Perceptual Differences Caused by Altering the Elevation of Early Room Reflections

Master's Thesis in MSc Programme Sound and Vibration

LEON MÜLLER

DEPARTMENT OF ARCHITECTURE AND CIVIL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2021
www.chalmers.se

Master's thesis 2021

Perceptual Differences Caused by Altering the Elevation of Early Room Reflections

Supervisor & Examiner: Jens Ahrens, Division of Applied Acoustics

Cover: Loudspeaker array setup for one of the performed listening experiments.

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Architecture and Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
Gothenburg, Sweden 2021

Perceptual Differences Caused by Altering the Elevation of Early Room Reflections

LEON MÜLLER
Department of Architecture and Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology

Abstract

In recent years, spatial audio became more and more relevant for consumer applications such as immersive multimedia playback and video games. Thereby, an ongoing challenge is to authentically reproduce the spatial auditory impression of acoustic spaces. In this context, it has not yet been investigated how accurately the vertical positions of early reflections need to be reproduced in order to result in a plausible spatial room impression.

The aim of this thesis is to evaluate the ability of the human hearing system to distinguish between early reflections with different elevation angles as well as to quantify the auditory differences caused by changing an early reflection's elevation. Therefore, impulse responses of an isolated ceiling reflection were measured and, in combination with spatial impulse responses of two different rooms, used to perform both loudspeaker- and headphone-based listening experiments. The results of these experiments show that changing an early reflection's elevation angle can lead to clearly perceivable differences depending on the reflection strength, the reflection position, the change in elevation, the acoustic environment and the source signal. Thereby, the loudspeaker based reproduction method with custom HRTFs resulted in larger perceived spatial differences than the headphone based method using generalized HRTFs. Furthermore, it was found that the change in interaural cross-correlation caused by altering a reflection's elevation angle is strongly related to the amount of perceived spatial differences.

The outcomes of this thesis suggests that, under certain conditions, the vertical positions of early reflections need to be reproduced in order to achieve an authentic spatial room impression. On the other hand, it was shown that compressing all elevated reflections of a spatial room impulse response to the horizontal plane does not result in perceivable differences if the SRIR is sufficiently diffuse and does not contain pronounced ceiling reflections.

Keywords: spatial audio, auditory perception, psychoacoustics, elevated reflections, spatial decomposition method, room acoustics, spatial impulse responses, loudspeaker array, microphone array, direction of arrival

Contents

1

Introduction

Since the middle of the last century, stereophonic sound reproduction is the standard for consumer audio applications and for a long time, surround sound systems could only be found in cinemas and ambitious home theaters. However, in recent years more and more consumer products like soundbars or even desktop computers implemented multi-speaker setups that promise some kind of immersive audio experience. Simultaneously, headphone based spatial audio reproduction gained in relevance as virtual reality headsets, headphones with integrated tracking systems and game consoles with binaural audio support became accessible to the mainstream consumer market. Thereby, one ongoing challenge is to plausibly reproduce the spatial audio impression of real or simulated acoustic spaces. I.e. for video games, it is desirable that the sound scene a user perceives via headphones matches the environment simulated in the game, including an authentic spatial reproduction of all sound sources and room reflections.

In order to reproduce such a sound field at a specified position in a room for an arbitrary source signal, the room's impulse response, i.e. the transfer path from a source position to a receiver position, has to be measured or simulated. For monaural or stereo applications, this procedure is rather simple and can be conducted by measuring the room response with one or two microphones. For spatial audio on the other hand, not only the time and frequency structure of the room response but also the exact incident directions of different room reflections are important. Such a spatial room impulse response can be measured with a microphone array, consisting of four or more microphones. Depending on the used processing method, a general rule of thumb is that a high spatial resolution of such a measurement requires a large number of microphones which is cost, processing and, depending on the measurement method, also time intensive. Thereby, one unanswered question is how accurately the spatial impulse response of a room has to be measured in order to authentically reproduce the acoustic impression at a listener position. From the current state of research in this field it is, to the author's best knowledge, not evident whether or not the exact elevation angle i.e. the vertical position of an early room reflection influences the spatial auditory perception and if humans perceive a difference when changing the original elevation of such a reflection. As the human detection accuracy for a change in elevation of a sound source is not as precise as the detection of a change in lateral displacement [1, Sec. 2.1], one could assume that the vertical resolution of a spatial room impulse response does not have to be as accurate as the horizontal resolution in order to achieve a plausible room impression.

In practice, this aspect is relevant since accurately determining the elevation angle of a reflection requires a three dimensional microphone array and thereby a significantly larger number of microphones than e.g. using a two dimensional microphone array which results in a decreased vertical localization accuracy. Recently developed microphone arrays such as the far-field equatorial array [2] follow this approach by assuming only horizontally-propagating sound fields and hence using a two dimensional, circular microphone arrangement which leads to a significantly reduced number of required microphones at the cost of compressing all elevated sound sources to the horizontal plane.

Motivated by the question if such a spatial compression of elevated reflections leads to noticeable auditory changes, the scope of this thesis was set to investigate under which conditions humans perceive a difference when altering the elevation of an early room reflection. Thereby, the objective was not to obtain precise detection threshold values but to rather gain fundamental insights on the elevation dependent auditory perception of early reflections which can then be used to estimate under which conditions an accurate spatial reproduction of elevated reflections might be required.

While this seems to be a rather fundamental psychoacoustic question, not much literature on this specific aspect of spatial hearing was found. In the last decades, several authors investigated the perceptual effects of isolated ceiling reflections on parameters such as localization accuracy [3], auditory envelopment [4] as well as timbral [5] and spatial aspects [6]. Recently, publications such as [7], [8], [9] and related work by the same authors evaluated different perceptual aspects of vertical reflections, often in a 'listening for entertainment' context based on three-dimensional loudspeaker reproduction systems such as Dolby Atmos and Auro 3D. Thereby, the authors of [7] used a listening experiment setup consisting of two loudspeakers in a semi-anechoic space whereby one of the loudspeakers was elevated to electroacoustically simulate an early ceiling reflection. This setup is, in its fundamental idea, quite similar to the loudspeaker based listening experiment performed in this thesis. However, while [7] evaluated perceived spatial and timbral differences as well as preference for different musical stimuli and speech, only cases with and without added ceiling reflection were compared to each other and it was not evaluated how a shift of the elevated reflection down to the horizontal plane affects the auditory impression. To the authors best knowledge, none of the previously performed studies actually investigated the effects of altering the elevation of an early room reflection.

2

Theory

2.1 Fundamental Principles of Spatial Hearing

The purpose of this section is to introduce the fundamental auditory concepts relevant in the context of this thesis. Naturally, this introduction can only present a mere fraction of the continuously growing state of knowledge on spatial hearing, more detailed insights can be obtained from relevant literature such as [1], [10] and [11].

2.1.1 The Auditory Space

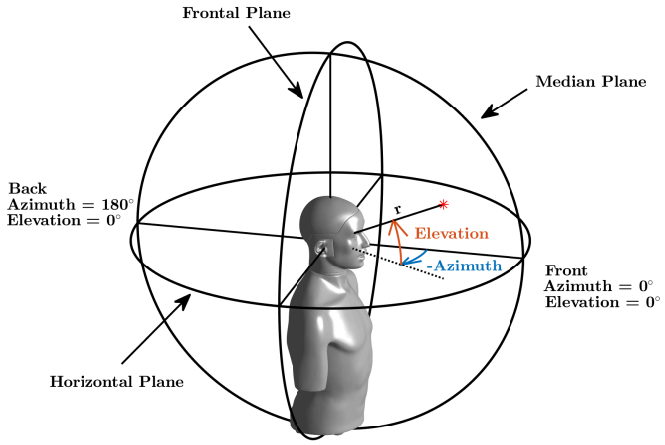


Figure 2.1: Head-related coordinate system which describes a source position by its azimuth angle, elevation angle and distance r relative to a listener's head center.

In order to describe the position of a sound source relative to a receiver position, it is common to use a head-related coordinate system as shown in Figure 2.1 which describes a source position by its azimuth and elevation angle in degree as well as its distance r. Thereby, the origin of the coordinate system is defined as the center between the upper margins of the two ear canal entrances [1, Sec. 1.3]. Since there are different symbols commonly used for azimuth (e.g. φ or θ) and elevation (e.g. δ), this thesis just uses the written out terms in order to avoid confusion. In addition to the spherical coordinates describing a source position, three different planes which

are orthogonal to each other and intersect at the origin allow to further classify the auditory space. While different authors use different terms for those planes, this report defines them as median, frontal and horizontal plane as proposed by [1] and shown in Figure 2.1.

2.1.2 Human Sound Source Localization

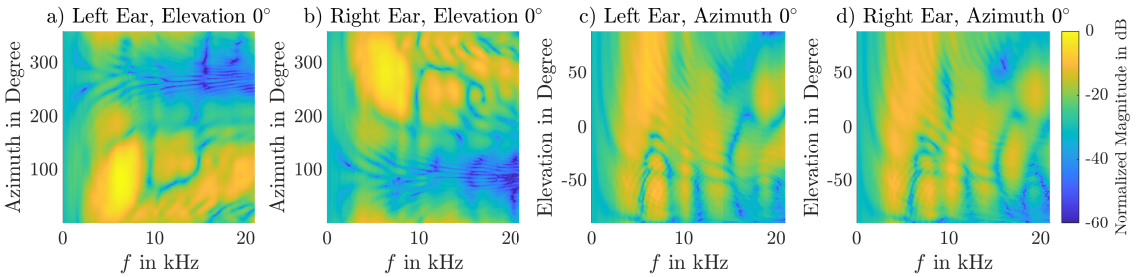


Figure 2.2: Normalized magnitude spectra of KEMAR HRTF set [12] for source positions on the horizontal plane (a and b) as well as for source positions on the forward median plane (c and d).

The human ability to localize a sound event in the auditory space relies on attributes of the sound waves arriving at both ears, so called auditory cues. Those cues can generally be divided into two groups i.e. attributes that differ between both ears, so called interaural cues such as interaural time or level differences, and attributes that affect both ear signals in the same way, so called spectral or monaural cues.

The fundamental mechanism responsible for the generation of those directional cues is the anatomy of the human head, torso and outer ear which alters a sound wave depending on its incident angle. E.g. a sound wave arriving from the left side reaches the right ear later and with less amplitude compared to the left ear. Additionally, the wave bends around the head and pinna which leads to spectral differences between both ears. A sound wave originating from the median plane on the other hand, i.e. with zero azimuth and a certain elevation, reaches both ears with the same level and at the same time. Assuming a symmetric head anatomy, the spectral modifications due to interferences caused by the outer ear, sound waves bending around the head, reflections from the torso etc. are also identical. In this case, the perception of elevation only relies on monaural cues which means that the human brain is able to recognize certain spectral characteristics of the sound arriving at both ears and assign those spectral cues to a perceived elevation. In general, those monaural cues are not as robust as interaural cues which means that the human ability to perceive changes in the elevation angle of an auditory event is lower than for changes in azimuth angle [1, Sec. 2.1]. However, interaural cues alone are not sufficient for a reliable discrimination of the front and back location of a sound source as well as its elevation [10, Sec. 5.1]. Without any spectral information, the interaural cues yield in a so called “cone of confusion”, which describes all possible sound source position that have the same distance to the left and right ear and therefore result in similar interaural differences [1, Sec. 2.5]. This effect can be observed in Figure 2.3, where

both the interaural level difference (ILD) and interaural time difference (ITD) are very similar in a specific cone shaped region centered at ± 90° azimuth angle and 0° elevation. This means that the key for the human ability to accurately localize a sound source is the combination of interaural and spectral cues as well as dynamic cues caused by head movements [13]. Thereby, the smallest difference in a specific attribute such as azimuth or elevation of an auditory event that is required to result in a perceived change of the sound localization is referred to as localization blur. As comprehensively described in [1, Sec. 2.1], this localization blur is not only dependent on the source's azimuth and elevation angle but also on source signal attributes such as frequency spectrum, level, distance and time structure.

A common approach to describe the modifications of sound waves arriving from different incident angles caused by the human anatomy is to measure or simulate the transfer path from an arbitrary source position to the entry points of both ear canals. This transfer path can be modelled as LTI system with a certain frequency response function which is referred to as head related transfer function (HRTF) and with a finite impulse response referred to as head related impulse response (HRIR). A HRTF can be obtained from measurements with an artificial head which is a manikin that, depending on the specific model, mimics the shape of the human head, outer ears and torso and contains two microphones placed in its ear canals. The artificial head used for all measurements in this thesis is a KEMAR model from GRAS Sound and Vibration as shown on the cover page as well as in Figure 3.1b. Instead of using artificial heads, HRTFs can also be measured by placing small microphones in a subject's ear canals or by using approaches such as the boundary element method (BEM) in order to simulate the transfer paths numerically. One important distinction in the context of HRTFs is whether a custom HRTF set was individually measured for a subject or if a standard HRTF, e.g. obtained from an artificial head which not necessarily fits to a subject's specific anatomy, is used. The latter one is referred to as generalized HRTF and has the downside that, due to inconsistencies between the generalized HRTF and the subjects individual HRTF, both spectral and interaural cues might not be accurate for the specific person. As the spectral cues are not as robust as the interaural cues, one common effect of generalized HRTFs is a degradation of the localization accuracy for elevated sources [14].

Figure 2.2 shows an example HRTF set obtained from placing a KEMAR head in an anechoic environment and measuring the transfer paths from different source positions placed on a sphere surrounding the head with 1° × 1° resolution [12]. Thereby, plots a and b show the left and right ear HRTFs for source positions on the horizontal plane and plots c and d visualize the left and right ear HRTFs for source positions on the forward median plane. When comparing plots a and b to each other it becomes obvious that, for sources on the horizontal plane i.e. with 0° elevation and varying azimuth angles, the left and right ear HRTFs differ significantly from each other except from the 0° and 180° azimuth positions where the source is centered in front or behind the head. This observation visualizes the previously introduced concept of interaural localization cues: if a sound wave impinges from a lateral incident angle, the time of arrival, level and spectrum differs between both ears. Due to the, to some extend, symmetrical human anatomy, the left and right

ear horizontal HRTFs resemble mirrored versions of each other.

For sound sources on the median plane, i.e. at 0° azimuth angle with varying elevation, the left and right ear HRTFs are almost identical as shown in Figure 2.2c and d. However, the HRTFs for both ears change with the elevation angle, those spectral changes which affect both ear signals in the same way are referred to as monaural localization cues.

2.1.3 Interaural Difference Measures

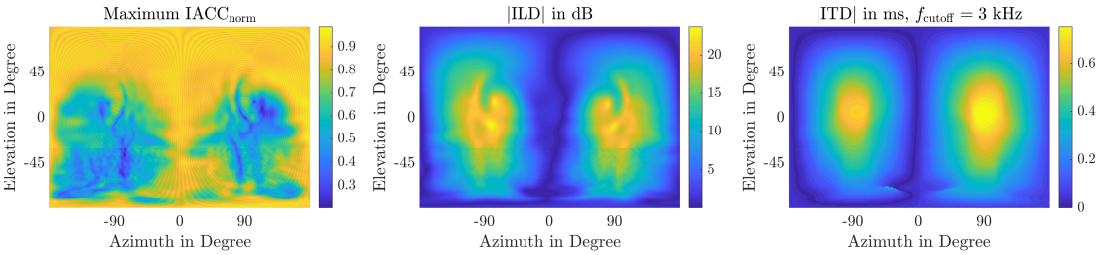


Figure 2.3: Absolute values of ILD, ITD and maximum normalized IACC calculated for KEMAR free field HRIR set.

A binaural sound scene such as an artificial head recording or a simulation obtained by filtering a source signal with a set of HRTFs can be analyzed regarding its interaural differences by comparing the two ear signals to each other. The most common measures for interaural differences are the interaural level difference (ILD), the interaural time difference (ITD) and the interaural cross-correlation (IACC), which will be briefly described in the following. Figure 2.3 shows the directional distribution of those three parameters calculated for the KEMAR HRIR set. Comparing those three parameters for different elevation angles on the median plane, i.e. at 0° azimuth angle, it becomes clear that interaural cues alone are not sufficient for an accurate sound source localization as all three parameters stay almost constant regardless of the elevation angle.

2.1.3.1 Interaural Cross-Correlation (IACC)

The interaural cross-correlation is a measure for the similarity of the two ear signals. The normalized IACC, which will be used in the following, can be calculated from the left and right ear signals $y_{\text{l}}(t)$ and $y_{\text{r}}(t)$ as

$$\text{IACC}_{\text{norm}}(\tau) = \frac{\int_{t=-\infty}^{+\infty} y_{\text{l}}(t) \cdot y_{\text{r}}(t+\tau)\, dt}{\sqrt{\int_{t=-\infty}^{+\infty} y_{\text{l}}^2(t)\, dt \cdot \int_{t=-\infty}^{+\infty} y_{\text{r}}^2(t)\, dt}} = \frac{\overline{y_{\text{l}}(t) y_{\text{r}}(t+\tau)}}{\widetilde{y}_{\text{l}} \cdot \widetilde{y}_{\text{r}}} \tag{2.1}$$

[10, eq. 2] [1, eq. 3.1], where τ is a lag parameter which is used to compensate a frequency independent delay between the both signals. $\widetilde{y}_{\text{l}}$ and $\widetilde{y}_{\text{r}}$ denote the RMS values of the left and right ear signals. Thereby, the two signals can be considered as identical regarding their IACC if $\max(|\text{IACC}_{\text{norm}}(\tau)|) = 1$, which is the case as long as both ear signals do not have frequency dependent level or phase differences to

each other. I.e. a frequency independent level difference or delay between y_l and y_r does not influence the maximum of the normalized IACC. In this case, some authors also refer to "coherent" signals [1, p. 202]. The lag τ for which the normalized IACC reaches its maximum can be considered as the overall delay between the two signals. As shown in the left plot of Figure 2.3, the maximum of the normalized IACC for a set of free-field HRIRs is almost independent of the elevation angle at 0° and 180° azimuth angle. This is expected since the human anatomy is relatively symmetric on the median plane hence the left and right ear signals are very similar independent on the elevation. For source positions which do not lie on the median plane however, the maximum of the normalized IACC varies both with azimuth and elevation angle.

In room acoustics, the IACC is also associated with the apparent source width (ASW), which describes the perceived audible impression of a spatially extended sound source. Thereby, the perceived apparent source width was found to be higher the less correlated the two ear signals are hence a large ASW value corresponds to a low IACC. Other room acoustic properties such as early reflections before 100 ms also increase the ASW [15]. Additionally, the interaural cross-correlation calculated for the 500 Hz, 1000 Hz and 2000 Hz octave band was found to be significantly negatively correlated to the perceived listener envelopment (LEV) of a concert hall [16], since more reflections in a room lead to a reduced IACC which makes the primary auditory event appearing more diffuse and more spatially extended [1, p. 280]. This can be considered as relevant in the context of this thesis since a difference in IACC between two stimuli with different elevated reflections could indicate a change in the spatial perception and several studies have shown that the auditory system can be remarkably sensitive regarding IACC changes [17][18].

2.1.3.2 Interaural Level Difference (ILD)

The interaural level difference describes the sound pressure level difference between the two signals of a single sound source arriving at both ears and can be calculated from the RMS values of both ear signals $\widetilde{y_\mathrm{l}}$ and $\widetilde{y_\mathrm{r}}$ as

$$\mathrm{ILD} = \left|20 \log_{10}\left(\frac{\widetilde{y_\mathrm{l}}}{\widetilde{y_\mathrm{r}}}\right)\right| \ . \tag{2.2}$$

Thereby, a change of ILD causes a lateral displacement of an auditory event. This is the fundamental concept behind techniques such as stereo amplitude panning, where a sound source gets positioned on the horizontal plane by sending it with different levels to the left and right loudspeaker or headphone channels. While an auditory event on the median plane causes almost no level differences between both ears, both the sound pressure decrease over a longer distance as well as the shielding of the head itself cause the signal arriving at the ear facing the sound source to have a higher amplitude than the signal arriving at the other ear. This effect is shown in Figure 2.3 where, depending on the azimuth angle, the ILD under free-field conditions reaches up to 26 dB at an azimuth angle of approximately 90°.

