


Master Thesis

Can you hear the shape of a concert hall? An audiovisual test in simulated 3D environments

for the achievement of the academic degree
Master of Science (M. Sc.)

Author: Jakob Greif 

Supervisor: Prof. Dr. Stefan Weinzierl
David Ackermann

Date: November 23, 2020

Submitted to: Technische Universität Berlin
Fakultät I - Geistes- und Bildungswissenschaften
Institut für Sprache und Kommunikation
Fachgebiet Audiokommunikation

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt gegenüber der Fakultät I der Technischen Universität Berlin, dass die vorliegende, dieser Erklärung angefügte Arbeit selbstständig und nur unter Zuhilfenahme der im Literaturverzeichnis genannten Quellen und Hilfsmittel angefertigt wurde. Alle Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind kenntlich gemacht.

Ich reiche die Arbeit erstmals als Prüfungsleistung ein. Ich versichere, dass diese Arbeit oder wesentliche Teile dieser Arbeit nicht bereits dem Leistungserwerb in einer anderen Lehrveranstaltung zugrunde lagen. Mit meiner Unterschrift bestätige ich, dass ich über fachübliche Zitierregeln unterrichtet worden bin und diese verstanden habe.

Berlin, den

..... Unterschrift

Abstract

The classical architectural types of concert halls have traditionally been ascribed a characteristic acoustic spatial impression, which is conditioned by the architectural shape. But, how easily can a *Vineyard*, a *Fan*, a *Horseshoe* or a *Shoe Box* actually be distinguished acoustically if the volume of the room, the reverberation time and the degree of wall diffusion are all identical? In order to investigate this question, a virtual audiovisual test environment was constructed, which enables test persons to visually grasp the shape of concert halls and the particularities of the architecture in a realistic way and to link this information with the binaural acoustic spatial impression. Participants were able to experience the visual impression of virtual concert halls through a stereoscopic virtual reality display and the acoustic impression through high-resolution non-individualized dynamic binaural room synthesis. They were asked to identify an acoustic impression and assign it to the architectural shape. For this test, we use digital models of the four concert halls of the above-mentioned types. Acoustic simulations were generated with a hybrid ray tracing method for each model with different variations of the reverberation time, wall scattering and volume. The visual test environment was designed in a game-like fashion using typical mechanics. This way, the entire listening test could take place in the virtual environment, allowing the setup to achieve a good level of immersion. 36 male and female subjects (aged 23 to 53 years) participated, half of whom had to undergo training before the test. The results showed that it was surprisingly difficult for untrained participants to recognize the shape of the concert halls with an accuracy above the guessing probability. However, a small number of the test participants were able to identify certain shape types with significantly higher probability. This shows that some participants were able to identify acoustic cues responding to the acoustic signature of the architectural shape and make an educated guess about the correct match of the visual and auditory impression. Training had a significant positive effect on the probability of making a correct choice. A smaller apparent source width was beneficial for trained and untrained participants, while a higher degree of listener envelopment increased the probability to make a correct choice for untrained participants. A significant effect of the reverberation time was revealed, suggesting that a shorter reverberation time helped participants to identify acoustic cues and assign the corresponding shape type correctly. This suggests that the acoustic spatial impression, produced by the three-dimensional pattern of early reflections, is more pronounced in a scenario with less diffuse reverberation. The results further validate the functionality of the complex test setup. The results as well as the methodology are valuable for further research on the audiovisual perception of space and the quality assessment of virtual acoustic environments.

Zusammenfassung

Den klassischen architektonischen Typen von musikalischen Aufführungsräumen wird traditionell ein charakteristischer akustischer Raumeindruck zugeschrieben, der durch die architektonische Form bedingt ist. Aber wie einfach lässt sich ein *Weinberg*, ein *Fächer*, ein *Hufeisen* oder eine *Schuhbox* tatsächlich akustisch unterscheiden, wenn das Volumen des Raumes, die Nachhallzeit und der Streugrad der Wände identisch sind? Um diese Frage zu untersuchen, wurde eine virtuelle audiovisuelle Testumgebung konstruiert, die es Testpersonen ermöglicht die Form von Konzertsälen und die Besonderheiten der Architektur visuell realitätsnah zu erfassen und diesen Eindruck mit dem binauralen akustischen Eindruck der Säle zu verknüpfen. Die Teilnehmer konnten den visuellen Eindruck von virtuellen Konzertsälen durch ein stereoskopisches Virtual Reality Display und den akustischen Eindruck über Arualisationen mit hochauflösender dynamischer Binauralsynthese erleben. Sie wurden gebeten, den akustischen Raumeindruck zu identifizieren und ihn der architektonischen Form der Konzertsäle korrekt zuzuordnen. Für diesen Test wurden vier Konzertsäle der oben genannten Typen in jeweils zwei Größen digital modelliert. Akustische Simulationen wurden mit einer hybriden Ray Tracing Methode für jedes Modell erstellt. Dabei wurde die Nachhallzeit, der Streugrades und das Volumen jeweils in zwei Stufen variiert. Die visuelle Testumgebung wurde nach dem Vorbild von Videospielen konstruiert und bedient sich ähnlicher Kontrollmechaniken. Auf diese Weise konnte der gesamte Hörtest in der virtuellen Umgebung stattfinden, sodass der Aufbau ein akzeptables Maß an Immersion zuließ. Es nahmen 36 männliche und weibliche Probanden im Alter von 23 bis 53 Jahren an dem Versuch teil. Die Hälfte musste zuvor ein Training absolvieren. Die Ergebnisse zeigen, dass es untrainierten Teilnehmern überraschend schwer fiel, die Form der Konzertsäle besser als mit Ratewahrscheinlichkeit zu erkennen. Ein kleiner Teil der Teilnehmer konnte jedoch manche der Räume heufig erkennen. Das Training hat einen signifikanten positiven Einfluss auf die Trefferwahrscheinlichkeit. Ein signifikanter Effekt konnte auch für die Nachhallzeit nachgewiesen werden, der aussagt, dass eine kürzere Nachhallzeit die Wahrscheinlichkeit erhöht, die Raumform richtig zu raten. Die Ergebnisse zeigen auch, dass eine große wahrgenommene Quellbreite die Trefferwahrscheinlichkeit verringert wohingegen eine hohes Maß and Umhüllung die Trefferquoten von untrainierten Probanden verbessert. Die Ergebnisse bestätigen, dass der komplexe Testaufbau die notwendigen akustischen und visuellen Eigenschaften der Konzärtsäle hinreichend abbilden konnte um die Hypothesen zu untersuchen. Die Ergebnisse und die Methodik sind wertvoll für zukünftige Forschung über die audiovisuelle Raumwahrnehmung und Qualität von virtuellen akustischen Umgebungen.

Acknowledgments

I want to thank Stefan Weinzierl for this interesting topic and his assistance to conceptualize the test design. I also want to thank Eckhard Kahle for his assistance in the construction and acoustic design of the concert hall models. Thanks to David Ackermann and Omid Kokabi for their assistance and guidance with the simulation and auralization tools. I also want to thank Jochen Steffens and Stefen Lepa for their assistance with the statistical analysis.

A special thanks goes to my Family, Friends for having my back.

Table of Contents

1	Introduction	1
1.1	Room acoustics and spatial impression	4
1.2	Echolocation	9
1.3	Related work and audiovisual interaction	11
1.4	Room acoustic simulation with Ray Tracing	12
1.5	Research goals and hypothesis	15
2	Method	16
2.1	Audiovisual listening test	17
2.1.1	Stimulus signal	18
2.1.2	Procedure and training	19
2.1.3	Graphical user interface	20
2.1.4	Test setup	21
2.1.5	Synchronization and control mechanic	23
2.2	Digital concert hall models	24
2.2.1	Vineyard	26
2.2.2	Shoe Box	28
2.2.3	Fan	30
2.2.4	Horseshoe	32
2.2.5	Model design in the virtual environment	34
2.3	Room acoustic simulation	35
2.3.1	Sound source and receiver	35
2.3.2	Source and listener position	36
2.3.3	Acoustic properties of the audience area	38
2.3.4	Acoustic properties of the organ	38
2.3.5	Test condition: reverberation time	39
2.3.6	Test condition: scattering coefficients	40
2.4	Dynamic auralization	41
2.5	Statistical analysis	43
3	Results	46
3.1	Mean correct response count of untrained participants	46
3.2	Mean correct response count of trained participants	47
3.3	Response count of individual participants	48
3.4	Mean response count of shape specific misinterpretations	50
3.5	Regression analysis	51
3.5.1	Mixed effects model with test conditions as predictors	51
3.5.2	Mixed model predictions	53
3.5.3	Mixed effects model with ASW and LEV predictors	57
4	Discussion	60
4.1	Critical assessment and possible sources of error	65
5	Conclusion	66
	References	68
	List of Figures	74

List of Tables	75
6 Appendix	76

1 Introduction

Concert Halls, Theaters and Opera Houses are often architecturally significant symbols for cities around the world. They exist in many different shapes and sizes but all serve the same purpose. The halls set the environment for cultural events and artistic performances to be experienced by large audiences. They passively intensify the auditory experience by amplifying and enriching the sound with reverberation. The reverberation has temporal, tonal, energetic and spatial characteristics, which shape the experience of the visitors. It is well established, that different concert halls produce different experiences of acoustic spatial impression. An important part of the spatial impression is linked to the early arriving sound reflections. The specific temporal and directional pattern of the early reflections has a great influence on the perceived spatial impression and is primarily produced by the architectural shape of the auditorium. In this Master thesis, the question is raised if people are able to identify this spatial impression and guess the corresponding shape of the concert hall. The scientific foundation of this work draws from two fields of research:

- 1 The perceptual aspects of acoustic spatial impression in concert halls is well studied. Spatial impression is an important qualitative aspect of how musical pieces and human speech are perceived in enclosed space. It describes the three-dimensional aspects of the acoustic sound distribution and is linked to the overall quality rating. The architectural shape of a concert hall defines the paths of early reflections, which are primarily responsible for the perceived acoustic spatial impression. Some objective acoustic measurements can help to unfold this complex auditory process.
- 2 Being able to perceive the form of the physical environment through hearing is known as a method called *Echolocation* and is most commonly used by blind individuals. Research on this subject has found evidence, that blind as well as sighted people are able to detect objects and surfaces merely by hearing sound reflections and interpreting the spatial changes they produce in the immediate sound environment. This process is partly intuitive but can also actively be learned. The active use of this method is known to work best within short distances ($d_{\text{critical}} = 2,5 \text{ m}$) to the reflecting surface and self generated sound as well as the ability to physically move around. Even though these constraints can not be met in the investigated concert hall setting, the underlying principle lets reason to believe that it is possible to hear the shape of concert halls on the basis of similar cognitive processes.

To investigate this subject, one has to acknowledge, that the perception of space is dominated by vision unless this sense is impaired. Thus, the auditory perception of

acoustic spatial impression has to be treated as a multimodal perception concept. In order to construct a test environment to investigate, whether it is possible to identify the acoustic spatial impression of an architectural space, both the auditory and the visual impression of space need to be displayed accurately and in a natural way. As stressed by Weinzierl and Vorländer (2015), such an investigation relies on state-of-the-art technologies for room acoustic simulation and auralization in order to have total control over the architectural form, objective acoustic parameters, the audio content and the position and directivity of the source and listener addressed. This allows to influence the properties of the investigated concert halls and helps to unfold the multidimensional profile of perceptual features of room acoustical environments. The listening test, constructed for this work, also made use of a stereoscopic Virtual Reality Display in order to present the architectural features of the halls in a comprehensible and natural way.

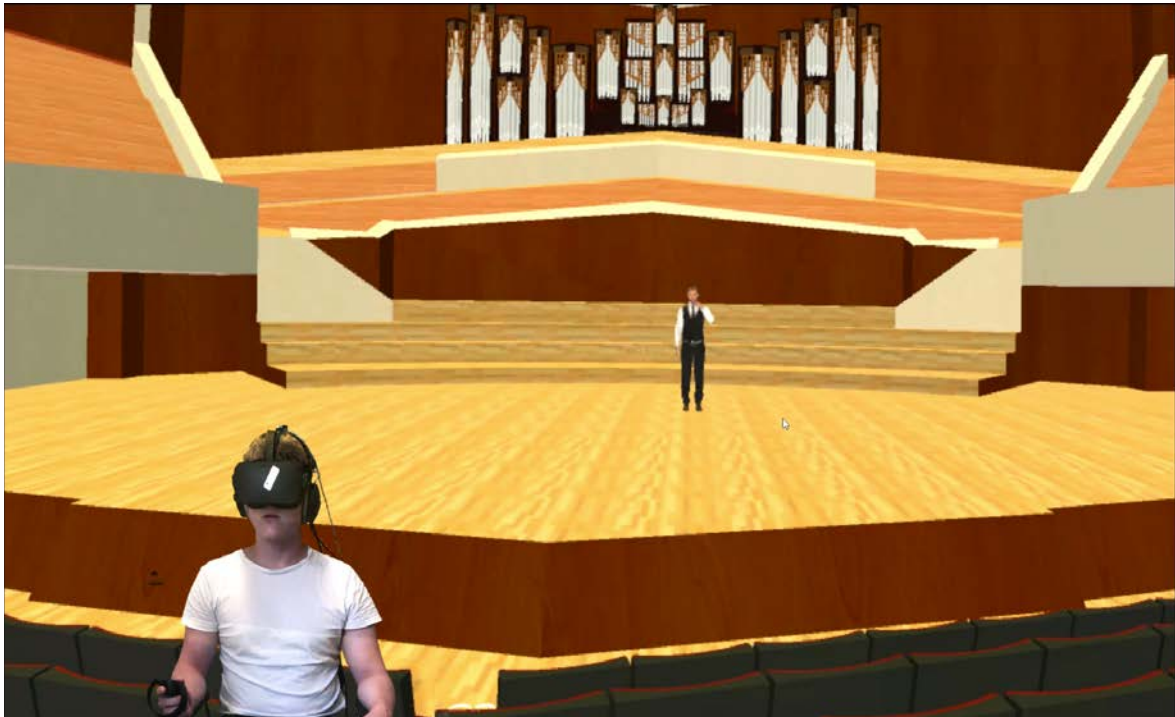


Fig. 1: In-game view of the virtual Vineyard hall. The point of view perspective of the shown test participant is displayed.

In the test, participants were placed in the first block of the main tribune of virtual concert hall models and one male speaker was placed on the stage. Even though the graphical resolution and detail of the models does not rise to the level of photorealism, it enables the viewer to perceive the proportions and the architectural shape as well as the distance to the sound source and the models surfaces with sufficient accuracy. Some graphical additions were added to the visible models to enhance the immersion and provide visual anchors for better size impression. Together with three-dimensional dynamic acoustic auralizations, this technological effort allows to give participants the

audiovisual experience of sitting in a concert hall. Both the Virtual Reality headset and the auralization method used head tracking, which enabled participants to gather information on the spatial characteristics of the concert halls in any head orientation. An “in-game” perspective of the test environment in the Vineyard hall is shown in figure 1. In a Four Alternative Forced Choice (4-AFC) format, participants were asked to select one of four auditory stimuli, consisting of auralizations of the differently shaped concert halls with similar acoustic conditions. One concert hall was visually presented and defined as the target. Participants had to find the auralization of the concert hall they were virtually sitting in by matching the acoustic impression with the visual impression of the performance space. The aim of this research project is to investigate the auditory cues involved in the auditory perception of space and evaluate the discrepancy between the visually expected and actually perceived acoustic impression. It is expected that people have consciously or subconsciously memorized characteristic spatial impressions of basic room shapes or specific architectural elements and are able to match the correct auralization based on expectation or instinct. Digital models of concert halls were specifically designed to resemble the classical shape types: *Shoe Box*, *Vineyard*, *Fan* and *Horseshoe* in their most basic form. Room acoustic simulations of the models, in the form of Binaural Room Impulse Responses (BRIR), were generated with the hybrid acoustic simulation tool *RAVEN* (Schröder and Vorländer, 2011) and auralized through dynamic binaural room synthesis using the *SoundScape Renderer* (Ahrens et al., 2008). The acoustic conditions were predefined with two settings for the parameters: Reverberation Time (RT), Wall Scattering (SC) and Room Volume (V). The exact measures were chosen to resemble common conditions in real concert halls. As a Result, 32 acoustic auralizations were created and compared in the listening test so that each auralization was set as the target stimulus once.

The acoustic stimuli were theoretically accurate acoustic simulations of the digitally modeled concert halls and differ primarily due to the shape. As the excitation signal, a anechoic speech recording of a professional speaker was used which was placed 1 m to the right of the center of the stages. The listener was placed in the first block of the main tribune. A “point of view” perspective of the virtual test environment with the visible speaker on the stage is shown in figure 1. In this listening test only one sound signal and one relative listening position was investigated but different scenarios could be implemented. The participants were randomly split into two subject groups, one of which received a training before the test and the other was untrained. This research project on the audiovisual space perception in concert halls can help to understand this multimodal perceptual ability. Knowledge on the auditory perception of space can help to improve technological tools in the field of acoustics and virtual acoustics. At the same time, proof of the expedience of the test environment is a promising outcome for further similar research tasks.

1.1 Room acoustics and spatial impression

Acoustic planning procedures rely on well established acoustic measurements, e.g. Reverberation Time (RT), Strength (G) and Clarity (C80) which are derived from measured impulse responses and are defined specifically for performance spaces in the ISO 3382-1:2009 standard. These attempt to describe perceptual properties of the room acoustic impression, which has temporal, tonal, energetic and spatial dimensions. A number of qualitative perceptual features, including other perceptual dimensions, were gathered by expert listeners in a round table discussion and concluded in the Spatial Audio Quality Inventory (SAQI) (Lindau et al., 2014). This set of 48 descriptive terms stresses the subjective complexity of the perceivable acoustic impression and the need to investigate the underlying physical properties further. Especially the spatial aspects of the reverberation need further contemplation, considering that the room itself does not produce sound but only shapes the perceived appearance of the auditory content. Binaural features, derived from Binaural Room Impulse Responses (BRIR), relate more closely to the signals perceived by a listener at the ear canals. These are especially relevant for measures of the perceived spatial impression (SI) since they are able to capture differences of the two signals, received at both ears. The perception of acoustic space is subtle and has to be treated as a multidimensional concept which depends partially on the sound content, the properties of the sound source, the resemblance of the measurement procedure and the actual listening situation as well as on the individual expertise and preference. In this work, neither the acoustic quality, nor specific acoustic cues are rated in the listening test, but the higher order cognitive ability to interpret the spatial cues produced by concert halls of different architectural shapes is investigated. Objective measurements are used to interpret the test results and predict the relevant acoustic cues which were used by participants to fulfill the test task.

Due to the nature of the human hearing perception, attention is predominantly drawn towards the sound source content and localization of the source position. The acoustical environment is rather perceived through timbre changes and reverberation than through spatial attributes. Still, spatial impression needs to be considered as an important acoustic attribute in concert hall acoustics, since it influences the quality judgment strongly, as shown by Marshall (1967). Spatial impression has been connected to different attributes throughout the literature: spaciousness, source broadening, objective envelopment, apparent source width (ASW), subjective diffusion, ambiance and others. Especially early reflections, which arrive at the listener position in the first 80 ms after the direct sound are known to shape the apparent spatial dimensions of the sound source. Barron and Marshall (1981) describe spatial impression as a broadening of the sound source and the music gaining body and fullness, as the level of the early lateral

reflection is increased. With strong lateral reflections, the listener experiences the sensation of being enveloped by the sound and being in a three dimensional space filled with music, instead of looking at the it. The spatial dimension of sound perception, entails more detailed attributes as proposed by Rumsey (2002). With a scene based approach he determined source and room specific terminology for spatial attributes constituting SI which can be found below:

1. **Source-related terms:** Source width, source depth, source height, source distance, source envelopment, source localization accuracy
2. **Room-related terms:** Room width, room depth, room size, room volume, room envelopment, level of room reverberation

A descriptive sketch of some of those attributes can be found in Figure 2 which is derived as a single source scene from Mason et al. (2004).

SI as described above is dependent on the time of arrival, direction of incidence, frequency content and loudness of the early reflections generated by the walls, floor and ceiling of a room. The according quality measure for concert halls, named "spatial responsiveness" (SR) by Marshall (1967), is closely related to room shape as the defining condition for the early reflection structure. SR is said to make the difference between "good" and "great" concert hall acoustics, even if the compared halls satisfy the desired monaural acoustic attributes. As a matter of fact, in a quality study on concert hall acoustics by Beranek (2004) two-thirds of the 15 best ranked concert halls are "shoe box" shaped. Arguably room reverberation as well as other monaural acoustic characteristics can have somewhat overlapping effects with the SI. Considering, that early reflections are produced by the sound bouncing of the room boundary surfaces with acoustic properties, not only the relative angle but loudness and the frequency proportions of the reflections are determined by the acoustic parameters of the wall, ceiling and floor materials, as well as the room volume, shape and the positioning of the the source and the listener. All of which also influence the RT, Clarity (C80), Early Decay Time (EDT) and other acoustic features. This aggravates a clear perceptual auditory stream segregation of source-related and room-related spatial acoustic cues. The temporal delay, angle of incidence and frequency content have shown to produce different perceptual effects. In an effort to simplify this research task, the effects of specific single lateral and overhead reflections where among others investigated by Barron and Marshall (1981). He found that the threshold of audibility for a lateral reflection with an azimuth angle of $\phi = 40^\circ$ lies around $\Delta L_p = -20$ dB compared to the direct sound and for 5 ms to 20 ms delay and then declines by 0,06 dB/ms. An additional ceiling reflection with an elevation angle of $\theta = 40^\circ$, a delay of 32 ms and a level difference to the direct sound $\Delta L_p = -2$ dB raises the threshold by 2,5 dB. The angle description can be seen in Figure 2. Frontal and overhead reflections tend to

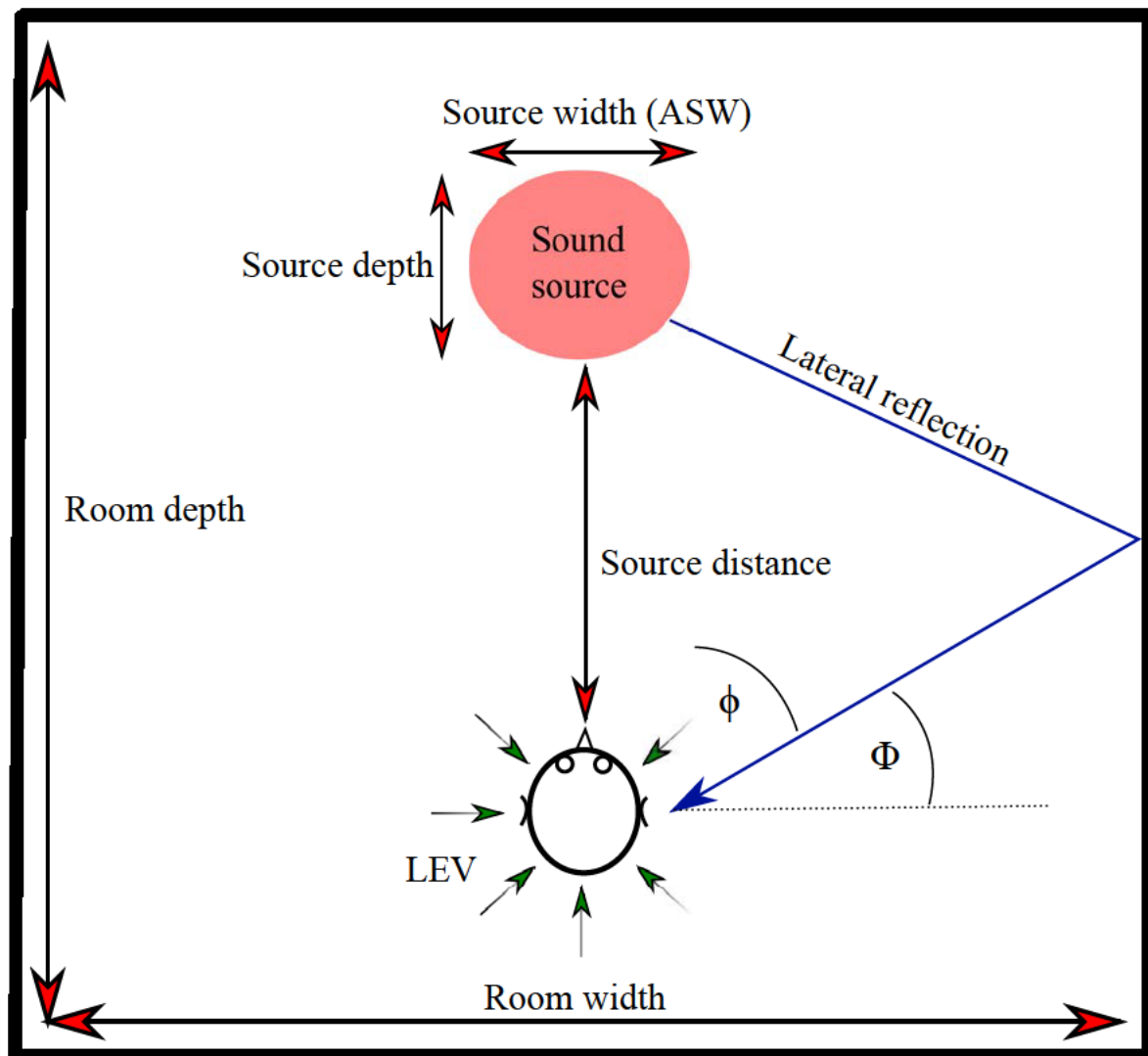


Fig. 2: Sketch of spatial attributes in an acoustic space with a single sound source. Threedimensionality has to be taken into account but is not shown in this sketch. Source and room height correspond in a similar fashion on the vertical axis.

have a more intense effect on image shift (source localization above or below the actual source position) and spectral coloration. Subjective effects dependent on the time delay threshold of single reflections were determined by Schubert (1966) and are shown in Table 1. The temporal forward masking threshold of early reflections was shown to be 7 ms. The perceptual manifestation of the audibility of the reflection is due to spectral coloration in the early and temporal aspects for latter reflection (Buchholz, 2007). The effect of echo disturbance is signal and time dependent as discovered by Dietsch and Kraak (1986) and uncommon in acoustically optimized concert halls.

The two most dominant perceivable subcategories of SI, are Listener Envelopment (LEV) and Apparent Source Width (ASW). The ASW is defined as the perceived width of a sound producing entity and was shown to depend on the angle of incidence and the energy of lateral arriving early reflections in comparison to the total early

Tab. 1: Subjective effects dependent on the delay threshold of a single lateral reflection at $\theta = \phi = 40^\circ$ for the lateral reflection and $\phi = 0^\circ$ for the frontal reflection (after Schubert (1966))

Delay (ms)	Frontal reflection $\phi = 0^\circ$	Lateral reflection $\phi = 40^\circ$
0	Loudness	Apparent source width or image shift
5	Tone coloration	
10-20		Tone coloration
30-60		Spatial impression
> 80	Echo disturbance	Echo disturbance (LEV)

arriving sound energy (Reichardt and Schmidt, 1966). Besides the delay, the angle of incidence also has be considered. Johnson and Lee (2019) discovered the perceivable threshold angle for reflections conditioning the ASW to be $\phi = 40^\circ$ and 130° . LEV is defined as the degree to which a listener has the experience of being enveloped by sound. Bradley and Soulodre (1995) found that late arriving lateral energy has a higher contribution to LEV while early arriving lateral reflections tend to increase the ASW. Morimoto et al. (2001) could show, that reflections arriving from behind the listener increase the LEV, with later reflections being more effective than early ones. Potter (1993) classified the different factors and dependencies influencing SI into main categories as shown in figure 3.

