somewhat along the entire horizontal plane in order to avoid a strong weighting of one side. But, as explained in the theory, the auditory system is not equally sensitive to angle shifts over the entire horizontal plane. The localization blur for sound incidence from the sides is much higher than from the front. Logically, if a reflection is placed around 90° instead of 20°, for example, this has an effect on the maximum possible angle shift of that reflection without causing a change of perception. In this context, this could mean that one or more reflections have no influence at all on the threshold found, since they are in a range of lower sensitivity and therefore no difference is perceptible. Although it was possible for the subjects in this study to move their heads and thereby shift the reflections to areas of higher sensitivity, only a few did so frequently. Furthermore, moving the head could cause other reflections to move to less sensitive areas. It would therefore be very interesting to see how the threshold measured here would change if one or more reflections were located in less or more sensitive areas.

The biggest peculiarity of this experiment, however, is the way the reflections are shifted. This was done for all reflections at the same time, but not with the same

step size. To the best of the author's knowledge, no experiment has ever been performed where multiple reflections were shifted in such a way as to seek the maximum possible shift that would produce a perceptible difference. The starting position of the adaptation process is also unusual. The fact that the reflections are lateral at the beginning of the adaptation process instead of coming from the front, for example, prevents the pure increase of spatial information, i.e. sound incidence from directions other than the frontal incidence direction, from leading to a change in perception. But since here the reflections already have many lateral incidence angles at the beginning, the spatial impression does not change that much. This was also reported by many test persons, who felt the change between the signals more as a weighting inside the room than the spatial impression of the room itself. However, this effect should be investigated in more detail through experiments. For this purpose, it would be useful to carry out a similar experiment with different adaptation starting points. In any case, the subjectively perceived differences between the signals should also be queried, preferably between each adaptation step. In this way, it can be determined which effects usually occur with which combination of angles. Also experiments would be interesting, where all angles are shifted with the same step size, or also start the adaptation process from different directions.
All the above limitations except the ones regarding the angle shift are true for the discrimination test as well. This test is especially interesting regarding the plausibility of the reproduction. Even though a clear difference was spotted between the original and the mirrored case, it makes sense that also the mirrored room should be a plausible room. Here, it would be intriguing to examine the combination of the reflection angles with the visual position of the person in a room. Especially the transition from a plausible to an implausible reproduction, when the person moves towards a wall, but the reflection does not come from the wall, but from the other, mirrored side, would be exciting to investigate. This could give more hints to how the visual and the acoustical information work together and why and when there is a mismatch between the information that leads to the conclusion a reproduction, either the visual or the acoustical, is not plausible anymore.

5.6 Effects of the design

The test design has a big influence on the results of an experiment. It does not only determine what we actually measure (for example different points on a psychometric function), but also whether there can be systematic error or biases in the measurements simply because of the procedure. There are a lot of factors in the design that can influence the outcome of the experiment, some on larger scale, for example what kind of test was chosen, but also on the smaller scale such as the decision what step size to use. Here, the consequences of the decisions taken with regard to the correctness of the results themselves, the reliability of them and the possibility to compare the obtained thresholds with other work will be discussed.

The main effect of choosing an ABX threshold test is that the process will converge to one specific point on the psychometric function, in this case the 70.7% point. This is very important for comparing the results with related work since the angle shift

will be different for a different point on the psychometric function. The threshold for the 80% point, for example will be higher than the one for the 70.7%, because the requirement of how reliable the signals have to be distinguished is higher for a higher point on the psychometric function. As described in section 2.4, the convergence point is determined by the choice of the transformed staircase method. Therefore the obtained thresholds here can only be compared to other work to a limited extend if they used a different test design because their convergence point on the psychometric function is very likely to be different from the one here. This also applies for the comparison that was made in the discussion when comparing the threshold with the findings in [4]. They used a 3-up-1-down method which results, in combination with their chosen step size, at a convergence point of 73.03%, so the results can not be compared directly. Important in this context are additionally the findings from [14]. They showed that the convergence point of transformed staircase methods are not only dependent on the chosen method (for example 1-up-2-down) but also on the step size ratio between the UP step and the DOWN step as well as dependent on the spread of the psychometric function. In this case of a 1-up-2-down method with an equal step size of both steps, the true convergence point decreases with increasing spread of the psychometric function from the assumed 70.7% point down to around 60% for a widely spread psychometric function. Since the width of the psychometric function is not known for this effect, the deviation from the assumed convergence point is unknown. It is, however, very likely that the true convergence point is smaller than assumed here. To avoid this, the authors in [14] give step size ratios for different methods that can be used to target specific point on the psychometric function. Also, the averaging process to obtain the threshold from the measurement curves has an influence on the resulting threshold. Here, it was decided to use the average over the last four reversals. Generally, the more averaging points the more accurate is the estimation but in this case the number was not higher because of the strong fluctuations in the first half of the experiments which would have resulted in much higher thresholds for some participants which did not seem reasonable compared to the second half of the experiment. Here, also the decreasing of the step size plays a role. In this work the step size was halved when an upper reversal was found. This way, an average over an even number of reversals was necessary to make up for this. Having in mind all the factors explained here, it is clear that the final threshold obtained from an experiment is dependent on many factors and that it is important in future work to choose the variables accordingly so that the results are representative and comparable to related work.