2.1.3.3 Interaural Time Difference (ITD)

The interaural time difference specifies the overall difference in time of arrival between the left and right ear signals of a single auditory event. While there are several approaches to calculate the ITD from HRIR data or binaural recordings described in literature, such as onset time difference or temporal moment of maximum correlation, none of them can be considered as 'ground truth' [19]. The approach used for the ITD results shown in the right plot of Figure 2.3 simply analyzed a low pass filtered version of the HRIR regarding the lag τ that leads to a maximum normalized IACC following

$$\mathrm{ITD} = \arg\max(|\mathrm{IACC}_{\mathrm{norm}}(\tau)|) \ . \tag{2.3}$$

Similar to the ILD, the interaural time difference is responsible for the perception of lateral displaced auditory events whereby nowadays, most of the authors on this subject agree that the ITD is in fact the most important attribute for perceived lateral displacement [1, p. 141]. As described earlier, analyzing just the interaural time difference does not allow an accurate elevation or front to back distinction since there are concentric areas with an identical ITD value centered around the $\pm$ 90° azimuth and 0° elevation position as shown in Figure 2.3.

2.1.4 Spatial Perception of Reflections

So far, all described aspects of spatial hearing primarily assumed free field conditions, i.e. a single sound wave arriving at the receiver position from one incidence angle. However, in practice it is more likely that even a single sound source emits sound waves that reach a listener's ears from different directions and with different delays due to room reflections. Nevertheless, when placing a sound source in a room that fulfills certain acoustic requirements the listener still localizes the sound source at only one position, even though the sound arriving at the listener position contains signals from multiple image sources with different locations. This phenomenon can be explained by the precedence effect, also known as the law of the first wavefront [1]. Thereby, one can distinguish between three different scenarios:

If the delay between a reflection and the direct signal is smaller than approximately 1 ms, a fusion of both sounds to a single auditory event occurs. Thereby, the perceived location of this single event depends on the delay between the two signals, the larger the delay between direct and reflected sound the more the perceived location shifts towards the position of the sound source. This effect is also referred to as summing localization [10, p. 3].

If the delay between direct and reflected sound is bigger than 1 ms but smaller than the so called echo threshold, the precedence effect occurs. This means that again, both sounds are fused to a single auditory event but, opposed to the previously described scenario of a delay below 1 ms, the direct sound has the localization dominance over the reflection in this case hence the auditory event is mainly perceived from the position of the direct sound source. While the reflection is not perceived as a separate sound source, it still contributes to the overall impression of the au-

ditory event as it leads to a change in volume, can cause spectral changes due to interferences between both signals also referred to as coloration and it results in an increased perceived spaciousness [10, p. 3]. Thereby, two significant effects in the context of binaural listening are the binaural decoloration and binaural dereverberation, which describe the ability of the human auditory system to both compensate for the coloration and increased reverberation caused by early reflections. A simple example for these effects is that both the perceived reverberation and coloration of an auditory event containing a direct signal as well as reflection are clearly larger when plugging one ear than when listening to the signal with both ears [10, pp. 4, 362]. This observation is particularly important for the experiments performed later in this thesis as, due to the binaural decoloration, the perceived difference in timbre between two auditory events with reflections at different elevation angles is likely to be compensated by the listeners auditory system. Without any form of compensation, an elevated reflection would always sound different in timbre than a non elevated reflection due to the different HRTFs applied for both cases.

Lastly, if the delay between direct sound and reflection exceeds the echo threshold the reflection is perceived as echo i.e. as a separated auditory event with a different apparent location than the direct sound. The exact value of this echo threshold is strongly dependent on the source signal, e.g. for short impulses it can be only 1 ms, for a continuous speech signal it is approximately 50 ms and for classical music it can reach up to 80 ms [10, p. 3]. Besides the overall delay time and the type of source signal, the echo threshold also depends on other parameters such as volume. E.g. for a speech signal with a reflection delay less than 32 ms, the reflection level can even be up to 5 dB higher than the source signal without becoming an audible echo [1, p. 226]. This effect was firstly described by Haas in 1951 and is therefore also known as “Haas effect”. However, it is still a subject of discussion how these fundamental effects occur in a diffuse sound field depending on the spatial position of a reflection. Additionally, the presence of other reflections i.e. the reverberation time of a room influences whether or not a particular reflection is perceived as echo. In general, additional reflections between the primary sound and a specific reflection decrease the likeliness of this particular reflection becoming audible due to masking effects [1, p. 274].

2.2 Spatial Room Impulse Responses

In order to simulate reflections with varying elevation angles in a realistic diffuse field scenario, the sound field at a specific position in a room for a given sound source location can be reproduced by measuring and processing so called spatial room impulse responses.

Assuming that a room or any other system is linear and time-invariant, the system's behaviour can be modeled in time and frequency domain by determining its impulse response (IR). Thereby, a room impulse response (RIR) describes the transfer path from a specific source position to a specific receiver position including all occurring room reflections. If a RIR is measured with a single omnidirectional microphone, i.e. all reflections are summed with the same weight regardless of from which direction

they are arriving, one also refers to a pressure room impulse response (PRIR). Such a pressure room impulse response does not contain any spatial information besides the individual reflection path lengths. A spatial room impulse response (SRIR) on the other hand does not only contain the pressure over time signal found in a PRIR but it also includes additional spatial information which can be used to map individual parts of the pressure signal, i.e. individual transmission paths, to different incident angles, so called directions of arrival (DOAs). In order to determine such an SRIR, several pressure RIRs have to be measured at different, spatially distributed receiver positions e.g. by using a microphone array consisting of different microphone positions arranged in a specific geometry. Thereby, at least four microphone positions in a non-planar arrangement are required to explicitly determine the three-dimensional incident vector of a sound wave arriving at the microphone array, the exact number of required microphone positions generally depends on the desired localization accuracy. The optimum geometry of such a microphone array differs for specific applications and signal processing methods used to obtain spatial information from the measured set of spatially distributed pressure impulse responses. The SRIR processing approach used in this thesis is the spatial decomposition method, which will be described in the following section.

2.2.1 Spatial Decomposition Method (SDM)

The Spatial Decomposition Method (SDM) is a method to assign each sample of a spatial room impulse response to an incident direction vector, which can also be interpreted as decomposing the SRIR into a limited number of image-sources [20].

Therefore, impulse responses of a room have to be measured or simulated at different spatially distributed positions, e.g. by using a microphone array. From this set of PRIRs, the direction of arrival for each sample of a reference pressure RIR, which can either be one of the array microphone signals or the signal from a separate microphone placed in the center of the array, gets assigned to a direction of arrival. This can either be achieved by using so called pseudo intensity vectors which use the sound intensity at each microphone to determine the DOA, or by comparing the time differences of arrival (TDOA) of the signal at each microphone position. The latter approach is the method implemented in the publicly available SDM toolbox [21], which was used as foundation for the SDM calculations performed in this thesis. This directional information is then used to divide the original pressure RIR into a subset of impulse responses for different sound incident angles. This set of impulse responses can then either be used for a loudspeaker array based spatial audio reproduction, i.e. by placing a number of loudspeakers around the listener and assigning the SRIR for the corresponding source position to each loudspeaker, or for headphone based spatial audio reproduction by convolving the SRIR for each DOA with the HRIR of the corresponding incident angle. The fundamental concept of the SDM method is visualized in Figure 2.4.

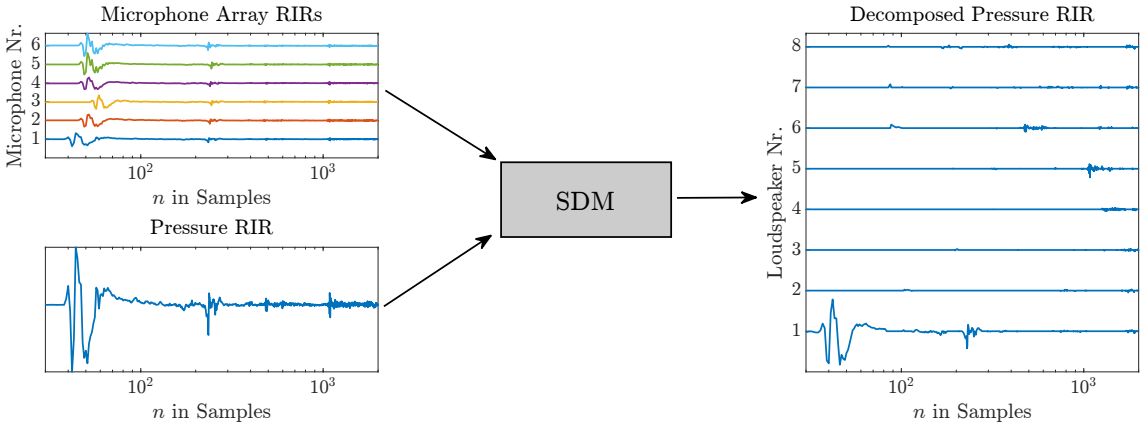


Figure 2.4: Visualization of SDM method: A single channel pressure RIR gets decomposed into a set of SRIRs for different incident angles i.e. virtual playback loudspeaker positions by analyzing the spatial information obtained from a set of microphone array signals.

2.2.2 DOA Estimation Using the Time Difference of Arrival Method

A straightforward method to estimate the direction of arrival for each sample of a spatial room impulse response is to analyze the time difference of arrival between the different microphone signals. Therefore, the different microphone signals get sectioned using a small analysis window and then cross-correlated in order to find the individual delays that maximize the cross-correlation values between the windowed sections of the different microphone signals. Thereby, the window length has to be at least as long as the largest run-time difference between the different microphone capsules. Assuming that only one broadband sound event arrives within an analysis window, the incident angle of this event can be calculated from the estimated TDOAs and the known microphone positions using a least squares solution as described in [20]. If multiple sound events arrive within the same analysis window, the corresponding reference PRIR sample gets assigned to the DOA of the strongest sound event within this window. Moving the window sample wise over the entire SRIR results in a matrix containing the three dimensional DOA for each sample of the spatial room impulse response. Optionally, a moving average filter can be applied afterwards to smooth those DOAs in order to avoid a fixed sound source changing its position over time due to the limited spatial resolution of this method [22]. Since the echo density of an impulse response increases over time, the late parts of an IR contain a large number of sound events per analysis window which then result in randomly fluctuating DOAs for the diffuse parts of an SRIR.

2.2.3 DOA Quantization

Apart from limitations due to finite computational precision and sampling resolution, the previously described DOA estimation method can result in infinitely many different DOAs. For a loudspeaker array based reproduction method however, the amount of available playback loudspeakers is limited and hence those DOAs have

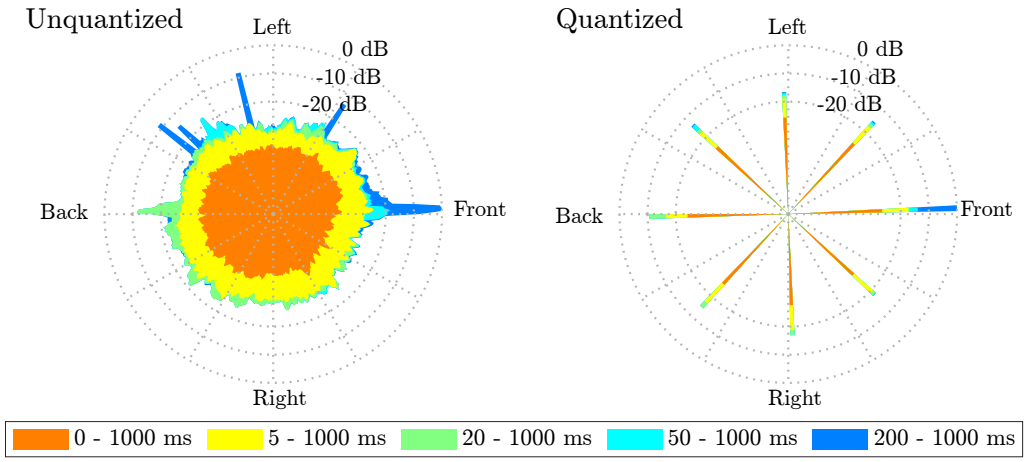


Figure 2.5: Unquantized and quantized spatio-temporal visualization of "Big Hall" SRIR on horizontal plane, calculated using the SDM Toolbox [21].

to be quantized to a finite number of loudspeaker positions. The same applies for a headphone based reproduction, where the HRIR set only contains data for a limited number of incident angles. An additional concern for headphone based rendering is that using all the unquantized DOA information with a high resolution HRTF set can result in a spread of specular reflections to multiple incident directions which then leads to spatial and timbral distortions [22]. Therefore, the DOAs for a binaural rendering often get quantized to a spherical array of virtual loudspeaker positions such as a Lebedev grid. [22] showed that such a grid of 14 virtual loudspeaker positions is sufficient for a plausible headphone based reproduction as no perceptual improvement was found when using more than 14 quantization positions on a Lebedev grid.

A commonly used approach to quantize SRIRs is the Nearest Loudspeaker Synthesis (NLS) method [23], which simply shifts each DOA to the nearest loudspeaker position and thereby results in one single impulse response for each loudspeaker. The sum of these individual impulse responses is equal to the original pressure RIR. This approach was used for the SRIR calculations performed in this thesis and is visualized in Figure 2.5, where the left plot shows the directional energy distribution on the horizontal plane using unquantized DOAs while the right plot shows the same SRIR quantized to eight loudspeaker positions on the horizontal plane.

An alternative approach to quantize a set of continuous DOAs to discrete loudspeaker positions is the Vector Based Amplitude Panning (VBAP) method [24] which, instead of mapping each DOA to the nearest loudspeaker position and thereby potentially causing perceivable localization shifts, aims to keep each DOA at its original position by assigning different amplitudes of it to the nearest three loudspeakers. This approach is similar to classical stereo panning where a sound source is intended to be perceived from a position between two loudspeakers gets assigned to both the left and right loudspeaker with different gains. However, one downside of VBAP is that it can cause a spread of single auditory events as, depending on the used loudspeaker layout, a single sound source can appear from three different loudspeakers and therefore might not be perceived from a single point anymore which leads to a

reduced perceived clarity [23]. This is especially critical if the individual loudspeakers have a relatively large distance to each other. Besides NLS and VBAP, there are also more advanced approaches like virtual layout optimization [25] and density based spatial clustering [26]. However, it was found that the NLS method produces sufficiently adequate results for the purpose of this thesis.

2.2.4 SDM Equalization

Splitting a single pressure room impulse response into several IRs for different incident angles i.e. loudspeaker channels leads to a reduced impulse density per channel. These decomposed SRIRs then often reassemble a sum of individual impulses rather than a continuous signal which can cause clicking artifacts. Those wide band transients then result in an increase of white noise components which is also referred to as spectral whitening. While these effects can partly be compensated by simple low-pass filtering, a more advanced SDM post equalization method described in [23] aims to correct these spectral changes by comparing the sum of the synthesized SDM IR spectra to the spectrum of the original pressure room impulse response for overlapping short-time Fourier-transform frames and thereby generating time varying filters for each rendered channel. This is especially useful for a loudspeaker based reproduction method with a limited amount of channels. For binaural rendering, the RTmod method recently introduced in [22] equalizes the reverberation tail by dividing the rendered BRIRs into fractional octave bands and modifying their energy envelopes to match the original PRIR.

2.2.5 Spatial Audio Playback

In order to finally play back a source signal virtually placed in the measured room, one can either use a loudspeaker based or a headphone based reproduction method. For the loudspeaker based playback, the quantized and optionally equalized SRIRs can be directly convolved with a source signal and the resulting N audio signals can be played back through N different loudspeakers which should be placed at the same angle relative to the listener as the DOA quantization positions. For a headphone based binaural reproduction, the SRIRs get convolved with the left and right ear HRIRs for the corresponding DOAs and are then separately summed up for both ears. The outcomes of this summation are two IRs, one for each ear, which, when convolving them with a single channel source signal and listening to the results via headphones, auralizes the measured acoustic environment. This combination of SRIRs and HRIRs is also referred to as binaural impulse response (BRIR), since it approximates the impulse responses one would obtain when placing an artificial head at the receiver position in the room of interest and measure the IRs from a source position to the ear canal entry points of the artificial head. In this context, the DOA quantization positions are also referred to as virtual loudspeakers.

2.2.6 SDM Limitations

Since its introduction in 2013, the SDM method gained popularity for spatial room impulse response encoding due to its rather simple functionality in time domain

and a publicly available Matlab implementation [21]. However, one fundamental limitation of the SDM method is the assumption that the surrounding sound field consists of a succession of broadband specular events and that direction of sound propagation is the average of all sound waves arriving simultaneously at the microphone array. As the method assigns only one DOA to each PRIR sample, two reflections arriving at the same time can not be accurately reproduced since they are both mapped to the same DOA. Other methods like HO-SIRR overcome this limitation by dividing a higher-order spherical harmonic input signal into different directional sectors which are then analyzed separately [27]. An additional limitation of SDM is that low-frequency signal components with wavelengths that exceed the DOA analysis window are divided into separate temporal segments and might thereby get assigned to varying incident directions based on the broad band DOA estimation. This can lead to time-varying distortions as well as colouration of the output signal which can be partly compensated by applying adaptive filtering [27].

3

Methods

The foundation of this thesis are two different listening experiments of which the first one used a loudspeaker based reproduction method in an anechoic environment while the second experiment relied on a headphone based reproduction method with generalized HRTFs. The following section describes the measurement and selection process of the evaluated stimuli as well as the setup of both experiments and the used data evaluation methods.

3.1 Simulations and Preliminary Experiments

(a) (b)

Figure 3.1: Preliminary SDM measurement with microphone array (a) and BRIR measurement with strong ceiling reflection (b).

In order to develop a meaningful listening experiment setup to evaluate the influence of changing a reflection's elevation on the human auditory perception, several preliminary experiments were performed. Thereby, the first step was to simulate a single ceiling reflection using methods of geometrical room acoustics in Matlab. Therefore, a simple setup with one direct path and one ceiling reflection path was considered. The ceiling reflection was modeled by designing an IIR filter emulating the absorption of a rough concrete surface in combination with additional all-pass filters in order to simulate the reflection's phase alterations. The reflection path was attenuated following the inverse square law and delayed according to the calculated run-time difference. The resulting two transfer path IRs were then summed and convolved with HRIRs corresponding to the incident angle of the direct and

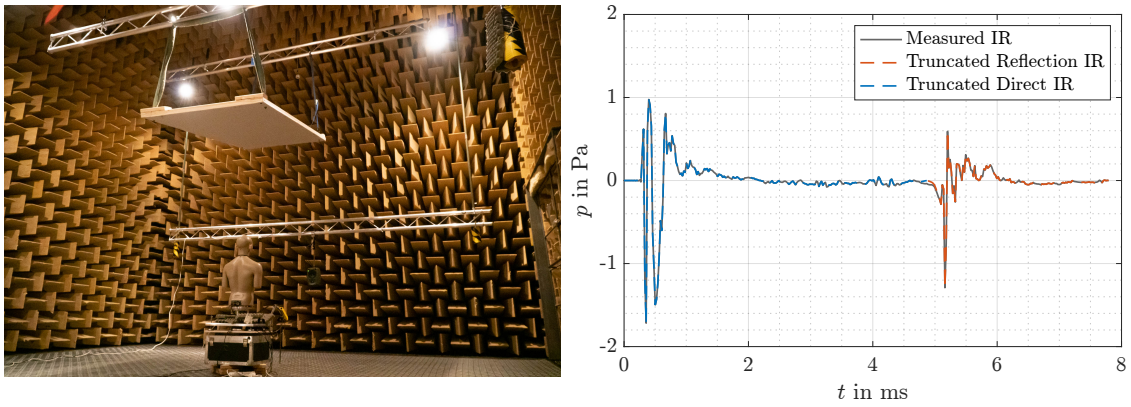
the reflection path which yielded in a set of BRIRs for a simple binaural sound scene with a direct sound source positioned in front of the listener position as well a single ceiling reflection positioned at an arbitrary elevation angle. Convolving these BRIRs with a source signal and listening to the results via headphones gave an initial impression on the range of audible differences when comparing different reflection elevation angles. Thereby, it was found that, depending on the level and position of the reflection as well as the type of source signal, there are indeed perceivable differences when changing the reflection's elevation angle using a headphone based reproduction with nonindividualized HRTFs. However, these differences were in general not very pronounced which led to the conclusion that the actual listening experiments should be designed in a way such that even very small differences between two stimuli can be identified.

While these initial simulations were a good starting point for this thesis, it is obvious that a real ceiling reflection is more complex than delaying and filtering a single copy of the source signal. Therefore, it was decided that the actual listening experiment should be based on stimuli obtained from acoustic measurements rather than simulations. In order to evaluate a useful way to measure and reproduce a spatial room impulse response including a strong ceiling reflection, a series of pilot measurements in a medium sized storage room with and without a 60 cm $\times$ 90 cm $\times$ 1 cm gypsum plate placed directly above the receiver position was conducted both using an open microphone array consisting of six positions with a radius of 25 mm as well as with a KEMAR artificial head. Both measurements setups are shown in Figure 3.1. While the results of these measurements allowed to directly compare artificial head or SDM rendered BRIRs with and without an early ceiling reflection to each other, this approach did not allow to change the position or level of the ceiling reflection except from manually modifying the SDM decomposed SRIR which might lead to spatial and timbral degradations. Therefore, it was found that, for the actual listening experiments, the elevated reflection should be measured separately in an anechoic environment and then manually be added to a measured SRIR. The exact procedure of these reflection measurements are described in Section 3.2.

One final question that had to be answered before conducting the actual listening experiments was whether or not a headphone based reproduction method is suitable for this purpose. Thereby, the main concern was that, as no custom HRTFs for the participants were available and measuring them would exceed the scope of this thesis, a generalized HRTF set had to be used for the binaural renderings which might lead to a reduced localization accuracy especially for elevated sound sources [14]. In this context, the easiest way to implement an experiment with individualized HRTFs is to use a loudspeaker array based reproduction method. This eliminates the need for any HRTFs applied to the stimuli as, when physically placing multiple loudspeakers around the participant and assigning different sound sources to them, the listeners automatically use their own HRTFs for the localization. In order to estimate if such a loudspeaker based reproduction bears any advantages in the context of this thesis, a multi-channel loudspeaker array consisting of five surround speakers and one elevated speaker was set up in a listening room. By playing back several of the previously measured stimuli via this setup it was found that the difference between an elevated reflection and the same reflection mapped to the horizontal plane seems

more perceivable using a loudspeaker based setup compared to a headphone based reproduction with generalized HRTFs. In order to evaluate the difference between both reproduction methods, it was decided to finally perform two listening experiments, i.e. one using a loudspeaker array and one using a headphone based sound reproduction. However, the previously described setup consisting of six loudspeakers was not found to be very immersive and it was unclear in what sense the reflections of the used listening lab distort the spatial perception of the stimuli. Therefore, it was decided to use a larger number of speakers in an anechoic environment for the actual loudspeaker based experiment as described in Section 3.5.

3.2 Ceiling Reflection Measurements



(a) (b)

Figure 3.2: Reflection measurement setup (a) and obtained impulse responses (b).

While the simulations of a single ceiling reflection described in Section 3.1 were sufficient for the preliminary experiments, it was decided that, for the final listening tests, the elevated reflection should be as authentic as possible hence a real ceiling reflection had to be measured.

Originally, it was intended to directly measure the SRIR of rooms modified to have strong ceiling reflections i.e. by placing a reflecting plate above a receiver position. However, during the preliminary experiments it became clear that measuring an isolated reflection in an anechoic environment and then manually adding this reflection to the SRIR of a real room allows more flexibility both in position and strength of the reflection. Therefore, a 2.40 m x 1.20 m gypsum plate was mounted on a wooden frame and installed in the anechoic chamber of the Division of Applied Acoustics as shown in Figure 3.2a. The plate was centered 1.3 m above the receiver position with an upwards tilt of approximately 12°. According to basic geometrical room acoustics calculations, this results in a ceiling reflection with an elevation angle of approximately 84° when referring to the point on the plate where the incident angle equals the reflection angle. The transfer path from a Genelec 8030 source loudspeaker to a B&K Type 4190 free-field measurement microphone was measured using the sweep deconvolution method [28] with a logarithmic sweep ranging from 50 Hz to 20 kHz and 20 measurement repetitions. The microphone signal was pre-conditioned using

a B&K Type 1708 signal conditioning amplifier which was set to a linear frequency response, an Antelope Orion 32 audio interface was used to record the microphone signal at 48 kHz / 24 bit and to output the sweep signal. In order to obtain a delay compensated reference signal for the deconvolution, the sweep signal was fed back to a second input of the audio interface. In addition to the single microphone recording, the ceiling reflection was also measured using an open array consisting of 6 microphone positions as well as using a KEMAR artificial head at 360 different azimuth angles as shown in Figure 3.2a. While those two additional measurements were not of further interest for this thesis, they might be useful for related projects. In order to obtain just the reflection path IR, the full IR was truncated and windowed, as shown in Figure 3.2b. Thereby, the original delay between direct and reflected sound (ca. 5 ms $\approx$ 1.7 m path-length difference) was preserved by zero padding the truncated reflection IR.

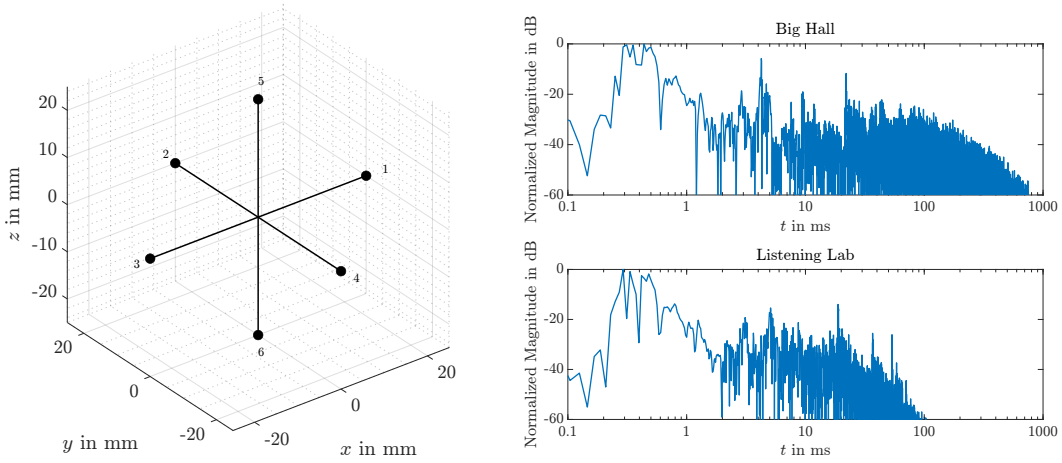
Using a single reflection measured at one position and then manually changing its elevation angle to render the different stimuli for the listening test means that the measured reflection IR is only authentic at its originally measured position. I.e. physically changing the elevation of a reflection would potentially lead to a different tilting angle and hence variations in effects such as scattering while virtually changing the position of the reflection in the SRIR rendering results in exactly the same reflection arriving from a different incident angle. However, as the goal for these experiments was to only change the position of a reflection without altering its timbre, this inaccuracy is actually an advantage in the context of this study.

3.3 SRIR Decomposition

(a) (b)

Figure 3.3: Spherical panoramas of “Big Hall” (a) and “Listening Lab” (b).

In order to simulate elevated early reflections in a realistic acoustic environment, spatial room impulse of a small, acoustically treated listening lab and a big hall were processed. Pictures of both rooms are shown in Figure 3.3. These SRIR measurements were performed by Ahrens as described in [29] by using an open array consisting of 6 microphone positions placed on a spherical grid with a radius of 2.5 cm as shown in Figure 3.4a. Figure 3.4b shows the measured pressure RIRs which are equal to the first microphone signal, Section 3.3 lists the essential room acoustic parameters of booth rooms obtained by analyzing the PRIR with the ITA-



(a) (b)

Figure 3.4: Microphone array geometry (a) and normalized magnitude of "Big Hall" and "Listening Lab" PRIRs on logarithmic time axis (b).

Toolbox [30]. Additionally, the table includes a Δ_{Ceiling} parameter which indicates the delay between the direct sound and the first strong natural ceiling reflection.

These two spatial room impulse responses were then processed using the optimized version of the SDM toolbox provided by [22], which was adapted to the requirements of this thesis by changing the included binaural synthesis method to a NLS loudspeaker rendering. For the SDM DOA estimation, a window length of 36 samples with a 16 samples smoothing window was used. Neither the original SDM post-equalization nor the RTMod equalization as proposed in [22] were applied since it was found that neither method resulted in a significant improvement for the used SRIRs and source stimuli and, as the purpose of these experiments is to investigate differences between stimuli rendered with exactly the same SRIRs and varying, manually added elevated reflections, the overall plausibility and potentially timbral inaccuracies of the decomposed SRIRs compared to the real room IRs were not considered as extremely significant.

For each room, two different SDM decomposed SRIRs were rendered i.e. one where all DOAs were quantized to eight virtual loudspeakers on the horizontal plane and one where an additional virtual top loudspeaker with 90° elevation was included in the quantization grid. This allowed to evaluate whether or not there is a perceivable difference between a "natural" SDM decomposed SRIR without any added strong ceiling reflections when mapping all reflections to the horizontal plane opposed to mapping the elevated reflection components of the measured room to a single elevated loudspeaker. Spatio-temporal analyses of these four decomposed SRIRs are included in Appendix A.

3.4 Stimuli Rendering

Based on the decomposed SRIRs described in Section 3.3 and the isolated reflection measurements from Section 3.2, a total of 100 different stimuli with two different

Room Name	$T30_{1kHz}$	$C80_{1kHz}$	EDT_{1kHz}	$\Delta_{Ceiling}$
Big Hall	1.43 s	7.53 dB	1.17 s	45.8 ms ≈ 15.7 m
Listening Lab	0.10 s	46.09 dB	0.17 s	3.1 ms ≈ 1 m

Table 3.1: Acoustic parameters of measured rooms. The T30, C80 and EDT values were calculated using the ITA-Toolbox [30]

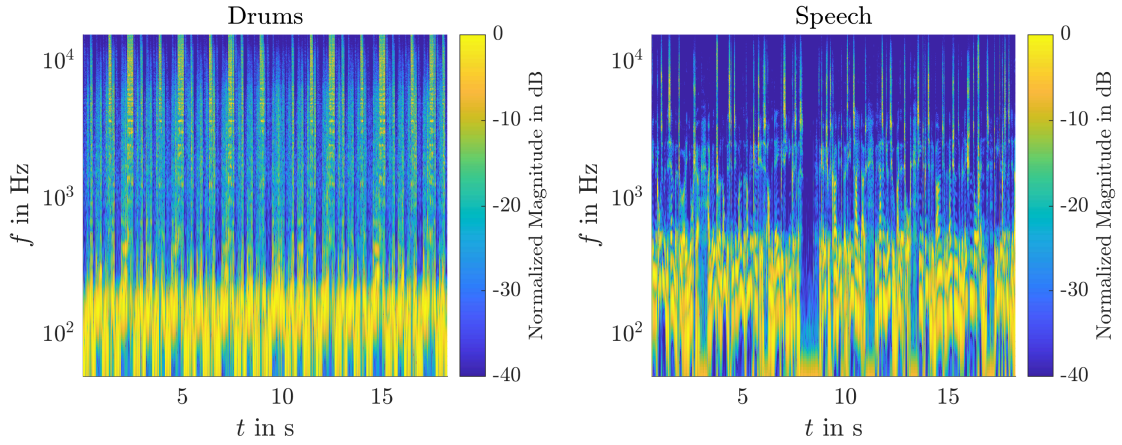


Figure 3.5: Source Signal Spectrograms

source signals and varying reflection positions, reflection levels and SRIR room components were rendered. This resulted in 50 different stimuli comparisons whereby for each comparison, the only difference between the two compared stimuli was the reflection position.

During the preliminary experiments, it was found that a drum recording seems to be a suitable source signal in the context of these experiments as it contains impulsive broad band components and is a well known recognizable sound. Therefore, one channel of the anechoic drum recording included in the Matlab Audio Toolbox was used as main source stimulus. Additionally, it was decided to also include the SQAM 50 speech signal from the EBU sound quality assessment material recordings for subjective tests [31] in the evaluation. Spectrograms of both unprocessed source signals are shown in Figure 3.5. In order to achieve a similar perceived loudness, both signals were normalized to the same RMS level.

In total, three different room scenarios were evaluated: The direct signal with added reflection but without any added SRIR, this case is in the following also referred to as “dry”, the direct signal with added reflection and the “big hall” SRIR as well as the direct signal with added reflection and the “listening lab” SRIR. With the exception of one special case described in the last paragraph of this section, the SRIR room components were always mapped to eight virtual loudspeakers on the horizontal plane.

For this combination of two source signals and three different room scenarios, 45° and 10° elevated reflections on the median plane as well as 45° elevated reflections on the frontal plane were compared to a corresponding 0° elevated reflection. Additionally, a 45° elevated reflection on the median plane was compared to a 45° elevated

reflection on the frontal plane. All these cases were evaluated with different reflection levels which were determined as described in Section 3.4.1, whereby the range of reflection levels to evaluate for each scenario was set based on preliminary experiments. In order to limit the overall experiment duration to a acceptable range, only up to four different reflection levels per comparison were considered.

Lastly, the relevance of elevated reflections for a measured SRIR without any added strong ceiling reflections was evaluated by comparing both the "big hall" and "listening lab" SRIRs quantized to eight loudspeakers on the horizontal plane to the corresponding SRIRs quantized to eight speakers on the horizontal plane and one additional top speaker with 90° elevation (see Section 3.2 and Appendix A).

3.4.1 Reflection Level

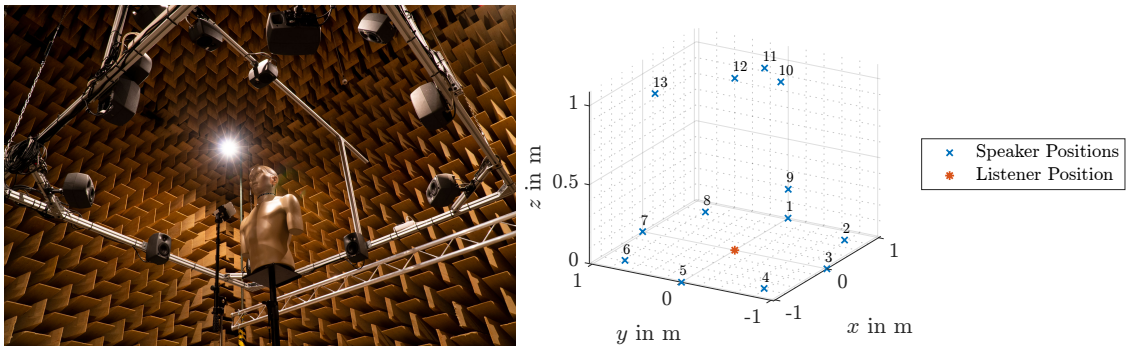
In order to specify the level of the added reflection, it was decided to use the RMS of the signals obtained from convolving the source signal with the SDM decomposed SRIRs and the reflection IR as reference. Thereby, the reflection level L_{ref} was determined by comparing the RMS of the isolated reflection signal $y_{\mathrm{ref}}(n)$ which equals the source signal convolved with the reflection IR to the signal mapped to the center speaker $y_{\mathrm{c}}(n)$ following

$$L_{\mathrm{ref}} = 20 \cdot \log_{10} \frac{\frac{1}{N_{\mathrm{ref}}} \sqrt{\sum_{n=1}^{N_{\mathrm{ref}}} |y_{\mathrm{ref}}(n)|^2}}{\frac{1}{N_{\mathrm{c}}} \sqrt{\sum_{n=1}^{N_{\mathrm{c}}} |y_{\mathrm{c}}(n)|^2}} \ . \tag{3.1}$$

For the stimuli with no room, $y_{\mathrm{c}}(n)$ corresponds to the source signal while for stimuli with added room, $y_{\mathrm{c}}(n)$ corresponds to the source signal convolved with the part of the SRIR which was quantized to the virtual center speaker position. This means that the experiment in its simplest form can be accurately reproduced by playing back a delayed and by L_{ref} attenuated version of the direct signal via an elevated reflection speaker.

3.5 Loudspeaker Based Listening Experiment

3.5.1 Setup



(a) (b)

Figure 3.6: Loudspeaker array setup (a) and geometry (b).

Speaker Nr.	1	2	3	4	5	6	7	8	9	10	11	12	13
Azimuth	0°	315°	270°	225°	180°	135°	90°	45°	0°	0°	0°	0°	0°
Elevation	0°	0°	0°	0°	0°	0°	0°	0°	10°	45°	60°	90°	45°
r	1.03 m	1.10 m	1.03 m	1.10 m	1.03 m	1.10 m	1.03 m	1.10 m	1.05 m	1.26 m	1.19 m	1.09 m	1.26 m

Table 3.2: Spherical coordinates of loudspeaker positions.

As shown in Figure 3.6a, the loudspeaker based listening test setup consisted of 13 Genelec 8020 loudspeakers installed in the anechoic chamber of the Division of Applied Acoustics. Thereby, speakers 1-8 were arranged in a circle with a radius of approximately 1 m on the horizontal plane, speakers 9-12 were positioned at an elevation angle of 10°, 45°, 60° and 90° on the median plane and an additional speaker was positioned at 45° elevation and 90° azimuth, i.e. to the left of the participant position. Figure 3.6b and Table 3.2 show the exact geometry as well as the spherical loudspeaker coordinates. Speaker 11 was installed but not actually used for the listening experiment. The multi-channel audio playback and listening test routines were implemented in Matlab and an Antelope Orion 32 audio interface was used for the D/A conversion. Additionally, a smartphone with a custom GUI was set up to remotely control the Matlab process and thereby act as experiment user interface.

3.5.1.1 Collateral Reflections

Initially, it was planned to seat the participants on a chair placed in the center of the array as shown in Figure 3.7. However, as the listening test environment for this experiment should be as anechoic as possible it was decided to measure the transfer paths from the individual array speakers to a microphone placed at the receiver position in order to determine if the chair causes significant reflections which might influence the experiment. Indeed, these measurements revealed that the chairs armrests and seating surface cause reflections and thereby interferences which

Figure 3.7: Initial loudspeaker array setup with chair.

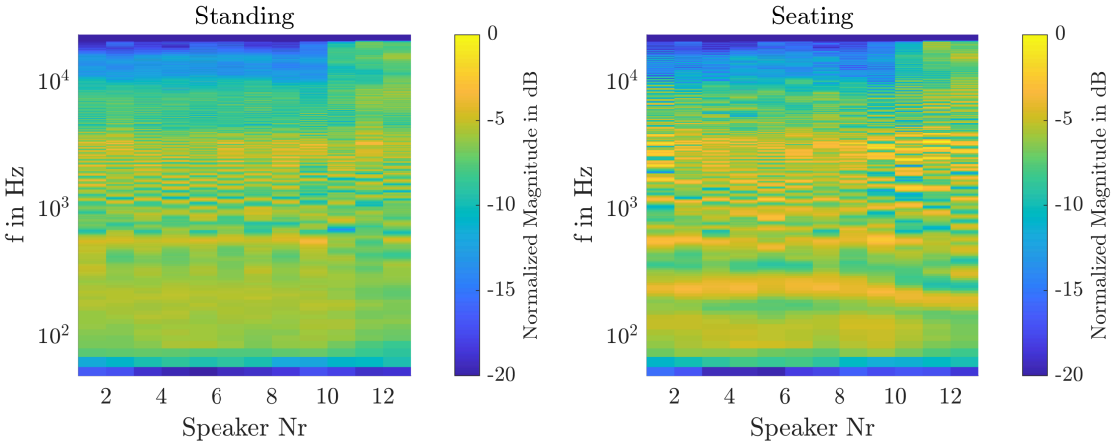


Figure 3.8: Normalized transfer functions between individual loudspeakers and participant position for both standing and seating setup. The right plot shows that, for the seating setup, e.g. the first interference peak around 250 Hz varies in frequency by more than 50 Hz depending on the loudspeaker position.

vary depending on the loudspeaker position. As shown in Figure 3.8, especially changes in the elevation angle lead to variations in the run time difference and ratio between direct and reflected sound waves hence the frequency and magnitude of the interference changes. This is problematic in the context of this experiment as any spectral differences between the individual loudspeakers might cause unwanted tonal differences between stimuli with different reflection positions. This could then lead to participants sensing differences which are not caused by the stimuli themselves but by these interferences due to additional reflections. In order to keep the amount of unwanted reflections as low as possible, it was decided to let the participants stand on a small wooden plate instead of seating them down. While the wooden plate and also the array construction itself still caused unwanted reflections, this way at least additional reflections from the chair and the angled participant legs were avoided.

3.5.1.2 Calibration

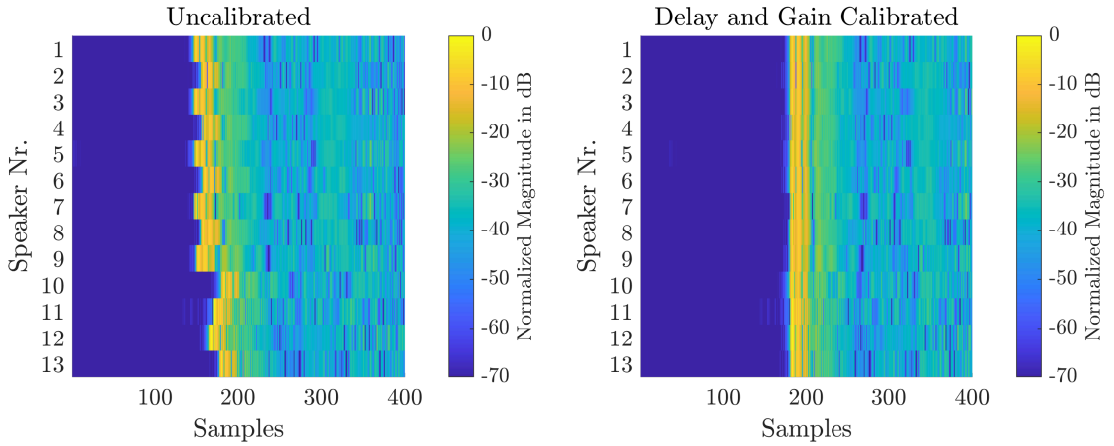


Figure 3.9: Uncalibrated and delay and gain compensated loudspeaker IRs for standing position.

The loudspeaker array was calibrated in delay and amplitude by measuring the transfer path from each speaker to a microphone placed at the participant position using the sweep deconvolution method. The resulting IRs where then analyzed regarding on-set time and magnitude peak, from these results correction gain and delay values for each speaker were calculated. A second measurement which included those calibration values confirmed that all speakers afterwards matched in volume and phase at the receiver position. The overall playback volume was set to a comfortable level of approximately 64 dB_A and all participants were instructed to always keep their head oriented towards the center speaker and to avoid head movements while listening to the stimuli.

3.5.2 Listening Test Procedure

As the preliminary experiments showed that the perceived differences between the individual stimuli can be expected to be relatively small, it was decided to use the ABX test method in order to determine if participants are able to hear a difference between two different reflection positions. This means that the participants had to listen to three different sounds A, B and X whereof sound A and B were two different stimuli and sound X corresponded either to sound A or B. The participants then had the task to determine if X equals to sound A or sound B. This approach has the advantage that the participants are forced to make a binary decision and, given a sufficient number of participants, the statistical evaluation of the results provides reliable evidence whether the participant group perceived a difference between two stimuli. Both the order of experiments and the assignment of the different stimuli to A, B and X was randomized.

In addition to the standard ABX evaluation, two continuous sliders were added to the experiment user interface in order to query the perceived tonal and spatial difference between the two sounds A and B. Thereby, the scale was set from “no perceived difference” at the 0% slider position to “large perceived difference” at the

100% slider position. All participants were instructed to use this scale relative to the stimuli they heard during a training phase at the beginning of the experiment which consisted of 10 comparisons including all "extreme cases" of large and small differences between sounds A and B. Hence if a participant reported 100% perceived spatial difference during the actual experiment, this means that the perceived difference was as large as the maximum perceived difference during the training phase. In addition to establishing a reference for the range of differences between the stimuli, this training phase was also used to familiarize the participants with the user interface.

A commonly used implementation of the ABX test is to play the A, B and X sounds only once per trial in a fixed order directly after each other and conducting several trials per comparison. However, for the listening tests performed in this thesis the participants were allowed to decide on their own when and in which order they want to listen to the ABX stimuli and only one trial per stimuli comparison was conducted. This has the disadvantage that it is unclear how often each participant actually switched between A, B and X which could potentially lead to a bias and, from a statistical point of view, it would of course have been beneficial to get multiple results per stimuli comparison from each participant. However, including the training phase the listening test consisted of 60 different stimuli comparisons and the experiment already took ca. 30 minutes per participant. Given the fact that the participants had to stand for the entire time and the task of identifying small auditory differences can be very exhausting, it was decided to not extend the experiment duration by repeating individual stimuli comparisons. While enforcing a fixed ABX playback order instead of giving the participants control over which sound is currently played could eventually have led to more comparable ABX results, additionally deciding on how large the perceived spatial and tonal differences were after listening to the sounds only once would have been an extremely different task for the participants hence one could expect unreliable results.

3.5.3 User Interface

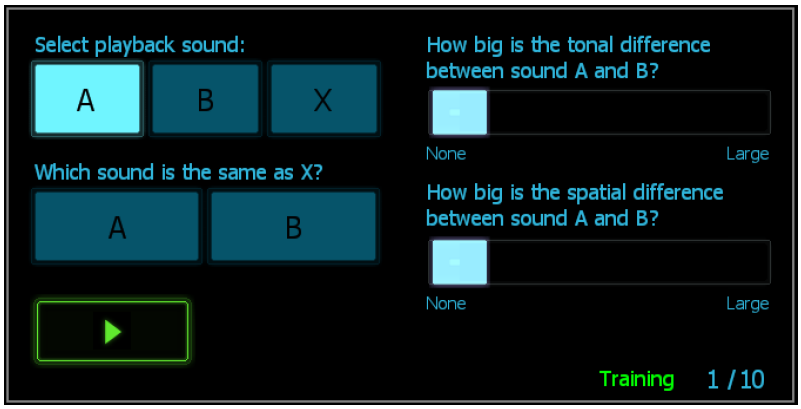


Figure 3.10: Listening Experiment User Interface

The experiment user interface itself was implemented using Lemur[1], a commercial iOS app that allows the creation of custom MIDI control interfaces. This app was installed on a smartphone placed at the participant position and sent MIDI commands to a Matlab PC which handled the multi-channel audio playback. By using a smartphone as user interface, the amount of unwanted reflections from the screen was significantly less than if a setup including a laptop or computer screen would have been used.

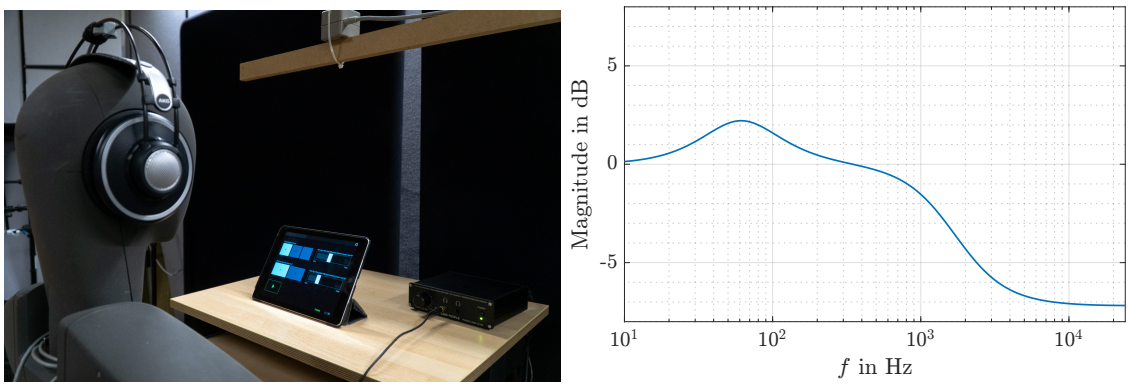
As shown in Figure 3.10, the GUI itself consisted of three buttons which allowed the user to select which sound they want to listen to, a play and pause button, a pair of buttons to answer the ABX test, two sliders to state the perceived tonal and spatial difference as well as a counter which showed the overall experiment progress and an indicator whether or not the participants are currently in the training phase. Once a participant listened to all three stimuli, made a selection for the ABX comparison and touched both sliders once, a "next" button appeared allowing the user to proceed to the next test. This ensured that the participants could not accidentally finish a test without fulfilling all the tasks.

3.5.4 Participants

The loudspeaker based experiment was performed by a group of 25 different participants consisting of students from the "Sound and Vibration" Master's programme, PhD students in the field of applied acoustics and a small number of subjects without an academic background in acoustics. All of the participants reported to have normal hearing, 22 participants stated to have a background in acoustics and 11 participants claimed to have experience with critical listening in the context of spatial audio.

[1] `https://apps.apple.com/us/app/lemur/id481290621`, accessed 12.05.2012

3.6 Headphone Based Listening Experiment



(a) (b)

Figure 3.11: Headphone based listening experiment setup (a) and applied equalization curve (b).

In order to evaluate if the results from the loudspeaker based experiment are also valid for headphone based spatial audio reproduction, a second listening experiment was performed. Therefore, the same stimuli used for the loudspeaker based experiment were convolved with HRIRs for the corresponding loudspeaker position and then played back using AKG K-702 headphones, an Antelope Orion 32 audio interface and a Lake People G109 headphone preamplifier.

The used HRIR data was obtained from measurements of the division's own KEMAR artificial head conducted by the ITA Aachen [12]. In order to compensate timbral unbalances caused by this specific combination of HRTFs and headphones, the playback channel was was equalized by ear using a biquad multi-band EQ as shown in Figure 3.11b.

Similar to the first experiment, the participants were instructed to always keep their head orientated towards the virtual center speaker. However, in order to ensure a realistic spatial audio impression and to account for small head movements the participants' head position was tracked using a Polhemus Patriot magnetic tracking system and the whole binaural scene was rotated according to the participants current head orientation. The binaural playback was implemented in Matlab using the same remote control application as described in Section 3.5.3. For the loudspeaker based experiment, a 4.7" smartphone was used as control interface in order to reduce unwanted reflections from the screen. However, for the headphone based experiment reflections from the surrounding are not relevant hence a larger 9.7" tablet was used in order to provide a more comfortable user interface as shown in Figure 3.11a. The experiment took place in an acoustically treated listening lab and, unlike in the loudspeaker based listening experiment, the participants were allowed to adjust the playback volume to a comfortable level during the first ten training tests. This means that different participants possibly chose slightly different playback levels based on their personal preference and that these levels do not match the playback volume of the loudspeaker based experiment. However, as all participants stated

to have normal hearing one can assume that the individually selected comfortable playback levels were in a similar range as the approximately 64 dB_A chosen for the loudspeaker based experiment.

The headphone based based experiment was performed by a group of 13 different participants consisting of students from the "Sound and Vibration" Master's programme and PhD students in the field of applied acoustics. All of the participants reported to have normal hearing and stated to have a background in acoustics, 9 participants claimed to have experience with critical listening in the context of spatial audio. 10 of the 13 participants already performed the loudspeaker based experiment before.

3.7 Data Analysis

3.7.1 ABX Test Evaluation

ABX tests are widely used in the audio engineering community to evaluate whether or not audible differences between two stimuli exist. While there is no standardized ABX method that would specify certain implementation aspects, e.g. if the participants have control over the stimuli playback order or if multiple trials per stimuli comparison and participant are performed, the fundamental evaluation of the ABX results is always similar. For the following ABX results analysis method, [32] was used as the main reference.

The basic concept of an ABX test is that the participants listen to three different stimuli "A", "B" and "X" where A and B are two different sounds while X corresponds to either A or B. Thereby, the assignment of the two different stimuli to A, B and X has to be randomized for each trial. The listener's task is then to identify whether sound X is the same as stimulus A or the same as stimulus B. If there is a, for the specific listener, perceivable difference between sound A and B, the participant will most likely make a correct selection. This case is also referred to as correct ABX identification. Putting the number of correct ABX identifications in relation to the absolute number of trials for a specific stimuli comparison yields in a percentage of correct identifications.

If there is no audible difference between A and B, the participant will simply guess which should lead to binomially distributed answers when repeated over a large number of trials. I.e. when using two stimuli with no audible difference, assigning them randomly to A, B and X and performing a large amount of trials, one would expect that 50 % of the results are correct and 50 % are false identifications. This is an important characteristic of ABX test: If there is no perceivable difference between the two A and B stimuli, an ideal ABX experiment would still yield in 50 % correct identifications. If an ABX result shows 0 % correct identifications, this does not mean that none of the participants perceived a difference but that all participants managed to constantly identify the wrong answer. This would indicate some kind of bias, e.g. the participants misunderstood the task or something went wrong during the stimuli randomization. From these basic considerations one can conclude that, given a sufficient number of trials and an unbiased experiment setup,

100 % correct ABX identifications for a specific stimuli comparison means that there is an audible difference while 50 % correct identifications indicate that there is no perceivable difference. However, in practice it is unlikely that the percentage of correct identifications exactly equals 50 % or 100 %, hence a method is needed to determine a minimum percentage of correct identifications required to state, with a certain confidence, that there is an audible difference between two stimuli.

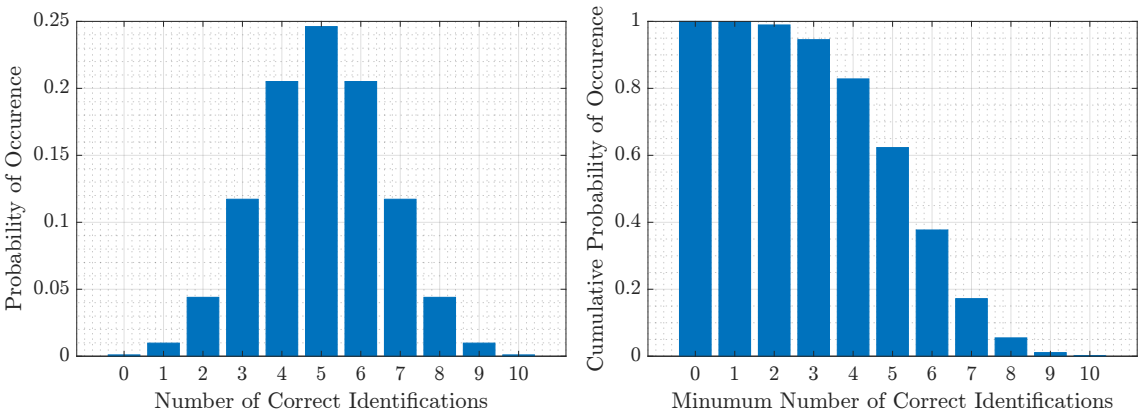


Figure 3.12: Binomial probability mass function and inverse binomial cumulative distribution function for $n = 10$ trials.

Assuming that the ABX experiment is randomized and that therefore the participants' answers are random and uncorrelated to the stimuli when there is no audible difference, one can use the binomial distribution to determine such a confidence threshold. Figure 3.12 shows the binomial probability mass function as well as the inverse cumulative distribution function for $n = 10$ trials, calculated as described in [32]. Thereby, the binomial probability mass function indicates the probability of a specific number of correct responses to occur for random binomial data with n trials. A simple example for this would be flipping a coin, where one side of the coin is considered as "correct" and the other side as "false". According to the binomial probability mass function, the chance of getting exactly 5 out of 10 correct coin flips is 0.25 i.e. 25% while chance of getting exactly 10 out of 10 correct results is only 0.1 %. However, to determine a threshold of how many correct answers are at least required in order to exceed a certain confidence threshold, the probability for a specific amount or more correct answers is of interest. This probability can be obtained from the inverse binomial cumulative distribution function as shown in the right plot of Figure 3.12, which simply represents the sum of all binomial probabilities equal or above a certain number of correct responses. E.g. the chance of getting 0 or more correct responses is 100 % since it equals the probability of getting exactly 0 correct responses plus the probability for getting exactly 1,2, ... 10 correct responses. The cumulative probability of getting 8 or more correct responses is only 5.5 % which, in the context of ABX tests, means that 8 or more correct identifications for 10 trials only occur with a 5.5 % probability when the answers are binomial distributed. I.e. the chance that there was no audible difference between the two stimuli and 8 or more out of 10 participants guessed correctly is only 5.5 %. In other words, the probability of the answers being not binomial distributed equals 100 % - 5.5 % = 94.5 %, which corresponds to the confidence level of an audible

difference for 8 or more correct IDs out of 10 trials in an ABX test. This way, a threshold that determines the minimum number of correct IDs for a given number of trials in order to state with a 95 % confidence that there is an audible difference between two stimuli can be calculated. For the two experiments performed in this thesis, these thresholds lie at 17 or more correct IDs or 68 % correct identifications for the loudspeaker based experiment with 25 participants and 9 or more correct IDs for the loudspeaker based experiment with 13 participants, which equals to 68 % correct identifications.

For completeness, it should also be stated that there is a minimum threshold below which the results can not be considered as binomial distributed since, as described before, 0 out of 10 correct identifications means that all participants managed a wrong identification which indicates a strong negative correlation between the answers and the stimuli. This limit can be determined in a similar way as the 95 % confidence threshold described earlier but, as none of the obtained ABX test results fall below this limit, it will be omitted in the presentation of the results. Besides this fundamental evaluation method, more advanced ABX analysis approaches using signal detection theory [32] exist but will not be applied in this thesis.

3.7.2 Perceived Difference Values

Since all participants used the same set of training stimuli as reference for the range of perceived differences, the arithmetic mean and 95% confidence intervals as well as median values were directly calculated from the participants' perceived difference answers without applying any kind of normalization. Additionally, relative probability histograms of the perceived difference values were compared for different scenarios as presented in Section 4.2.5. Relative probability is in this context also referred to as relative frequency and simply indicates the number of results that fall into a specific category divided by the total number of results [33]. E.g. assuming a participant reported a perceived spatial difference in the range between 0% and 10% for 20 out of 50 stimuli comparisons, the relative probability for this range is $20/50 = 0.4$. Visualizing the relative probability in form of histograms allows a straightforward comparison of answer distributions for different scenarios.

3.7.3 Difference in Interaural Cross-Correlation

In order to analyze the exact stimuli the participants heard during the listening experiment including possible collateral reflections caused by the loudspeaker setup, a KEMAR artificial head was placed at the listener position and all stimuli were binaurally recorded through the loudspeaker array. For the headphone based experiment, the rendered binaural stimuli for a 0° head rotation were used for further analyses.

For all stimuli from both experiments, the normalized interaural cross-correlation between both ear signals was calculated as described in Equation (2.1) by using the Matlab `xcorr()` function. For the evaluation of the results, the change in maximum IACC between both stimuli for each test was calculated as

$$\Delta\mathrm{IACC}_{\mathrm{norm}} = |\max(|\mathrm{IACC}_{\mathrm{norm},1}(\tau)|) - \max(|\mathrm{IACC}_{\mathrm{norm},2}(\tau)|)| \tag{3.2}$$

which corresponds to the Matlab expression

```
dIACC =  abs(max(abs(xcorr(yl1, yr1, 'normalized'))) - max(abs(
   xcorr(yl2, yr2, 'normalized'))))
```

This IACC difference can be considered as important in the context of this thesis since, as described in Section 2.1.3.1, the amount of correlation between both ear signals allows to draw conclusions on the apparent source width, listener envelopment and diffuseness of a room. Without going more into detail on the specific relation between the IACC and those parameters, it can be noted that a change in the maximum normalized IACC between two stimuli can indicate a difference in their spatial impression. Related research such as [18] and [34] showed that indeed, the human auditory system is remarkably sensitive regarding changes in the interaural cross-correlation. Depending on parameters such as the overall IACC and the source signal, it has been shown that listeners can detect the difference between an IACC of 1.00 and 0.99, i.e. an IACC fluctuation of 0.01 in an otherwise diotic noise [17]. These findings legitimate the approach to treat the change in maximum normalized IACC between two stimuli as possible indicator for the amount of perceived differences when varying an early reflection's elevation angle.

4

Results

4.1 Loudspeaker Based Experiment

4.1.1 Graphical Representation of the Results

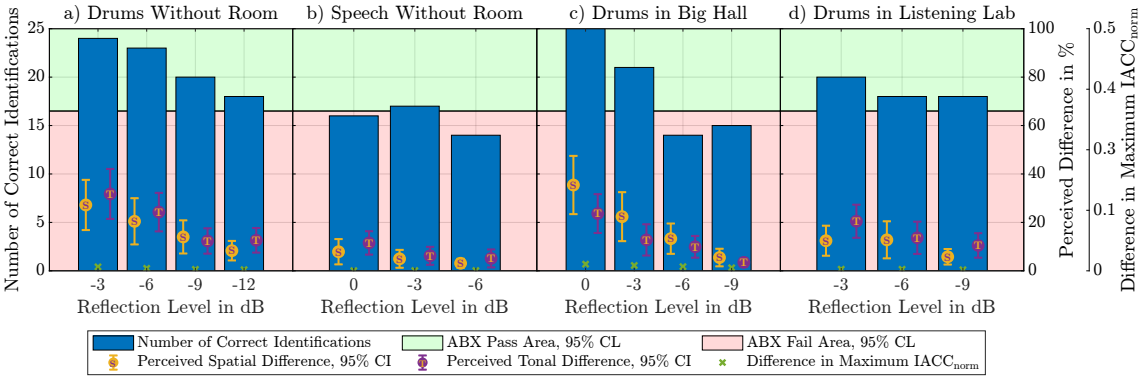


Figure 4.1: Loudspeaker based listening test results for comparison between a reflection with an elevation angle of 45° and 0° at an azimuth angle of 0° with varying reflection levels, source stimuli and room IRs.

Figure 4.1 shows the loudspeaker based experiment results comparing reflections with an elevation of 45° and 0° at an azimuth of 0° with varying reflection levels, source stimuli and room IRs. The sum of correct ABX identifications of all 25 participants is represented by the blue bar graphs. Thereby, the green area behind the bars represent the regions in which, with a 95% confidence level, the sum of correct identifications of all participants is large enough to state that, statistically, the participant group identified a difference between the both compared stimuli. For 25 participants, this threshold lies at 17 correct identifications (see Section 3.7.1). If a bar lies in the red area, this means that statistically not enough participants correctly identified the X stimulus to state that the entire group of participants perceived a difference in this comparison. Of course not only this threshold is of relevance, i.e. if all 25 participants correctly identified the X stimulus it is reasonable to assume that the perceived difference between both stimuli is larger than if 17 of 25 participants made a correct identification, even though in both cases the sum of correct identifications exceeded the 95 % confidence threshold.

Additionally, Figure 4.1 contains error-bars which represent the arithmetic mean as well as the 95% confidence interval of the perceived spatial and tonal differences

between both reflection positions which was obtained from the two difference sliders in the experiment's user interface (see Section 3.5.3).

The last parameter represented in Figure 4.1 is the difference in the maximum normalized IACC between the two compared stimuli, measured by placing a KEMAR artificial head at the participant position and analyzing the recorded signals (see Section 3.7.3). The absolute values for all plotted parameters, i.e. the number of correct identifications, the perceived difference and the IACC difference can be read off the figure's three different Y axes.

For the evaluation of the results, it is reasonable to take all those parameters into account as especially the combination of the sum of correct identifications and the average perceived spatial difference allows a meaningful interpretation on how much of an effect a change in reflection position had on the spatial auditory perception of the presented sounds.

4.1.2 Elevated Reflections on the Median Plane

4.1.2.1 45° Elevated Reflection on the Median Plane

Figure 4.1 shows the loudspeaker based results for the most fundamental comparison between a reflection on the median plane, i.e. with 0° azimuth in front of the participant, switching between an elevation angle of 0° and 45°. In the 0° case, the reflection signal was played back via the same loudspeaker as the direct signal.

Looking at Figure 4.1a it is obvious that, for all tested reflection levels, the group of participants was able to correctly identify the difference between both reflection elevations when listening to the drum signal without any added reverberation. However, the number of correct identifications drops with decreasing reflection level but even at -12 dB the number of correct IDs still exceeds the 95 % confidence threshold. The perceived spatial and tonal difference sliders show a very similar result as their mean values drop with decreasing reflection level. At -12 dB, they are both in a range below 20 % which, in combination with the number of correct identifications barely exceeding the 95% confidence threshold, can be interpreted so that the difference between both reflection positions is very close to the perceptional threshold. However, it is noteworthy that, for this case, the perceived tonal and spatial differences are very similar to each other which means that, from these results, there is no clear distinction if an elevated reflection on the median plane causes more spatial or more tonal differences.

For a speech signal without added room reverberation on the other hand, the perceived difference between 0° and 45° elevation of a reflection on the median plane is not as pronounced, as shown in Figure 4.1b. For reflection levels of 0 dB and -6 dB, the sum of correct identifications is below the 95% confidence threshold, at -6 dB the number of correct IDs is slightly above this threshold. The latter is unexpected as one would assume that, if a group of participants can not distinguish two stimuli at a reflection level of 0 dB, the difference at a reflection level of -3 dB should neither be audible. This specific result might be an outcome of inaccuracies in the ABX method. Since the perceived spatial and tonal difference values for this comparison are in the same range as the values for the -6 dB case shown in the same

plot, one can assume that, even though the number of correct identifications reaches the 95 % confidence region for a -6 dB reflection level, this might be a result of the participants by chance guessing the correct answer instead of actually perceiving a difference.

Adding spatial reverberation from the "big hall" SRIR to the drum signal increases the threshold for a correct identification from -12 dB to -3 dB compared to the drum signal with no added room as shown in Figure 4.1c. At a reflection level of -6 dB, the average answers for the perceived spatial and tonal differences are still comparatively high even though only 14 of 25 participants identified the stimuli correctly. This could indicate that either the participants which gave a correct answer reported a high perceived difference or that some of the participants thought they perceived a spatial or tonal difference but still could not identify the correct stimulus for sound X. An other interesting observation is that, for the drum signal with added "big hall" SRIR, the mean value of the perceived spatial difference slightly exceeds the one of the perceived tonal difference whereas in the case of drums without added room, the mean value of the tonal difference is slightly higher. As the confidence intervals of both values overlap each other it is not possible to draw accurate conclusions from this observation, but one can at least make the assumption that the added diffuse reflections of the big hall could mask the tonal differences between both reflection positions.

The results for the drum signal with added room components from a small listening lab seem to support this theory as for this case, where the room response contains less diffuse reflections than the big hall, the perceived tonal difference values are again slightly higher than the perceived spatial difference. Apart from this, the number of correct identifications is slightly lower for this case than for the drum signal with no added room but it still exceeds the 95% confidence threshold for all tested reflection levels.

For all comparisons shown in Figure 4.1, the difference in IACC is almost zero and, independent on the source signal and added room, does not significantly vary for different reflection levels of the same stimulus type. This is to be expected as the only difference between the two compared stimuli for this case is the elevation on the median plane which means that, apart from some reflections by the loudspeaker array itself, inaccuracies in the positioning of the artificial head and asymmetries in its HRTFs, exactly the same signal should arrive at both ears, independent from the elevation angle.

From this first evaluation one can already draw several assumptions, which are to be confirmed by the following results. Firstly, a decreasing reflection level leads to a smaller perceived difference between different elevation angles. Additionally, added room reflections seem to raise the threshold of how loud a reflection needs to be in order to lead to a noticeable difference when its elevation is changed. Thereby, a room with a longer reverberation time seems to result in a stronger masking of both the overall perceived difference and of the tonal differences compared to the spatial differences. Lastly, Figure 4.1 already reveals that spatial differences between different reflection positions are also perceived when there is no difference in interaural cross-correlation, which means that a change in IACC between two

stimuli alone is not an exclusive indicator for perceived spatial differences.

4.1.2.2 10° Elevated Reflection on the Median Plane

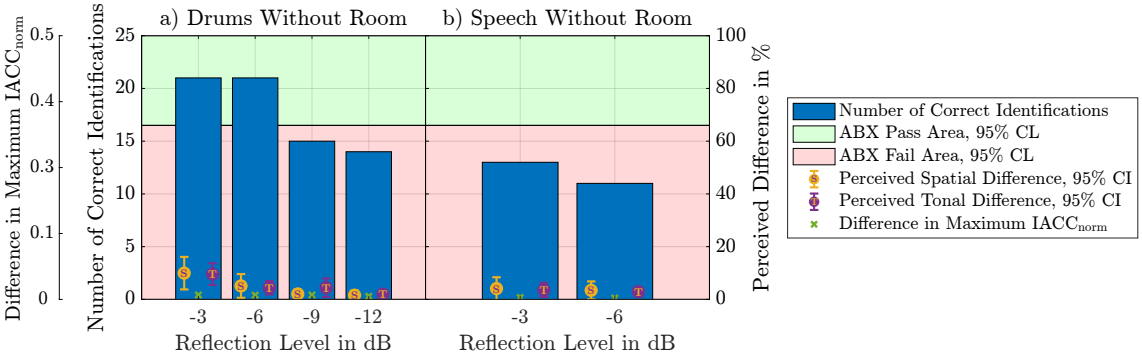


Figure 4.2: Loudspeaker based listening test results for comparison between a reflection with an elevation angle of 10° and 0° at an azimuth angle of 0° with varying reflection levels and source stimuli.

Figure 4.2 shows the loudspeaker based listening test results for a 10° elevated reflection on the median plane compared to a reflection at 0° elevation. In order to restrain the total amount of comparisons to a reasonable level, this test was only conducted for the drums and speech signal without added room. Comparing Figure 4.2 to the results of a 45° elevated reflection on the median plane from Figure 4.1 shows that, at least in this case, a smaller difference in elevation angle results in smaller audible differences which is reflected both in a higher reflection level threshold for the drum signal without added room shown in Figure 4.2a as well as significantly smaller perceived spatial and tonal difference values. For the speech signal, the difference between 45° elevation and 10° elevation is not as significant as, presumably, the participants did not perceive a pronounced difference in neither of the two cases.

4.1.3 45° Elevated Reflection on the Frontal Plane

Figure 4.3 shows the loudspeaker based experiment results for the comparison between a 45° and a 0° elevated reflection at an azimuth angle of 90°. This means that the reflection in this case was located to the left of the participant. Thereby, the participant group managed to correctly identify all tested stimuli with a confidence level of at least 95%, even for reflection levels as low as -12 dB. Comparing this to the results for a 45° elevated on the median plane shown in Figure 4.1 clearly indicates that, at least for the evaluated stimuli and used the loudspeaker setup, a difference in elevation of a lateral early reflection causes a larger perceptual difference than a difference in elevation of a frontal early reflection. This is also visible in the perceived spatial difference values which are overall higher than their counterparts for the reflection on the median plane and, in contrast to the case of a frontal reflection, are always higher than the perceived tonal difference values. While the analysis of the stimuli with an added early reflection on the median plane did not reveal a significant difference in interaural cross correlation, Figure 4.3 shows that

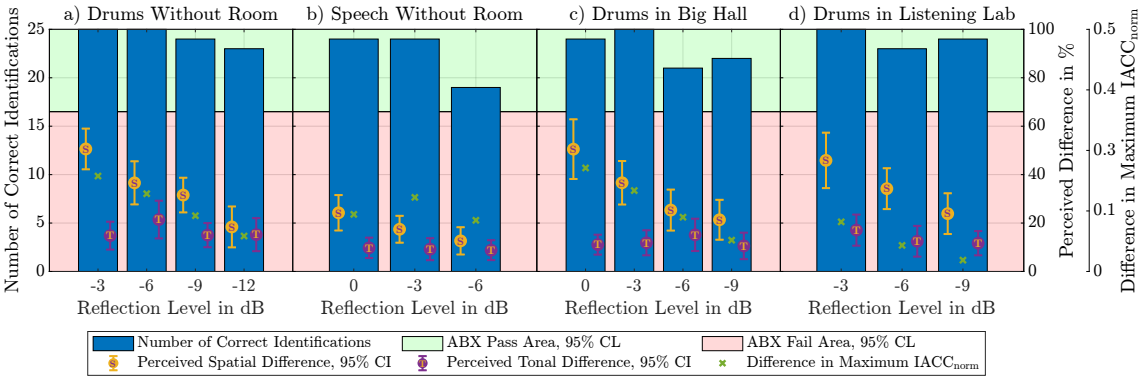


Figure 4.3: Loudspeaker based listening test results for comparison between a reflection with an elevation angle of 45° and 0° at an azimuth angle of 90° with varying reflection levels, source stimuli and room IRs.

changing the elevation of an early reflection on the frontal plane results in a noticeable IACC difference. Thereby, the IACC difference decreases with decreasing reflection level which is to be expected as a lower reflection level means that the two compared stimuli with different reflection elevation angles are more similar to each other. Additionally, the results show a clear correlation between IACC difference, perceived spatial difference and the number of correct identifications. Whether or not this correlation indicates an actual causality will be discussed in Section 5.5.

4.1.4 45° Elevated Reflection Altering Between Frontal and Median Plane

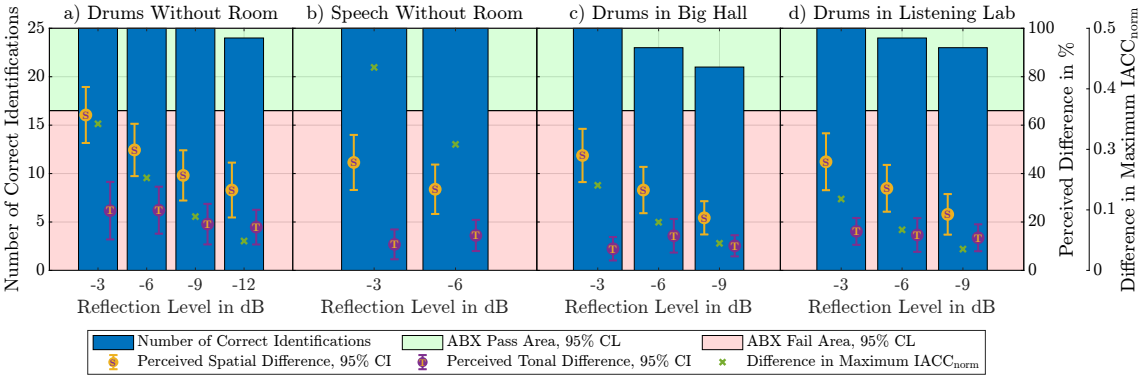


Figure 4.4: Loudspeaker based listening test results for comparison between a reflection with an azimuth angle of 0° and 90° at an elevation angle of 45° with varying reflection levels, source stimuli and room IRs.

As shown in Figure 4.4, changing the azimuth angle of a 45° elevated reflection by 90° results in clearly audible differences for all evaluated reflection levels, source signals and room scenarios. Thereby, the amount of correct ABX identifications always exceeded the 95 % confidence threshold and the perceived difference values indicate that the spatial differences predominate the tonal differences. This is to be expected, since a lateral displacement leads to a change in interaural cues which are

in general more robust than monaural cues. This can also be seen from the difference in maximum normalized IACC between each of the two compared stimuli which, for these comparisons, correlates very well with the perceived spatial differences. However, as the purpose of this thesis is to evaluate the effect of changing a reflection's elevation angle, these results for a change in azimuth are only of secondary interest.

4.1.5 Natural SRIR Without added Reflection

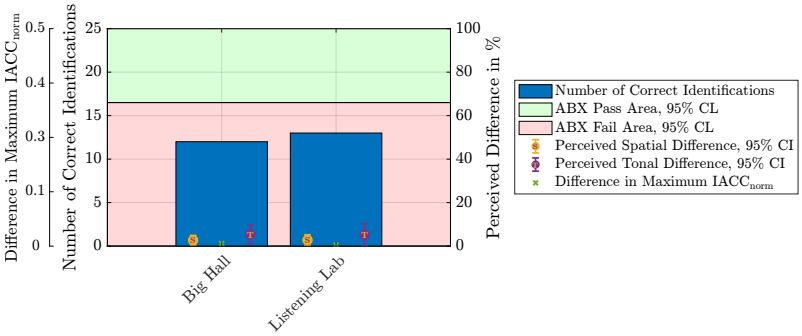


Figure 4.5: Loudspeaker based listening test results for comparison between SDM SRIR rendered with added loudspeaker at 90° elevation and SDM SRIR rendered only for loudspeakers on the horizontal plane. Only the drums signal was evaluated and just the natural room IRs of Big Hall and Listening Lab were used.

The last and maybe in practice most relevant comparison is whether or not there is an audible difference when assigning the natural reflections of a measured SRIR without added strong ceiling reflection to an elevated loudspeaker compared to mapping them to the horizontal plane. Figure 4.5 shows the results for this comparison including two different rooms, thereby only the drum signal was evaluated. Both the number of correct ABX identifications and the perceived spatial and tonal difference values indicate that, using a loudspeaker based reproduction method, the natural elevated reflections in both rooms are not strong enough to result in a perceivable difference when playing them back through a single loudspeaker at an elevation angle of 90° instead of quantizing them to the horizontal plane.

4.2 Comparison between Loudspeaker Based and Headphone Based Results

In order to limit the extend of this thesis, it was decided to directly compare the headphone based results to the results from the loudspeaker array experiment instead of including a detailed description of the isolated headphone results at this point. In general, this comparison of loudspeaker and headphone based experiments can also be interpreted as comparing a reproduction method using individualized HRTFs to nonindividualized HRTFs. Thereby, only the results which were found the most significant are shown i.e. the percentage of correct identifications and mean values of perceived spatial difference for both reproduction methods. The

detailed result plots for the headphone based listening experiment are attached in Appendix B.

Following the method described in Section 3.7.1, the minimum number of correct identifications for a 95% confidence level equals to 17 out of 25 for the loudspeaker based experiment and 9 out of 13 for the headphone based experiment. In the following, the correct identifications are presented in percent instead of stating the absolute number of correct answers. As 17/25 = 68% and 9/13 = 69% are very close to each other, it was decided to use 69% correct identifications as the common threshold for the 95% confidence level for both experiments. Thereby, it is important to keep in mind that, due to the fact that the loudspeaker based experiment had 25 participants while the headphone based experiment included only 13 subjects, the statistical significance of both experiment results differs. Also, since 10 of the 13 participants of the headphone based experiment already participated in the loudspeaker based experiment one to two weeks before, it is possible that a certain training effect occurred and that those participants therefore were more successful in the headphone based experiment. In order to quantify this possible training effect, Section 4.2.5 compares the amount of correct ABX identifications of both experiments for those participants.

However, the purpose of this experiments was not to find a comprehensive answer on which reproduction method is more suitable for critical spatial audio listening but to evaluate if both methods lead to comparable results and if there is a general trend in the perceived differences. For this purpose, the acquired data should have a certain meaningfulness.

4.2.1 Elevated Reflections on the Median Plane

4.2.1.1 45° Elevated Reflection on the Median Plane

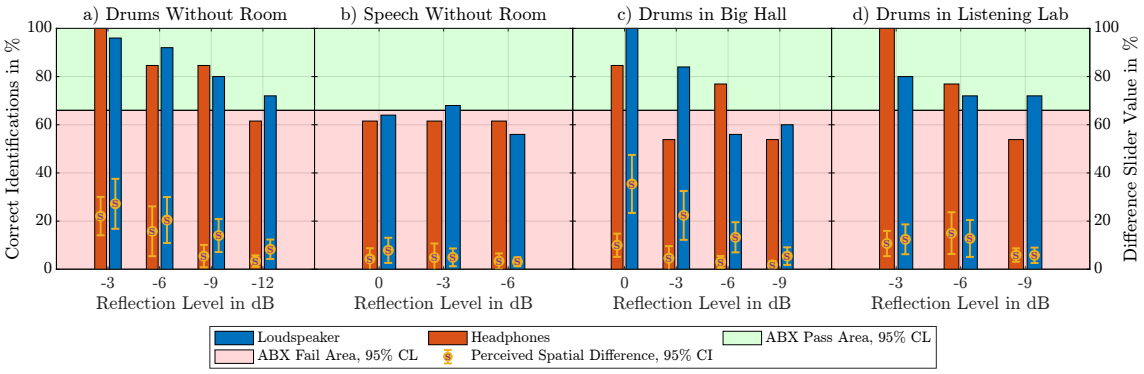


Figure 4.6: Loudspeaker and headphone based listening test results for comparison between a reflection with an elevation angle of 45° and 0° at an azimuth angle of 0° with varying reflection levels, source stimuli and room IRs.

Figure 4.6 shows the results from the loudspeaker and headphone based listening experiments comparing stimuli with an added 45° elevated early reflection on the median plane against a reflection with 0° elevation. As it can be seen, the results from the headphone based experiment generally follow a similar trend as the ones

from the loudspeaker experiment but there are some noticeable differences. From the results of the drum source signal without added room shown in Figure 4.6a, it is apparent that the threshold for a correct identification lies below -12 dB reflection level for the loudspeaker based experiment and between between -9 dB and -12 dB for the headphone based experiment. At a -12 dB reflection level, not enough participants of the headphone based test achieved a correct ABX identification to pass the 95% confidence threshold. The error bars in the plot show the perceived spatial difference values, thereby the values which overlay the blue bars represent the results from the loudspeaker based experiment and the error bars which overlay the orange bars stand for the perceived spatial difference of the headphone based experiment. It can be seen that, for the drum signal without added room, the perceived spatial difference in the headphone based experiment is always below the perceived spatial difference in the loudspeaker based listening test. Combining these results with the previous observation that the reflection level threshold for a correct identification is lower for the loudspeaker based experiment allows the conclusion that, at least for these specific stimuli, experiment setups and participant groups, the loudspeaker based reproduction method resulted in more perceivable differences when comparing the drum stimulus with an added 45° elevated early reflection on the median plane to a reflection with 0° elevation than the headphone based reproduction method.

The results for the drum signal with added reverberation from the big hall shown in Figure 4.6c are similar to the ones for the drum signal without added room. For these comparisons, the ABX evaluation of the headphone based experiment shows some inconsistencies as the percentage of correct identifications at a reflection level of -6 dB exceeds the 95% confidence threshold while the percentage of correct identifications for a reflection level of -3 dB lies below this threshold. However, given the fact that the perceived spatial difference values for the headphone based experiment at a reflection level of -6 dB are close to zero, as are the perceived tonal difference values shown in Figure B.1c, it can be assumed that the ABX test results contain an abnormal number of false positive identifications. This leads to the conclusion that, also for the drum signal with added big hall, the loudspeaker based reproduction method results in more perceivable differences between the different reflection positions than the headphone based method. This gets confirmed by the perceived spatial difference values which are significantly lower for the headphone based experiment than for the loudspeaker based listening test.

The results for the drum stimulus comparisons with added room components from the listing lab shown in Figure 4.6 are not as clear. While the ABX test results at -9 dB reflection level show a significantly higher percentage of correct identifications in the loudspeaker based experiment, the perceived spatial difference values do not differ that much between both methods.

Lastly, the results for the speech signal without added room shown in Figure 4.6b suggest that neither of the two reproduction methods resulted in a significant audible difference between the two different reflection positions.

Figure 4.7: Loudspeaker and headphone based listening test results for comparison between a reflection with an elevation angle of 10° and 0° at an azimuth angle of 0° with varying reflection levels and source stimuli.

4.2.1.2 10° Elevated Reflection on the Median Plane

Figure 4.7 shows the results for an added 10° elevated early reflection on the median plane compared to a reflection with 0° elevation. While, for the drum stimulus with no added room, a reflection level of -6 dB led to enough correct ABX identifications to exceed the 95% confidence level in the loudspeaker based experiment, none of the evaluated reflection levels produced a significant number of correct identifications in the headphone based experiment. The perceived spatial difference values for this case show a similar result. This indicates that, for the evaluated setups, a larger difference in elevation between two reflection positions is required to perceive a noticeable difference in the headphone based reproduction method than for the loudspeaker based method when using the same reflection level. For the dry speech signal on the other hand, neither of the two methods resulted in a significant perceptional difference between 0° and 10° early reflection elevation angle.

4.2.2 45° Elevated Reflection on Frontal Plane

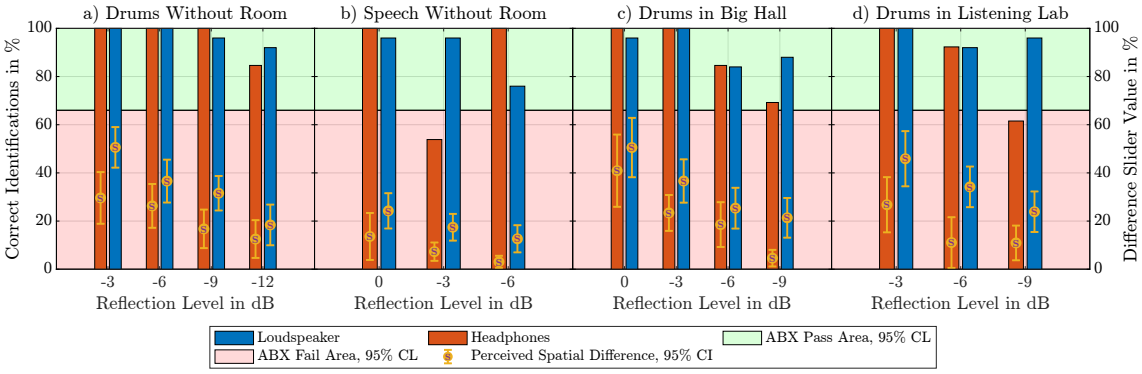


Figure 4.8: Loudspeaker and headphone based listening test results for comparison between a reflection with an elevation angle of 45° and 0° at an azimuth angle of 90° with varying reflection levels, source stimuli and room IRs.

For reflections on the frontal plane, i.e. at an azimuth angle of 90°, the headphone

based experiment yielded in similar ABX results as the loudspeaker based experiment when comparing a 45° elevated reflection to a non elevated reflection. As shown in Figure 4.8, for the drum signal with no added room as well as for the drums with added "big hall" SRIR the participant groups of both experiments were able to correctly identify enough stimuli to exceed the 95 % confidence threshold, even at -12 dB reflection level.

For the speech signal without added room on the other hand, the loudspeaker based experiment resulted in more correct identifications than the headphone based experiment. The headphone based ABX results show an inconsistency for these specific comparisons as the number of correct IDs exceeds the 95 % confidence threshold for 0 dB and -6dB reflection level but not for -3 dB. However, as the perceived spatial difference values for the headphone based -6 dB reflection comparisons are almost zero, it can be assumed that there was no significant perceivable difference in this case and that the 100 % correct headphone based ABX identifications for the -6 dB reflection level represent a statistical outlier caused by the moderate number of participants. An alternative explanation could be that there was some undetected bias in the headphone based experiment which allowed the participants to correctly identify these specific stimuli without actually perceiving a difference which is, to the authors best knowledge, rather unlikely.

Furthermore, the overall perceived spatial difference for the loudspeaker based experiment is noticeably higher than for the headphone based experiment, independent on the source signal or room. While this deviation of perceived spatial difference between the both experiment setups is also apparent for elevated reflections on the median plane, the results for a 45° elevated reflection on the frontal plane show, independent of the stimuli type, the most consistent differences between both methods.

4.2.3 45° Elevated Reflection Altering Between Frontal and Median Plane

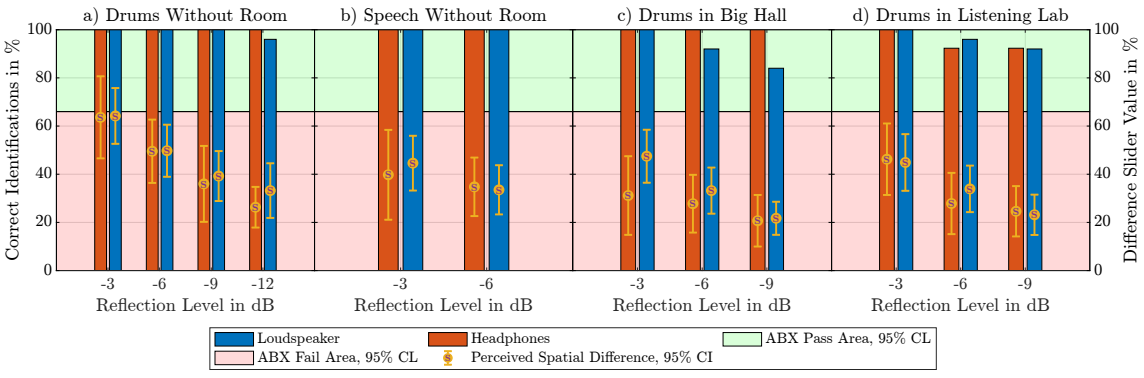


Figure 4.9: Loudspeaker and headphone based listening test results for comparison between a reflection with an azimuth angle of 0° and 90° at an elevation angle of 45° with varying reflection levels, source stimuli and room IRs.

As shown in Figure 4.9, altering the azimuth angle of a 45° elevated reflection

between 0° and 90° results in clearly noticeable differences, independent on the source signal, added room or reproduction method. For this case, there is no pronounced overall difference in the ABX results or perceived spatial difference values between the loudspeaker and the headphone based experiment which can most likely be explained by the fact that mainly interaural cues are responsible for the perception of lateral displacement. These interaural cues can be considered as relatively robust which means that, even in the headphone based experiment with generalized HRTFs, a change in an elevated early reflection's azimuth angle is clearly perceivable.

4.2.4 Natural SRIR Without added Reflection

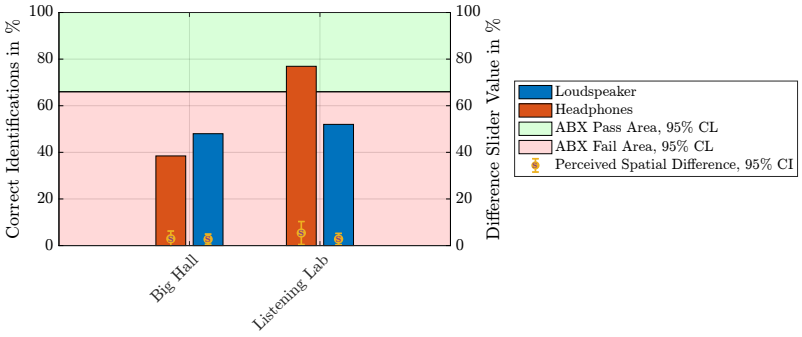


Figure 4.10: Loudspeaker and headphone based listening test results for comparison between SRIRs quantized with extra 90° elevated loudspeaker and SRIRs quantized to eight loudspeakers on the horizontal plane. Only the drums signal was evaluated and just the natural SRIRs of "big hall" and "listening lab" without an added ceiling reflection were used.

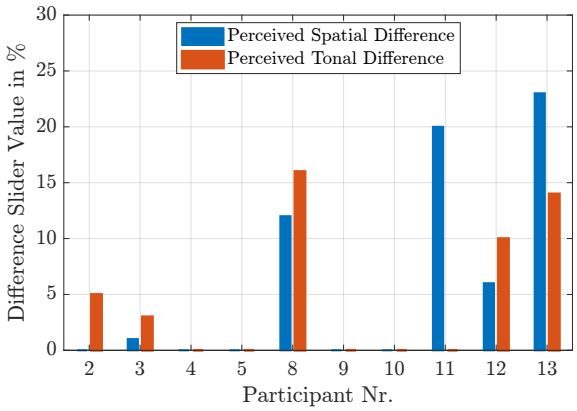


Figure 4.11: Perceived difference answers of participants who achieved a correct ABX identification in the headphone based experiment when comparing the Listening Lab RIR with and without added virtual loudspeaker at 90° elevation.

Figure 4.10 compares the results from both listening experiments for the two SRIRs with and without their "natural" elevated reflections played back via a separate elevated loudspeaker using the drum signal as sound source. While neither the number

of correct ABX identifications nor the perceived difference answers for the "big hall" SRIR show a perceivable stimuli difference for both reproduction methods, the headphone based experiments results for the "listening lab" SRIR show a sufficiently high number of correct ABX identifications to exceed the 95 % confidence threshold. Interestingly, the loudspeaker based correct ABX identifications don't exceed this threshold which would mean that, for this scenario, the headphone based method yields in stronger perceivable differences. In order to confirm that the high number of correct headphone based ABX identifications actually corresponds to a perceivable difference between the two stimuli, the individual perceived spatial and tonal difference responses of the participants who achieved a correct ABX identification in the headphone based experiment were evaluated as presented in Figure 4.11. Indeed, those perceived difference values show that three out of thirteen participants correctly identified the ABX stimuli and thereby stated that they perceived a spatial difference of more than 10 %. Assuming that those three participants did not guess the correct ABX identification and accidentally set the perceived difference sliders to nonzero values, these results indicate that there is a small perceivable difference between the "listening lab" SRIR's elevated reflections mapped to an elevated virtual loudspeaker compared to quantizing all reflections to the horizontal plane when using a headphone based reproduction method with this specific experiment setup.

4.2.5 Consistency and Differences Between Both Reproduction Methods

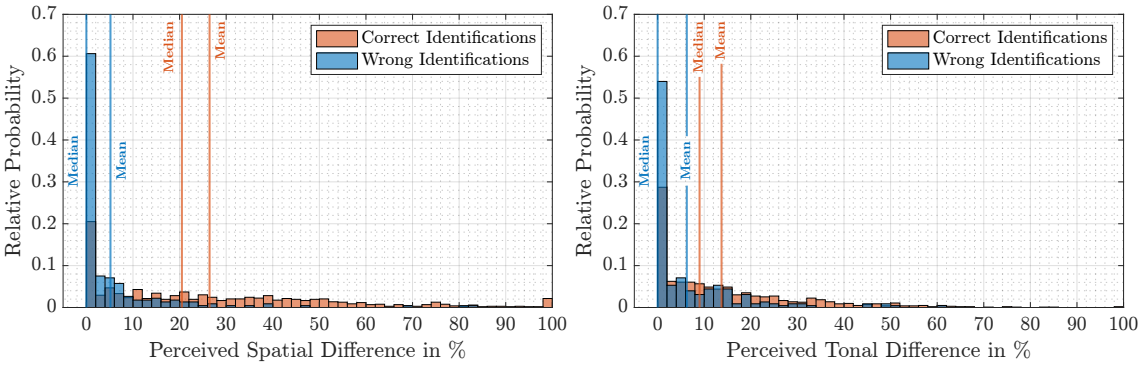


Figure 4.12: Relative probability histogram of perceived spatial and tonal differences separated by correct and wrong ABX identifications for loudspeaker based experiment.

One possible method to evaluate the consistency of the perceived spatial and tonal difference values is to differentiate between the perceived difference answers for comparisons where the participants could not correctly identify the difference in the ABX test and the results where the participants achieved a correct ABX identification. Figure 4.12 and Figure 4.13 show the relative probability distribution (see Section 3.7.2) of the perceived difference values for both experiments, separated between answers for correct and wrong ABX identifications.

In theory, a wrong ABX identification means that a participant did not hear a difference and hence one would expect the perceived spatial and tonal difference

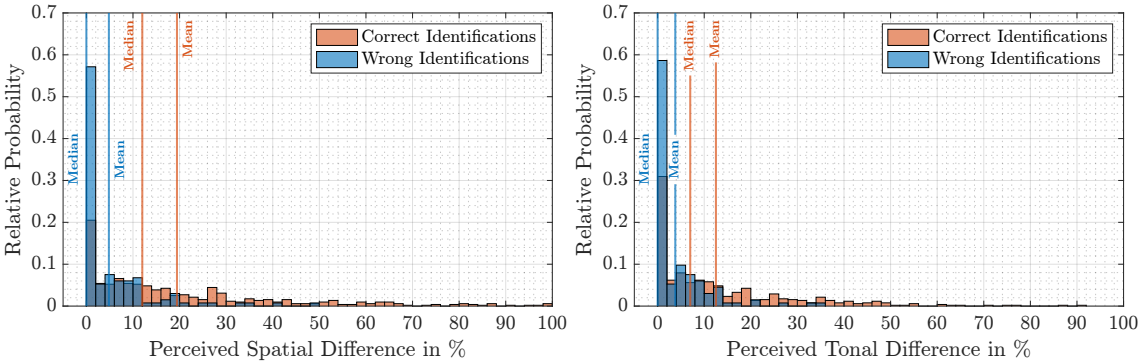


Figure 4.13: Relative probability histogram of perceived spatial and tonal differences separated by correct and wrong ABX identifications for headphone based experiment.

values to be zero for this case. As it can be seen in both figures, the median of the perceived spatial and tonal differences for both experiments is indeed zero for answers with false ABX identifications. This means that, most of the times, the participants set the spatial and tonal difference sliders to zero when they could not correctly identify the ABX stimuli. However, there are still several perceived difference values in the range below 15 % as well as some scattered outliers in the higher difference range. The latter increases the arithmetic mean of the perceived difference values for incorrect ABX identifications to around 5% for the spatial differences and 4 % respectively 6 % for the tonal differences.

There are a number of possible explanations for these deviations. The outliers in the higher difference region could be caused by errors in the participant's ABX answers, e.g. a participant heard a difference between the compared stimuli, set the difference sliders to a high value but then accidentally selected the wrong answer in the ABX test. The higher number of perceived difference values in the range from 1 % to 10 % could partly be caused by the fact that the participants had to move both difference sliders in the GUI once before they could continue to the next test. Thereby, it is possible that, even though a participant did not hear a difference, the slider values were not set exactly back to zero because the GUI forced the user to move the sliders once. Additionally, it is of course thinkable that the participants deceptively thought they heard a minor difference and therefore set the deference sliders to small values but then failed in correctly identifying the ABX stimuli.

When a participant correctly identified the X stimulus in the ABX test, this could mean that an actual difference between the both compared sounds was perceived. In that case, one would expect at least one of the two perceived difference values to be above zero. On the other hand, it is also possible that the participant did not perceive a difference but randomly made a correct identification. In this case, the perceived difference values should be zero. Looking at the distribution of perceived difference values for the comparisons where a participant made a correct ABX identification as shown in Figure 4.12 and Figure 4.13 confirms this assumption. Indeed, these probability distributions contain a significant amount of tests where participants made a correct identification but reported no spatial or tonal difference which

means that the participants guessed the correct X stimulus. As these plots differentiate between spatial and tonal differences, it is of course also possible that a participant only reported to hear a difference in one of those two attributes and set the other attribute to zero.

Comparing the distribution of perceived differences between the correct and wrong identifications, it is obvious that both the mean and the median of perceived differences is higher for the tests with a correct identifications than for the wrong identifications. This shows that the perceived difference values are actually correlated to the participants' ability to make a correct ABX identification. While it is not possible to make a precise statement about the overall experiment accuracy or potential biases, these results at least least indicate that the perceived difference values must have a certain significance.

Comparing the mean and median values of perceived spatial difference for correct identifications between both reproduction methods, it is apparent that both the mean and the median values are up to 10% higher for the loudspeaker array experiment than for the headphone based tests. For the tonal difference, those values are in the same range independent from the reproduction methods. This means that the participants perceived larger spatial differences for the stimuli comparisons where they managed a correct identification when played back via the loudspeaker array than for the headphone based experiment. This observation matches the results from Section 4.2 where, for most of the comparisons, the perceived spatial differences during the loudspeaker experiment were higher than for the headphone based test.

As 10 out of 13 participants from the headphone based listening experiment also performed the loudspeaker array test, an other approach to evaluate the consistency of the results and possible deviations between both reproduction methods is to compare the number of correct ABX identifications between the two reproduction methods for each of those 10 participants. Of course this analysis should be treated with caution as the performed ABX test did not contain repetitions of each comparison and hence one can not derive individual detection thresholds. However, comparing the total number of correct identifications per participant for a data set of 50 different comparisons should at least give a general indication about if there is a significant difference in the participants identification accuracy between loudspeaker array and binaural sound reproduction. Figure 4.14 shows the number of correct IDs using both reproduction methods for each participant as well as the arithmetic mean of correct identifications for all 10 participants.

As it can be seen, there is no pronounced trend regarding which reproduction method is more suitable to correctly identify the stimuli in the ABX tests. Some participants achieved more correct IDs in the loudspeaker array test, other participants were more efficient in the binaural experiment. For most of the participants, the difference in the total number of correct IDs between both methods is below five, which can be considered as not significant due to the randomness involved in the ABX method. Overall, the arithmetic mean of correct identifications for the loudspeaker setup is 1.7 correct IDs higher than for the binaural experiment. This difference is not large enough to state that the binaural method is less suitable to identify differences

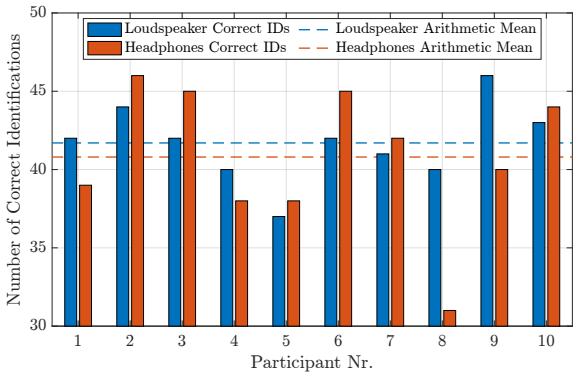


Figure 4.14: Comparison of total number of correct ABX identifications for participants who performed both experiments.

between different elevated reflections, it is also questionable if the arithmetic mean is a good measure for this evaluation. Instead, one could also compare how many participants improved in the binaural test and vice versa. Following this approach, one finds that six participants achieved more correct identifications in the headphone based experiment and four subjects achieved more correct identifications in the loudspeaker based method.

While these results show that there is no major discrepancy between the overall number of correct identifications of both experiments, it is still possible that some training effect occurred for the group of participants who already performed the loudspeaker based experiment before. This would mean that the overall detection accuracy in the headphone based experiment could be lower when using a group of untrained participants instead.

4.2.6 Difference in IACC

Figure 4.15 presents the maximum normalized IACC values obtained from analyzing the dummy head measurements of the loudspeaker setup as well as the binaural signals used for the headphone based listening tests as described in Section 3.7.3. Thereby, subplots a - c show the values for different reflection positions, the data in each plot is grouped regarding source signal and added SRIR. The values of each group are arranged so that the two stimuli which are compared to each other are subsequent. This becomes clearer when looking at Figure 4.15d, which shows the detailed maximum $\mathrm{IACC}_{\mathrm{norm}}$ values for the drum signal rendered with the "big hall" SRIR and an added reflection at an elevation angle of 45° with varying levels and an azimuth angle altering between 0° and 90°.

Looking at Figure 4.15a, the maximum normalized IACC for signals with a reflection on the median plane stays relatively constant independent from the reflection's elevation angle and level. For the dry signals with no added room, the IACC is close to one for all reflection elevation angles and levels. This is expected since the used KEMAR HRTFs are almost symmetric for sources on the frontal plane hence very similar signals arrive at both ears, in this case independent from the elevation angle (see Section 2.1.3.1). However, even those small IACC changes between the

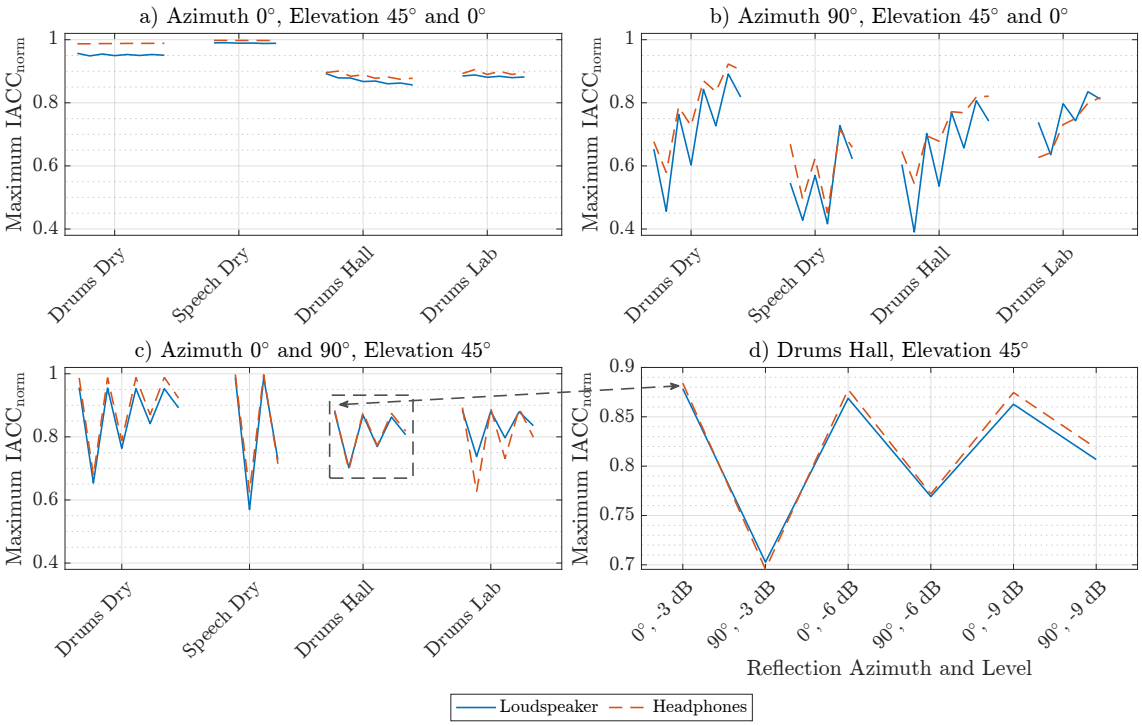


Figure 4.15: Maximum normalized IACC values for both experiment setups, divided in stimuli type and reflection positions.

individual stimuli could, under the right circumstances, result in audible differences as shown in [17].

In general, the IACC values obtained from the loudspeaker measurements are often slightly lower than the ones of the binaural signals. This could be caused by unwanted collateral reflections of the loudspeaker array setup or due to the dummy head being not perfectly placed in the center of the array. Also, the speakers consisting of two drivers and an enclosure potentially result in a larger apparent source width than a signal convolved with a single HRTF which, again, could cause a lower IACC for the speaker setup. Furthermore, the difference between the IACC values of the loudspeaker based measurement and the binaural rendering is generally bigger for the drum signal than for the speech signal, this could be explained by the fact that the drum signal contains more high frequency content than the speech signal which is more likely to be affected by reflections and positioning inaccuracies in the loudspeaker setup.

Comparing the overall maximum $\mathrm{IACC}_{\mathrm{norm}}$ values between the different room scenarios confirms that, as described in Section 2.1.3.1, the interaural cross-correlation decreases with an increasing room diffuseness. I.e. stimuli rendered with the "big hall" SRIR, which has the longest reverberation time of all compared rooms, have in general the lowest maximum $\mathrm{IACC}_{\mathrm{norm}}$ values. According to [18], this overall decrease in IACC generally leads to a higher detection threshold for IACC differences. The stimuli with added "big hall" SRIR and a strong reflection on the frontal plane show the lowest maximum IACC which could indicate that they are perceived as most enveloping [16]. This indeed corresponds to subjective impressions from

listening to the different stimuli.

Summarized, it can be stated that the change of maximum normalized IACC between two stimuli varies with all other previously mentioned relevant parameters like the reflection level, position, elevation change, the source signal and the room. As it was found that the amount of perceived differences for a changing reflection depends on those attributes, it is apparent why there is a relation between the change in maximum IACC and the audible differences.

5

Discussion

5.1 Spatial and Tonal Differences

One fundamental task for understanding the effect of elevated reflections on the human spatial auditory perception is to determine if changing a reflection's elevation angle results in more spatial or more tonal audible differences. In theory, it would be expected that the effect of binaural decoloration significantly reduces the tonal differences introduced by the listeners HRTF when changing the position of a single reflection in a perfectly anechoic environment. However, as discussed in Section 3.5.1.1, the setup used for the loudspeaker based experiment can not be considered to be perfectly anechoic due to collateral reflections from the loudspeaker array which lead to varying comb filtering effects depending on the playback loudspeaker position. Whether or not the auditory system is able to compensate for these kind of differences is currently unclear. As generalized HRTFs were used for the headphone based experiment, one could assume that the binaural decoloration does not work as good as if individualized HRTFs would have been used. However, to the authors best knowledge this has not been investigated yet.

Nevertheless, the evaluation of the perceived spatial and tonal difference values for both experiment shows that, independent on the reproduction method, changing the elevation angle of a reflection on the frontal plane overall results in slightly higher perceived tonal differences than spatial differences. Assuming that the arithmetic mean of the perceived difference values is a meaningful measure, the perceived tonal differences are, relative to the spatial differences, minimally higher for the headphone based experiment than for the loudspeaker based experiment. Even though this deviation in the results of both reproduction methods is only subtle it would, in theory, make sense that the headphone based method results in more tonal differences than the loudspeaker based method since the used generalized HRTFs might not match to the listeners anatomy and therefore a change in elevation angle might rather be perceived as tonal change than as a spectral elevation cue.

On the other hand, both experiments showed that changing a reflection's elevation angle on the frontal plane results in significantly higher perceived spatial than tonal differences. This could be explained by the fact that, for lateral reflections, altering the elevation angle results in a change of interaural differences. As shown in Figure 2.3, a sound wave impinging on the frontal plane with 0° elevation generally leads to the highest possible ILD and ITD as well as the lowest IACC under free field conditions. An elevation angle of 45° on the other hand leads to relatively low

ILD and ITD values as well as a high IACC. The obtained results suggest that this large change in interaural cues results in strong perceived spatial differences which then predominate the tonal differences.

Lastly, it can be observed that, throughout all comparisons in both experiments, the perceived tonal differences are slightly lower compared to the spatial differences when adding the "big hall" SRIR to the stimuli. This could indicate that adding a significant amount of reverberation masks the tonal differences introduced by changing a reflection's elevation angle in addition to overall increasing the detection threshold. This corresponds to the findings of [35], where it was shown that the level of a room's reverberant field has an effect on the timbral contribution of an individual reflection.

5.2 Detection Threshold

The presented results show that the strength of a reflection strongly influences whether or not changes in the reflection's elevation angle are audible. Thereby, it was found that this detection threshold depends on the reflection position whereby lateral displaced reflections generally result in a noticeably lower detection threshold than reflections on the median plane. The amount of change in elevation also affects the detection threshold as it was shown than a 10° elevation change results in less audible differences than a 45° change. Additionally, both the source signal as well as the amount of room reverberation and the used reproduction method influence this threshold. Due to this complexity, it is not possible to determine universal detection thresholds for a change in a reflection's elevation angle from the obtained data. However, repeating the experiment with a denser grid of evaluated reflection positions, a wider range of reflection levels and a set of frequency limited test signals would eventually allow to identify general reflection position- and frequency-dependent detection thresholds. While such an experiment would require a large number of stimuli comparisons per participant, its results could potentially be used to develop a model to predict if quantizing all elevated reflections of a specific SRIR to the horizontal plane would yield in perceivable differences.

5.3 Influence of Stimuli Type

From both the ABX test evaluation and the perceived difference values it is obvious that, independent on the reflection position or reproduction method, a drum signal always yields in larger perceived differences between two reflection positions than a speech signal at the same reflection level. This is expected since the drum signal contains more high frequency components and is more impulsive than the speech signal which makes higher frequency spectral cues as well as timbral differences more audible. For both reproduction methods, the speech signal in fact resulted in so small differences that, even at a reflection level of 0 dB and without added room, a 45° elevation change on the median plane did not result in a statistically relevant perceivable difference. For a lateral displaced reflection altering in elevation on the frontal plane on the other hand, a clear difference between two reflection

positions was also audible using the speech signal. This means that the relevance of a reflection's elevation angle for the spatial perception of a room is dependent on the source signal whereby it was shown that the perception of a drum signal is more critical regarding changes in a reflection's elevation angle than the perception of a speech signal.

5.4 Influence of Room

Independent on the stimulus type, reflection position or reproduction method, the experiment results suggest that adding a spatial room impulse response to the compared stimuli decreases the perceived spatial and tonal differences between two reflection positions and therefore increases the detection threshold. Thereby, it was found that a room with a longer reverberation time like the evaluated "big hall" has a stronger masking effect than a small room with less diffuse reflections like the "listening lab". This corresponds to findings in the context of other psychoacoustic effects like for example the echo threshold, which also decreases when additional reflections are present between the direct sound and a specific reflection [1, p. 274]. As shown in Figure 3.4b, the RIR of the "big hall" contains more energy than the "listening lab" RIR in the range before the added ceiling reflection occurs around 5 ms which could cause a partial masking of the elevated reflection. In general, it can be assumed that the spatial and tonal differences caused by changing a single reflection's elevation angle are less severe when many other early and late reflections are present since the influence of this single reflection on the overall perceived sound is less than in a free-field scenario.

While related research proved that the localization accuracy for brief impulsive tones is not significantly affected when changing a room's reverberation time, it has also been shown that the localization accuracy for steady state noise is remarkably decreased when increasing the amount of reverberation [36]. This suggests that adding a large room such as the "big hall" to the stimuli affects the speech signals more than the rather impulsive drum signal which could then lead to a lower reflection localization accuracy for the speech which would finally result in less perceivable differences. However, since the speech signal with added reverberation was not evaluated in the performed listening tests, this theory can not be confirmed from the results.

5.5 Relation of Perceived Differences and IACC

While the results show that perceivable spatial and tonal differences even occur when the maximum normalized IACC between two compared stimuli is exactly the same, it is evident that a large IACC difference between two stimuli corresponds to large audible differences. Comparing the results between reflections on the median and frontal plane suggests that, even though spectral cues also allow the discrimination of two sounds with different reflection elevation angles, a change of interaural cues results in a larger perceivable difference. This is in accordance with literature such as [18], [34] and [17] which confirmed that the human auditory system can be highly

sensitive to changes in the IACC. This outcome is especially relevant in the context of headphone based sound reproduction, where nonindividualized HRTFs can result in less robust spectral cues and hence the influence of interaural differences on the spatial perception might be considered as more prominent.

Additionally, it was confirmed that adding a diffuse room does decrease the overall IACC but does not significantly reduce the IACC difference between two stimuli with varying reflection positions, which, according to [18], means that a change in IACC is less perceivable than without any added room.

Even though IACC differences are not the only indicators for perceivable differences in the context of this thesis, one can assume that analyzing a spatial sound scene with altering reflection positions regarding changes in the maximum normalized IACC and comparing those changes to the just noticeable IACC differences found in the previously mentioned literature would allow to at least identify scenarios where a clear audible difference due to a change in a reflection's position is likely. However, analyzing just the IACC without taking monaural cues into account can not rule out that differences are perceived in cases where the IACC stays constant as it was shown for reflections on the median plane.

6

Conclusion

The aim of this thesis was to evaluate the ability of the human auditory system to distinguish between early reflections with different elevation angles. The obtained listening test results show that, under certain conditions, changing an early reflection's elevation angle results in clearly perceivable spatial and tonal differences both for a loudspeaker and headphone based reproduction method. Thereby, it was determined that the amount of perceived differences depends on:

The reflection level
: A stronger reflection signal results in larger perceived differences.

The lateral reflection angle
: Changing the elevation of a reflection on the frontal plane results in significantly stronger perceived differences than a reflection on the median plane.

The change of elevation
: Changing a reflection's elevation angle by 10° produces less audible differences than an elevation change of 45°.

The source signal
: Using a drum recording as source signal clearly yields in more perceivable differences than using a speech signal.

The room
: The perceived differences are noticeably higher in an anechoic environment than in a diffuse sound field. Thereby, a longer reverberation time resulted in less audible differences.

The reproduction method
: The amount of perceived differences was found to depend on the specific reproduction method. Thereby, it can be assumed that parameters such as the number of speakers for a loudspeaker based reproduction or the used HRTFs for a headphone based method influence the overall detection accuracy.

Furthermore it was found that, if changing a reflection's elevation angle causes a large IACC difference, this is a strong indicator for perceivable differences. Thereby, the difference in maximum IACC is directly affected by the reflection level, position, elevation change, the source signal and the room. However, it is not possible to state that the amount of perceived differences between two reflection positions directly depends on the resulting IACC difference since it was shown that a change in elevation can also be perceived when the IACC stays constant.

These findings suggest that, for the sake of an accurate reproduction of a spatial

sound scene, it is, under certain conditions, necessary to measure and reproduce the original three-dimensional DOAs of all strong room reflections. Furthermore, the data indicates that an accurate spatial reproduction is especially important for dry rooms with strong and eventually laterally displaced elevated early reflection. For large rooms with no pronounced ceiling reflections such as the investigated "big hall", quantizing all elevated reflections to the horizontal plane did not result in audible differences.

However, the conducted experiments only investigated whether or not differences between varying reflection elevation angles can be perceived and it was not evaluated if the participants found one of the two compared stimuli to be more plausible. This means it is possible that, even though the participants perceived differences when changing a strong early reflection's elevation angle, both stimuli were in the end perceived as equally plausible. Nevertheless, in a non-formal interview after the experiments many participants stated that, especially for the loudspeaker based setup, the stimuli which included an elevated reflection often resulted in a more enveloping perception compared to the cases with no elevated reflection. This aspect could be further investigated in a follow up experiment. The same applies to the comparison of the two reproduction methods. Thereby, the results only show that the detection accuracy for both methods is in a similar range but it is certainly possible that the overall spatial impression using a headphone based reproduction method is different than using a loudspeaker setup.

References

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

[2] J. Ahrens, H. Helmholz, D. L. Alon, and S. V. A. Gari, "The Far-Field Equatorial Array for Binaural Rendering," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2021, pp. 421–425. DOI: 10.1109/ICASSP39728.2021.9414368.

[3] B. Rakerd and W. M. Hartmann, "Localization of sound in rooms, II: The effects of a single reflecting surface," *The Journal of the Acoustical Society of America*, vol. 78, no. 2, pp. 524–533, Aug. 1985. DOI: 10.1121/1.392474.

[4] H. Furuya, K. Fujimoto, Y. Takeshima, and H. Nakamura, "Effect of early reflections from upside on auditory envelopment.," *Journal of the Acoustical Society of Japan (E)*, vol. 16, no. 2, pp. 97–104, 1995. DOI: 10.1250/ast.16.97. [Online]. Available: http://joi.jlc.jst.go.jp/JST.Journalarchive/ast1980/16.97?from=CrossRef.

[5] S. Bech, "Perception of Reproduced Sound: Audibility of Individual Reflections in a Complete Sound Field, II," in *Audio Engineering Society Convention 99*, Oct. 1995. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=7673.

[6] S. Bech, "Spatial aspects of reproduced sound in small rooms," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 434–445, Jan. 1998. DOI: 10.1121/1.421098.

[7] T. Robotham, M. Stephenson, and H. Lee, "The effect of a vertical reflection on the relationship between preference and perceived change in timbral and spatial attributes," *140th Audio Engineering Society International Convention 2016, AES 2016*, pp. 1–9, 2016.

[8] R. Wallis and H. Lee, "The effect of interchannel time difference on localization in vertical stereophony," *AES: Journal of the Audio Engineering Society*, vol. 63, no. 10, pp. 767–776, 2015. DOI: 10.17743/jaes.2015.0069.

[9] R. Wallis and H. Lee, "Vertical stereophonic localization in the presence of interchannel crosstalk: The analysis of frequency-dependent localization thresholds," *AES: Journal of the Audio Engineering Society*, vol. 64, no. 10, pp. 762–770, 2016. DOI: 10.17743/jaes.2016.0039.

[10] J. Blauert, *The technology of binaural listening*. 2013, pp. 1–511. DOI: 10.1007/978-3-642-37762-4.

[11] K. Brandenburg, F. Klein, A. Neidhardt, U. Sloma, and S. Werner, *The Technology of Binaural Understanding*, J. Blauert and J. Braasch, Eds., ser. Mod-

ern Acoustics and Signal Processing. Springer International Publishing, 2020, pp. 623–663. DOI: 10.1007/978-3-030-00386-9.

[12] H. Braren and J. Fels, "A High-Resolution Head-Related Transfer Function Data Set and 3D-Scan of KEMAR," pp. 1–6, 2020. DOI: 10.18154/RWTH-2020-11307.

[13] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization.," *Journal of Experimental Psychology*, vol. 27, no. 4, pp. 339–368, 1940. DOI: 10.1037/h0054629.

[14] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993. DOI: 10.1121/1.407089.

[15] L. L. Beranek, "Concert Hall Acoustics—2008," *J. Audio Eng. Soc*, vol. 56, no. 7/8, pp. 532–544, 2008. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14398.

[16] J. Y. Jeon and J. You, "Effect of sound strength and IACC on perception of listener envelopment in concert halls," *20th International Congress on Acoustics 2010, ICA 2010*, vol. 3, pp. 2363–2366, 2010.

[17] K. J. Gabriel and H. S. Colburn, "Interaural correlation discrimination: I. Bandwidth and level dependence," *The Journal of the Acoustical Society of America*, vol. 69, no. 5, pp. 1394–1401, May 1981. DOI: 10.1121/1.385821.

[18] B. Rakerd and W. M. Hartmann, "Localization of sound in rooms. V. Binaural coherence and human sensitivity to interaural time differences in noise," *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 3052–3063, Nov. 2010. DOI: 10.1121/1.3493447.

[19] A. Andreopoulou and B. F. G. Katz, "Identification of perceptually relevant methods of inter-aural time difference estimation," *The Journal of the Acoustical Society of America*, vol. 142, no. 2, pp. 588–598, Aug. 2017. DOI: 10.1121/1.4996457.

[20] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *AES: Journal of the Audio Engineering Society*, vol. 61, no. 1-2, pp. 17–28, 2013. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16664.

[21] S. Tervo and J. Pätynen, *SDM Toolbox for Matlab*. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/56663-sdm-toolbox.

[22] 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, Jan. 2021. DOI: 10.17743/jaes.2020.0063.

[23] S. Tervo, J. Pätynen, N. Kaplanis, M. Lydolf, S. Bech, and T. Lokki, "Spatial Analysis and Synthesis of Car Audio System and Car Cabin Acoustics with a Compact Microphone Array," *AES: Journal of the Audio Engineering Society*, vol. 63, no. 11, pp. 914–925, 2015. DOI: 10.17743/jaes.2015.0080.

[24] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," vol. 144, no. 5, pp. 357–360, 1997. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=7853.

[25] O. Puomio, J. Pätynen, and T. Lokki, "Optimization of virtual loudspeakers for spatial room acoustics reproduction with headphones," *Applied Sciences (Switzerland)*, vol. 7, no. 12, 2017. DOI: 10.3390/app7121282.

[26] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96, AAAI Press, 1996, pp. 226–231.

[27] L. Mccormack, A. Politis, O. Scheuregger, and V. Pulkki, "Higher-order processing of spatial impulse responses," *Proc. 23rd International Congress on Acoustics*, no. 1, pp. 4909–4916, 2019.

[28] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," *Journal of the Audio Engineering Society*, Feb. 2000. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=10211.

[29] J. Ahrens, "Perceptual Evaluation of Binaural Auralization of Data Obtained from the Spatial Decomposition Method," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, Oct. 2019, pp. 65–69. DOI: 10.1109/WASPAA.2019.8937247.

[30] M. Berzborn, R. Bomhardt, J. Klein, J.-G. Richter, and M. Vorländer, "The ITA-Toolbox: An Open Source MATLAB Toolbox for Acoustic Measurements and Signal Processing," 43th Annual German Congress on Acoustics, Kiel (Germany), 6 Mar 2017 - 9 Mar 2017, Mar. 2017. [Online]. Available: http://publications.rwth-aachen.de/record/687308.

[31] European Broadcasting Union, "EBU Tech 3253 - SQAM: Sound Quality Assessment Material Recordings for Subjective Tests," 2008. [Online]. Available: https://tech.ebu.ch/docs/tech/tech3253.pdf.

[32] J. Boley and M. Lester, "Statistical Analysis of ABX Results Using Signal Detection Theory," *Audio Engineering Society Convention 127*, 2009. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=15022.

[33] A. M. Mood, F. A. Graybill, and D. C. Boes, *Introduction to the theory of statistics*, 3rd ed. McGraw-Hill New York, 1974.

[34] M. J. Goupell and W. M. Hartmann, "Interaural fluctuations and the detection of interaural incoherence: Bandwidth effects," *The Journal of the Acoustical Society of America*, vol. 119, no. 6, pp. 3971–3986, Jun. 2006. DOI: 10.1121/1.2200147.

[35] S. Bech, "Timbral aspects of reproduced sound in small rooms. I," *The Journal of the Acoustical Society of America*, vol. 97, no. 3, pp. 1717–1726, Mar. 1995. DOI: 10.1121/1.413047.

[36] W. M. Hartmann, "Localization of sound in rooms," *The Journal of the Acoustical Society of America*, vol. 74, no. 5, pp. 1380–1391, Nov. 1983. DOI: 10.1121/1.390163.

A

Analyses of SDM Decomposed SRIRs

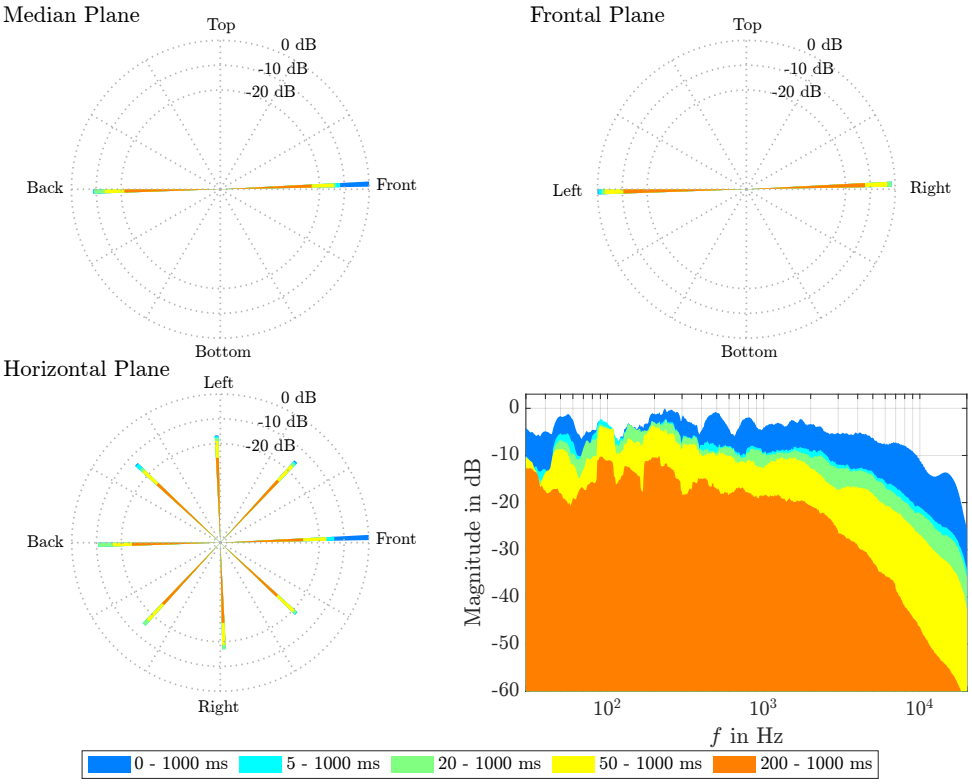


Figure A.1: Spatio-temporal visualization of "Big Hall" SRIR using eight virtual loudspeakers on the horizontal plane calculated using the SDM Toolbox [21].

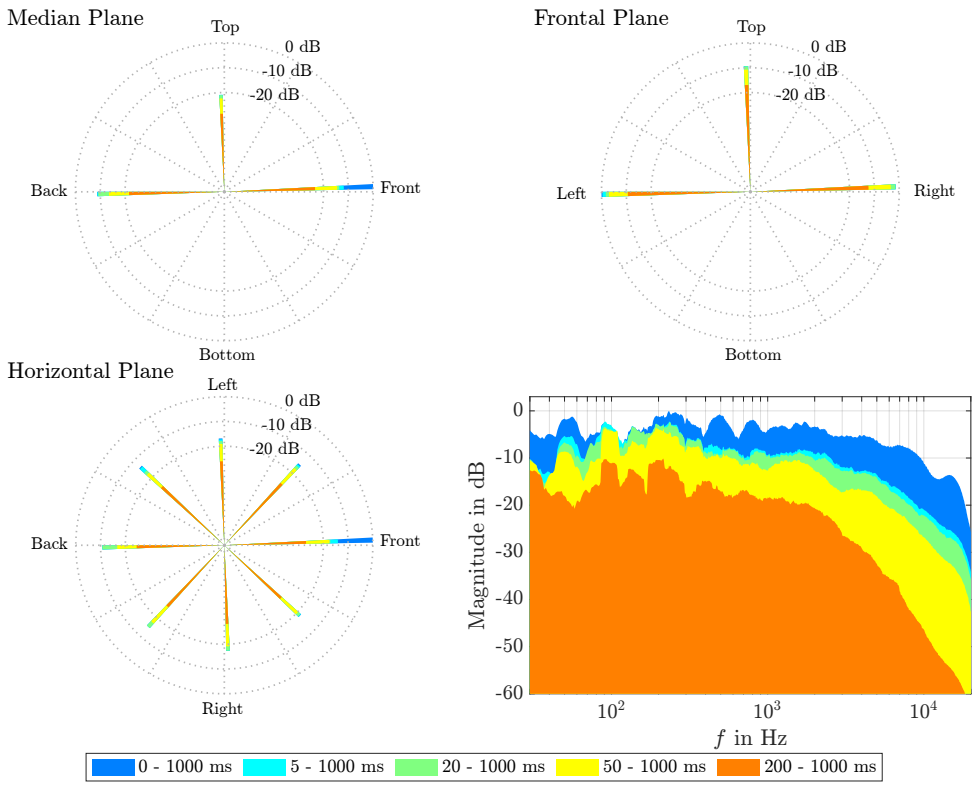


Figure A.2: Spatio-temporal visualization of "Big Hall" SRIR using eight virtual loudspeakers on the horizontal plane and one extra virtual loudspeaker with 90° elevation calculated using the SDM Toolbox [21]

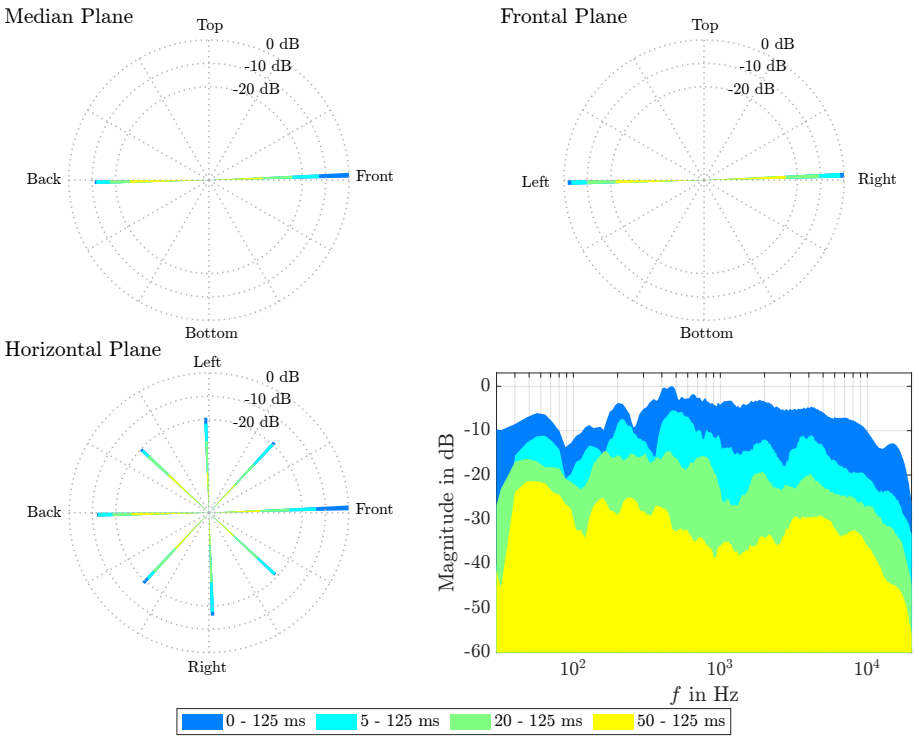


Figure A.3: Spatio-temporal visualization of "Listening Lab" SRIR using eight virtual loudspeakers on the horizontal plane calculated using the SDM Toolbox [21].

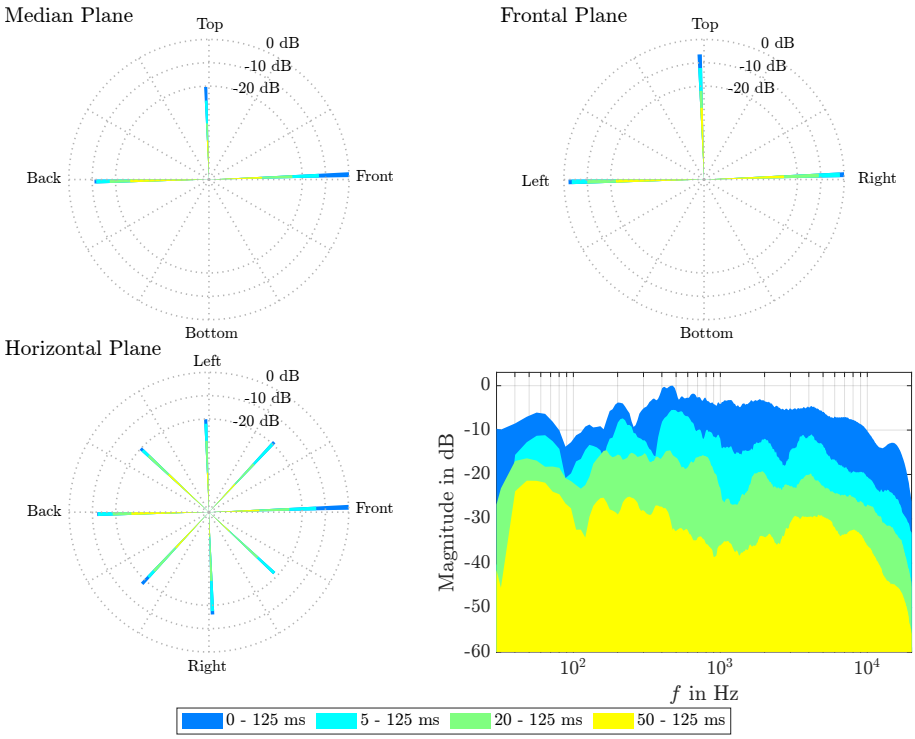


Figure A.4: Spatio-temporal visualization of "Listening Lab" SRIR using eight virtual loudspeakers on the horizontal plane and one extra virtual loudspeaker with 90° elevation calculated using the SDM Toolbox [21].

B

Headphone Based Experiment Results

B.1 Elevated Reflections on the Median Plane

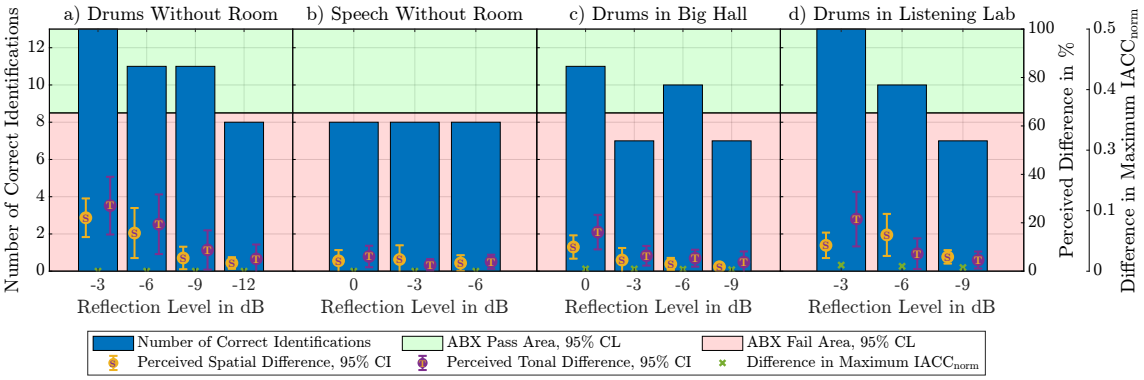


Figure B.1: Headphone based listening test results for comparison between a reflection with an elevation angle of 45° and 0° at an azimuth angle of 0° with varying reflection levels, source stimuli and room IRs.

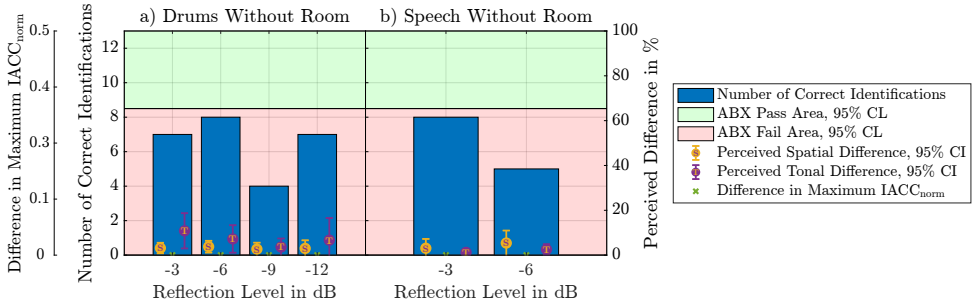


Figure B.2: Headphone based listening test results for comparison between a reflection with an elevation angle of 10° and 0° at an azimuth angle of 0° with varying reflection levels and source stimuli.

B.2 Elevated Reflections on the Frontal Plane

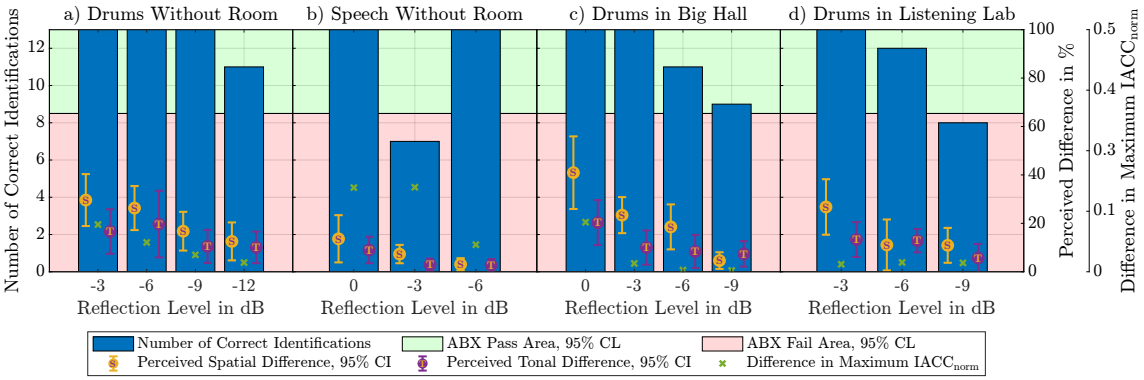


Figure B.3: Headphone based listening test results for comparison between a reflection with an elevation angle of 45° and 0° at an azimuth angle of 90° with varying reflection levels, source stimuli and room IRs.

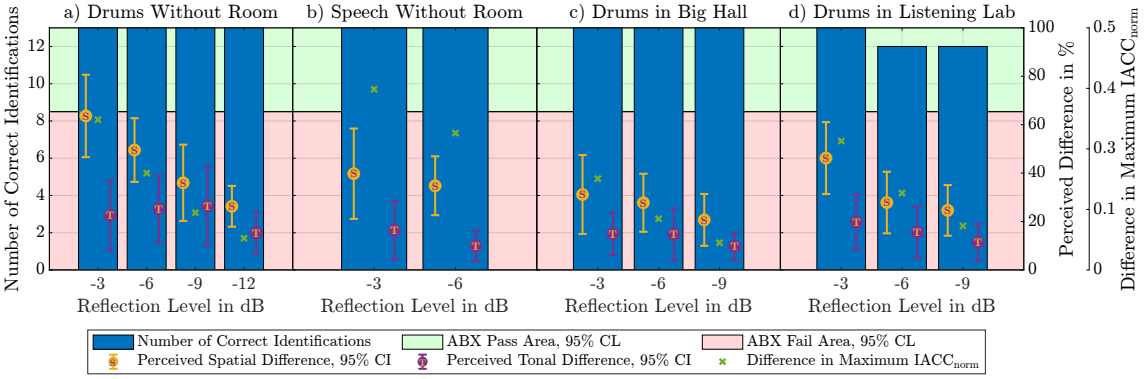


Figure B.4: Headphone based listening test results for comparison between a reflection with an azimuth angle of 0° and 90° at an elevation angle of 45° with varying reflection levels, source stimuli and room IRs.

B.3 Natural SRIRs with and without added Elevated Loudspeaker

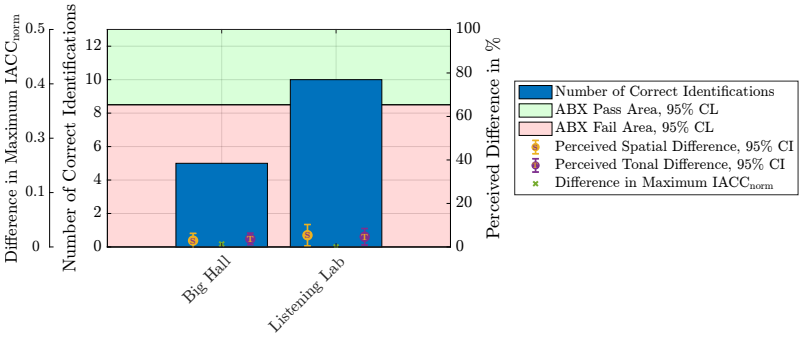


Figure B.5: Headphone based listening test results for comparison between SDM SRIR rendered with added loudspeaker at 90° elevation and SDM SRIR rendered only for loudspeakers on the horizontal plane. Only the drums signal was evaluated and just the natural IRs of “Big Hall” and “Listening Lab” were used.