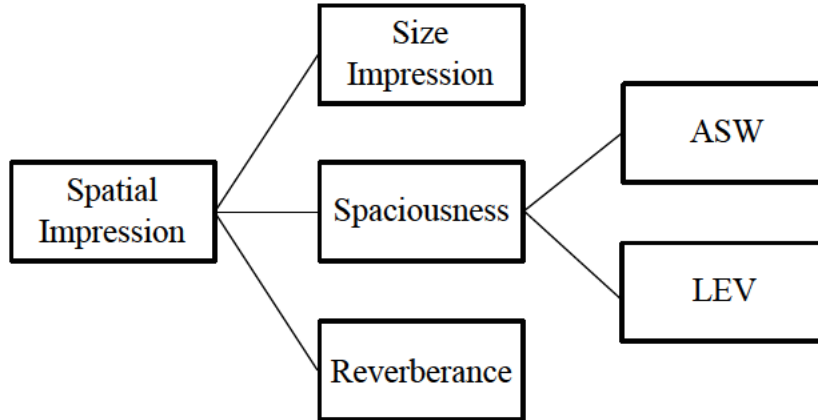


Fig. 3: Terminology connection of the specific aspects of SI

Barron and Marshall (1981) proposed a physical measure, derived from subjective test data, to predict the ASW called *Lateral Energy Fraction* L_f . This measure sets the lateral arriving energy in relation to the total arriving energy and can be derived from measurements with a figure of eight microphone as described in ISO 3382-1:2009.

The attempts to describe the subjective degree of the ASW as linearly dependent on the listening level but does not consider the frequency dependence. Low frequencies contribute well to ASW and low frequency of late reverberation can support the impression of LEV (Potter, 1993). Low frequency absorption of audience areas to the sides of the listener instead have a negative effect on the ASW and LEV of a concert hall.

Frequency components above 1,5 kHz were in return shown to have little influence on the ASW considering the auditory mechanisms involved and a linear weighting model is proposed. The just noticeable difference of the ASW and LEV was investigated in a subjective listening test and binaural measurements of real concert halls by Witew et al. (2005). In a paired comparison test, participants were asked to rate the perceived ASW and LEV and results were compared with the L_f , calculated from measurements with a figure of eight microphone. The results show, that the Lateral Energy Fraction could not accurately describe the perceived ASW and that it is highly dependent on the stimulus signal. A more convenient and accurate way to predict subjective ASW is through the Interaural Cross Correlation Coefficient (IACC), derived from the Interaural Cross-Correlation Function (IACF) (see equation: 1 and 2). Here, p_L and p_R are the two BRIRs at the left and right ear canal and $-1 \text{ ms} < \tau < 1 \text{ ms}$ as described in ISO 3382-1:2009.

$$IACF_T(\tau) = \frac{\int_{t_1}^{t_2} p_L(t)p_R(t + \tau)dt}{[\int_{t_1}^{t_2} p_L^2(t)dt \int_{t_1}^{t_2} p_R^2(t)dt]^{\frac{1}{2}}} \quad (1)$$

$$IACC = \max|IACF_T(\tau)| \quad (2)$$

IACC was endorsed as a measure on a psycho-physical basis for the perception of ASW by Potter (1993). Okano et al. (1998) found a good correlation between subjective impression of ASW and the arithmetic average of $[1 - IACC_{E3}]$, at the 500, 1 k and 2 kHz octave band, combined with the strength factor G_{low} at frequencies between 125 Hz 250 Hz. The subscript “E” stands for “early” and describes that $IACC_E$ is derived from the BRIR and the time range from 0 – 80 ms after the direct sound. Ando later found an orthogonal relationship between the subjective preference and the $[1 - IACC(A)]$ value, which describes the A-weighted IACC Ando (2012). Hidaka found, that the $[1 - IACC_{L3}]$ (late) for the time range from $SI80ms$ to 35 000 ms can be used to predict the LEV (Hidaka et al., 1995). These binaural acoustic features are useful to measure and compare the spaciousness of concert halls but are not able to describe all subjective characteristics of the acoustic impression in a concert hall setting. (Griesinger, 1997) proposed a hypothesis, that the spatial impression of the background (space behind the listener) is cognitively separated from the foreground and conditioned primarily by reflections arriving at least 120 ms after the direct sound. This was verified by Kahle (1995), who also found, that early lateral energy decreases the level of spaciousness in some acoustic scenes rather than increasing it. Most of the referenced research was investigated in laboratory test setups. A room acoustic scene though entails multiple interdependent auditory effects that make up the spatial impression as a whole. Individual participants will likely focus their attention on different auditory cues and reach individual interpretations. In this work, the uncertainty is countered by generating a realistic concert hall scene with defined acoustic conditions,

which were similar for all participants. Also, the acoustic and architectural parameters RT, SC and Volume of the compared concert halls were similar to guarantee that the relevant acoustic cues were primarily produced by the hall shape. Several research projects have investigated the audiovisual interaction of space perception.

1.2 Echolocation

The human hearing system is capable of capturing information on the surrounding environment, mainly through entities emitting sound but also through sound reflected at surfaces and objects. The knowledge of this ability dates back to 1749 when Diderot, D trans. by Jourdain (1916) described the “amazing ability” of blind people avoiding obstacles. What was first called “Facial Vision” and believed to be a sensory effect of the facial skin was later discovered to be an auditory effect (Supa et al., 1944). Gathering information on our physical surrounding that does not emit sound itself is a basic part of the human perceptual experience, but often overseen because the localization of discrete sound sources and the visual sense take a dominant role in the assessment of our surroundings and behavioral adjustment. It is often said, that the ears guide the attention of the eyes, which supposes that we seek visual confirmation if an auditory event outside of the current field of vision catches our attention. Echolocatory information becomes a conscious tool, when vision is impaired due to a lack of light or blindness. In fact, the ability of using echolocation is most commonly known for blind individuals that learn this particular skill, either in a natural way, or in specific school programs. Even though the extend of the echolocation ability seems to vary between persons, some are quite outstanding. Thaler et al. (2011) tested two blind subjects that use echolocation in their daily lives during hiking, mountain biking and playing basketball by generating clicking noises and interpreting the reflected sound. The mechanism using self generated sound is common but also external sound sources like tapping a cane on the floor or independent external sources can be used (Burton, 2000). Yet, since the ability of echolocation increases with learning, one can speculate that familiar sounds are more reliable. Rojas et al. (2010) found short transient sounds around 10 ms to be most effective. It was also discovered, that high frequency components of the generated sound have a bigger impact on the localization accuracy, especially when the obstacle, that is supposed to be located, is further from the subject (Rowan et al., 2013). Further Research from Schenkman and Nilsson (2010) suggests that the detection accuracy of a reflecting surface by blind and blindfolded sighted subjects increases when the duration of the excitation signal increases from 5 ms to 500 ms. They also found that echolocation works slightly better in normal than anechoic rooms, suggesting that the echolocation process profits from additional spatial information generated by the reverberation and environmental sounds. The operable range showed to be below 2,5 m, which relates to a reflection delay of about 7 ms. Inter-

estingly, Buchholz (2007) found a similar threshold of reflection masking by the direct sound, while earlier reflections are perceptually integrated and merged with the direct sound. Spectral coloration and loudness cues seemed to be most present in this temporal span. Many researches, testing the echolocation abilities of sighted blindfolded and genuinely blind subjects, found that blind individuals are often slightly more accurate in the given tasks (Santani Teng, 2011). It is assumed, that the visual cortical areas are recruited for auditory processing, in blind people. Some neural studies support this theory (Thaler et al., 2013). Echolocation is used actively and consciously by people who are dependent on it. The acoustic cues known to be used for echolocation are summarized by Kolarik et al. (2014) and listed below with some addition by the author:

1. **Energy:** The reflection of the excitation signal increases the energy of the sound at the ears. This responds in the sensation of loudness and yields the information of the direction of the reflecting surface, through the Interaural Level Difference (ILD). The level of the echo also holds information about the sound absorption of the reflecting surface. In contrast, research has shown, that the level of the reflection has considerably less influence than pitch and time dependent cues (Schenkman and Nilsson, 2011).
2. **Time delay:** Depending on the distance to the reflecting surface, the time delay between the excitation signal and the reflection will be perceived as such. This is dominant for delays above 30 ms. For small delays 1 – 30 ms the dominant audible effect is a small change in pitch. The so called “time separation pitch” is inversely related to the the delay (Bilsen, 1966) .
3. **Frequency response:** Constructive and destructive interference between the signal and the reflection result in ripples in the perceived sound spectrum known as comb filter effects. This may be perceived as a change in timbre or pitch. The timbre change is also dependent on the absorption and scattering conditions of the reflecting surface.
4. **Binaural cues:** Depending on the position of the target surface, the two ear signals differ. These produce all the localization cues, that an external source would. Since the reflection carries less energy than the excitation signal and is likely to be partially masked by it, the binaural cues are subtle but audible in the operable range.
5. **Reverberation:** In a reverberant environment, the reverberation pattern will be altered by an obstacle in close range. The short distance likely increases the level of late reverberation in the direction of the object. Reverberation further is used as a strong indicator for room size perception (Hameed et al., 2004).

The strength of the acoustic cues described above are dependent on the specific situation and occur concurrently. Through movement the perceived cues can change and therefore augment the information received by the echolocator, which gives a slight advantage to the echolocation accuracy Rosenblum et al. (2000). Some research has shown, that echolocation experts as well as untrained sighted subjects can gather information on the shape and texture of objects through echolocation quite well Hausfeld et al. (1982). Little research has been done on whether humans can confidently recognize the shape of a larger architectural space through the acoustics or echolocation mechanisms. Daniel Kish, one of the most prominent blind echolocators, described the ability to “see” the three-dimensional surface geometry, shape, size and the height of a large auditorium in a Ted Talk ¹. Considering the research described above it becomes evident that echolocation is more focused on close object recognition and nearby walls or the shape of the floor. This information is obviously more important for a blind person to navigate than the actual shape of the room they are in. Thus, it has to be assumed that cognitive echolocation mechanisms play a minor role in recognizing the shape of a large concert hall especially with an external unknown sound source. Still, described research suggests, that the auditory space perception has unexplored potential and that the acoustic identification of a concert hall shape can be possible, at least with sufficient training. Additionally, the acoustic cues which are theorized to be important for the perception of acoustic spatial impression are similar to the ones, used in echolocation.

1.3 Related work and audiovisual interaction

Many research projects have used virtual audiovisual environments to test specific sensory abilities and aspects of the audiovisual perception. This approach shows great promise for many different research questions but relies on extensive technological effort. Maempel and Horn (2017) constructed a virtual concert hall by means of binaural synthesis and a stereoscopic projection on a cylindrical screen. This test environment enables research on audiovisual perception of performance spaces but can only display a visual field of 161°. A group at RWTH Aachen constructed an interactive cave-like environment that enables research projects of many kinds (Schröder et al., 2010). This virtual reality method has the advantage, that multiple subjects can simultaneously partake in audiovisual listening tests. Vorländer et al. (2015) provided a good overview on the technical approaches and demands for audiovisual virtual reality environments in the context of architectural acoustics. An often investigated subject making use of audiovisual virtual environments is the perception of distance and room size. An audiovisual interaction of the expected room size was shown to correlate with the reverberation time but a general underestimation was present (Maempel and Jentsch,

¹<https://youtu.be/uH0aihGWB8U>

2013). An audiovisual underestimation of the source distance was not revealed but the acoustic distance variation showed to contribute more than an optical variation. In virtual reality environments, using head-mounted stereoscopic displays, perception of distance has shown to be somewhat compressed (Grechkin et al. (2010); Willemsen et al. (2008)). Objects seem to be closer than intended. Finnegan et al. (2017) explored an audiovisual matchmaking task to counter this compression effect by moving the sound source in a simulation to match with the visually expected position. A well-known interaction between the auditory and visual sense is known as the *ventriloquist effect* (Thurlow and Jack, 1973). This effect occurs, when auditory and visual information is simultaneously presented but positioned in separate locations. It refers to the perception of sound originating from a different direction other than the true direction, when a relating visual cue is presented. The audiovisual perception then merges the perceived sound source position with the position of the visually perceived sound source. Valente et al. (2012) investigated the perceived auditory width and the interaction of the visual perception of a performing ensemble in a audiovisual virtual environment. The results showed, that the ASW, measured by the IACC, influenced the width judgment significantly. They suggested that participants were not able to differentiate the actual width of the performing ensemble from the perceived ASW, influenced by early lateral reflections. These results coincide with findings about the ventriloquist effect. Valente and Braasch (2010) investigated the acoustic expectation of ASW and LEV under the influence of a visual impression of virtual rooms. They found, that the seating position, which inevitably changes the egocentric perspective, had a strong influence on the expected ASW and LEV. It was suggested, that participants expected a higher degree of ASW or LEV, when the distance to the sound source was greater and a larger part of the room volume was visible. Additionally, the acoustic expectation was influenced by the presence of a video recording of the actual performer in comparison to when the sound source was represented by an image of a loudspeaker. These results suggest, that the visual sensory information influences the acoustic expectation strongly and needs to be considered in audiovisual experiments.

1.4 Room acoustic simulation with Ray Tracing

Common reverberation tools, used in media production are sufficient for many applications. To create a reverberant environment, that lets the user experience the full three-dimensional acoustic space, requires a three-dimensional recording and playback method. For this work, room acoustic simulations were generated with a hybrid ray tracing method and auralized with dynamic non-individualized binaural room synthesis. Ray tracing is an efficient method to simulate the three-dimensional characteristics of reverberation in a realistic and plausible way. For low frequencies, up to the Schroeder frequency, a numerical wave-based calculation delivers more accurate re-

sults about the modal structure of a simulated soundfield. For higher frequency sounds and more complex architectural environments the computational costs of these methods exceeds practicability. ray tracing method uses the simplification of Geometrical Acoustics (GA) to calculate the sound propagation in an arbitrary digitally generated room model. In GA, the description of the soundfield is reduced to traveling particles representing spherical waves with an infinitely small opening angle. In principle, a high number of particles with incoherent frequency but identical energy are dispersed into a virtual room. For each particle the intensity loss through air absorption, depending on the room temperature and humidity and air pressure, is traced and calculated individually. When hitting a wall, the particles energy content are attenuated depending on the predefined absorption properties of the reflecting surface and the reflection angle is calculated as either specular or diffuse, depending on the predefined scattering properties. If the scattering factor s of the reflecting wall material is above 0, the particles energy is weighted with $(1 - s)$ and the reflection angle shifted according to Lambert's cosine law (Vorländer and Mommertz, 2000). The particles are eventually detected at a chosen listener position represented by a sphere. The detected particles are examined for their frequency and energy content, as well as their direction of incidence. With this data, the corresponding room Impulse response (RIR) can be concatenated. With a set of Head Related Transfer Functions (HRTF) a Binaural Room Impulse Response (BRIR) can be generated which carries all spatial information of the acoustic scene as well as the directional filtering properties of the human head, torso and pinna. Through convolution with an anechoic sound signal, this method attempts to generate the exact binaural audio stream received at the listeners ear canals. However, the human psychoacoustic perception has different temporal sensitivities for different parts of the reverberation phenomenon. Especially the early part of the process after the excitation is very important for the perception of the spatial characteristics of the room (see section 1.1). Especially the direct sound and the early reflections need to be accurate in their temporal and spectral information, since they influence the source localization, spatial impression and many other relevant acoustic features. A stochastic ray tracing method, is used to calculate the reverberation tail, but has a uncertain error rate when calculating the complete early reflection pattern. The late reverberation is still also a major part of the auditory experience, but evaluated by the human hearing system in a much lower temporal resolution (Blauert, 1971).

Therefore the early and late part of the impulse responses are calculated with different methods and merged afterwards. In this thesis the acoustic simulation tool *RAVEN* was used which relies on a combination of an Image Source (IS) model, to accurately calculate the early reflection pattern, and a stochastic ray tracing technique to calculate the late reverberation (Schröder and Vorländer, 2011). The basic concept of the IS model, shown in figure 4, is to find the correct reflection path by mirroring the sound source at the reflecting wall and using the corresponding secondary source

(S2) position for the calculation of the reflected sound ray. This guarantees that all early reflections are fully received at the listening position and represented in the simulated RIR. Diffraction was not considered in this listening test, because the listening positions are not directly exposed to diffracted or shadowed sound energy.

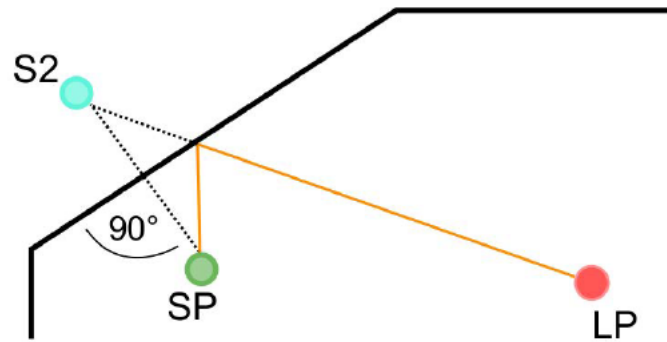


Fig. 4: Descriptive sketch of the Image Source Model in 2D with the secondary source (S2) being mirrored at the reflecting wall. The correct reflection path, plotted as the orange line, can be found by drawing the line between S2 and the Listener Position LP.

The direct sound and the early reflections are then overlapped with the late reverberation to generate the full Impulse Response (IR). For the calculation of the late reverberation, the reflection angle assumed to be diffuse. It is then shifted by a random amount, weighted with s and the energy attenuated accordingly. The late reverberation is measured by dispersion of a finite number of particles from a sound source sphere with even distribution on the spheres surface. The energy distribution of the particles can be manipulated by an arbitrary directivity pattern to simulate the desired type of the sound source. The particles paths are traced until their amount of energy falls below a chosen threshold or the particle is detected on the reception sphere at the desired listening position. The traced particles are best described as a sound rays traveling in straight lines. The process is split in 31 intervals of thirds, with each particle representing only one frequency band. The direction and time of arrival, as well as the frequency and energy content of all particles, detected on the reception sphere, are recorded. The information of all particles are concatenated at the correct time slot and merged to gain the full frequency spectrum and directional and temporal information of the simulated acoustic scene. In the final step, these directional impulse responses are convoluted with a HRTF dataset to generate BRIRs. This BRIR can be calculated at any listening position and a desired head orientation. The virtual auditory scene can be played with any anechoic audio stream as input for the sound source.

1.5 Research goals and hypothesis

From the described research on the subject of echolocation, it becomes clear, that human beings have a certain intuitive capacity to gather information on their physical surroundings by hearing sound reflections and interpreting spatial changes they produce in the energetic, temporal and tonal structure of the perceived sonic environment. However, this mechanism requires attention and training to be accurately used. It is unclear, if echolocation cues play a role in the chosen virtual acoustic scene. Still, it lays the grounds to suggest, that some people will be able to access similar cognitive mechanisms and hear the shape of the concert halls. Research on spatial impression in concert halls suggests, that differences in ASW, LEV and other acoustic features will be audible between the differently shaped halls, when the reverberation time, degree of wall diffusion and the volume are identical. These effects were so far primarily investigated with respect to quality ratings. Whether the participants can recognize the shape of the performance space by interpreting the spatial characteristics of the reverberation is the aim of this thesis. Research on audiovisual interaction and expectation have shown, that the visual impression influences the expectation and perception of the acoustic impression. Therefore, the creation of a realistic audiovisual three-dimensional environment is essential for this project. It is of interest, if the visual perception of an architectural space provides an expectation of the acoustic impression and allows test participants to match the two modalities. Further it is of interest, if the acoustic conditions defined by the wall materials and the room volume influence the ability to assign the correct auralization to a visually presented concert hall. In this exploratory study, with multiple test conditions, an exact assessment of the responsible auditory cues is not the aim. It is rather an attempt to close in on the boundaries of auditory space perception and to narrow this complex field of research. The following hypotheses are raised:

- H0 Human beings are not able to identify the shape of concert halls by interpreting the acoustic spatial impression.
- H1 Human beings are able to identify the shape of concert halls by interpreting the acoustic spatial impression.
- H2 The acoustically relevant attributes *volume*, *reverberation time* and the degree of *wall scattering* influence the acoustic spatial impression and the ability to identify the shape of concert halls.
- H3 Training increases the probability to correctly identify the shape of concert halls through the acoustic spatial impression.

2 Method

To test the hypotheses, an audiovisual listening test was constructed. The test was subject to two basic requirements: Test participants needed to be able to hear the acoustic characteristics of the concert halls conditioned by the shape and at the same time, see the halls in a natural way to gain a clear visual impression of the shape and size. The room acoustics of the concert halls under investigation, were simulated using a ray tracing method and with different acoustic conditions, described section 2.3. Non-individualized dynamic auralizations of these room acoustic simulations were generated to make the rooms audible through headphones (see section 2.4). The visual component of the test was generated using a head-mounted Virtual Reality (VR) display. The test setup was targeted to generate a comprehensible, immersive virtual environment that enabled the test participants to experience the concert hall simulations visually and audibly. Many research projects have successfully tested perceptual abilities in virtual audiovisual environments, which supports the approach (Llorach et al., 2018). Four concert halls were investigated with the prominent shape types: *Vineyard*, *Fan*, *Shoebbox* and *Horseshoe/ Theater*. The models were based on existing concert halls, but recreating specific halls in detail was not the aim. They were designed to represent the basic architectural features of the halls in a realistic way. In order to investigate the influence of acoustic conditions, the volume (V), reverberation time (RT) and wall scattering (SC) were defined to be tested in two levels each. The Volume, as an architectural condition, was chosen to be $V_{\text{big}} = 20\,000\text{ m}^3$ and $V_{\text{small}} = 5000\text{ m}^3$. V_{big} lies within the normal range of many prominent concert halls (comp. Beranek (2004)). V_{small} is an atypically small size for concert halls due to unfavorable acoustic conditions and a low seating capacity. Yet, echolocation abilities are more accurate when the distance to the reflecting surface is short. In hope to see this effect, the substantial size difference was chosen. The reverberation time and degree of wall diffusion are conditioned by the absorption coefficients $\alpha(f)$ and scattering factors $\gamma(f)$ of the surface materials (see section 2.3). They were each chosen to emulate common acoustic conditions found in real concert halls. The RT was chosen to be $\text{RT}_{\text{low}} = 1,5\text{ s}$ and $\text{RT}_{\text{high}} = 2,5\text{ s}$. Most large concert halls have reverberation times in the range of 2 to 4 s. In the small models (V_{small}) a higher RT will generate an unnatural loudness. This could cause the late reverberation to partially mask the early reflections and weaken the spatial impression (comp. Bradley and Soulodre (1995)). The wall scattering coefficients were calculated on a theoretical basis and are further described in section 2.3.6. The acoustic of each concert hall type was simulated in each combination of the described acoustic conditions and sums up to eight test conditions for each shape type. To realize this, a specific surface material was calculated for each

condition and shape, which yields 32 acoustic simulations. The test conditions are listed in table 2.

Tab. 2: Test conditions calculated for each concert hall.

Condition	Ind.	Volume	RT	SC
1		big	low	low
2		big	low	high
3		big	high	low
4		big	high	high
5		small	low	low
6		small	low	high
7		small	high	low
8		small	high	high

2.1 Audiovisual listening test

The listening test was designed in a way, that subjects had to match acoustic auralizations with the visual impression of a concert hall. This approach was taken, since the objective of this test was to investigate a realistic situation where a person is sitting in the audience of a concert hall and is trying to identify the hall shape by matching the perceived with the expected acoustic impression. In the Virtual Reality (VR) environment, participants could easily identify the hall shapes visually and try to detect acoustic cues that match with the architectural features. The obvious approach here would be to present the subjects with one auralization and ask them to choose the matching concert hall out of different visible models. Pretests have shown, that quickly switching between the visible halls leads to discomfort and even nausea when using the VR headset. This could possibly be avoided by modifying the VR system and optimizing the modeling process but could not be achieved for this work. Instead, the test task was inverted. Participants were visually placed inside of one of the hall models and presented with four acoustic stimuli to hear. Each acoustic stimulus consisted of one auralization of the four concert halls described in section 2.2, while one auralization was calculated for the exact model that was visually presented. This test, resembling a classical 4-AFC format, was repeated 32 times by each participant, so that each hall was set as the target stimulus once in all test conditions. Accordingly, the visible model changed after each iteration to match the target hall for the next round. In each round, the test conditions V, RT and SC were identical in all four auditory stimuli. The size of the audience area and the organ was also set to be similar (see section 2.3). This way, the stimuli which were compared were only dissimilar in the acoustic effects of the hall shape. The test subjects were asked to select the hall, that matches the concert hall that is seen and were thus trying to hear the shape of the concert halls. The subjects could repeatedly listen to all stimuli in any order until a decision was made. The sequence of the target hall for the 32 test rounds was

randomized as well as the order of the stimuli in each round.

A “point of view” perspective of the virtual test environment is shown in figure 5. The test was conducted in a between subjects design, where half of the participants received a training before the test and the other half did not. Participants were randomly assigned to the groups.

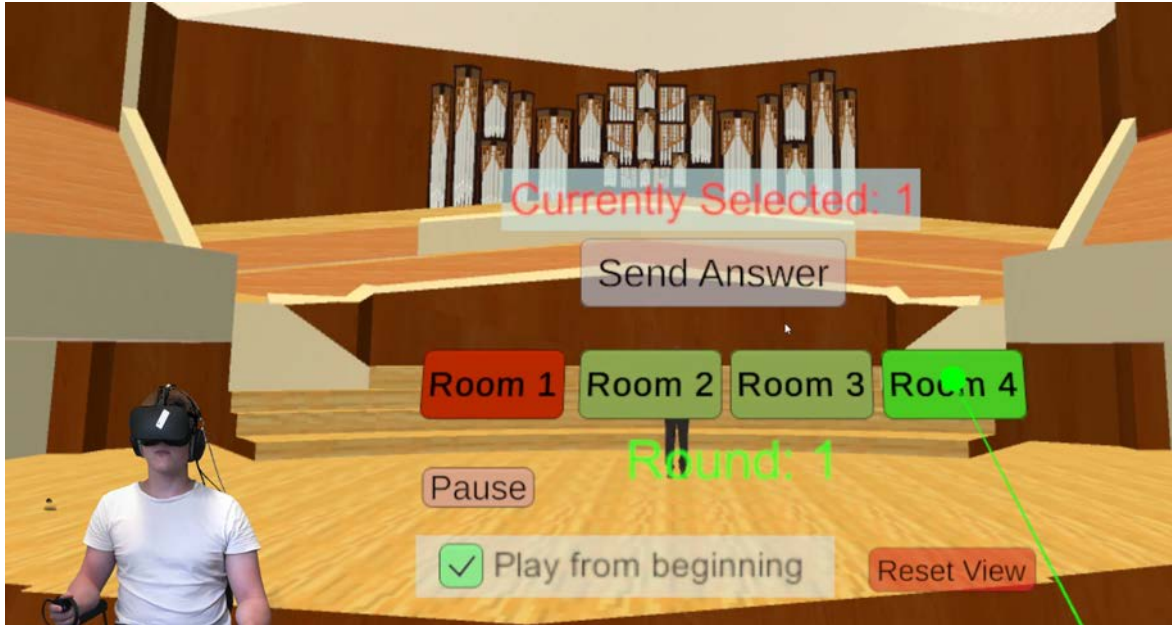


Fig. 5: Point of view of a trial participant virtually sitting in the vineyard hall. The control menu is shown in front of the participant. The selection is done with a laser beam tracked to a hand held controller

2.1.1 Stimulus signal

As a stimulus signal, different sources were tested in various pretests. Due to the number of acoustic conditions under investigation, only one type of sound signal and one source position could be tested to keep the test duration reasonable. It was decided to use a speech signal, instead of an instrument signal or noise bursts. In the context of concert halls with one sound source on stage, using a piano or violin sample would seem appropriate. Research on echolocation (see section 1.2) has suggested, that self generated, transient rich sounds produce the best accuracy in echolocation tasks. It can therefore be expected, that the necessary acoustic cues will be most pronounced with familiar sound signals. Seraphim (1961) discovered that the auditory perception of early reflections is much more sensitive to the delay of the reflection with speech signals, than with most musical signals. One can argue, that speech is the most common type of sound signal. Psycho acoustic effects, like the famous “Cocktail Party Effect” Arons (1992) suggest, that the human psycho acoustic information retrieval mechanism is most trained in perception of speech. This is supported by the fact, that the sensitivity of the human hearing system corresponds with the frequency spectrum of

human speech Blauert (1997). Additionally, speech signals usually contain broadband transients and pauses which allow to assess the temporal aspects of the reverberation. A speech signal was chosen, that fits into the concert hall setting and at the same time provides suitable complexity. The signal consisted of an anechoic recording of Cicero's 3rd Cataline Oration, spoken by a professional speaker in the German language (Böhm et al., 2019). The signal was recorded with a sound pressure level of 91 dB in 1 m distance in an anechoic chamber.

The virtual acoustic scene is played by the SSR using the on board audio device as output. In order to set the correct gain factor to achieve a desired sound pressure level of the speech signal in the virtual concert hall, the level of the signal, produced at the headphones, was measured with the FABIAN dummy head (Lindau et al., 2006). BRIRs of a measurement scene were created to mimic the original anechoic recording setup of the stimulus signal with 1 m distance between source and receiver and no reflecting walls. The sound signal was then played with same software and hardware setup, later used in the listening test. Since the impedance of the FABIAN recording system is known, the corresponding level of the sound source was measured and corrected for a set gain factor. The level of the speech signal was calculated with the method proposed in the DIN (2012). This way the actual level of the source as it was perceived through the headphones could be adjusted. It was set to 88 dB in the big and 85 dB in the small concert halls. The level reduction was chosen by subjective assessment of the loudness by the author to prevent the test participants to experience hearing fatigue. In the small halls, the condition RT_{high} produced an especially unpleasant loudness, which was countered by the reduction of the level by 3 dB compared to the large halls. It can also be assumed, that a speaker would adjust the intensity of their speech in a small reverberant environment.

2.1.2 Procedure and training

Before the test, each participant received a written description of the test task (see Appendix Sec. 6), as well as a form to describe their personal experience with the subject of room acoustics. After reading the written description, subjects were given a verbal explanation of the task, and the handling of the controller and headset. Additionally, the floor layouts of the four concert hall types were shown in printed form and the architectural features discussed shortly. Participants were seated on a revolving chair and asked to adjust the VR headset and headphones to their comfort. In a first step, participants had to synchronize the visual and acoustic environment as described in section 2.1.5. The VR environment and handling is unfamiliar to many people. To make sure, that all participants understood the test task and handling and to avoid errors in the selection process, an example test had to be executed first. This

example test consisted of two randomly chosen rounds of the main test including the auditory stimuli. Participants were specifically instructed to concentrate on the control mechanics and not evaluate the acoustic characteristics in this step. After the example test, participants were asked to start a training session. Half of the participants were chosen to receive a training with auditory stimuli. The other half could only look at all concert halls to gain a good visual impression, but were not given any auditory stimuli to learn.



Fig. 6: Sketch of the virtual training menu. The RT button has a toggle function switching between RT_{low} (dry) and RT_{high} (wet).

The training consisted of a scene, where participants could see the concert halls and simultaneously hear the corresponding auralizations. This way, participants were able to learn and memorize acoustic cues produced by the different concert hall shapes in the different test conditions. The control menu of the training is shown in figure 6. Pretests have shown, that the time spent inside the VR environment has to be kept as low as possible. Some participants reported a feeling of nausea or dizziness after about 20 min. To reduce the time spent in the VR environment, the scattering condition was excluded from the training menu. This reduced the amount of stimuli to be learned in the training by half. All conditions heard in the test task, were set to the low scattering condition (SC_{low}). The decision was founded on observations during various pretests. Different test subjects were asked to rate the amount of audible difference between concert hall shapes in the different condition settings. Verbally given answers have shown, that high scattering resulted in the least audible differences between the concert halls. In the training scene, subjects could listen to the concert halls in all conditions in any order. They were explicitly asked to listen and compare the four different hall shapes in each condition and remember the auditory cues that distinguish them. The training was completed, when participants heard all halls in all conditions and expressed confidence for the given task.

2.1.3 Graphical user interface

The Interface for the test was created and controlled using typical Virtual Reality game mechanics further described in section 2.1.5. A sketch of the main test menu consisting of virtual buttons and info text is shown in figure 7. With the buttons: “Room

1, Room 2, Room 3, Room 4”, the different acoustic stimuli could be selected and played. The currently playing stimulus was marked by the button changing color from green to red. Additionally, the current selection was shown in a text field (“Currently Selected: X”) above the menu. Only after listening to all stimuli, the decision could be confirmed by pressing the “Send Answer” button. The current selected stimulus was saved as the participants decision for the current round, when pressing “Send Answer”. Whenever a decision was made this way, the round was completed and the next round initialized. For each iteration, the visible concert hall was changed to the target hall of the next round. The acoustic stimuli where changed to the next condition and the stimulus “Room 1” started. An additional option to pause the auditory stimulus during the rounds was implemented for convenience. In standard settings, whenever a stimulus was selected, the sound sample was played from the beginning. An option to switch between the stimuli without restarting the audio stream at the beginning was provided but rarely used by the participants (“Play from beginning”). This option was implemented to provide an additional option to avoid listening fatigue. An option to readjust the synchronization of the visible and acoustic environments was also provided in the main test menu (“Reset View”). This was necessary in case of a participant needing to take a break and readjust the VR display or the headphones.

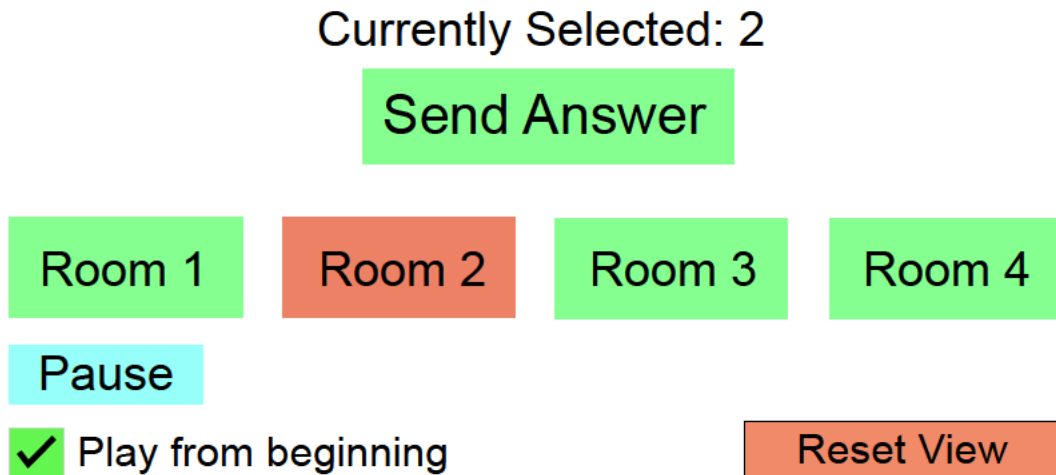


Fig. 7: Sketch of the main test menu constructed in the VR environment. In this example, the stimulus “Room 2” is selected.

2.1.4 Test setup

The test setup was created with two parallel running Systems connected through a local Ethernet network. The VR environment was built using the game development software *UNITY*² and ran on a powerful PC suited for this task. A *Oculus Rift* headset without the optional accessory headphones was used to present the visual environment. With this VR technology, participants where supposed to be immersed

²UNITY 2018 1.19

into the virtual concert halls. Therefore all control options had to be built into the VR environment, enabling participants to run the complete test without the need to communicate with the real world. The auralizations were rendered with a separate, Linux PC and the *SoundScape Renderer* (SSR) software (Ahrens et al., 2008). This software loads the beforehand simulated BRIR datasets into an internal memory and provides a convolution engine to play an arbitrary anechoic sound signal. All spatial and acoustic information is contained in the dataset. The SSR is able to interchange the BRIRs for any desired head orientation and generate the impression of a static sound field. A separate head-tracker³ was used to measure the head orientation of the participants with the tracker mounted on top of the headphones. The SSR uses this information to select the correct BRIRs and crossfade between them with low latency. In combination with the VR headset this system is able to generate the impression of a stable audiovisual environment, that allows free head movement on a fixed seating position. The communication between the two systems was enabled via OSC and TCP Protocol. A structural plan of the test setup is shown in figure 8.

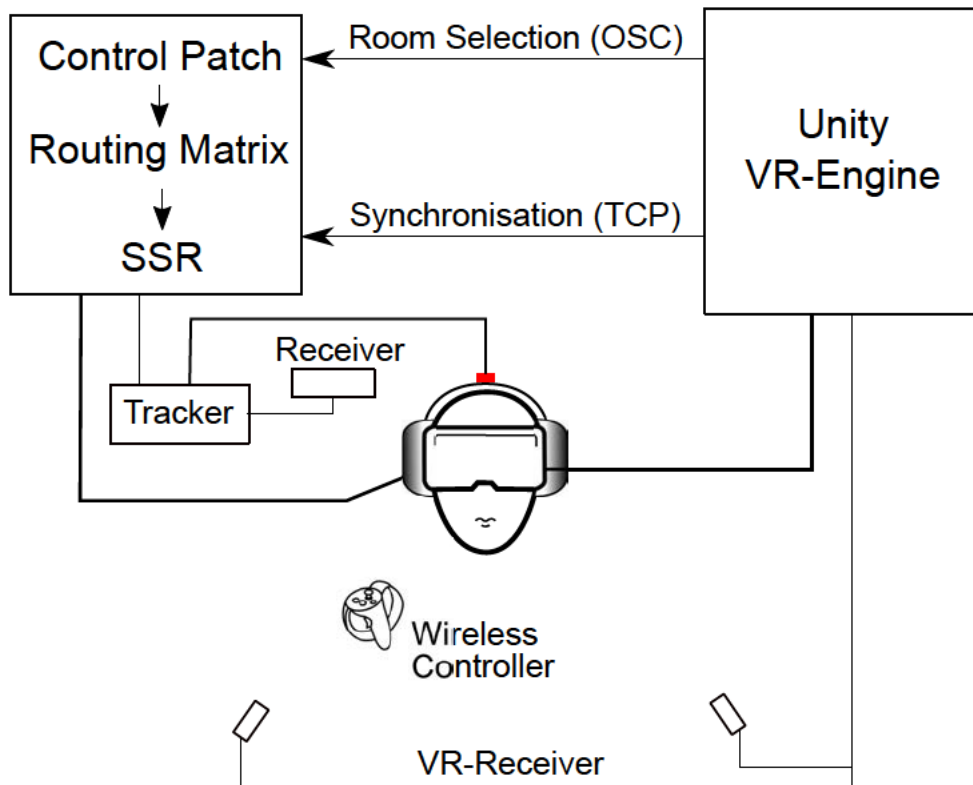


Fig. 8: Sketch of the test setup. Lines represent connecting cables of different types. Boxes represent technical devices.

All BRIR datasets of the simulated concert halls were loaded into the SSR system, creating an internal input channel for each hall through the *Jack Audio* software. The

³Polhemus Patriot FCC Class B

halls can be selected by routing the input signal internally to the according channel of the SSR. This is realized with a *Pure Data* control patch. The control patch receives an Open Sound Control (OSC) control signal from the UNITY system and routes the input signal to the according channel (see technical documentation and Digital Appendix⁴).

2.1.5 Synchronization and control mechanic

The synchronization of the two systems had to be accurate, so that the sound source position matches with the visual source position represented by a virtual speaker figure on the stage of the virtual concert halls. Therefore, the direction of the coordinate system of the VR environment had to be matched with the coordinate system of the SSR. The direction of the coordinate system of the VR environment in the x-y-plane can be set internally to the current frontal head orientation ($\phi = \theta = 0^\circ$) measured by the VR headset. The height of the virtual environment is set by matching the listener position, represented by the “Oculus Camera Object” in UNITY with the head position of the participant wearing the VR-headset. The SSR allows a reset of the tracked frontal head orientation through a Transmission Control Protocol (TCP) command. When the command is received, the current orientation in the x-y plane is set as the frontal view direction with $\phi = \theta = 0^\circ$. Both coordinate systems were set at the same time, which synchronizes the two tracking systems. This is done in the beginning of the listening test with the scene, shown in figure 9. A sphere with crosshairs is presented in front of the participant. The sphere is attached to the headset to generate a reference point for the frontal head direction. Subjects were asked to place the sphere on the horizon by looking straight ahead and hold a trigger on the controller. To eliminate the possibility of a sudden unwanted reset, a loading bar would fill as long as the trigger was held. The synchronization was triggered, when the loading indicator reached the end of the shown bar. A later reset was only possible in the main test, in case that a participant had to take a break or change the position of the VR headset or the headphones.

A typical game mechanic of VR games was used, where virtual control menu was projected in front of the participant and selections could be done using a handheld *Oculus Touch* controller (see Fig. 10). The controller used, has similarities to a pistol handle with multiple buttons. The two buttons used are called *grip* and *trigger*. The grip button can be pressed with the middle finger while the trigger button is pressed with the pointing finger. The control menu was tracked to the main Axis of the VR-headset guaranteeing, that the menu could be seen in front of the participant in all head orientations. On the push of the grip button, the control menu appeared together with a virtual laser pointer originating at the controller. By pointing the laser at a

⁴Digital Appendix/Listening Test SSR Project



Fig. 9: Start screen for synchronization.

virtual button in the menu and pressing the trigger button on the controller, selections could be made. When releasing the grip button the test menu would disappear. This is an intuitive way to handle the control menu while at the same time enabling test participants to hide the menu for a greater immersion effect.



Fig. 10: Handheld *Oculus Touch* controller with grip and trigger button held.

2.2 Digital concert hall models

The main objective in the design of the concert halls was to construct models that fit the given shape type in its most basic form. The definition of the shape types is mostly colloquial, but a basic descriptions of the layouts were given by Long et al. (2006) and are shown in figure 11. A basic layout for the vineyard shape could not be found and was added by the author.

Still, existing concert halls assigned to a certain shape type have different architectural elements that set each other apart. Especially modern concert halls can often not easily be identified as a certain shape type. To guarantee that the halls are representative for the assigned types, some prominent examples of existing concert halls were chosen as reference. The models basic proportions and some specific features where designed

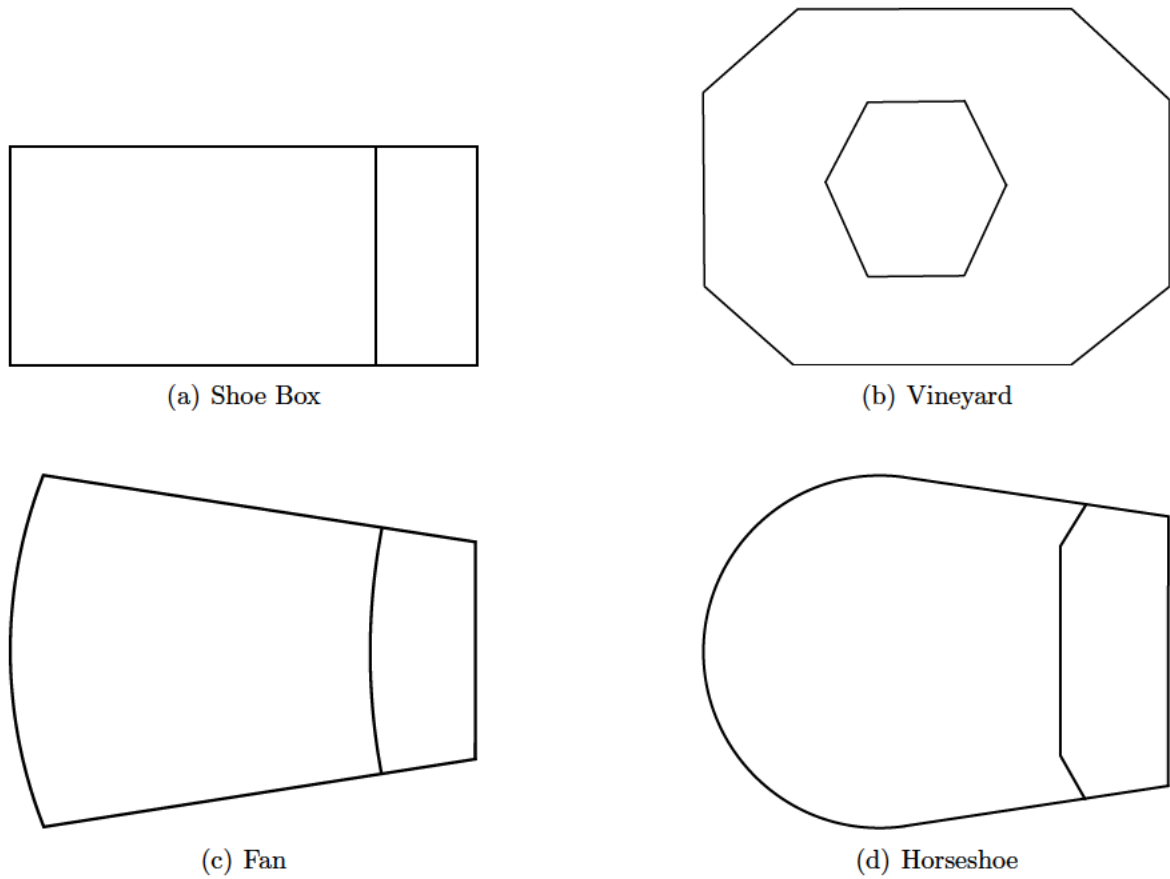


Fig. 11: Descriptive sketches of the layout of the concert hall shape types

by modeling the major similarities of the reference halls. Detailed design elements with small surfaces reduce the frequency to which the ray tracing method can work accurately (see Chapter 4 in: Schröder (2011)). Surely, many acoustically optimized halls contain specific reflection surfaces to guarantee a sufficient amount of early reflections spread over the audience area. This and other specific acoustic manipulation, that can normally be found in real concert halls, was not introduced here to preserve the acoustic effect of the hall shape. Chairs, doors and a speaker were modeled only into the visible models but had no effect in the acoustic simulation. The models were constructed in the 3D modeling software *Sketch Up*⁵. Here all surfaces were assigned with a wall material and the size and basic proportions were adjusted. The four concert hall models generated for this project are described below. The Groundplan and cross section plots entail the early reflections of 1st order. These can be expected to shape the perceived spaciousness the most.

⁵Sketch Up Make 2015

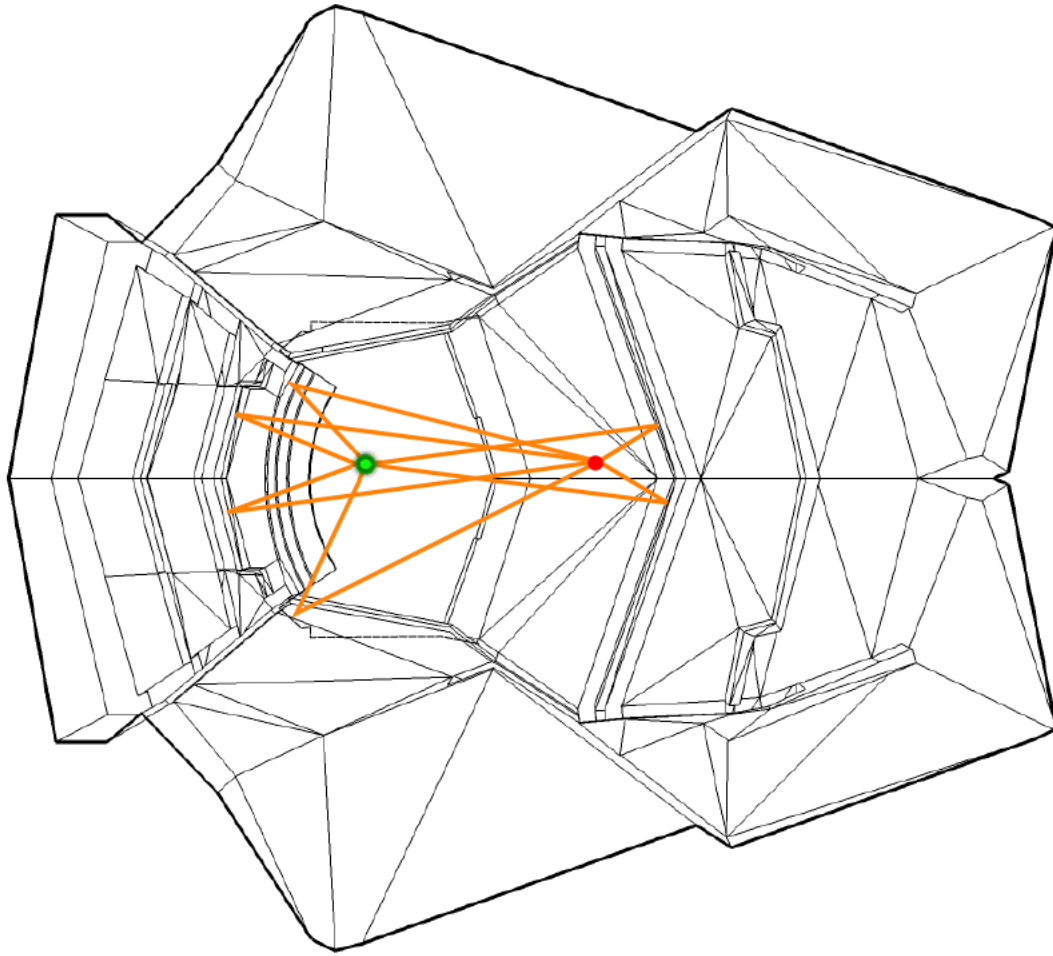
2.2.1 Vineyard

The prominent architectural feature of a Vineyard hall, is the stage positioned in the middle with the tribunes rising from each side. This acoustically convenient shape is architecturally challenging, since the infrastructure for musicians and staff to have access the stage would need to be below the floor or otherwise interrupt a tribune. Additionally, some reflective walls close to the stage are essential for the musicians to hear each other and practice their art as well as possible. The architectural solutions for this problem differ significantly, especially in the shape of walls around the stage. Reflector panels above the stage are often used to generate specific reflections to improve the acoustic conditions for the musicians. For this research the vineyard hall is modeled to closely resemble the *Berlin Philharmony* as a well renowned concert hall for its specific vineyard shape. Groundplans of the original hall were used to build the model and maintain the original proportions.

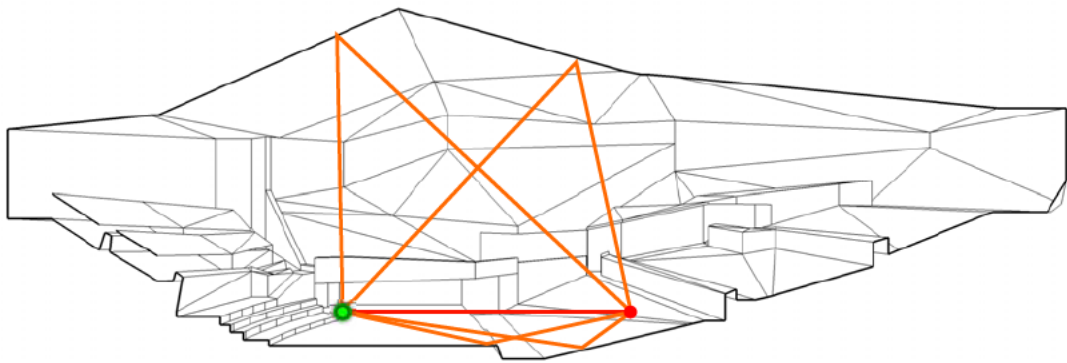


Fig. 12: Wide shot of the visible in-game model of the Vineyard hall.

The hall is built with a main tribune of three seated blocks and a back tribune with two seated blocks. The tribunes to the side of the hall rise up in a steeper angle and reach almost to the beginning of the ceiling. The stage is encapsulated to the sides by a lowered area with a small overhanging balcony. The chosen listener position is in the third row of the first main tribune block and the source position is positioned in the back center of the stage. The ceiling of many vineyard halls is shaped in a convex way with the highest point being above the stage. In the original Berlin Philharmony the ceiling is convex as well. In a polygon model, round surfaces can not be modeled without many small plane surfaces. In a ray tracing simulation this can lead to unrealistic holes in the reflection pattern and explodes the calculation workload. In this model the ceiling is therefore estimated by larger plane surfaces. The groundplan and cross section of the model including reflection paths of the 1st order reflections can be seen in figure 13.



(a) Floor plan



(b) Cross section Vineyard

Fig. 13: Groundplan and cross section perspective of the **Vineyard** hall with lateral, ceiling and floor reflections of 1st order. The sound source position marked as a dot green and the listener position marked red. The early reflections are represented by the orange lines.

2.2.2 Shoe Box

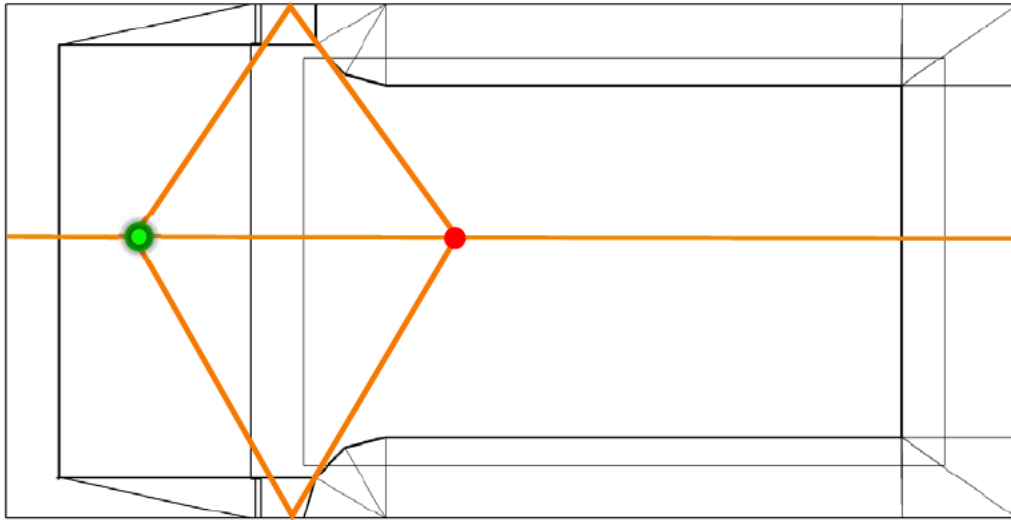
The shoe box shape is already well described by its name. Usually, halls with this shape are squared with the longitudinal side being significantly longer and the ceiling being high. The proportions of this shoe box model were chosen to be $4 \times 8 \times 3$, by using the average proportions of three prominent concert halls of that shape. The concert halls used as reference were the *Avery Fisher Hall* in New York, the *Sala Sao Paulo* in Sao Paulo and the *Herkulesaal* in Munich. All of these halls were evaluated by music experts as having good or excellent acoustics (Beranek, 2004). All entail one or more balconies on the side and back wall and an organ above the stage. Even though they can easily be recognized as being shaped like a shoe box, the details differ. The Avery Fisher hall has a rising tribune and the stage is tapered in its width and height. In the Sala Sao Paulo, the main audience area has a rising part towards the sidewalls and a small rising tribune behind the stage. The sidewalls are also supported by big pillars. The Herkulesaal also has pillars holding up the single balcony and a sloping stage. It becomes evident, that the walls of these halls are structured by the wall elements and pillars. This helps to generate a more diffuse sound field and counter low frequency modes as well as flutter echos.



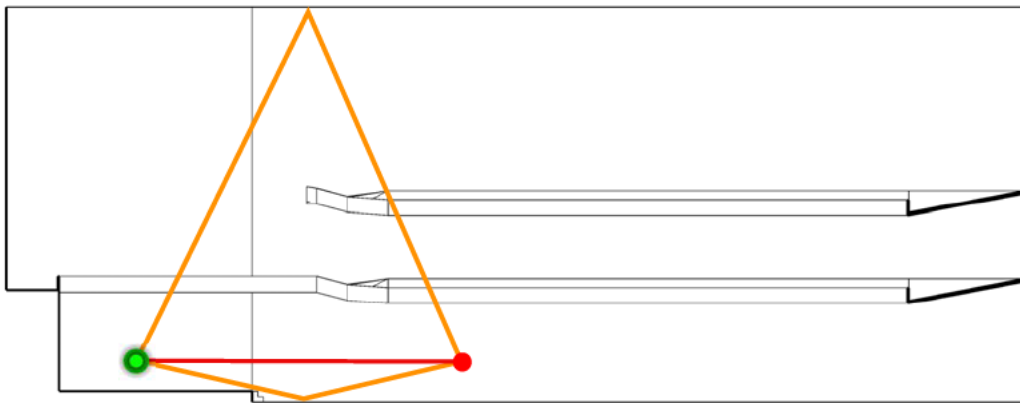
Fig. 14: Wide shot of the visible in-game model of the Shoe Box hall.

The model used in this project is reduced to the basic shape to guarantee for the desired reflections and therefore has unstructured sidewalls. Since only one listening position is taken into account and low frequency diffraction was not calculated with the chosen ray tracing method, structured walls and pillars were not modeled here. These features could also corrupt the correct calculation of the early reflection pattern. Wall structure with an explicable acoustic effect is furthermore accounted for by the wall scattering conditions. An organ with specific acoustic properties was placed above the stage and

can be seen in figure 14. An organ has high scattering factors and is an inevitable part of a standard shoe box shaped concert hall. It does not influence the reflection pattern of the 1st reflections, which would corrupt the comparability between the halls without an organ. In the acoustic simulation model, the organ was implemented as a plane surface with the measured acoustic properties of an existing organ. The two balconies are also acoustically important, since they produce stronger reflected acoustic energy from the sides. The groundplan and cross section of the model including the paths of early reflections of first order are shown in figure 15.



(a) Groundplan Shoe Box



(b) Cross section Shoe Box

Fig. 15: Groundplan and cross section perspective of the **Shoe Box** hall. The sound source position marked as a dot green and the listener position marked red. The early reflections are represented by the orange lines.

2.2.3 Fan

The significant features of the fan shaped hall are the angled sidewalls opening up like a fan and the rising main tribune. This shape is atypical for concert halls but often used for lecture halls and theaters. It is known for having positive effects on speech intelligibility but being problematic for music performances. Prominent halls of that shape differ especially in the height and angle of the ceiling but also in the shape of the stage walls. In most concert halls with a fan shape, many reflection surfaces have to be installed to guide the reflections for an even distribution over the audience area. Also absorbing and diffusing material on the back walls of the hall are sometimes used to counter for undesirable interference and focusing effects.

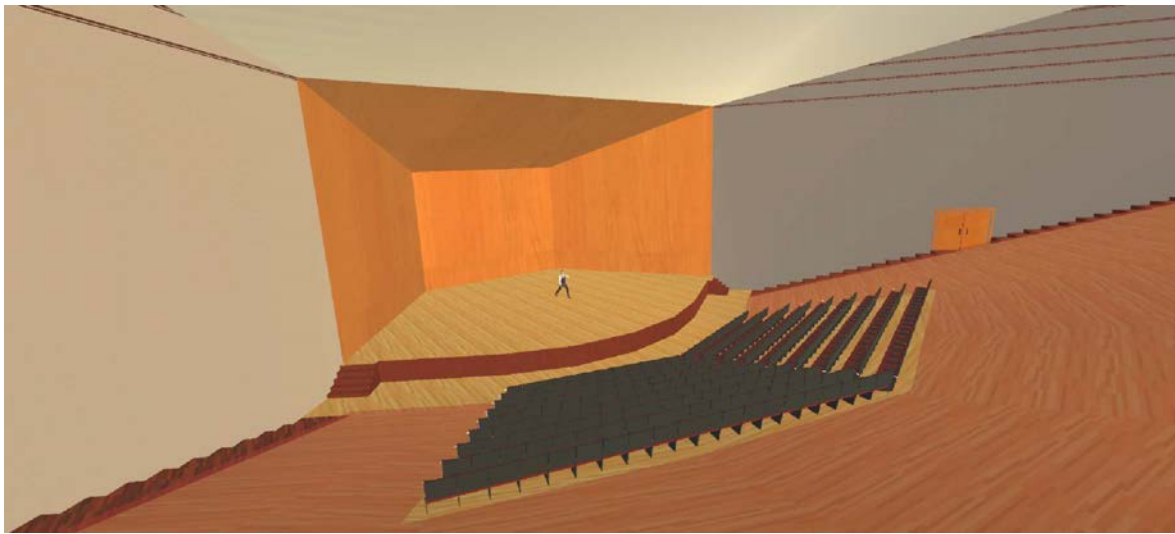
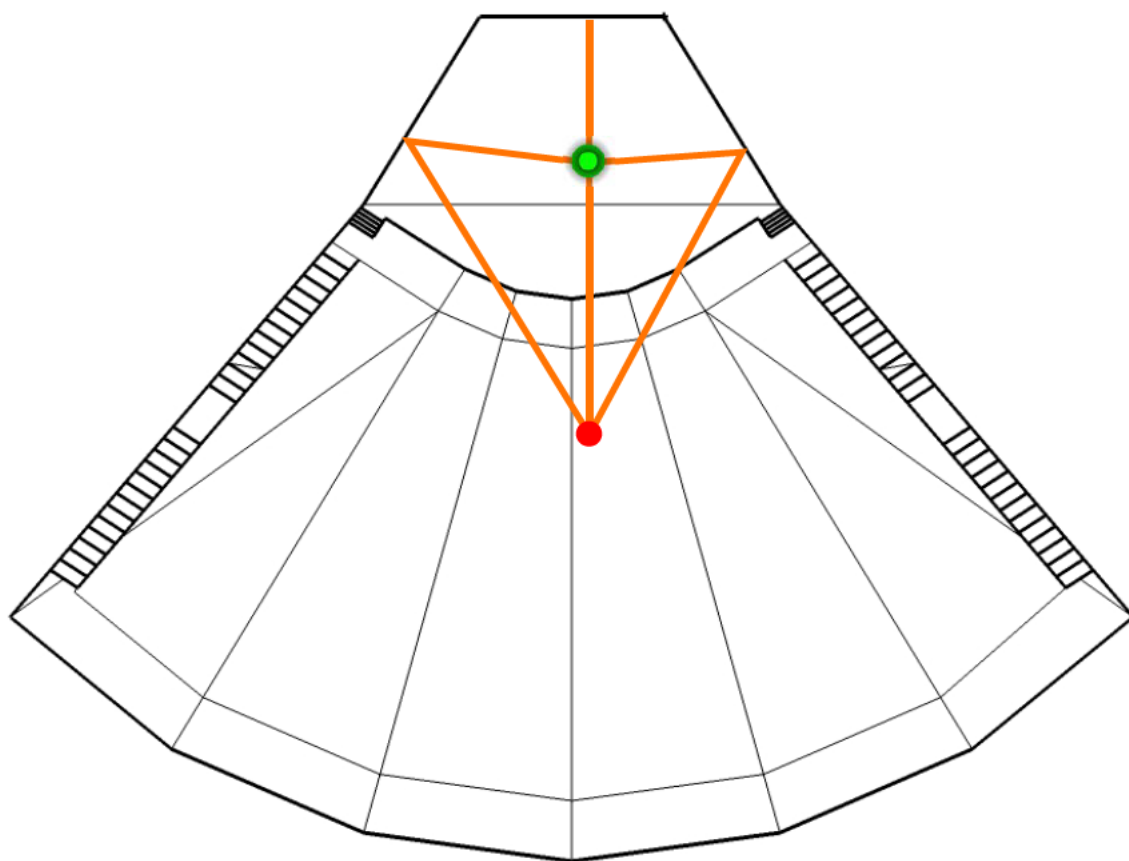
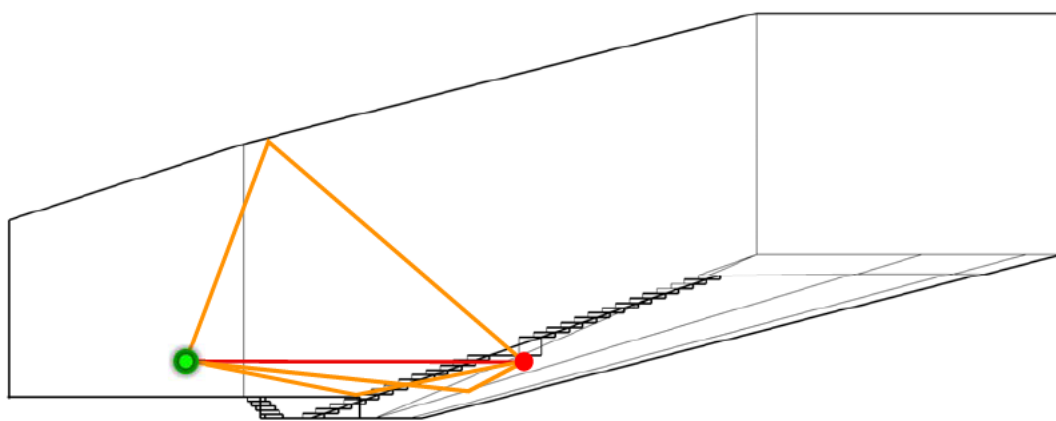


Fig. 16: Wide shot of the visible in-game model of the Fan hall.

The model constructed here was based on three prominent concert halls with a fan shape. The *Kleinans Music Hall* in Buffalo, NY is most closely related to the model used in this trial. The trapezoid shape of the stage and the low ceiling, rising almost parallel to the tribunes, are strong acoustic factors. The *Festspielhaus* in Freiburg, Germany has a higher ceiling and a wider stage but carries the elementary architectural features described above. The *Aula Magna* in Caracas also has a strong trapezoid shape but the ceiling is covered with reflector panels to generate a longer RT and guide the ceiling reflections down onto the audience. The floor plan and cross section of the model, with the reflections of 1st order are shown in figure 17.



(a) Groundplan Fan



(b) Cross section Fan

Fig. 17: Groundplan and cross section perspective of the **Fan** hall. The sound source position marked as a dot green and the listener position marked red. The early reflections are represented by the orange lines.

2.2.4 Horseshoe

The distinct horseshoe shape is most common for theaters and operas of the 19th century. The model, built for this project, is designed using the *Semperoper* in Dresden, Germany, the *Teatro Alla Scala* in Milan, Italy, and the *Teatro di San Carlo* in Naples, Italy as reference. The prominent architectural feature is the oval shape of the wall is surrounding the tribune up to the edges of the stage. The wall is densely populated with overhanging balconies that reach up until close below the ceiling. The main Floor tribune is rising in a small angle towards the back and the stage is usually of a box shape. The wall, surrounding the main tribune is usually interrupted with scuncheons in a similar fashion as in the shoe box halls, to counter the focusing effect of the curvature. The proportions of the oval and the height were taken from the mean of the above described halls. Most Horseshoe halls also contain an orchestra pit in front of the stage. It was left out, since the effect on the early reflections should be negligible.

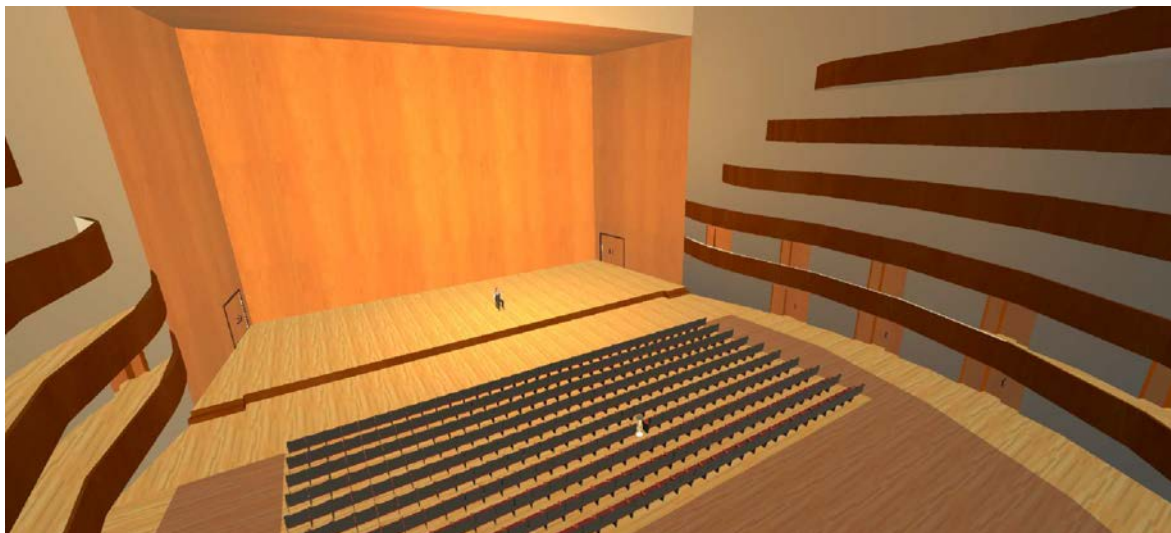
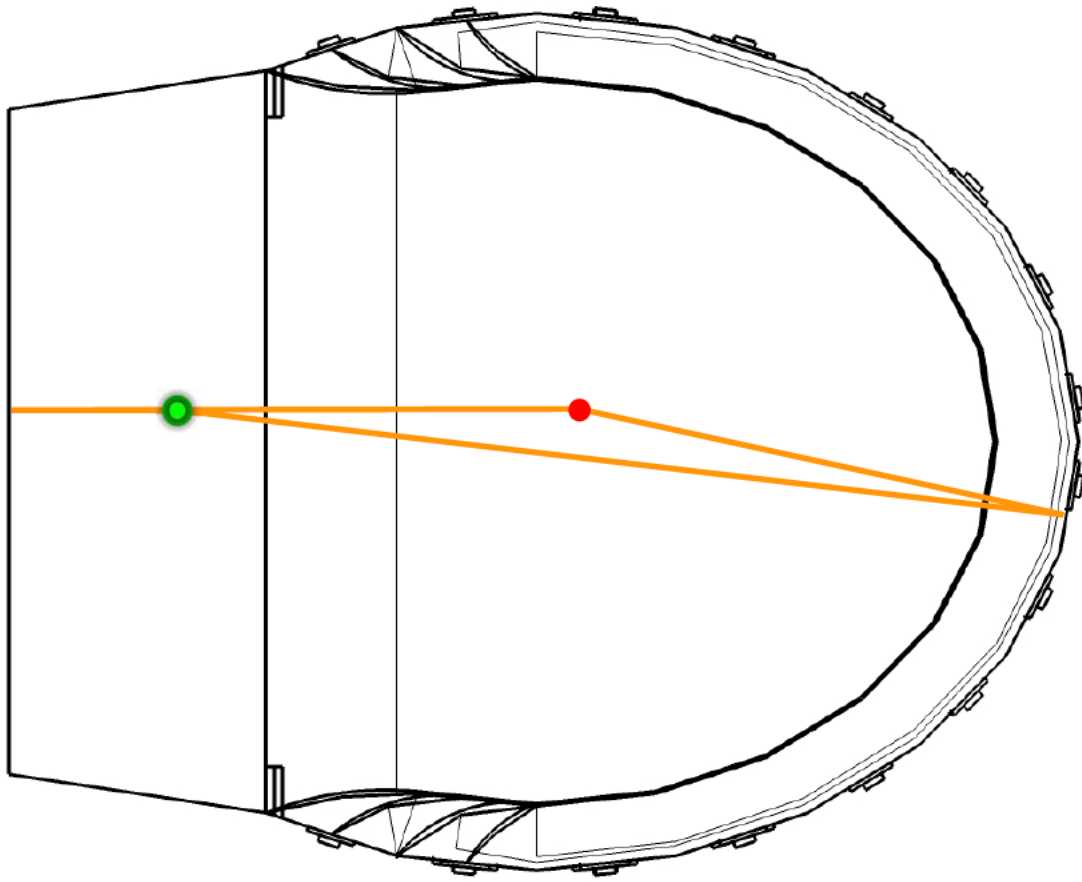
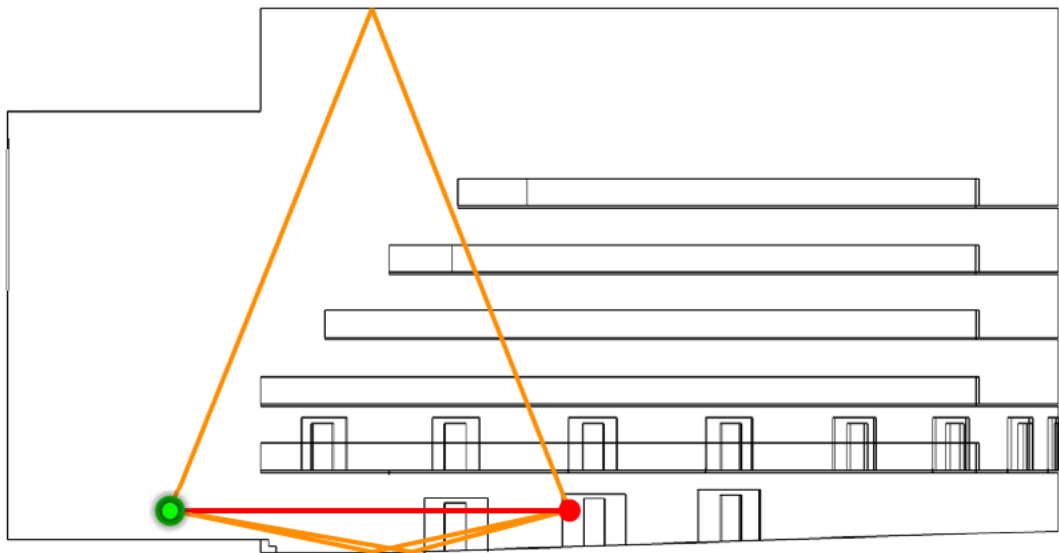


Fig. 18: Wide shot of the visible in-game model of the Horseshoe hall.

The oval back wall of this particular shape has a specific acoustic effect, that needs to be accounted for. Reflections get focused onto a point in the back center of the audience. In an attempt to avoid this undesired strong acoustic cue, scuncheons were placed on the side and back of the curved wall. Additionally, the listening position was set to be outside of the focus point.



(a) Groundplan Horseshoe



(b) Cross section Horseshoe

Fig. 19: Groundplan and cross section perspective of the **Horseshoe** hall with ceiling and floor reflections of 1st order. The sound source position marked as a green dot and the listener position as a red dot. The early reflections are represented by the orange lines.

2.2.5 Model design in the virtual environment

The computational demand for VR games is relatively high in order to run smoothly, since all motion rendering has to be done twice, once for each screen of the stereoscopic display. Because of the screens being very close to the eyes, the frame rate should be around 90 fps to avoid discomfort. The frame rate does automatically adjust to avoid lagging. With very detailed graphical design and lighting, the computational demand is increased significantly, which can lead to lower frame rates. In order to reach an immersion effect, the VR environment has to entail a certain degree of detail. The models created in Sketch Up and used for the acoustic simulation are relatively simple and hard to interpret without visual anchors. In an attempt to show the exact halls for which the auralizations were calculated and still present a realistic visual impression, some visual details were introduced after the models were imported into UNITY, such as high definition textures on the surfaces of the models. Doors and decorative moulding were placed on sidewalls to ease comprehensibility of the architectural features. The seat rows on the tribunes take up the main floor space of the concert halls. Around the listener position, rows of theater seats were modeled, so that the test subject would get the impression of sitting in the audience of the virtual halls. To cover all of the seated areas with visible seat rows, showed to decrease the frame rate and result in serious discomfort, when wearing the VR headset. Therefore, seat rows were only placed in the area around the virtual seat of the participant. All other seated areas were layered with a half transparent red surface to mark the seated areas, which are acoustically relevant. Since the impression of size is an important part of this test scenario, the additional visual features described, were similar in size in all models. These elements could therefore be used as visual references for the participants to judge the distance to the sound source and walls, and support the overall size impression. For the sound source, a model of a male person was placed on the stage as the “speaker”. The head of the speaker figure was set at the sound source position facing the participant.

2.3 Room acoustic simulation

The room impulse responses (RIR) of the above described concert hall models were simulated using a hybrid method described in section 1.4. The environmental conditions, shown in table 3, were defined. These conditions influence the speed of sound and the air absorption and are set to standard settings.

Tab. 3: Environmental conditions used in acoustic Simulations.

Environmental Condition	Value
Room Humidity	50 %
Temperature	20 °C
Air Pressure	101,3 kPa

The simulation preferences, listed in table 4, are parameters influencing the resolution and accuracy of the simulated BRIRs. The number of used particles has to be sufficient to be able to capture the full spectrum of acoustic characteristics. It depends on the volume and architectural complexity of the model. The number of particles used here, has shown to result in a sufficient density of the energy spectrum at the receiver position of the big models guaranteeing to be sufficient for the small models as well. The Image Source model was set to calculate up to 3rd order reflections. With a sufficient number of particles higher order reflections are likely to be calculated accurately with the ray tracing method of the hybrid simulation tool. The detection sphere radius and cutoff threshold were set to the recommended standard settings and the filter length was set to be at least double of the reverberation time.

Tab. 4: Simulation preferences

Simulation Condition	Value
Number of Particles	400000
Image Source Order	3
Energy Cutoff Threshold	60 dB
Detector Sphere Radius	0,5 m
Filter Length	6 s

2.3.1 Sound source and receiver

As the sound source, it was decided to use a speech signal further described in section 2.1.1. To match this signal type in the simulation, a measured directivity pattern of the vocal tract of a singer was applied to the sound source. The directivity pattern was taken from the directivity database provided by the RAVEN toolbox and measured by a team at TU-Berlin (Weinzierl et al., 2017). This pattern determines the energy content of the particles dispersed from the sound source sphere in the ray tracing simulation (see section 1.4). The sound source direction was static, and facing the receiver in the same height and a defined distance. An open source database of high resolution HRTFs

measured with the FABIAN dummy head were used in the simulation to represent the receivers head (Brinkmann et al., 2013). The receivers view direction in relation to the sound source position is described by a polar coordinate system with azimuth ϕ and elevation angles θ . The central point of the system is located between the ear canals of the head and the frontal head orientation, facing the sound source is represented by the $\phi = \theta = 0^\circ$ angle. A sketch of this system is shown in figure 20. The FABIAN database provides HRTFs for source positions with elevation angles from -64° to 90° and full range of azimuth angles with a resolution of 1° . The polar coordinate system shown here will be used for further descriptions.

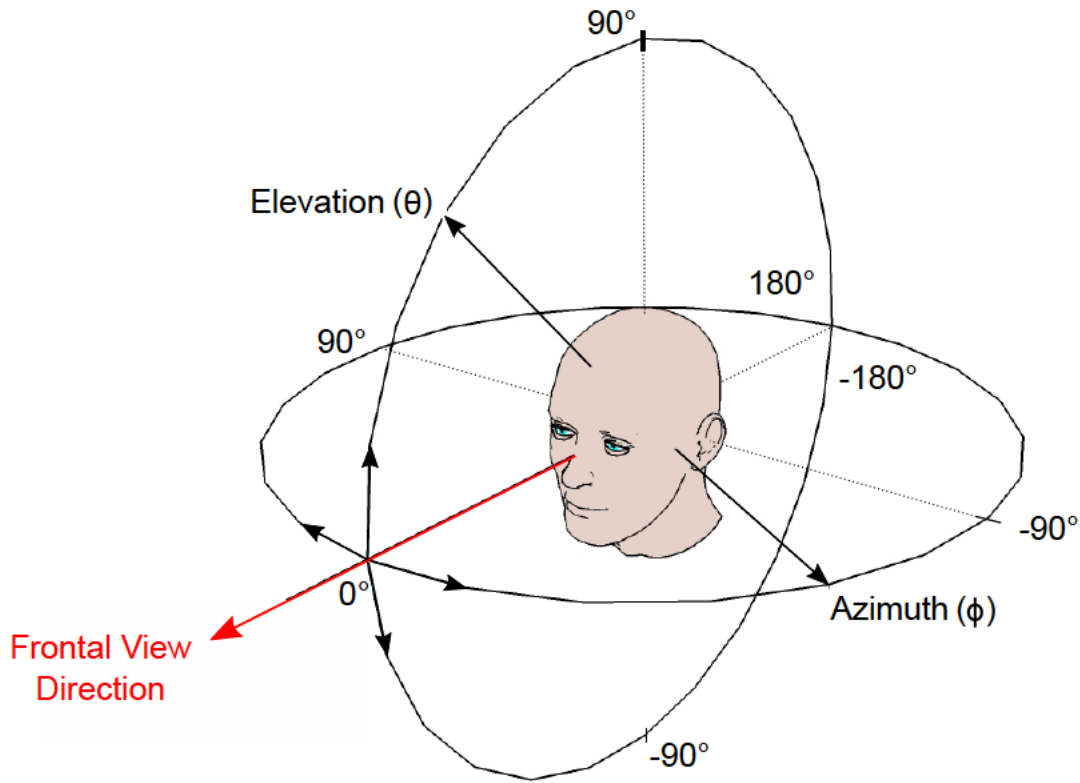


Fig. 20: Sketch of the coordinate system of the receiver head with azimuth and elevation angles described.

2.3.2 Source and listener position

The acoustic conditions: reverberation time and wall scattering were predefined by the test design. The acoustic spatial impression is also conditioned by the source (SP) and listening position (LP). A problem occurs with the placement of the SP and LP when comparing differently shaped concert halls. Due to the height of the stage and shape of the tribune, the height of the SP and LP in a comparable seating position differs naturally. To eliminate potential movement of the sound source, when comparing the acoustic stimuli, the relationship between the SP and LP for this test has to be identical in all halls. It was decided to use one sound source, placed in a

central position of the stage and the LP being placed at a distance of $d_{\text{small}} = 10$ m for the small and $d_{\text{large}} = 15$ m for the big halls on the longitudinal x-axis of the hall. These distances guarantee the listening position to be in the diffuse sound field which begins at a distance of the reverberation radius r_H . This distance is dependent on the volume V , the speed of sound c_S and the Reverberation Time T_{60} (see equation 3). The source and listener position was moved 1 m towards the right of the central room axis (y-axis) to counter potential interference effects in the center of the symmetric halls. Using the example of a shoe box model with $V = 20\,000\text{ m}^3$, the above described placements are illustrated in figure 21.

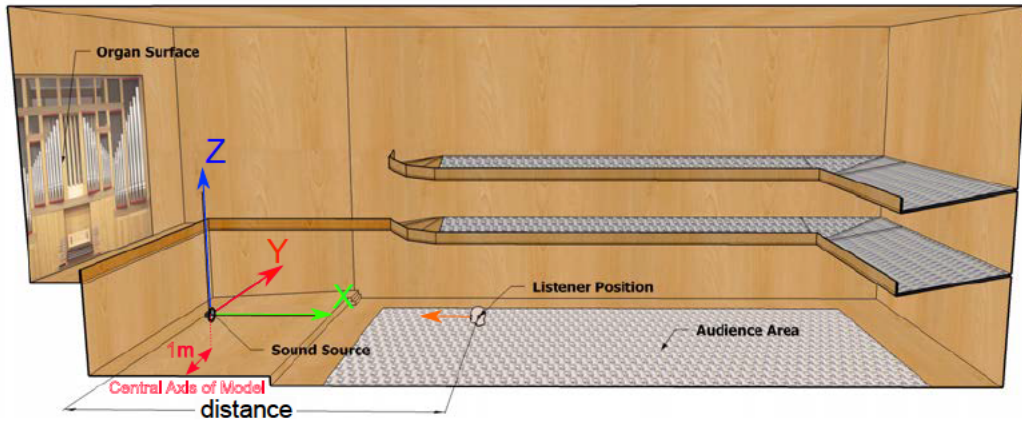


Fig. 21: Shoe box hall with placement of SP and LP and description of acoustically relevant simulation surfaces.

Due to the architecture of the halls, the SP and LP had to be adjusted in height and the position on the longitudinal axis of the models individually to ensure a reasonable positioning in the architectural context. In the Vineyard and Fan hall, the tribune is rising towards the back wall. This allows the SP to be placed at a height of 1,7m above the stage floor, relating well to the height of a standing person, while the LP can be placed at a height of 1,2m above the floor of the audience area. This way the LP is at the right height to correspond with the height of a sitting listeners head and the test participants can be placed as though they were sitting in a seat row of the main tribune. In the Shoe Box the main floor is flat and in the Horseshoe hall only slightly rising, while the stage is elevated. This would lead to the LP being high above the natural head position of an audience member, when the speaker is standing. In the architectural context, as well as the virtual environment, this effect would be undesirable. To strike a balance, it was decided to lower the SP in those two halls to 1,2m corresponding to a seated speaker on stage, while the LP is at the height of 1,7m, corresponding to a person standing on the audience area floor. Since participants were sitting on a revolving chair, a visual model of a bar stool was modeled to prevent the participants from floating in the air.

$$r_h = \sqrt{\frac{V}{c_s T_{60}}} \quad (3)$$

2.3.3 Acoustic properties of the audience area

In concert halls, the main absorption area limiting the RT is provided by cushions on the seats and the audience. Most concert halls therefore have upholstered seats, specifically chosen for their acoustic properties. To mimic this effect, a audience area with static acoustic properties was placed in each concert hall model. The area was of equal size for each group of similar volume, but shaped to covert the tribunes of the individual shape types. The fan shaped hall offers the smallest possible area size, due to the shape. This size was then used for the seated areas in all other halls. For the big halls an audience area size of $1150 \text{ m}^2 \pm 20 \text{ m}^2$ and for the small halls $460 \text{ m}^2 \pm 10 \text{ m}^2$ was therefore chosen. The number of occupied seats also changes the absorption of the seated area and a certain variability is usually accounted for. For the models, described here, a seated area with the acoustic conditions shown in figure 22 was used. We show the absorption and scattering coefficient over frequency. These acoustic properties were taken from the standard material dataset of the RAVEN toolbox and resemble a semi-occupied upholstered seated area.

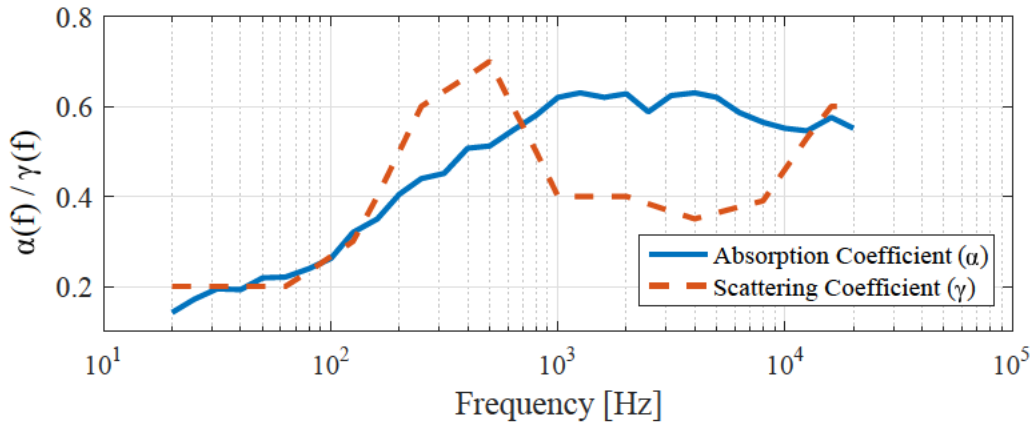


Fig. 22: Acoustic properties of the seated audience area used in all concert hall models for the acoustic simulation. Absorption and scattering coefficient shown over frequency.

2.3.4 Acoustic properties of the organ

In two of the four models, the Shoe Box and Vineyard model, an organ was placed behind the stage. The organ is especially relevant for the Shoe Box hall, since the high scattering of an organ counters eventual flutter-echos and acoustic interference between the back wall and the stage wall. In vineyard halls, the position of the organ varies but is commonly placed between the back of a tribune and the beginning of the ceiling. The acoustic properties were set to be similar for both halls and the organ

area size similar for both groups of volume. The acoustic properties of the organ area were taken from the standard material dataset of the RAVEN toolbox and are shown in figure 23. We show the absorption and scattering coefficient over log frequency. The organ size was set to be 150 m^2 for the big halls and 60 m^2 for the small concert halls.

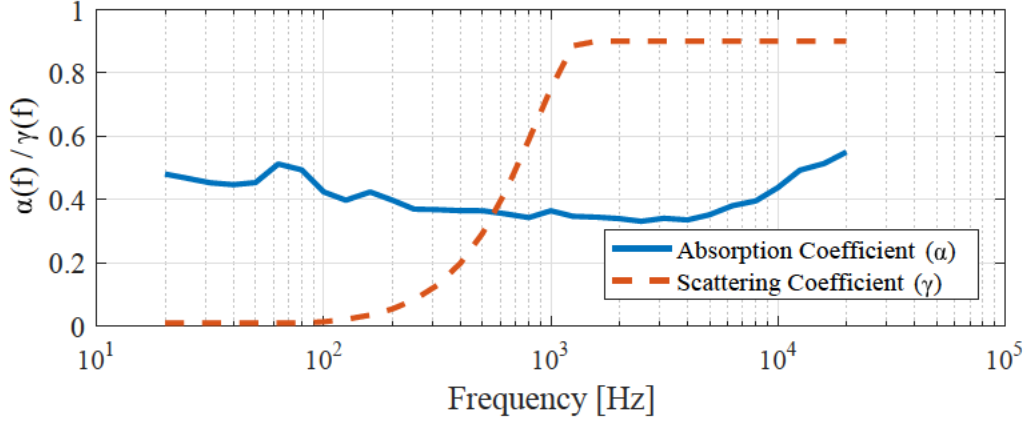


Fig. 23: Acoustic properties the of organ used in the Shoe Box and Vineyard hall. Absorption and scattering coefficient shown over frequency.

2.3.5 Test condition: reverberation time

With all predefined acoustic conditions described above, the RT can only be changed by adjusting the absorption coefficients $\alpha(f)$ of the remaining surfaces. This was done in an iterative process, by fitting the RT of each individual hall to a set RT curve by adjusting the absorption coefficients of the main wall material. A function from the ITA-toolbox was used for this task (Berzborn et al., 2017). The RTs are calculated from the RIR measurements, produced by a ray tracing simulation with a spherical sound source with omnidirectional directivity. The source and receiver position for these measurements were identical with the ones, used for the auralizations and as described in section 2.3.2. All surfaces apart from the audience and organ area were set to be the same material. The $RT(f)$, used for this listening test, were taken from simulated measurements generated in a study by Weinzierl et al. (2018). The RTs were shifted to be $T60_{500 \text{ Hz}} = 1,5 \text{ s}$ and $2,5 \text{ s}$. In figure 24, the target RTs are shown over log frequency.

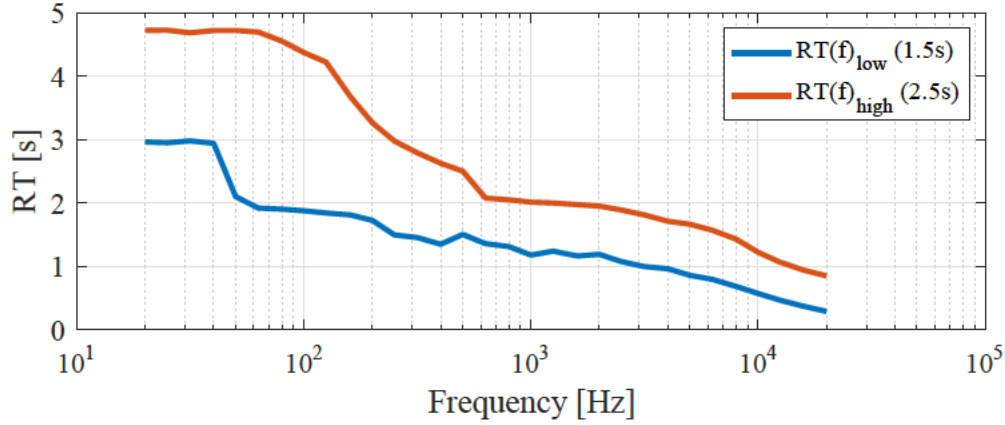


Fig. 24: Reverb times (T_{60}) defined for the test conditions RT_{low} and RT_{high} over frequency.

2.3.6 Test condition: scattering coefficients

According to the test design, two different settings for the scattering behavior were set for the wall materials. The sound scattering of a diffusing surface is dependent on the structure and depth of the surface material, as well as the angle of arriving sound energy. The Scattering Coefficient $\gamma(f)$ defines the amount of acoustic energy reflected in a non-specular compared to the specular reflected energy. In the simulation software used, scattering directivity is not accounted for. Therefore the scattering was estimated with a theoretical approach for random surfaces as proposed by Embrechts et al. (2001). With this approach, shown in equation 4, the angle of incidence θ_i , the r.m.s. height s of a random rough surface and the wave number k is given as input.

$$\gamma(f) = 1 - e^{-4s^2k^2 \cos^2 \theta_i} \quad (4)$$

The angle of incidence was set to $\theta_i = 45^\circ$ and the r.m.s. height was set to $s = 1$ cm and 5 cm for the two acoustic conditions, resulting in the scattering coefficients shown in figure 25. The stronger scattering setting rises in a steep slope between 100 Hz – 1 kHz whereas the lower scattering setting starts rising at 1 kHz up to 4 kHz.

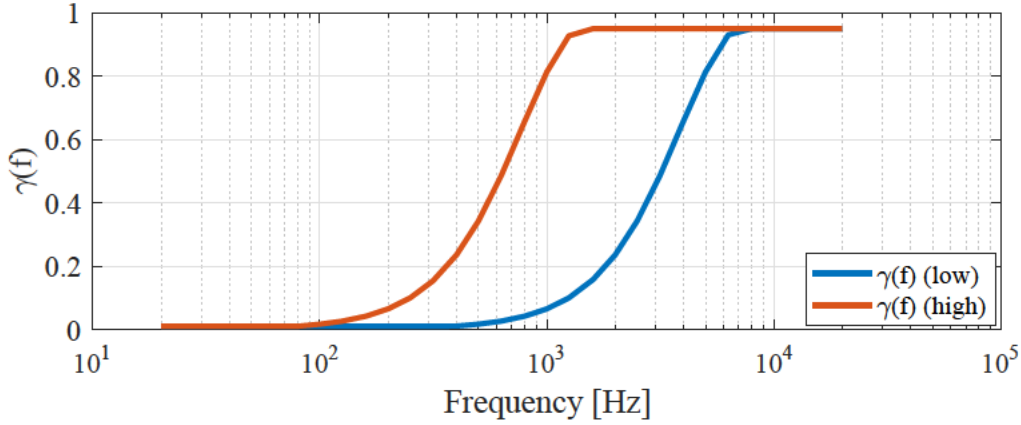


Fig. 25: Scattering coefficients defined for test conditions SC_{low} and SC_{high} over log frequency.

2.4 Dynamic auralization

Dynamic auralizations were generated by calculating Binaural Room Impulse Responses (BRIR) for the different concert halls and using the *SoundScape Renderer* (SSR) as the software framework for real-time non-individualized binaural room synthesis for headphone reproduction (Ahrens et al., 2008). For this method, the early part of BRIRs were to be calculated for each concert hall model with head orientations in a sufficient resolution. The diffuse reverberation tail is represented by a single pair of BRIRs for each auralization. The SSR enables the auralization of the calculated impulse responses with an arbitrary anechoic input signal. All spatial information of the reverberation and the directional head and torso related impulse responses of the FABIAN dummy head are contained in the BRIRs. The SSR generates two ear signals for the the virtual acoustic scene, through convolution of the input signal with the pair of BRIRs relating to the current head orientation. It provides a convolution engine that crossfades between the BRIRs, depending on the head orientation of the listener. The head orientation is measured with an external head-tracker. This real-time adaptation of the position of the virtual acoustic environment to the head orientation generates the impression of a stable acoustic environment. Head movement also increases the localization accuracy as shown by Minnaar et al. (2001) and enhances the externalization of speech stimuli in reverberant environments as shown by Hendrickx et al. (2017).

The resolution of the BRIR datasets has to be chosen in consideration of the localization accuracy of the human hearing system to prevent the experience of the sound source jumping around the supposed virtual source position. The human auditory localization accuracy is highly dependent on the location of the sound source in comparison to the head orientation as well as the temporal and spectral properties of the sound signal and the acoustic conditions of the surrounding space. In the horizontal plane, the localization accuracy is relatively high due to the information derived from the

Intraural Time and Level Difference (ITD, ILD) (Blauert, 1997). Also, localization of sound sources in front of the subject is much more accurate than for sources behind the listeners head. The localization blur in the frontal lateral plane was found to be as low as 1° (Letowski and Letowski, 2011). Sound sources above or below the horizontal plane are localized mainly through monaural cues, produced by the sound waves being diffracted around the listeners torso, head and pinna. This spectral information is less pronounced than the ILD and ITD and more dependent on the sound signal. The localization blur in the median plane was found to be around 4° (white noise) and going up to 17° (continuous speech of an unknown person) Blauert (1997). Reverberation has shown to worsen the localization accuracy (Kopčo and Shinn-Cunningham, 2002). For the described listening test, the resolution of the BRIRs was set to be of $\phi = 1^\circ$ in the azimuth and $\theta = 10^\circ$ in the elevation angle of the head orientation. The maximum elevation angles for which BRIRs were calculated is $\theta = -30^\circ$ below, to $\theta = 60^\circ$ above the horizontal plane. Head movement, outside of these boundaries, would require uncomfortable and unnatural tilt of the neck and are therefore not expected. Pretests have shown no effect of a source shift induced by head movements. The sufficient latency of the tracking system was proven by pretests with white noise. No switching effects or movement of the sound source could be detected, guaranteeing that audible differences would only depend on the content of the BRIRs.

A dataset of BRIRs were calculated for each concert hall in the resolution grid of the defined head orientations, using the VSE toolbox, built by Böhm (2015). The result is a set of 3600 BRIRs packed in a SOFA file format including the azimuth and elevation angles for each BRIR. The early and late part of the BRIR is split at the mixing time calculated for each room and condition after Lindau et al. (2012). The late reverberation is saved in a single BRIR since it can be expected to appear as a diffuse sound field arriving with the same energy from all directions. The two parts of the BRIR are simultaneously convoluted in the frequency range with the input signal by the SSR, creating the binaural auralization. The 32 datasets of BRIRs were loaded into the SSR system and are played as described in section 2.1.4. A second partition of the SSR was generated to apply headphone compensation for the *Bayerdynamic DT770 Pro* headphones. This technical application of the auralization method raises to the standard found by Lindau and Weinzierl (2011), to provide plausible state-of-the-art room acoustic simulations

2.5 Statistical analysis

Descriptive statistics

In a first step, the count of correct choices of each participant was summed and the arithmetic average, confidence intervals ($p < 0.05$) as well as the inter quantile range calculated for each shape type and both trained and untrained participants. All averaged results are compared with the guessing probability (GP) of 25 %. The wrong but shape specific choices are summed and mean hit rates calculated in a similar fashion. This leads to a confusion matrix that shows, which of the concert halls were often mistaken for one another. Additionally the histogram of the hit rates is shown to analyze the distribution of the response data over the participants. Additionally, the sum of correct selections for each participant and shape type are analyzed. One-way-ANOVA tests were done to test significance. In the particular test setup, with different acoustic test conditions, this contemplation can not reveal the effects of the test conditions and which acoustic features were useful to identify the concert hall shapes. Therefore a linear regression model was calculated as described below.

Mixed model analysis

In an attempt to investigate the influence of the test conditions on the response and to gain information on why correct choices were made, regression models are calculated on basis of the gathered response data. The chosen models are Generalized Linear Mixed Models (GLMM) with a probit link function and maximum likelihood Laplace approximation. GLMMs are extensions of linear mixed models that include fixed and random effects but also allow an individual link functions to correspond to the specific type of target variable. In this case the response variable is binary (correct [1], incorrect [0]), which is why the probit function was used. The models are calculated with *R-Statistics*⁶ and can account for fixed, interaction and random effects for complex data sets with multiple linear predictor variables. The effect parameters are fitted with a maximum likelihood approximation since the underlying integral over the random effect space can not be solved analytically. For each calculated effect a coefficient is calculated which describes the direction and strength of the influence of the predictor on the predicted outcome. Fixed and interaction effects describe the direct influences of the chosen predictors on the predicted outcome, while random effects account for different intercepts between participants. Participants might use the same or different acoustic cues to make their decision but have different sensibility or cognitive weighting on the multiple acoustic cues present. Interaction effects account for the reciprocal influence of multiple predictors and help to understand multilevel dependencies. For

⁶R Statistic 3.6.2

example, the volume might influence the effect of the reverberation time. A small concert hall with long reverberation develops a loud strong reverberant characteristic, while the same reverberation time in a large hall does not necessarily produce the same loudness of the reverberation. A interaction effect can reveal this reciprocal influence and reach a higher accuracy in the prediction. Two GLMMs are calculated in this work. The first uses the test conditions RT, SC and Volume as well as the shape type and training as the predictor variables. The condition parameters are dummy coded to be represented by binary values [0,1] and are therefore categorical predictor variables. The exact representation of the dummy coding is shown in table 5. This model relates to the test conditions shaping the appearance of the acoustic stimuli. In the second model, the $[1 - \text{IACC}_{E3}]$ as a measure for the ASW and $[1 - \text{IACC}_{L3}]$ as a measure for the LEV (see Sec. 1.1) are used as the predictor variables. These measures are calculated from the BRIRs after ISO 3382-1:2009 and for the head orientation in the horizontal plane $\theta = \phi = 0^\circ$. The model therefore relates to the signals received at the ear canals and perceptual cues which are not directly related to the test conditions. This model in specific is generated to analyze, if the perceptual acoustic cues known to influence the spaciousness and are used to explain the perceived acoustic spatial impression.

Tab. 5: Dummy coding for test conditions

Variable	0	1
Volume	small	big
Reverberation Time	low	high
Wall Scattering	low	high
Training	no	yes

Both models were approached as proposed by Peugh (2010). First the models are calculated with all main effects and all possible interaction effects with and without random effects. If the model with random effects enhances the models predictive capability significantly, the random effect is present and needs to be included. This was tested with a ANOVA and was found to be significant ($p < 0.05$) in both models. The random effect was calculated over the “subject ID” to account for random intercept between individual subject responses. The models with random intercept are then reduced by insignificant interaction effects to generate a more effective model. The enhancement of the reduced model can be shown by a reduction of the Akaike information criterion (AIC). The goodness of fit of the reduced models is calculated with the approach proposed by Nakagawa et al. (2017) which estimates pseudo R^2 values for mixed models. These describe the amount of variance in the data, which can be explained by the models. R^2_{marginal} which describes the variance explained without random effects and $R^2_{\text{conditional}}$, which describes the explained variance including the random effect were

calculated for both models.

To analyze the different effects, the model predictions are calculated for different fixed values of the predictor variables. The predicted mean help to interpret the direction and strength of the different effects. It has to be noted, that the calculated intervals of the prediction can not describe the 95 % confidence intervals because the method uses the mean of the random effect variance to account for the random intercept. Therefore the returned intervals are much larger and describe prediction intervals and not confidence intervals. The significance of the calculated effects is therefor not corrupted by overlapping of the prediction intervals. A higher number of participants yielding more observations would reduce the span of the intervals.

3 Results

36 subjects participated in the listening test. 9 female and 27 male participants with 23 to 53 years of age, were split into two subject groups of 18 participants. One group received a training as described in section 2.1.2. Of the 36 participants with a mean age of 33, 6 reported a slight hearing impairment. Most participants reported some or extensive experience with listening tests and good or expert knowledge in room acoustics. Though only very few participants reported to visit concert halls more than once in three months. Three participants were practicing acousticians currently working in the field of room acoustic planning. The surrounding test environment had a quiescent sound pressure level of 46 dB_{SPL} averaged over 1 min. The participants spent round 35 to 45 min in the VR environment, with the main test taking about 30 min in total. Four participants had to pause the test due to discomfort, but could complete it after a short break.

3.1 Mean correct response count of untrained participants

The response of the listening test is a binary value since the selections can either be correct (1) or incorrect (0). In each iteration of the test, four choices were given, which yields a guessing probability (GP) of 25 %. Therefore, the overall rate of correct responses has to be evaluated in comparison to the guessing probability. In figure 26 the mean of correct response rate as well as the 95 % confidence intervals of all untrained test participants are shown.

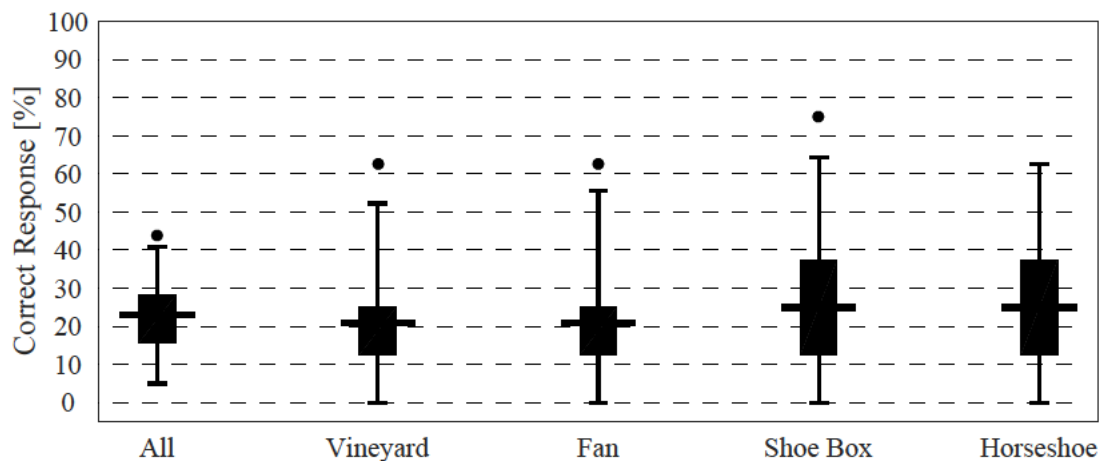


Fig. 26: Rate of correct responses of untrained test participants. Boxes show means and upper and lower inter quantile range, whiskers show the 95 % confidence intervals. Dots mark individual participants reaching a hit rate above the confidence interval.

Untrained test participants could in mean only identify 22,9 % of the concert halls correctly. One participant reached a correct response rate above the upper confidence

interval with 43,7 % of the halls guessed correctly. The mean hit rate of the Vineyard and Fan hall, lie around 20 %. The mean hit rate of the Shoe Box and Horseshoe hall correspond exactly with the guessing probability of 25 %. Outliers marked as dots above the whiskers, show that some participants where able to guess 60 – 75 % of the Vineyard, Fan and Shoe Box hall correctly. A more detailed analysis of individual response rates for the different concert halls shapes is shown in chapter 3.3.

3.2 Mean correct response count of trained participants

Trained participants performed slightly better in the given task, than untrained participants. The mean rate of correct responses as well as the 95 % confidence intervals are shown in figure 27. Trained participants were in mean able to guess 30 % which is a slight increase compared to the untrained participants of 7 %. Two test participants could identify 62,5 % and one 59,4 % of the concert halls correctly. Still, of trained participants, most response rates were within the range of the guessing probability. The Vineyard and Fan hall where identified with a mean rate of 30 % improving in comparison to untrained participants by 7,1 %. For the Shoe Box hall, no increase in comparison to untrained participants could be seen. 43 % of the Horseshoe halls could be identified by trained participants, showing the biggest increase compared to the untrained test subjects. One participant was able to identify all of the Vineyard halls in all acoustic conditions. One trained participant was only able to identify one stimulus correctly. With this outlier excluded from the dataset, the mean correct response rate rises to 31,3 %.

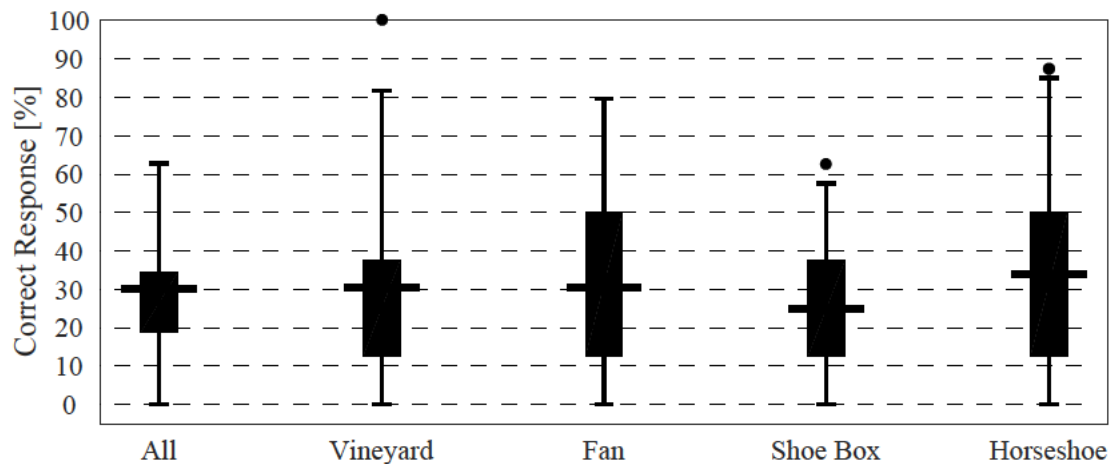


Fig. 27: Rate of correct responses of trained test participants. Boxes show means and show upper and lower inter quantile range, whiskers show the 95 % confidence intervals. Dots mark individual participants reaching a hit rate above the confidence interval.

Both the response rates of trained and untrained subjects lie in close range to the guessing probability. In fact the confidence intervals of summed results (All) and results summed for the individual concert hall shapes overlap with the guessing probability. This is a good indicator that the over all population of participants could not identify the concert hall shape by matching the acoustic spatial impression with the visual impression. Still, some untrained and some trained subjects were able to reach impressive response rates more than double the guessing probability for individual shapes, which is further analyzed in the next chapter. A T-test was not applicable, since variances were not similar. Therefore a One-Way ANOVA was calculated to analyze the significance of the training effect. The difference between the means of the hit rate distribution could only be shown for the rates, summed over all shapes (All) and a significance level of $p = 0.06$. In this specific 4-AFC test design, the choices were not only conditioned by the shape of the concert halls, but on the volume, RT and SC as well. Therefore, the full spectrum of effects are not yet captured with this approach. Random effects also have to be taken into account, since individual subjects might have different perceptual sensitivity towards spatial acoustic cues. Therefore, an examination of individual subject responses and a mixed effect model analysis was done as described below in Sec. 3.5.

3.3 Response count of individual participants

In figure 28 the summed correct response count of individual trained and untrained participants for the different concert hall shapes are shown. Each shape was set as the test target exactly eight times, once for each test condition. The data shows most values distributed around the guessing probability of 25 % (2 of 8 choices correct) or lower. Considering the upper confidence intervals of the mean response count (see Sec. 3.1, 3.2) a response rate above 55 % is treated as significant. This corresponds for participants guessing more than 4 of the 8 halls correctly. Of untrained participants, a total of 4 participants were able to identify at least one of the hall shapes with a higher hit rate. One was able to identify the Vineyard and Fan hall 5 times correct. Two participants were able to identify the Shoe Box hall 5 and 6 times correctly and one participant was able to identify the Horseshoe hall 5 of the 8 times. Of trained participants a total of 7 participants were able to identify at least one of the hall shapes 5 times or more. One participant was able to identify all Vineyard halls correctly and 7 of the eight Horseshoe halls but could not confidently identify the Fan and Shoe Box hall. Two participants performed well and were able to identify 4 to 7 of all halls correctly and reached a overall hit rate of 62,5 %. Two participants were able to identify the Fan hall 6 times and one participant 7 times, but could not identify the other halls with confidence. One was able to identify only the Vineyard hall 5 times.

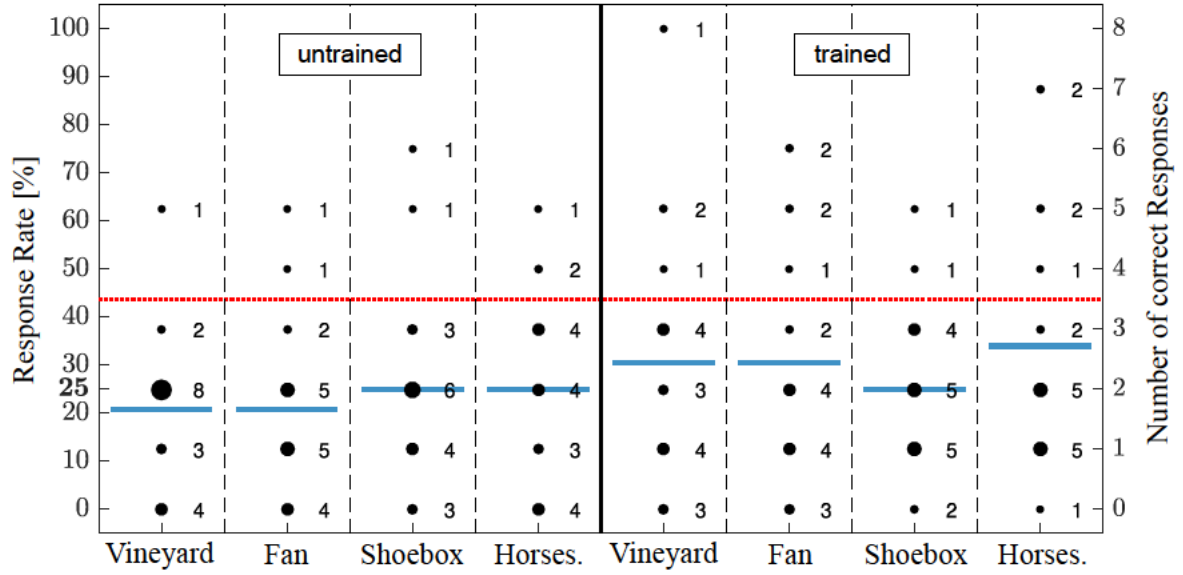


Fig. 28: Sum of correct responses by individual participants for each shape type. The dot size and the number to the right of every dot mark the number of participants, who made correct selections with the same rate. The red line marks the 43,75 % significance threshold and the blue lines mark the mean response rate for each shape type.

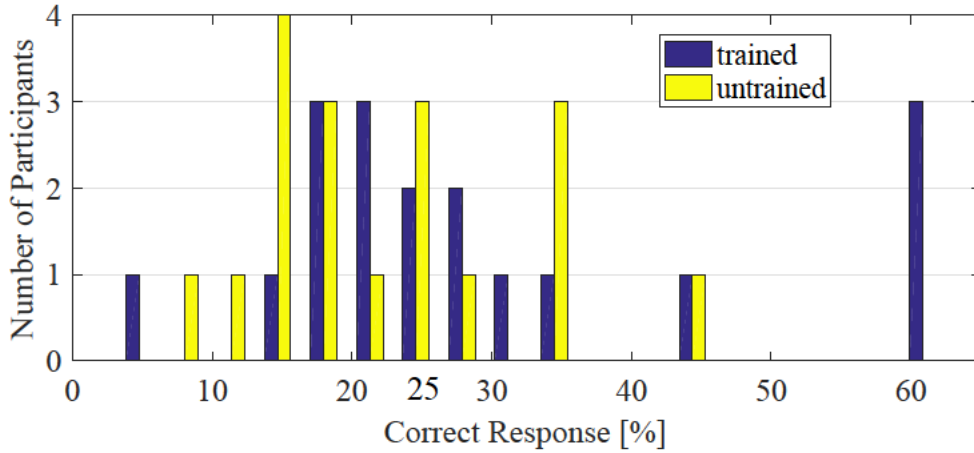


Fig. 29: Total correct response count in % of individual untrained and trained participants.

The data shows that some participants were able to identify one or more halls with a high probability, while the mean response rates show no significant difference from the GP. This suggest, that a small number of participants were able to identify certain acoustic cues produced by specific hall shapes with some accuracy intuitively or were able to memorize them after the training session. The majority of trained and untrained participants were not able to reach significantly higher hit rates than with GP. This can relate to hearing fatigue, so that some participants were only able to concentrate on some acoustic cues, while other cues remained undetected. As an additional way to analyze this data, a histogram of the mean response rates is shown in figure 29. From

this representation, it becomes evident, that only three trained participants were able to identify the halls correctly with a hit rate above 55 %. The majority of test subjects could only identify one or two halls well, while the others could not be identified well.

3.4 Mean response count of shape specific misinterpretations

The Confusion Matrix, shown in figure 30, displays the mean of the individual shape specific response count of untrained participants in %. The guessing probability for each selectable concert hall auralization is 25 % including the target hall and miss-matches have to be compared to the GP as well. Most of the response rates lie within close range of the guessing probability. One-Way ANOVA tests were calculated to test the significance of the difference between the hit rates. For the results of untrained participants, three significant observations could be made: The Vineyard hall as target was in mean 43 % of the times mistaken for the Horseshoe which yields a significant difference to the correct selection of the Vineyard hall. The Horseshoe was also significantly more often chosen than the Fan hall (12,5 %), when the Vineyard hall was set as the target. With the Fan hall as the target, the Horseshoe hall was chosen significantly more often (34,7 %) than the Vineyard hall, with only 12,5 %. These results suggest a strong difference in those Halls between the expected and perceived acoustic impression.

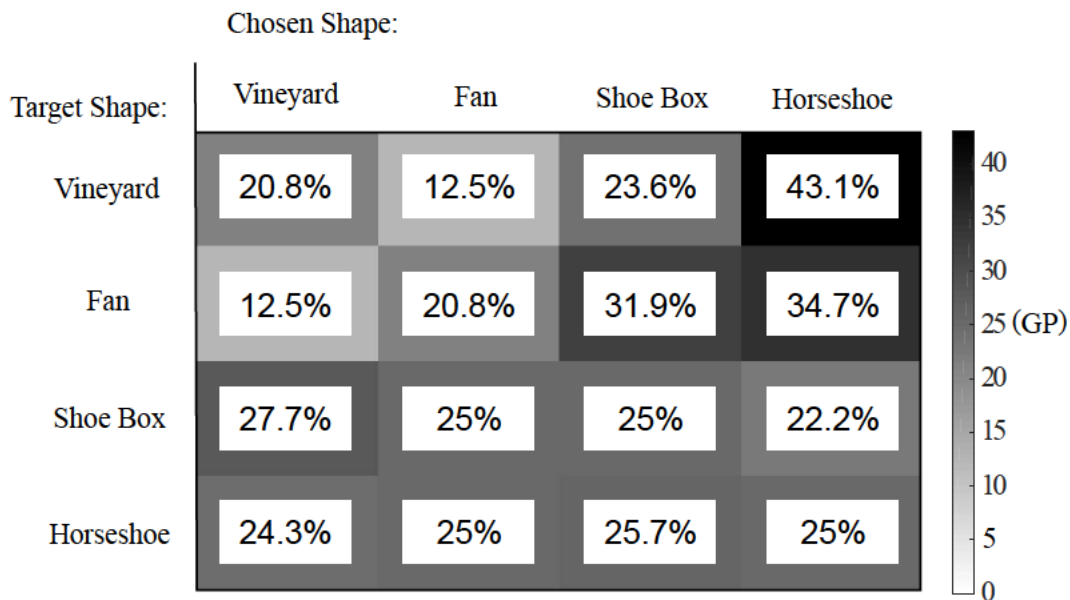


Fig. 30: Confusion matrix of mean response rates of untrained test participants.

The effect of the training, already observed earlier in this chapter can also be observed in the difference between the confusion matrices displayed here. The confusion matrix, for trained participants is shown in figure 31. In the same manner as described above, One-Way ANOVA tests were evaluated to test significant differences of the hit rate distributions. No significant difference could be found, even though, correct choices were

more likely. Comparing the two matrices reveals some interesting results. The misinterpretation of the Vineyard and Fan for the Horseshoe was not repeated by trained participants. Especially the Fan and Horseshoe seemed to produce individual acoustic cues, which could be learned by participants and make more accurate distinctions. In the contrary, the Shoe Box was often misinterpreted as the Fan hall by trained participants, while untrained participants misinterpreted the Fan for the Shoe Box more often.

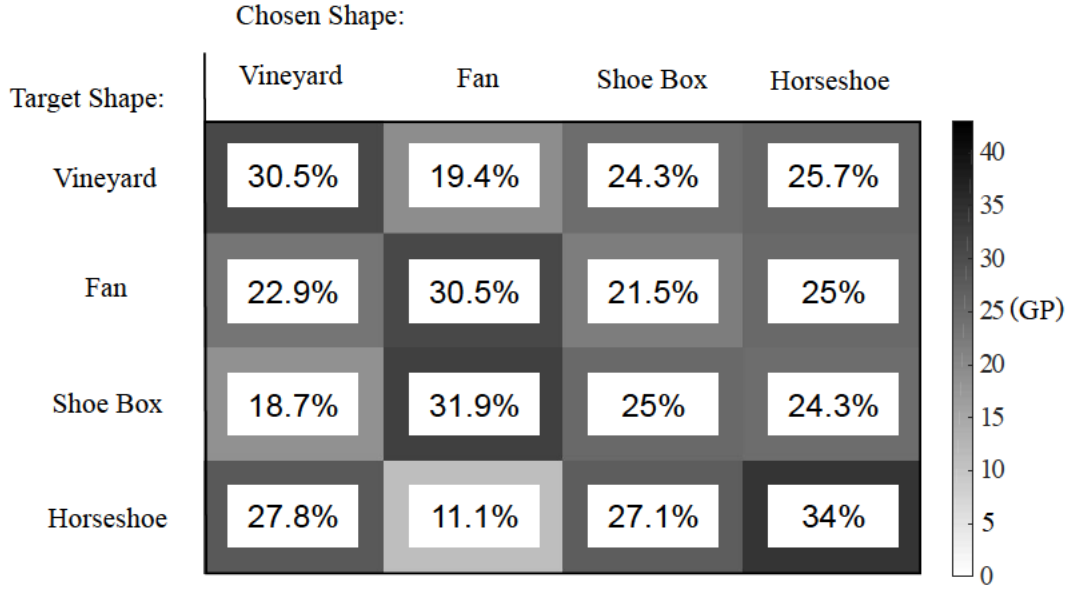


Fig. 31: Confusion matrix of mean response rates of trained test participants.

The results shown so far describe the response count for the different shape types but do not take interaction effects between the different test conditions and random effects between the individual subjects into account. These were further evaluated with a Regression model (see Sec. 3.5).

3.5 Regression analysis

3.5.1 Mixed effects model with test conditions as predictors

First a Generalized Linear Mixed Effects Model (GLMM) was calculated including all variables that conditioned the stimuli as predictors. These were the reverberation time (RT), wall scattering (SC), the volume (V) and the hall shape. Additionally the factor if the participant received a training was included. A significant increase of the model fit could be reached with a random effect for variances between subjects. All non significant interaction effects were excluded and the model recalculated. This yielded a decrease in the AIC value indicating an increase in the model quality. All significant ($p < 0.05$ and $p < 0.1$) interaction effects and main effects are shown in table 6. The other interaction effects could not reach a significance. A special aspect of this model needs to be mentioned. The shape types are included as nominal factors. In a mixed

model analysis, the model always calculates the response probability in relation to the default factor value. In this case, the Vineyard was set as the default, so all shape related results have to be interpreted with this effect in mind.

Tab. 6: Main effects and significant interaction effects, revealed by GLMM model. Significance ($p < .05$) is marked with * and ($p < 0.01$) with **. The coefficient describes the direction of the effect. The z-value describes the distance of the effect score from the mean, measured in standard deviation units.

Stimuli Model	Coefficient	Std. Error	z-value	p-value (sig)
RT	-0.511	0.258	-1.981	0.04 *
SC	0.104	0.269	0.386	0.699
Volume	0.280	0.284	0.988	0.32
Training	0.528	0.156	3.395	0.0007 **
Fan	-0.21	0.296	-0.724	0.47
Shoe Box	-0.05	0.29	-0.172	0.86
Horseshoe	-0.143	0.288	-0.5	0.62
SC \times Training	-0.512	0.167	-3.064	0.002 **
Volume \times Shoe Box	-0.70	0.413	-1.707	0.09
RT \times Shoe Box	0.715	0.35	2.041	0.041 *
RT \times Horseshoe	0.92	0.34	2.680	0.007 **
RT \times Volume \times Fan	-0.828	0.478	-1.730	0.08
RT \times Volume \times Horseshoe	-1.162	0.474	-2.449	0.01 *
SC \times Volume \times Shoe Box	0.884	0.479	1.844	0.06

Random Effect (Subjects)	Variance	Standard Error
Intercept	0.084	0.29

The variance explained by the model was examined with the R-squared values calculated to be: $R^2_{marginal} = 0.08$ and $R^2_{conditional} = 0.15$. The full model therefore only accounts for 15 % of the variance in the data. This is not surprising, since only very few participants were able to cope the difficult test task. The model likely explains the grounds for the decision of those few. The residuals are normally distributed, as shown by a Sharpio-Wilk test. The residual variance is similar in all test conditions confirming that the heterogeneity assumption is valid (see Appendix 6). A significant main effect could be shown for the condition of the reverberation time (RT) and the factor “Training”. The sign of the coefficient suggest that participants could better identify the hall shapes with the lower reverberation time and with training. The calculated training effect is highly significant but no interaction effect between training and a shape factor could be revealed. The training does therefore not help to identify specific concert hall shapes but increases the overall probability of making a correct choice. No significant main effect could be found for the wall scattering (SC) and volume condition as well as the individual shapes. This reveals that the volume and wall scattering did not directly influence the accuracy of the predictions and no hall could be identified especially well. From section 3.3 it became clear that some participants

were able to identify certain shapes with confidence. The well identified shape type varies between participants, which explains that no shape specific main effects were found. The first highly significant interaction effect was found to be between training and SC. This effect suggests that participants with training, could identify the shapes better in the low scattering setting. This effect is no surprise, since the options for the scattering conditions was taken out of the training session, to reduce the duration of the test and avoid learning fatigue. The high significance also shows that the training could only increase the probability to find the halls with acoustic conditions, which were learned. Significant interaction effects were also found between the test conditions volume and RT and the individual shapes. These suggest, that the Volume and RT had a significant influence on the probability to identify the Horseshoe and Shoe Box. They show that a high reverberation time increased the chance to identify these shapes, in comparison to the Vineyard hall, where a low RT was beneficial. Three-way interaction effects can not be well interpreted by the coefficient value. These effects are therefore analyzed in section 3.5.2.

3.5.2 Mixed model predictions

The models can be used to predict the correct response probability for defined values of the predictor conditions. The response is predicted, with what appears to be the upper and lower confidence interval. The prediction method takes the random effects into account, when calculating the standard errors. In this case, the intervals must to be treated as prediction intervals. Significance of the observed effects is therefore not undermined by overlapping prediction intervals. The degree of overlap merely represent the chance, that the predicted outcome falls within the same range, when predictors change. These predictions shown below can be used to evaluate the measured effects further. A plot of the residual distribution and additional predictions for other combination of fixed predictor parameters are shown in the Appendix. These show merely redundant information.

It becomes evident, that training raises the probability for all concert hall shapes in all conditions besides SC_{high} . In figure 32, the predicted probabilities are shown for the shapes and both states of SC and the fixed test conditions RT_{high} and V_{big} . With low wall scattering, the predicted hit rates for trained participants are higher, while for high wall scattering, the probability for trained and untrained participants are almost identical. This suggests that participants needed to learn all acoustic cues in order to be able to memorize the acoustic spatial impression and guess the shape types correctly. The condition SC_{high} seemed to influence the acoustic impression in a way, that the learning effect with low wall scattering could not improve the accuracy in the high scattering condition. A shape type specific training effect was not present and can not be observed here. An close to significant three-way interaction effect between SC,

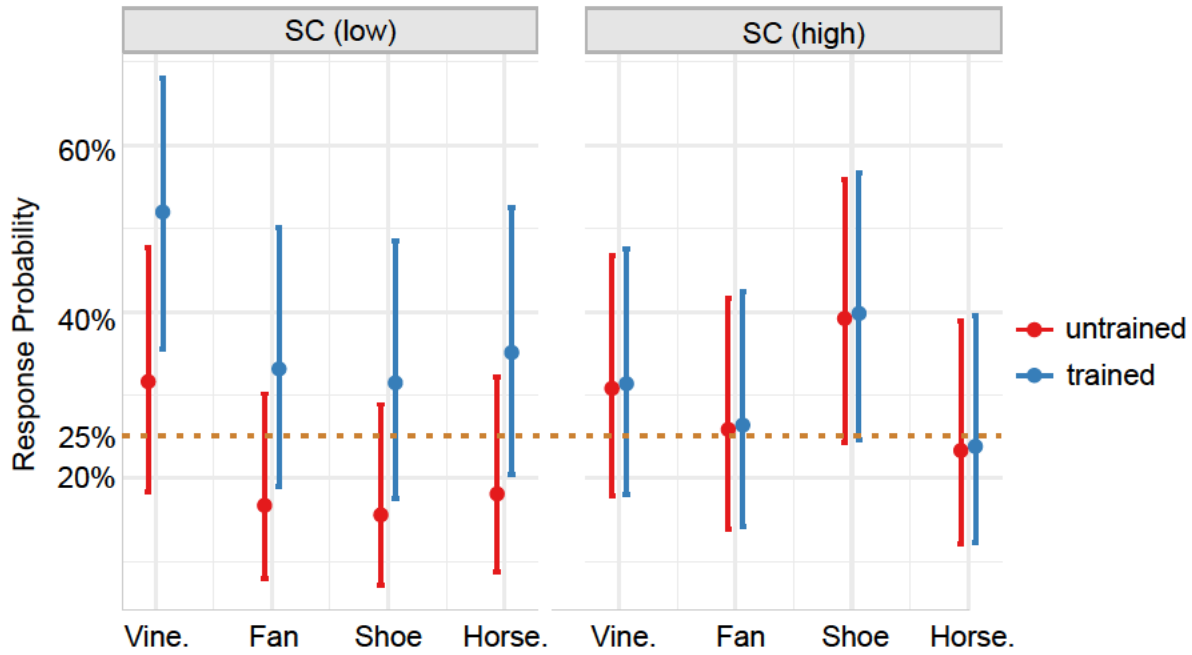


Fig. 32: Mean predicted probabilities for the conditions RT_{high} and V_{big} and both stages of the wall scattering (SC). The dashed line marks the guessing probability and the intervals show prediction intervals.

volume and the probability to detect the Shoe Box over the Vineyard was revealed by the model. This effect can be seen for untrained participants, which could identify the Shoe Box with more accuracy in the high scattering condition. A similar but small increase can be seen for the Fan and Horseshoe.

In figure 33, the predicted response probability for the conditions RT_{high} and SC_{low} are shown. The most prominent effect of the volume can be seen for the Vineyard hall. The small Vineyard hall seemed to be very hard to identify, even with a training, while the big Vineyard hall could be detected better than the other shape types. In the contrary, the Horseshoe and Shoe Box hall could be detected better with smaller volume. The difference between the predicted probabilities of the Vineyard and Shoe Box hall, due to the volume respond to the almost significant ($p < 0.1$) interaction effect found for this relation (see Tab. 6). In this case the effect is especially pronounced, since the RT_{high} condition was chosen for the predictions. A low RT decreases the probability to identify the two shape types, which can be observed in Fig. 34. Here predicted probabilities for the conditions V_{small} and SC_{low} and the two states of the RT are shown. The difference in the predicted probabilities conditioned by the reverberation time are small. Interestingly the effect of the RT on the Vineyard and Horseshoe hall is contradictory. In the small halls, a lower RT helps to identify the Vineyard while it decreases the probability to identify the Horseshoe hall. A higher RT inverts this effect which corresponds to the highly significant effect revealed by the model (see Tab. 6). The RT had also had a significant effect on the probability to detect the Shoe Box hall, which could be identified better with the longer RT.

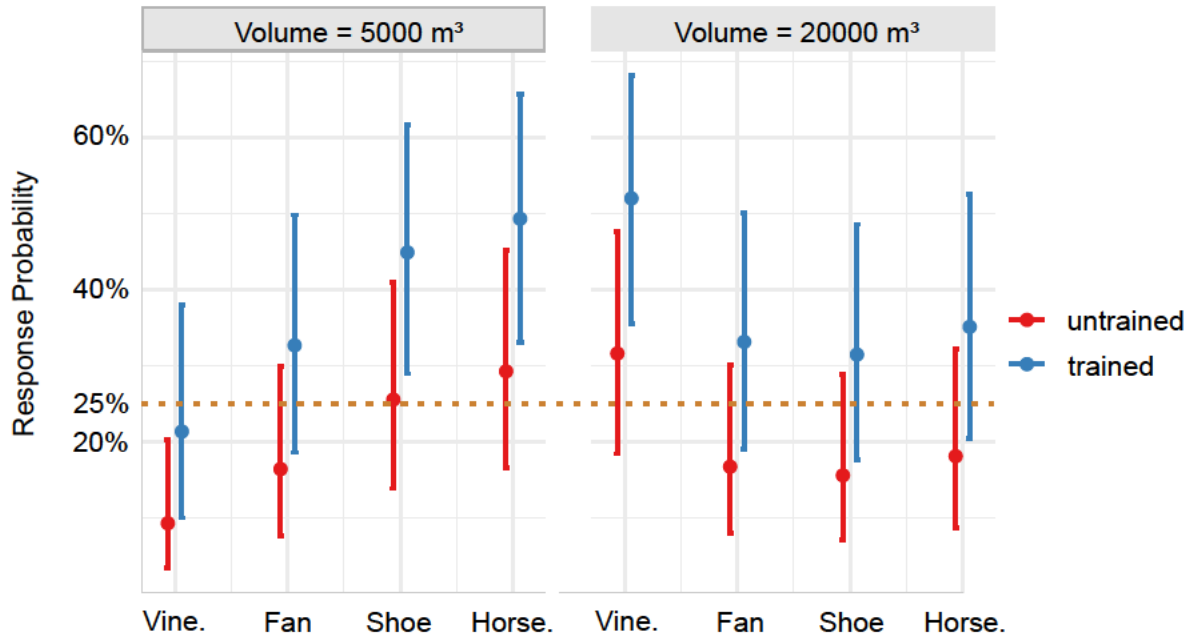


Fig. 33: Mean predicted probabilities for the test condition RT_{high} and SC_{low} and both stages of the volume. The dashed line marks the guessing probability and the intervals show prediction intervals.

The model predictions, shown above suggest a shape specific interaction effect of volume and the RT. To clarify this, the predicted probabilities for trained participants and SC_{low} are shown in figure 35. Especially with the longer reverberation time, the probability to detect the Horseshoe over the Vineyard is strongly affected by the volume. The small Vineyard is predicted to be detected only with guessing probability, while the large Vineyard could be detected very well. The inverse is the case for the Horseshoe hall, which responds to the highly significant interaction effect revealed by the model (see Tab. 6). A similar plot for untrained participants does not reveal new information, since the training raised the probability for all hall types and no shape specific training effect could be found in the data.

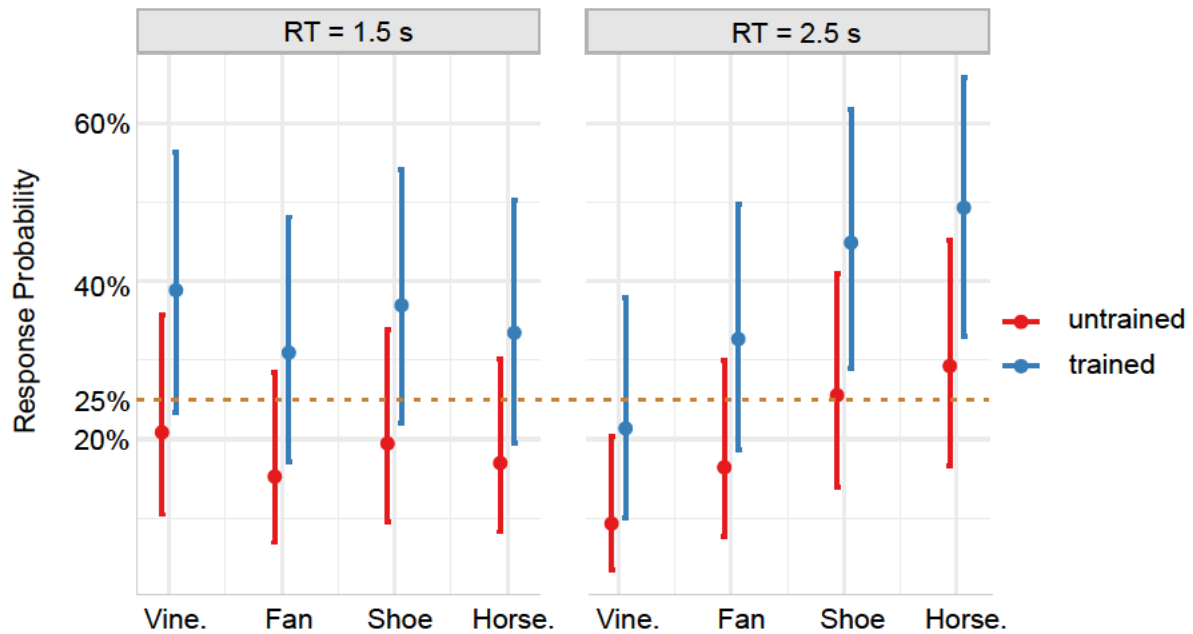


Fig. 34: Mean predicted probabilities for the conditions V_{small} and SC_{low} and both stages of the reverberation time (RT). The dashed line marks the guessing probability and the intervals show prediction intervals.

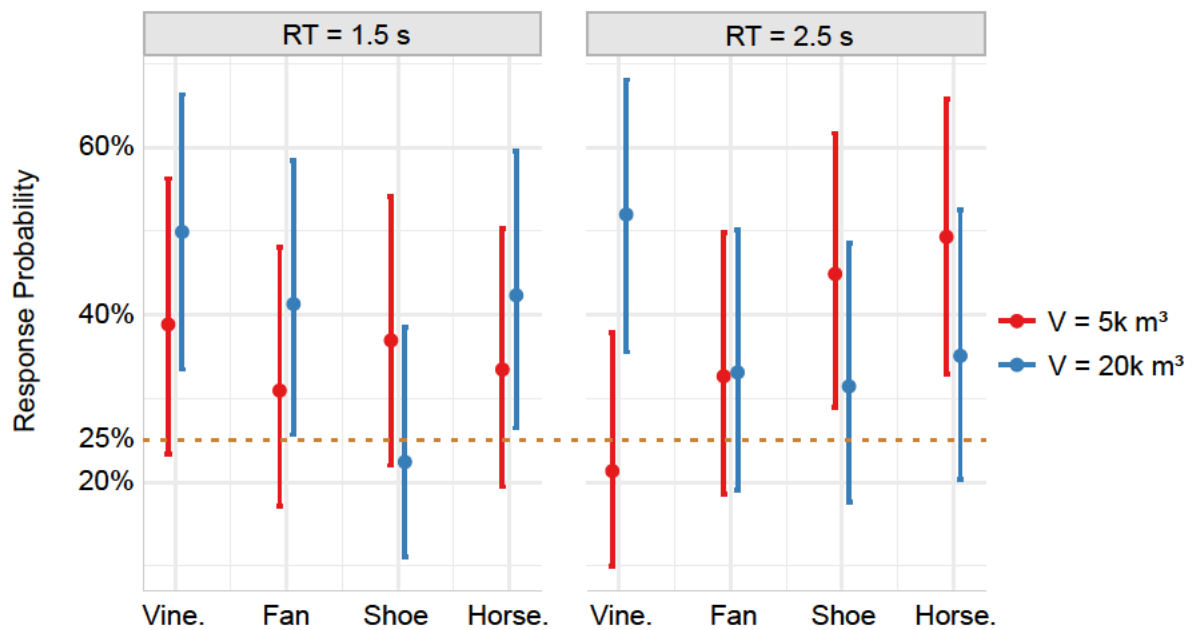


Fig. 35: Mean predicted probabilities for participants that received a training and the condition SC_{low} and both stages of the reverberation time (RT). The dashed line marks the guessing probability and the intervals show prediction intervals.

3.5.3 Mixed effects model with ASW and LEV predictors

A second model was generated including the binaural acoustic features $[1 - \text{IACC}_{E3}]$ and $[1 - \text{IACC}_{L3}]$ for the frontal head orientation $\theta = \phi = 0$ as measures of the spaciousness (ASW and LEV). The $[1 - \text{IACC}_{E3}]$ is calculated as the arithmetic average of the IACC, for the first 80 ms after the direct sound, over the 500 – 2 kHz octave band. It was found to be related to the perceived apparent source width (ASW) (Okano et al., 1998). The $[1 - \text{IACC}_{L3}]$ is calculated in a similar way but for the time interval from 80 ms to 3,5 s and was found to be a measure for the listener envelopment (LEV) (Hidaka et al., 1995). The model also included the variable “Training” as well as interaction effects between training and the IACC measures. A significant effect could be revealed for both the ASW and the LEV. A close to significant ($p < 0.1$) effect for “Training” and an interaction effect between training and the LEV measure could be revealed. The results are shown in table 7. The variance explained by the model was examined with the R-squared values calculated to be: $R^2_{\text{marginal}} = 0.03$ and $R^2_{\text{conditional}} = 0.10$. The full model can therefor only account for 10 % of the variance in the data.

Tab. 7: Fixed Effects revealed by GLMM for ASW and LEV measures. Significance ($p < .05$) is marked with *. The coefficient describes the direction of the effect. The z-value describes the distance of the effect score from the mean, measured in standard deviation units.

SI Model	Coefficient	Std. Error	z-value	p-value (sig)
Training	5.1018	3.0047	1.698	0.0895
$[1 - \text{IACC}_{E3}]$ (ASW)	-0.6351	0.3046	-2.085	0.04 *
$[1 - \text{IACC}_{L3}]$ (LEV)	6.0298	2.6198	2.302	0.02 *
Training \times $[1 - \text{IACC}_{L3}]$	-5.6482	3.5069	-1.611	0.09

Random Effect	Variance	Standard Error
Intercept	0.079	0.29

These results confirm the assumption that the ASW and LEV, are important acoustic cues of the perceived spaciousness and were use for the identification of the concert hall shapes. The predicted probability with this model is shown for the ASW measure in figure 36. The results show that a low degree of ASW measure increases the predicted probability to make a correct selection. This is a surprising result since a wide ASW was expected to assist in the identification of the concert hall shapes. Instead, a wide perceived source width decreases the audibility of the shapes. This suggests, that a wide perceived sound source, produced by early lateral reflections generates less pronounced spatial cues that could assist in the auditory perception of the shapes.

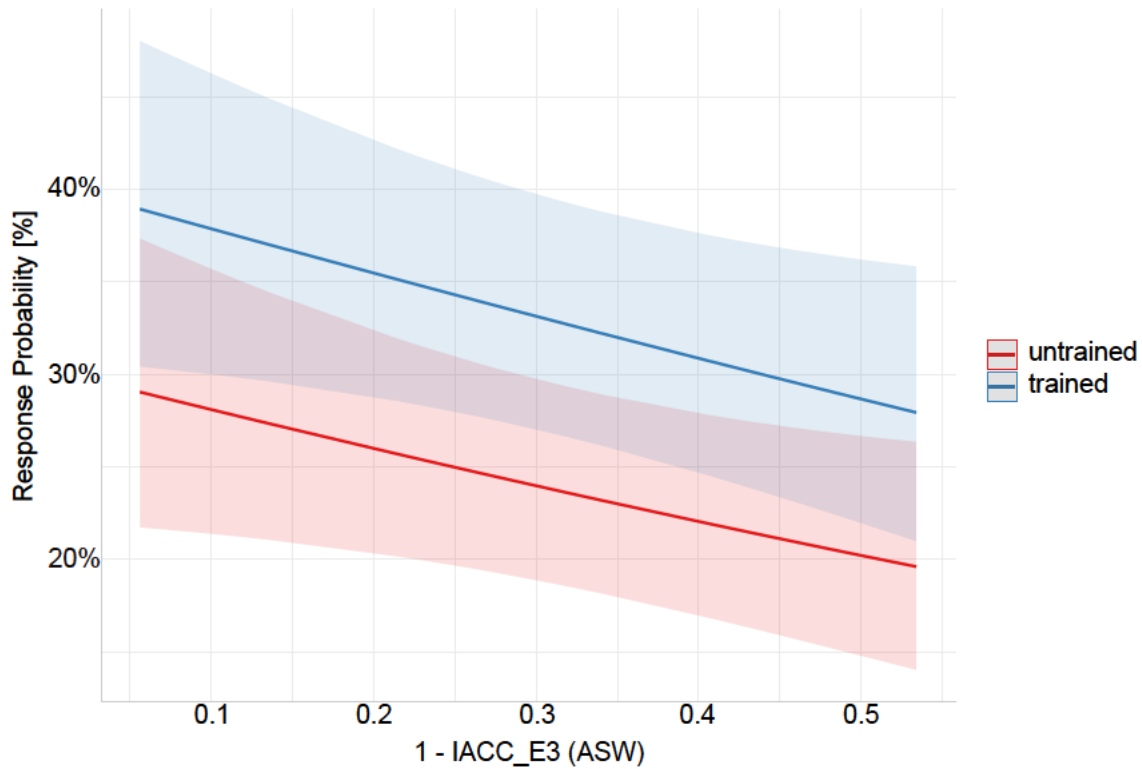


Fig. 36: Predicted probabilities for trained and untrained participants over $[1 - \text{IACC}_{E3}]$. High values of the IACC describe the sensation of a wide ASW. The semitransparent areas describe the prediction intervals. $[1 - \text{IACC}_{E3}]$ is set to its average.

Predictions over the $[1 - \text{IACC}_{L3}]$, used as a measure for the perceived LEV, are shown in figure 37. The response probability rises with the level the of LEV measure, which was shown to be a significant effect. An interesting interaction of the LEV measure and training can be observed as well. Untrained participants could identify the shapes with very low accuracy (15 %) when the degree of LEV was low. With higher level of envelopement, the response probability of untrained participants rises. Trained participants show no specific sensitivity to this LEV measure. This suggests, that participants learned to identify different acoustic cues during the training session. It is likely that specific spatial acoustic cues were used by trained participants, while untrained participants were reliant on intuition, which is supposedly more sensitive to the degree of LEV. The interaction effect for this theory was not significant ($p < 0.1$) and has to be interpreted accordingly. The fact that these measures influenced the participants interpretation validates the chosen simulation method and the visual presentation of the test scene. It shows, that the acoustic cues, which are known to constitute the spatial impression strongly, were present and could be used by a small number of participants to identify the shapes of the concert halls. An additional model was calculated including the common monaural acoustic measures: strength (G),

clarity (C80), definition (D80) and early decay time (EDT), but no significant effects could be revealed this way.

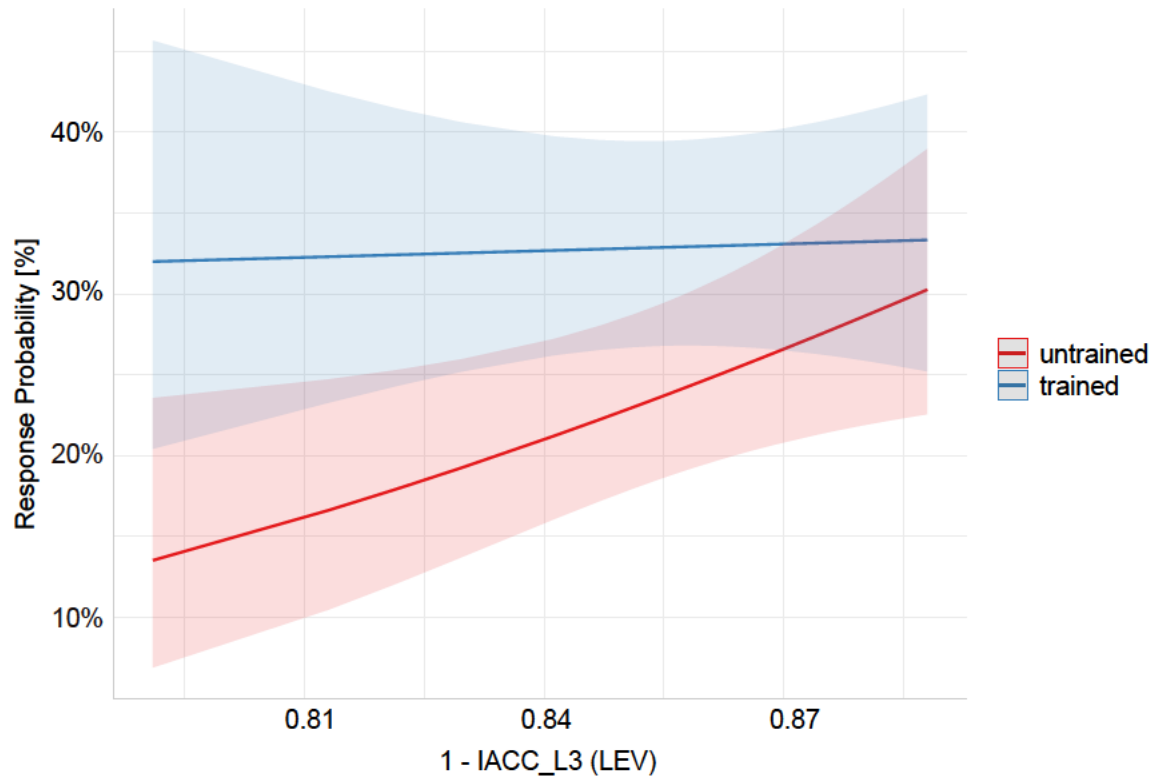


Fig. 37: Predicted probabilities for trained and untrained participants over $[1 - \text{IACC}_{L3}]$. High values of the IACC describe the sensation of a wide ASW. The semitransparent areas describe the prediction intervals. $[1 - \text{IACC}_{E3}]$ is set to its average

4 Discussion

Untrained participants were in mean able to identify 22,9 % of the halls correctly which lies below the guessing probability of 25 %. This result clearly states, that the majority of test participants were not able to confidently identify the shape of concert halls by interpreting the acoustic spatial impression in the given setup. Therefore the alternative hypothesis (H1) has to be rejected (see Sec. 1.5). Some participants were knowledgeable in the field of room acoustics and experienced visitors of concert performances. Only one untrained participant was able to guess 43,7 % of the halls correctly which lies above the upper confidence interval of the distribution. This participant did not report specific expertise but could be especially gifted in the task. The mean correct response rates for the individual shape types could not reveal significant results, suggesting, that in fact none of the concert hall shape types could confidently be identified by the majority. This result has to be interpreted with the specific acoustic scene in mind. The spatial impression changes with more sound sources, other sound signals and at other listening positions.

Participants, who received a training before the test, were on average able to identify 30 % of the concert halls correctly. An one-way ANOVA test revealed, that this response count is not significantly higher than the guessing probability, thus a significant improvement through the training session can not be shown this way. An evaluation of the individual response count could reveal that some untrained and some trained participants were able to hear certain shapes with relatively high accuracy. Nevertheless, most of these participants could only identify one of the shape types with an accuracy above 55 %. This suggests that the test participants either had different sensitivities for the acoustic cues related to their individual expertise or used different acoustic cues to make their predictions. Generalized Linear Mixed Models were used, to investigate the influence the test conditions and the training. A significant random effect between participants was found, which confirms the assumption that individual sensitivities need to be accounted for.

With this approach, a highly significant training effect could be revealed. Training showed to increase the correct response probability for all shape types. An interaction effect between training and the degree of wall scattering also showed high significance. This effect has to be interpreted with the design of the training session in mind. The condition SC_{high} was not included in the training session to reduce its duration. The results show, that the concert hall shapes could not be identified better than with guessing probability in this untrained condition. This suggests, that memorizing the acoustic cues with low scattering did not improve the accuracy to detect concert halls with higher scattering. It is worth noticing, that the author was able to reach hit rates

around 80 % after multiple training sessions, suggesting that a different training design could increase the training effect. It can further be evaluated through the confusion matrix of the shape specific response rates (see Figs. 30 & 31). As an example, untrained participants often falsely identified the Horseshoe hall for the Vineyard hall. The Fan hall was also often falsely identified for the Horseshoe hall. These mean response rates were significantly higher than the guessing probability. Trained participants did not repeat this mistake in the same manner, suggesting that the acoustic cues responsible for the mismatch were audible and could be learned.

The GLMMs revealed, that the reverberation time influenced the response probability significantly. A correct identification showed to be more likely with a lower RT. Significant interaction effects of the volume and the RT were found for specific shape types. Especially the probability to detect the Vineyard and Horseshoe was highly influenced by the volume and reverberation time. The small Vineyard hall with a long RT was apparently very hard to identify, while the large Vineyard hall could be identified a lot better. With the lower RT, this difference is almost not present. In an attempt to explain this effect, the architectural characteristics and resulting patterns of early reflections must be considered (see Sec. 2.2), as well as the acoustic features, influenced by the test condition. The visual impression of the halls has probably influenced the acoustic expectation and has to be taken into account as well. The confusion matrices showed, that the Vineyard was often mistaken for the Horseshoe hall. The results from the GLMM let reason to believe that this was especially the case in the conditions V_{small} , RT_{high} and SC_{low} .

Tab. 8: Acoustic Measurements for the Vineyard and Horseshoe hall in different test conditions and SC_{low} , measured at 1 kHz

	Vineyard				Horseshoe			
	RT_{low}		RT_{high}		RT_{low}		RT_{high}	
Volume	small	big	small	big	small	big	small	big
G [dB]	-3.2	-2.2	2	-0.8	-0.04	-5.4	1.4	-4.1
C80 [dB]	-0.05	5.3	2.9	2.9	4	5.5	1.3	2.6
D80 [%]	0.5	0.8	0.7	0.7	0.7	0.8	0.6	0.6
EDT [ms]	1.8	1.2	1.2	1.6	1.1	1.5	1.5	1.8
1 – $IACC_{E3}$	0.49	0.34	0.34	0.34	0.11	0.05	0.23	0.07
1 – $IACC_{L3}$	0.84	0.87	0.87	0.87	0.81	0.82	0.88	0.84

The small halls with a long reverberation time generate the most unusual acoustic conditions of all possible combinations of the test conditions. Here, early reflections are less dominant and likely partially masked by the late reverberation. It can be expected, that the spatial impression is therefore less pronounced in these circumstances, which aggravates the identification of spatial acoustic cues. The responses therefore depen-

dent on other acoustic parameters. This can be further observed, with the acoustic measures strength (G), clarity (C80), definition (D80) and the early decay time (EDT), of the Vineyard and Horseshoe, shown in table 8 , as well as the reflection patterns, shown in figure 38. A complete list of acoustic measures for all halls and all conditions is shown in the Appendix section 6. An unexpected variation of the clarity (C60) in the small halls with low and high RTs can be observed. With a low RT, the small Horseshoe produces a high degree of clarity and a low degree of LEV while the Vineyard generates a lower clarity and a higher degree of LEV. In this test condition, both halls are predicted to be identified with a higher probability than the guessing probability. With the longer reverberation time, the small Horseshoe produces a higher clarity than the Vineyard and a slightly lower degree of LEV. It can be assumed, that subjects expected a higher clarity from the small Horseshoe hall than from the small Vineyard hall, however, in this specific test condition it is the opposite. Interestingly, Witew et al. (2005) found that subjects judged listener envelopment as the inverse of the clarity C80, suggesting that a clear sample produces a low perceived envelopment. A similar effect can be the cause of the misinterpretation of the Horseshoe and Vineyard hall in this particular condition and the two shape types.

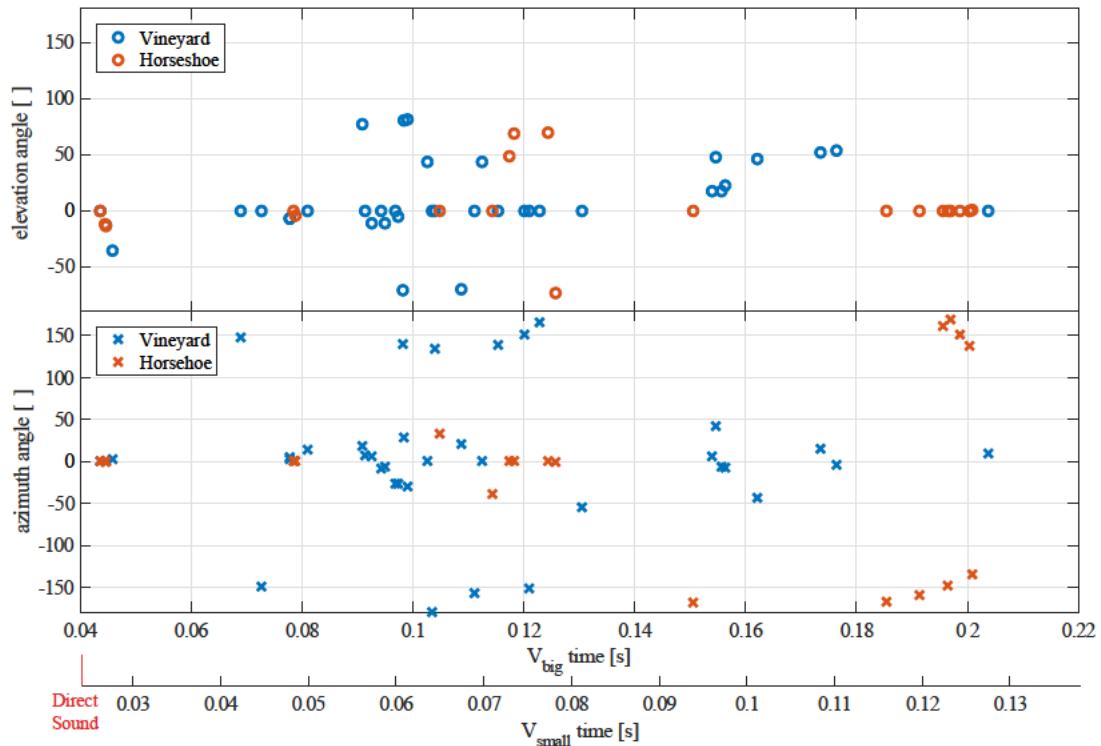


Fig. 38: Early reflections with azimuth (bottom panel) and elevation (top panel) angles of Vineyard and Horseshoe hall, calculated by the IS model (1st and 2nd order).

Figure 38 shows the early reflection pattern of the Horseshoe and Vineyard hall. The plot shows the angles of incidence of the early reflections, calculated with the image

source model for the frontal head orientation ($\phi = \theta = 0^\circ$) over time. Each data point marks a single reflection. A second x-axis describes the time axis for the same halls with smaller volume. The reflection patterns are identical, but arrive with shorter delay. It becomes evident that the Vineyard produces many frontal lateral reflections between ($-40^\circ < \phi < 40^\circ$) and back wall reflections ($150^\circ < \phi < -150^\circ$) well within the first 80 ms. In the Horseshoe similar back wall reflections arrive after the 80 ms threshold, when the volume is big. The amount of frontal early reflections, is lower. With the small volume, the back wall reflections arrive earlier and could therefore have a bigger influence the spatial impression.

The listening position was chosen to be central in the first block of the main tribune. The rising tribunes of the Vineyard and Fan hall with seat rows behind the listener produce the visual impression, of close reflecting surfaces behind the listener. Participants might have expected strong reflections from behind producing the sensation of envelopment. The Horseshoe hall in the contrary does have a wide stage and no wall behind the listener. It can be assumed, that participants expected a higher clarity and lower degree of envelopment due to the shape of the virtual Horseshoe hall. Still, the curvature of the back wall produces similar back wall reflections as the vineyard shape, but with a larger delay (180 ms in V_{big}). The assumption can be drawn, that these back wall reflections were used as an acoustic cue to identify the Vineyard. In the test condition ($SC_{\text{low}}, V_{\text{small}}, RT_{\text{high}}$), the back wall reflections of the Horseshoe arrive around 90 ms after the direct sound and might therefore have a stronger influence on the spatial impression. Reflections arriving from behind the listener were shown to increase the perceived LEV, with later reflections being more effective (Morimoto et al., 2001). This further supports the suggestion, that the horseshoe produced a unexpected high degree of envelopment in this particular test condition, which led to the misinterpretation.

The GLMM model including IACC based LEV and ASW measures showed interesting results about the influence of these parameters of the perceived spaciousness. A low degree of ASW was predicted to produce a higher chance for a correct response, than a high degree of ASW. It can be assumed, that a wide apparent source suppresses other spatial acoustic cues. Strong frontal lateral reflections tend to merge with the direct sound and increase the ASW. The perceptual integration over these reflections might lead to masking of reflections arriving in the same temporal range but different angles of incidence. A low ASW is produced, if less frontal lateral reflections are present. In this case other reflections, which are not perceptually merged with the direct sound, produce more pronounced spatial information. Especially late reflections seemed to be beneficial for the identification of the hall shapes. The degree of LEV did not have any influence for trained participants, while untrained participants showed to

benefit from a high degree of LEV. In fact a low degree of LEV caused the predicted response probability to lie well below the guessing probability. This interesting effect suggests, that the sensation of envelopment is used as an intuitive acoustic cue, which is cognitively matched with internally memorized acoustic characteristics. Trained participants were able to detect specific acoustic cues and did not have to rely on intuition.

This effect might relate to the evolutionary development of the auditory sense. The human senses are tuned by evolutionary selection to avoid predators, food selection and social exchange (Haselton et al., 2015). The auditory perception of space is a subtle sensory ability, which seems to be of minor importance and is therefore unnoticed by most people. Still the research on echolocation (see Sec. 1.2) has shown, that the possibilities of this ability are far greater than one might expect. The result of this research project suggest, that the LEV cue was most important for untrained participants, which had to rely on intuition to identify the concert hall shapes through the acoustic impression. This lets reason to believe, that the auditory perception of space behind the listener is more accurate, because this part of the environment can not be seen. Instead an intuitive assessment of the environment behind the listener is left to auditory perception. This coincides with the hypothesis proposed by Griesinger (1997), that frontal and background spatial impression can be cognitively separated in a foreground and background stream and that the background SI is the dominant form. A more focused study on this theory could be done using similar methods and focusing more closely on the effect of reflections arriving from behind and the LEV cue. In consideration of echolocation mechanisms, an investigation on the audibility of close objects, especially behind the listener, could further help to understand this effect.

The full spectrum of results show that the complex test setup was able to display the audiovisual cues necessary to investigate the multimodal perception of space in the concert hall setting. The concert hall models produced individual audible characteristics, which could be identified and learned by some participants. The results suggest, that the ability to hear the shape of a concert hall is dependent on the individual expertise and sensitivity towards spatial acoustic cues.

4.1 Critical assessment and possible sources of error

The generated test setup showed to be able to display the visual and auditory cues necessary to make adjudicated predictions for some participants. Still, the technological methods can not fully cover all possible acoustic effects. Especially objects in the close range around the spectator in a real concert hall, could produce acoustic cues, which can likely be subconsciously or actively perceived. The simulation and auralization method needs to be highly accurate in order to provide all natural acoustic cues. In this test a non-individualized HRTF data-set was used to provide the binaural signals. Individual HRTF measurements could help to increase the prediction accuracy.

A more detailed and graphically optimized virtual surrounding could produce other acoustic expectations and also influence the results. Other audiovisual interaction effects of the size, color, brightness and other perceptual parameters could be taken into account to further investigate the acoustic expectation. For this work, the visual design of the models was chosen to be as similar as possible in all halls. It was attempted to reduce the influence of specific visual cues but to preserve a naturalistic immersive experience. The same textures were used on the walls, ceiling and floor of the models. Due to the shape, these surfaces have different sizes and therefor generate different lighting conditions. The visual impression of the brightness and color pattern therefore varies slightly between the halls.

A feedback driven training session could have produced a higher training effect. The training stimuli had to be reduced by half to protect the test subjects and prevent learning fatigue. Additionally, the 4-AFC format could have produced listening fatigue after a number of test iterations and therefor influence the concentration of participants. A reduced test design could prevent this effect.

A technologically advanced Virtual-Reality headset could likely reduce the stress for participants generated by low frame-rates and the weight of the headset and headphones. Still, most participants were able to perform the test without a noticeable lack of concentration.

5 Conclusion

In this work, an immersive audiovisual listening test was created to test if people can hear the shape of concert halls. Test participants were placed in four concert halls by means of a virtual reality display and high resolution binaural auralizations. They were asked to find the acoustic simulation that matches with the hall they were virtually sitting in. The generated data was analyzed, using predictive regression models to test the influence of different acoustic conditions. The main results are summarized in the list below:

1. Untrained participants were on average not able to confidently identify the concert hall shapes by comparing and interpreting the acoustic impression in the investigated setting. Nevertheless, a small number of trained and untrained participants was able to identify certain shape types with some accuracy.
2. Training helped participants to identify the shapes better. The GLMM revealed a significant effect of the training, but could only explain a small part of the variance in the data. However, the overall mean hit rate of trained participants could not reach a significant improvement from the guessing probability. This suggests that only a few participants were able to memorize the acoustic spatial impressions of the concert halls.
3. The ASW and LEV were important acoustic cues for the discrimination of the spatial impressions and the identification of the shape types. A high degree of ASW reduced the probability of identifying the shapes correctly, while a high degree of LEV helped untrained participants to identify the concert hall shapes.
4. The volume and reverberation time had a strong influence on the probability to identify of the Vineyard and Horseshoe hall in particular. It is assumed, that these parameters can influence the sensation of envelopment depending on the specific shape. Whether this relates to the clarity (C80) needs to be investigated further.

The results show, that the simulation and auralization method was able to produce acoustic cues that some participants were able to detect and make an educated guess about the shape of the concert hall. The immersive visual environment enabled participants to experience the visual impression of the concert halls in a natural and comprehensive way and match the acoustic spatial impression with the architectural features. The work has shown that acoustic spatial perception must not only be treated as a means to identify acoustic quality but as a multimodal cognitive ability. A better understanding of the boundaries of this ability is important for developments in the

field of acoustics and virtual acoustics. Further research could reveal, whether different source signals and multiple sound sources produce additional spatial information and influence the outcome of a similar test design. Especially the relationship between the sound source position and the listener position as well as the position of the listener in relation to the architectural space are likely to be important in this quest. It could for example be useful to find the critical distance to objects and walls, at which they are consciously or subconsciously detected in different acoustic scenes. In the virtual concert hall environments, created for this work, participants were placed in a seat row of the main tribune. A high reflective wooden back of the seat close behind the listeners head, or the sound absorbing seat cushions and the acoustic shadow of the seat row in the front can influence the immediate sound field around the listener. Echolocation experiments could be used to investigate, if the form and substance of surfaces in the close environment can be perceived through acoustic cues in a similar setting. These cues could probably influence the acoustic spatial impression more strongly, than previously believed. For this work, the immediate environment around the listener was not acoustically simulated in detail. It is unclear if the simulation and auralization method are able to produce the necessary auditory cues to enable the perception of close objects. Experiments with a similar methodology and test tasks, related to those used in echolocation research, could provide an interesting way to evaluate the quality of simulation and auralization tools. This will further help to develop new approaches to increase the authenticity of simulated virtual acoustic environments. Spatial reverberation tools for real time applications (e.g. video-games) have a high computational demand. Reducing the simulation effort while maintaining important audible characteristics could help to decrease the required workload. A better understanding of the auditory spatial perception and scene related dependencies could help to reduce this workload. In the context of spatial reverberation tools for artistic projects, the results of this work are useful as well. If people are not able to identify the shape of a performance space through spatial acoustic cues, an arbitrary room model could be used to generate acoustic simulations for three-dimensional musical productions, while avoiding audiovisual discrepancy. However, the results have shown, that the room volume and the reverberation time must be adjusted, when doing so.

References

- Ahrens, Jens; Matthias Geier; and Sascha Spors (2008): “The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods.” In: *124th Audio Engineering Society Convention*.
- Ando, Yoichi (2012): *Concert hall acoustics*, vol. 17. Springer Science & Business Media.
- Arons, Barry (1992): “A review of the cocktail party effect.” In: *Journal of the American Voice I/O Society*, **12**, pp. 35–50.
- Barron, M. and A.H. Marshall (1981): “Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure.” In: *Journal of Sound and Vibration*, **77**, pp. 211–232.
- Beranek, Leo L (2004): *Concert halls and opera houses : music, acoustics, and architecture*. 2. ed. Springer Verlag.
- Berzborn, Marco; Ramona Bomhardt; Johannes Klein; Jan-Gerrit Richter; and Michael Vorländer (2017): “The ITA-Toolbox: An Open Source MATLAB Toolbox for Acoustic Measurements and Signal Processing.” 43th Annual German Congress on Acoustics, Kiel (Germany).
- Böhm, Christoph (2015): *Entwicklung einer Versuchsumgebung zur Auralisation von virtuellen Konzerträumen für Musiker*. Master’s thesis, Technische Universität Berlin.
- Böhm, Christoph; Felicitas Fiedler; and Stefan Weinzierl (2019): “An Anechoic Recording of Cicero’s 3rd Cataline Oration: Italian, Latin and German.” private communication.
- Bilsen, FA (1966): “Repetition pitch: Monaural interaction of a sound with the repetition of the same, but phase shifted, sound.” In: *Acta Acustica United with Acustica*, **17**, pp. 295–300.
- Blauert, J. (1971): “Localization and the Law of the First Wavefront in the Median Plane.” In: *Journal of the Acoustic Society of America*, **50**, pp. 466–470.
- Blauert, Jens (1997): *Spatial hearing: the psychophysics of human sound localization*. MIT press.
- Bradley, John S and Gilbert A Soulodre (1995): “The influence of late arriving energy on spatial impression.” In: *The Journal of the Acoustical Society of America*, **97**, pp. 2263–2271.
- Brinkmann, Fabian; Alexander Lindau; Stefan Weinzierl; Gunnar Geissler; and Steven van de Par (2013): “A high resolution head-related transfer function database including different orientations of head above the torso.” In: *Proceedings of the AIA-DAGA 2013 Conference on Acoustics*.
- Buchholz, Jörg M (2007): “Characterizing the monaural and binaural processes underlying reflection masking.” In: *Hearing research*, **232**, pp. 52–66.

- Burton, G. (2000): “The role of the sound of tapping for nonvisual judgment of gap crossability.” In: *Journal of Experimental Psychology: Human Perception and Performance*, **26**, pp. 900–916.
- Diderot, M, D trans. by Jourdain (1916): “Letter on the blind for the use of those who see.” In: *The Open Court Series of Classics of Science and Philosophy*, pp. 68–218.
- Dietsch, L and W Kraak (1986): “Ein objektives Kriterium zur Erfassung von Echostörungen bei Musik-und Sprachdarbietungen.” In: *Acta Acustica united with Acustica*, **60**, pp. 205–216.
- DIN, EN (2012): “60268-16: Elektroakustische Geräte—Teil 16: Objektive Bewertung der Sprachverständlichkeit durch den Sprachübertragungsindex.” In: *Beuth, Berlin*.
- Embrechts, Jean-Jacques; Dominique Archambeau; and Guy-Bart Stan (2001): “Determination of the scattering coefficient of random rough diffusing surfaces for room acoustics applications.” In: *Acta Acustica united with Acustica*, **87**, pp. 482–494.
- Finnegan, D. J.; E. O’Neill; and M. J. Proulx (2017): “An approach to reducing distance compression in audiovisual virtual environments.” In: *2017 IEEE 3rd VR Workshop on Sonic Interactions for Virtual Environments (SIVE)*. pp. 1–6.
- Grechkin, Timofey; Tien Nguyen; Jodie Plumert; James Cremer; and Joseph Kearney (2010): “How does presentation method and measurement protocol affect distance estimation in real and virtual environments?” In: *Transactions on Applied Perception*, **7**.
- Griesinger, David (1997): “The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces.” In: *Acta Acustica united with Acustica*, **83**, pp. 721–731.
- Hameed, Sharaf; Jyri Pakarinen; Kari Valde; and Ville Pulkki (2004): “Psychoacoustic Cues in Room Size Perception.” In: *Audio Engineering Society Convention 116*.
- Haselton, Martie G; Daniel Nettle; and Damian R Murray (2015): “The evolution of cognitive bias.” In: *The handbook of evolutionary psychology*, pp. 1–20.
- Hausfeld, Steven; Roderick P Power; Angela Gorta; and Patricia Harris (1982): “Echo perception of shape and texture by sighted subjects.” In: *Perceptual and Motor Skills*, **55**, pp. 623–632.
- Hendrickx, Etienne; et al. (2017): “Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis.” In: *The Journal of the Acoustical Society of America*, **141**, pp. 2011–2023.
- Hidaka, Takayuki; Leo L Beranek; and Toshiyuki Okano (1995): “Interaural cross-correlation, lateral fraction, and low-and high-frequency sound levels as measures of acoustical quality in concert halls.” In: *The Journal of the Acoustical Society of America*, **98**, pp. 988–1007.
- ISO 3382-1:2009 (2009): *Acoustics—Measurement of Room Acoustic Parameters—Part 1: Performance Spaces*. Standard, International Organization for Standardization, Geneva, CH.

- Johnson, Dale and Hyunkook Lee (2019): “Perceptual threshold of apparent source width in relation to the azimuth of a single reflection.” In: *The Journal of the Acoustical Society of America*, **145**, pp. 272–276.
- Kahle, Eckhard (1995): “Validation d’un modèle objectif de la perception de la qualité acoustique dans un ensemble de salles de concerts et d’opéras.” In: *Unpublished Ph. D. dissertation, Université du Maine, Le Mans*.
- Kolarik, Andrew J; Silvia Cirstea; Shahina Pardhan; and Brian CJ Moore (2014): “A summary of research investigating echolocation abilities of blind and sighted humans.” In: *Hearing research*, **310**, pp. 60–68.
- Kopčo, Norbert and B Shinn-Cunningham (2002): “Auditory localization in rooms: Acoustic analysis and behavior.” In: *Proceedings of the 32 nd EAA International Acoustics Conference of the European Acoustics*.
- Letowski, Tomasz and Szymon Letowski (2011): “Localization error: Accuracy and precision of auditory localization.” In: *Advances in sound localization*, pp. 55–78.
- Lindau, A.; et al. (2014): “Spatial Audio Quality Inventory (SAQI).” In: *Acta Acust. United Acust.*, **100**, pp. 984–994.
- Lindau, Alexander; Linda Kosanke; and Stefan Weinzierl (2012): “Perceptual evaluation of model-and signal-based predictors of the mixing time in binaural room impulse responses.” In: *Journal of the Audio Engineering Society*, **60**, pp. 887–898.
- Lindau, Alexander and Stefan Weinzierl (2011): “Assessing the plausibility of virtual acoustic environments.” In: *Proc. of the EAA Forum Acusticum*. Aalborg, pp. 1187–1192.
- Lindau, Alexander; Stefan Weinzierl; and HJ Maempel (2006): “FABIAN-An instrument for software-based measurement of binaural room impulse responses in multiple degrees of freedom.” In: *24. Tonmeistertagung*, pp. 621 – 625.
- Llorach, Gerard; Giso Grimm; Maartje ME Hendrikse; and Volker Hohmann (2018): “Towards Realistic Immersive Audiovisual Simulations for Hearing Research.” In: *Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia*. pp. 33 – 40.
- Long, M.; M. Levy; and R. Stern (2006): *Architectural Acoustics*. Applications of modern acoustics. Elsevier Science.
- Maempel, Hans-Joachim and Michael Horn (2017): “The Virtual Concert Hall - a research tool for the experimental investigation of audiovisual room perception.” In: *International Journal on Stereo and Immersive Media*.
- Maempel, Hans-Joachim and Matthias Jentsch (2013): “Auditory and Visual Contribution to Egocentric Distance and Room Size Perception.” In: *Building Acoustics*, **20**, pp. 383–401.
- Marshall, A. H. (1967): “A note on the importance of room cross-section in concert halls.” In: *Journal of Sound and Vibration*, **5**, pp. 100–112.

- Mason, Russell; Tim Brookes; and Francis Rumsey (2004): “Spatial impression: measurement and perception of concert hall acoustics and reproduced sound.” In: *Proceedings of the International Symposium on Room Acoustics*.
- Minnaar, P.; S. K. Olesen; F. Christensen; and H. Moller (2001): “The Importance of Head Movements for Binaural Room Synthesis.” Proc. of the International Conference on Auditory Displays.
- Morimoto, Masayuki; Kazuhiro Iida; and Kimihiro Sakagami (2001): “The role of reflections from behind the listener in spatial impression.” In: *Applied Acoustics*, **62**, pp. 109–124.
- Nakagawa, Shinichi; Paul CD Johnson; and Holger Schielzeth (2017): “The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded.” In: *Journal of the Royal Society Interface*, **14**, pp. 2–13.
- Okano, Toshiyuki; Leo L Beranek; and Takayuki Hidaka (1998): “Relations among interaural cross-correlation coefficient (IACC E), lateral fraction (LF E), and apparent source width (ASW) in concert halls.” In: *The Journal of the Acoustical Society of America*, **104**, pp. 255–265.
- Peugh, James (2010): “A practical guide to multilevel modeling.” In: *Journal of school psychology*, **48**, pp. 85–112.
- Potter, J. M. (1993): *On the binaural modelling of spaciousness in room acoustics*. Ph.D. thesis, Technical University of Delft.
- Reichardt, v W and W Schmidt (1966): “Die hörbaren Stufen des Raumeindrucks bei Musik.” In: *Acustica*, **17**.
- Rojas, Juan Antonio Martínez; Jesús Alpuente Hermosilla; Rocío Sánchez Montero; and Pablo Luis López Espí (2010): “Physical analysis of several organic signals for human echolocation: Hand and finger produced pulses.” In: *Acta acustica united with acustica*, **96**, pp. 1069–1077.
- Rosenblum, Lawrence D; Michael S Gordon; and Luis Jarquin (2000): “Echolocating distance by moving and stationary listeners.” In: *Ecological Psychology*, **12**, pp. 181–206.
- Rowan, Daniel; et al. (2013): “Identification of the lateral position of a virtual object based on echoes by humans.” In: *Hearing research*, **300**, pp. 56–65.
- Rumsey, Francis (2002): “Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm.” In: *Journal of the Audio Engineering Society*, **50**, pp. 651–666.
- Santani Teng, David Whitney (2011): “The acuity of echolocation: Spatial resolution in the sighted compared to expert performance.” In: *Journal of visual impairment & blindness*, **105**, p. 20.
- Schenkman, Bo N and Mats E Nilsson (2010): “Human echolocation: Blind and sighted persons’ ability to detect sounds recorded in the presence of a reflecting object.” In: *Perception*, **39**, pp. 483–501.

- Schenkman, Bo N and Mats E Nilsson (2011): “Human echolocation: pitch versus loudness information.” In: *Perception*, **40**, pp. 840–852.
- Schröder, Dirk (2011): *Physically based real-time auralization of interactive virtual environments*, vol. 11. Logos Verlag Berlin GmbH.
- Schröder, Dirk and Michael Vorländer (2011): “RAVEN: A real-time framework for the auralization of interactive virtual environments.” In: *Forum Acusticum*. Aalborg Denmark, pp. 1541–1546.
- Schröder, Dirk; et al. (2010): “Virtual reality system at RWTH Aachen University.” In: *Proceedings of the international symposium on room acoustics (ISRA)*, Melbourne, Australia.
- Schubert, P (1966): “Untersuchungen über die Wahrnehmbarkeit von Einzelrückwürfen bei Musik.” In: *Technische Mitteilung RFZ*, **3**, pp. 124–127.
- Seraphim, H.P. (1961): “Über die Wahrnehmbarkeit mehrerer Rückwürfe von Sprachschall.” In: *Acta Acustica united with Acustica*, **11**, pp. 80–91.
- Supa, Michael; Milton Cotzin; and Karl M Dallenbach (1944): ““Facial vision”: the perception of obstacles by the blind.” In: *The American Journal of Psychology*, **57**, pp. 133–183.
- Thaler, Lore; Stephen R Arnott; and Melvyn A Goodale (2011): “Neural correlates of natural human echolocation in early and late blind echolocation experts.” In: *PLoS one*, **6**, p. e20162.
- Thaler, Lore; Jennifer L Milne; Stephen R Arnott; Daniel Kish; and Melvyn A Goodale (2013): “Neural correlates of motion processing through echolocation, source hearing, and vision in blind echolocation experts and sighted echolocation novices.” In: *Journal of Neurophysiology*, **111**, pp. 112–127.
- Thurlow, Willard R and Charles E Jack (1973): “Certain determinants of the “ventriloquism effect”.” In: *Perceptual and motor skills*, **36**, pp. 1171–1184.
- Valente, Daniel L and Jonas Braasch (2010): “Subjective scaling of spatial room acoustic parameters influenced by visual environmental cues.” In: *The Journal of the Acoustical Society of America*, **128**, pp. 1952–1964.
- Valente, Daniel L; Jonas Braasch; and Shane A Myrbeck (2012): “Comparing perceived auditory width to the visual image of a performing ensemble in contrasting bi-modal environments.” In: *The Journal of the Acoustical Society of America*, **131**, pp. 205–217.
- Vorländer, Michael and Eckard Mommertz (2000): “Definition and measurement of random-incidence scattering coefficients.” In: *Applied acoustics*, **60**, pp. 187–199.
- Vorländer, Michael; Dirk Schröder; Sönke Pelzer; and Frank Wefers (2015): “Virtual reality for architectural acoustics.” In: *Journal of Building Performance Simulation*, **8**, pp. 15–25.
- Weinzierl, S.; S. Lepa; and D. Ackermann (2018): “A measuring instrument for the auditory perception of rooms. The Room Acoustical Quality Inventory (RAQI).” In: *Journal of the Acoustical Society of America*, **144**, pp. 1245–1257.

- Weinzierl, Stefan and Michael Vorländer (2015): “Room Acoustical Parameters as Predictors of Room Acoustical Impression: What Do We Know and What Would We Like to Know?” In: *Acoustics Australia*, **43**, pp. 41–48.
- Weinzierl, Stefan; et al. (2017): “A Database of anechoic microphone array measurements of musical instruments.” private communication.
- Willemsen, Peter; Amy A Gooch; William B Thompson; and Sarah H Creem-Regehr (2008): “Effects of stereo viewing conditions on distance perception in virtual environments.” In: *Presence: Teleoperators and Virtual Environments*, **17**, pp. 91–101.
- Witew, Ingo B; Gottfried K Behler; and Michael Vorländer (2005): “About just noticeable differences for aspects of spatial impressions in concert halls.” In: *Acoustical science and technology*, **26**, pp. 185–192.
- Zuur, Alain F; Elena N Ieno; and Chris S Elphick (2010): “A protocol for data exploration to avoid common statistical problems.” In: *Methods in ecology and evolution*, **1**, pp. 3–14.

List of Figures

1	In-game view of Vineyard	2
2	Spatial attributes in acoustic scene	6
3	Terminology of spatial impression	7
4	Image source model	14
5	POV perspective of listening test	18
6	GUI of training menu	20
7	GUI of test menu	21
8	Sketch of test setup	22
9	Synchronization screen	24
10	Controller	24
11	Sketch of floor layout	25
12	Wide shot of Vineyard	26
13	Groundplan and cross section of Vineyard	27
14	Wide shot of Shoe Box	28
15	Groundplan and cross section of Shoe Box	29
16	Wide shot of in-game Fan hall	30
17	Groundplan and cross section of Fan	31
18	Wide shot of in-game Horseshoe	32
19	Groundplan and cross section of Horseshoe	33
20	Coordinate system of receiver	36
21	Source and listener position	37
22	Acoustic properties of audience	38
23	Acoustic properties of organ	39
24	Test condition: everberation time	40
25	Test condition: wall scattering	41
26	Average response count of untrained participants	46
27	Average response count of trained participants	47
28	Sum of participant responses over shape	49
29	Total response count histogram	49
30	Confusion matrix of untrained participants	50
31	Confusion matrix of trained participants	51
32	Mixed model predictions (RT_{high} , V_{big})	54
33	Mixed model predictions (RT_{high} , SC_{low})	55
34	Mixed model predictions (V_{small} , SC_{low})	56
35	Mixed model predictions (SC_{low} and trained participants)	56
36	Mixed model predictions for ASW	58
37	Mixed model predictions for LEV	59
38	ER pattern of Vineyard and Horseshoe	62
39	Residuals over fitted values of condition model	76
40	Residuals over fitted values of spatial impression model	77

List of Tables

1	Subjective effects of reflection delay	7
2	Test conditions calculated for each concert hall.	17
3	Environmental conditions used in acoustic Simulations.	35
4	Simulation preferences	35
5	Dummy coding for test conditions	44
6	Mixed model effects	52
7	Spatial impression mixed model	57
8	Acoustic measurements	61

6 Appendix

Residual plots for mixed model

The validity of the model can be shown through the residual plots below as proposed by Zuur et al. (2010). The variances are equally spread which confirms the homoscedasticity assumption.

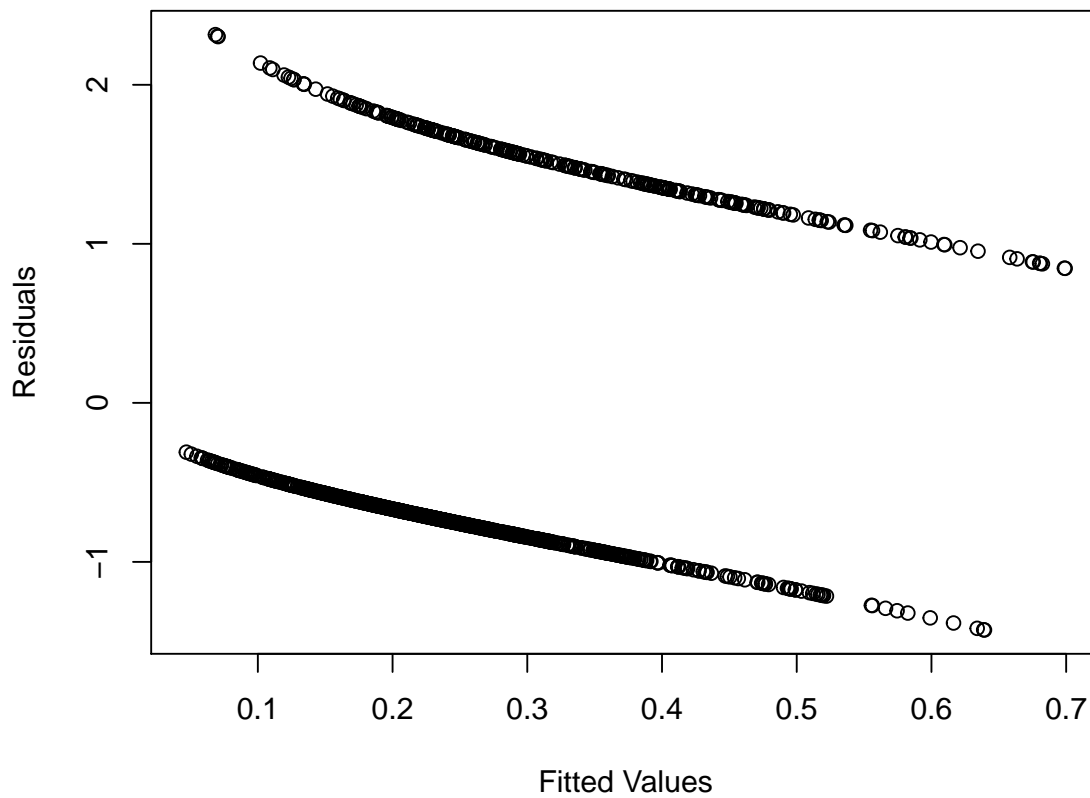


Fig. 39: Residuals vs fitted values of stimuli model including test conditions. Equal spread confirms the homogeneity assumption. The atypical scatter plot is a result of the binary response data.

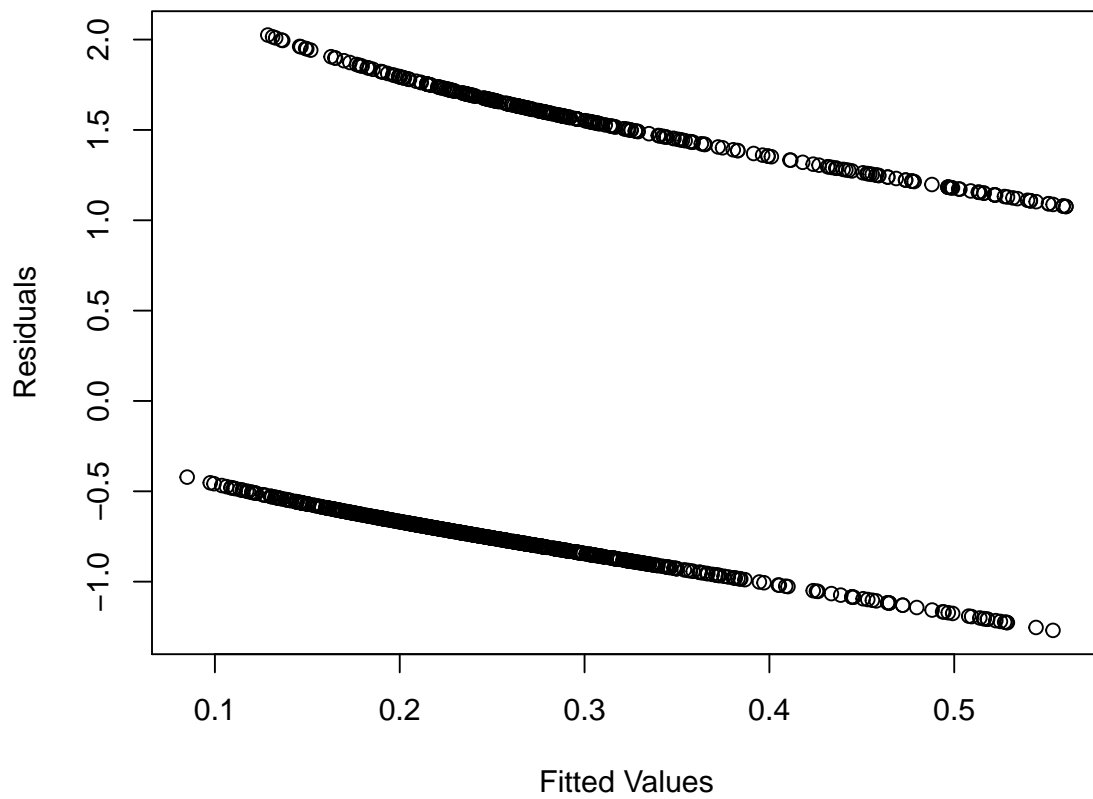


Fig. 40: Residuals vs fitted values of spatial impression model including ASW and LEV measures. Equal spread confirms the homogeneity assumption. The atypical scatter plot is a result of the binary response data.

Fragebogen:

Probanden-ID: _____

Alter: _____

Geschlecht: Weiblich: ☐ Männlich: ☐ Andere: ☐

- Leiden Sie unter Hörschäden oder anderen Einschränkungen Ihres Hörvermögens, die Ihnen bekannt sind?

JA: ☐ NEIN: ☐

Wenn „JA“ welche:

Bitte bewerten Sie Ihre persönliche Expertise nach eigener Einschätzung:

- Haben Sie Erfahrung mit jeglicher Art von Hörversuchen?

JA: ☐ NEIN: ☐

Wenn „JA“, bitte bewerten Sie ihre Erfahrung auf der Skala:

Keine Erfahrung 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 Viel Erfahrung

- Bitte bewerten Sie ihre persönliche Expertise im Bereich der **Akustik**:

Keine Erfahrung 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 Experte

Wie häufig besuchen Sie im durchschnitt Konzerte o.Ä. in Konzert-/Theatersälen:

Öfter als 1x im Monat: ☐

1x im Monat: ☐

1x in 3 Monaten: ☐

1x in 6 Monaten: ☐

Weniger als 1x in 6 Monaten: ☐

Erläuterungen zu dem Hörversuch

In diesem Hörversuch geht es um die Frage: **Können Menschen die Form eines Konzertsaaes hören?**

Bei dem Versuch werden Sie mithilfe einer Virtual-Reality Brille in virtuellen Konzertsälen platziert. Jeder Saal hat eine eindeutige Form. Es werden insgesamt vier Säle untersucht. Diese sind vom Typ: **Weinberg, Fächer, Schuhgeschachtel und Hufeisen/Theater**. Zum besseren Verständnis schauen Sie sich bitte die Grundrisse auf dem separaten Zettel an.

Über Kopfhörer werden Ihnen dreidimensionale raumakustische Simulationen der Räume vorgespielt. Über einen Headtracker wird Ihre Kopfbewegung verfolgt, sodass Sie sich nach Belieben in den Räumen umschauen und diese mit beliebigen Blickrichtungen abhören können.

Der Test wird insgesamt 32 mal wiederholt. Bei jeder Wiederholung wird Ihnen jeweils nur ein Raum visuell gezeigt. Dazu stehen Ihnen vier akustische Stimuli zur Auswahl. Jeder Stimulus stellt eine akustische Raumsimulation der vier Raumtypen dar. Nur einer der vier Auswahlmöglichkeiten entspricht dabei dem Raum, den Sie sehen. Ihre Aufgabe ist es, diese Raumsimulation zu finden.

Bei den 32 Wiederholungen werden jeweils unterschiedliche raumakustische Parameter untersucht. Dazu zählt, das Volumen, die Nachhallzeit und die Streuung der Wandmaterialien. Diese sind pro Wiederholung bei allen vier Auswahlmöglichkeiten gleich. Es unterscheiden sich jeweils nur die Form der simulierten Räume.

Für Fragen wenden Sie sich bitte an den Versuchsleiter.

Anmerkungen:

- Sie können die Akustischen Stimuli in jeder Blickrichtung umschalten, da das Menü immer vor Ihnen erscheint. Dies ist bei der Aufgabe, den richtigen Stimulus auszuwählen hilfreich. Bitte hören Sie die Räume also mit verschiedenen Blickrichtungen ab.
- Bitte achten Sie darauf, dass Ihre Wahl auch korrekt verbucht wird. Wenn Sie auf „Send Answer“ klicken, wird derjenige Stimulus als Auswahl gespeichert, der aktuell spielt. Dieser ist Rot markiert und schriftlich unter dem Button vermerkt („Current Selection: X“). Falls Sie einmal einen falschen stimulus ausgewählt haben, teilen Sie dies bitte dem Versuchsleiter mit. Die Änderung wird dann nachträglich vorgenommen.

Einverständniserklärung zur Teilnahme am Hörexperiment

Name: _____

[für Studenten AKT] Matrikelnummer: _____

Versuchsdauer: _____

Ich bin ausführlich und verständlich über Wesen und Bedeutung des Hörexperiments aufgeklärt worden und hatte ausreichend Zeit, Fragen zu stellen. Mir ist bekannt, dass ich die Teilnahme jederzeit ohne Nachteile und ohne Angabe von Gründen abbrechen kann. Hiermit erkläre ich mein Einverständnis am beschriebenen Hörexperiment teilzunehmen und stimme einer anonymisierten Veröffentlichung der Ergebnisse zu. Auch diese Einwilligung kann ich jederzeit ohne Angabe von Gründen widerrufen. Die Versuchsdauer wird für Studierende des Studiengangs Audiokommunikation und -technologie durch die Versuchsleitung angerechnet.

Berlin, den _____

(Ort, Datum)

(Versuchsteilnehmer_in)

(Versuchleitung)

Acoustic measurements for all concert halls in all test conditions

Acoustic attributes at 1 kHz									
	Volume	RT	SC	G	C50	C80	D50	D80	EDT
Vineyard	big	low	low	-2,22	0,90	5,36	0,55	0,77	1,20
	big	low	high	-3,91	1,05	3,94	0,56	0,71	1,21
	big	high	low	-0,79	-0,98	2,90	0,44	0,66	1,56
	big	high	high	-2,38	-1,19	1,34	0,43	0,58	1,82
Fan	big	low	low	-3,35	5,65	5,96	0,79	0,80	1,69
	big	low	high	-4,29	2,52	5,42	0,64	0,78	1,22
	big	high	low	-1,58	2,45	2,76	0,64	0,65	2,02
	big	high	high	-2,56	-0,06	2,51	0,50	0,64	1,76
Shoe Box	big	low	low	-3,77	5,69	8,25	0,79	0,87	0,61
	big	low	high	-4,13	0,83	3,14	0,55	0,67	1,22
	big	high	low	-2,21	3,79	5,83	0,71	0,79	1,02
	big	high	high	-2,46	-1,37	0,73	0,42	0,54	1,78
Horseshoe	big	low	low	-5,43	4,10	5,50	0,72	0,78	1,52
	big	low	high	-4,79	0,44	2,64	0,53	0,65	1,30
	big	high	low	-4,08	1,41	2,60	0,58	0,65	1,81
	big	high	high	-3,24	-2,00	-0,05	0,39	0,50	1,84
Vineyard	small	low	low	3,58	2,59	4,35	0,64	0,73	1,09
	small	low	high	2,00	0,59	2,92	0,53	0,66	1,23
	small	high	low	4,81	0,66	2,20	0,54	0,62	1,60
	small	high	high	3,54	-1,47	0,66	0,42	0,54	1,71
Fan	small	low	low	2,20	1,46	1,98	0,58	0,61	1,38
	small	low	high	1,67	1,38	3,39	0,58	0,69	1,22
	small	high	low	3,98	-0,93	-0,43	0,45	0,48	1,85
	small	high	high	3,32	-0,84	1,01	0,45	0,56	1,75
Shoe Box	small	low	low	1,13	2,37	5,20	0,63	0,77	0,82
	small	low	high	1,74	0,02	2,59	0,50	0,64	1,22
	small	high	low	2,49	0,52	3,06	0,53	0,67	1,18
	small	high	high	3,36	-2,22	0,19	0,38	0,51	1,74
Horseshoe	small	low	low	-0,04	2,15	3,98	0,62	0,71	1,14
	small	low	high	1,03	-0,68	2,14	0,46	0,62	1,25
	small	high	low	1,43	-0,26	1,31	0,48	0,57	1,55
	small	high	high	2,65	-2,74	-0,24	0,35	0,49	1,76

Acoustic measurements of all halls in all test conditions measured at the 1 kHz octave band. G, C50, C80, D50 and D80 in [dB] EDT in [ms]