Another decision that not only has a direct influence on the results but also on the test participants is the number of reversals that is used before the test is finished. Usually a minimum number of 6 to 8 reversals is recommended to get reliable threshold results[15]. Of course, the more reversals there are the more accurate the results will be, but more reversals will also result in a longer test and thus be annoying for the participants. Also, a threshold test can be very frustrating because the task seems to be impossible to solve which can result in slackening concentration and thus mistakes and higher thresholds. With the chosen 8 reversals some participants already needed a long time of up to 47 minutes and reported to be very frustrated. Early uncertainties also lengthen the test because of the decreasing step

size at upper reversals, so test subjects that face uncertainties early on even have to do a longer test which could make them even more uncertain and increasing the frustration level. This did not seem to show in the results, however, it is important to keep in mind that the environment and level of comfortless or rather the lack of the latter, can lead to non-representative results.

6

Conclusion

The purpose of thesis was to find the maximum azimuth angle shifts of five early room reflections that just lead to a change in perception of that room. To obtain this threshold an adaptive ABX threshold listening test was conducted with 20 participants. During the adaptation process the azimuth angles were shifted closer to their original position in case of correct answers and further away for wrong answers and thus converging to the point of just noticeable difference. The possible range for the adaptation process lay between the maximum shift where all five reflections come from the left side and their reference position. The mean of the stimulus level at the last four reversals for each person were used as the threshold estimate. It was shown that there was nearly no difference in the found median threshold for the 20 participants between the speech and the drum signal, even though the speech signal had more outliers suggesting that reflections in more impulsive signals might be easier to hear, even though the actual threshold is at the same angle shift. It was also shown that inexperienced listeners were responsible for almost all outliers in the measurement which leads to the assumption that inexperienced listeners (i.e. the majority of the population) probably tend to be less sensitive to angle shifts than the median found here suggests.

The found median thresholds for the five reflections lay in a range between 4.3° for the first reflection and 17.1° for the fourth reflection and the reported main perceived difference between the rooms with the reference and shifted reflection angles was a change in the center of gravity even though some participants reported a change in localization as well. The found maximum shift of 17.1° is approximately 5° smaller than the found threshold in a similar experiment for one shifted reflection. The could mean that effects of several reflections with smaller angle shifts somehow add up in the auditory system thus leading to a changed perception even though anyone of the shifted reflections alone would not yield to a perceptual change. However, this difference could also have been caused by the test design only, resulting in the second assumption that the threshold found here is defined only by the reflection with the biggest angle shift. However, the angle shifts found here represent the highest stimulus intensity where no difference between the original and the adjusted room impulse response could be heard, which does not mean that a room with a greater angle shift of the reflections can not still be rated as a plausible reproduction of that room. This is an important note regarding the relevance of the found thresholds and the initial motivation for this thesis. In order to examine how much the angles can be changed without disturbing the listeners experience other tests, for example plausibility tests, are needed.

In summary, the joint perception of several reflections in a room impulse response depends on many factors, such as the number of reflections, their time delays to the direct sound as well as between the reflections themselves, their relative levels, the angles of incidence and many other factors. Furthermore, these factors can cause a lot of different perceptual effects while some of them can be more important in the question of how much of an angle shift of reflections can be tolerated from the listener. A general statement about the resulting perception of multiple reflections in room impulse responses is therefore extremely difficult and requires much further research in this area.

Bibliography

[1] J. Blauert, *Spatial hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 1996. DOI: 10.7551/mitpress/6391.001.0001.

[2] S. E. Olive and F. E. Toole, "The detection of reflections in typical rooms," *Journal of the Audio Engineering Society*, November 1988.

[3] M. Barron, "The subjective effects of first reflections in concert halls—the need for lateral reflections," *Journal of Sound and Vibration*, vol. 15, no. 4, pp. 475–494, April 1971, ISSN: 0022460X. DOI: 10.1016/0022-460X(71)90406-8.

[4] O. C. Gomes, N. Meyer-Kahlen, W. Lachenmayr, and T. Lokki, "Perceptual consequences of direction and level of early reflections in a chamber music hall," 2022.

[5] J. Blauert and J. Braasch, Eds., *The Technology of Binaural Understanding*, Modern Acoustics and Signal Processing, Cham: Springer International Publishing, 2020, ISBN: 978-3-030-00386-9. DOI: 10.1007/978-3-030-00386-9.

[6] F. Brinkmann, H. Gamper, N. Raghuvanshi, and I. Tashev, "Towards encoding perceptually salient early reflections for parametric spatial audio rendering," *Journal of the Audio Engineering Society*, May 2020.

[7] J. Ahrens, "Auralization of omnidirectional room impulse responses based on the spatial decomposition method and synthetic spatial data," in *ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom: IEEE, May 2019, pp. 146–150, ISBN: 978-1-4799-8131-1. DOI: 10.1109/ICASSP.2019.8683661.

[8] L. Müller and J. Ahrens, "Perceptual differences for modifications of the elevation of early room reflections," *Journal of the Audio Engineering Society*, August 2022.

[9] L. Müller, "Perceptual differences caused by altering the elevation of early room reflections," Master's Thesis, 2021.

[10] S. Tervo, J. P. Tynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *Journal of the Audio Engineering Society*, vol. 61, no. 1, 2013.

[11] S. Tervo, *SDM Toolbox*. (https://www.mathworks.com/matlabcentral/fileexchange/56663-sdm-toolbox), MATLAB Central File Exchange. Retrieved August 26, 2022.

[12] S. V. Amengual Garí, J. M. Arend, P. T. Calamia, and P. W. Robinson, "Optimizations of the spatial decomposition method for binaural reproduction," *Journal of the Audio Engineering Society*, vol. 68, no. 12, pp. 959–976, January 14, 2021, ISSN: 15494950. DOI: 10.17743/jaes.2020.0063.

[13] H. Fastl and E. Zwicker, *Psychoacoustics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, ISBN: 978-3-642-51765-5 978-3-540-68888-4. DOI: 10.1007/978-3-540-68888-4.

[14] M. A. García-Pérez, "A cautionary note on the use of the adaptive up–down method," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. 2098–2107, October 2011, ISSN: 0001-4966. DOI: 10.1121/1.3628334.

[15] H. Levitt, "Transformed up-down methods in psychoacoustics," *The Journal of the Acoustical Society of America*, vol. 49, no. 2, pp. 467–477, February 1971, ISSN: 0001-4966. DOI: 10.1121/1.1912375.

[16] J. M. Arend, S. V. A. Garí, C. Schissler, F. Klein, and P. W. Robinson, "Six-degrees-of-freedom parametric spatial audio based on one monaural room impulse response," *Journal of the Audio Engineering Society*, vol. 69, no. 7, pp. 557–575, November 11, 2021, ISSN: 15494950. DOI: 10.17743/jaes.2021.0009.

[17] J. A. Bierer, S. M. Bierer, H. A. Kreft, and A. J. Oxenham, "A fast method for measuring psychophysical thresholds across the cochlear implant array," *Trends in Hearing*, vol. 19, December 29, 2015, ISSN: 2331-2165, 2331-2165. DOI: 10.1177/2331216515569792.

[18] X. Zhong, W. Guo, and J. Wang, "Audible threshold of early reflections with different orientations and delays," *Computers, Materials & Continua*, vol. 61, no. 3, pp. 18–22, 2019, ISSN: 1546-2226. DOI: 10.32604/sv.2018.03900.

[19] J. Boley and M. Lester, "Statistical analysis of ABX results using signal detection theory," *Journal of the Audio Engineering Society*, 2009.

DEPARTMENT OF CIVIL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY